INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Autarquia Associada à Universidade de São Paulo

Evaluation of breast cancer molecular subtypes using artificial intelligence in micro-FTIR hyperspectral images

MATHEUS DEL VALLE

Tese apresentada como parte dos requisitos para obtenção do Grau de Doutor em Ciências na Área de Tecnologia Nuclear – Materiais.

Orientadora:
Profa. Dra. Denise Maria Zezell

Coorientador:
Prof. Dr. Emerson Soares Bernardes

São Paulo
2023

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Autarquia Associada à Universidade de São Paulo

Evaluation of breast cancer molecular subtypes using artificial intelligence in micro-FTIR hyperspectral images

Versão Corrigida

Versão Original disponível no IPEN

MATHEUS DEL VALLE

Tese apresentada como parte dos requisitos para obtenção do Grau de Doutor em Ciências na Área de Tecnologia Nuclear – Materiais.

Orientadora:
Profa. Dra. Denise Maria Zezell

Coorientador:
Prof. Dr. Emerson Soares Bernardes

São Paulo
2023

Fonte de Financiamento: FAPESP (CEPID 05/51689-2, 17/50332-0), CAPES (Finance Code 001, PROCAD 88881.068505/2014-01) e CNPq (INCT-465763/2014-6, PQ-309902/2017-7, 142229/2019-9)

Como citar:

DEL-VALLE, M. ***Evaluation of breast cancer molecular subtypes using artificial intelligence in micro-FTIR hyperspectral images.*** 2023. 126 f. Tese (Doutorado em Tecnologia Nuclear), Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, São Paulo. Disponível em: ‹http://repositorio.ipen.br/› (data de consulta no formato: dd/mm/aaaa)

Ficha catalográfica elaborada pelo Sistema de geração automática da Biblioteca IPEN, com os dados fornecidos pelo(a) autor(a).

del Valle, Matheus

Evaluation of breast cancer molecular subtypes using artificial intelligence in micro-FTIR hyperspectral images / Matheus del Valle; orientadora Denise Maria Zezell; coorientador Emerson Soares Bernardes. – São Paulo, 2023.

126 f.

Tese (Doutorado) - Programa de Pós-Graduação em Tecnologia Nuclear (Materiais) – Instituto de Pesquisas Energéticas e Nucleares, São Paulo, 2023.

1. breast cancer subtype. 2. biomarker level. 3. micro-FTIR imaging. 4. machine learning. 5. convolutional neural network. I. Zezell, Denise Maria, orient. II. Soares Bernardes, Emerson, coorient. III. Título.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to:

- Professor Denise Maria Zezell, my advisor, for her trust, knowledge, academic dedication, and friendship. She is a brave and intelligent woman and a role model for any researcher.

- Professor Emerson Soares Bernardes, my co-advisor, for providing ideas and guidance on this important subject of breast cancer.

- My parents, Sueli and Alfredo, for their unwavering support throughout this journey. This title is as much theirs as it is mine.

- My family, friends, and especially my girlfriend, Marina, for their understanding in times of absence and for the motivational support to keep moving on.

- My colleagues at the Laboratory of Biophotonics, especially Moises, Pedro, Ariane and Amanda for their teachings, partnership, and friendship.

- My colleagues at the Lasers and Applications Center, for all the shared knowledge.

- USP and IPEN for the opportunities and infrastructure provided, and for their contribution to science and education.

- FAPESP (CEPID 05/51689-2, 17/50332-0), CAPES (Finance Code 001, PROCAD 88881.068505/2014-01) and CNPq (INCT-465763/2014-6, PQ-309902/2017-7, 142229/2019-9) for their financial support.

- Colleagues from my former job at the Health Innovation Techcenter of Hospital Israelita Albert Einstein, for sharing with me their GPU workstation (a.k.a. Vivi) and supporting my academic activities.

- Everyone who contributed to this research.

Thank you!

RESUMO

DEL-VALLE, Matheus. ***Avaliação de subtipos moleculares de câncer de mama utilizando inteligência artificial em imagens hiperespectrais por micro-FTIR***. 2023. 70 p. Tese (Doutorado em Tecnologia Nuclear) – Instituto de Pesquisas Energéticas e Nucleares – IPEN – CNEN/SP. São Paulo.

