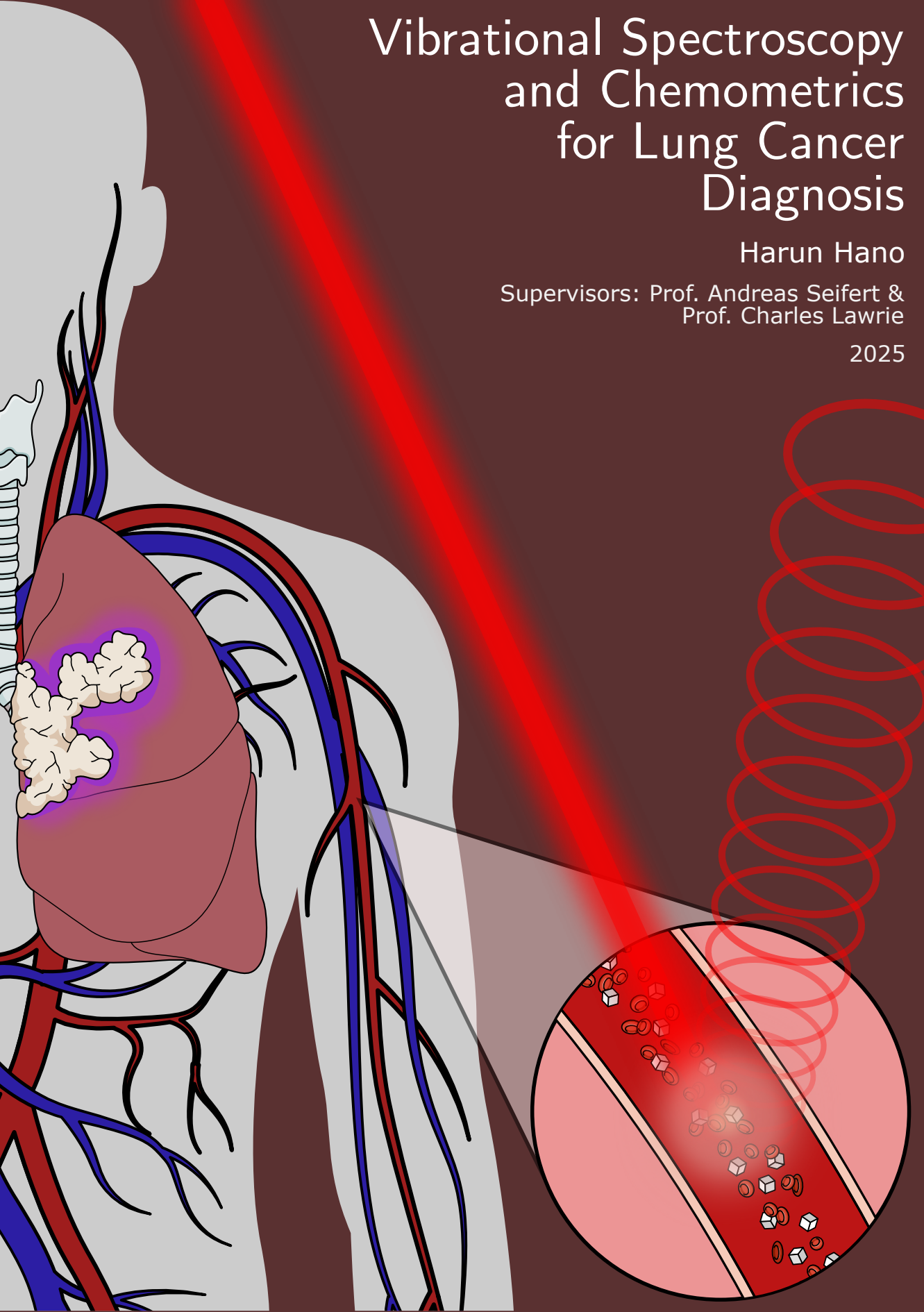
Vibrational Spectroscopy and Chemometrics for Lung Cancer Diagnosis

Harun Hano

Supervisors: Prof. Andreas Seifert &
Prof. Charles Lawrie

2025



eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Vibrational Spectroscopy and Chemometrics for Lung Cancer Diagnosis

Harun Hano

Directed by Prof. Andreas Seifert and Prof. Charles Lawrie

A thesis presented for the degree of Doctor of Philosophy

March 2025

"Nobody's Free Until Everybody's Free"

– Fannie Lou Hamer, **1971**.

In solidarity with all those who continue to face oppression under governmental, systemic, and institutional injustices. May this work stand as a reminder that true freedom is only achieved when it is shared by all.

Acknowledgements

And just like that... this chapter comes to a close.

It has been a long learning process, one for which I am deeply grateful, a journey where I have grown, made mistakes, and learnt from them.

There are so many people to thank, and the list could go on... but above all: My sister, Güler (which means "the one who smiles" in Turkish, and if you knew her, you would think the language itself invented that word just for her). She has cared for me through every high and low, always lifting me up. And my true friend Hülya, from my homeland, who has stood by me through every success, failure, and everything in between. She is truly one of a kind, intellectual, joyful, and irreplaceable.

So many beautiful souls have touched my life with their kindness: The "Katuak eta Diskak" crew (Sofia, Diana, Raquel, Oksana, Ignacio, Mikel, Bao); Suhail, habibi, carrying his Palestinian soul; Lucia, the fiercest Catalan defender, ever-ready with her gin tonic; Natalia, as known as Momo, the most caring and always being inclusive; Mauro, effortlessly chic and the funniest; and many others with hearts of gold. I feel incredibly fortunate to have them in my life, and I could not be more grateful.

Basque Country, this is where my heart belongs. When I first arrived, I struggled to adapt. The city felt foreign, and I wondered if I would ever fit in... But over time, it has become my new home, thanks to the warm embrace of its people. They have made every effort to understand me, even blending their culture with that of a typical foreigner. The secret is patience. Give them time, give yourself time, and soon you will find some of the kindest souls around you.

As we all know, without foreigners, a place can feel dull. It is the blending of cultures that brings true vibrancy. In this spirit, I must also thank those who have added colour to my life: Vangelis, my one and only Greek friend, the kindest soul, and Matias, the ever-passionate Argentinian, forever in love with empanadas.

Lastly, my time with the Nanoengineering team in CIC nanoGUNE has been

an incredible experience. I am grateful to every member, especially Renata, who has always strived to be the best version of herself, Ursula, whose beautiful heart remains with me even though she is no longer here, and Andreas, who has always been calm, constructively critical when needed, and supportive at every step. I also extend my thanks to Charles and his group at Biogipuzkoa Health Research Institute for their collaborative partnership during this journey.

I have learnt so much during this time, and now I close this chapter with the Latin phrase: *Ancora imparo: I am still learning.* Because isn't that the point? We are all just... learning.

Summary

In this thesis, the potential of vibrational spectroscopy, combined with chemometrics, as a rapid diagnostic tool for lung cancer detection is explored. Lung cancer is one of the leading causes of cancer-related deaths worldwide, and early detection is crucial because late-stage diagnosis greatly reduces survival rates. Current screening methods can introduce false positive results, while invasive biopsies come with significant risks. This research addresses these challenges by using Raman and Fourier-transform infrared (FTIR) spectroscopy to identify cancer-specific molecular changes in human blood plasma. The proposed approach, which combines photonic techniques with chemometrics, is investigated to establish a foundation for rapid and accurate in vitro diagnostics, potentially reducing the need for invasive procedures in the future.

The research is built around three interconnected publications. The first publication focuses on improving prediction models to classify spectroscopic data accurately, with finally achieved figures of merit in the order of 90%. The second publication identifies key spectral biomarkers, primarily associated with changes in proteins, lipids, and nucleic acids, through feature selection methods. The third publication combines data from both Raman and FTIR using data fusion techniques, boosting diagnostic accuracy to 99.2%, far outperforming single-method approaches.

This work also outlines data processing steps and feature selection methods, along with rigorous cross-validation procedures. These measures ensure model robustness and allow the findings to be reliably generalized to new patient samples. Moreover, specific spectral features are linked to changes in biochemical and metabolic parameters commonly observed in lung cancer.

In conclusion, this research demonstrates that vibrational spectroscopy, especially when combined with chemometrics, holds strong potential as a rapid and accurate tool for lung cancer screening. These straightforward photonic-based methods, requiring minimal sample preparation yet enabling direct biochemical profiling, further support their practicality for cost-effective and portable clinical deployment. Future studies should prioritize validating this approach in clinical settings, performing larger-scale trials, and developing

portable diagnostic platforms.

Resumen

En esta tesis, se explora el potencial de la espectroscopía vibracional, combinada con quimiometría, como herramienta de diagnóstico rápido para la detección del cáncer de pulmón. El cáncer de pulmón es una de las principales causas de muerte relacionadas con el cáncer a nivel mundial, y su detección temprana es crucial, ya que un diagnóstico en etapas avanzadas reduce drásticamente las tasas de supervivencia. Los métodos actuales de cribado pueden generar falsos positivos, mientras que las biopsias invasivas conllevan riesgos significativos. Esta investigación aborda estos desafíos mediante el uso de espectroscopía Raman y de infrarrojo por transformada de Fourier (FTIR) para identificar cambios moleculares específicos del cáncer en el plasma sanguíneo humano. El enfoque propuesto, que combina técnicas fotónicas con quimiometría, se estudia para sentar las bases de un diagnóstico in vitro rápido y preciso, con el potencial de reducir la necesidad de procedimientos invasivos en el futuro.

La investigación se estructura en torno a tres publicaciones interconectadas. La primera publicación se centra en mejorar los modelos de predicción para clasificar datos espectroscópicos con precisión, logrando finalmente métricas de rendimiento del orden del 90%. La segunda publicación identifica biomarcadores espectrales clave, principalmente asociados a cambios en proteínas, lípidos y ácidos nucleicos, mediante métodos de selección de características. La tercera publicación combina datos de Raman y FTIR utilizando técnicas de fusión de datos, elevando la precisión diagnóstica al 99.2%, superando ampliamente los enfoques que emplean un único método.

Este trabajo también detalla los pasos de procesamiento de datos, los métodos de selección de características y procedimientos rigurosos de validación cruzada. Estas medidas garantizan la robustez de los modelos y permiten generalizar los hallazgos a nuevas muestras de pacientes con fiabilidad. Además, se vinculan características espectrales específicas con alteraciones en parámetros bioquímicos y metabólicos comúnmente observados en el cáncer de pulmón.

En conclusión, esta investigación demuestra que la espectroscopía vibracional, especialmente al combinarse con quimiometría, posee un gran potencial como herramienta rápida y precisa para el cribado del cáncer de pulmón. Es-

tos métodos fotónicos, sencillos y que requieren una preparación mínima de la muestra —permitiendo un perfilado bioquímico directo— respaldan su practicidad para aplicaciones clínicas portátiles y rentables. Futuros estudios deberían priorizar la validación clínica de este enfoque, la realización de ensayos a mayor escala y el desarrollo de plataformas de diagnóstico portátiles.

Özet

Bu tezde, titreşim spektroskopisinin kemometri ile birleştirilerek akciğer kanseri tespiti için hızlı bir tanı aracı olarak potansiyeli araştırılmıştır. Akciğer kanseri, dünya çapında kansere bağlı ölümlerin önde gelen nedenlerinden biridir ve geç evre teşhis hayatta kalma oranlarını önemli ölçüde düşürdüğü için erken teşhis kritik önem taşımaktadır. Mevcut tarama yöntemleri yanlış pozitif sonuçlara yol açabilirken, invaziv biyopsiler önemli riskler barındırır. Bu çalışma, insan kan plazmasındaki kansere özgü moleküler değişiklikleri belirlemek için Raman ve Fourier dönüşümlü kızılötesi (FTIR) spektroskopisi kullanarak bu zorlukları ele almaktadır. İnvaziv prosedürlere olan ihtiyacı azaltma potansiyeli taşıyan, hızlı ve doğru in vitro tanı için bir temel oluşturmayı amaçlayan bu yaklaşımda, fotonik teknikler ile kemometri yöntemlerinin kombinasyonu incelenmiştir.

Araştırma, birbiriyle bağlantılı üç yayın üzerine kuruludur. İlk yayın, spektroskopik verileri doğru şekilde sınıflandırmak için tahmin modellerinin iyileştirilmesine odaklanmış ve nihai olarak %90 seviyesinde başarı ölçütleri elde edilmiştir. İkinci yayın, özellikle proteinler, lipitler ve nükleik asitlerdeki değişikliklerle ilişkili temel spektral biyobelirteçleri özellik seçim yöntemleriyle tanımlamıştır. Üçüncü yayın ise Raman ve FTIR verilerini veri füzyon teknikleriyle birleştirerek tanısal doğruluğu %99,2'ye yükseltmiş ve tek yöntemli yaklaşımları açık ara geride bırakmıştır.

Bu çalışmada ayrıca veri işleme adımları, özellik seçim yöntemleri ve titiz çapraz doğrulama prosedürleri detaylandırılmıştır. Bu adımlar, model sağlamlığını garanti altına almakta ve bulguların yeni hasta örneklerine güvenilir şekilde genellenebilmesini sağlamaktadır. Ayrıca, belirli spektral özelliklerin akciğer kanserinde sıklıkla gözlemlenen biyokimyasal ve metabolik parametre değişiklikleriyle ilişkilendirilmesi sağlanmıştır.

Sonuç olarak, bu araştırma titreşim spektroskopisinin (özellikle kemometri ile birleştirildiğinde) akciğer kanseri taramasında hızlı ve doğru bir araç olarak güçlü bir potansiyele sahip olduğunu göstermektedir. Minimal numune hazırlığı gerektiren ancak doğrudan biyokimyasal profil oluşturmayı mümkün kılan bu basit fotonik temelli yöntemler, düşük maliyetli ve taşınabilir klinik uygula-

malar için pratikliği desteklemektedir. Gelecek çalışmalar, bu yaklaşımın klinik ortamlarda doğrulanmasına, geniş ölçekli denemeler yapılmasına ve taşınabilir tanı platformlarının geliştirilmesine öncelik vermelidir.

Dissemination of results

Peer-reviewed Publications as First Author

- Harun Hano, Beatriz Suarez, Jose Manuel Amigo, Charles H. Lawrie, and Andreas Seifert. Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy. *Microchemical Journal* 209, 112496, (2025).
- Harun Hano, Beatriz Suarez, Charles H. Lawrie, Andreas Seifert. Fusion of Raman and FTIR Spectroscopy Data Uncovers Physiological Changes Associated with Lung Cancer. *International Journal of Molecular Sciences*, 25(20):10936, (2024).
- Harun Hano, Beatriz Suarez, Charles H. Lawrie, Andreas Seifert. Raman Spectroscopy Detects Biochemical Signatures in Non-Small Cell Lung Cancer. *IEEE Xplore*, 1-2, (2024).
- Harun Hano, Charles H. Lawrie, Beatriz Suarez, Alfredo Paredes Lario, Ibone Elejoste Echeverría, Jenifer Gómez Mediavilla, Marina Izaskun Crespo Cruz, Eneko Lopez, Andreas Seifert. Power of Light: Raman Spectroscopy and Machine Learning for the Detection of Lung Cancer. *ACS Omega* 9, 12:14084–14091, (2024).

Further Publications

- Kristina Ashurbekova, Evgeny Modin, Harun Hano, Karina Ashurbekova, Iva Saric Jankovic, Robert Peter, Mladen Petravić, Andrey Chuvilin, Aziz Abdulagatov, and Mato Knez. *In Situ* Investigation of Thermally Induced Surface Graphenization of Polymer-Derived Ceramic (PDC) Coatings from Molecular Layer (MLD) Deposited Silicon-Based Preceramic Thin Films. *Chemistry of Materials* 35 (19), 8092-8100, (2023).
- Fethullah Güneş, Ahmet Aykaç, Mustafa Erol, Çağlar Erdem, Harun Hano, Begüm Uzunbayır, Mustafa Şen, Arzum Erdem. Synthesis of hierarchical hetero-composite of graphene foam/α-Fe_2O_3 nanowires

and its application on glucose biosensors. *Journal of Alloys and Compounds* 895, 162688, (2022).

Conferences

- Harun Hano, Beatriz Suarez, Charles H. Lawrie, Andreas Seifert. Raman Spectroscopy Detects Biochemical Signatures in Non-Small Cell Lung Cancer. *Optical MEMS and Nanophotonics*, Spain, (2024).
- Harun Hano, Beatriz Suarez, Charles H. Lawrie, Andreas Seifert. Non-Invasive Lung Cancer Detection by Multivariate Analysis of Vibrational Spectroscopy Data. *19th European Technology Platform on Nanomedicine (ETPN)*, Italy, (2024).
- Harun Hano, Andreas Seifert. Systematic Assessment of Feature Selection Methods with PLS-DA Model for Photonic In Vitro Detection of Lung Cancer. *21th Euroanalysis*, Switzerland, (2023).
- Harun Hano, Andreas Seifert. Comparison of Classification Models for Lung Cancer Detection Using Raman Spectroscopy. *11th Colloquium Chemiometricum Mediterraneum (CCM XI)*, Italy, (2023).
- Harun Hano, Ahmet Aykaç. Development of Glucose Biosensor from Hierarchical Graphene/α-Fe_2O_3 Nanocomposites. *3rd International Students Science Congress*, Türkiye, (2019).

Glossary

Alb Albumin.

ALP Alkaline Phosphatase.

ALT Alanine Aminotransferase.

AST Aspartate Aminotransferase.

ATR Attenuated Total Reflection.

AUC Area Under the Curve.

CCD Charge-Coupled Device.

Cr Creatinine.

EDTA Ethylenediaminetetraacetic Acid.

EMSC Extended Multiplicative Signal Correction.

FP Fingerprint.

FR Feature Reduction.

FS Feature Selection.

FTIR Fourier-Transform Infrared.

GFR Glomerular Filtration Rate.

GGT Gamma-Glutamyl Transferase.

Hb Hemoglobin.

HLDF High-Level Data Fusion.

LDA Linear Discriminant Analysis.

LDCT Low-Dose Computed Tomography.

LDH Lactate Dehydrogenase.

LLDF Low-Level Data Fusion.

LR Logistic Regression.

LVs Latent Variables.

LWs Loading Weights.

M Metastasis.

MCH Mean Corpuscular Hemoglobin.

MCHC Mean Corpuscular Hemoglobin Concentration.

MCV Mean Corpuscular Volume.

MLDF Mid-Level Data Fusion.

MSC Multiplicative Scatter Correction.

MWU Mann-Whitney U.

N Nodal Involvement.

NB Naive Bayes.

NCDs Non-Communicable Diseases.

PCA Principal Component Analysis.

PCs Principal Components.

PLS Partial Least Squares.

PLS-DA Partial Least Squares Discriminant Analysis.

RBCs Red Blood Cells.

RCs Regression Coefficients.

RDW Red Cell Distribution Width.

RF Random Forest.

RFE Recursive Feature Elimination.

ROC Receiver Operating Characteristic.

SHAP SHapley Additive exPlanations.

SNV Standard Normal Variate.

SVM Support Vector Machine.

T Tumor Stage.

VIP Variable Importance in Projection.

WBCs White Blood Cells.

Contents

Acknowledgements III

Summary V

Resumen VII

Özet IX

Dissemination of results XI

Glossary XV

1 Overview **1**

1.1 Introduction 1

1.2 Theoretical background 4

1.2.1 Biofluids 4

1.2.2 Vibrational spectroscopy 5

1.2.3 Data analysis 10

1.3 Materials and methods 25

1.3.1 Sample collection and preparation 25

1.3.2 Data collection and preprocessing 26

1.3.3 Data analysis strategies 27
1.4 Results and discussions 31
1.4.1 Publication 1: Supervised learning algorithms 31
1.4.2 Publication 2: Feature selection methods 42
1.4.3 Publication 3: Data fusion strategies 52
1.4.4 Vibrational frequencies and bands assignment 61
1.4.5 Model validation with newly registered patients 65
1.4.6 Biochemical analysis 70
1.5 References 79

2 Conclusions **89**

A Publications as a first author **93**

A.1 Supervised learning algorithms 94
A.2 Feature selection methods 103
A.3 Data fusion strategies 120

Chapter 1

Overview

1.1 Introduction

Global mortality patterns have shifted considerably in recent decades, largely due to advances in healthcare and improved living standards. Non-communicable diseases (NCDs), particularly cardiovascular disease and cancer, now rank as the predominant causes of death worldwide. Ischemic heart disease remains the leading cause of mortality globally. It accounts for 13% of all deaths and claimed over nine million lives in 2021 [1, 2]. However, cancer-related fatalities are rising at a concerning rate. Lung cancer has become the leading cause of cancer deaths worldwide, underscoring the growing impact of NCDs. As illustrated in Figure 1.1, in 2020 alone, around 2 million deaths were reported, accounting for nearly 25% of all cancer-related fatalities [3]. Despite medical improvements, the overall five-year survival rate for lung cancer stands at only 18.1%, which is significantly lower than that for breast, prostate, and colorectal cancers [4]. Later-stage diagnoses worsen the prognosis. Stage 1 cases have a five-year survival rate of up to 67%, but it plummets to a mere 23% for stage 3 [5].

Smoking remains the principal risk factor. Other contributors include exposure to secondhand smoke, environmental pollutants, and occupational car-

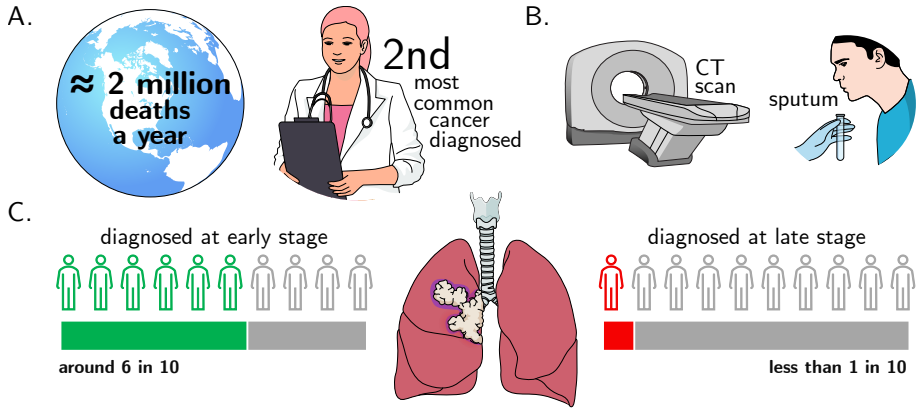


Figure 1.1: Lung cancer management: (A) disease prevalence, (B) conventional diagnostic methods, (C) survival rates with early detection.

cinogens, such as asbestos (a fibrous mineral linked to respiratory illnesses). These factors highlight the importance of timely detection. Early diagnosis offers more effective treatment options, including surgery, radiation therapy, and targeted drug therapies. It can also lower healthcare costs by reducing the need for aggressive interventions and prolonged hospital stays [1, 3].

Conventional diagnostic methods, although improved in past decades, still face notable limitations. Low-dose computed tomography (LDCT) screening is recommended for high-risk individuals and can detect early-stage lung cancer. But it often produces high false-positive rates, which can lead to unnecessary invasive procedures, overdiagnosis, incidental findings, and additional radiation exposure. Chest X-rays, though widely accessible, sometimes fail due to low resolution. Bronchoscopy and needle biopsies provide accurate diagnoses but are invasive and carry procedural risks. Moreover, these traditional techniques rely on interpretation by radiologists, introducing the possibility of human error and inter-observer variability [6, 7].

Chemometrics offers a promising way to address these limitations. It uses algorithms to detect subtle differences that might elude human observers, thus improving diagnostic accuracy. These algorithms can process large, complex datasets—such as medical images, genetic information, and clinical records—to

reveal patterns and correlations that conventional approaches might miss. Blood samples further enhance diagnostic capabilities. They contain a vital source of diagnostic information, such as proteins, lipids, carbohydrates, DNA, and RNA, all of which can serve as potential biomarkers. However, standardization remains a challenge, and more research is needed to improve the predictive power of blood-based tests [8].

Vibrational spectroscopy has emerged as powerful tool in medical diagnosis, especially Raman and Fourier-Transform Infrared (FTIR) spectroscopy, serving as molecular fingerprints for detecting subtle biochemical changes associated with lung cancer. Raman spectroscopy relies on inelastic light scattering to reveal molecular structures, while FTIR spectroscopy depends on the absorption of infrared radiation. Both methods enable non-destructive, label-free analysis of biological specimens, generating detailed spectral datasets. When integrated with chemometrics, these techniques can help identify lung cancer–specific molecular signatures in blood samples, potentially at earlier stages than conventional diagnostic methods allow [6, 8, 9]. Therefore, the primary aim of this thesis is to distinguish lung cancer from healthy controls by integrating vibrational spectroscopy with chemometrics. Raman and FTIR spectroscopy are applied to analyze human blood plasma samples, and they reveal subtle molecular signatures of lung cancer that current diagnostic approaches might overlook. Early detection is the ultimate goal, as it can improve patient outcomes.

This thesis is guided by several interrelated objectives, each supported by specific hypotheses. The first objective is to optimize machine learning models for analyzing spectroscopic data. It includes a systematic evaluation and refinement of supervised learning algorithms. These models incorporate a dimensionality reduction technique and feature selection method to enhance classification performance. The second objective is to identify potential spectral biomarkers that might indicate lung cancer through feature selection methods. It focuses on recognizing subtle vibrational changes in blood plasma, reflecting alterations in proteins, lipids, or nucleic acids linked to the disease. The third objective aims to increase diagnostic capability by combining data from Raman and FTIR spectroscopy through data fusion methods at low, mid, and high levels. This

multi-modal approach seeks to surpass the diagnostic performance of any single spectroscopic technique alone.

Finally, this work explores possible links between identified spectral features and the underlying biological changes and suggests that vibrational spectroscopy can classify human blood plasma samples, in near real time, as indicative of either lung cancer or healthy state. This rapid, real-time assessment may be adaptable to clinical environments in the future.

1.2 Theoretical background

1.2.1 Biofluids

Biofluids play vital roles in human physiology and possess significant diagnostic value. The body produces several biofluids, each characterized by unique compositions and clinical insights. These include blood, urine, saliva, cerebrospinal fluid, sweat, tears, seminal fluid, bile, peritoneal fluid, and synovial fluid as illustrated in Figure 1.2. Biofluid-based testing is particularly advantageous when frequent analyses are required or when tissue biopsies pose challenges. It requires only small sample volumes—often just a few microliters—which are more patient-friendly and accessible. Among all biofluids, blood remains the gold standard, as it circulates through the entire body and carries biomarkers from various tissues and organs. Its collection, handling, and analysis also follow widely established clinical protocols [10].

Biofluids hold particular relevance in addressing global health issues linked to NCDs, including diabetes, cardiovascular disease, chronic lung disease, and cancer. Conventional diagnostic methods often involve lengthy and expensive procedures, such as in vivo or in vitro testing, pathological assessments, or microbiological analyses. In response, spectroscopic techniques—such as vibrational, fluorescence, and nuclear magnetic resonance—have emerged as promising tools for fast, non-destructive characterization of body fluids [11]. These methods facilitate more efficient and patient-friendly diagnostic approaches and may support earlier detection and improved disease management.

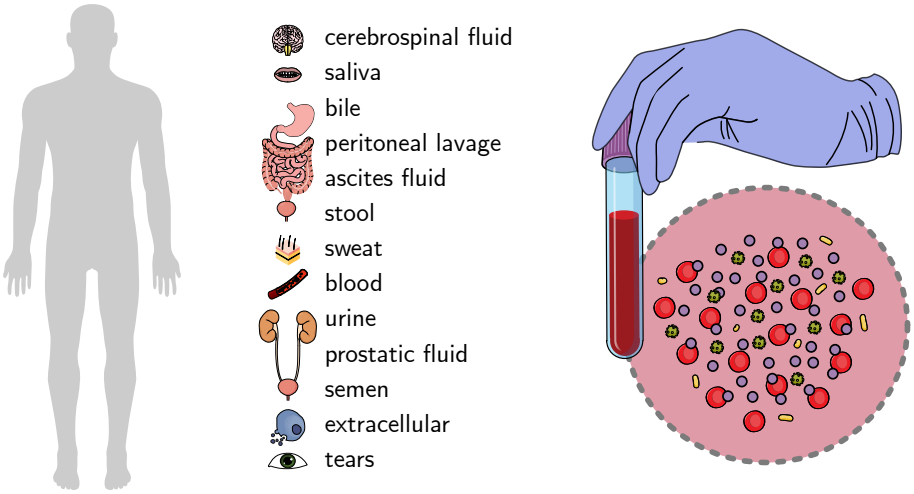


Figure 1.2: Schematic representation of various biofluids in the human body.

1.2.2 Vibrational spectroscopy

Vibrational spectroscopy is an analytical method that probes the energetic transitions between quantized vibrational states in molecules. It examines how electromagnetic radiation interacts with molecular bonds, focusing on changes in dipole moments or polarizabilities caused by molecular vibrations or the movement of atomic groups. These changes, induced by the absorption or scattering of electromagnetic radiation, provide valuable information about a sample's molecular structure and composition [12].

By identifying characteristic "fingerprints" of molecular groups, vibrational spectroscopy can detect proteins, lipids, nucleic acids, and carbohydrates, as well as unknown compounds (see Figure 1.3). This technique shows particular promise in medical diagnostics. First, it can detect biochemical changes before they become visible through conventional histopathology. Second, it offers automated, objective sample classification, which can augment or partially automate traditional methods [13].

Raman and FTIR spectroscopy are frequently employed vibrational spec-

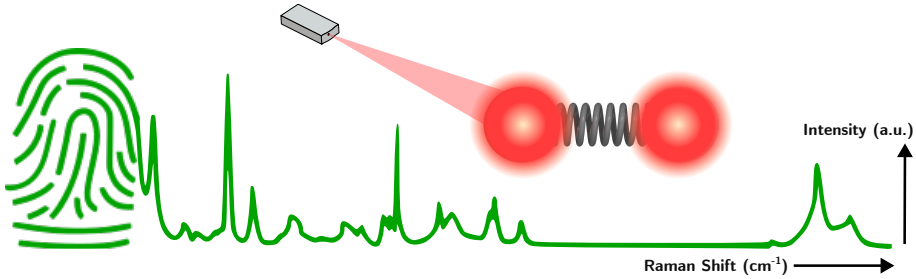


Figure 1.3: Representative Raman spectrum of a biological sample displaying characteristic biomolecular fingerprint peaks.

troscopic techniques in biomedical research. Raman spectroscopy detects the inelastic scattering of monochromatic light, while FTIR measures the absorption of infrared radiation by molecules. Both provide complementary insights into molecular vibrations and structures, with several key advantages for clinical applications [13]:

1. Less invasive: These methods often require only small sample volumes and can be performed non-destructively.
2. Rapid: Spectra can be acquired quickly, enabling potential real-time diagnostics.
3. Objectivity: They provide an objective measure of sample composition, reducing inter-observer variability.
4. High information content: The resulting data capture subtle biochemical changes associated with disease.
5. Versatile: They can be applied to various biological samples, including tissues, body fluids, and cell cultures.

Raman spectroscopy

Raman spectroscopy, discovered by C.V. Raman in 1928, a non-destructive analytical technique, is based on the inelastic scattering of monochromatic light.

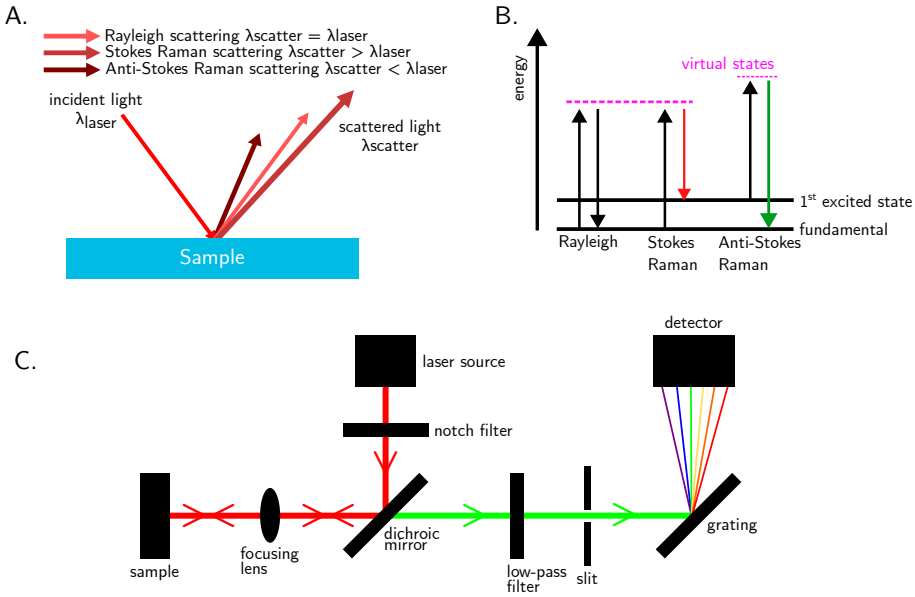


Figure 1.4: Principles and instrumentation of Raman spectroscopy: (A) schematic identifying light scattering after laser exposure on a sample surface, (B) energy level diagram for Rayleigh, Stokes Raman, and anti-Stokes Raman scattering, (C) typical setup for a dispersive Raman spectrometer.

When incident light interacts with molecules, a small fraction (approximately one in a million) undergo inelastically scattered photons with shifted frequencies corresponding to specific molecular vibrations (see Figure 1.4A). Raman effect occurs due to the interaction between incident photons and the electron cloud of the molecule. This interaction induces a temporary distortion of the electron cloud, followed by the reemission of a photon. If the molecule returns to a different vibrational state than its initial state, the emitted photon will have a different energy from the incident photon. This energy difference, measured as a shift in frequency, is characteristic of specific molecular bonds and structures (see Figure 1.4B) [14].

This label-free method requires minimal sample preparation and can be performed in real time, making it well suited for biofluid analysis. In the case of human blood plasma, a rich source of potential cancer biomarkers, Raman

spectroscopy shows promise for early lung cancer detection, tumor margin assessment, and treatment monitoring [15]. Each molecule exhibits a distinct set of vibrational modes, collectively referred to as its "molecular fingerprint," which allows the technique to identify and quantify chemical species without labeling.

Modern Raman spectrometers operate through a precise sequence of optical components to analyze material composition (see Figure 1.4C). The process starts with a monochromatic light source, passing through a filter to ensure only the desired excitation wavelength reaches the sample, where it interacts with molecules, generating scattered light. While Rayleigh scattering is filtered out by a notch filter, Raman scattering carries molecular vibrational information. Raman-scattered light is collected and directed through a slit to control spatial resolution before being dispersed by a grating into its component wavelengths. A detector, typically a (CCD), captures the resulting spectrum, providing a molecular fingerprint for material identification and characterization. [16, 17].

Fourier-transform infrared spectroscopy

FTIR spectroscopy is an another analytical technique in chemical analysis and biomedical diagnostics. This method is based on the principle that molecules absorb specific frequencies of infrared light that are characteristic of their chemical structure. When a sample is exposed to infrared radiation, molecules absorb the energy at frequencies that match their vibrational modes. These vibrations can involve changes in bond length (stretching) or bond angle (bending). The resulting spectrum represents a "fingerprint" of an unknown sample, with absorption peaks corresponding to the frequencies of vibrations between the bonds in the molecule. Electromagnetic spectrum in FTIR is typically divided into three main infrared regions: Near-IR (13000-4000 cm^{-1}), Mid-IR (4000-400 cm^{-1}), and Far-IR ($<$400 cm^{-1}). Mid-IR region is of most interest as it contains the characteristic bands for identification of functional groups [18, 19].

FTIR spectrometer works with an interferometer to modulate infrared light (see Figure 1.5). The process begins with an IR source emitting a broad spectrum of infrared radiation. This light is directed to a beamsplitter, which di-

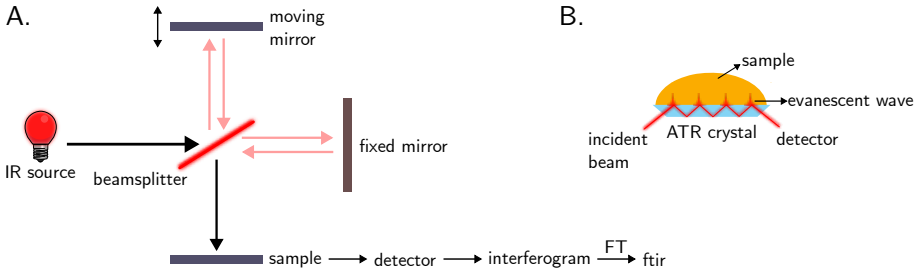


Figure 1.5: (A) Optical diagram of a Michelson interferometer in FTIR, and (B) attenuated total reflection (ATR) FTIR setup.

vides the beam into two paths. One beam is directed to a fixed mirror, while the other is directed to a moving mirror. The moving mirror introduces a variable path difference between the two beams. When these beams recombine at the beamsplitter, they create an interference pattern. This recombined beam then passes through the sample and reaches the detector. The detector records the intensity of the beam as a function of the mirror position, producing an interferogram. The resulting interferogram is then converted into a conventional IR spectrum through Fourier transform [20].

FTIR spectroscopy provides extensive biochemical information by detecting and quantifying proteins, lipids, nucleic acids, and carbohydrates. Each type of molecule produces characteristic absorption bands. For instance, proteins commonly exhibit strong amide I (1700–1600 cm^{-1}) and amide II (1600–1500 cm^{-1}) peaks, corresponding to C=O stretching and N–H bending, respectively. Lipids typically show C–H stretching bands near 2900 cm^{-1}, and nucleic acids display phosphate absorptions around 1080 cm^{-1} [21]. This capacity for rapid, non-destructive, and comprehensive analysis makes FTIR a compelling complement to Raman spectroscopy in early and accurate diagnosis.

1.2.3 Data analysis

Data preprocessing

Data preprocessing is a critical step in spectroscopic analysis for diagnostic applications, significantly influencing the quality and interpretability of the results, as indicated in Figure 1.6. Its principal aim is to remove or minimize signals unrelated to the analyte or target property, thereby enhancing subsequent analyses. However, the choice of preprocessing methods requires care to avoid introducing artifacts or obscuring relevant features.

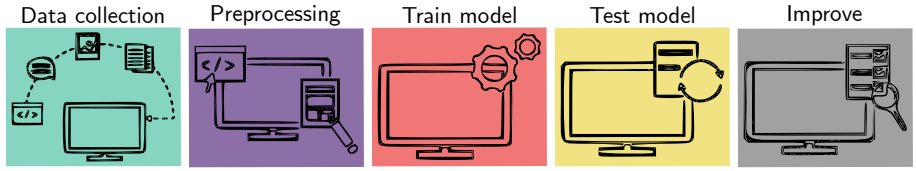


Figure 1.6: Iterative workflow of data processing using chemometrics for analytical chemistry based on photonic data applied to medical diagnostics.

Smoothing, often achieved using Savitzky-Golay filters, reduces noise while preserving key spectral features. Light-scattering corrections—such as Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and Extended Multiplicative Signal Correction (EMSC)—help normalize spectra to a common scale by mitigating multiplicative and additive effects (see Figure 1.7). Baseline corrections further enhance spectral quality by removing background signals; methods include rubber-band, polynomial fitting, and iterative techniques like asymmetric or automatic-weighted least squares [22].

Spectral differentiation, typically in the form of first or second derivatives, enhances subtle features and removes baseline shifts or offsets [23]. Normalization approaches, such as Amide I normalization (for protein analysis) or vector normalization, address intensity variations by scaling spectra to a consistent reference [22]. Meanwhile, scaling methods—like standardization or autoscaling—ensure equal weighting of all variables, typically by centering on the mean and dividing by the standard deviation [24]. Outlier detection, using jack-knife or Z-score methods, identifies anomalous spectra that might otherwise skew

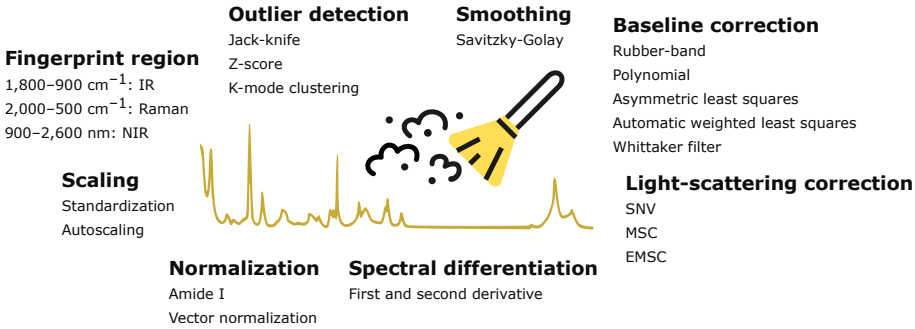


Figure 1.7: Overview of commonly used preprocessing techniques for spectral data. SNV: standard normal variate, MSC: multiplicative scatter correction, EMSC: extended multiplicative signal correction, IR: infrared, NIR: near infrared.

results. Finally, K-mode clustering (a variant of K-means for categorical data) can group spectra with similar characteristics, potentially revealing underlying chemical or structural patterns in complex datasets [25].

The order in which preprocessing steps are applied is critical and can significantly change model performance. Generally, the optimal sequence can vary depending on the specific characteristics of the dataset and the analytical goals. Data preprocessing often involves experimenting with various combinations to find the most effective approach for a given application, and even these preprocessing methods may depend on or be developed for specific statistical distributions of the data.

Unsupervised learning

Unsupervised learning techniques are essential for extracting patterns from high-dimensional spectral data without predefined labels. They help uncover hidden structures and reduce data complexity while retaining key spectral features. Common models such as Principal Component Analysis (PCA), k-means clustering etc. are frequently used in spectral analysis to group similar data points and reduce dimensionality. These models do not require labeled data, making them well-suited for exploratory analysis [26].

PCA is one of fundamental unsupervised techniques in spectral analysis.

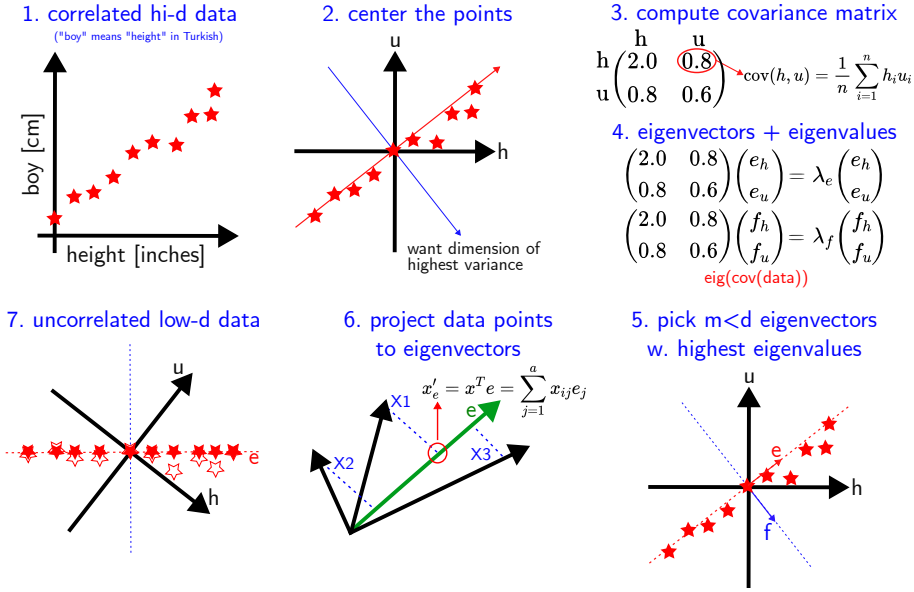


Figure 1.8: Schematic illustration of PCA process for dimensionality reduction in spectral data analysis.

1

It transforms spectral data into a new coordinate system defined by orthogonal vectors called principal components (PCs) (see Figure 1.8). These components are arranged by the amount of variance they explain. PCA begins by centering the data and calculating the covariance matrix. Eigendecomposition of this matrix yields eigenvectors and eigenvalues, which form the basis of the new coordinate system. By selecting the top eigenvectors, PCA creates a lower-dimensional representation of the data that retains the most critical information. Typically, the first few principal components capture the majority of meaningful spectral variation. This transformation enables dimensionality reduction while preserving significant variations in the data [26]. In general, PCA is useful for identifying dominant spectral features, distinguishing overlapping peaks, and revealing chemical or structural similarities among samples.

Supervised learning

Supervised learning is an essential machine learning approach that operates on prelabeled datasets, unlike unsupervised learning, which infers patterns from unlabeled data. These methods train algorithms to predict or classify features based on known outcomes. The objective is to develop a function f that maps input spectral data X to output labels or values Y, expressed as $Y = f(X)$. Here, X represents a matrix of spectral intensities at various wavelengths or wavenumbers, while Y represents the target variable, such as analyte concentration (in regression tasks) or sample class (in classification tasks). During the training phase, the algorithm learns to associate patterns in the labeled spectral data, thereby enabling predictions on new, unseen samples [27].

A key challenge in supervised learning arises from the high dimensionality of spectral data, which often exceeds the number of samples. This high dimensionality can introduce multicollinearity among variables [28]. Effective models balance the need to fit the training data against the requirement to generalize well to new data, often relying on dimensionality reduction or regularization to mitigate overfitting. The choice of learning algorithm depends on whether Y is continuous or categorical, as well as on the complexity of the relationship between spectral features and the target variable. Classifiers are fundamental to supervised learning for the categorization of samples based on their spectral features. These algorithms learn decision boundaries in the high-dimensional space, allowing for the prediction of sample classes.

1. Linear Discriminant Analysis (LDA)

   LDA is a dimensionality reduction technique that finds a linear combination of features to separate two or more classes. Unlike PCA, which focuses on maximizing variance, LDA aims to maximize the separability between classes. It does this by maximizing the ratio of between-class variance to within-class variance to ensure better class separability [29]. This approach makes LDA particularly effective for classification tasks, especially when dealing with high-dimensional data and overlapping classes. LDA projects the data onto a lower-dimensional space while preserving class discrimination, which can improve classification accuracy and reduce

computational complexity.

The algorithm operates by computing two scatter matrices: the between-class scatter matrix $\mathbf{S}_B$ and the within-class scatter matrix $\mathbf{S}_W$, which are defined as

$$\mathbf{S}_B = \sum_k N_k(\mu_k - \bar{x})(\mu_k - \bar{x})^T \tag{1.1}$$

$$\mathbf{S}_W = \sum_k \sum_{i \in k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{1.2}$$

where N_k is the sample size of class k, μ_k is the sample mean of class k, $\bar{x}$ is the global sample mean vector of all data (overall mean), x_i is a data point in class k. Known as Fisher's criterion, LDA can be computed by eigendecomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$, with the eigenvectors corresponding to the largest eigenvalues forming the columns of $\mathbf{W}$. Alternatively, LDA can also be formulated using a least squares approach, though this method does not directly optimize Fisher's criterion.[29].

2. Support Vector Machine (SVM)

SVM addresses the fundamental task of classifying data points into distinct classes. In its simplest form, given a set of data points belonging to two classes, SVM aims to find the optimal hyperplane that separates these classes. A hyperplane divides the original d-dimensional space into two half-spaces. A dataset is described as linearly separable if each half-space contains points from only one class [27]. When the dataset is linearly separable, it is possible to define all the points with label $y_i = -1$ if $h(x_i) \leq -1$, and $y_i = 1$ if $h(x_i) \geq 1$ [30]. The function $h(x)$ acts as a linear classifier (or linear discriminant) that assigns a class label y to any point x according to

$$y = \begin{cases} 1, & \text{if } h(x) \geq 1, \\ -1, & \text{if } h(x) \leq -1 \end{cases}$$

Hyperplane function is defined as follows

$$h(x) = w^T x + b \tag{1.3}$$

where w is the normal vector to the hyperplane and b is the offset or distance of the hyperplane. If any arbitrary point that lies on the hyperplane, then

$$h(x) = w^T x + b = 0 \tag{1.4}$$

While the directed distance of a point to the hyperplane is calculated as

$$r = \frac{w^T x_i + b}{||w||} \tag{1.5}$$

the margin, which is the total distance from two support vectors of different classes to the hyperplane, is calculated as

$$\gamma = \frac{2}{\|w\|}$$

3. Naive Bayes (NB)

   NB is a probabilistic classification algorithm rooted in Bayes' theorem. It calculates posterior probabilities for each class given the input features, then assigns each data point to the class with the highest posterior probability. Formally, NB seeks to determine the most likely class for a given spectral data point under the assumption that the features (e.g., intensities at different wavelengths) are conditionally independent given the class label. Although this assumption is often unrealistic in practice—particularly for spectral data, where correlated wavelengths are common—NB frequently performs well in many applications due to its simplicity and robustness. Its computational efficiency stems from straightforward probability calculations and non-iterative nature [31].

   Bayes' theorem underpins NB, expressed as [30]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.6}$$

   where

- A is the event (or class) of interest
- B is the observed data or evidence
- $P(A|B)$ is the posterior probability: the likelihood of event A occurring given that B is true
- $P(B|A)$ is the likelihood: the probability of observing B given that A is true
- $P(A)$ is the prior probability of A occurring
- $P(B)$ is the marginal likelihood (or evidence), representing the total probability of observing B across all possible events

The "naive" aspect comes from the assumption of feature independence. This assumes that the intensity at one wavelength does not affect the intensity at another wavelength. Normally, this assumption does not hold for spectral data, as intensities at different wavelengths are often correlated. Peaks are tied to specific molecular vibrations, meaning that dependencies between features (wavelengths) are critical for accurate classification. However, this classifier was still applied despite these theoretical limitations, and it performed adequately in this context. While the independence assumption might suggest that the model is not feasible for spectral data, several factors can explain why it worked. NB is known for being robust, even when the independence assumption is violated [32]. Additionally, feature selection (FS) and preprocessing techniques, like dimensionality reduction (e.g., PCA), can help reduce feature interdependencies. Moreover, simplicity and computational efficiency made NB a practical choice for a quick and interpretable classification.

4. Logistic Regression (LR)

   LR is a classification algorithm used to assign observations to discrete classes. LR calculates the probability that an input spectrum belongs to a particular class by applying the logistic (sigmoid) function [27]. The hypothesis is defined as:

$$h_\theta(x) = g(\theta^T x)$$

where g is the sigmoid function, which is defined as below:

$$g(z) = \frac{1}{1 + e^{-z}}$$

where $g(z)$ is the output between 0 and 1 (probability estimate), z is the input to the function (the algorithm's prediction, e.g. $mx + b$), and e is the base of the natural logarithm.

The class label is predicted using the rule:

$$y = \begin{cases} 1, & \text{if } h_\theta(x) \geq 0.5 \\ 0, & \text{if } h_\theta(x) < 0.5 \end{cases}$$

5. Random Forest (RF)

   RF is an ensemble learning method that has found significant application in spectral analysis. Spectroscopic measurements often result in datasets with numerous features, corresponding to different wavelengths or frequencies. RF navigates this high-dimensionality by considering random subsets of features at each decision point, thereby reducing overfitting and improving generalization. Moreover, it demonstrates robustness to noise and outliers, which are common challenges in spectral measurements due to instrument variations or sample heterogeneity.

   The working mechanism involves an ensemble of decision trees, each trained on a random subset of the spectral data. For each tree, the algorithm randomly selects a subset of spectral features at each node and determines the best split based on these features. This process continues until stopping conditions are met, such as reaching a maximum tree depth, failing to satisfy minimum sample requirements for splitting, or failing to meet the minimum samples required in a leaf node. Trees are typically grown to full depth by default unless restricted by these hyperparameters. RF captures various aspects of spectral data by creating multiple trees in this manner. To classify a new spectrum, each decision tree in the forest makes a prediction. Final classification is determined by a majority vote among all trees. This ensemble approach allows RF to effectively handle non-linear relationships in spectral data [30].

Each tree $f_i(x)$ outputs a class label from the set $\mathcal{C}$. Mathematically, RF model can be expressed as:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i=1}^{n} 1\big(f_i(x) = c\big) \tag{1.7}$$

where x denotes the input, $f_i(x)$ is the prediction of the i-th tree, and $1(\cdot)$ is an indicator function counting votes for class c. The final prediction $\hat{y}$ is the class with the highest vote among all n trees [33, 34].

6. Partial Least Squares (PLS)

   PLS is a multivariate linear statistical analysis method designed to model the relationship between two data sets. In classification tasks, categorical variables are first recoded into continuous dummy variables. Consider a binary classification problem where the outcome variable $\mathbf{Y}$ is recoded as 0 or 1. Given a predictor matrix $\mathbf{X} \in \mathbb{R}^{N \times J}$, where N is the number of samples and J is the number of predictor variables, the objective of PLS is to find latent variables that maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$ [35, 36].

   PLS decomposes $\mathbf{X}$ and $\mathbf{Y}$ into the following forms:

   $$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \quad \mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}$$

   where: $\mathbf{T}$ and $\mathbf{U}$ are scores, $\mathbf{P}$ and $\mathbf{Q}$ are loadings, $\mathbf{E}$ and $\mathbf{F}$ are residuals. PLS algorithm proceeds iteratively to find latent variables as follows:

   (a) **Weights (w)**
   Weight vector $\mathbf{w}$ captures directions of maximum covariance between $\mathbf{X}$ and $\mathbf{Y}$:
   $$\mathbf{w} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}}$$
   Normalization is applied to ensure $\|\mathbf{w}\| = 1$.

   (b) **Scores (t and u)**
   X- and Y-space score vectors $\mathbf{t}$ and $\mathbf{u}$ are then given as:

   $$\mathbf{t} = \mathbf{X}\mathbf{w} \quad \mathbf{u} = \mathbf{Y}\mathbf{c}$$

(c) **Loadings (p and q)**
Loadings are computed as:

$$\mathbf{p} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{t}}{\mathbf{t}^{\mathrm{T}}\mathbf{t}}, \quad \mathbf{q} = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}}$$

(d) **Regression Coefficients (b)**
Regression coefficients are computed as:

$$\mathbf{b} = \mathbf{w}\left(\mathbf{p}^{\mathrm{T}}\mathbf{w}\right)^{-1}\mathbf{q}^{\mathrm{T}}$$

(e) **Prediction**
Predicted values for a new dataset $\mathbf{X}_{\text{new}}$ are obtained using:

$$\hat{\mathbf{y}} = \mathbf{X}_{\text{new}}\mathbf{b}$$

PLS, particularly in its orthogonal score variant, provides a powerful framework for feature selection in high-dimensional datasets. The components derived from PLS models—particularly weights, loadings, and regression coefficients—can be effectively implemented in feature selection processes. These components identify the most relevant features for classification: 1) weights ($\mathbf{w}$) directly represent the importance of each variable in constructing the PLS components. 2) loadings ($\mathbf{p}, \mathbf{q}$) indicate how much each original variable contributes to the PLS components and how important each component is for prediction. 3) regression coefficients ($\mathbf{b}$) provide a direct measure of the impact of each variable on the predicted outcome.

Feature selection

1. **Variable importance in projection (VIP):** scores takes into account the outputs from PLS model, specifically utilizing scores, weights and loadings. VIP quantifies the relative importance of each predictor variable in explaining the variance in the response variable, thus serving as a critical tool for FS in PLS-based models.

   VIP score for the j^{th} variable is given as:

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \cdot \mathrm{SSY}_f \cdot J}{\mathrm{SSY}_{\mathrm{total}} \cdot F}} \tag{1.8}$$

Where w_{jf} is the weight value for the j^{th} variable in the f^{th} component, and SSY_f $\left(b_f^2 \mathbf{t}_f' \mathbf{t}_f\right)$ represents the sum of squares of the variance in the response variable explained by the f^{th} component, J is the number of $\mathbf{X}$ variables, $\mathrm{SSY}_{\mathrm{total}}$ $\left(b^2 \mathbf{T}'\mathbf{T}\right)$ is the total sum of squares explained of the dependent variable, F is the total number of components. VIP score (VIP_j) evaluates the contribution of each variable to the variance explanation in the PLS model. Variables with VIP scores greater than 1 are considered to have a significant impact on the model, while those below this threshold may be candidates for exclusion [37].

2. **Statistical analysis**

   Statistical analysis in FS employs mathematical methods to examine variable contributions through hypothesis testing, quantifying the significance of each feature associated with the target variable. For each feature, a null hypothesis (H0) posits no significant relationship with the target variable, while an alternative hypothesis (Ha) suggests a meaningful association. Statistical tests, like Mann-Whitney U (MWU) test, are then employed to evaluate these hypotheses. The first output of those tests is p-value, a probability measure that quantifies the likelihood of observing the data (or more extreme data) under the assumption that the null hypothesis is true. Smaller p-values indicate stronger evidence against null hypothesis, suggesting that feature may be relevant to the model. P-value serves as a metric for ranking features based on their statistical significance [8].

   Implementing statistical analysis for FS involves an iterative process. It begins by calculating p-values for all features in the dataset, then those values are used to rank features from most to least significant. Based on this ranking, a subset of features is selected, either by choosing those below a certain p-value threshold or by selecting a predetermined number of top-ranked features.

While p-values are useful for FS, relying on them alone can be misleading. Effect size and correlation coefficients provide crucial complementary information. Effect size quantifies the magnitude of relationships between variables, offering a measure of practical significance that p-values cannot capture. Correlation coefficients, like Pearson's r or Spearman's rho, indicate the strength and direction of associations between features and the target variable. This approach helps identify features that are not only statistically significant but also have meaningful impact on the target variable. It can also highlight potentially important features in smaller datasets where achieving statistical significance is challenging [8].

3. **Model-agnostic approach**

   Model-agnostic methods like SHAP (SHapley Additive exPlanations) are versatile black-box approaches to FS that can be applied to any computational models without affecting the internal structure. SHAP is a robust framework derived from game theory that assesses the contribution of each feature to individual predictions. This method is particularly valuable because it is model-agnostic, meaning it can be applied to any model for a consistent and comprehensive way to interpret model behavior. By computing Shapley values, SHAP quantifies the importance of each feature in the decision-making process, independent of model complexity [8].

   Shapley value $\phi_i(v)$ is computed by aggregating its marginal contributions over all possible coalitions of features. It is defined as:

$$\phi_i(v) = \frac{1}{M} \sum_{S \subseteq \mathcal{M} \setminus \{i\}} \binom{M-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S)) \tag{1.9}$$

   where M is the total number of features, S is a subset of features excluding i, and $v(S)$ represents the model's evaluation function over the subset S. The term $v(S \cup \{i\}) - v(S)$ captures the marginal contribution of feature i to coalition S [38].

   Shapley values are computed as a unified measure of feature importance, representing the average marginal contributions of features across all possible combinations. For example, to compute the Shapley value for a fea-

ture, all possible outcomes are considered by replacing the feature value with that from another instance and comparing the original prediction with the new one.

Data fusion

Data fusion is an analytical approach that integrates information from multiple data sources to improve the performance of predictive models. The core principle is to integrate diverse spectral datasets, each contributing distinct analytical strengths, to achieve a more holistic and detailed compositional understanding than any single method could provide independently. Three primary data fusion techniques are widely recognized: Low-Level Data Fusion (LLDF), Mid-Level Data Fusion (MLDF), and High-Level Data Fusion (HLDF).

LLDF represents the most straightforward approach, involving the direct concatenation of data matrices from different analytical techniques to create a single, comprehensive dataset for subsequent analysis. This method preserves the original structure of each dataset, with fusion occurring at the raw/preprocessed data level without FS or feature reduction (FR) methods. The process aligns data from different spectroscopic methods along a shared dimension, typically the sample or observation axis. For instance, when Raman and FTIR spectra are collected for a set of samples, LLDF merges these datasets side-by-side, maintaining the original number of samples while expanding the feature space to include variables from both techniques [39–41].

MLDF addresses some limitations inherent in LLDF by focusing on reducing the dimensionality of each dataset independently or selecting the most prominent features before concatenation. This approach offers a balance between preserving relevant information and mitigating the curse of dimensionality. MLDF typically involves applying dimension reduction techniques such as PCA or PLS to each dataset. The resulting scores or selected components, which capture the most significant variations or predictive power, are then concatenated to form a new, reduced-dimension matrix for subsequent analysis [39–41].

HLDF, also known as decision-level fusion, takes a distinct approach by combining the outputs or predictions of individual models rather than com-

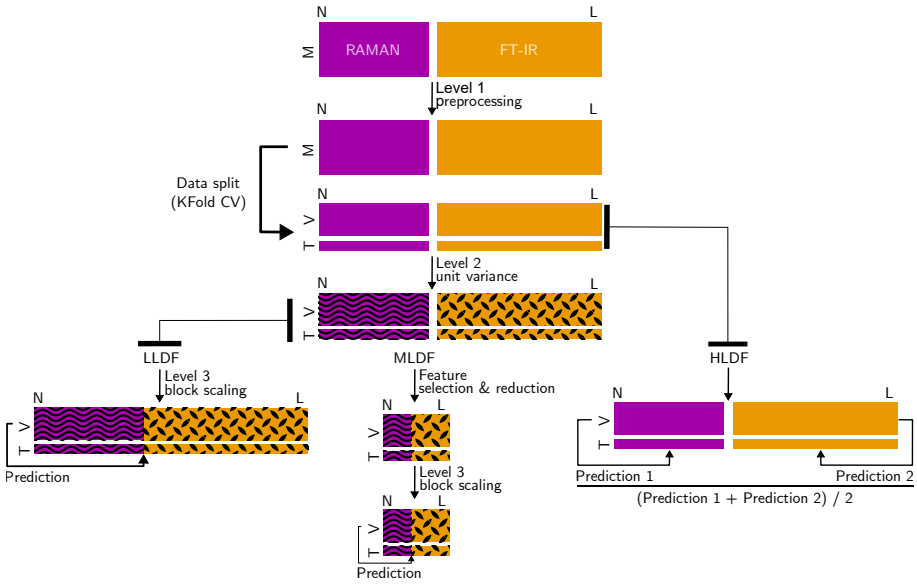


Figure 1.9: Multi-level data fusion workflow integrating Raman and FTIR spectroscopy for lung cancer detection, illustrating low-level (LLDF), mid-level (MLDF), and high-level (HLDF) fusion strategies with associated preprocessing and model validation steps.

plete datasets or extracted features. This method develops independent prediction models for each technique and then merges their outputs to form a final decision or prediction. HLDF allows for flexibility in model selection, as different types of models can be applied to different datasets based on their specific characteristics or the nature of the analytical problem. A key advantage is to circumvent scaling issues between different spectroscopic techniques, making it particularly useful when dealing with heterogeneous data or complex relationships between variables from different techniques [39–41].

- **Level 1, level 2 and level 3 data processing:** Data processing at those stages shown in Figure 1.9 constitute a hierarchical approach to prepare spectroscopic dataset for further analysis. Level 1 processing, as discussed in Section 1.2.3, aims to improve data quality by correcting measurement-specific artifacts and enhancing the signal-to-noise ratio. In the image, this is applied to both Raman and FTIR data blocks,

potentially involving techniques such as baseline correction, scatter correction, etc. Level 2 processing, labeled as "unit variance", focuses on intra-block scaling to equalize the contributions of variables within each block. This often involves mean-centering or scaling to unit variance, ensuring comparability across variables with different magnitudes or units. Level 3 processing, denoted as "block scaling", addresses inter-block relationships and is crucial for balancing the influence of each data block in the final model. This step prevents larger or more variable blocks from dominating the analysis. The image illustrates how these three levels of processing are sequentially applied to prepare the data for different fusion strategies. This systematic approach ensures that the complementary information from different spectroscopic techniques is optimally integrated for subsequent multivariate analysis.

Performance evaluation metrics

Model evaluation is fundamental in chemometrics, providing quantitative measures and enabling rigorous comparisons across different mathematical operations. Evaluation metrics assess predictive capability of a model, its generalization to unseen data, and its robustness across various operational conditions. Common classification metrics include [42]: (Tp: true positive, Tn: true negative, Fp: false positive, Fn: false negative, Tpr: true positive rate, Fpr: false positive rate)

1

- **Accuracy:**

$$\text{Accuracy} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn}} \tag{1.10}$$

- **Precision:**

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp} + \text{Fp}} \tag{1.11}$$

- **Sensitivity (Recall):**

$$\text{Sensitivity} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}} \tag{1.12}$$

- **Specificity:**

$$\text{Specificity} = \frac{\text{Tn}}{\text{Tn} + \text{Fp}} \tag{1.13}$$

- **F1 score:**

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1.14}$$

- **AUC-ROC:**

$$\text{AUC-ROC} = \int_0^1 \text{Tpr}(\text{Fpr}), d(\text{Fpr}) \tag{1.15}$$

1.3 Materials and methods

1.3.1 Sample collection and preparation

Peripheral blood samples, collected in ethylenediaminetetraacetic acid (EDTA) tubes, underwent immediate centrifugation at 2095xg (4°C, 15 minutes). The resultant human blood plasma was separated, aliquoted, and stored at -80°C for subsequent analysis (see Figure 1.10A).

A glass substrate covered by aluminum foil was employed as the base material in Raman analysis, chosen for its high reflectivity, stability, flexibility, low background signal and cost-effectiveness [9]. A critical step involved cleaning the covered substrate using a nitrogen blow, effectively removing dust particles that might interfere with spectral quality. Fresh human blood plasma samples, stored at -80°C, were rapidly processed to minimize potential degradation during thawing. A volume of 1 µL from each sample was precisely deposited onto the prepared substrate. The samples were then dried in a vacuum desiccator for approximately 5 minutes, a duration empirically determined to achieve optimal dryness without risking sample degradation or contamination (see Figure 1.10B).

FTIR sample preparation focused on a rigorous cleaning procedure to maintain the ATR crystal in pristine condition. The crystal was treated with 20% sodium dodecyl sulfate solution for 10 minutes to remove organic or inorganic residues. Multiple rinses with distilled water followed, with careful removal of

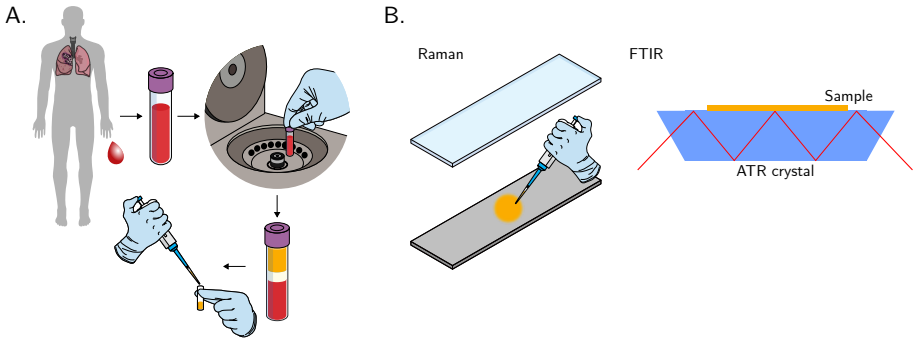


Figure 1.10: Sample handling process: (A) sample collection and (B) sample preparation for Raman and FTIR spectroscopy.

liquid using disposable optical cleaning tissue to prevent scratching the delicate surface, then meticulously dried using nitrogen flow. A final cleaning step utilized specialized optical element cleaning paper with isopropanol, ethanol, or acetone to remove any remaining organic residues. This thorough cleaning protocol was essential to prevent cross-contamination between samples and ensure the acquisition of high-quality spectra. Prior to sample application, a background spectrum of the clean ATR crystal exposed to air was recorded. Fresh human blood plasma samples (1 µL) were then applied to the crystal, and the drying process was monitored through consecutive spectral acquisitions (see Figure 1.10B). This approach allowed for precise determination of complete sample drying, typically occurring around 10 minutes post-application. To ensure data reliability and account for potential sample heterogeneity, 10 spectra were collected for each dried sample.

1.3.2 Data collection and preprocessing

Raman measurements were conducted using Renishaw inVia™ confocal Raman microscope. The system operated with 785 nm laser at 73 mW output power, utilizing 50x long-distance objective for laser focusing and signal collection, and 1200 L/mm spectrometer grating. This configuration optimized signal strength and signal-to-noise ratio while minimizing sample damage. For each sample, 25 spectra were acquired from the droplet's periphery to capture the

concentrated biomolecules resulting from the coffee ring effect. Each spectrum comprised 20 accumulations with 1 s exposure time.

FTIR measurements were performed using Bruker Vertex 70 spectrometer in ATR mode. Spectra were recorded at 4 cm^{-1} resolution with 100 s sampling time per measurement. To ensure data consistency and reliability, samples were completely dried before analysis to avoid water band interference. Ten spectra were collected for each sample to guarantee result stability and reproducibility.

Spectral data from both Raman and FTIR spectroscopy underwent careful preprocessing to enhance signal quality across samples. For Raman spectra, initial preprocessing involved the elimination of random cosmic ray interference using the zap function in the Renishaw WiRE 5.4 software. Subsequently, two preprocessing techniques were applied: asymmetric Whittaker baseline correction ($\lambda = 100$, $p = 0.01$) to remove baseline drifts and distortions [43], followed by SNV transformation to correct for variations due to sample thickness, scattering, and instrumental response [44]. In contrast, FTIR spectra required only SNV transformation, as baseline correction was deemed unnecessary and lacking physical justification for this technique. To obtain a single representative spectrum per subject, 25 Raman spectra and 10 FTIR spectra were averaged for each individual. This averaging process serves to reduce random noise, improve signal-to-noise ratio, and provide one comprehensive spectral signature for each subject. The differential approach in preprocessing between Raman and FTIR data highlights the technique-specific considerations necessary in spectral analysis, ensuring that each dataset is optimally prepared for subsequent analytical procedures while maintaining the integrity of the spectral information.

1.3.3 Data analysis strategies

This research utilized Spyder Integrated Development Environment (IDE), a comprehensive scientific platform for Python (version 3.9.12). Spyder provides a range of tools essential for data analysis, including console, code editor, and debugging capabilities. The study employed several well-established Python libraries: Numpy and Pandas for data processing, Matplotlib for generating high-quality result visualizations, Scipy for optimizing data processing

algorithms, and Sklearn for model design.

Publication 1: Supervised learning algorithms

This section employed a range of predictive models as shown in Figure 1.11 to analyze the spectral dataset derived from Raman spectroscopy. Five classifiers - LDA, SVM, NB, LR, and RF - were tested in different configurations. First, these classifiers were combined with PCA to reduce data dimensionality while preserving variance. Second, the classifiers were integrated with both PCA and Fisher Score feature selection. Third, the classifiers were applied directly to the preprocessed data. Additionally, Partial Least Squares-Discriminant Analysis (PLS-DA) was utilized as a standalone model to handle the multicollinearity often present in spectroscopic data.

- **Fisher score:** Fisher discriminant ratio is a FS technique that evaluates the discriminative power of each feature within dataset. It maximizes the

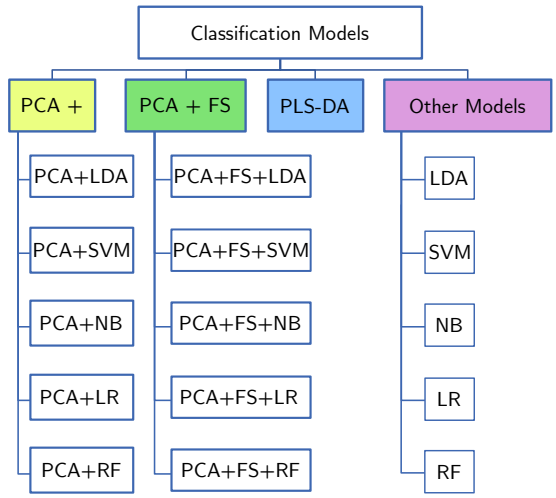


Figure 1.11: Classification models used for spectral data analysis. PCA was combined with LDA, SVM, NB, LR and RF for dimensionality reduction, with FS further refining inputs. PLS-DA addressed multicollinearity, while classifiers were also tested independently on preprocessed data.

separation between different classes. The score is computed as the ratio of the between-class variance to the within-class variance. Mathematically, S_i for i-th feature is given by

$$S_i = \frac{\sum n_j(\mu_{ij} - \mu_i)^2}{\sum n_j \rho_{ij}^2}$$

where μ_{ij} and ρ_{ij} represent mean and variance of i-th feature within j-th class, respectively, n_j denotes the number of instances in class j, and μ_i is the overall mean of i-th feature. Higher score indicates that features provide significant class separation, suggesting its importance in classification. By ranking features based on their scores, only those with the highest scores are selected, thereby improving model performance by focusing on the most discriminative features [45].

Publication 2: Feature selection methods

Figure 1.12 illustrates a systematic approach through FS methods. The process begins with preprocessed data, which is divided into train and test sets. Train set, comprising 85% of the total data, undergoes 5-fold cross-validation to ensure model robustness. This part involves feature ranking and selection. Various techniques, including PLS-DA, VIP scores, Shapley values, and statistical analysis, are employed to evaluate and rank features. This ranking process is repeated 500 times, allowing for a comprehensive assessment of feature importance. FS then occurs based on percentile thresholds, ranging from 1% to

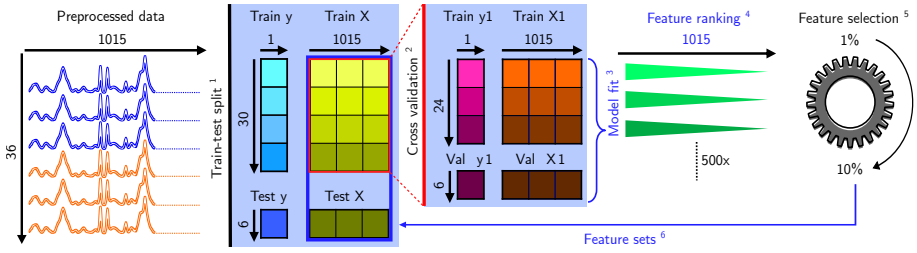


Figure 1.12: Flowchart illustrating the framework for evaluating feature selection methods in spectral data analysis.

10%. The most frequently selected features across all iterations are identified, ensuring only the most consistently selected features are retained. These features are then applied to both train and test sets, forming the basis for logistic regression model. This iterative and selective approach aims to identify the most relevant spectral features for lung cancer.

Publication 3: Data fusion strategies

The workflow illustrated in Figure 1.9 presents multi-level data fusion strategies. The process starts with row-based preprocessing of the spectra, detailed in Section 1.2.3, followed by stratified 6-fold cross-validation to partition data into train and test sets. This approach helps balance bias and variance while mitigating overfitting through repeated data partitioning. After data splitting, unit variance scaling is applied to ensure each block has equal variance. This method standardizes datasets with different units or scales, promoting equal contribution across the analysis. The scaling factor is calculated as $1/n_{\text{block}}$, where n_{block} is the number of variables in each block.