O câncer de mama é o mais incidente no mundo. A avaliação do subtipo molecular e seus biomarcadores tem um papel fundamental para o prognóstico. Os biomarcadores utilizados são os Receptores de Estrogênio (ER), de Progesterona (PR), de tipo 2 do fator de Crescimento Epidérmico Humano (HER2), e Ki67. Com base nestes, os subtipos são classificados como Luminal A (LA), Luminal B (LB), subtipo HER2 e Triplo-Negativo (TNBC). O padrão-ouro desta análise é a histologia e imuno-histoquímica, técnicas semiquantitativas que apresentam variações inter-laboratorial e inter-observador. A técnica de micro-espectroscopia no Infravermelho por Transformada de Fourier (FTIR), que fornece imagens hiperspectrais com informações bioquímicas de tecidos biológicos, é aplicada em conjunto de inteligência artificial (IA) para avaliação de cânceres. Nesta tese, foram utilizadas vinte amostras de duas linhagens celulares de câncer de mama, BT-474 e SK-BR-3, para definição do número ótimo de varreduras co-adicionadas para técnicas de aprendizado de máquina (ML). Foram utilizados os modelos de Análise Discriminante Linear (LDA), Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA), K-Vizinhos Mais Próximos (KNN), Máquinas de Vetores de Suporte (SVM), Floresta Aleatória (RF) e Aumento de Gradiente Extremo (XGB). Sessenta imagens hiperespectrais de 320x320 pixels foram coletadas de trinta pacientes de biópsias humanas de mama em um microarranjo, cada qual contendo um núcleo de Câncer de mama (CA) e um de Tecido Adjacente (AT). Foram desenvolvidos métodos automatizados para organização e pré-processamento dos dados em unidimensionais (1D) e bidimensionais (2D) baseados em agrupamento K-Médias. Os dados foram utilizados para treinamento de dois novos modelos de aprendizado profundo para avaliação de subtipo de câncer de mama: CaReNet-V1, Rede Neural Convolucional (CNN) 1D; e CaReNet-V2, CNN 2D. Todos os modelos de ML alcançaram desempenhos semelhantes com os grupos b256_064 (256 varreduras de fundo e 64 varreduras de amostra), b256_128 e b128_128, onde a melhor acurácia de 0.995 foi apresentada pelo modelo XGB. O b256_064 foi estabelecido como o ideal dentre os três devido ao menor tempo de aquisição. O método baseado em K-Médias possibilitou o pré-processamento e organização totalmente automatizado, melhorando a qualidade dos dados e otimizando o treinamento das CNN. A CaReNet-V1 classificou com eficácia CA e AT (acurácia de teste dos espectros individuais de 0,89), além dos subtipos HER2 e TNBC (0,83 e 0,86), apresentando maiores dificuldades para LA e LB (0,74 e 0,68). O modelo possibilitou a avaliação dos números de onda que mais contribuíram para as predições, fornecendo uma relação direta com o conteúdo bioquímico das amostras. A CaReNet-V2 demonstrou melhor desempenho que a 1D, com acurácias de teste acima de 0,84, e possibilitou a predição dos níveis de ER, PR e HER2, onde os valores limítrofes apresentaram menor desempenho (acurácia mínima de 0,54). A regressão da porcentagem de Ki67 demonstrou erro médio absoluto de 3,6%. Por outro lado, sua avaliação de impacto por número de onda foi inferior ao 1D. Assim, este estudo aponta as técnicas de IA por imagens por micro-FTIR como potenciais para prover informações adicionais aos relatórios patológicos, servindo ainda como técnicas de triagem de pacientes.

Palavras-chave: subtipo câncer mama; nível biomarcador; imagem micro-FTIR; varreduras co-adicionadas; aprendizado máquina; rede neural convolucional.

ABSTRACT

DEL-VALLE, Matheus. ***Evaluation of breast cancer molecular subtypes using artificial intelligence in micro-FTIR hyperspectral images***. 2023. 70 p. Tese (Doutorado em Tecnologia Nuclear) – Instituto de Pesquisas Energéticas e Nucleares – IPEN – CNEN/SP. São Paulo.

Breast cancer is the most incident cancer worldwide. The evaluation of molecular subtypes and their biomarkers plays an essential role in prognosis. The biomarkers used are Estrogen Receptor (ER), Progesterone Receptor (PR), Human Epidermal growth factor Receptor-type 2 (HER2), and Ki67. Based on these, subtypes are classified as Luminal A (LA), Luminal B (LB), HER2 subtype, and Triple-Negative Breast Cancer (TNBC). The gold standard for this analysis is histology and immunohistochemistry, semi-quantitative techniques that present inter-laboratory and inter-observer variations. The Fourier Transform Infrared micro-spectroscopy (micro-FTIR), which provides hyperspectral images with biochemical information of biological tissues, is applied together with artificial intelligence (AI) for cancer evaluation. In this thesis, twenty samples of two breast cancer cell lines, BT-474 and SK-BR-3, were used to define the optimal number of co-added scans for machine learning (ML) techniques. Linear Discriminant Analysis (LDA), Partial Least Squares Discriminant Analysis (PLS-DA), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB) models were used. Sixty hyperspectral images of 320x320 pixels were collected from thirty patients of a human breast biopsies microarray, each containing a breast cancer (CA) and an adjacent tissue (AT) core. Automated methods based on K-Means clustering were developed for data organization and pre-processing to one-dimensional (1D) and two-dimensional (2D) data. The dataset was used to train two new deep learning models for breast cancer subtype evaluation: CaReNet-V1, a 1D Convolutional Neural Network (CNN); and CaReNet-V2, a 2D CNN. All ML models achieved similar performances with the b256_064 (256 background scans and 64 sample scans), b256_128, and b128_128 groups, where the best accuracy of 0.995 was presented by the XGB model. The b256_064 was established as the ideal among the three due to the shortest acquisition time. The K-Means-based method enabled fully automated preprocessing and organization, improving data quality and optimizing CNN training. CaReNet-V1 effectively classified CA and AT (individual spectra test accuracy of 0.89), as well as HER2 and TNBC subtypes (0.83 and 0.86), with greater difficulty for LA and LB (0.74 and 0.68). The model enabled the evaluation of the most contributing wavenumbers to the predictions, providing a direct relationship with the biochemical content of the samples. CaReNet-V2 demonstrated better performance than 1D, with test accuracies above 0.84, and enabled the prediction of ER, PR, and HER2 levels, where borderline values showed lower performance (minimum accuracy of 0.54). The Ki67 percentage regression demonstrated an absolute mean error of 3.6%. On the other hand, its impact evaluation by wavenumber was inferior to 1D. Thus, this study indicates image-based AI techniques using micro-FTIR as potential providers of additional information to pathological reports, also serving as patient biopsy screening techniques.