In LLDF, preprocessed data matrices are concatenated to maximize within-block variance and span the full range of measured variables. Soft block scaling is then applied, adjusting each column by multiplying standard deviation by the fourth root of the number of features. This ensures balanced contributions from all blocks and prevents dominance by larger blocks. In MLDF, FS or/and FR is/are incorporated prior to scaling. FS employs regression coefficients (RCs) from PLS-DA to rank features by their absolute values, evaluated over 100 iterations with 6 folds. Selection frequency identifies the most significant features, which are refined iteratively. FR utilizes PCs analysis, incrementally adding components until optimal model performance is reached. In HLDF, model predictions are combined rather than data, averaging predictions from separate logistic regression models trained on each block. This method treats each block equally, bypassing block scaling, and derives final predictions by averaging probabilities and applying thresholds. Model performance is validated through averaged accuracy scores across folds [46–48].

1.4 Results and discussions

1.4.1 Publication 1: Supervised learning algorithms

Figure 1.13 presents a comprehensive analysis between lung cancer patients (purple) and healthy controls (green). The visualization comprises six distinct score plots that showcase the data from different analytical methods.

In Raman analysis, PCA results are displayed in Figure 1.13A-B. Figure 1.13A presents the relationship between PC1 and PC2, where PC1 accounts for the largest variance at 42.12%, followed by PC2 at 15.4%. Together, these first two components effectively capture the dominant patterns in the data because they represent the two most significant orthogonal directions of variation, with each subsequent component necessarily capturing less variance due to increasingly restrictive orthogonality constraints. Figure 1.13B explores less conventional but meaningful combination of PC2 and PC5, where PC5 contributes 5.77% of the variance. PC5 was selected due to its high Fisher score as said earlier despite its lower variance contribution. It is worth noting that PC3 and PC4, though not visualized, account for 11.84% and 8.67% of the variance respectively but showed less discriminatory power.

The analysis extends to PLS, represented in Figure 1.13C-D through latent variables (LVs). PLS specifically seeks to maximize the covariance between data matrix and response variables, providing an supervised approach to classification. Latent variables represent linear combinations of the original variables that best explain the between-group differences while maintaining the correlation with the classification outcome.

Figure 1.14 shows typical Raman spectra of human blood plasma acquired using a 785 nm excitation wavelength. The main graph (A) provides an overview of averaged spectra, while three magnified regions (B, C, D) allow for a closer examination of specific spectral features.

The overall spectral patterns appear similar between the two groups at first glance. However, a more careful look at the zoomed-in sections reveals subtle differences in peak intensities. These variations, though difficult to detect

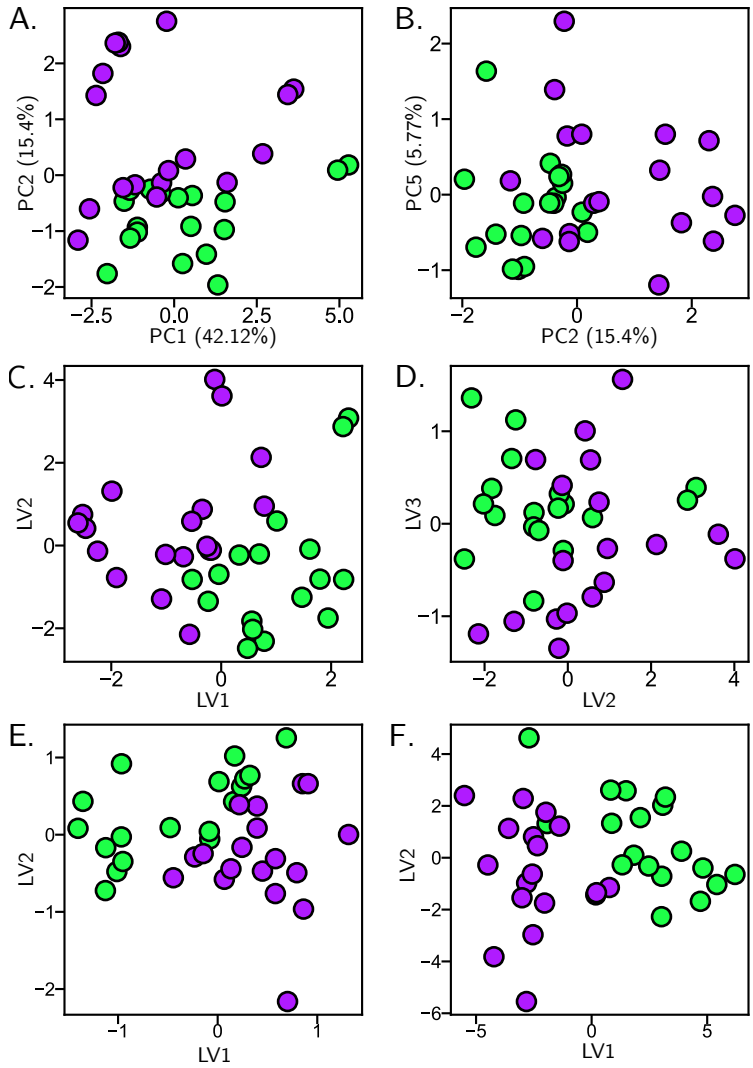


Figure 1.13: 2D score plots illustrating the distinction between lung cancer patients (purple) and healthy controls (green). (A) PC1-PC2, (B) PC2-PC5, (C) LV1-LV2, and (D) LV2-LV3 are derived from Raman spectroscopy, as detailed in Publication 1. (E) LV1-LV2 is obtained from FTIR spectroscopy, and (F) LV1-LV2 is generated from the fused Raman and FTIR data, both of which are discussed in Publication 2.

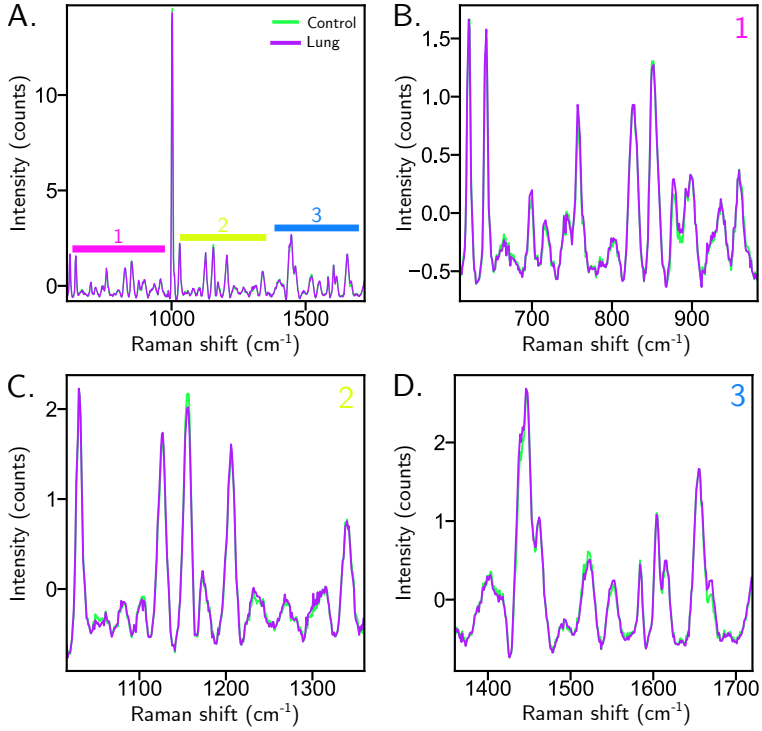


Figure 1.14: Selected spectral regions of mean Raman spectra per class. (A) full range, (B) 610 to 990 cm^{-1}, (C) 1016 to 1360 cm^{-1} and (D) 1360 to 1720 cm^{-1}.

visually, likely reflect changes in the molecular structures of the plasma samples. Figure 1.14B shows the 610-990 cm^{-1} range, capturing ν-ring breathing, skeletal modes, and stretches like ν(CC), ν(C=O) etc. linked to proteins and lipids. Notably, the 990-1016 cm^{-1} range was visually deemphasized during zoomed-in presentations, as its prominent signal could obscure other relevant spectral features. Figure 1.14C examines distinct vibrational stretching and bending modes as well as deformation modes and amide III region. Figure 1.14D targets the 1360–1720 cm^{-1} range for protein secondary structures through C=O and C=C stretching, amide I and CH_2/CH_3 deformation modes associated with proteins (α-helix, aromatic amino acids) and lipids.

It is important to remember that biological variability and experimental factors may also influence these results. The small differences could come from disease-related molecular changes, but they could also be due to environmental factors, life style, eating habits, etc. This inherent uncertainty in spectral data makes it challenging to interpret molecular differences between groups just by looking at the graphs. What is seen is a mix of molecular signatures, the limits of the measurement precision and the uncertainty of the entire data structure. Given these challenges in distinguishing subtle spectral differences linked to lung cancer, chemometrics becomes a crucial tool. By analyzing the entire dataset to obtain the entire spectral information, classification and prediction algorithms can detect patterns and connections that might be hidden by noise and variability.

Loadings presented in Figure 1.15A and B explain the spectral features that drive the separation in PCA and PLS, respectively. These plots illustrate how specific Raman shifts contribute to each principal component and each latent variable, effectively bridging the gap between the original high-dimensional spectral data and the reduced-dimensional space used in the analysis.

In Figure 1.15A, loadings of PC1, PC2, and PC5 reveal the bands that account for the most significant variations in the dataset. PC1 and PC2, typically capturing the majority of the variance, highlight spectral regions where differences are mostly pronounced. Loadings at each band indicate the strength and direction of the contributions to the respective component. Strong positive or negative values suggest that these particular bands play a crucial role

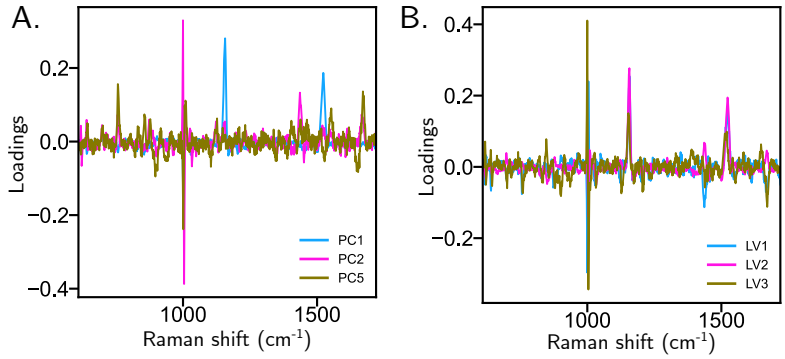


Figure 1.15: Loading plots based on (A) PCA and (B) PLS.

in defining the characteristics of that specific component in biomolecules. PC5 chosen because of its high Fisher score as mentioned in Section 1.3.3 while accounting for less overall variance, also reveal other subtle spectral features that still contribute to the discrimination. These could represent hidden molecular changes associated with the disease that are not immediately apparent in the first component/s. Moreover, although PC3 and PC4 captured higher variance than PC5, they showed less discriminatory power and were therefore not included in the plot. This is because PCs are derived solely from the directions of highest data variance, independent of any class labels. Therefore, it is not uncommon for PCs that capture smaller portions of variance to carry more discriminative power if they align with the features distinguishing each class. In this case, PC5 captures the least variance, but is more distinctively associated with specific class differences.

Figure 1.15B displays loadings (LV1, LV2, and LV3) derived from PLS. Since PLS directly correlates the spectral data with class labels (e.g., cancer vs. healthy), these loadings also reflect which vibrational features are most critical for classification. However, the contribution to each Raman band varies depending on specific latent variable, similar to how it changes in PCA loadings, as some bands may be dominated by particular biomolecules in one component (e.g., LV3), but these changes might also appear in other components (e.g., LV2 or PC2) with less significance. For example, the band around 1000 cm^{-1} may show a pronounced contribution in LV3 in PLS, while this same band

could be more influential in PC2 in PCA. This variability across components is important because it highlights how different vibrational modes, representing distinct biomolecular structures, contribute to each component or variable.

Model performance was primarily evaluated through different approaches based on a feature selection method and multiple classifiers. The first approach in Figure 1.16A, utilizing the first 10 PCs with classifiers, reveals optimal performance within 6-8 PCs. PCA-LDA and PCA-SVM demonstrated superior performance with accuracies of 0.85 ± 0.13 and 0.84 ± 0.16 respectively at 7 PCs. LDA is likely effective here because the majority of features might assume Gaussian-distributed classes with identical covariance. LDA also finds boundaries easily when classes are linearly separable or have high within-class cohesion. Moreover, PCA was applied before LDA not only to reduce dimensionality but also to regularize the problem, preventing overfitting by stabilizing the inversion of within-class covariance matrix and minimizes the risk of overfitting. PCA-LR also showed comparable performance (0.83 ± 0.15) with 6 PCs, while PCA-RF and PCA-NB required different optimal component numbers, achieving 0.80 ± 0.17 (8 PCs) and 0.77 ± 0.14 (6 PCs) respectively.

The second approach, Figure 1.16B, incorporated Fisher score feature selection following PCA, introducing a more targeted feature selection process. This method achieved comparable classification accuracies while requiring fewer principal components, demonstrating more efficient feature utilization. PCA-FS-SVM achieved 0.85 ± 0.14 with only 5 PCs, while PCA-FS-LDA and PCA-FS-LR achieved similar performance (0.84 ± 0.14) using 6 and 5 PCs, respectively. Applying Fisher score helps maximize class separability rather than just variance, improving the decision boundary of SVM. Since PCA has already removed much redundancy, only a marginal improvement is expected from further feature selection. Notably, PCA-FS-NB showed improved efficiency, requiring only 3 PCs to achieve 0.81 ± 0.15, though PCA-FS-RF showed slightly decreased performance at 0.77 ± 0.16 with the same number of components.

Standalone classifiers in Figure 1.17, without dimensionality reduction or feature selection, revealed distinct performance patterns. LDA demonstrated the highest model performance with 0.84 ± 0.14. NB also achieved comparable performance (0.82 ± 0.13) despite its algorithmic simplicity, while RF main-

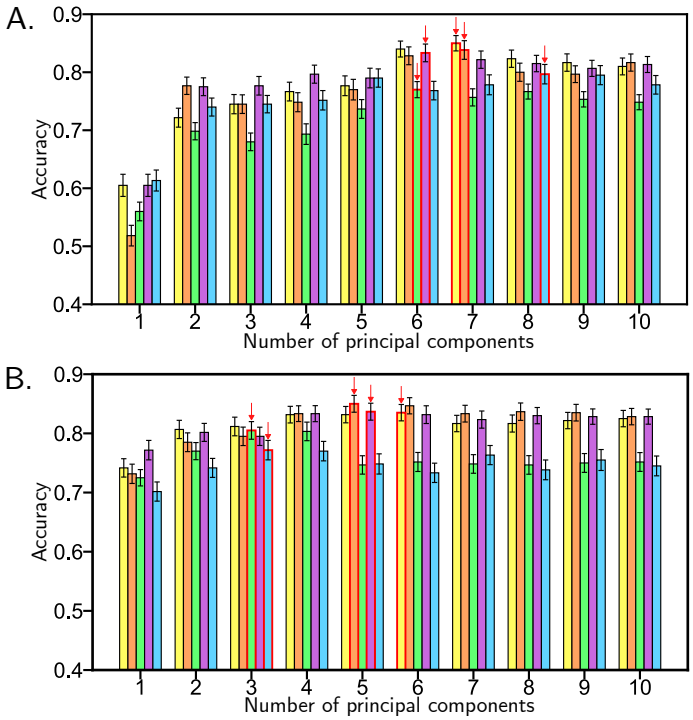


Figure 1.16: Comparative analysis of classifiers using different number of principal components (PCs) and a feature selection method. (A) Classification accuracy using sequential inclusion of the first 10 PCs with five different classifiers: LDA (yellow), SVM (orange), NB (green), LR (purple), and RF (blue). (B) Classification accuracy employing Fisher score-selected PCA components with the same classifiers. Red arrows indicate the highest accuracy for each classifier. Error bars represent standard errors.

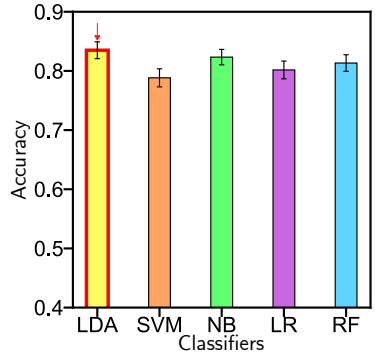


Figure 1.17: Comparison of classification algorithms in original feature space. Performance assessment of five classifiers applied directly to spectroscopic data without FR or FS. Error bars represent standard errors.

tained consistent accuracy (0.81 ± 0.14) through its ensemble approach. SVM and LR showed slightly reduced performance (0.79 ± 0.15 and 0.80 ± 0.15 respectively), suggesting potential limitations when processing high-dimensional datasets. Non-linear SVM kernels could improve results while having more flexible decision boundaries for the separation, but the limited dataset increases the risk of overfitting, so they were not explored further. This limitation is particularly notable as high-dimensional data often contains redundant and unnecessary information, which can lead to unsatisfactory outcomes. To address this challenge, implementing FR or FS generally improves the performance of classifiers by reducing noise, redundancy, and multicollinearity, ultimately resulting in high model performance.

Unlike PCA, PLS-DA in Figure 1.18 specifically incorporates class information during latent variable construction, optimizing the feature space for classification rather than just variance explanation. This approach allows to identify class-discriminative patterns more efficiently than standard PLS regression, making it particularly suitable for binary classification tasks in cancer detection. The analysis revealed optimal performance with just three latent variables, achieving a mean accuracy of 0.82 ± 0.14. This performance with minimal components highlights its ability to capture relevant class-specific spectral variations while reducing dimensionality.

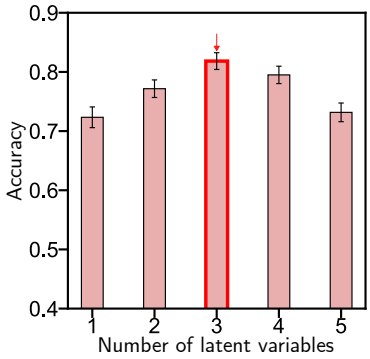


Figure 1.18: Classification performance of PLS-DA as a function of latent variables. Accuracy plot demonstrates optimal model performance achieved with three latent variables. Error bars represent standard errors.

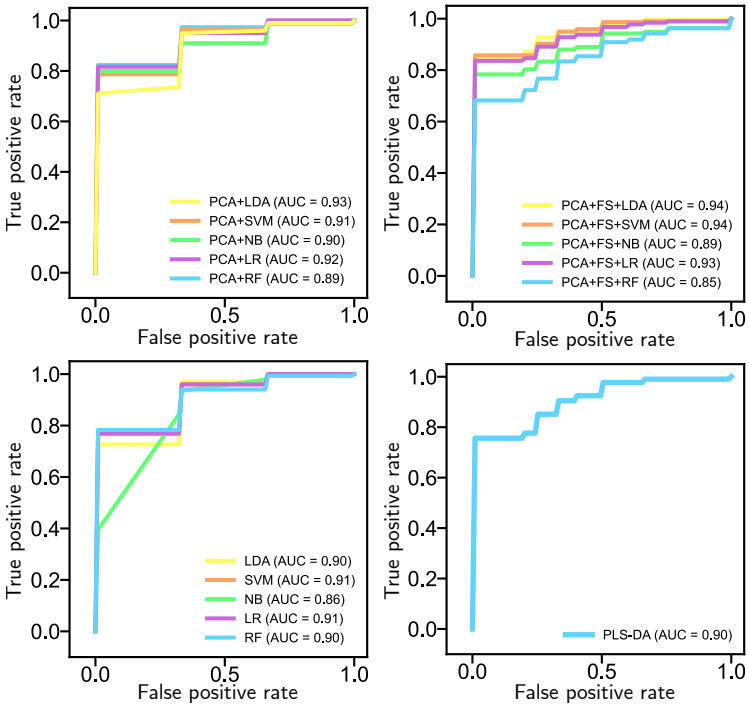


Figure 1.19: ROC curves of (A) PCA + classifiers, (B) PCA + FS + classifiers, (C) only classifiers and (D) PLS-DA.

Table 1.1: Performance comparison of classification models based on accuracy and AUC scores.

Models	**PCs/LVs**	**Accuracy**	**AUC**
PCA+LDA	7	0.85 ± 0.13	0.93
PCA+SVM	7	0.84 ± 0.16	0.91
PCA+NB	6	0.77 ± 0.14	0.90
PCA+LR	6	0.83 ± 0.15	0.92
PCA+RF	8	0.80 ± 0.17	0.89
PCA+FS+LDA	6	0.84 ± 0.14	0.94
PCA+FS+SVM	5	0.85 ± 0.14	0.94
PCA+FS+NB	3	0.81 ± 0.15	0.89
PCA+FS+LR	5	0.84 ± 0.14	0.93
PCA+FS+RF	3	0.77 ± 0.16	0.85
LDA	-	0.84 ± 0.14	0.90
SVM	-	0.79 ± 0.15	0.91
NB	-	0.82 ± 0.13	0.86
LR	-	0.80 ± 0.15	0.91
RF	-	0.81 ± 0.14	0.90
PLS-DA	3	0.82 ± 0.14	0.90

Figure 1.19 compares different classification approaches through ROC curves based on FR and FS methods. Figure 1.19A, PCA combined with LDA, shows 0.93, while LR follows closely with 0.92, and SVM reaches 0.91. Lower scores for NB and RF, at 0.90 and 0.89 respectively, suggest that those may not fully utilize PCs as effectively as the rest do. The reason behind this is that, NB classifier is not linear, but can sometimes act as linear in specific feature spaces only if the likelihood factors $p(x_i|C_k)$ are from exponential families. However, one should not expect that the variances of each feature (Raman shift intensity) will be same for both lung cancer patients and healthy controls, therefore decision rule does not simplify to a linear boundary. RF, in contrast, classify by dividing the feature space into discrete, non-linear regions (or "boxes"), making decisions based on majority voting within these regions. This approach allows RF to capture complex, non-linear patterns, but it may lead to overfitting in linearly separable data.

Figure 1.19B explores Fisher score-based feature selection in refining the principal components. Both PCA-FS-LDA and PCA-FS-SVM achieve 0.94, indicating that selecting the most relevant components for classification improves model performance. Fisher score ranks features based on their separation between classes, and the improved results highlight that focusing on the top-ranked components helps the classifiers concentrate on the most discriminative components. Table 1.1 further quantifies these observations in terms of classification performances. For example, PCA-FS-RF, which uses only 3 PCs, achieves an accuracy of 0.77 ± 0.16 with an AUC of 0.85, showing that the balance between FR and maintaining enough relevant information can influence model performance.

Figure 1.19C shows the performance of standalone classifiers applied without dimensionality reduction or feature selection. LDA, LR, and RF maintain similar AUC values between 0.90 and 0.91, which indicates that these models can handle high-dimensional spectral data alone. SVM also performs similarly with 0.91, while NB stays behind with 0.86. This difference in performance may arise from how each classifier deals with the original, high-dimensional data space. NB, which assumes feature independence, faces challenges with correlated spectral features, while the other classifiers can better handle such complexity. Moreover, applying PCA or PCA combined with FS did not improve the performance of NB. One reason for this is that NB assumes conditional independence between features given the class, represented as

$$p(x_i|C_k) = p(x_i|x_{i+1}, \ldots, x_n, C_k)$$

which means that the features do not need to be independent but are treated as such once conditioned on the class label. Here, $p(x_i|C_k)$ represents the probability of observing the feature x_i, given that the instance belongs to class C_k while $p(x_i|x_{i+1}, \ldots, x_n, C_k)$ shows the probability of x_i given all other features $x_{i+1}, \ldots, x_n$, in addition to the class C_k. The assumption made by NB is that, given the class C_k, the value of x_i does not depend on the values of the other features, $x_{i+1}, \ldots, x_n$. Thus, the formula essentially states that knowing the values of other features does not change the conditional probability of x_i, as long as the class is known [49, 50]. However, PCA does not address this conditional independence assumption and, in fact, may worsen performance by removing

features with small variance that could still hold discriminative power between classes. Since PCA merely rotates the data and reduces dimensionality based on variance, it may not resolve the issue of correlated features in the way that would benefit the assumptions in NB.

PLS-DA in Figure 1.19D incorporates class labels during the extraction of LVs, which maximizes the covariance between the spectral data and the class variables. AUC of 0.90 means that it can effectively capture the relevant information for classification. Unlike traditional PLS, which is designed for continuous response variables, PLS-DA adapts the method for classification by treating the class labels as categorical variables. This ensures that LVs derived through PLS are optimized for class separation. The key point of PLS-DA lies in its ability to handle collinearity among spectral features (wavelengths or wavenumbers), projecting them into latent variables that are more useful for classification.

1.4.2 Publication 2: Feature selection methods

Figure 1.20 presents comprehensive spectra derived from feature selection methods, indicating different mathematical operations. Each method exhibits varying model performance in terms of accuracy, as depicted in Figure 1.21. When certain Raman bands show high importance across multiple methods, it strongly suggests these spectral regions are truly significant for the classification task. This aligns with the described approach in Figure 1.12 where features are ranked and selected based on their consistent importance across 500 iterations, using percentile thresholds from 1% to 10%. The varying patterns visible in all three methods across different bands indicate varying levels of molecular information content, with some regions showing notably higher importance than others. These patterns reflect the underlying molecular vibrations and structural changes that distinguish between different sample classes in the dataset.

Building on these insights, PLS regression is utilized for handling complex, high-dimensional spectral data while selecting the most important features and reducing the whole column features to a smaller set of uncorrelated variables

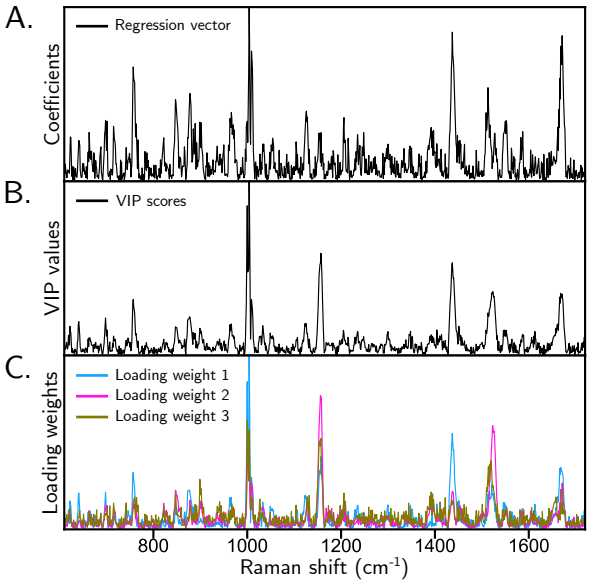


Figure 1.20: Comparison of three feature selection methods: average results over all iterations from (A) regression coefficients, (B) VIP scores and (C) loading weights.

through covariance maximization. The mathematical framework provides loading weights (LWs) that show feature-response covariance, and RCs that measure feature-response associations. VIP scores further enhance this framework by providing cumulative feature weights using the scores, loading weights, and loadings from PLS. Later, complementary feature selection methods - including Shapley values and statistical analyses - were implemented to create a robust selection strategy as defined in Section 1.3.3.

SHAP demonstrated superior performance with an accuracy of 0.835 $\pm$ 0.013, establishing itself as the most promising method. The strength lies in its ability to identify feature importance across different predictive models, which enhances the decision-making process. This comprehensive search into each feature likely contributes to high accuracy observed, as models can be fine-tuned with a deeper understanding of data patterns. However, high computational demands and large data requirements may limit its use in resource-constrained environments. Balancing its interpretability advantages with resource efficiency

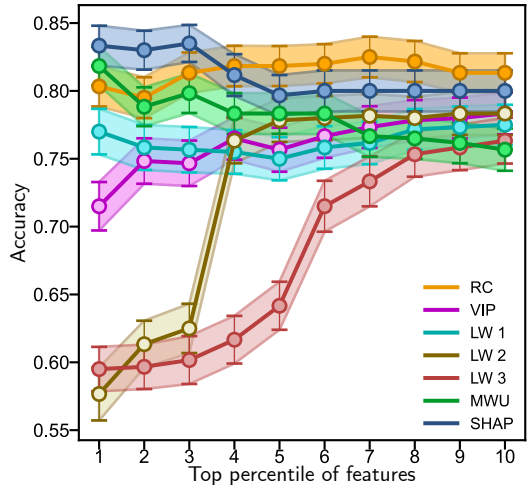


Figure 1.21: Model performances based on feature selection methods as a function of top-ranked feature percentiles (1% to 10%). RC: regression coefficients, VIP scores, LW1: loading weight 1, LW2: loading weight 2, LW3: loading weight 3, MWU: Mann-Whitney U test, SHAP: shapley values.

is essential for its practical application [51].

1

RCs from PLS followed closely with an accuracy of 0.825 ± 0.014, offering a compelling balanced model performance. The magnitude of these coefficients directly indicates the importance of variables for specific responses, allowing for straightforward feature selection based on coefficient values. These findings align with theoretical expectations, as PLS is designed to handle multicollinear data while maintaining interpretability. The approach offers a particularly attractive balance between computational efficiency and accuracy, especially when compared to more complex methods requiring extensive parameter optimization. RCs allow for independent variable selection for each response in a multivariate PLS framework, providing a more streamlined approach compared to running multiple univariate models. This capability makes them particularly valuable for applications in spectroscopic analysis where multiple responses often need to be modeled simultaneously while identifying response-specific relevant features. Moreover, VIP scores and loading weights (LW1, LW2 and LW3) showed consistent but slightly lower performance, with accuracies rang-

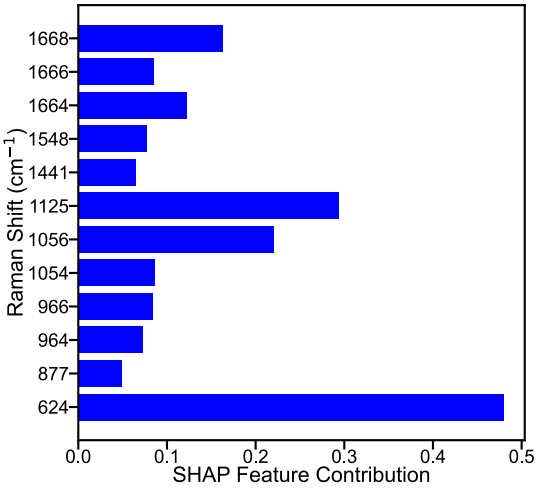


Figure 1.22: SHAP feature importance based on the absolute values associated with the selected wavenumbers.

ing from 0.763 to 0.783. These methods, while less accurate than SHAP or RCs, provide valuable complementary information within PLS framework. The baseline approach without any feature selection method yielded an accuracy of 0.805 ± 0.061, highlighting the potential benefits of implementing strategic feature selection methods.

MWU also achieved significant performance, 0.818 ± 0.014, through its non-parametric statistical framework to identify discriminative spectral features between classes. It evaluates each spectral feature by comparing the distributions of intensity values between classes through rank-based analysis, generating U statistics and corresponding p-values that quantify the degree of class separation. Features are then ranked by their p-values, with lower p-values indicating stronger discriminative power, enabling the identification of the most relevant spectral regions for classification. The main benefit comes from its distribution-free nature and robustness to outliers, making it particularly suitable for spectroscopic data where signal intensities often deviate from normality. MWU test presents strong but slightly lower performance compared to SHAP which can be attributed to its univariate approach - while it effec-

tively captures individual feature differences, it does not account for potential feature interactions that may contribute to class discrimination. SHAP analysis as shown in Figure 1.22 identified peaks at 624 cm^{-1}, 877 cm^{-1}, 964–966 cm^{-1}, 1054–1056 cm^{-1}, 1125 cm^{-1}, 1441 cm^{-1}, 1548 cm^{-1}, and 1664–1668 cm^{-1} as important for classification, with their selection based on their presence within the top 3% of the dataset. These findings align with our previous discussion.

Window-based approach

In this section, a new approach is applied to address the inherent ambiguities in full-spectrum analysis due to the issue of multicollinearity and spectral redundancy in adjacent Raman shifts. This approach employs a systematic segmentation process of the spectral range into discrete windows, specifically targeting regions associated with established vibrational modes of molecules. The fundamental principle behind this approach stems from the understanding that not all spectral regions contribute equally to the biochemical signature of the samples, and some regions may introduce artifacts rather than meaningful information. More specifically, when any feature selection methods are applied to continuous spectral data, they often select clusters of highly correlated neighboring wavelengths that essentially represent the same molecular vibration. This phenomenon, known as spectral correlation or spectral covariance, can lead to overemphasis of certain vibrational modes while potentially overlooking other significant spectral markers. Partitioning the spectrum into discrete windows based on known molecular vibrational assignments ensures that the selected features represent distinct biochemical information from different bands.

RCs were selected due to their robust characteristics in handling specific spectral regions, particularly when analyzing discrete spectral windows. As illustrated in Figure 1.23, the segmentation process begins with the identification of significant regions through a peak detection algorithm. The spectrum undergoes baseline correction, thereby establishing a standardized reference point for subsequent analysis. Local maxima are identified through comparative analysis of adjacent points, specifically evaluating ranges of 10 points to detect significant spectral features. A threshold criterion of 0.1% of the adjusted maximum

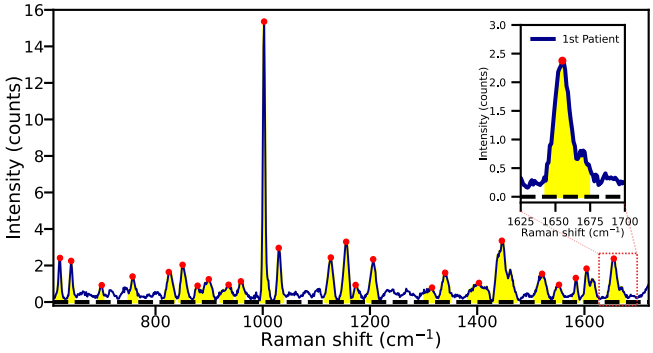


Figure 1.23: Raman spectrum for the first patient, showing the selected windows.

spectral intensity is applied to discriminate between significant peaks and noise. This method allows us to identify areas under the curve with pronounced peaks, focusing on spectral regions where biochemical changes (e.g., proteins, lipids, and nucleic acids) are likely to dominate.

The optimal number of LVs, as shown in Figure 1.24, was evaluated by incrementally adding each variable and observing changes in classification accuracy. The analysis implemented a rigorous hold-out technique with 100 iterations using 85:15 data partition ratio. The results exhibited characteristic behavior: a steep initial ascent indicating effective capture of discriminatory spectral features, later on the model approaches optimal complexity. This asymptotic behavior reflects the diminishing returns of additional components, where further complexity fails to enhance discriminatory power. The optimal number of components, marked by the blue dashed line, was selected at the point where accuracy reached its maximum value, providing the most efficient model complexity for classification.

Figure 1.25 presents the consistently selected Raman bands due to their biochemical relevance and contribution to model performance. This method differs from traditional methods, where all features are treated equally, sometimes leading to noise from less significant spectral regions. This approach resulted in the identification of seven key features (1671, 1436, 878, 760, 1517, 966 and 1206 cm^{-1}) that demonstrated consistent significance across multiple

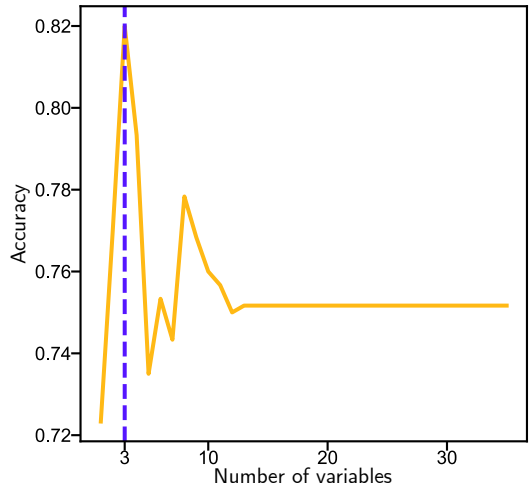


Figure 1.24: Optimization of latent variables based on classification performance.

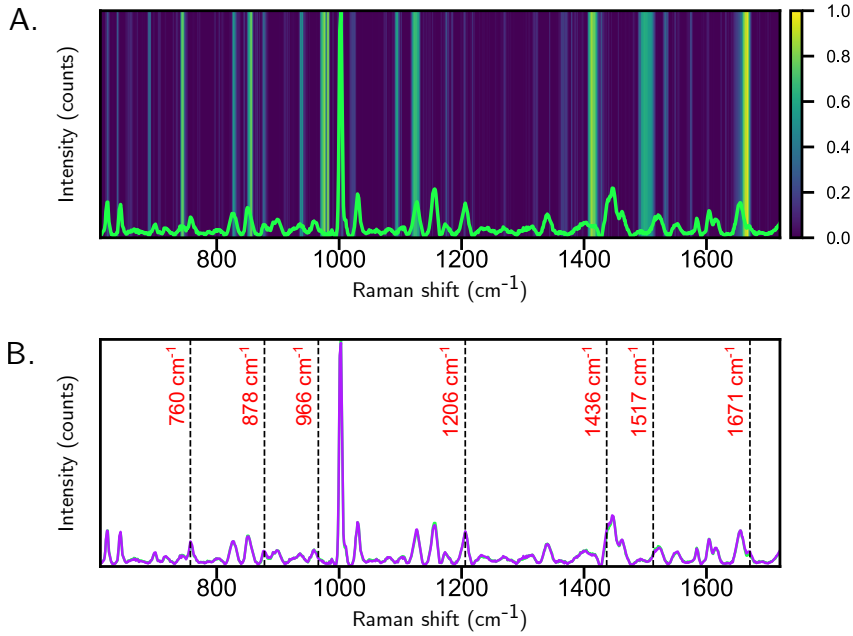


Figure 1.25: (A) Heatmap illustrating the frequency of selected features across feature selection methods. (B) Features selected using a window-based approach with regression coefficients, focusing on key areas that differentiate the groups. In both panels, purple represents lung cancer patients, and green indicates healthy controls.

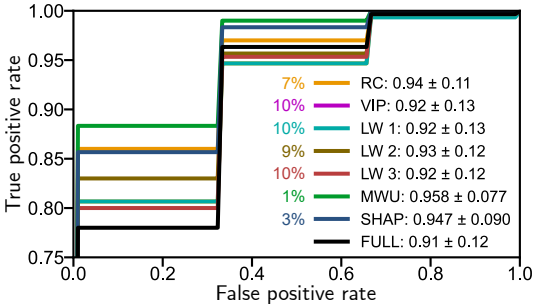


Figure 1.26: Performance assessment based on feature selection methods via ROC curves. The percentages represent the proportion of selected data, also explained in Figure 1.21, where each method demonstrated high performance in terms of accuracy.

iterations. These specific frequencies of distinct molecular vibrations represent biochemically relevant spectral features. The efficiency is validated by classification metrics, specifically an accuracy of 0.813 ± 0.015 and an AUC score of 0.914 ± 0.127, utilizing only these selected bands. This approach demonstrates superior performance compared to traditional methods by focusing on biochemically relevant spectral regions while minimizing the influence of noise and non-informative spectral features.

Moreover, ROC analysis reveals a clear performance among the feature selection methods as seen in Figure 1.26. MWU achieved the highest (0.958 ± 0.077), indicating superior ability to distinguish between classes while maintaining both high sensitivity and specificity. This exceptional performance was achieved with only 1% of the total features, suggesting highly efficient feature selection. SHAP (0.947 ± 0.090, 3% features) and PLS RCs (0.94 ± 0.11, 7% features) demonstrated comparable discriminative power. Although SHAP achieves high AUC score, its standard error of ± 0.090 places it within a range that overlaps with other methods, which also exhibit similarly wide error margins. This overlap suggests that while SHAP's performance is strong, its high AUC score alone may not be as definitive in our dataset.

VIP scores (0.92 ± 0.13, 10% features) and loading weights (LW1: 0.92 ± 0.13, LW2: 0.93 ± 0.12, LW3: 0.92 ± 0.12, using 10%, 9%, and 10% features respectively) exhibited slightly lower values despite utilizing larger feature sub-

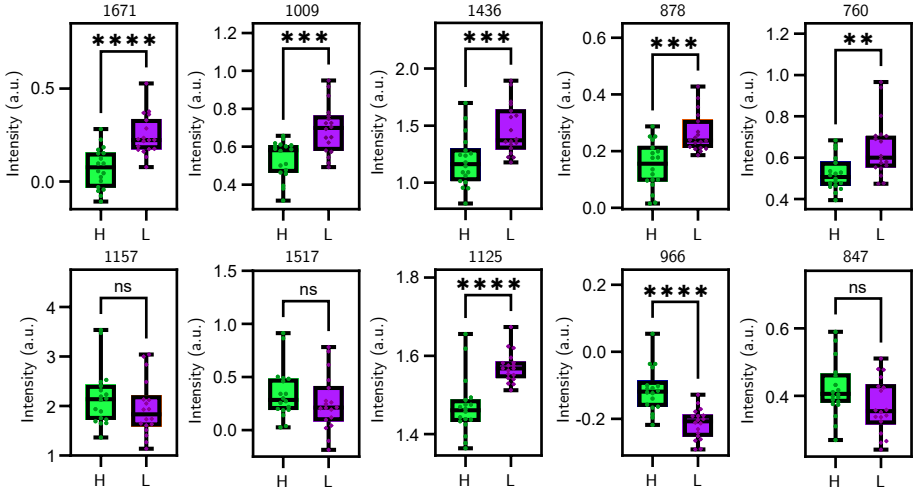


Figure 1.27: Distribution of key Raman bands across healthy controls (H) and lung cancer patients (L), displayed as box-and-whisker plots. Features are ranked by selection frequency. Statistical significance was determined using MWU test (GraphPad Prism 10.2.1). ns (non-significant) $p > 0.05$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

sets. While these scores still indicate strong classification performance - significantly above the random classifier baseline of 0.5 - the reduced AUC-ROC values suggest these methods may be more susceptible to false positive classifications or may miss some true positive cases. This performance becomes particularly relevant in spectroscopic analysis, where false positives could lead to misidentification of chemical signatures. High AUC-ROC values (>0.92) across all methods demonstrate robust classification performance, with each method significantly outperforming random classification. The superior performance of MWU with only 1% feature set suggests it captures the most discriminative spectral features while maintaining optimal sensitivity and specificity in the classification task.

Figure 1.27 and Table 1.2 reveal several selected features as explained in Section 1.3.3, with statistically significant Raman bands, suggesting their potential relevance as disease biomarkers for lung cancer.

Raman bands around 966 cm^{-1} and 1125 cm^{-1} emerge as particularly no-

Table 1.2: Analysis of the 10 most stable and prominent Raman bands associated with lung cancer.

Feature Count	Band Position (cm^{-1})	p-value	Effect Size	Correlation Coefficient
2858	1671	1.06e-04	-1.56e+00	-1.99e-01
2673	1009	7.53e-04	-1.35e+00	1.46e-01
2488	1436	9.46e-04	-1.21e+00	-4.34e-01
2386	878	5.97e-04	-1.45e+00	-1.14e-02
2240	760	2.52e-03	-1.08e+00	4.64e-02
2050	1157	1.59e-01	3.70e-01	3.95e-01
1829	1517	2.61e-01	3.48e-01	4.01e-01
1797	1125	6.27e-05	-1.68e+00	-1.23e-01
1761	966	2.76e-05	1.70e+00	3.68e-01
1597	847	1.03e-01	6.20e-01	-2.45e-01

table, both displaying high selection frequency and strong statistical relevance with large effect sizes and strong correlations. Effect size, often referred to as the standardized mean difference or Cohen's d, plays a critical role in clarifying the magnitude of difference in spectral intensity between the groups. Here, values of 1.70 and -1.68 for the 966 cm^{-1} and 1125 cm^{-1} bands respectively highlight substantial biological distinctions between these groups. While p-values convey whether a difference is statistically significant, they don't inform us about the size or importance of that difference. Reporting both values provides a more complete picture: p-values help confirm that observed differences are unlikely to be due to chance, while effect sizes reveal that these differences are not only statistically significant but practically meaningful. Also, correlation coefficients, ranging from -1 to 1, measure the direction and strength of the linear relationship between spectral intensities and disease status. Positive correlation (0.368) at 966 cm^{-1} suggests increased spectral intensity while negative correlation (-0.123) at 1125 cm^{-1} indicates decreased intensity. This pattern is consistent across other significant bands, such as 1436 cm^{-1} showing a strong negative correlation (-0.434) with an effect size of -1.21, suggesting systematic biochemical alterations. Extremely low p-values ($p \leq 0.0001$) for these bands establish the statistical reliability of these observations, minimizing the likeli-

hood that these differences occurred by chance.

Other bands, despite frequently selected, exhibit higher p-values, such as 1157 cm^{-1}, 1517 cm^{-1} and 847 cm^{-1} both with $p > 0.05$. Effect sizes for these bands, 0.37, 0.34 and 0.62 respectively, imply only minor differences with less statistical certainty. Additionally, their high correlation coefficients do not provide strong diagnostic contrast, suggesting limited utility as biomarkers. These bands, despite being selected frequently, may not hold diagnostic value and may require further evaluation to determine their utility.

Table 1.2 highlights variability in effect sizes and correlation coefficients across selected bands, supporting trends associated with cancer. For instance, while the correlation for 1009 cm^{-1} (0.146) is modest, it suggests a possible positive trend, whereas 878 cm^{-1} shows near-zero correlation (-0.0114), indicating limited association. The bands with high effect sizes and low p-values, such as 966 cm^{-1}, 1125 cm^{-1} and 1671 cm^{-1} appear most indicative of molecular changes linked to lung cancer. Conversely, bands like 1157 cm^{-1} and 1517 cm^{-1} show high selection frequency but may require further validation due to their higher p-values.

1

1.4.3 Publication 3: Data fusion strategies

The fusion of Raman and FTIR spectral data provides a bi-modal approach to enhance spectroscopic analysis. Each technique offers complementary insights: Raman spectroscopy focuses on changes in polarizability, while FTIR detects dipole moment alterations. In Figure 1.28A-B, mean spectra and standard errors for both Raman and FTIR highlight key vibrational features. Variations are evident, especially in spectral regions known to correspond to molecular vibrations associated with lung cancer.

Figure 1.28C-D display RCs of PLS used in feature selection prior to block scaling, showing different contribution due to their variances. This imbalance could have led to a biased feature selection, where one dataset would dominate the results. To address this issue, after the datasets were concatenated, block scaling was applied prior to feature selection, adjusting the variance between Raman and FTIR blocks to ensure comparable contributions from both meth-

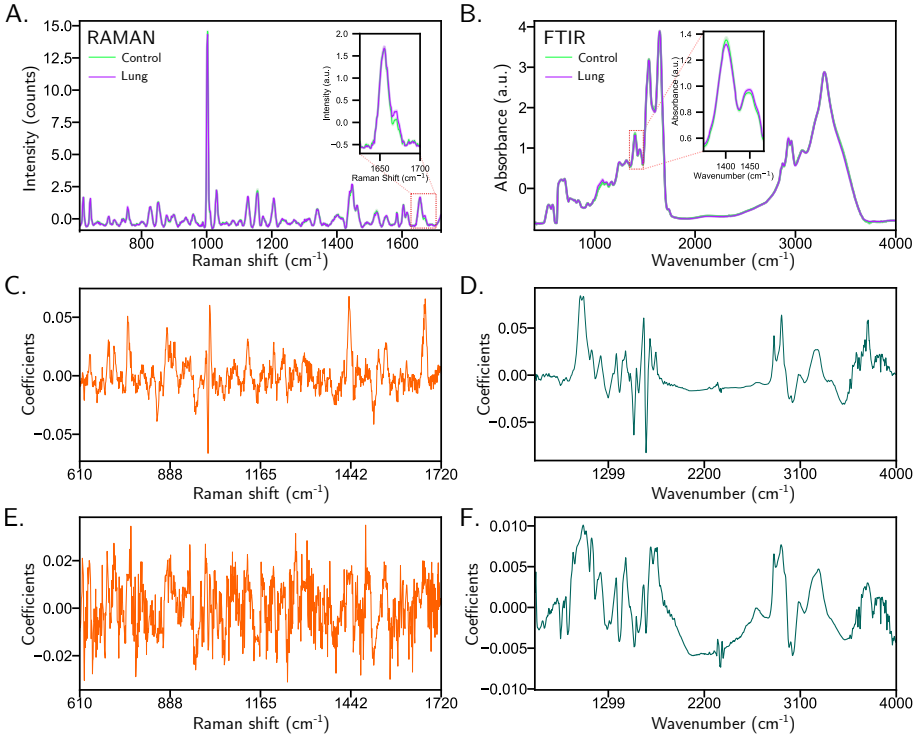


Figure 1.28: (A-B) Raman and FTIR spectra displaying the mean and standard error, alongside corresponding regression coefficients before (C-D) and after (E-F) block scaling, demonstrating the influence of this scaling technique on spectral analysis. The color scheme distinguishes between data types: Raman data is shown in orange (C, E), and FTIR data in dark cyan (D, F).

ods. After scaling, as shown in Figure 1.28E-F, RCs indicate more balanced influence from both methods. In this context, block scaling serves to prevent one spectral dataset from overshadowing the other.

The apparent fluctuations in the RCs observed in Figure 1.28C-E may be attributed to the inherent multicollinearity in spectroscopic data. Spectral variables frequently exhibit linear correlations due to overlapping molecular vibrations, structured signal patterns or instrumental resolution limits, which leads to noise-like fluctuations in RCs. This multicollinearity sometimes makes it challenging to precisely determine which specific spectral features are most strongly related to the classification outcome, as neighboring wavelengths exhibit similar behavioral patterns. When multiple variables are highly correlated, the regression model fails to distinguish their individual contributions, potentially leading to unstable coefficient estimates that appear noisy. This effect persists both before and after scaling, although scaling helps balance the overall contribution between Raman and FTIR data blocks. The presence of experimental noise, combined with the inherent multicollinearity of spectroscopic data, further complicates the interpretation of loading patterns without machine learning, particularly in regions where spectral peaks overlap.

Figure 1.13E, based on FTIR data, shows a distinct clustering pattern similar to that seen in Raman-based score plot. However, FTIR exhibits slightly different discrimination patterns due to specific vibrational modes. These differences enable FTIR to capture molecular signatures that Raman may overlook or represent differently, resulting in distinct cluster positions and separations. It is important to note that these positions are somewhat arbitrary, as the placement of each point can shift depending on the choice of PCs used for the plot.

Figure 1.13F represents the fusion of both Raman and FTIR analysis, where data are fused and block-scaled. The complementary nature of both techniques provides more robust and comprehensive complementary analysis, capturing different aspects of the molecular composition. While these score plots represent simplified visualizations of complex multidimensional data, they demonstrate high discrimination achieved through the fusion of multi-spectroscopic techniques.

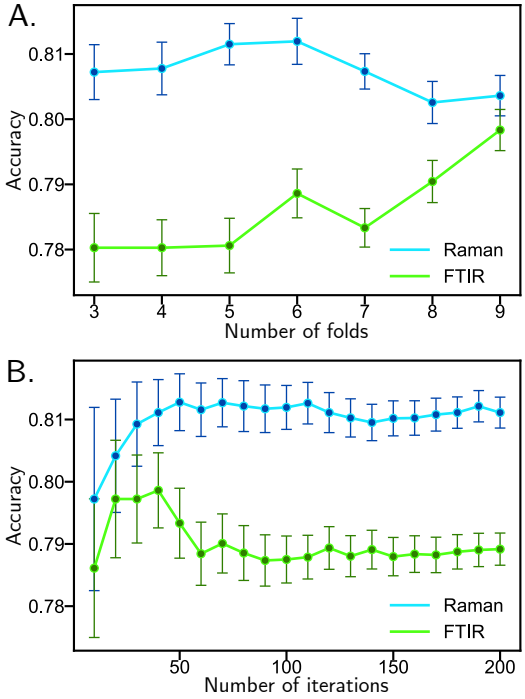


Figure 1.29: Parameter optimization for model assessment: classification accuracy versus (A) number of folds and (B) number of iterations.

Advancing from various predictive models and feature selection methods, a comprehensive strategy emerges in this last phase through data fusion as detailed in Section 1.3.3. The combined datasets from Raman and FTIR offer a unified molecular perspective, refining analytical outcomes and increasing model robustness. Data fusion is organized into three primary types: LLDF directly links full datasets, covering all measured variables; MLDF reduces dimensionality or feature selection prior to merging, retaining critical features while simplifying data complexity; and high-level fusion merges model predictions from each modality, enhancing predictive strength by combining distinct yet complementary spectroscopic insights.

The optimization process was continued through number of cross-validation folds and number of iterations as seen in Figure 1.29. The relationship between

accuracy and increasing fold numbers demonstrates varying optimal points for different methods. Six folds were selected as a balanced compromise to ensure consistent validation across both analyses, enabling direct performance comparison between techniques. Similarly, the iteration analysis reveals how model performance converges with repeated test runs, where 100 iterations were chosen based on the observed stabilization of accuracy metrics, with additional iterations producing negligible improvements in model robustness.

Table 1.3: Model performance of Raman and FTIR without data fusion. Full: full range for FTIR (400-4000 cm^{-1}), FP: fingerprint region for FTIR (400-1800 cm^{-1}), FS: feature selection, FR: feature reduction.

Option	Data Fusion	Spectral Range for FTIR	Method	Data Range	Number of Features	Accuracy
1	No fusion	Full	1. Raman	100%	1015	0.8119±0.0035
			2. Raman (FS)	5%	51	0.8539±0.0056
			3. Raman (FR)	6PCs	6	0.8378±0.0060
			4. FTIR	100%	1868	0.7886±0.0037
			5. FTIR (FS)	4%	75	0.8425±0.0058
			6. FTIR (FR)	8PCs	8	0.7928±0.0068
		FP	7. FTIR	100%	727	0.7567±0.0033
			8. FTIR (FS)	1%	8	0.8419±0.0057
			9. FTIR (FR)	5PCs	5	0.7633±0.0068

Table 1.3 outlines individual performances of Raman and FTIR spectroscopy. For FTIR, the full range analysis shows that FS substantially improves accuracy from 0.79 to 0.84 with only 75 features (4% selection). FP range achieves a similar increase, where selecting just 8 features (1%) improves accuracy from 0.76 to 0.84. Conversely, FR for FTIR yield limited improvement, with accuracies of 0.79 and 0.76 in the full and FP ranges, respectively. Raman spectroscopy (610-1720 cm^{-1}), beginning at 0.81 accuracy with 1015 features, also benefits from FS, achieving 0.85 with only 5% of the data (51 features), while FR maintains a strong 0.84 accuracy with only 6 PCs. FS demonstrates higher performance by identifying and retaining only the most discriminative original spectral bands while completely eliminating irrelevant features, whereas FR transforms the data into the components that may obscure distinctive spectral markers crucial for classification.

Table 1.4 reveals the impact of (LLDF) on Raman and FTIR data. For the

Table 1.4: Model performance of Raman and FTIR with low level data fusion.

Option	Data Fusion	Spectral Range for FTIR	Method	Data Range	Number of Features	Accuracy
2	LLDF	Full	10. Raman + FTIR	100%	2883	0.8625±0.0035
			11. Raman + FTIR + FS	6%	173	0.9922±0.0015
			12. Raman + FTIR + FR	6PCs	6	0.8711±0.0034
		FP	13. Raman + FTIR	100%	1742	0.8592±0.0037
			14. Raman + FTIR + FS	10%	175	0.9497±0.0039
			15. Raman + FTIR + FR	5PCs	5	0.8681±0.0031

full range, LLDF yields an initial accuracy of 0.86 using 2883 combined features, with FS significantly enhancing accuracy to 0.99 with only 6% (173 features). FS to retain only the most discriminative spectral bands from each technique results in exceptionally high classification performance. FR, though yielding fewer components (6 PCs), leads to a smaller accuracy increase, reaching 0.87. In the FP range, LLDF achieves 0.86 with 1742 features. FS here improves performance to 0.95, using 10% of features (175 selected), while FR yields a modest increase to 0.87 with only 5 PCs.

Table 1.5: Model performance of Raman and FTIR with mid level data fusion.

Option	Data Fusion	Spectral Range for FTIR	Method	Data Range	Number of Features	Accuracy
3	MLDF	Full	16. Raman (FS) + FTIR (FS)	5% + 4%	126	0.8472±0.0039
			17. Raman (FR) + FTIR (FR)	6PCs + 8PCs	14	0.8425±0.0034
		FP	18. Raman (FR) + FTIR (FR)	5% + 1%	59	0.7972±0.0035
			19. Raman (FR) + FTIR (FR)	6PCs + 5PCs	11	0.8583±0.0032

MLDF also revealed distinct patterns across spectral ranges. As shown in Table 1.5. MLDF with FS at 5% for Raman and 4% for FTIR (Full) achieves a high accuracy of 0.85 while reducing the feature set to 126. FR with 6 PCs for Raman and 8 PCs for FTIR (Full) yields a comparable accuracy of 0.84, further reducing the feature count to just 14. Notably, FS outperforms FR slightly, suggesting it may be more effective for maintaining accuracy in MLDF within the full range. MLDF using FS (5% Raman and 1% FTIR (FP)) yields a moderate accuracy of 0.80 with 59 features. In contrast, FR (6 PCs for Raman and 5 PCs for FTIR) enhances accuracy to 0.86 with only 11 features, indicating that dimensionality reduction in FR is particularly advantageous for FP range.

While MLDF showed some improvement in model performance compared to individual spectroscopic techniques, the inconsistent results between FS and FR suggest that processing each spectral dataset separately before fusion may lead to loss of important correlations between the complementary techniques.

Table 1.6: Model performance of Raman and FTIR with high level data fusion.

Option	Data Fusion	Spectral Range for FTIR	Method	Data Range	Number of Features	Accuracy
4	HLDF	Full	20. Raman + FTIR	100%	1015 / 1868	0.8383±0.0024
			21. Raman (FS) + FTIR (FS)	5% / 4%	51 / 75	0.8131±0.0023
			22. Raman (FR) + FTIR (FR)	6PCs / 8PCs	6 / 8	0.8383±0.0024
		FP	23. Raman + FTIR	100%	1015 / 727	0.8319±0.0024
			24. Raman (FS) + FTIR (FS)	5% / 1%	51 / 8	0.7989±0.0035
			25. Raman (FR) + FTIR (FR)	6PCs / 5PCs	6 / 5	0.8319±0.0024

In the final analysis, HLDF was evaluated as shown in Table 1.6. Raman with FTIR (Full) achieves an accuracy of 0.84. FS, which reduces Raman to 51 features (5%) and FTIR (Full) to 75 features (4%), slightly decreases accuracy to 0.81. However, FR maintains 0.84 accuracy with only 6 PCs from Raman and 8 from FTIR (Full), indicating that FR more effectively preserves critical information in HLDF than FS. However, Raman with FTIR (FP) achieves 0.83. FS selects the most prominent features, 51 and 8 from Raman and FTIR (FP), respectively, but reduces accuracy to 0.80. FR proves more effective, maintaining the accuracy of 0.83 with 6 Raman and 5 FTIR (FP) PCs. These results show that FR better preserves the complementary information in HLDF compared to FS.

Expanding from feature selection methods to data fusion strategies also marks a significant enhancement in classification task, merging distinct yet complementary diagnostic features from both Raman and FTIR spectral datasets. This transition from single-method to multi-method addresses the inherent limitations of individual method by combining data from another analytical method with additional molecular information. The following analysis, as illustrated in Figure 1.30, demonstrates how integrating both data sources improves model performance based on AUC-ROC scores, underscoring the advantage of multi-modal fusion in capturing a more comprehensive diagnostic profile.

Figure 1.30 provides a comparative overview across different analytical tech-

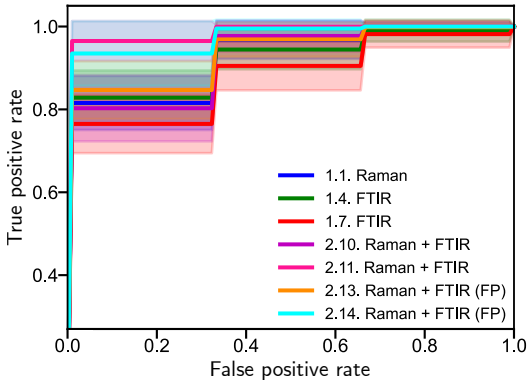


Figure 1.30: AUC-ROC scores for the selected data fusion configurations, corresponding to Table 1.3 and Table 1.4.

niques. Raman spectroscopy achieves 0.924 ± 0.030 while FTIR (Full) reaches almost similar AUC of 0.918 ± 0.033. However, FTIR (FP) demonstrates a slightly reduced but still substantial value with 0.881 ± 0.039. LLDF of Raman with FTIR (Full) maintains comparable performance, 0.923 ± 0.030, but when enhanced through FS, the fused set achieves 0.983 ± 0.016. Similarly, the same method, now with FTIR (FP), yields an AUC of 0.934 ± 0.030, which further improves to 0.972 ± 0.029 when FS is applied in LLDF. These results suggest that, while each individual method is effective on its own, fusing Raman and FTIR data—especially when FS is applied—increases discriminative power and enhances diagnostic precision.

Moreover, Figure 1.31 illustrates the selected features derived from LLDF because of their high model performance, as shown in the Table 1.4. This figure is based on a systematic approach that applied recursive feature elimination (RFE) to an initial set of 173 features, ultimately refining the selection to the 20 most impactful ones. These selected features represent the most predictive spectral bands, as determined by combining Raman and FTIR. Box-and-whisker plots provide a focused view of specific biochemical groups and blend information from multiple data sources, enhancing the discriminatory power of specific features possibly linked to lung cancer. Data fusion approaches in spectroscopy takes into account the complementary nature of spectral datasets from differ-

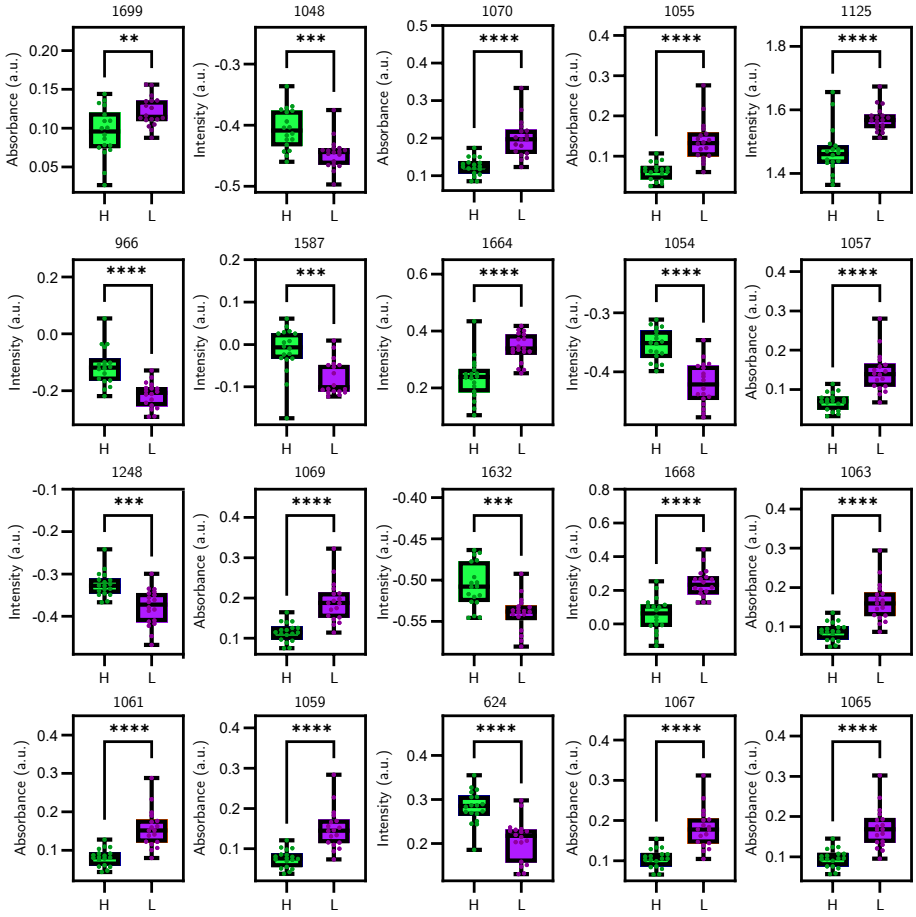


Figure 1.31: Box-and-whisker plots showing the distribution of the top 20 features based on feature importance, as determined through data fusion analysis. All features exhibit significant differences between healthy controls (H) and lung cancer patients (L) (MWU test, ns (non-significant) $p > 0.05$,* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.). Analyses were performed using GraphPad Prism version 10.2.1 for Windows.

ent analytical methods by capturing subtle yet consistent spectral variations in these methods.

It is evident that several bands exhibit strong discriminations between the groups, marked by high statistical significance in Figure 1.31. For instance, 1055–1070 cm^{-1} and 1699 cm^{-1} from FTIR; 1125 cm^{-1}, 1587 cm^{-1}, and 1632–1668 cm^{-1} from Raman show high significance, suggesting these bands are especially pertinent in identifying lung cancer-associated biochemical alterations because significant features stand for their potential as strong biomarkers. The comparison highlights the variability in spectral response, pointing to the most robust bands for diagnostic application.

1.4.4 Vibrational frequencies and bands assignment

This section delivers a comprehensive biochemical interpretation of Raman and FTIR spectral data as shown in Table 1.7, focusing on vibrational assignments for each strategy as previously explained in Section 1.3.3. Each strategy is designed to capture an extensive range of vibrational information and its biochemical significance, thereby deepening our understanding of lung cancer.

Machine learning models in Section 1.3.3 identified key vibrational groups associated with lung cancer. Peaks in Figure 1.15A at 619–624 cm^{-1} and 641–643 cm^{-1}, corresponding to τ(CC) in phe and tyr, were captured across different principal components. The peak at 619–624 cm^{-1} appears in PC1, PC2, and PC5 with a regression coefficient of -0.021, while the peak at 641–643 cm^{-1} is only present in PC1 and PC5 with no significant regression coefficient. This suggests that although these vibrational modes contribute to high variance in the spectral data, they have limited direct influence on classification. The $\delta(CH_{ring})$ of trp (756–758 cm^{-1}), present in PC1, PC2, and PC5, exhibited a regression coefficient of 0.051, suggesting a moderate positive association with cancer detection. In contrast, the $\delta(CH_{ring})$ of tyr (822 cm^{-1}), isolated to PC5 (regression coefficient: -0.021), indicated subtle variance with a weak negative correlation to lung cancer. The ring breathing mode of phe (1000–1004 cm^{-1}), spanning PC1, PC2, and PC5, showed a stronger negative regression coefficient (-0.067), underscoring its suppression in cancerous samples.

Table 1.7: Raman and FTIR spectral band assignments for human blood plasma[a] [9, 18, 52–57]

Raman shift (cm^{-1})	**Raman assignment**	**Wavenumber (cm^{-1})**	**FTIR assignment**
619–624	τ(CC): phe	579	γ(C = O): amide VI
641–643	τ(CC): tyr	700	δ(OCN), γ(NH): amide IV, amide V
698–701	ν(CS): met	833–1078	ν(CO), $\nu(PO_2^-)$
756–758	$\delta(CH)_{ring}$: trp	1169	ν(CO): esters
822	$\delta(CH)_{ring}$: tyr	1246–1315	ν(CN), δ(NH): amide III
855–856	$\delta(CH)_{ring}$: tyr	1400–1450	$\delta(CH_3)$: proteins, side chains
874–878	ν(CN), ν(CC): proteins, lipids; trp, arg	1537	ν(CN), δ(NH): amide II
897–901	ν(C − O − C)	1651–1699	ν(C = O): amide I
939	$\nu(NC_\alpha C)$, ν(CC), ν(CO): α-helix, lipids, glycogen	1740	ν(C = O): lipids
966	ν(CC), ν(CN), $\delta(CH_2)$	2868	$\nu(CH_3)$: proteins, lipids
1000–1004	Ring breathing: phe	2930	$\nu(CH_2)$: proteins, lipids
1029–1033	δ(CH): phe	2961	$\nu(CH_3)$: proteins, lipids
1048–1054	δ(= CH), ν(CC), ν(CO): proteins (phe), collagen, glycogen	3071	ν(NH): amide B
1104	ν(CC): the gauche bonded chain	3294	ν(NH): amide A, ν(OH)
1123–1127	ν(CC), ν(CO), ν(CN): lipids, glycogen, proteins		
1156–1157	ν(CC), ν(CO), δ(CH): lipids, glycogen, proteins (tyr)		
1204–1210	ν(CPh): trp, phe		
1232–1269	δ(= CH), amide III: lipids (unsat. fa), proteins (α-helix)		
1397–1404	$\nu(COO^-)$: asp, glu		
1436–1438	$\delta(CH_3)$, $\delta(CH_2)$: proteins (aliph. AA), lipids		
1513–1528	ν(C = C): carotenoids		
1548–1553	ν(C = C): trp		
1587–1589	ν(C = C): phe, trp		
1604–1606	ν(C = C): phe, trp		
1619	ν(C = C): tyr, trp		
1666–1671	Amide I: α-helix		

[a] phe: Phenylalanine; tyr: Tyrosine; met: Methionine; trp: Tryptophan; arg: Arginine; glu: Glutamic acid; asp: Aspartic acid.

Protein secondary structures were highlighted by the amide III band (1232–1269 cm^{-1}, PC5; regression coefficient: 0.023, 0.014) and the amide I region (1666–1671 cm^{-1}, PC2/PC5; 0.066), the latter strongly linked to α-helix conformations. Lipid-related features, such as ν(CC) (1123–1127 cm^{-1}, PC2/PC5; 0.031) and ν(C = C) of carotenoids (1513–1528 cm^{-1}, PC1/PC2/PC5; −0.042), revealed altered lipid metabolism, with carotenoids exhibiting a negative classification influence. Carbohydrate and hydrocarbon changes were marked by ν(C − O − C) modes (897–901 cm^{-1}, PC2/PC5; none) which was excluded from regression due to weak discriminative power. Notably, $\nu(NC_{\alpha}C)$, ν(CC), ν(CO) in α-helix (939 cm^{-1}, PC5; none) and ν(C = C) in phe, trp (1587–1589 cm^{-1}, PC2/PC5; −0.022) were uniquely captured by specific PCs, emphasizing their localized variance contributions. Modes with none in regression coefficients, such as ν(CS) in met (698–701 cm^{-1}) or ν(C = C) in tyr, trp (1619 cm^{-1}), lacked statistical relevance in supervised classification, despite their variance in unsupervised PCA.

Feature selection methods described in Section 1.3.3 enable a deeper investigation of vibrational groups, also including those that may be less prominent, but still hold significant relevance in identifying lung cancer. The analysis incorporates multiple selection approaches, including PLS-DA, VIP scores, Shapley values, and statistical tests, to determine the most relevant spectral features (see Figure 1.27 and Table 1.2). Notably, the bands at 1125 cm^{-1}, 966 cm^{-1}, and 1671 cm^{-1} emerge as key spectral markers, exhibiting statistically significant alterations between healthy and lung cancer samples. The peak at 1125 cm^{-1}, assigned to ν(CC), ν(CO), ν(CN) in lipids, glycogen, and proteins, demonstrates a substantial effect size (−1.68) and negative correlation, indicating biochemical modifications associated with the disease. Similarly, the 966 cm^{-1} band, linked to ν(CC), ν(CN), $\delta(CH_2)$, shows a strong positive effect size (1.70) and correlation with lung cancer. The amide I band at 1671 cm^{-1}, representative of α-helix conformations, exhibits a pronounced reduction, reflecting structural protein changes. Additional relevant features include $\delta(CH_3)$, $\delta(CH_2)$ at 1436 cm^{-1}, ν(CN), ν(CC) at 878 cm^{-1}, and ring breathing modes of phe (1009 cm^{-1}) and tyr (847 cm^{-1}), all contributing to alterations in protein conformation. Window-based feature selection further refines these findings, emphasizing spectral regions such as 1436 cm^{-1} ($\delta(CH_3)$, $\delta(CH_2)$: proteins (aliph. AA),

lipids), 1514 cm^{-1} (ν(C = C): carotenoids), 760 cm^{-1} ($\delta(CH)_{ring}$: trp), 1206 cm^{-1} (ν(CPh): trp, phe), 1671 cm^{-1} (Amide I: α-helix), 878 cm^{-1} (ν(CN), ν(CC): proteins, lipids; trp, arg) and 966 cm^{-1} (ν(CC), ν(CN), $\delta(CH_2)$), highlighting their potential role in capturing the molecular heterogeneity of lung cancer. These vibrational features, identified through a systematic selection process over 500 iterations, provide a detailed spectral profile that differentiates between healthy and cancerous states, reinforcing their biochemical significance within the dataset.

Data fusion strategies, as explained in Section 1.3.3, were essential for integrating the complementary vibrational information from Raman and FTIR spectroscopy. Among the fusion strategies, LLDF with FS demonstrated the highest model performance, as indicated in Figure 1.31 and Table 1.4, achieving an accuracy of 0.9922 ± 0.0015, effectively identifying the most discriminative spectral markers. Although MLDF also incorporated FS, LLDF provided superior classification accuracy by capturing the most relevant vibrational features. In LLDF, FS was applied to extract the most informative spectral markers while maintaining the structural integrity of the fused dataset. For Raman spectroscopy, the key vibrational bands included 624 cm^{-1} (τ(CC) in phe), 966 cm^{-1} (ν(CC), ν(CN), $\delta(CH_2)$), 1048-1054 cm^{-1} (δ(= CH), ν(CC), ν(CO) in proteins (phe), collagen, glycogen), 1125 cm^{-1} (ν(CC), ν(CO), ν(CN) in lipids, glycogen, proteins), 1248 cm^{-1} (δ(= CH) and amide III in lipids and proteins (α-helix)), 1587 cm^{-1} (ν(C = C) in phe, trp), and 1632-1668 cm^{-1} (amide I: α-helix). Similarly, in FTIR spectroscopy, the most critical spectral markers were 1055-1070 cm^{-1} (ν(CO), $\nu(PO_2^-)$) and 1699 cm^{-1} (ν(C = O) in amide I), indicating significant protein alterations. The integration of these Raman and FTIR markers provided a comprehensive spectral profile capable of distinguishing lung cancer from healthy states with high accuracy. Raman markers captured essential protein and lipid modifications across multiple vibrational modes, while FTIR markers highlighted crucial changes in molecular bonding within proteins, particularly in the fingerprint region. This fusion of vibrational data reinforced the diagnostic power of the combined spectroscopic approach, offering a robust biochemical fingerprint of lung cancer that encompasses a wide range of molecular structural changes associated with the disease state.