Key words: breast cancer subtype; biomarker level; micro-FTIR imaging; co-added scans; machine learning; convolutional neural network.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1D	One-dimensional
2D	Two-dimensional
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AT	Adjacent Tissue
ATR	Attenuated Total Reflectance
CA	Cancer
$\textbf{CAF}_2$	Calcium Fluoride
CE	Cross-Entropy
CNN	Convolutional Neural Network
CV	Cross-Validation
D-FFNN	Deep Feedforward Neural Network
Dev	Development
DL	Deep Learning
DNA	Deoxyribonucleic Acid
EMSC	Extended Multiplicative Scatter Correction
ER	Estrogen Receptor
ECD	Extracellular Domain
FCN	Fully Convolutional Network
FFN	Feedfoward Neural Network
FFPE	Formalin Fixed and Paraffin Embedded

FN	False Negative
FP	False Positive
FPA	Focal Plane Array
FTIR	Fourier transform Infrared
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GMP	Global Max Pooling
Grad-CAM	Gradient-weighted Class Activation Mapping
GT	Ground Truth
GPU	Graphics Processing Unit
H2O	Water Vapor
HER2	Human Epidermal Growth Factor Receptor 2
IHC	Immunohistochemistry
IR	Infrared
KNN	K-Nearest Neighbors
LA	Luminal A
LB	Luminal B
LDA	Linear Discriminant Analysis
Low-e	Low-emissivity
MAE	Mean Absolute Error
ML	Machine Learning
MSE	mean squared error
MSC	Multiplicative Scatter Correction

NIPALS	nonlinear iterative partial least squares
NN	Neural Network
PC	Principal Component
PCA	Principal Component Analysis
PLS-DA	Partial Least Squares-Discriminant Analysis
PR	Progesterone Receptor
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RF	Random Forest
RGB	Red, Green and Blue
RMSE	Root-Mean-Square Error
RNA	Ribonucleic Acid
SD	Standard Deviation
SG	Savitzky-Golay
SGD	Stochastic Gradient Descent
SNR	Signal-to-Noise Ratio
SNV	Standard Normal Variate
SVM	Support Vector Machines
TN	True Negative
TNBC	Triple-Negative Breast Cancer
TNM	Tumor, Node, Metastasis
TP	True Positive

VGG	Visual Geometry Group
XGB	Extreme Gradient Boosting

CONTENTS

1 INTRODUCTION

Cancer is a group of diseases where abnormal cells grow uncontrollably, go beyond their usual boundaries to invade adjoining parts of the body and/or spread to other organs [3]. The mutations that lead to abnormal cells may be due to interaction with an external carcinogenic agent, as physical, chemical, and infectious, or may result from spontaneous mutations during cell division, with unknown cause [4].

In 112 out of 183 countries, cancer is the first or second leading cause of premature death. It is estimated 19.3 million new cases and 9.9 million deaths from cancer in 2020. Female breast cancer is the most incident cancer with 11.7%, or 2.3 million, of new cases in 2020, aside from 6.9%, or 690 thousand, of new deaths [5]. In Brazil, the estimate for 2020 is 66 thousand of new breast cancer cases, which correspond to 29.7% of the total new cancer cases in women [6].

The classification for the breast cancer can follow different parameters, as stage, grade, and molecular subtypes. The molecular subtypes classification plays an import role in the breast cancer treatment, sorting patients with divergent prognosis and helping to select an appropriate and specific therapy [7]. Subtypes are defined using the expression levels of Ki67 biomarker and three hormone receptors: estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). The four subtypes and their usual treatments are [8, 9]:

- Luminal A (ER and/or PR positive, HER negative, Ki67 low) – endocrine therapy.
- Luminal B (ER and/or PR positive, HER variable, Ki67 high) – endocrine therapy and chemotherapy; if HER2 positive, anti-HER2 therapy may also be used.
- HER2 subtype (ER and PR negative, HER positive, Ki67 usually high) – chemotherapy and anti-HER2 therapy
- Triple-negative (ER, PR and HER negative, Ki67 usually high) – chemotherapy.

Classifications are widely performed by histology and immunohistochemistry semiquantitative techniques, however, several issues affect its assessment quality, such as interlaboratory (antibodies, detection systems, and protocols used) and interobserver variations [10]. Fourier Transform Infrared (FTIR) spectroscopy has been studied as a further cancer evaluation technique in the past years, not only to overcome the variations, but also to provide additional information, helping to improve assessment quality [11, 12], once the FTIR spectrum contain lots of biochemical information, such as lipids, proteins, nucleic acids and carbohydrates contents [13].

With the consolidation of FTIR spectroscopy imaging, which provides thousands of spectra in a single acquisition, machine learning approaches stood out as powerful tools for many diagnostics, including cancer classification [14, 15].

Several protocols for using FTIR to analyze biological samples have been published [13, 16, 17], standardizing acquisition and processing parameters. Despite that, the number of co-added scans is minimally commented, without considering its effects in machine learning classifications. Parameters optimization [18] were studied, but no biological samples or machine learning techniques were applied. An empirical study was conducted with brain tissue samples [19], limited to using only one clustering technique (K-means) and without varying the background scans number.

Artificial intelligence techniques have grown exponentially in the past decade, where deep learning approaches, a subarea of machine learning, were in the spotlight [20]. Researches have applied deep learning in several spectral domains [21], including biospectroscopy/biospectral imaging [22] and vibrational spectroscopy [23]. To the date, there is no study using FTIR and deep learning for breast cancer classifications assessing, being limited to malignant vs benign diagnosis using ATR-FTIR single spectra acquisition [24] or blood serum [25], and morphological comparison with chemical histology techniques [26].

In this way, there is still a lack of consensus for a systematic approach to define an optimal number of co-added scans regarding a machine learning classification task; and of studies using recent deep learning techniques to evaluate breast cancer, their molecular subtypes and biomarkers expression levels using micro-FTIR hyperspectral images. A complete automated analysis tool could provide extra information for the pathology report and act as a biopsy screening technique, speeding up the patient assessment and prioritization process.