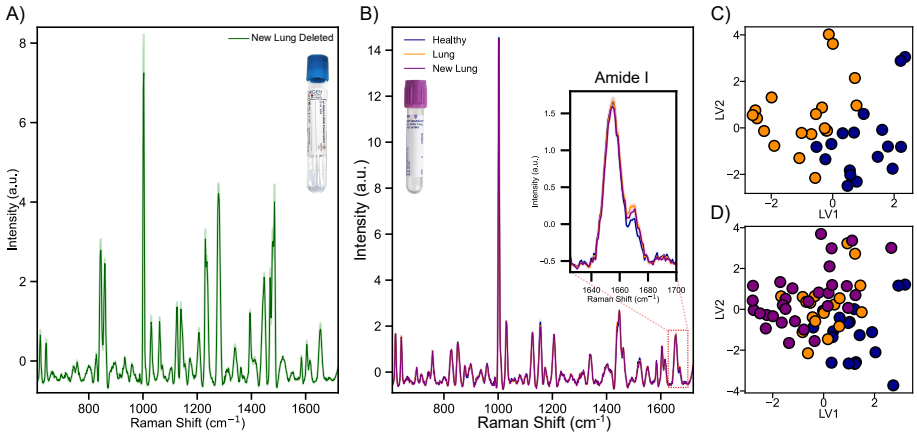


Figure 1.32: (A) Mean Raman spectra (green) of samples collected in cf-DNA/cf-RNA preservative tubes, later discarded due to inconsistent anticoagulant effects. Shaded regions indicate standard errors. (B) Mean Raman spectra of healthy (blue), previous lung cancer patients (orange), and newly registered lung cancer (purple) samples, all collected in BD Vacutainer EDTA tubes. The inset highlights the Amide I region. (C) Score plot showing group separation between healthy and lung cancer patients. (D) Score plot including newly registered samples, indicating that the new cohort clusters similarly to the previous lung cancer group.

1.4.5 Model validation with newly registered patients

This research began with the first 18 lung cancer patients registered in the hospital. Over time, our dataset expanded to include 60 newly acquired samples. However, 30 of these had to be excluded in the analysis because they were drawn using specialized cf-DNA/cf-RNA preservative tubes that contain proprietary anticoagulants. To avoid confounding spectral differences stemming from tube chemistry rather than disease state, only the remaining 30 samples collected in BD Vacutainer EDTA tubes were retained for analysis. Figure 1.32A illustrates the discarded spectra, highlighting the significant variation likely introduced by different tube formulations, while Figure 1.32B shows spectra acquired from BD Vacutainer EDTA tubes used consistently throughout the study.

In Figure 1.32A, the mean spectra of human blood plasma samples collected in the cf-DNA/cf-RNA preservative tubes show noticeably broader sig-

nal variability, as reflected in the broader shaded error regions. This suggests that proprietary anticoagulants or other tube additives might be altering the biochemical composition of the samples or introducing additional noise. Conversely, Figure 1.32B highlights the newly registered patients (purple) whose blood was collected in BD Vacutainer EDTA tubes—the same type used for both the healthy (blue) and initial lung cancer (orange) patients. Here, the mean spectra and their small standard errors are more closely aligned with those seen in earlier datasets, indicating better reproducibility and less variance. By retaining only these EDTA-based samples, we minimize technical biases and help ensure that any observed spectral differences are linked to the disease rather than blood collection methods.

Figures 1.32C and 1.32D show score plots that help visualize how the newly registered patients compare with the existing cohorts. In Figure 1.32C, healthy controls (blue) are clearly separated from the original lung cancer patients (orange). When the new patients (purple) are introduced in Figure 1.32D, they cluster mainly in the same region as the original lung cancer group, reflecting the preservation of the core spectral differences initially identified even when new data are introduced, providing additional support for the model's generalizability. This outcome suggests that the model effectively accommodates new data and highlights the consistency of these biochemical markers in a growing patient population

To assess how well each classification model generalizes to newly registered patients, we first fit and parameter-optimize the model on the existing old dataset. Next, the new data are transformed using the same way, and prediction is performed with the parameters obtained from the training step (Fig. 1.33). Moreover, several dimensionality reduction and feature selection strategies were tested. Figure 1.33A illustrates the application of PCA as a dimensionality reduction step before classification: the old set is transformed into a lower-dimensional space (principal components), and the new set is projected onto the same space for prediction. Figure 1.33B follows a similar approach using PLS-DA, where latent variables replace principal components. In Figure 1.33C, the PLS-DA model is developed by extracting RCs and intercept terms from the old dataset and applying them directly to the new data. Lastly, Figure 1.33D

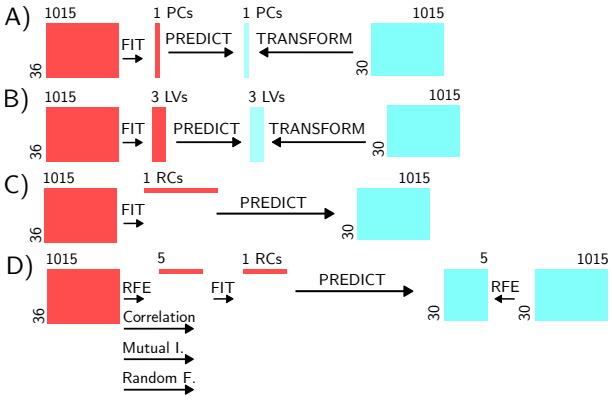


Figure 1.33: Schematic of the four workflows for evaluating newly registered samples with models trained on the old dataset: (A) PCA-based dimensionality reduction, (B) PLS-DA using latent variables, (C) direct use of regression coefficients from PLS-DA, and (D) multi-method feature selection prior to PLS-DA.

shows a procedure that begins with feature selection through a combination of methods—RFE, Spearman rank correlation (suitable for non-parametric data), mutual information (to capture potential non-linear relationships), and RF (to assess feature relevance)—before fitting a PLS-DA model to classify the new patients.

In practice, each workflow from Figure 1.33 was used to predict disease status in the 30 newly registered patients, all of whom were clinically confirmed as cancer cases. PCA-based approach correctly identified 18 of 30 patients as cancer. Employing PLS-DA led to 20 correctly diagnosed patients, while using the RCs from PLS-DA yielded 26 cancer identifications. Remarkably, the multi-step feature selection (RFE, Spearman rank correlation, mutual information, and RF) paired with the RCs from PLS-DA diagnosed all 30 patients as cancer by focusing on only five discriminative Raman bands. Here, applying feature selection enhances model performance by removing irrelevant or noisy features that obscure meaningful signals in both old and new dataset. By focusing on a smaller subset of highly discriminative features, classification becomes less prone to overfitting, as it learns from variables that are most indicative of the disease state.

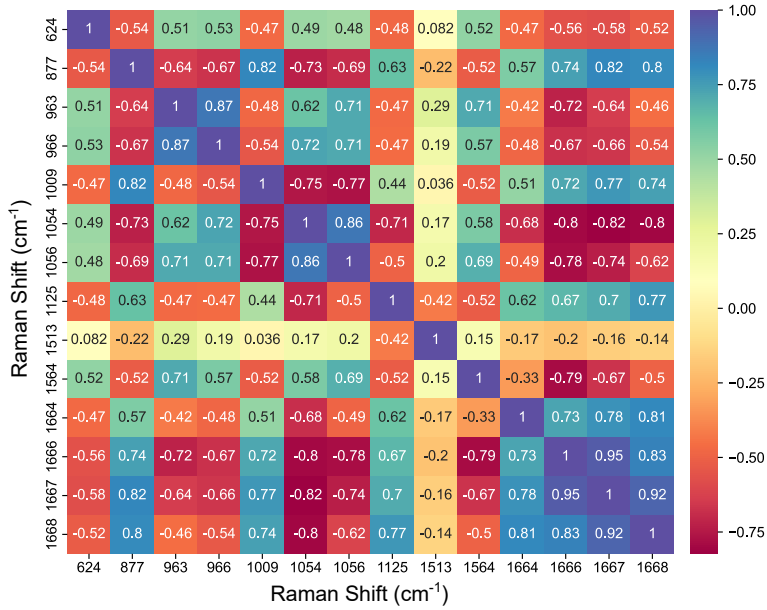


Figure 1.34: Correlation matrix of the selected Raman bands from the final feature set, where the color scale spans from -1 (strong negative correlation) to +1 (strong positive correlation).

After identifying the most discriminative bands through feature selection, correlation matrix was constructed to better understand the relationships among these spectral features. High positive correlation between any two Raman shifts often indicates they originate from similar biochemical bonds or vibrational modes within the sample. Conversely, if two adjacent shifts exhibit little or no correlation, this suggests that despite their close proximity on the spectra, they capture distinct molecular information. For instance, in Figure 1.34, the correlation between the bands at 1513 and 1564 cm^{-1} is relatively low, implying that these peaks, though close in spectral range, likely arise from different vibrational groups. In contrast, bands around 1664–1668 cm^{-1} show a strong correlation, suggesting they belong to closely related vibrational modes (e.g., the Amide I region). Overall, these observations confirm that the most predictive features in a classification model are not necessarily those that are spectrally adjacent, but rather those that capture the most diagnostically relevant chemical differences.

To assess how model performance changes with the addition of new dataset, four scenarios were tested. First, the model was trained only with the original 36 samples, consisting of 18 healthy controls and 18 lung cancer patients. Second, the dataset was expanded by adding 30 newly registered lung cancer patients, increasing the total number of samples to 66. With the larger dataset, the model could potentially learn more and improve its performance. Third, feature selection methods, as described previously, were performed on this expanded dataset to narrow down the most relevant bands, which reduced the feature size and led to better classification. Lastly, highly correlated spectral bands were averaged, specifically those in the ranges 963–966, 1054–1056, and 1664–1668 cm^{-1}, to avoid the redundancy that might inflate the performance.

Following the procedures described above, PLS-DA with 3 LVs was then applied to classify lung cancer in each of the four scenarios (Fig. 1.35). When using only the original 36 samples (18 healthy + 18 lung cancer) and 1015 features, the model achieved an accuracy of 0.8372 ± 0.0034. Expanding the dataset by adding 30 newly registered cancer patients (66×1015) raised the accuracy to 0.8600 ± 0.0020. Further applying feature selection (to reduce the matrix to 66×14) improved performance to 0.8739 ± 0.0015. Finally,

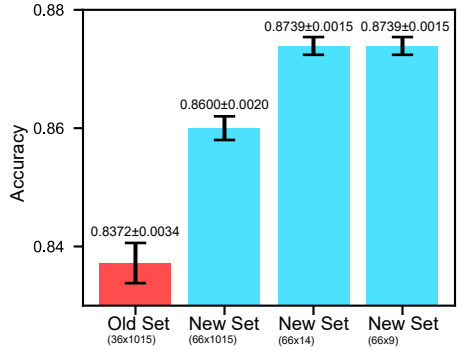


Figure 1.35: PLS-DA classification accuracy for four scenarios: (i) original dataset, (ii) expanded dataset with 30 additional lung cancer samples, (iii) expanded dataset after feature selection, and (iv) expanded dataset after averaging highly correlated features. Error bars represent standard errors. The final configuration maintains high accuracy while reducing feature redundancy.

averaging highly correlated Raman shifts (resulting in 66×9) produced the same accuracy of 0.8739 ± 0.0015. Although merging these correlated bands did not further increase accuracy, it removed redundancy and simplified the model while maintaining its diagnostic performance.

1

1.4.6 Biochemical analysis

A total of 30 newly registered lung cancer patients were evaluated in conjunction with Raman measurements described earlier. Their biochemical data consists of 32 parameters monitored for potential alterations associated with lung cancer (Figure 1.36). However, the figure includes data points from only 28 patients, as biochemical data are fully available for this subset. Each point is shown in either green or red. Green dots indicate values within the reference ranges provided on the top of each plot, whereas red dots denote measurements outside these intervals. A detailed list of all these parameters is provided below. The names ending with "_1" represent alternative reference categories used during clinical screening.

It is important to clearly mention that the last six biochemical parameters have two different reference ranges, as indicated in the parameter list below

and shown in Figure 1.36. Accordingly, the last six plots in the figure employ two sets of thresholds: green and blue markers represent values within their respective reference ranges, whereas red and orange markers indicate values outside these ranges. The remaining parameters have only a single reference range, represented by green markers for values within range and red markers for those outside the range.

The above-mentioned parameters are detailed below alongside the reference intervals indicated in parentheses:

- Glucose (70–110 mg/dL)
- Glomerular Filtration Rate (GFR) (>60 mL/min)
- Aspartate Aminotransferase (AST) (0–31 U/L)
- Alanine Aminotransferase (ALT) (0–33 U/L)
- Bilirubin (0.0–1.1 mg/dL)
- Calcium (8.8–10.2 mg/dL)
- Proteins (6.6–8.7 mg/dL)
- Albumin (Alb) (3.7–5.1 g/dL)
- Sodium (135–145 mEq/L)
- Chloride (93–110 mEq/L)
- Lactate Dehydrogenase (LDH) (135–250 U/L)
- Hemoglobin (Hb) (12–15.3 g/dL)
- Mean Corpuscular Volume (MCV) (80–97 fL)
- Mean Corpuscular Hemoglobin (MCH) (27–33 pg)
- Mean Corpuscular Hemoglobin Concentration (MCHC) (32–36 g/dL)
- Red Cell Distribution Width (RDW) (11.5–15.6%)

- Platelets ($140 - 400 \times 10^3$ µL)
- White Blood Cells (WBCs) ($3.8 - 10 \times 10^3$ µL)
- Neutrophils_1 ($1.6 - 7.5 \times 10^3$ µL)
- Lymphocytes (19–48%)
- Lymphocytes_1 ($0.9 - 3.5 \times 10^3$ µL)
- Monocytes (3.5–12%)
- Eosinophils (0.5–7%)
- Eosinophils_1 ($0 - 0.6 \times 10^3$ µL)
- Immature Granulocytes (0.0–0.5%)
- Immature Granulocytes_1 ($0.0 - 0.0 \times 10^3$ µL)
- Creatinine (Cr) (0.4–1 mg/dL) (0.7–1.20 mg/dL)
- Gamma-Glutamyl Transferase (GGT) (6–42 U/L) (10–71 U/L)
- Alkaline Phosphatase (ALP) (35–104 U/L) (40–129 U/L)
- Red Blood Cells (RBCs) ($4.3 - 5.6 \times 10^6$ µL) ($3.8 - 5 \times 10^6$ µL)
- Hematocrit (40–50%) (35–46%)
- Neutrophils (32–36%) (40–75%)

Parameters such as bilirubin, chloride, and monocytes fall within established reference ranges, as evidenced by the prevalence of green dots in Figure 1.36. This consistency suggests preserved physiological function in key areas: bilirubin reflects adequate hepatic detoxification [58], while chloride level indicates maintained electrolyte balance, essential for cellular and neuromuscular activity [59]. Similarly, renal markers—GFR and creatinine—predominantly remain within normal limits for most patients, underscoring sufficient kidney function to manage metabolic waste and potential treatment-related toxicity [60]. Monocytes, critical for innate immune responses, reinforces systemic stability in a subset of patients [61].

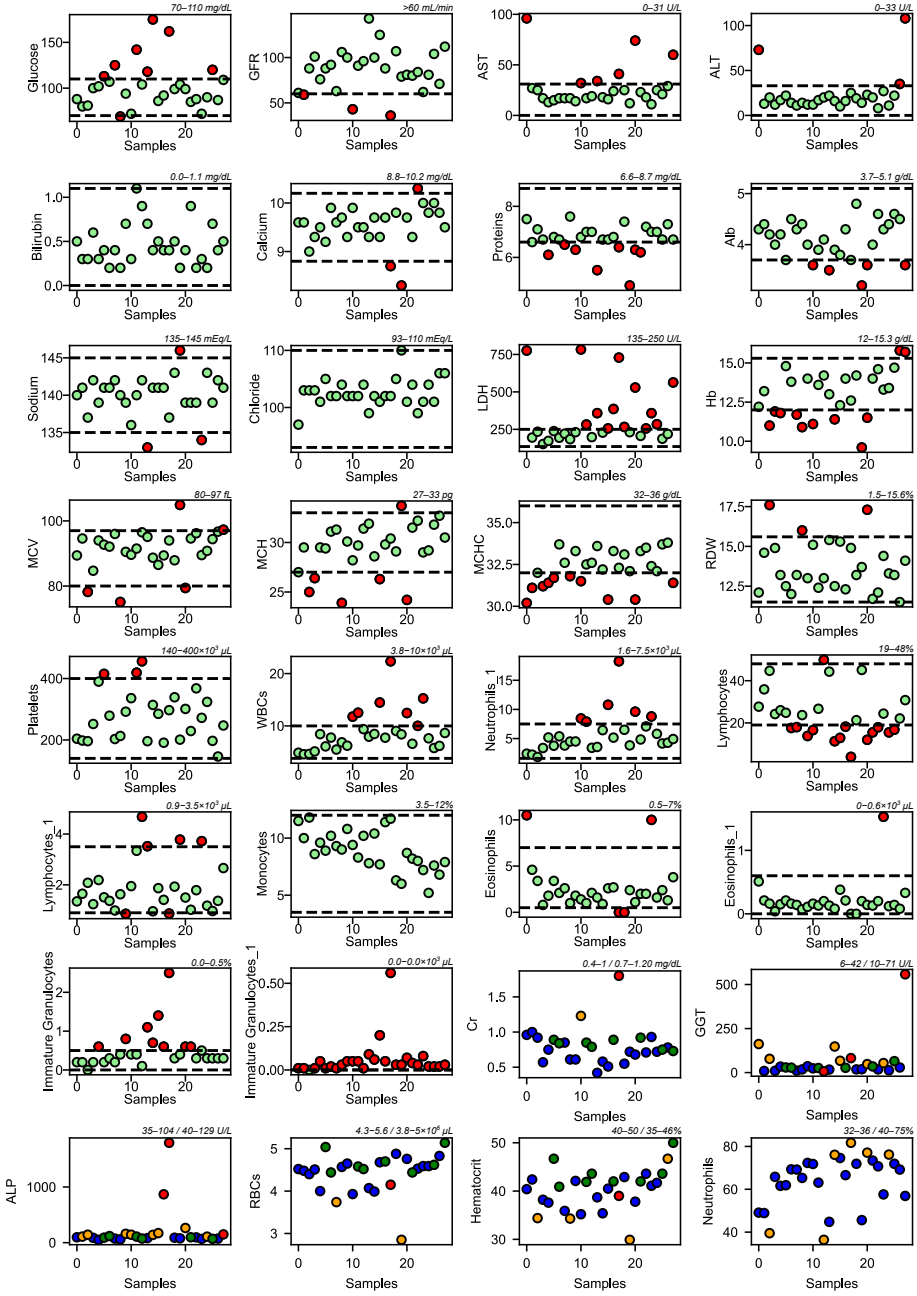


Figure 1.36: Biochemical parameter profiles of newly registered lung cancer patients.

Conversely, several parameters exhibit frequent deviations, marked by clusters of red dots. Immature granulocytes_1 and LDH show notable elevations, likely reflecting heightened cellular turnover or inflammatory cascades driven by malignancy [62, 63]. Reduced Hb and MCHC signal anemia, a common sequela of chronic disease or disrupted red blood cell production [64]. Variability in proteins may point to nutritional deficits or metabolic dysregulation, frequently observed in cancer-associated cachexia [65]. ALP deviations, observed in a subset of patients, raise suspicion of hepatic or biliary compromise, potentially implicating metastatic involvement or cholestasis [66]. These findings indicate that biochemical profiles among newly registered lung cancer patients display a degree of heterogeneity. The observed deviations across hematological, inflammatory, hepatic, and metabolic markers may result from a combination of disease-specific effects, physiological responses, and individual variability.

In addition to biochemical parameter evaluation, Raman spectral data were analyzed using multivariate classification methods to further delineate clinical characteristics of the cohort. Figure 1.37 illustrates the spectral classification results, with score plots (Figure 1.37A, C, E, G, I) derived from latent variables LV1 and LV2, and corresponding regression coefficients (Figure 1.37B, D, F, H, J) highlighting influential spectral regions for each clinical class: tumor staging (T), lymph node involvement (N), metastasis (M), gender, and age. It is important to note that this analysis was based on data from only 24 patients, as the remaining four patients lacked complete clinical class information for the parameters mentioned above.

The score plots demonstrate varying degrees of separation across clinical categories. For T in Figure 1.37A, classes display partial clustering, suggesting that tumor size or local spread influences the measured spectral profiles. Class 3 tends to occupy higher LV1 values, while Class 1 and 2 remains near the upper or central part. Classes 4 appear scattered, reflecting greater heterogeneity in their biochemical signatures. In Figure 1.37C (N), classes 1, 2, 3, and N/A also form distinct clusters, with class 2 often positioned separately from the other groups. This pattern suggests that nodal involvement may be linked to identifiable biochemical features in Raman data. Figure 1.37E (M) compares the metastatic status. Medical doctors sometimes use letters to further classify

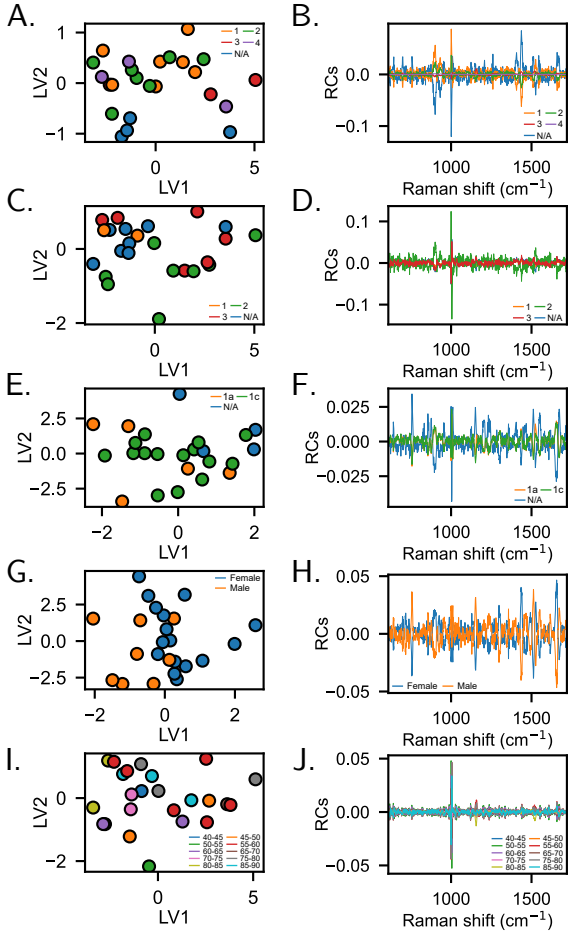


Figure 1.37: Multivariate analysis of Raman spectroscopic data grouped by clinical parameters. Figure A, C, E, G, and I (score plots) illustrate how patient spectra cluster according to tumor stage, nodal status, metastatic status, gender, and age, respectively. Figures B, D, F, H, and J (regression coefficients) highlight the specific Raman shifts most responsible for separating these classes in each respective comparison.

metastatic stages, where 1a indicates cancer spread to the other lung, while 1c denotes spread to multiple body parts. In this analysis, 1c produces more discrete clusters along the LV1 axis compared to 1a, although some overlap remains. N/A group occupies a region distinct from both 1a and 1c, indicating that unknown metastatic status can also exhibit a distinct spectral profile. Figure 1.37G (gender) presents some localized clustering, yet it is minor overall, implying that there may be gender-related biochemical differences in this dataset. Finally, Figure 1.37I (age) shows points corresponding to multiple age brackets, with no clear gradient or grouping across LV1–LV2 space. The overlapping distributions of different age ranges indicate that age-based variation is not strongly distinguishable by these spectroscopic methods.

Regression coefficients for T (Figure 1.37B) peak at 1001 cm^{-1} (phenylalanine) for classes N/A, 1, 2, and 4, and at 1156 cm^{-1} (lipids/proteins) for class 3, while minima occur at 638 cm^{-1} (τ(CC) in tyr: N/A), 1648 cm^{-1} (amide I: class 1), 1460 cm^{-1} (CH_2/CH_3: class 2), 1692 cm^{-1} (amide I: class 3), and 1291 cm^{-1} (amide III: class 4), indicating metabolic dysregulation across amino acid, lipid, and protein pathways. Figure 1.37D reveals nodal involvement correlates with uniform peaks at 1005 cm^{-1} across all classes and class-specific minima (946, 978, 1079, 1470 cm^{-1}), suggesting lymphatic spread disrupts amino acid and metabolic pathways. Figure 1.37F shows metastatic classes (1a, 1c) and N/A with maximal regression coefficients at 1004 cm^{-1} and class-specific minima at 910 cm^{-1} (N/A), 1480 cm^{-1} (1a), and 1447 cm^{-1} (1c), indicating metastatic progression disrupts amino acid and metabolic regulation. In contrast, Figure 1.37H (gender) shows identical peaks (1656 cm^{-1}) and lowest values (1276 cm^{-1}) for both males and females, suggesting minimal biochemical differences between sexes. Figure 1.37J (age) shares a common peak at 1005 cm^{-1} across all age groups but displays widely varying lowest values (622–1642 cm^{-1}), implying age-related biochemical changes are subtle or inconsistent compared to other factors.

Analysis of biochemical data grouped by clinical parameters complements Raman spectroscopic findings (Figure 1.38). Similar to the analysis of spectral data described previously, biochemical parameters were evaluated using the same multivariate method, with score plots (Figures 1.38A, B, C, D, E)

illustrating patient clustering according to T, N, M, gender, and age.

Figure 1.38A (T) demonstrates distinct biochemical clustering: class 1 aligns with negative LV1 values, classes 2 and N/A overlap centrally, class 3 shifts toward positive LV2, and class 4 exhibits dispersion, reflecting tumor heterogeneity. Figure 1.38B (N) reveals moderate clustering, with class 3 concentrated at negative LV2, distinct from overlapping classes 1, 2, and N/A, suggesting biochemical divergence in advanced nodal disease. Figure 1.38C (M) separates 1c and 1a groups centrally, while N/A clusters at negative LV2, indicating biochemical distinctions tied to metastatic status. Figure 1.38D (gender) shows moderate sex-based separation, implying hormonal or metabolic influences on biochemical profiles. Figure 1.38E (age) displays minimal clustering across groups, with biochemical alterations dominated by cancer pathology rather than age-related variability.

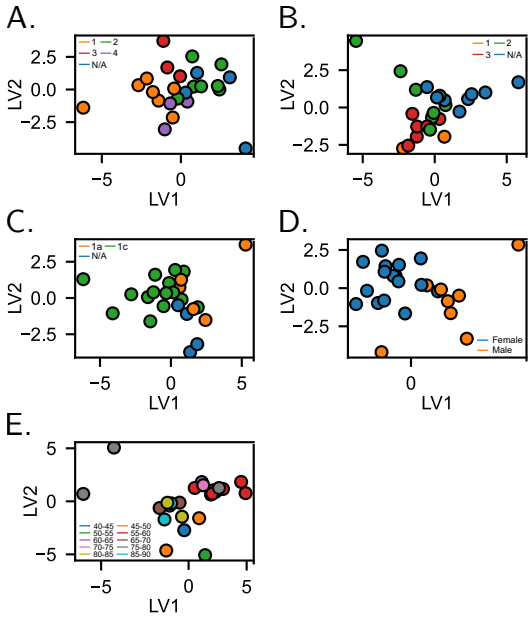


Figure 1.38: Clustering patterns of biochemical data across clinical parameters. Figure A, B, C, D, and E (score plots) illustrate how biochemical data cluster according to T, N, M, gender, and age, respectively.

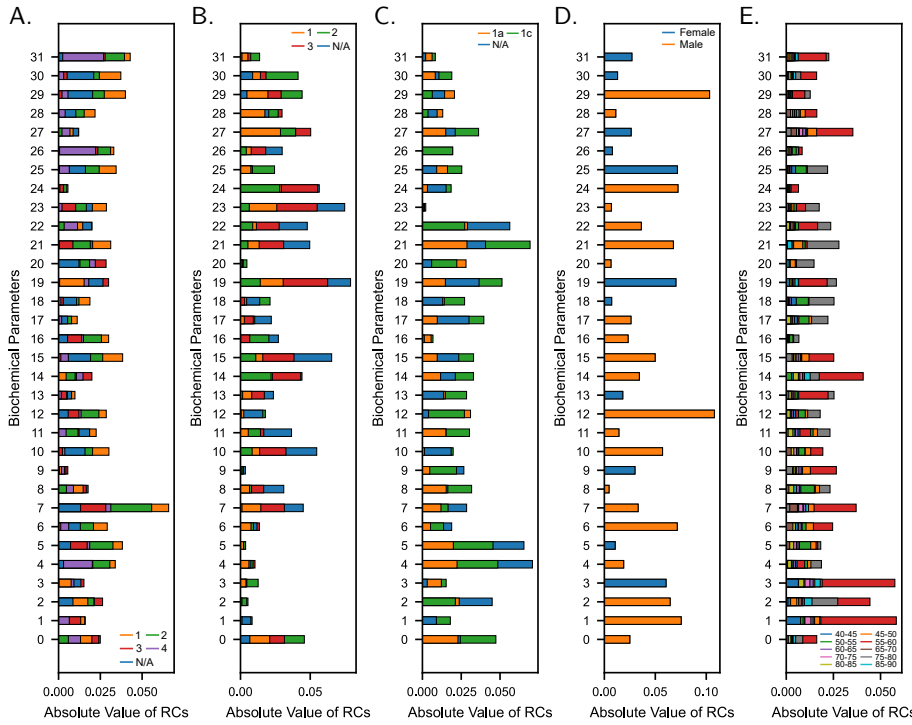


Figure 1.39: Absolute values of regression coefficients for biochemical parameters across clinical classifications: (A) T: 1–4, N/A, (B) N: 1–3, N/A, (C) M: 1a, 1c, N/A, (D) gender: male and female, (E) age: 5-year brackets (40–45 to 85–90). Numerical labels (0–31) correspond to biochemical parameters: LDH (0), Hb (1), RBCs (2), Hematocrit (3), MCV (4), MCH (5), MCHC (6), RDW (7), WBCs (8), Platelets (9), Neutrophils (10), Neutrophils_1 (11), Monocytes (12), Eosinophils (13), Eosinophils_1 (14), Lymphocytes (15), Lymphocytes_1 (16), Immature Granulocytes (17), Immature Granulocytes_1 (18), Calcium (19), Sodium (20), Chloride (21), Proteins (22), Alb (23), Bilirubin (24), ALP (25), GGT (26), AST (27), ALT (28), Cr (29), MDRD (30), Glucose (31). N/A denotes missing or unclassified data. Coefficients reflect parameter-specific contributions to clinical classifications.

Figure 1.39A (T) shows that RDW dominates stage-specific associations, with the highest coefficients in class 2 (0.056) and class 3 (0.028), reflecting progressive inflammatory or erythropoietic dysregulation. Class 1 (0.066) shows the strongest overall RDW association, while class 4 (0.031) and N/A (0.013) exhibit weaker effects. Bilirubin contributes minimally across classes ($\leq +0.005$), underscoring limited hepatic involvement in tumor staging. Figure 1.39B (N) reveals calcium as the dominant contributor, with the highest coefficients in N/A (0.079) and class 3 (0.062), suggesting metabolic dysregulation or bone involvement in advanced nodal disease. Class 1 (0.031) and class 2 (0.014) exhibit progressively weaker calcium associations. Platelets show minimal influence across all classes (≤ 0.0035), indicating negligible alterations in hemostatic balance. Figure 1.39C (M) identifies MCV as the dominant contributor, with the highest coefficient in N/A (0.071), suggesting erythrocyte size alterations in unclassified metastatic cases. 1c (0.049) shows stronger MCV associations than 1a (0.022). Albumin exhibits negligible contributions across all groups (≤ 0.0022), indicating stable hepatic/nutritional status regardless of metastatic burden. Figure 1.39D (gender) shows monocytes with identical high coefficients for both male and female (0.108), suggesting shared monocyte-related immune or inflammatory mechanisms regardless of sex. White blood cells (WBCs) exhibit minimal influence (0.0046), indicating negligible sex-specific alterations in leukocyte counts. Figure 1.39E (age) identifies Hb with the strongest age-specific association in the 55–60 bracket (0.058), suggesting pronounced erythrocyte-related shifts in this cohort. Other age groups exhibit weaker Hb contributions ($\leq +0.018$), with no progressive trend across decades. WBCs show uniformly minimal influence (≤ 0.0065), indicating negligible age-dependent leukocyte alterations. The isolated Hb peak in 55–60-year-olds highlights transient biochemical dynamics, while WBCs stability across ages underscores their limited role in age-related profiling within this dataset.

1.5 References

[1] Aida Budreviciute et al. “Management and Prevention Strategies for Non-communicable Diseases (NCDs) and Their Risk Factors”. In: *Frontiers in*

Public Health 8 (2020). ISSN: 2296-2565. DOI: 10.3389/fpubh.2020.574111.

[2] Mariachiara Di Cesare et al. "The Heart of the World". In: *Global Heart* (Jan. 2024). DOI: 10.5334/gh.1288.

[3] Rajesh Sharma. "Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050". In: *International Journal of Clinical Oncology* 27 (Jan. 2022). DOI: 10.1007/s10147-021-02108-2.

[4] Mary F. Henningfield and Alex A. Adjei. "Lung Cancer Awareness Month—A Lot of Progress, But More Work Needs to Be Done". In: *Journal of Thoracic Oncology* 12.11 (2017), pp. 1603–1605. ISSN: 1556-0864. DOI: https://doi.org/10.1016/j.jtho.2017.09.091.

[5] Roberto Gasparri, Angela Sabalic, and Lorenzo Spaggiari. "The Early Diagnosis of Lung Cancer: Critical Gaps in the Discovery of Biomarkers". In: *Journal of Clinical Medicine* 12.23 (2023). ISSN: 2077-0383. DOI: 10.3390/jcm12237244.

[6] Harun Hano et al. "Fusion of Raman and FTIR Spectroscopy Data Uncovers Physiological Changes Associated with Lung Cancer". In: *International Journal of Molecular Sciences* 25.20 (2024). ISSN: 1422-0067. DOI: 10.3390/ijms252010936.

[7] Mohammad A. Thanoon et al. "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images". In: *Diagnostics* 13.16 (2023). ISSN: 2075-4418. DOI: 10.3390/diagnostics13162617.

[8] Harun Hano et al. "Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy". In: *Microchemical Journal* 209 (2025), p. 112496. ISSN: 0026-265X. DOI: 10.1016/j.microc.2024.112496.

[9] Harun Hano et al. "Power of Light: Raman Spectroscopy and Machine Learning for the Detection of Lung Cancer". In: *ACS Omega* 9 (12 Mar. 2024), pp. 14084–14091. DOI: 10.1021/acsomega.3c09537.

[10] Vanessa Santos et al. “Liquid Biopsy: the Value of Different Bodily Fluids”. In: *Biomarkers in Medicine* 16.2 (2022), pp. 127–145. DOI: 10.2217/bmm-2021-0370.

[11] Kelly Virkler and Igor K. Lednev. “Analysis of body fluids for forensic purposes: From laboratory testing to non-destructive rapid confirmatory identification at a crime scene”. In: *Forensic Science International* 188.1 (2009), pp. 1–17. ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2009.02.013.

[12] Carlos R. Baiz et al. “Vibrational Spectroscopic Map, Vibrational Spectroscopy, and Intermolecular Interaction”. In: *Chemical Reviews* 120.15 (2020). PMID: 32598850, pp. 7152–7218. DOI: 10.1021/acs.chemrev.9b00813.

[13] Vera Balan et al. “Vibrational Spectroscopy Fingerprinting in Medicine: from Molecular to Clinical Practice”. In: *Materials* 12.18 (2019). ISSN: 1996-1944. DOI: 10.3390/ma12182884.

[14] Daniel R. Neuville, Dominique de Ligny, and Grant S. Henderson. “Advances in Raman Spectroscopy Applied to Earth and Material Sciences”. In: *Reviews in Mineralogy and Geochemistry* 78.1 (Jan. 2014), pp. 509–541. ISSN: 1529-6466. DOI: 10.2138/rmg.2013.78.13.

[15] Kenny Kong et al. “Raman spectroscopy for medical diagnostics — From in-vitro biofluid assays to in-vivo cancer detection”. In: *Advanced Drug Delivery Reviews* 89 (2015). Pharmaceutical applications of Raman spectroscopy – from diagnosis to therapeutics, pp. 121–134. ISSN: 0169-409X. DOI: 10.1016/j.addr.2015.03.009.

[16] Nguyen Quy Dao. “Dispersive Raman Spectroscopy, Current Instrumental Designs”. In: *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, 2006. ISBN: 9780470027318. DOI: 10.1002/9780470027318.a6402.