2 OBJECTIVE

The main objective of this research is to develop an artificial intelligence approach using micro-FTIR hyperspectral images to be a potential technique for breast cancer molecular subtype evaluation. Specific objectives are:

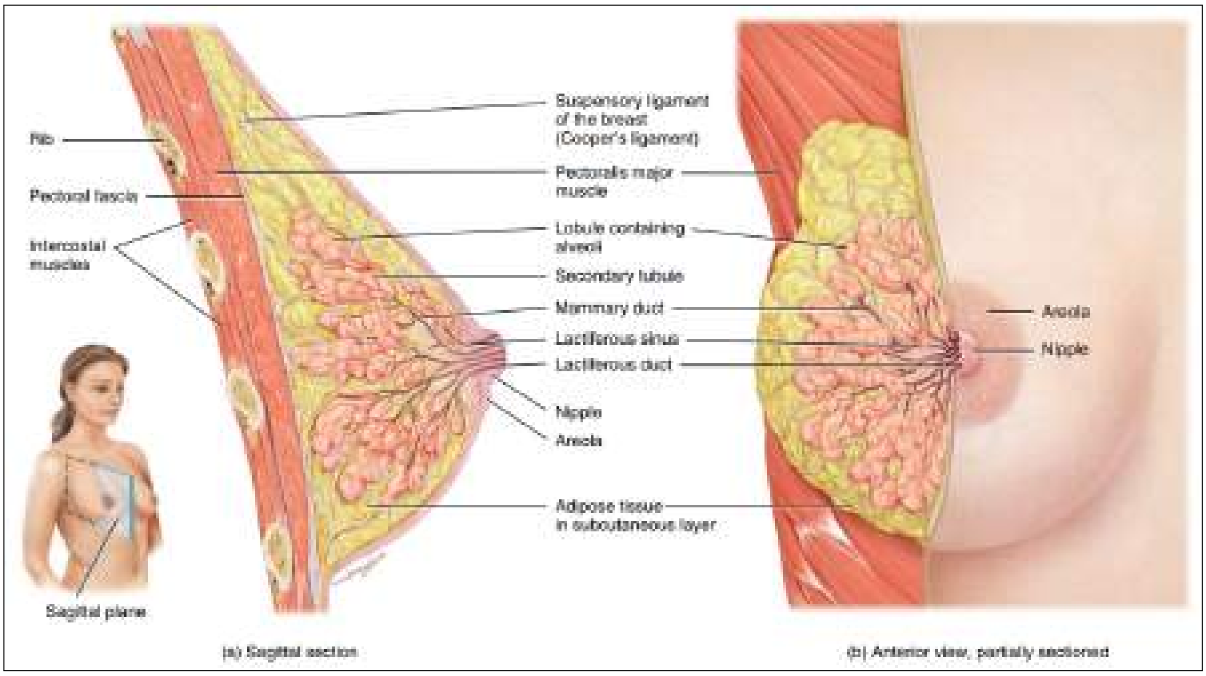
- Define an optimal co-added scans number for machine learning tasks.
- Fully automate the data organization and preprocessing
- Develop and evaluate 1D and 2D deep learning models.
- Differentiate breast cancer from adjacent tissue (AT).
- Classify subtypes (Luminal A, Luminal B, HER2 and TNBC)
- Predict biomarkers (ER, PR, HER2, Ki67) levels.

3 BACKGROUND

3.1 Breast Morphology and Physiology

The breasts are projections attached to the pectoralis major and serratus anterior muscles by the deep fascia, a layer of dense and irregular connective tissue [27]. The skin is the outermost layer of the breast, linked with the superficial fascia. The deep and the superficial fascias envelops the breast parenchyma, a structure composed of glandular epithelium, Cooper's ligament, and adipose tissue (Figure 1) [28].

Figure 1 – Breast structure representation.



Source: Tortora and Derrickson, 2014 [27].

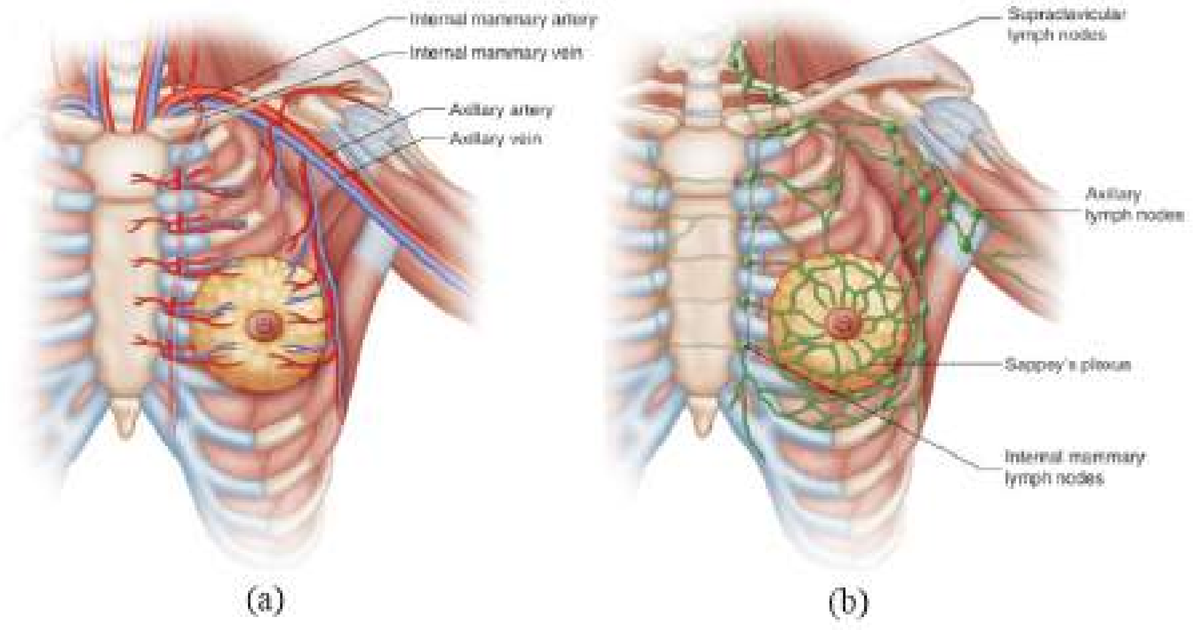
Glandular epithelium forms the mammary gland, a modified sudoriferous gland to produce milk. Each mammary gland consists of 15 to 20 fat separated lobes, which are composed of several lobules. Although present in both male and females, mammary glands are normally functional only in female. The lobules are formed by milk-secreting glands called alveoli, consisted of cuboidal epithelium. After the production in the alveoli, milk passes through secondary tubules and then into mammary ducts. The ducts are surrounded by smooth muscle-like cells, termed myoepithelial cells, which help to propel milk. As it get closer to the nipple, the mammary ducts expand and form the lactiferous sinuses, a structure lined with stratified squamous epithelium where milk can be stored before exiting by a lactiferous duct [27–30].

Cooper's ligaments are fibrous bands of connective tissue (stroma) between the skin and fascia, providing support to the breast. These ligaments become looser with aging or excessive strain. Adipose tissue constitutes the remainder part of the breast, where its portion increases with aging and after menopause [27, 28].

Nipple and areola epidermis are covered by keratinized, stratified squamous epithelium, containing papillae that allow blood perfusion to the surface, which give the pigmented aspect to the nipple and areola. During puberty and pregnancy, the pigmented aspect increases and the areola enlarges. The areola has sebaceous and apocrine sweat glands, in addition to accessory glands, called Montgomery glands, that can secrete milk [31, 32].

Arterial blood comes to the breast through three sources: anterior perforators of the internal mammary artery, responsible for approximately 60% of the breast supply, axillary artery branches, responsible for approximately 30% of the supply, and lateral branches of intercostal arteries, responsible for the 10% remaining supply. The blood supply to the breast skin depends on the subdermal plexus. Venous drainage of the breast mimics the arterial supply. The lymphatic drainage is performed by the Sappey's plexus. Approximately 97% is drained to the axilla, while the remaining 3% is drained to the internal mammary lymph nodes (Figure 2) [28, 32].

Figure 2 – Representation of (a) the arterial supply and venous drainage, and (b) the lymphatic drainage of the breast.



Source: McGuire, 2019 [28].

The sensory nerve supply to the breast is mainly performed by lateral cutaneous branches of the third to sixth intercostal nerves, while the nipple and areola nerve supply comes from the fourth intercostal nerve [28]. The nipple and areola sensory innervation also plays an important rule for breast feeding. When the infant sucking occurs, it stimulates the hypothalamus, sending nerve impulses to the posterior pituitary gland, which secretes oxytocin, a hormone to induce the milk release [31, 33].

Breast development is mainly stimulated by estrogen and progesterone, steroid hormones derived from cholesterol and produced in the ovary. Estrogen is responsible for the stromal tissue development, ductile system growth, and deposition of fat. Progesterone is required for lobules and alveoli development, inducing alveolar cells to proliferate, enlarge, and become secretory. Although the progesterone activity, alveoli are only able to secrete milk after further stimulus from prolactin, a hormone produced in the pituitary gland [34, 35].

3.2 Breast Cancer

The classification for the breast cancer can follow different parameters, as the histopathology diagnosis, stage, grade, and molecular subtypes. Histopathologically, the breast cancer involves two major groups, based on the ductal-lobular system: *in situ* and invasive. In situ carcinoma is divided in ductal carcinoma in situ, where the proliferation of cells is restricted to the ductal-lobular system, without the basement membrane invasion, and lobular carcinoma in situ, where more than half of the alveoli in a lobular unit are distended and distorted. Invasive carcinoma can also be divided in two: no special type and special subtype. The invasive ductal carcinoma no special type includes tumors that cannot be categorized as one of the special rare types, which have specific definitions, as lobular, tubular, and papillary invasive carcinomas [36–40].

The staging system is defined by the tumor, node, metastasis (TNM) system. “Tumor” is related to the primary tumor type and size, “node” is associated to the lymph node status, basically the histological evaluation of size and extension pattern of the excised lymph node, and “metastasis” is essentially if the tumor has spread to different sites.

Grade is a semi-quantitative method obtained by histological evaluation where each feature receives a score. The features and their scores are: tubule and gland formation (more than 75% = 1, 10 to 75% = 2, less than 10% = 3), nuclear pleomorphism (mild = 1, moderate = 2, significant = 3), and mitotic count (depending on the microscope field area, from 1 to 3). The final grade is determined as “Grade I” for 3-5 total score, “Grade II” for 6-7 total score, and “Grade III” for 8-9 total score [36, 39].

Subtypes are defined using the expression levels of Ki67 biomarker and three hormone receptors: estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). The four subtypes are Luminal A (LA), Luminal B (LB), human epidermal growth factor receptor 2 subtype (HER2), and triple-negative breast cancer (TNBC).

ER is a intracellular receptor, present in two forms: ERα and ERβ. ERα is expressed in 15 to 30% of luminal cells, and its binding to estrogen can stimulate proliferation of mammary cells. ERβ is expressed in both myoepithelial and stromal cells, and plays an important role in alveolar differentiation. PR is also present in two forms: PRA and PRB. PRA is expressed in the luminal epithelium, and can be related to progesterone-induced ductal lateral branching. PRB is expressed in both luminal and myoepithelial cells, being related to alveologenesis [35, 41]

HER2 receptor is a tyrosine kinase related receptor, part of Her/ErbB2/Neu transmembrane receptors. Structurally, these receptors have three distinct regions: an extracellular domain (ECD), a membrane spanning region, and a cytoplasmic tyrosine kinase domain. Metalloproteinases can cleave the HER2 protein, resulting in the production of a truncated membrane-bound fragment (p95) and release of the ECD into the serum. Increased levels of ECD may predict response to hormones and chemotherapy. Overexpression of HER2 in tumor cells has been linked to increased angiogenesis, which is essential for tumor survival and metastases. This effect is due to the ability of HER2 to modulate the balance between proangiogenic and antiangiogenic factors [42].