[17] C. D. Allemand. “Design Criteria for a Raman Spectrometer”. In: *Appl. Opt.* 9.6 (1970), pp. 1304–1311. DOI: 10.1364/AO.9.001304.

[18] Ahmed Fadlelmoula et al. “Fourier Transform Infrared (FTIR) Spectroscopy to Analyse Human Blood over the Last 20 Years: A Review towards Lab-on-a-Chip Devices”. In: *Micromachines* 13.2 (2022). ISSN: 2072-666X. DOI: 10.3390/mi13020187.

[19] Brian J. Goodfellow Sandra Magalhães and Alexandra Nunes. “FTIR spectroscopy in biomedical research: how to get the most out of its potential”. In: *Applied Spectroscopy Reviews* 56.8-10 (2021), pp. 869–907. DOI: 10.1080/05704928.2021.1946822.

[20] M. Olga Guerrero-Pérez and Gregory S. Patience. “Experimental methods in chemical engineering: Fourier transform infrared spectroscopy—FTIR”. In: *The Canadian Journal of Chemical Engineering* 98.1 (2020), pp. 25–33. DOI: 10.1002/cjce.23664.

[21] Abdul Rohman et al. “The use of FTIR and Raman spectroscopy in combination with chemometrics for analysis of biomolecules in biomedical fluids: A review”. In: *Biomedical Spectroscopy and Imaging* 8 (Jan. 2020), pp. 1–17. DOI: 10.3233/BSI-200189.

[22] Nils Kristian Afseth, Vegard Herman Segtnan, and Jens Petter Wold. “Raman Spectra of Biological Samples: A Study of Preprocessing Methods”. In: *Applied Spectroscopy* 60.12 (2006). PMID: 17217584, pp. 1358–1367. DOI: 10.1366/000370206779321454.

[23] Christopher Rowlands and Stephen Elliott. “Automated algorithm for baseline subtraction in spectra”. In: *Journal of Raman Spectroscopy* 42.3 (2011), pp. 363–369. DOI: 10.1002/jrs.2691.

[24] Piotr Gromski et al. “The influence of scaling metabolomics data on model classification accuracy”. In: *Metabolomics* 11 (June 2015), pp. 684–695. DOI: 10.1007/s11306-014-0738-7.

[25] Akanksha Sharma and Vishal Sharma. “Chapter 9 - Raman spectroscopy combined with chemometrics”. In: *Chemometrics*. Ed. by Fabiano André Narciso Fernandes, Sueli Rodrigues, and Elenilson Godoy Alves Filho. Elsevier, 2024, pp. 197–222. ISBN: 978-0-443-21493-6. DOI: 10.1016/B978-0-443-21493-6.00009-5.

[26] Kanishka Tyagi et al. “Chapter 3 - Unsupervised learning”. In: *Artificial Intelligence and Machine Learning for EDGE Computing*. Ed. by Rajiv Pandey et al. Academic Press, 2022, pp. 33–52. ISBN: 978-0-12-824054-0. DOI: 10.1016/B978-0-12-824054-0.00012-5.

[27] Lorna Ashton Carlos A. Meza Ramirez Michael Greenop and Ihtesham ur Rehman. “Applications of machine learning in spectroscopy”. In: *Applied Spectroscopy Reviews* 56.8-10 (2021), pp. 733–763. DOI: 10.1080/05704928.2020.1859525.

[28] Tom Howley et al. “The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data”. In: *Knowledge-Based Systems* 19.5 (2006). AI 2005 SI, pp. 363–370. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2005.11.014.

[29] Hong Zhou. “Linear Discriminant Analysis”. In: *Data Mining and Machine Learning* (2020).

[30] Zhi-Hua Zhou. “Support Vector Machine”. In: *Machine Learning*. Singapore: Springer Singapore, 2021, pp. 129–153. ISBN: 978-981-15-1967-3. DOI: 10.1007/978-981-15-1967-3_6.

[31] Tao Pan et al. “Visible and Near-Infrared Spectroscopy Combined With Bayes Classifier Based on Wavelength Model Optimization Applied to Wine Multibrand Identification”. In: *Frontiers in Nutrition* 9 (2022). ISSN: 2296-861X. DOI: 10.3389/fnut.2022.796463.

[32] Autcha Araveeporn and Puntipa Wanitjirattikal. “Comparison of Machine Learning Methods for Binary Classification of Multicollinearity Data”. In: *Proceedings of the 2024 7th International Conference on Mathematics and Statistics*. ICoMS ’24. New York, NY, USA: Association for Computing Machinery, 2024, 44–49. ISBN: 9798400707223. DOI: 10.1145/3686592.3686600.

[33] Robin Genuer and Jean-Michel Poggi. “Random Forests”. In: *Random Forests with R*. Cham: Springer International Publishing, 2020, pp. 33–55. ISBN: 978-3-030-56485-8. DOI: 10.1007/978-3-030-56485-8_3.

[34] Adele Cutler, D. Richard Cutler, and John R. Stevens. "Random Forests". In: *Ensemble Machine Learning: Methods and Applications*. Ed. by Cha Zhang and Yunqian Ma. New York, NY: Springer New York, 2012, pp. 157–175. ISBN: 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7_5.

[35] Svante Wold, Michael Sjöström, and Lennart Eriksson. "PLS-regression: a basic tool of chemometrics". In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001). PLS Methods, pp. 109–130. ISSN: 0169-7439. DOI: 10.1016/S0169-7439(01)00155-1.

[36] Roman Rosipal and Nicole Krämer. "Overview and Recent Advances in Partial Least Squares". In: *Subspace, Latent Structure and Feature Selection*. Ed. by Craig Saunders et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 34–51. ISBN: 978-3-540-34138-3.

[37] Tahir Mehmood, Solve Sæbø, and Kristian Hovde Liland. "Comparison of variable selection methods in partial least squares regression". In: *Journal of Chemometrics* 34.6 (2020). e3226 cem.3226, e3226. DOI: 10.1002/cem.3226.

[38] Luke Merrick and Ankur Taly. *The Explanation Game: Explaining Machine Learning Models Using Shapley Values*. 2020. arXiv: 1909.08128 [cs.LG].

[39] Marina Cocchi. "Chapter 1 - Introduction: Ways and Means to Deal With Data From Multiple Sources". In: *Data Fusion Methodology and Applications*. Ed. by Marina Cocchi. Vol. 31. Data Handling in Science and Technology. Elsevier, 2019, pp. 1–26. DOI: 10.1016/B978-0-444-63984-4.00001-6.

[40] Agnieszka Smolinska et al. "Chapter 3 - General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences". In: *Data Fusion Methodology and Applications*. Ed. by Marina Cocchi. Vol. 31. Data Handling in Science and Technology. Elsevier, 2019, pp. 51–79. DOI: 10.1016/B978-0-444-63984-4.00003-X.

[41] Silvana M. Azcarate et al. "Data handling in data fusion: Methodologies and applications". In: *TrAC Trends in Analytical Chemistry* 143 (2021), p. 116355. ISSN: 0165-9936. DOI: 10.1016/j.trac.2021.116355.

[42] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. "A Review of Evaluation Metrics in Machine Learning Algorithms". In: *Artificial Intelligence Application in Networks and Systems*. Ed. by Radek Silhavy and Petr Silhavy. Cham: Springer International Publishing, 2023, pp. 15–25. ISBN: 978-3-031-35314-7.

[43] Sung-June Baek et al. "Baseline correction using asymmetrically reweighted penalized least squares smoothing". In: *Analyst* 140 (1 2015), pp. 250–257. DOI: 10.1039/C4AN01061B.

[44] Yiming Bi et al. "A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation". In: *Analytica Chimica Acta* 909 (2016), pp. 30–40. ISSN: 0003-2670. DOI: 10.1016/j.aca.2016.01.010.

[45] Quanquan Gu, Zhenhui Li, and Jiawei Han. "Generalized Fisher Score for Feature Selection". In: *CoRR* abs/1202.3725 (2012). arXiv: 1202.3725.

[46] Maria P. Campos and Marco S. Reis. "Data preprocessing for multiblock modelling – A systematization with new methods". In: *Chemometrics and Intelligent Laboratory Systems* 199 (2020), p. 103959. ISSN: 0169-7439. DOI: 10.1016/j.chemolab.2020.103959.

[47] M. Silvestri et al. "A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines". In: *Chemometrics and Intelligent Laboratory Systems* 137 (2014), pp. 181–189. ISSN: 0169-7439. DOI: 10.1016/j.chemolab.2014.06.012.

[48] Puneet Mishra et al. "Recent trends in multi-block data analysis in chemometrics for multi-source data integration". In: *TrAC Trends in Analytical Chemistry* 137 (2021), p. 116206. ISSN: 0165-9936. DOI: 10.1016/j.trac.2021.116206.

[49] Matthias Schonlau. "The Naive Bayes Classifier". In: *Applied Statistical Learning: With Case Studies in Stata*. Cham: Springer International Publishing, 2023, pp. 143–160. ISBN: 978-3-031-33390-3. DOI: 10.1007/978-3-031-33390-3_8.

[50] D. Prabha et al. “A Survey on Alleviating the Naive Bayes Conditional Independence Assumption”. In: *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. 2022, pp. 654–657. DOI: 10.1109/ICAISS55157.2022.10011103.

[51] Guy Van den Broeck et al. “On the Tractability of SHAP Explanations”. In: *CoRR* abs/2009.08634 (2020). arXiv: 2009.08634.

[52] Alla Sinica et al. “Raman spectroscopic discrimination of normal and cancerous lung tissues”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 219 (2019), pp. 257–266. ISSN: 1386-1425. DOI: 10.1016/j.saa.2019.04.055.

[53] Jolanta Bujok et al. “Applicability of FTIR-ATR Method to Measure Carbonyls in Blood Plasma after Physical and Mental Stress”. In: *BioMed Research International* 2019.1 (2019), p. 2181370. DOI: 10.1155/2019/2181370.

[54] Ondřej Vrtělka et al. “Vibrational and chiroptical analysis of blood plasma for hepatocellular carcinoma diagnostics”. In: *Analyst* 148 (12 2023), pp. 2793–2800. DOI: 10.1039/D3AN00164D.

[55] Ketan Gajjar et al. “Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer”. In: *Analyst* 138 (14 2013), pp. 3917–3926. DOI: 10.1039/C3AN36654E.

[56] Kelvin W. C. Poon et al. “Quantitative reagent-free detection of fibrinogen levels in human blood plasma using Raman spectroscopy”. In: *Analyst* 137 (8 2012), pp. 1807–1814. DOI: 10.1039/C2AN35042D.

[57] Hao Shen et al. “Protocol for determining protein dynamics using FT-IR spectroscopy”. In: *STAR Protocols* 4.4 (2023), p. 102587. ISSN: 2666-1667. DOI: 10.1016/j.xpro.2023.102587.

[58] Vasimahmed Lala, Muhammad Zubair, and David A. Minter. “Liver Function Tests”. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing, 2025.

[59] Isha Shrimanker and Sandeep Bhattarai. “Electrolytes”. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing, 2025.

[60] Verena Gounden, Harshil Bhatt, and Ishwarlal Jialal. "Renal Function Tests". In: *StatPearls*. Treasure Island, FL: StatPearls Publishing, 2025.

[61] Valerie E. Espinoza and Prabhu D. Emmady. "Histology, Monocytes". In: *StatPearls*. Treasure Island, FL: StatPearls Publishing, 2025.

[62] Nuran Cetın et al. "Immature granulocytes as biomarkers of inflammation in children with predialysis chronic kidney disease". In: *Pediatric nephrology (Berlin, Germany)* 38.1 (2023), 219—225. ISSN: 0931-041X. DOI: 10.1007/s00467-022-05530-4.

[63] Erika Poggiali et al. "Lactate dehydrogenase and C-reactive protein as predictors of respiratory failure in CoVID-19 patients". In: *Clinica Chimica Acta* 509 (2020), pp. 135–138. ISSN: 0009-8981. DOI: https://doi.org/10.1016/j.cca.2020.06.012.

[64] Madhu Badireddy, Krishna M. Baradhi, and Andrea Wilhite (Hughes). *Chronic Anemia (Nursing)*. StatPearls Publishing, Treasure Island (FL), 2023.

[65] Michele Petruzzelli and Erwin F. Wagner. "Mechanisms of metabolic dysfunction in cancer-associated cachexia". In: *Genes & Development* 30 (2016), pp. 489 –501.

[66] Dhruv Lowe et al. "Alkaline Phosphatase". In: *StatPearls*. Treasure Island, FL: StatPearls Publishing, 2025.

Chapter 2

Conclusions

This research strengthens the growing evidence that vibrational spectroscopy, combined with chemometrics, can be an effective non-invasive tool for detecting lung cancer. By analyzing human blood plasma, this thesis delves into the chemical changes linked to lung cancer and highlights the potential of spectroscopic methods for improving early diagnosis and treatment.

The first study (**Publication 1**) demonstrated that PCA + LDA and PCA + FS + SVM achieved the highest classification performance, with accuracy around 0.85 ± 0.14 and AUC scores above 0.93. Notably, LDA alone also maintained strong performance, making it a viable standalone option without dimensionality reduction. Additionally, PLS-DA provided reliable classification, further supporting its potential for spectral-based diagnostics. These results highlight that while advanced feature extraction enhances predictive accuracy, simpler models like LDA and PLS-DA remain robust for linearly separable dataset and flexible for different diagnostic applications.

The second study (**Publication 2**) identified key wavenumbers—1125 cm^{-1}, 966 cm^{-1}, and 1671 cm^{-1}—as highly relevant spectral markers for lung cancer. These vibrational bands, consistently selected across different chemometric models, are linked to molecular alterations primarily in proteins, lipids,

and nucleic acids. Although SHAP (accuracy: 0.835 ± 0.013, AUC-ROC: 0.947 ± 0.090) and MWU (accuracy: 0.818 ± 0.014, AUC-ROC: 0.958 ± 0.077) exhibited strong performance in feature selection, the analysis primarily focused on RCs (accuracy: 0.825 ± 0.014, AUC-ROC: 0.94 ± 0.11) due to their simplicity in optimization, ease of implementation, and comparable classification performance. While the exact biochemical mechanisms behind these spectral shifts require further investigation, the strong statistical significance of these selected features suggests their potential as robust biomarkers for lung cancer detection.

The third study (**Publication 3**) demonstrated that combining Raman and FTIR spectroscopy with data fusion techniques significantly improves classification performance. FS and FR, when further applied, enhanced accuracy, highlighting their importance in optimizing classification models. Key protein-related vibrations at 1125 cm^{-1}, 1587 cm^{-1}, and 1632–1668 cm^{-1} from Raman spectroscopy, along with 1055–1070 cm^{-1} and 1699 cm^{-1} from FTIR spectroscopy, were identified as crucial indicators. Among the techniques, LLDF with FS achieved the highest accuracy (0.9922 ± 0.0015), significantly outperforming models using Raman (0.8539 ± 0.0056) or FTIR alone (0.8425 ± 0.0058). FR in fusion-based models, while effective (0.8711 ± 0.0034), did not surpass the performance of FS, indicating that retaining selected wavenumbers enhances classification efficiency. These findings highlight that integrating complementary spectral data, especially with FS, provides superior diagnostic accuracy for lung cancer detection.

Moreover, the findings presented in Figures 1.37 underscore the clinical utility of Raman spectroscopy for rapid, non-invasive classification of lung cancer progression. Distinct spectral signatures (1001–1005 cm^{-1}, 1156 cm^{-1}, etc.) correlate strongly with tumor staging, nodal involvement, and metastatic spread, offering a potential alternative to invasive biopsies or time-consuming biochemical assays. While this work does not establish diagnostic prediction, the ability to detect class-specific metabolic disruptions (e.g., amino acid dysregulation in metastasis) in minutes, using only a droplet of blood plasma, highlights that the spectroscopic methods promise as a photonic tool for real-time monitoring of cancer-associated biochemical dynamics.

In conclusion, this thesis represents a major step forward in developing non-invasive, spectroscopic methods for diagnosing lung cancer. Future research should focus on using larger datasets, improving our understanding of the molecular changes involved, and integrating these techniques into clinical practice. These efforts could lead to earlier detection, better treatment outcomes, and a clearer understanding of how lung cancer develops.

Appendix A

Publications as a first author

A.1 Supervised learning algorithms

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. **Power of Light: Raman Spectroscopy and Machine Learning for the Detection of Lung Cancer**. ACS Omega 2024, 9 (12), 14084-14091. DOI: 10.1021/acsomega.3c09537

Impact Factor: 3.7

ACS OMEGA

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

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 non-destructive 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

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

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.

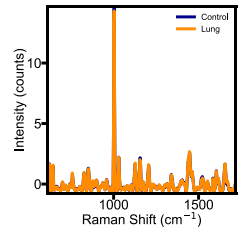


Figure 1. Averaged Raman spectra per class.

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

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

ACS Omega http://pubs.acs.org/journal/acsodf Article

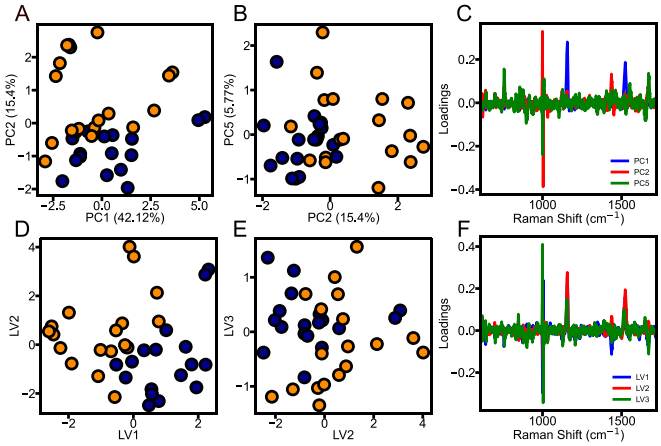


Figure 2. 2D visualization of the discrimination between lung cancer patients (orange) and healthy controls (blue). (A) Score plots of PC1−PC2, (B) PC2−PC5, (D) LV1−LV2, and (E) LV2−LV3. (C) Corresponding loadings of PC1, PC2, and PC5, and (F) loadings of LV1, LV2, and LV3.

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

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

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

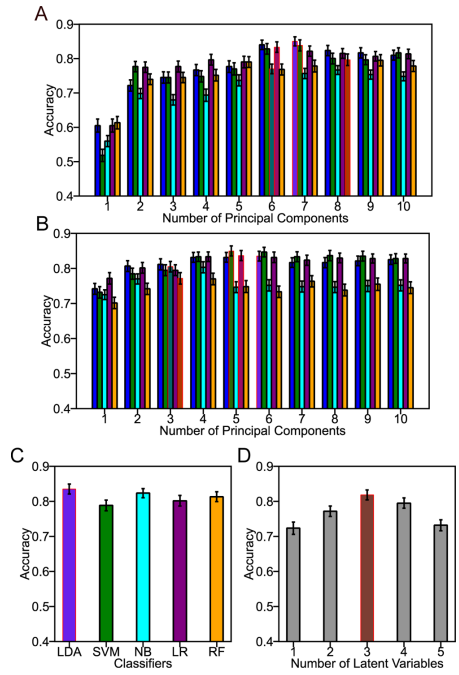


Figure 3. Accuracy of comparative classification performance. Hatched bars indicate the highest accuracy presented in each graph per classifier. The error bars reflect the standard errors. (A) PCA with the first 10 PCs as inputs for various classifiers: LDA (blue), SVM (green), NB (cyan), LR (purple), 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.

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.

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 +

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

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.

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

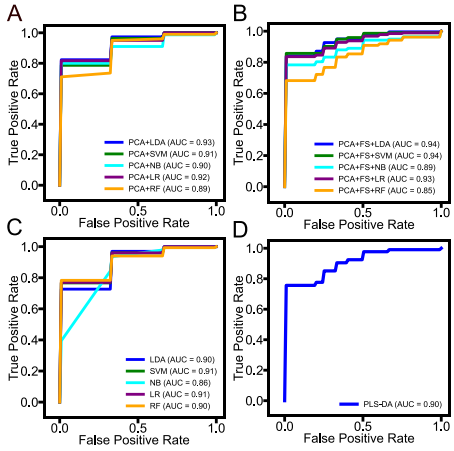


Figure 4. ROC curves based on various classification strategies. (A) PCs with classifiers. (B) Selected PCs by Fisher score with classifiers. (C) Only classifiers. (D) Selected LVs from PLS.

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

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

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.

https://doi.org/10.1021/acsomega.3c09537
ACS Omega 2024, 9, 14084−14091

A.2 Feature selection methods

Harun Hano, Beatriz Suarez, Jose Manuel Amigo, Charles H. Lawrie, and Andreas Seifert. **Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy**. Microchemical Journal 2025, 209, 112496. DOI: 10.1016/j.microc.2024.112496

Impact Factor: 4.9

Journal Pre-proof

Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy

Harun Hano, Beatriz Suarez, Jose Manuel Amigo, Charles H. Lawrie, Andreas Seifert

PII: S0026-265X(24)02609-2
DOI: https://doi.org/10.1016/j.microc.2024.112496
Reference: MICROC 112496

To appear in: *Microchemical Journal*

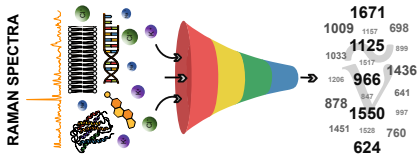
Received date : 23 July 2024
Revised date : 28 November 2024
Accepted date : 16 December 2024

Please cite this article as: H. Hano, B. Suarez, J.M. Amigo et al., Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy, *Microchemical Journal* (2025), doi: https://doi.org/10.1016/j.microc.2024.112496.

Graphical Abstract

Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy

Harun Hano,Beatriz Suarez,Jose Manuel Amigo,Charles H. Lawrie,Andreas Seifert



Highlights

Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy

Harun Hano,Beatriz Suarez,Jose Manuel Amigo,Charles H. Lawrie,Andreas Seifert

- Raman spectroscopy combined with chemometrics enables effective *in vitro* analysis of lung cancer.
- Model performance is improved in terms of accuracy through feature selection methods.
- Certain wavenumbers are highlighted as potential diagnostic biomarkers for lung cancer.

Rapid noninvasive lung cancer screening via discriminative wavenumbers in Raman spectroscopy

Harun Hano[a,b,*], Beatriz Suarez[c,d], Jose Manuel Amigo[e,f], Charles H. Lawrie[d,f,g,h] and Andreas Seifert[a,f,**]

[a]CIC nanoGUNE BRTA, 20018 San Sebastián, Spain
[b]Department of Physics, University of the Basque Country (UPV/EHU), 20018 San Sebastián, Spain
[c]Faculty of Nursing and Medicine, University of the Basque Country (UPV/EHU), 48940 Leioa, Spain
[d]Biogipuzkoa Health Research Institute, 20014 San Sebastián, Spain
[e]Department of Analytical Chemistry, University of the Basque Country (UPV/EHU), 48940 Leioa, Spain
[f]IKERBASQUE - Basque Foundation for Science, 48009 Bilbao, Spain
[g]Sino-Swiss Institute of Advanced Technology (SSIAT), University of Shanghai, 201800 Shanghai, China
[h]Radcliffe Department of Medicine, University of Oxford, OX3 9DU Oxford, UK

ARTICLE INFO

Keywords:
lung cancer
vibrational spectroscopy
chemometrics
binary classification
feature selection

ABSTRACT

Lung cancer still presents a major health challenge due to its high mortality rate, but early diagnosis can contribute to a favorable prognosis. Here an innovative approach is introduced that combines Raman spectroscopy with chemometrics to analyze human blood plasma and identify specific wavenumbers associated with the disease. The approach is to thoroughly examine each wavenumber in the Raman spectra and determine the most discriminant ones, which helps clinical researchers in their analyses. In particular, the peaks around 1125 cm^{-1} and 966 cm^{-1} were identified as important and showed significant correlations with lung cancer using a regression-based, statistical, and model-agnostic approach. In addition, multiple peaks were identified indicating potential biomarkers based on proteins and lipids that may have a high correlation with lung cancer. This promising study rep-resents an important step towards developing efficient, noninvasive diagnostic tools for lung cancer that justify further validation and research in clinical settings.

1. Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, with approximately 1.8 million annual deaths. This major challenge is compounded by late diagnoses, leading to limited treatment options, especially with a 5-year survival rate of about 67% for stage I and 23% for stage III [1, 2]. Modern computer-aided diagnostic procedures may respond to this concerning fact and urgent clinical need in the development of effective, noninvasive diagnostic methods that not only facilitate diagnosis in high-risk groups but also pave the way for personalized treatment options [3, 4]. Current clinical diagnostic methods, including low-dose computed tomography (LDCT), chest X-rays, and invasive procedures such as bronchoscopy, needle biopsies etc., have seen advancements over the past few decades [5]. However, these techniques still exhibit several limitations. These include high false-positive rates, overdiagnosis, incidental findings, increased patient distress, low resolution in detecting early-stage tumors, and risks associated with invasive procedures.[6]

In this context, chemometrics based on Raman spectroscopy emerges as a promising tool for early diagnosis. Raman spectroscopy offers several advantages for cancer diagnostics. It enables label-free, non-destructive analysis of biofluids with minimal sample preparation, providing real-time information about subtle biochemical changes associated with disease, thereby addressing many of the shortcomings of conventional methods by improving specificity and reducing

*Corresponding author. CIC nanoGUNE BRTA, 20018 San Sebastián, Spain. Department of Physics, University of the Basque Country (UPV/EHU), 20018 San Sebastián, Spain.
**Corresponding author. CIC nanoGUNE BRTA, 20018 San Sebastián, Spain. IKERBASQUE - Basque Foundation for Science, 48009 Bilbao, Spain.
h.hano@nanogune.eu (H. Hano); a.seifert@nanogune.eu (A. Seifert)
https://www.nanogune.eu/en/nanogune/people/harun-hano (H. Hano); https://www.nanogune.eu/es/nanogune/personas/andreas-seifert (A. Seifert)
ORCID(s): 0000-0001-8559-8563 (H. Hano); 0000-0001-5849-4953 (A. Seifert)

false-positive rates[7]. Furthermore, when coupled with chemometrics, it enables the analysis of complex datasets, allowing the detection of subtle patterns that might otherwise go unnoticed.

Applying chemometrics to advanced analytical techniques like Raman spectroscopy offers a powerful tool for disease monitoring [8]. The observed Raman shift is directly proportional to the frequency of molecular vibrations, providing a window into the molecular constitution of biological fluids [9]. Human blood plasma, in particular, a complex biological fluid consisting of proteins, lipids, nucleic acids, carbohydrates, etc., is ideally suited for such an analysis [10]. Raman spectroscopy is ideal for identifying and characterizing different spectral signatures in blood plasma, which are crucial for associating specific molecular changes with lung cancer [11].

Chemometrics extracts information from chemical systems using data-driven methods and offers a promising way to improve lung cancer detection. By analyzing data from vibrational spectroscopy, chemometrics makes an important contribution to identifying specific vibrational groups associated with the disease, which can lead to early diagnosis and the development of targeted treatments [12, 13, 14].

Partial least squares-discriminant analysis (PLS-DA) simplifies the complexity in high-dimensional datasets by reducing the number of predictors to a smaller set of uncorrelated components. In PLS-DA, key elements such as loading weights (LWs)—reflecting the covariance of each feature with response, and the vector of regression coefficients (RCs)—as a single measure of association between each feature and response, play crucial roles. At the same time, both provide a suitable approach for assessing the relevance and importance of each feature in the model. In addition, variable importance in projection (VIP), a cumulative measure of the weight of each feature, is another tool to determine the significance of each feature while using the scores, loading weights, and loadings generated by PLS-DA. In this way, VIP scores complement the PLS-DA framework and provide further insight into which features contribute significantly to high model performance [15, 16, 17].

To increase the performance of classification models, feature selection methods refine spectral datasets by eliminating irrelevant and redundant features, thus reducing complexity while preserving data integrity [18].

Shapley Additive exPlanation (SHAP)-based feature selection with XGBoost learner (SHAP-XGBoost) has proven to be a very powerful method. XGBoost outperforms other learners like Random Forest and Decision Trees by sequentially adding models to correct previous errors, offering efficient regularization, parallel processing, and superior handling of sparse data [19]. Although this approach is widely used in various fields to clarify the predictions of models by evaluating the mean marginal contribution of input features and explaining the decision-making process, it is still little explored in analytical chemistry, especially for spectral datasets from Raman spectroscopy [20]. SHAP values offer a transparent and intuitive way to understand how individual features contribute to model prediction, which is very important in healthcare for understanding decisions made by predictive models for medical diagnostics. SHAP values can aid in feature selection by identifying the most influential features for a given prediction task. By focusing on the features with the highest SHAP values, one can prioritize and potentially simplify the model without sacrificing predictive performance.

Straightforward approaches to reduce high-dimensional datasets are regression analysis or common statistical methods, which are often favored due to their clear interpretability and concise application [21, 22]. In that sense, statistical analysis also serves as a feature selection method by quantitatively evaluating each individual feature regarding classification [23, 24]. The process begins with a hypothesis test in which the null hypothesis (H_0) claims no significant correlation between the feature and the target, while the alternative hypothesis (H_1) suggests a significant correlation. The result is quantified by the p-value, which measures the probability of observing the obtained test statistics [25]. The p-value should not be used as the sole indicator of decision-making in feature selection because it only describes how consistent the data is with the null hypothesis. Therefore, this interpretation should be supported by the inclusion of additional metrics such as the effect size [26], the correlation coefficient [27], or visual representations [28, 29] that allow a more direct assessment of the uncertainty and magnitude of classification.

The present study demonstrates how the detection of lung cancer can be improved by combining Raman spectroscopy with feature selection methods such as regression-based, statistical, and model-agnostic approaches, focusing on subtle vibrational changes in spectra from human blood plasma and their assignment to key wavenumbers. By carefully linking the spectral data with the tumor characteristics, the aim is to show whether these wavenumbers can correlate with the tumor types present in patients. The overall objective is to determine whether there is a correlation between vibrational patterns and the presence of lung cancer through improved classification performance for early diagnosis.

2. Experimental Section

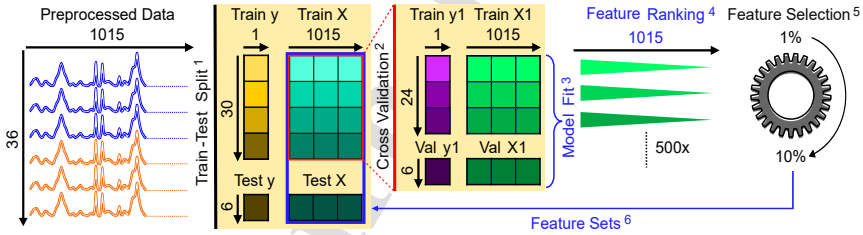
2.1. Sampling and Measurements

This study includes a total of 36 human subjects, eighteen of whom belong to the healthy control group, and eighteen were diagnosed with non-small cell lung carcinoma (NSCLC). Eighteen blood samples were taken from the patients in the Oncology Department of the University Hospital Donostia (San Sebastián, Spain). Samples were collected in ethylenediaminetetraacetic acid (EDTA) tubes and blood plasma was prepared within 1 h of phlebotomy according to the standard protocol. In addition, plasma from 18 healthy donors was collected retrospectively from the Basque Biobank (Bioef). Sampling was performed under the Declaration of Helsinki and with the approval of the local ethics committees (CEIC Euskadi approval number: PI2019170).

Before analysis, aliquots of 1 µL from each subject were deposited onto aluminum foil attached to a microscope slide and then air-dried for 5 min. Aluminum foil offers high reflectivity, stability, flexibility, a low background signal, and cost-effectiveness, which makes it an ideal substrate for enhancing Raman signals in diverse analytical settings [30, 31]. To increase the signal strength while minimizing possible damage to the samples, the laser with a wavelength of 785 nm of the confocal Raman microscope inVia from Renishaw was selected and a spectrometer grating with 1200 L/mm was chosen. Output power was set to 73 mW, and the beam was focused using a 50x long-distance objective. The analysis comprised 20 accumulations, each with an exposure duration of 1 s [32].

2.2. Data Preprocessing

Spectra were collected at 25 different points on the periphery due to the strong accumulation of biomolecules at the edge of the droplet during the drying process. Random and unavoidable cosmic rays were removed using a simple zap function embedded in the Renishaw WiRE 5.4 software. In addition, two preprocessing methods were applied: asymmetric Whittaker baseline correction (lam:100, p:0.01) and standard normal variate (SNV) transformation. Baseline correction minimizes baseline drifts and distortions, addressing asymmetric features and increasing signal-to-noise [33]. SNV transformation was employed to eliminate the interference of baseline drifts due to sample thickness, scattering, instrumental response, etc., without altering the spectral characteristics [34]. Finally, those 25 spectra per sample were averaged to create a single representative spectrum per subject. This last step is crucial for diminishing random noise, improving signal-to-noise ratio, and capturing a comprehensive spectral signature per patient [35].



Scheme 1: Flowchart depicting the technical framework for lung cancer detection using spectral datasets.

2.3. Model Development

The flowchart depicted in Scheme 1 outlines a structured process for analyzing and interpreting spectral datasets to improve the identification of lung cancer through feature selection. This framework unfolds sequentially, as described subsequently:

1. Preprocessed data were segmented into training (85%) and test (15%) sets randomly to develop a robust predictive model and facilitate an impartial evaluation of model performance on new, unseen data.
2. Stratified k-fold cross-validation with 5 folds—a bias-variance trade-off value due to the model's accuracy—was applied to the train set [36]. Accuracy represents the proportion of correct predictions relative to the total number of predictions. This strategy ensures that every subset in train and validation accurately reflects the total dataset, thus mitigating model overfitting and creating reliable performance.

3. Different feature selection methods were applied to derive:
   (a) LWs and RCs based on PLS-DA,
   (b) VIP scores,
   (c) SHAP values,
   (d) p-values from statistical analysis.
4. Features were initially ranked based on the absolute values of LWs, RCs, VIP, SHAP, and p-values, arranged from the highest to the lowest. This systematic approach, which comprises 100 iterations with 5-fold repetition, results in each feature being assessed 500 times. Features were selected each time based on specific percentiles as a threshold, namely 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, and 10%.
5. For the selected features, the frequency was calculated by counting how often they had been selected across all iterations for each percentile. Based on the features ordered by occurrence, the most significant features were extracted again by the same percentiles used in the ranking process. This ensures that only those features that are considered important are retained, those that were selected most frequently.
6. The selected features were applied to both train and test sets from the beginning to ensure that the logistic regression was trained and predicted based on these pertinent features, thereby improving the classification performance.

After data segmentation and stratification, three latent variables were used in PLS-DA based on classification accuracy as indicated in Figure S1 in the Supporting Information. LWs determine how the original data matrix X is transformed into scores (T), essentially defining a subspace for analysis. In contrast, RCs link these transformed scores directly to the response variable (y), outlining the exact predictive model. While the weights identify the directions in the feature space, the RCs quantify their impact on the predictions. Therefore, both LWs and RCs are an essential part of the feature selection process, with both fulfilling different but complementary functions [15, 21].

SHAP is a powerful tool for explaining the output of machine learning models. It assigns so-called Shapley values to each feature in the dataset to determine their contribution to the output of the model. These values help understand the importance of each feature and allow them to be ranked according to their importance for the predictions of the model. This local description is significant because it provides a more accurate representation than global models, as it captures the individual variability among data observations [37, 38].

As a more pragmatic and straightforward method, the Mann-Whitney U test (MWU), a nonparametric statistical test, was used to provide robust insights into each feature in spectra, irrespective of their erratic distribution in each iteration. This test is particularly advantageous when the data is not truly normal, maintaining the integrity of the type I error rate and keeping it efficient and reliable [38]. The resultant p-values are critical for determining the statistical significance of features; however, relying solely on p-values will not provide a comprehensive understanding of the differences observed. To remedy this limitation and improve the interpretative power of the analysis, additional metrics were incorporated. Specifically, the effect size is estimated using Hedge's g, which is a corrected version of Cohen's d, particularly suitable for a limited sample size. This adjustment is necessary to provide more accurate metrics of the extent of the differences between the classes [39]. Furthermore, Spearman's rank correlation coefficient analysis was conducted to assess the relationship between the classes [40].

3. Results and Discussion

3.1. Complete Evaluation

Figure 1 shows the classification performance obtained through the different feature selection methods. Figure 1A includes the result obtained without feature selection, which yielded an accuracy of 0.805 ± 0.061. Data selection for the receiver operating characteristic (ROC) curve in Figure 1B was based on feature percentiles where the models exhibited the highest accuracy. AUC-ROC curve evaluates a binary classifier performance by measuring its ability to separate classes, with higher values indicating better distinction. Particularly notable in this analysis were MWU and SHAP, which stood out for their performance in identifying key wavenumbers relevant to the classification of spectral data.