Ki67 is a non-histone nuclear cortex protein present in the polymerase I-dependent ribosomal RNA synthesis, named by the Kiel Univeristy after the 67th antibody clone able to detect it. Different from other cell proliferation methods, such as the analysis of thymidine uptake and percent of cells in S-phase via flow cytometry, Ki67 is a nuclear marker expressed in all phases of the cell cycle. Expression of Ki67 reaches the peak in the mitosis phase [43, 44].

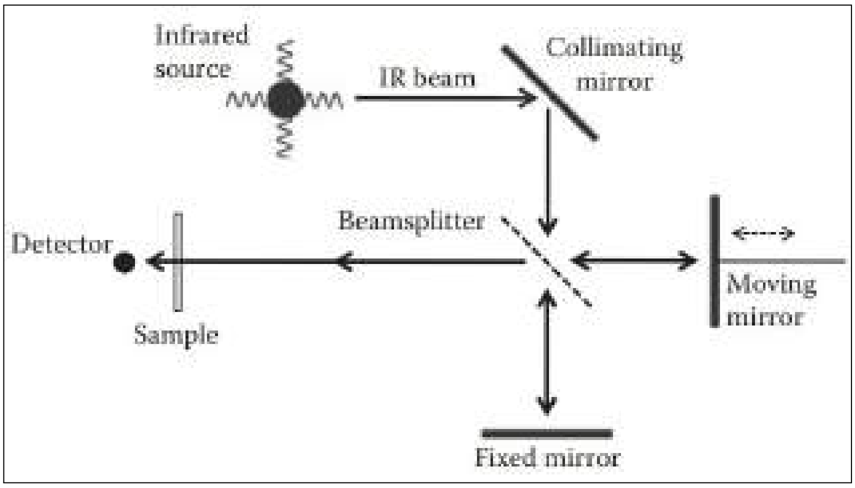
LA (ER/PR$+$, HER$-$, Ki67 low) is reported as the least aggressive subtype. Usually treated with endocrine therapy, as it is frequent chemoresistant and bear a higher risk of late recurrence. LB (ER/PR$+$, HER2$-/+$, Ki67 high) is more aggressive than LA, however it has a wider range of treatment possibilities, once it is affected by chemotherapy, and in case of HER+, anti-HER2 therapy may also be used. HER2 subtype (ER$-$, PR$-$, HER2$+$, Ki67 usually high) and TNBC (ER$-$, PR$-$, HER2$-$, Ki67 usually high) are the most aggressive ones, as they present a high proliferation rate and frequently leads to early recurrence of the disease, mostly before five years. HER2 subtype exhibits rapid tumor growth, increased risk of postoperative recurrence, and poor response to conventional chemotherapy. TNBC usually shows the worst prognosis as it lacks of drug targets, presenting a high chemoresitance [8, 9, 45]. TNBC and HER2 subtypes also leads to brain metastasis in 25-46% and 11-48% of the cases, respectively, against 8-15% in Luminal subtypes [46].

The biomarkers expression levels that classify the molecular subtypes are measured by immunohistochemistry (IHC) assays. IHC is a semiquantitative technique that is influenced by interlaboratory (antibodies, detection systems, and protocols used) and interobserver variations [10]. Fourier Transform Infrared (FTIR) spectroscopy has been studied as a further cancer evaluation technique in the past years, not only to overcome the variations, but also to provide additional information to the pathology analysis [11, 12]

3.3 Fourier Transform Infrared spectroscopy

Fourier Transform Infrared (FTIR) spectroscopy is a technique to characterize a sample in terms of biochemical content by measuring the vibrations of molecular bonds with an electric dipole moment. To perform that, firstly, a mid-infrared light source, usually globar- or synchrotron-based, passes through an interferometer (Figure 3). The interferometer splits the light beam from the source in two, causing them to travel different paths. Then, light beams are recombined into one beam due to the interference characteristic of the light wave and leaves the interferometer. A moving mirror causes an optical path difference between the two beams, providing an interferogram in relation to the mirror displacement.

Figure 3 – Optical diagram based on the Michelson interferometer.



Source: Smith, 2011 [47]