SHAP, which has the highest accuracy of 0.835 ± 0.013, excels at identifying feature importance and works across different predictive models, offering insights that go beyond conventional methods. However, its high computational demands and large data requirements can limit its use in resource-limited settings. Balancing interpretability and resource efficiency is essential for its practical application [41]. Contrary to accuracy values, the AUC-ROC scores

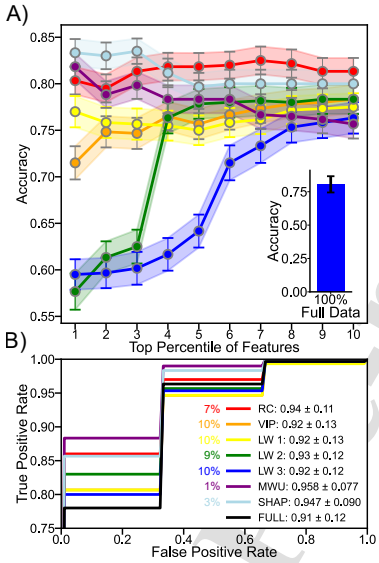


Figure 1: Complete model evaluations based on feature selection methods: (A) Accuracy plot as a function of percentiles of occurrence of features. Error bars indicate the standard error of averaged values, calculated over multiple iterations. Color codes used in Figure 1A are the same as those in Figure 1B. (B) ROC curves for the percentiles of the highest accuracy with the integrated AUC-ROC values in the legend.

are statistically less outstanding in cross-comparison among the methods. SHAP, for example, results in the highest AUC score of 0.947, but its standard error of $\pm$ 0.090 does not distinguish its performance from the other methods that have comparably high error bars, and hence, overlapping average values. This suggests that while SHAP's feature selection capabilities are promising, they may not be as definitive as the high AUC score alone might indicate in our dataset.

In contrast, RCs from PLS-DA, which achieve an accuracy of 0.825 $\pm$ 0.014, represent a more comprehensible yet potent alternative. Despite its slightly lower accuracy compared to SHAP, RCs offer notable advantages in feature selection due to their simplicity and interpretability. In parallel, VIP scores, with an accuracy of 0.783 $\pm$ 0.015, fulfill a crucial function within the PLS-DA framework by leveraging associated weights, loadings, and scores, to identify and prioritize significant wavenumbers. Furthermore, LW1, LW2, and LW3 exhibit consistent performance with accuracies of 0.775 $\pm$ 0.014, 0.783 $\pm$ 0.015, and 0.763 $\pm$ 0.016, respectively, highlighting the significance of each wavenumber on their influence on scores as shown in Figure 2. However, fluctuations in results and performance were observed leading to lower figures of merit compared to RCs.

Although MWU has a competitive accuracy of 0.818 $\pm$ 0.014 compared to PLS-based models, it remains vital for hypothesis testing and data distribution analysis. Despite an almost similar prediction accuracy to other models, MWU provides important insight into the statistical basis of the data and thus complements the general model evaluation.

To sum up the results related to the performance of different feature selection methods, we see that the AUC-ROC scores did not show statistically significant differences among the methods, but accuracy proved to be a more relevant metric for our balanced dataset. Notably, accuracy highlighted statistically significant disparities between the methods, as demonstrated in Figure 1A. Both SHAP and MWU achieved outstanding results with a substantially

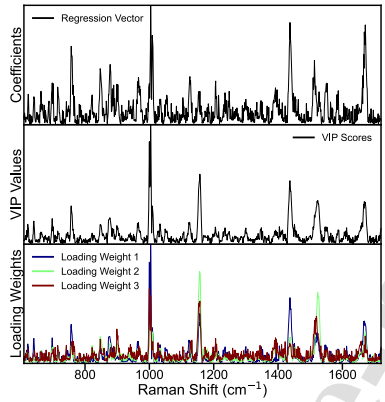


Figure 2: Plot of RCs, VIP scores, and LWs based on their absolute cumulative values following the completion of all iterations.

reduced number of features. Therefore, these results emphasize the need to consider multiple evaluation metrics for a holistic understanding.

3.2. Window-Based Evaluation

To assign significant features in spectra with broadened peaks, window-based feature selection was adopted. This method focuses on specific regions, considering the integral under the curve as explained in detail in Figure S2 of the Supporting Information. To the best of our knowledge, this approach has not been thoroughly explored in the literature, as most research has predominantly applied feature selection methods to entire spectra.

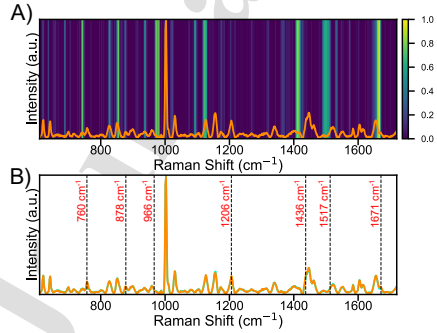


Figure 3: (A) Heatmap of data frequencies across all feature selection methods. (B) Window-based feature selection through RCs. Color codes are orange for lung cancer patients and turquoise for healthy controls.

Figure 3A illustrates a heatmap derived from all feature selection methods applied in this study, revealing the significant wavenumbers based on the number of selections across all iterations and percentiles. This process highlights spectral features that consistently stand out as significant. Before examining possible ambiguities, it is important

to point out that the wavenumbers in Figure 4 are discussed using box-and-whisker plots. In addition, the detailed statistical analysis in Table 1, which includes p-values, effect sizes, and correlation coefficients, provides deeper insight into their significance. While all features contribute to some degree, Table 1 highlights the most frequently selected ones, aligning with our aim to identify the potentially most contributive features in the analysis. This additional analysis is intended to further validate the significance of those selected features. However, feature selection to the entire spectrum may lead to ambiguities, as not all highlighted areas in the heatmap correspond to identifiable vibrational groups, which could lead to misinterpretation of the data. RCs were chosen for this window-based selection due to their stable behavior when dealing with spectral regions, which contrasts with possible fluctuations in MWU and SHAP. SHAP sometimes assigns zero values in scenarios where models do not provide a prediction or when no features are present, which is important in spectral analysis where the contribution of each wavenumber is critical [42]. Therefore, RCs provide a more reliable and straightforward approach for targeted spectral regions, ensuring meaningful and interpretable features. Figure 3B highlights the features using window-based selection, focusing on the classification performance of the top 3% of frequently selected wavenumbers illustrated in Figure S3 in the Supporting Information. The dashed lines indicate the features that achieve the highest model accuracy after applying RCs to these spectral regions. The selected wavenumbers also match these frequently chosen wavenumbers in the full spectrum in Figure 3A, but with fewer features, which validates the efficacy of the method.

Table 1: Statistical analysis was performed on the first and most stable 10 wavenumbers frequently identified in lung cancer detection.

Total Count	Raman Shift (cm^{-1})	Band Assignments	p-value	Effect Size	Correlation Coefficient
2858	1671	amide I	1.06e-04	-1.56e+00	-1.99e-01
2673	1009	ring breathing mode of phenylalanine	7.53e-04	-1.35e+00	1.46e-01
2488	1436	bending or scissoring of CH_2	9.46e-04	-1.21e+00	-4.34e-01
2386	878	C–C stretching	5.97e-04	-1.45e+00	-1.14e-02
2240	760	ring breathing in proteins	2.52e-03	-1.08e+00	4.64e-02
2050	1157	C–C/C–N stretching mode	1.59e-01	3.70e-01	3.95e-01
1829	1517	carotenoids (C=C)	2.61e-01	3.48e-01	4.01e-01
1797	1125	C–C, C–O, C–N stretching	6.27e-05	-1.68e+00	-1.23e-01
1761	966	C–C stretching	2.76e-05	1.70e+00	3.68e-01
1597	847	ring breathing mode of tyrosine	1.03e-01	6.20e-01	-2.45e-01

Using the window-based method can allow for a more nuanced approach, respecting vibrational regions instead of single wavenumbers. This targeted selection contrasts with the holistic method, where critical vibrational signatures could be overlooked due to a lack of region-specific analysis. Notably, only the seven wavenumbers around 1436 cm^{-1}, 1517 cm^{-1}, 760 cm^{-1}, 1206 cm^{-1}, 1671 cm^{-1}, 878 cm^{-1}, and 966 cm^{-1} were selected through the window-based approach [36, 43]. Focusing only on these wavenumbers yields an accuracy of 0.813 ± 0.015 and an AUC-ROC score of 0.914 ± 0.127, indicating that they are more important in feature selection than others that may only contribute noise. These findings highlight the advantages of a selective, contextually grounded feature selection method compared to conventional, full-range selection methods in Raman spectral analysis, which ensure a focus on biochemically relevant information.

3.3. Statistical Analysis

Box-and-whisker plots, as shown in Figure 4, show important differentiations between healthy controls and lung cancer patients at selected wavenumbers. These wavenumbers, which are ordered by frequency of occurrence, reveal varying degrees of statistical significance, shedding light on their potential as diagnostic markers.

In the analysis, 966 cm^{-1} is the most statistically significant wavenumber with high effect size and substantial correlation coefficient, indicating a strong correlation with lung cancer. 1125 cm^{-1} also shows a significant result, despite a smaller correlation coefficient. These wavenumbers were identified through rigorous evaluations due to their high frequency of occurrence and notable differences in the spectral dataset. In contrast, 1157 cm^{-1}, 1517 cm^{-1}, and 847 cm^{-1} exhibit higher p-values, indicating their limited utility in distinguishing lung cancer from healthy controls.

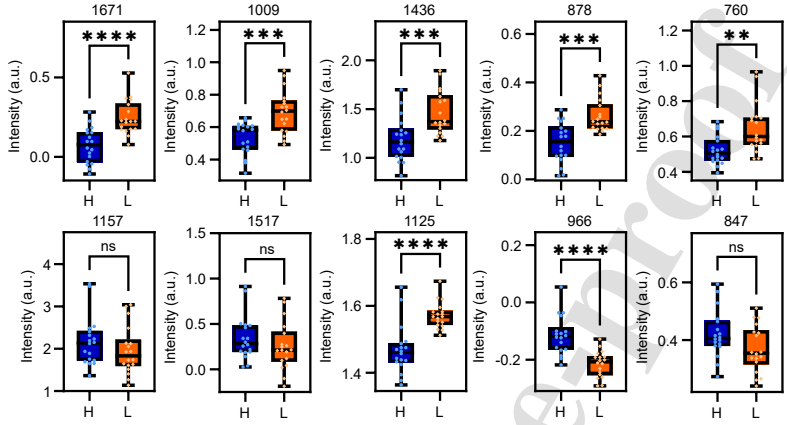


Figure 4: Box-and-whisker plots representing the distribution of significant wavenumbers through complete and window-based evaluations. The wavenumbers are arranged in descending order of their selection frequency. H: healthy controls, L: lung cancer patients. Statistical analysis was performed using the Mann-Whitney U test in GraphPad Prism version 10.2.1 for Windows. ns $p > 0.05$,* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

This lack of significance, despite being frequently selected, requires closer examination to determine their diagnostic value as potential biomarkers.

A look at Table 1 reveals that 1125 cm^{-1} and 966 cm^{-1} have exceptionally low p-values and show effect sizes of -1.68 and 1.70 respectively, indicating different vibrational behaviors indicative of pathological changes. The effect size, a measure of the magnitude of the difference between the two distributions, healthy controls, and lung cancer patients, underscores the robustness and biological relevance of Raman spectra.

When wavenumbers are marked by higher p-values, the whole narrative changes. Although they were initially selected by the models as important classification features, statistical analysis casts doubt on their utility as reliable biomarkers. For instance, peaks around 1517 cm^{-1} and 1206 cm^{-1}, corresponding to Carotenoids (C=C) and C–Ph stretching in proteins, respectively, displayed no statistical significance as mentioned in Figure 4 and Table 1, even though they were selected in both comprehensive and window-based selection methods. This marked difference in the statistical results calls for a critical reassessment of their role and suggests that the frequency of selection alone should not be decisive for their inclusion as key diagnostic markers.

Furthermore, the correlation coefficients offer a further level of interpretation. For instance, 1671 cm^{-1}, with a substantial effect size (-1.56) and a negative correlation coefficient (-0.199), signifies that the intensity associated with this wavenumber decreases as cancer presence increases. This can be indicative of specific molecular changes that require further validation through biological correlation and clinical trials. On the other hand, the positive effect size (1.70) and correlation coefficient (0.368) at 966 cm^{-1} indicate an increase in a certain feature associated with lung cancer, highlighting complex and diverse biochemical reactions.

3.4. Biochemical Assignment

The investigation of selected wavenumbers goes beyond the frequency of occurrence to encompass their biochemical significance. For instance, 1125 cm^{-1} exhibits significant p-value and substantial effect size, indicating considerable alterations in lipids, glycogen, and proteins with C–C, C–O, C–N stretching. The positive and negative effect sizes corresponding to these wavenumbers may indicate distinct vibrational groups involved in cancer development. Additionally, 1664 - 1671 cm^{-1} due to amide I vibrational mode in proteins (α-helix), and 964 - 966 cm^{-1} with ring breathing, C–C stretching in proteins (tryptophan, valyl, prolyl) and lipids are remarkable due to their low p-values and reasonable effect sizes, underscoring their potential as markers for structural changes in proteins. However, the

importance of correlating these statistical indicators with clinical data cannot be overestimated, as the mere presence of statistical significance does not necessarily confirm the usefulness of a biomarker in clinical settings. Moreover, certain wavenumbers like 1157 cm^{-1} (C–C/C–N stretching mode), 1517 cm^{-1} (Carotenoids (C=C)), and 847 cm^{-1} (ring breathing mode of tyrosine) are frequently selected, but they appear statistically non-significant even though their consistent selection and associated vibrational activities suggest an underlying biochemical relevance that should not be overlooked. Additionally, the peak at 1009 cm^{-1}, which represents the ring breathing mode in phenylalanine, is the highest in the Raman spectra, was frequently selected in the models, and hence, shows statistical significance, but does not rank first. This observation underscores that the most prominent peak in the spectra does not always correlate with the highest diagnostic value, advocating for a comprehensive evaluation beyond peak prominence to identify true biochemical markers for lung cancer detection [36, 43].

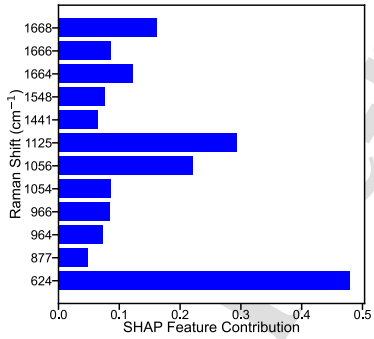


Figure 5: Feature importance in SHAP based on the absolute values generated from selected wavenumbers.

Interestingly, the peaks around 624 cm^{-1}, 877 cm^{-1}, 964 - 966 cm^{-1}, 1054 - 1056 cm^{-1}, 1125 cm^{-1}, 1441 cm^{-1}, 1548 cm^{-1}, and 1664 - 1668 cm^{-1} selected by SHAP for their classification performance based on the incidence within the 3% level of the dataset also confirm our reasoning, as mentioned earlier in the text. In Figure 5, SHAP analysis highlights the peaks around 624 cm^{-1} (C–C twisting in proteins, specifically tyrosine) and 1054–1056 cm^{-1} (associated with =CH bending, C–C and C–O stretching in proteins like phenylalanine, as well as collagen and glycogen) as the most significant. The peak at 1548 cm^{-1} corresponds to C=C stretching in tryptophan, while bands around 1664 - 1671 cm^{-1} denote amide I in proteins (α-helix) and C=C stretching in lipids, reinforcing their relevance in analysis. The bands around 1436 - 1441 cm^{-1} denote the bending or scissoring vibration of CH_2 in lipids and the asymmetric bending of CH_3 in proteins. In addition, 877 - 878 cm^{-1} represents the stretching vibration of C–C in lipids and collagen [36, 43].

Taking into account additional hospital-based information provided by Biogipuzkoa Health Research Institute, a further approach was considered by focusing on possible associations between spectral features and biomarkers in non-small cell lung cancer (NSCLC). Patient records reveal mutations predominantly in epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) genes, such as exon 19 deletion, L861Q, and exon 19 insertion for EGFR, alongside translocation for ALK [44] The study examines whether there are correlations between spectral Raman shifts and these specific genetic mutations. This examination considers the probability that the spectral changes could correspond to genetic abnormalities in the lung cancer samples.

In the study using the bioconjugated Rh6G SERS (surface enhanced Raman scattering) nanotag for multiplex biosensing, the peaks at 1650 cm^{-1} correlate with the prognostic biomarker EGFR, which parallels the bands we observed around 1664 - 1671 cm^{-1}, indicating a possible link to the structural changes typical of EGFR mutations in cancer pathology [45, 46, 47]. Moreover, Raman peaks observed in lung adenocarcinoma tissue showed a significant increase at 1127 cm^{-1}. These alterations parallel our detected peaks at 1125 cm^{-1}, corresponding to specific biomolecular changes as mentioned above [48, 49, 50]. Additionally, the exon 19 deletion, one of the EGFR subtypes, may correspond to the peak around 757 - 760 cm^{-1} identified through our analysis [48]. These correlations indicate that

changes in ring breathing in proteins could be potentially associated with lung cancer. The findings from the SERS study on the detection of pleural effusions, which often indicate the progression of lung cancer, further enhance our analysis. They show distinct vibrational bands at 1555 cm^{-1} corresponding to vibrational groups around 1548 cm^{-1} in our study, reflecting our results, especially the band at 1054 - 1056 cm^{-1}, which is due to higher nucleic acid bases caused by abnormal metabolism of DNA and RNA [47, 51]. While our analysis revealed significant peaks such as 966 cm^{-1} and 1009 cm^{-1} corresponding to protein and phenylalanine vibrations, these have not been highlighted as dominant in other external studies, but are still consistent with structural changes observed in proteins associated with lung cancer pathology in the literature [47, 52]. These results can provide valuable information for future clinical research. Raman spectroscopy can contribute to a deeper understanding of physiological changes unveiled by vibrational groups in the spectra, which in turn could support the development of precise diagnostic and therapeutic tools [53].

4. Conclusions

The presented study highlights the great potential of combining Raman spectroscopy with machine learning models and variable selection strategies to improve lung cancer detection through the analysis of human blood plasma. Meticulous analysis of vibrational groups has led to the identification of specific wavenumbers around 1125 cm^{-1}, 966 cm^{-1}, and 1671 cm^{-1} that show significant associations with lung cancer, providing a new avenue for early, noninvasive diagnosis.

Using feature selection methods, our research reveals that certain wavenumbers, corresponding to distinctive vibrational groups, demonstrate strong correlations with lung cancer. Although the exact relationship between the spectral features and lung cancer is not yet fully understood, their consistent selection in various models and statistical analyses suggests their utility as potential diagnostic markers. Additionally, the study is investigating the effects of the mutations prevalent in non-small cell lung cancer on spectral Raman shifts. Although it is not possible to conclusively determine the direct causality between these mutations and the spectral variations, the identified wavenumbers are part of a broader, more intricate puzzle that requires further validation and exploration in the clinical setting.

Future research should focus on refining these findings through more extensive, diverse datasets and assessing how Raman spectroscopy can be integrated into current diagnostic processes. Advancing the connection between analytical chemistry and clinical oncology will bring us closer to creating more effective, noninvasive tools for lung cancer detection. In summary, while our study represents a significant advancement in the noninvasive identification of lung cancer, it also highlights the importance of continuous research and collaboration between scientists and clinicians to confirm and build on these findings, and to develop new tools for medical diagnostics that will advance personalized medicine.

Acknowledgement

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 Un-ion with the Next Generation EU funds.

References

[1] S. K. Arya, S. Bhansali, Lung cancer and its early detection using biomarker-based biosensors, Chemical Reviews 111 (2011) 6783–6809. doi:10.1021/cr100420s.

[2] R. Gasparri, A. Sabalic, L. Spaggiari, The early diagnosis of lung cancer: Critical gaps in the discovery of biomarkers, Journal of Clinical Medicine 12 (2023). doi:10.3390/jcm12237244.

[3] M. M. Ahsan, S. A. Luna, Z. Siddique, Machine-learning-based disease diagnosis: A comprehensive review, Healthcare 10 (2022). doi:10.3390/healthcare10030541.

[4] N. Caballé-Cervigón, J. L. Castillo-Sequera, J. A. Gómez-Pulido, J. M. Gómez-Pulido, M. L. Polo-Luque, Machine learning applied to diagnosis of human diseases: A systematic review, Applied Sciences 10 (2020). doi:10.3390/app10155135.

[5] H. Hano, B. Suarez, C. H. Lawrie, A. Seifert, Fusion of raman and ftir spectroscopy data uncovers physiological changes associated with lung cancer, International Journal of Molecular Sciences 25 (20) (2024). doi:10.3390/ijms252010936.
URL https://www.mdpi.com/1422-0067/25/20/10936

[6] R. Ledda, G.-C. Funk, N. Sverzellati, The pros and cons of lung cancer screening, European Radiology (2024). doi:10.1007/s00330-024-10939-6.

[7] S. F. El-Mashtoly, K. Gerwert, Diagnostics and therapy assessment using label-free raman imaging, Analytical Chemistry 94 (1) (2022) 120–142. doi:10.1021/acs.analchem.1c04483.

[8] D. Cialla-May, C. Krafft, P. Rösch, T. Deckert-Gaudig, T. Frosch, I. J. Jahn, S. Pahlow, C. Stiebing, T. Meyer-Zedler, T. Bocklitz, I. Schie, V. Deckert, J. Popp, Raman spectroscopy and imaging in bioanalytics, Analytical Chemistry 94 (2021) 86–119. doi:10.1021/acs.analchem.1c03235.

[9] W. R. Premasiri, J. C. Lee, L. D. Ziegler, Surface-enhanced raman scattering of whole human blood, blood plasma, and red blood cells: Cellular processes and bioanalytical sensing, The Journal of Physical Chemistry B 116 (2012) 9376–9386. doi:10.1021/jp304932g.

[10] P. Giamougiannis, C. L. M. Morais, R. Grabowska, K. M. Ashton, N. J. Wood, P. L. Martin-Hirsch, F. L. Martin, A comparative analysis of different biofluids towards ovarian cancer diagnosis using raman microspectroscopy, Analytical and Bioanalytical Chemistry 413 (2021) 911–922. doi:10.1007/s00216-020-03045-1.

[11] X. Chen, J. Gole, A. Gore, Q. He, M. Lu, J. Min, Z. Yuan, X. Yang, Y. Jiang, T. Zhang, C. Suo, X. Li, L. Cheng, Z. Zhang, H. Niu, Z. Li, Z. Xie, H. Shi, X. Zhang, M. Fan, X. Wang, Y. Yang, J. Dang, C. McConnell, J. Zhang, J. Wang, S. Yu, W. Ye, Y. Gao, K. Zhang, R. Liu, L. Jin, Non-invasive early detection of cancer four years before conventional diagnosis using a blood test, Nature Communications 11 (2020) 3475. doi:10.1038/s41467-020-17316-z.

[12] J. Chapman, V. K. Truong, A. Elbourne, S. Gangadoo, S. Cheeseman, P. Rajapaksha, K. Latham, R. J. Crawford, D. Cozzolino, Combining chemometrics and sensors: Toward new applications in monitoring and environmental analysis, Chemical Reviews 120 (2020) 6048–6069. doi:10.1021/acs.chemrev.9b00616.

[13] N. M. Ralbovsky, I. K. Lednev, Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 219 (2019) 463–487. doi:10.1016/j.saa.2019.04.067.

[14] M. Ishigaki, Y. Maeda, A. Taketani, B. B. Andriana, R. Ishihara, K. Wongravee, Y. Ozaki, H. Sato, Diagnosis of early-stage esophageal cancer by raman spectroscopy and chemometric techniques, Analyst 141 (2016) 1027–1033. doi:10.1039/C5AN01323B.

[15] G. Palermo, P. Piraino, H.-D. Zucht, Performance of pls regression coefficients in selecting variables for each response of a multivariate pls for omics-type data, Advances and Applications in Bioinformatics and Chemistry 2 (2009) 57–70, doi: 10.2147/aabc.s3619. doi:10.2147/aabc.s3619.

[16] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for pls, Journal of Chemometrics 7 (1993) 45–59. doi:10.1002/cem.1180070104.

[17] H. Abdi, Partial least squares regression and projection on latent structure regression (pls regression), WIREs Computational Statistics 2 (2010) 97–106. doi:10.1002/wics.51.

[18] P. Rana, P. Thai, T. Dinh, P. Ghosh, Relevant and non-redundant feature selection for cancer classification and subtype detection, Cancers 13 (2021). doi:10.3390/cancers13174297.

[19] H. Wang, Q. Liang, J. Hancock, T. Khoshgoftaar, Feature selection strategies: a comparative analysis of shap-value and importance-based methods, Journal of Big Data 11 (03 2024). doi:10.1186/s40537-024-00905-w.

[20] L. Bellantuono, R. Tommasi, E. Pantaleo, M. Verri, N. Amoroso, P. Crucitti, M. D. Gioacchino, F. Longo, A. Monaco, A. M. Naciu, A. Palermo, C. Taffon, S. Tangaro, A. Crescenzi, A. Sodo, R. Bellotti, An explainable artificial intelligence analysis of raman spectra for thyroid cancer diagnosis, Scientific Reports 13 (2023) 16590. doi:10.1038/s41598-023-43856-7.

[21] T. Mehmood, S. Sæbø, K. H. Liland, Comparison of variable selection methods in partial least squares regression, Journal of Chemometrics 34 (2020) e3226, e3226 cem.3226. doi:10.1002/cem.3226.

[22] I. M. Johnstone, D. M. Titterington, Statistical challenges of high-dimensional data, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367 (2009) 4237–4253. doi:10.1098/rsta.2009.0159.

[23] H. Cook, A. Crisford, K. Bourdakos, D. Dunlop, R. O. C. Oreffo, S. Mahajan, "spectromics": Holistic optical assessment of human cartilage via complementary vibrational spectroscopy for osteoarthritis diagnosis, medRxiv (2023) 2023.12.21.23300367doi:10.1101/2023.12.21.23300367.

[24] F. E. R. Woods, S. Chandler, N. Sikora, R. Harford, A. Souriti, H. Gray, H. Wilkes, C. Lloyd-Bennett, D. A. Harris, P. R. Dunstan, An observational cohort study to evaluate the use of serum raman spectroscopy in a rapid diagnosis center setting, Clinical Spectroscopy 4 (2022) 100020. doi:10.1016/j.clispe.2022.100020.

[25] F. Emmert-Streib, M. Dehmer, Understanding statistical hypothesis testing: The logic of statistical inference, Machine Learning and Knowledge Extraction 1 (2019) 945–961. doi:10.3390/make1030054.

[26] G. Sullivan, R. Feinn, Using effect size—or why the p value is not enough, Journal of graduate medical education 4 (2012) 279–282. doi:10.4300/JGME-D-12-00156.1.

[27] J. I. Githaiga, H. K. Angeyo, K. A. Kaduki, W. D. Bulimo, D. K. Ojuka, Quantitative raman spectroscopy of breast cancer malignancy utilizing higher-order principal components: A preliminary study, Scientific African 14 (2021) e01035. doi:10.1016/j.sciaf.2021.e01035.

[28] S. Bonovas, D. Piovani, On p-values and statistical significance, Journal of Clinical Medicine 12 (2023). doi:10.3390/jcm12030900. URL https://www.mdpi.com/2077-0383/12/3/900

[29] P. Basran, G. Palma, C. Baldock, P-values should not be used for decision making in the practice of clinical medical physics, Physical and Engineering Sciences in Medicine 44 (2021) 1003–1006. doi:10.1007/s13246-021-01068-1.

[30] S. Aitekenov, A. Sultangaziyev, A. Boranova, A. Dyussupova, A. Ilyas, A. Gaipov, R. Bukasov, Sers for detection of proteinuria: A comparison of gold, silver, al tape, and silicon substrates for identification of elevated protein concentration in urine, Sensors 23 (2023) 1605. doi:10.3390/s23031605.

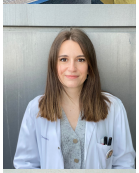
[31] L. Cui, H. J. Butler, P. L. Martin-Hirsch, F. L. Martin, Aluminium foil as a potential substrate for atr-ftir, transflection ftir or raman spectrochemical analysis of biological specimens, Anal. Methods 8 (2016) 481–487. doi:10.1039/C5AY02638E.

[32] A. Synytsya, M. Judexova, D. Hoskovec, M. Miskovicova, L. Petruzelka, Raman spectroscopy at different excitation wavelengths (1064, 785 and 532 nm) as a tool for diagnosis of colon cancer, Journal of Raman Spectroscopy 45 (2014) 903–911. doi:https://doi.org/10.1002/jrs.4581.

[33] S.-J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (2015) 250–257. doi:10.1039/C4AN01061B.

[34] Åsmund Rinnan, F. van den Berg, S. B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry 28 (2009) 1201–1222. doi:https://doi.org/10.1016/j.trac.2009.07.007.
[35] N. Blake, R. Gaifulina, L. D. Griffin, I. M. Bell, G. M. H. Thomas, Machine learning of raman spectroscopy data for classifying cancers: A review of the recent literature, Diagnostics 12 (2022) 1491. doi:10.3390/diagnostics12061491.
[36] H. Hano, C. H. Lawrie, B. Suarez, A. P. Lario, I. E. Echeverría, J. G. Mediavilla, M. I. C. Cruz, E. Lopez, A. Seifert, Power of light: Raman spectroscopy and machine learning for the detection of lung cancer, ACS Omega 9 (2024) 14084–14091. doi:10.1021/acsomega.3c09537.
[37] W. E. Marcílio, D. M. Eler, From explanations to feature selection: assessing shap values as feature selection mechanism, 2020, pp. 340–347. doi:10.1109/SIBGRAPI51738.2020.00053.
[38] F. Orcan, Parametric or non-parametric: Skewness to test normality for mean comparison, International Journal of Assessment Tools in Education 7 (2) (2020) 255–265. doi:10.21449/ijate.656077.
[39] J.-C. Goulet-Pelletier, D. Cousineau, A review of effect sizes and their confidence intervals, part i: The cohen's d family, The Quantitative Methods for Psychology 14 (2018) 242–265. doi:10.20982/tqmp.14.4.p242.
[40] J. Hauke, T. Kossowski, Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data, Quaestiones Geographicae 30 (2011) 87, copyright - Copyright Versita Jun 2011 Última actualización - 2023-12-04. doi:https://doi.org/10.2478/v10117-011-0021-1.
[41] G. V. den Broeck, A. Lykov, M. Schleich, D. Suciu, On the tractability of shap explanations, Journal of Artificial Intelligence Research 74 (2022) 851–886.
[42] D. Fryer, I. Strümke, H. Nguyen, Shapley values for feature selection: The good, the bad, and the axioms, IEEE Access 9 (2021) 144352–144360. doi:10.1109/ACCESS.2021.3119110.
[43] A. Sinica, K. Brožáková, T. Brůha, J. Votruba, Raman spectroscopic discrimination of normal and cancerous lung tissues, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 219 (2019) 257–266. doi:10.1016/j.saa.2019.04.055.
[44] P. Villalobos, I. I. Wistuba, Lung cancer biomarkers, Hematology/Oncology Clinics of North America 31 (2017) 13–29, lung Cancer. doi:https://doi.org/10.1016/j.hoc.2016.08.006.
[45] U. S. Dinish, G. Balasundaram, Y.-T. Chang, M. Olivo, Actively targeted in vivo multiplex detection of intrinsic cancer biomarkers using biocompatible sers nanotags, Scientific Reports 4 (2014) 4075. doi:10.1038/srep04075.
[46] S. Kaminaka, H. Yamazaki, T. Ito, E. Kohda, H. o Hamaguchi, Near-infrared raman spectroscopy of human lung tissues: possibility of molecular-level cancer diagnosis, Journal of Raman Spectroscopy 32 (2001) 139–141. doi:https://doi.org/10.1002/jrs.680.
[47] H. Wang, S. Zhang, L. Wan, H. Sun, J. Tan, Q. Su, Screening and staging for non-small cell lung cancer by serum laser raman spectroscopy, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 201 (2018) 34–38. doi:https://doi.org/10.1016/j.saa.2018.04.002.
[48] L. Wang, Z. Zhang, L. Huang, W. Li, Q. Lu, M. Wen, T. Guo, J. Fan, X. Wang, X. Zhang, J. Fang, X. Yan, Y. Ni, X. Li, Evaluation of raman spectroscopy for diagnosing egfr mutation status in lung adenocarcinoma, Analyst 139 (2014) 455–463. doi:10.1039/C3AN01381B.
[49] Y.-C. Ou, J. A. Webb, C. M. O'Brien, I. J. Pence, E. C. Lin, E. P. Paul, D. Cole, S.-H. Ou, M. Lapierre-Landry, R. C. DeLapp, E. S. Lippmann, A. Mahadevan-Jansen, R. Bardhan, Diagnosis of immunomarkers in vivo via multiplexed surface enhanced raman spectroscopy with gold nanostars, Nanoscale 10 (2018) 13092–13105. doi:10.1039/C8NR01478G.
[50] X. Li, T. Yang, C. S. Li, D. Wang, Y. Song, L. Jin, Detection of egfr mutation in plasma using multiplex allele-specific pcr (mas-pcr) and surface enhanced raman spectroscopy, Scientific Reports 7 (2017) 4771. doi:10.1038/s41598-017-05050-4.
[51] A. A. Kowalska, M. Czaplicka, A. B. Nowicka, I. Chmielewska, K. Kędra, T. Szymborski, A. Kamińska, Lung cancer: Spectral and numerical differentiation among benign and malignant pleural effusions based on the surface-enhanced raman spectroscopy, Biomedicines 10 (2022) 993. doi:10.3390/biomedicines10050993.
[52] S. Kim, B. H. Choi, H. Shin, K. Kwon, S. Y. Lee, H. B. Yoon, H. K. Kim, Y. Choi, Plasma exosome analysis for protein mutation identification using a combination of raman spectroscopy and deep learning, ACS Sensors 8 (2023) 2391–2400. doi:10.1021/acssensors.3c00681.
[53] X. Zhou, C. Chen, E. Zuo, C. Chen, X. Lv, Cross branch co-attention network multimodal models based on raman and ftir spectroscopy for diagnosis of multiple selected cancers, Applied Soft Computing 166 (2024) 112204. doi:10.1016/j.asoc.2024.112204.

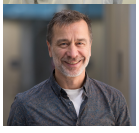
Harun is always passionate about diagnostic systems. His research primarily focuses on the synthesis of nanostructures, the application of vibrational spectroscopy and the development of machine learning models for disease diagnosis. During his M.Sc. studies in Nanoscience and Nanotechnology in Turkey, he worked on a national project developing glucose biosensors based on graphene foam/α-Fe_2O_3 nanocomposite. Currently, he is pursuing a Ph.D. in Physics of Nanostructures and Advanced Materials at the University of the Basque Country while working as a predoctoral researcher at CIC nanoGUNE in Spain. His doctoral research focuses on the diagnosis of lung cancer using spectroscopic methods supported by machine learning, contributing to advancements in biomedical diagnostics



Beatriz Suárez achieved her doctorate in Biomedicine and Applied Medicine from the University of Navarra (Pamplona, Spain) in 2021. She conducted her thesis at the Applied Medical Research Center, supported by a scholarship for industrial doctorates awarded by the Government of Navarra. Subsequently, she joined the Molecular Oncology group at Biogipuzkoa Health Research Institute (San Sebastian, Spain), where she began her research on identifying predictive biomarkers in liquid biopsies of patients with solid tumors. In 2022, she received the Margarita Salas grant from the University of the Basque Country (with European Union-Next Generation EU funds), allowing her to continue her liquid biopsy studies together with the Nursing Department of the Medicine Faculty at the University of the Basque Country (Bilbao, Spain).

Born in 1978. He obtained his PhD (Cum Laude) in Chemistry from the Autonomous University of Barcelona, Spain. He was employed, first as a post-doctoral student (2007 – 2009) and afterwards as an Associate Professor (2010 – 2019) at the Department of Food Science of the University of Copenhagen, Denmark. In 2017 he was, at the same time, Guest Professor at the Federal University of Pernambuco, Brazil. He is now a Distinguished Professor at the Department of Analytical Chemistry of the University of the Basque Country, Spain, and a Research Professor of IKERBASQUE, The Basque Foundation for Science. Current research interests include hyperspectral and digital image analysis and the application of Chemometrics (i.e. Machine and Deep Learning). He has authored more than 180 publications (150+ peer-reviewed papers, books, book chapters, proceedings, etc.) and has given more than 60 conferences and courses at international meetings. Jose has supervised or is currently supervising several MSc, PhD and Post Docs, and he is an editorial board member of four scientific journals within chemometrics, pharmaceutical sciences and analytical chemistry. Moreover, he received the "2014 Chemometrics and Intelligent Laboratory Systems Award" for his achievements in the field of Chemometrics and the "2019 Tomas Hirschfeld Award" for his achievements in the field of Near Infrared. He is editor of the book "Hyperspectral Imaging. Volume 32. Elsevier. ISBN: 9780444639776. In Data Handling in Science and Technology.

Charles Lawrie obtained his BA (Hons), in MA and doctorate from Trinity College, University of Oxford. After several posts as researcher and PI in the Medical Sciences division of Oxford University, he currently holds the post Ikerbasque research professor and Director of Oncology research at the Biogipuzkoa Research Institute in Donostia-San Sebastian, Spain. He also holds the position of University Research lecturer (University of Oxford) and is a Visiting Professor at Shanghai University.



Andreas Seifert has dedicated his career to optics and photonics. In brief: From space telescopes over optical microsystems to nanophotonics. After completing his doctorate in physics, he worked in the optical industry at Carl Zeiss, Germany, where he was responsible for X-ray space optics and synchrotron optics. He then worked at the University of Freiburg in the field of optical microsystems. Since 2015, Andreas Seifert has been Ikerbasque Research Professor at CIC nanoGUNE in San Sebastian, Spain. His research covers photonics and plasmonics in combination with nanotechnology, biomedical engineering and artificial intelligence, mainly for applications in medical diagnostics, food quality control and environmental monitoring

A.3 Data fusion strategies

Harun Hano, Beatriz Suarez, Charles H. Lawrie, and Andreas Seifert. **Fusion of Raman and FTIR Spectroscopy Data Uncovers Physiological Changes Associated with Lung Cancer**. International Journal of Molecular Sciences 2024, 25, 10936. DOI: 10.3390/ijms252010936

Impact Factor: 4.9

International Journal of
Molecular Sciences

MDPI

Article

Fusion of Raman and FTIR Spectroscopy Data Uncovers Physiological Changes Associated with Lung Cancer

Harun Hano [1,2,*], Beatriz Suarez [3,4], Charles H. Lawrie [4,5,6,7] and Andreas Seifert [1,5,*]

1 CIC nanoGUNE BRTA, 20018 San Sebastián, Spain
2 Department of Physics, University of the Basque Country (UPV/EHU), 20018 San Sebastián, Spain
3 Faculty of Nursing and Medicine, University of the Basque Country (UPV/EHU), 48940 Leioa, Spain; beatriz.suarez@ehu.eus
4 Biogipuzkoa Health Research Institute, 20014 San Sebastián, Spain; charles.lawrie@bio-gipuzkoa.eus
5 IKERBASQUE—Basque Foundation for Science, 48009 Bilbao, Spain
6 Sino-Swiss Institute of Advanced Technology (SSIAT), University of Shanghai, Shanghai 201800, China
7 Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DU, UK
* Correspondence: h.hano@nanogune.eu (H.H.); a.seifert@nanogune.eu (A.S.); Tel.: +34-943-574-045 (A.S.)

Abstract: Due to the high mortality rate, more effective non-invasive diagnostic methods are still needed for lung cancer, the most common cause of cancer-related death worldwide. In this study, the integration of Raman and Fourier-transform infrared spectroscopy with advanced data-fusion techniques is investigated to improve the detection of lung cancer from human blood plasma samples. A high statistical significance was found for important protein-related oscillations, which are crucial for differentiating between lung cancer patients and healthy controls. The use of low-level data fusion and feature selection significantly improved model accuracy and emphasizes the importance of structural protein changes in cancer detection. Although other biomolecules such as carbohydrates and nucleic acids also contributed, proteins proved to be the decisive markers found using this technique. This research highlights the power of these combined spectroscopic methods to develop a non-invasive diagnostic tool for discriminating lung cancer from healthy state, with the potential to extend such studies to a variety of other diseases.

Keywords: data fusion; vibrational spectroscopy; chemometrics; feature selection; photonic diagnostics

Citation: Hano, H.; Suarez, B.; Lawrie, C.H.; Seifert, A. Fusion of Raman and FTIR Spectroscopy Data Uncovers Physiological Changes Associated with Lung Cancer. *Int. J. Mol. Sci.* **2024**, *25*, 10936. https://doi.org/10.3390/ijms252010936

Academic Editor: Vasile Chiș

Received: 6 September 2024
Revised: 1 October 2024
Accepted: 8 October 2024
Published: 11 October 2024

1. Introduction

Lung cancer remains the most common cause of cancer death worldwide, with around 1.8 million deaths per year [1]. The prognosis is particularly poor, with a relative 5-year survival rate, as the diagnosis is made at a late stage, which limits the targeted treatment options [2]. Conventional diagnostic methods such as computed tomography (CT), sputum cytology, biopsy and bronchoscopy are often inadequate for early detection due to their cost, time-consuming nature and lack of sensitivity. Such limitations have paved the way for modern computer-aided diagnostic techniques, which represent a promising alternative as they offer less invasive and more patient-friendly methods without compromising diagnostic accuracy. These techniques have several advantages, including error reduction, faster results, greater efficiency and high reproducibility [3,4].

In recent years, the ability to collect large datasets with modern analytical instruments has proven extremely useful for analyzing, measuring and monitoring various biosamples [5]. In this context, vibrational spectroscopy, particularly Raman and Fourier-transform infrared spectroscopy (FTIR), stands out because of its high sensitivity to changes at the molecular level. These techniques are non-invasive, non-destructive and free of reagents and waste and provide detailed information on the composition and structural conformation of certain types of molecules [6]. FTIR spectroscopy measures the absorbance of infrared light by a sample, revealing insights into molecular vibrations, chemical bonds

Int. J. Mol. Sci. 2024, 25, 10936

and functional groups. In contrast, Raman spectroscopy measures the inelastic scattering of light and provides complementary information of the molecular structure of the biosample [7].

It has been shown that the integration of data from Raman and FTIR spectroscopy using data-fusion techniques can significantly improve the robustness and accuracy of prediction models [8]. Raman spectroscopy detects subtle molecular vibrations of molecules with polarizability, while FTIR spectroscopy provides detailed information about molecular bonds and functional groups that have a permanent dipole moment. By merging these datasets, researchers can achieve a unified view of the molecular composition, leading to more accurate and reliable analytical outcomes [9]. The combination of Raman and FTIR measurements can even be accomplished in a single spectroscopic instrument that combines both techniques such that the sample can be measured at the same time and same position by both spectroscopic methods [10]. Data-fusion strategies can be categorized into three types: low-level, mid-level and high-level data fusion. In low-level data fusion (LLDF), data matrices from several sources are directly linked together to create a comprehensive dataset that covers the entire range of measured variables. Mid-level data fusion (MLDF) addresses the problem of high dimensionality by selecting or reducing features from the data prior to fusion, reducing data complexity while preserving important information and enabling more efficient model training. High-level data fusion (HLDF) combines the predictive results of models developed for each data source and improves predictive accuracy by leveraging the strengths of each individual model [9,11,12].

Here, we present new developments for an effective computer-aided diagnostic approach for lung cancer by integrating Raman and FTIR spectroscopy with advanced data-fusion techniques. This research aims to identify the most indicative and discriminative biomolecular groups for the detection of lung cancer. The data from both spectroscopy techniques are combined to utilize the complementary information they provide to develop more reliable prediction models. Data fusion is applied and assessed for all three levels, low-, mid- and high-level fusion techniques. The results are further improved by implementing specific feature-selection methods.

2. Results and Discussion

2.1. Model Performance

Table 1 presents the performance of various configurations using Raman and FTIR spectroscopy data, and highlights the impact of FS, FR and different data-fusion strategies on model performance. A critical evaluation is carried out here, focusing on the most important comparative findings.

Without data fusion, Option 1, Raman spectroscopy consistently outperforms FTIR in terms of accuracy. Specifically, Raman spectroscopy with FS achieves a notably higher accuracy of 0.85 with fewer features compared to the full spectral range (0.81) and feature reduction (FR) (0.84). Similarly, FTIR spectroscopy saw its base accuracy improve from 0.79 to 0.84 with FS, while FR yields a comparable result (0.79). These results highlight the superior performance of FS in isolating the most relevant features and reducing noise, thus significantly enhancing model accuracy. Conversely, FR simplifies data complexity and provides moderate improvements but may overlook some vital information. The area under the curve (AUC) of the receiver operating characteristics (ROC), shown in Figure 1, further supports these accuracy results. Raman spectroscopy achieves the highest AUC value of 0.92, exceeding the AUC value for the entire spectral range of FTIR at 0.92. Remarkably, the fingerprint region (FP) in FTIR achieves an AUC of 0.88. Although these values are somewhat lower, the proximity of these values indicates that a significant portion of the diagnostic information in FTIR likely comes from the fingerprint region, indicating that while the full spectral region provides a broader context, the fingerprint region alone captures most of the critical diagnostic features. It is important to note that for Raman spectroscopy, the spectra were only collected in the spectral range of 610–1720 cm^{-1}, which is considered to contain most of the biological information and effectively serves as the

Int. J. Mol. Sci. 2024, 25, 10936

fingerprint region in this context, whereas in FTIR, we specifically distinguished between the full range (Full: 400–4000 cm^{-1}) and the fingerprint region (FP: 400–1800 cm^{-1}) as indicated in Table 1.

Table 1. Comparison of various data-fusion methods and their accuracy.

Option	Data Fusion	Spectral Range for FTIR	Method	Data Range	Number of Features	Accuracy
1	No Fusion	Full	1. Raman	100%	1015	0.8119 ± 0.0035
			2. Raman (FS)	5%	51	0.8539 ± 0.0056
			3. Raman (FR)	6PCs	6	0.8378 ± 0.0060
			4. FTIR	100%	1868	0.7886 ± 0.0037
			5. FTIR (FS)	4%	75	0.8425 ± 0.0058
			6. FTIR (FR)	8PCs	8	0.7928 ± 0.0068
		FP	7. FTIR	100%	727	0.7567 ± 0.0033
			8. FTIR (FS)	1%	8	0.8419 ± 0.0057
			9. FTIR (FR)	5PCs	5	0.7633 ± 0.0068
2	LLDF	Full	10. Raman + FTIR	100%	2883	0.8625 ± 0.0035
			11. Raman + FTIR + FS	6%	173	0.9922 ± 0.0015
			12. Raman + FTIR + FR	6PCs	6	0.8711 ± 0.0034
		FP	13. Raman + FTIR	100%	1742	0.8592 ± 0.0037
			14. Raman + FTIR + FS	10%	175	0.9497 ± 0.0039
			15. Raman + FTIR + FR	5PCs	5	0.8681 ± 0.0031
3	MLDF	Full	16. Raman (FS) + FTIR (FS)	5% + 4%	126	0.8472 ± 0.0039
			17. Raman (FR) + FTIR (FR)	6PCs + 8PCs	14	0.8425 ± 0.0034
		FP	18. Raman (FR) + FTIR (FR)	5% + 1%	59	0.7972 ± 0.0035
			19. Raman (FR) + FTIR (FR)	6PCs + 5PCs	11	0.8583 ± 0.0032
4	HLDF	Full	20. Raman + FTIR	100%	1015/1868	0.8383 ± 0.0024
			21. Raman (FS) + FTIR (FS)	5%/4%	51/75	0.8131 ± 0.0023
			22. Raman (FR) + FTIR (FR)	6PCs\8PCs	6/8	0.8383 ± 0.0024
		FP	23. Raman + FTIR	100%	1015/727	0.8319 ± 0.0024
			24. Raman (FS) + FTIR (FS)	5%\1%	51/8	0.7989 ± 0.0035
			25. Raman (FR) + FTIR (FR)	6PCs \5PCs	6/5	0.8319 ± 0.0024

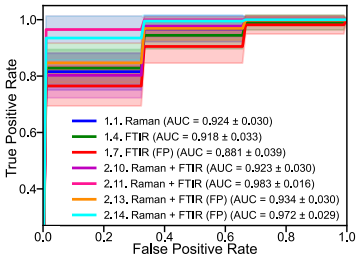


Figure 1. AUC-ROC scores for the selected data-fusion configurations, corresponding to Table 1.

LLDF, Option 2, which combines spectral Raman and FTIR data, demonstrates significant enhancements in model accuracy. When using the combined full spectral range, the accuracy reaches 0.86. Applying FS remarkably improves the accuracy to 0.99 with a substantially reduced number of features, highlighting the significant impact of FS in extracting the most discriminative information from both datasets. In contrast, FR achieves

Int. J. Mol. Sci. 2024, 25, 10936

an accuracy of 0.87, which is higher than using individual methods but not as high as FS. These results indicate that FS is more effective than FR in LLDF, as it accurately isolates critical features and thereby significantly improves model performance. The AUC-ROC values in Figure 1 also show an improvement in accuracy. For the fused data, AUC calculates to 0.92, which increases substantially to 0.98 when applying FS. Moreover, the fingerprint region (FP) alone demonstrates notable AUC values of 0.93 and 0.97 using FS. These high AUC values show that the combination of spectral Raman and FTIR data, especially in feature selection, significantly improves the discriminatory power of the model. The strong performance of the fingerprint region suggests that it contains most of the critical diagnostic features, making it a valuable component for improving prediction accuracy.

MLDF, Option 3, where FS or FR is applied to both Raman and FTIR datasets before they are combined, shows remarkable improvements in model performance. FS achieves an accuracy of 0.85, which illustrates a significant increase in performance through the selection of critical features. Interestingly, in the FP region, FR proves more effective than FS, achieving an accuracy of 0.86 compared to 0.80. Evidently, FR captures better essential information with fewer components in specific spectral regions. These results emphasize the importance of tailoring dimensionality-reduction techniques to the characteristics of the data, as FR appears to be better able to distill critical information into a reduced feature set in the FP region. MLDF thus benefits from a differentiated application of dimensionality-reduction strategies, whereby the choice between FS and FR should be determined by the specific data characteristics.

HLDF, Option 4, combines the predictions from individual models trained on the separate Raman and FTIR data blocks by averaging their predicted probabilities. For HLDF, the combination of the entire spectral range achieves an accuracy of 0.84. FS gives a slightly lower accuracy of 0.81 with fewer features, while FR maintains the same level of accuracy. In the FP region, both FS and FR achieve a similar accuracy of 0.83, indicating that the choice between FS and FR depends on the dataset and spectral range. These outcomes show that HLDF is an effective strategy for integrating predictions from different data sources, although the optimal dimensionality-reduction technique must be tailored to the specific characteristics of the data. Neither FS nor FR were consistently better than the others, underlining the importance of dataset-specific assessment.

2.2. Graphical Representations

The high performance of the model observed in Section 2.1 is visually explained by the score plots in Figure 2. Both Raman and FTIR show a clear class separation, each with a unique pattern, indicating a strong discriminating power. The combined score plot after data fusion and block scaling shows an even clearer separation of classes. The score plots do not reflect the whole truth, as only a pattern with greatly reduced information can be visually represented, but it can be seen that data fusion improves the classification for linear models and thus the predictive power. By integrating the complementary data with high holistic information from Raman and FTIR spectroscopy, this approach provides detailed and reliable spectroscopic information, making it a valuable tool for cancer diagnosis.

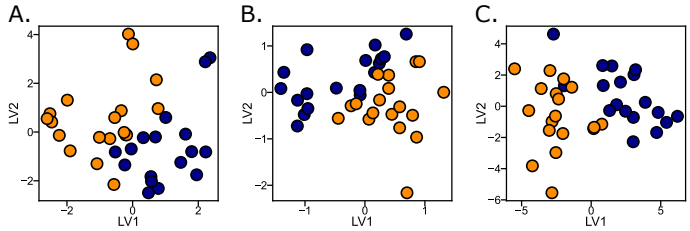


Figure 2. Score plots of (**A**) Raman, (**B**) FTIR and (**C**) the combined data from Raman and FTIR. Color codes are: orange for lung cancer patients and dark blue for healthy controls.

Int. J. Mol. Sci. 2024, 25, 10936

2.3. Biochemical Assignment

Figure 3 illustrates the vibrations assigned to specific biomolecular groups, showing their contribution to the observed separation and highlighting their significance in cancer diagnosis. The regression coefficients in Figure 3C–F provide insights into the specific molecular features responsible for the class separation and allow a deeper understanding of the underlying biochemical differences captured by each spectroscopic technique. Standard error in Figure 3A,B makes the variation in the spectra appear very small, almost invisible, indicating that the variability in the mean values is much smaller. This occurs because the standard error measures the precision of the sample mean—how much the sample mean is expected to vary from the true population mean—rather than the spread of individual data points.

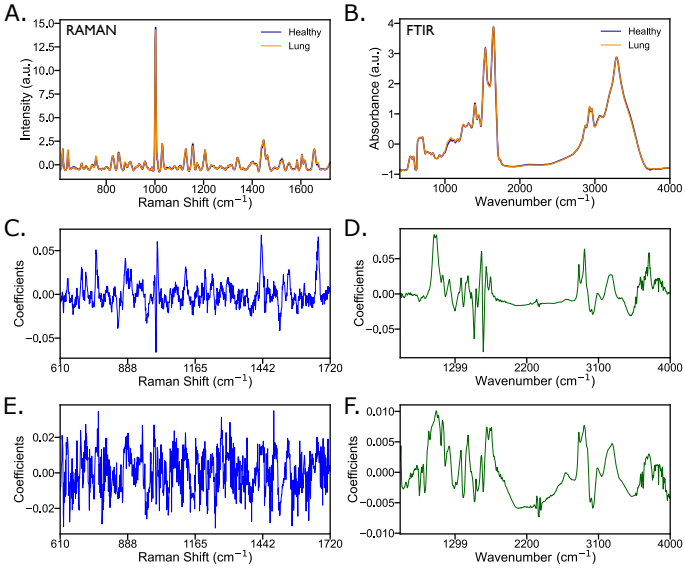


Figure 3. Raman and FTIR spectra with mean and standard error (**A**,**B**) and corresponding regression coefficients before (**C**,**D**) and after (**E**,**F**) block scaling, illustrating the impact of this technique on spectral data analysis. Color codes: Raman data in blue (**C**,**E**), FTIR data in green (**D**,**F**).

High model performance observed in Section 2.1 can be attributed to the most important features identified by these two spectroscopic methods. As shown in Figure 3, the regression vectors provide an overall scheme for these differences and highlight the key biochemical groups that contribute to segregation. The analysis of these vectors shows which specific molecular characteristics are most influential for the differentiation between the classes. Box-and-whisker plots based on statistical analysis per wavenumber, as shown in Figure 4 and as assigned in Table S1 in the Supplementary Information [13–19], show the most significant features (wavenumbers) selected for their high predictive power by LLDF of Raman and FTIR with FS, as found by method 2.11 in Table 1. From the initial 173 features selected, recursive feature elimination (RFE) was applied to narrow down to the most important 20 features. These plots show the distribution of specific biochemical groups, providing insight into their role in class separation. The reason for focusing on these 20 features is that this selection method yields the highest performance, as demonstrated by the results in Section 2.1.

Int. J. Mol. Sci. 2024, 25, 10936

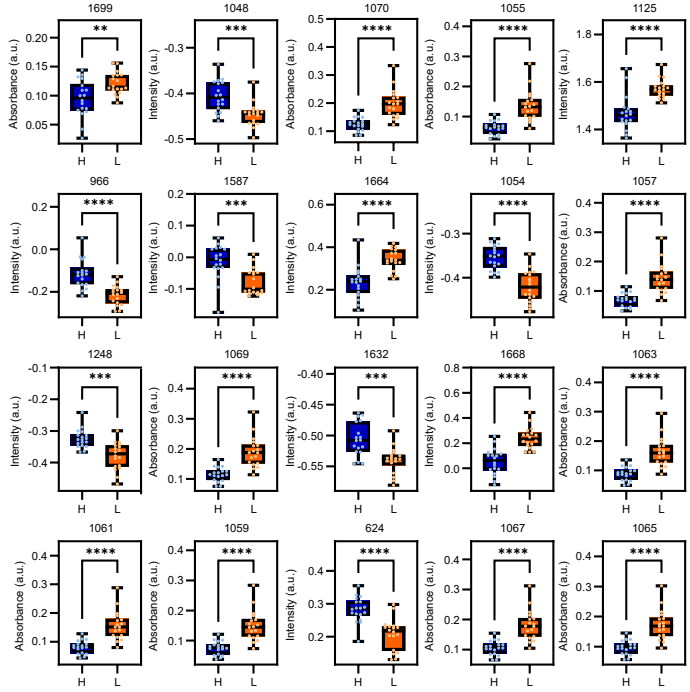


Figure 4. Box−and−whisker plots representing the top 20 wavenumbers obtained through data fusion, arranged in descending order of feature importance. All features show statistically significant differences between healthy controls (H) and lung cancer patients (L). Statistical analysis was performed using the Mann–Whitney U test in GraphPad Prism version 10.2.1 for Windows. ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

For Raman spectroscopy, peaks at 624 cm^{-1} are associated with C–C twisting in lipids and proteins (phenylalanine). The peak at 966 cm^{-1}, linked to CH_3 deformation, ring breathing, C–C stretching in proteins (tryptophan, valyl, prolyl) and lipids, is among the most critical features. Peaks between 1048–1054 cm^{-1} associated with =CH bending, C–C and C–O stretching in proteins (phenylalanine), collagen and glycogen, respectively, were highlighted as the most significant. The peaks at 1125 cm^{-1} also stand out due to C–C, C–O, C–N stretching in lipids, glycogen and proteins in lung cancer detection. The peak at 1248 cm^{-1} (protein C–N stretching) is significant for understanding biochemical differences. The peak at 1587 cm^{-1}, corresponding to C=C stretching in tryptophan, highlights protein structural changes. In addition, the amide I bands from proteins, found at 1632–1668 cm^{-1}, indicate changes in protein secondary structure, such as alpha-helices and beta-sheets, which are crucial for distinguishing lung cancer.

For FTIR spectroscopy, the vibrational band between 1055–1070 cm^{-1} is particularly significant. This range includes phosphate stretching bands, such as symmetric vibrations of PO_2^{-1} in phospholipids, and C–O symmetric vibrations. Moreover, the strong absorption band at 1699 cm^{-1} corresponds to C=O stretching vibrations in amide I, indicating significant protein changes in cancer patients. These specific vibrations originate from the

fingerprint region, although the entire spectral range was used, which emphasizes the importance of this region for cancer diagnostics.

While various biochemical groups may be significant for distinguishing between healthy controls and lung cancer patients, protein-related vibrations are particularly important. Peaks at 1125 cm^{-1}, 1587 cm^{-1} and 1632–1668 cm^{-1} in Raman spectroscopy as well as the strong absorption at 1055–1070 cm^{-1} and 1699 cm^{-1} in FTIR spectroscopy underline the crucial importance of protein structure changes in cancer diagnostics. Although other biomolecules such as carbohydrates and nucleic acids are also important, proteins appear to be the most important biochemical markers for accurate cancer detection, highlighting their crucial role in the diagnostic process.

3. Materials and Methods

3.1. Sample Collection and Preparation

In this study, 36 human subjects were examined: 18 healthy controls and 18 patients diagnosed with non-small cell lung carcinoma (NSCLC). Blood samples from NSCLC patients were collected at the Oncology Department of Donostia University Hospital (San Sebastián, Spain) and plasma was separated within one hour according to standard protocols. Additionally, plasma samples from 18 healthy donors were obtained retrospectively from the Basque Biobank, San Sebastián, Spain.

Raman analysis was performed by applying only 1 µL of human blood plasma from each subject to a slide-mounted aluminum foil and air-dried for 5 min. Aluminum foil was chosen for its high reflectivity, stability, flexibility, low background signal and cost efficiency, making it an ideal substrate for Raman signal amplification [13]. In contrast, the FTIR analysis was performed with an attenuated total reflectance crystal (ATR) as a contact sample method. A 1 µL aliquot of human blood plasma, corresponding to the amount used in the Raman analysis, was placed on the ATR crystal. The samples were completely dried before data collection to ensure consistent and reliable measurements.

3.2. Data Collection and Preprocessing

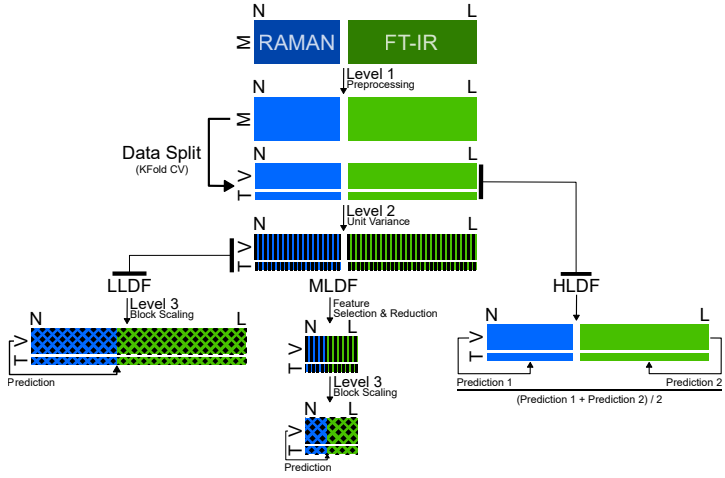
Raman measurements were performed with the Renishaw confocal Raman microscope inVia^{TM}, Wotton-under-Edge, England, UK, operated at a laser wavelength of 785 nm, laser power of 73 mW, a 50× long-distance objective for focusing the laser beam and collecting Raman signals and a spectrometer grating of 1200 L/mm. This setup was selected to optimize Raman signals and signal-to-noise (SNR) ratio while minimizing sample damage. Each sample underwent 20 accumulations, each lasting 1 s. Spectra were acquired from 25 different points at the droplet's periphery to capture the information from the high concentration of biomolecules that appear at the outer ring due to the coffee ring effect. Random cosmic ray interference was eliminated using the zap function in the Renishaw WiRE 5.4 software. Two preprocessing techniques were applied: asymmetric Whittaker baseline correction ($\lambda = 100$, $p = 0.01$) to remove baseline drifts and distortions, and standard normal variate (SNV) transformation to correct for variations due to sample thickness, scattering and instrumental response. Finally, 25 spectra per subject were averaged to obtain a single representative spectrum per subject. The averaging reduces random noise and improves SNR, providing a clear and comprehensive spectral signature for each subject [13].

FTIR measurements were carried out using the Bruker Vertex 70 spectrometer, Billerica, MA, USA, in ATR mode. The spectral resolution was set to 4 cm^{-1} and the sampling time for each measurement was 100 s. To avoid interference from water bands, the spectra were recorded after the sample was completely dried to ensure consistent and reliable data. For each sample, 10 spectra were collected to guarantee the stability and reproducibility of the results. Only the SNV transformation was applied to the data, as a baseline correction was not necessary and had no physical reason. Finally, 10 spectra were averaged for each subject to obtain a single representative spectrum.

Int. J. Mol. Sci. 2024, 25, 10936

3.3. Data Fusion and Model Building

The overall workflow shown in Scheme 1 describes the multi-level data-fusion strategy implemented in this study. It begins with row-based data preprocessing for Raman and FTIR spectra [20], as detailed in Section 3.2. The data are then partitioned into training and test datasets using stratified k-fold cross-validation with 6 folds, which was chosen as a trade-off between bias and variance due to the accuracy of the model (see Figure S1 in the Supplementary Information) and which mitigates overfitting by repeatedly partitioning the dataset into *k* subsets.



Scheme 1. Schematic representation of multi-level data fusion of Raman and FTIR data for lung cancer detection.

Later, unit variance scaling was applied to the preprocessed data after splitting. This column-based normalization technique adjusts the data so that each block has a unit variance and each variable within a block has the same variance equal to $1/n_{\text{block}}$, where n_{block} is the number of variables in the respective block. This method is particularly useful for datasets that contain variables with different units or scales, as it ensures that all variables contribute equally to the analysis [20–22].

In LLDF, the data matrices of the individual methods were directly concatenated after the preprocessing steps mentioned above in order to optimize the variations within the block and to cover the entire range of measured variables. Soft block scaling was applied by adjusting each data column with a scaling factor to ensure uniformity across a combined dataset. This factor was calculated for each column as the standard deviation of the column multiplied by the fourth root of the number of features. With this method, each scaled column has the same variance, and the sum of their variances is equal to the square root of the number of variables in the block. This approach balances the influence of individual blocks, particularly, different dynamics in the data, prevents larger blocks from dominating the analysis and ensures an equal contribution from all blocks [20,23,24].

In contrast, MLDF requires feature selection (FS) or feature reduction (FR) before block scaling. This approach retains the most important features of each method, reducing data complexity while preserving pertinent information. In MLDF, the identified key characteristics were fused and block-scaled as described above to ensure the consistency and balance of the contributions of all variables. The following steps outline the process: 1. FS was performed using regression coefficients (RCs) from partial least squares regression

Int. J. Mol. Sci. 2024, 25, 10936

(PLSR). Initially, features were ranked based on the absolute values of their RCs, arranged from highest to lowest. This systematic approach involved 100 iterations with 6 folds and makes sure that each feature was assessed 600 times. Features were selected in each iteration, based on specific percentiles, such as 1%, 2%, 3% and up to 10%. The selection frequency for each feature was calculated by counting how often it was selected across all iterations for each percentile. This process allowed us to identify the most significant features by their selection frequency. The features with the highest frequencies were then re-evaluated using the same percentile thresholds to ensure robustness. This method guarantees that only the most important features are retained. 2. FR was utilized with principal component analysis (PCA). The components were incrementally added, from 1 to 10, and model performance was calculated at each step. The highest accuracy was recorded with the specific combination of components that provided the best performance. This step ensures that the most informative components are used for model training, which further enhances the efficiency and effectiveness of the MLDF approach.

The most straightforward approach, HLDF, implies fusion of decisions and averages the predictions from individual data blocks, treating each block as equally important. Block scaling is therefore not an issue with HLDF, as no fused model per se, but only a fused decision is achieved; i.e., only the predictions of the individual models are combined [11]. Separate logistic regression models were trained for each method, and the predicted probabilities were then averaged. The final prediction was determined by applying a threshold to these averaged probabilities. Model performance was evaluated using accuracy scores, which were averaged across multiple folds for robust validation.

4. Conclusions

This study demonstrates the potential of combining Raman and FTIR spectroscopy with advanced data-fusion techniques to discriminate between healthy controls and lung cancer patients. By analyzing human blood plasma samples, we identified key protein-related vibrations that distinguish lung cancer patients, such as 1125 cm^{-1}, 1587 cm^{-1} and 1632–1668 cm^{-1} in Raman spectroscopy and 1055–1070 cm^{-1}, 1699 cm^{-1} in FTIR spectroscopy. These critical vibrational groups, identified through low-level data fusion supported by feature selection, significantly enhance model accuracy and highlight the importance of protein structural changes in cancer detection. Moreover, we were able to associate these spectral features with important biomarkers, providing insights into the underlying biochemical changes.

While other biomolecules such as carbohydrates and nucleic acids contribute to the classification, proteins were found to be the most decisive markers. The presented research underlines the potential of vibrational spectroscopy for the development of non-invasive diagnostic tools for the early detection of lung cancer. In particular, our research shows the power of a holistic approach—contrary to the analysis of single biomarkers—that takes into account the entirety of metabolic information that is accessible by Raman and FTIR spectrosopy.

This work provides a basis for further research into the molecular mechanisms of cancer and the development of new photonic diagnostic technologies based on large datasets that capture as much as possible the uncertainties associated with the method and the diversity of patients.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms252010936/s1.

Author Contributions: Conceptualization, H.H. and A.S.; methodology, H.H.; software, H.H.; validation, H.H. and A.S.; formal analysis, H.H.; investigation, H.H.; resources, B.S. and C.H.L.; data curation, H.H.; writing—original draft preparation, H.H.; writing—review and editing, H.H. and A.S.; visualization, H.H.; supervision, A.S.; project administration, A.S.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committees (CEIC Euskadi approval number: PI2019170).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chaitanya Thandra, K.; Barsouk, A.; Saginala, K.; Sukumar Aluru, J.; Barsouk, A. Epidemiology of lung cancer. *Contemp. Oncol./WspółCzesna Onkol.* **2021**, *25*, 45–52. [CrossRef]
2. Youlden, D.R.; Cramb, S.M.; Baade, P.D. The International Epidemiology of Lung Cancer: Geographical Distribution and Secular Trends. *J. Thorac. Oncol.* **2008**, *3*, 819–831. [CrossRef]
3. Thanoon, M.A.; Zulkifley, M.A.; Mohd Zainuri, M.A.A.; Abdani, S.R. A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *Diagnostics* **2023**, *13*, 2617. [CrossRef]
4. Prabhakar, B.; Shende, P.; Augustine, S. Current trends and emerging diagnostic techniques for lung cancer. *Biomed. Pharmacother.* **2018**, *106*, 1586–1599. [CrossRef]
5. Sultanbawa, Y.; Smyth, H.; Truong, K.; Chapman, J.; Cozzolino, D. Insights on the role of chemometrics and vibrational spectroscopy in fruit metabolite analysis. *Food Chem. Mol. Sci.* **2021**, *3*, 100033. [CrossRef]
6. Balan, V.; Mihai, C.T.; Cojocaru, F.D.; Uritu, C.M.; Dodi, G.; Botezat, D.; Gardikiotis, I. Vibrational Spectroscopy Fingerprinting in Medicine: From Molecular to Clinical Practice. *Materials* **2019**, *12*, 2884. [CrossRef]
7. Larkin, P. Chapter 1—Introduction: Infrared and Raman Spectroscopy. In *Infrared and Raman Spectroscopy*; Larkin, P., Ed.; Elsevier: Oxford, UK, 2011; pp. 1–5. [CrossRef]
8. Hayes, E.; Greene, D.; O'Donnell, C.; O'Shea, N.; Fenelon, M.A. Spectroscopic technologies and data fusion: Applications for the dairy industry. *Front. Nutr.* **2023**, *9*, 1074688. [CrossRef]
9. Cocchi, M. Chapter 1—Introduction: Ways and Means to Deal with Data from Multiple Sources. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Data Fusion Methodology and Applications; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 1–26. [CrossRef]
10. Arévalo, L.A.; O'Brien, S.A.; Lopez, E.; Singh, G.P.; Seifert, A. Design and Development of a Bimodal Optical Instrument for Simultaneous Vibrational Spectroscopy Measurements. *Int. J. Mol. Sci.* **2022**, *23*, 6834. [CrossRef]
11. Smolinska, A.; Engel, J.; Szymanska, E.; Buydens, L.; Blanchet, L. Chapter 3—General Framing of Low-, Mid-, and High-Level Data Fusion with Examples in the Life Sciences. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Data Fusion Methodology and Applications; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 51–79. [CrossRef]
12. Azcarate, S.M.; Ríos-Reina, R.; Amigo, J.M.; Goicoechea, H.C. Data handling in data fusion: Methodologies and applications. *TrAC Trends Anal. Chem.* **2021**, *143*, 116355. [CrossRef]
13. Hano, H.; Lawrie, C.H.; Suarez, B.; Lario, A.P.; Echeverría, I.E.; Mediavilla, J.G.; Cruz, M.I.C.; Lopez, E.; Seifert, A. Power of Light: Raman Spectroscopy and Machine Learning for the Detection of Lung Cancer. *ACS Omega* **2024**, *9*, 14084–14091. [CrossRef]
14. Sinica, A.; Brožáková, K.; Brůha, T.; Votruba, J. Raman spectroscopic discrimination of normal and cancerous lung tissues. *Spectrochim. Acta Part Mol. Biomol. Spectrosc.* **2019**, *219*, 257–266. [CrossRef] [PubMed]
15. Fadlelmoula, A.; Pinho, D.; Carvalho, V.H.; Catarino, S.O.; Minas, G. Fourier Transform Infrared (FTIR) Spectroscopy to Analyse Human Blood over the Last 20 Years: A Review towards Lab-on-a-Chip Devices. *Micromachines* **2022**, *13*, 187. [CrossRef]
16. Bujok, J.; Gąsior-Głogowska, M.; Marszałek, M.; Trochanowska-Pauk, N.; Zigo, F.; Pavľak, A.; Komorowska, M.; Walski, T. Applicability of FTIR-ATR Method to Measure Carbonyls in Blood Plasma after Physical and Mental Stress. *Biomed Res. Int.* **2019**, *2019*, 2181370. [CrossRef]
17. Vrtělka, O.; Králová, K.; Fousková, M.; Habartová, L.; Hříbek, P.; Urbánek, P.; Setnička, V. Vibrational and chiroptical analysis of blood plasma for hepatocellular carcinoma diagnostics. *Analyst* **2023**, *148*, 2793–2800. [CrossRef]
18. Gajjar, K.; Trevisan, J.; Owens, G.; Keating, P.J.; Wood, N.J.; Stringfellow, H.F.; Martin-Hirsch, P.L.; Martin, F.L. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: A novel diagnostic approach for ovarian cancer. *Analyst* **2013**, *138*, 3917–3926. [CrossRef] [PubMed]
19. 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. [CrossRef]
20. Campos, M.P.; Reis, M.S. Data preprocessing for multiblock modelling—A systematization with new methods. *Chemom. Intell. Lab. Syst.* **2020**, *199*, 103959. [CrossRef]

21. Silvestri, M.; Elia, A.; Bertelli, D.; Salvatore, E.; Durante, C.; Li Vigni, M.; Marchetti, A.; Cocchi, M. A mid level data-fusion strategy for the Varietal Classification of Lambrusco PDO wines. *Chemom. Intell. Lab. Syst.* **2014**, *137*, 181–189. [CrossRef]
22. Mishra, P.; Roger, J.M.; Jouan-Rimbaud-Bouveresse, D.; Biancolillo, A.; Marini, F.; Nordon, A.; Rutledge, D.N. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends Anal. Chem.* **2021**, *137*, 116206. [CrossRef]
23. Shaffer, R.E. Multi- and Megavariate Data Analysis. Principles and Applications, I. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Umetrics Academy, Umeå, 2001, ISBN 91-973730-1-X, 533pp. *J. Chemom.* **2002**, *16*, 261–262. [CrossRef]
24. Ríos-Reina, R.; Callejón, R.M.; Savorani, F.; Amigo, J.M.; Cocchi, M. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta* **2019**, *198*, 560–572. [CrossRef] [PubMed]

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
CIC nanoGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
BIOGIPUZKOA
EUSKAL OSASUN IKERKUNTZA
INVESTIGACIÓN VASCA EN SALUD
BASQUE HEALTH RESEARCH