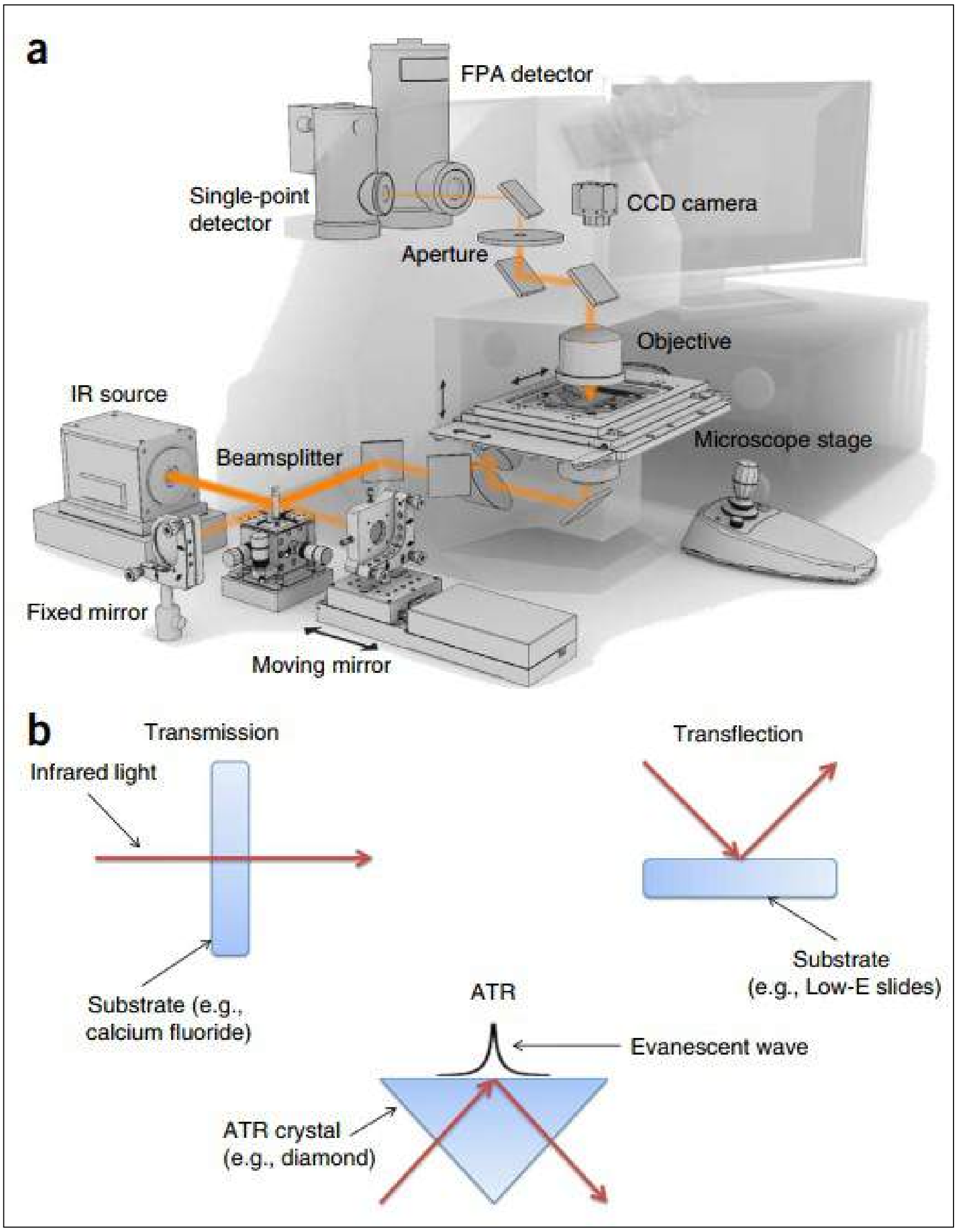
Applying the Fourier Transform to the interferogram, it is possible to calculate the light wavenumber by:

$$\tilde{\nu} = \frac{F}{2v} \tag{1}$$

where $\tilde{\nu}$ is the wavenumber in cm^{-1}, F is the frequency of the interferogram in Hertz, and v is the moving mirror velocity (assumed constant) in cm/s. This allows the entire spectrum to be obtained simultaneously, in contrast to traditional IR spectroscopy [47].

The light beam can interact directly to a sample, in liquid, solid or gaseous form; or a microscopy can be coupled to the spectroscopy equipment, enabling a micro-FTIR imaging (Figure 4 (a)). This can provide a single spectrum as a mean of the field of view of the single-point detector, or a hyperspectral image using a Focal Plane Array (FPA) or linear array of detectors. The light beam can interact with the sample in the microscopy slide by three main acquisition modes (Figure 4 (b)): transmission, transflection, and Attenuated Total Reflectance (ATR) [13].

Figure 4 – Representative image of the FTIR equipment acquisition. (a) Schematic instrumentation of a micro-FTIR. (b) Schematic representation of the three main acquisition modes.



Source: Baker et al, 2014 [13].

In transmission mode, the light beam passes through the sample fixed in a substrate with near none IR absorption, as in a calcium fluoride slide. In transflection, the light beam passes through the sample and is reflected by a IR-reflecting surface, such as a low-emissivity (low-e) slide, passing through the sample again. Calcium fluoride slides provides the best quality of spectra, but are expensive and mechanically unstable, while low-e slides are cheaper, however they add the electric field standing wave effect, where band intensity ratios and positions are severely affected depending on the thickness of the sample [48]. In ATR, the beam of light travels through a crystal of high refractive index, such as zinc selenide, germanium and diamond, and encounters the interface of the sample with lower refractive index. The total internal relfection, due to the refractive indexes difference, produces an evanescent wave, which interacts with the sample [13, 47, 49].

Each method presents convenience for some samples and challenges for others. The optimal usage of the FTIR is when detector noise exceeds all other noise sources. The signal-to-noise ratio (SNR) can be improved by optimizing the acquisition parameters, such as the mirror velocity or the co-added scans, where a single point is acquired several times and the equipment calculates the mean of them. Considering a simplified scenario, where the specific detectivity, which is given by the square of the detector area divided by the noise equivalent power of the detector, does not change with modulation frequency, doubling the mirror velocity will halve the acquisition time, but will also decrease the SNR in a factor of $\sqrt{2}$. On the other hand, doubling the co-added scans will recover the SNR of the original measurement [49].

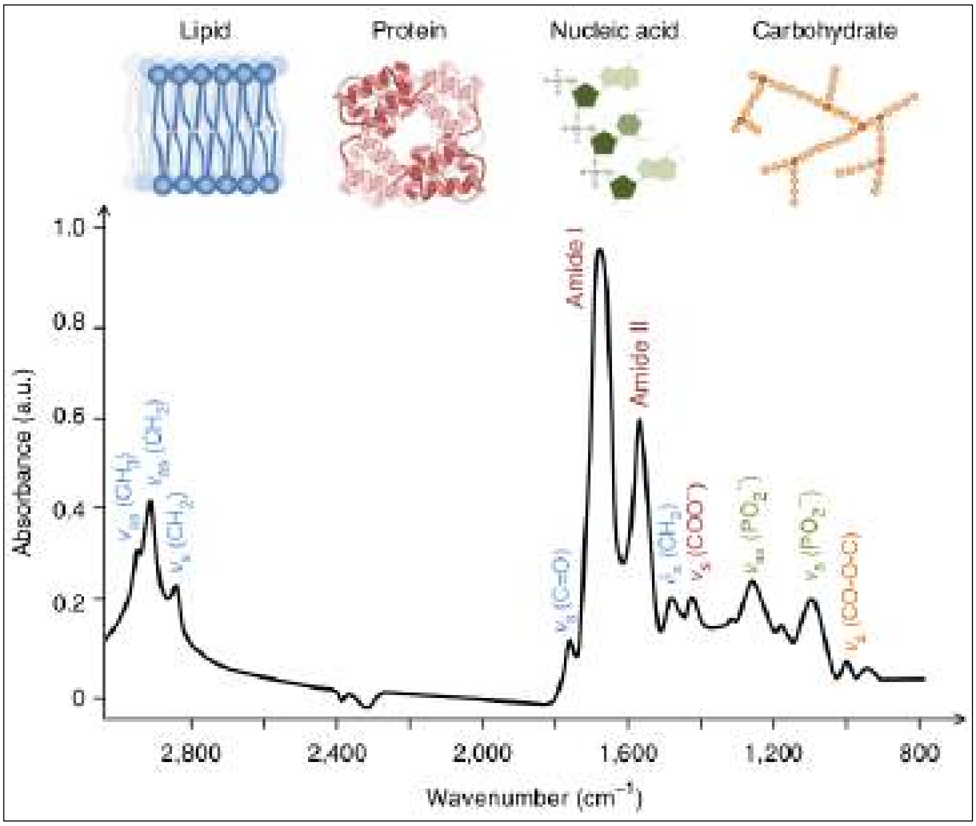
The intensity of the single-beam spectrum at any wavenumber, calculated from the interferometer measures, is proportional to the radiation reaching the detector. Using the Beer-Lambert Law, which is the fundamental law of quantitative spectroscopy, it is possible to measure the absorbance of the sample by each wavenumber:

$$A_i(\tilde{\nu}) = a_i(\tilde{\nu}) b c_i \tag{2}$$

where a_i is the absorptivity of each sample component i in the wavelength $\tilde{\nu}$, b is the sample thickness, and c_i is the concentration [49].

In this way, the measured absorbance is related to the vibrational modes of molecular bonds in each wavenumber. These vibrational modes provide an unique and label-free tool for characterizing the molecular content of a sample. For biological samples, the most import spectral regions is typically in the fingerprint region (1800-900 cm^{-1}), where lipids, protein, nucleic acids and carbohydrates content can be evaluated (Figure 5) [13].

Figure 5 – Typical biological spectrum showing biomolecular peak assignments.



Source: Baker et al, 2014 [13].

3.4 FTIR Data Preprocessing

To correctly evaluate the FTIR spectra, a series of preprocessing steps must be applied. This improves data quality, decreasing undesired signal contributions, i.e., not related to the target sample. Even though there is no unique preprocessing pipeline for all kinds of spectra and analysis, the steps must follow a logical order with adequate parameters, otherwise preprocessing may mask the signal of interest or add bias to the data [13, 16, 17]. General preprocessing steps are described next.

3.4.1 Quality test

Raw spectra data is evaluated by quality tests to identify anomalous or biased patterns. This can be performed by a SNR evaluation, where Amide I and II region (1700 to 1500 cm^{-1}) is used to check biological tissue signal, while 2000 to 1800 cm^{-1} is known as the dead region, without biological signal. A threshold may be applied to the ratio of the areas from these bands, biosignal/dead region, splitting target signal from noise [17]. The threshold may be manually defined by inspection or automatically calculated by thresholding algorithms, such as the Otsu method [50].

A more robust technique as Hotelling's T^2 vs Q residuals chart may be applied instead. The Q residuals, also known as the reconstruction loss, for an input sample x_i^T is given by [51]:

$$Q_i = x_i^T(I - P_A P_A^T)x_i \tag{3}$$

where P_A is the loadings matrix of the Principal Component Analysis (PCA) model with A components. Samples that have low Q-residuals are accurately depicted by the model, which means they have minimal orthogonal distances from their low-dimensional projections.

The Hotelling T^2 measures how much a sample deviates from the centroid of the data in the principal component space, indicating its level of outlier status. The T^2-contributions associated with it represent the degree of influence that each input variable has on the outlingness. This is calculated for each i-th sample as:

$$t_{\text{cont},i} = t_i \Lambda^{-1/2} P_A \tag{4}$$

where Λ is the diagonal matrix holding the A leading eigenvalues of $X^T X^2$.

In this way, it is possible to asses the abnormality of the data using the Hotelling T^2 vs Q residuals plot. Values far from the origin indicate outliers, where a confidence limit must be defined to evaluate the data, and may be set by the user according to the data, e.g. a 95% confidence interval [52].

3.4.2 Truncation

The fingerprint region, from 1800 to 900 cm^{-1}, is a crucial region for analyzing biomolecules. Within this range, various functional groups display distinct absorption peaks, including lipids (C=O symmetric stretching at approximately 1750 cm^{-1} and CH$_2$ bending at approximately 1470 cm^{-1}), proteins (amide I at approximately 1650 cm^{-1}, amide II at approximately 1550 cm^{-1}, and amide III at approximately 1260 cm^{-1}), carbohydrates (CO$-$O$-$C symmetric stretching at approximately 1155 cm^{-1}), nucleic acids (asymmetric phosphate stretching at approximately 1225 cm^{-1} and symmetric phosphate stretching at approximately 1080 cm^{-1}), glycogen (C$-$O stretching at approximately 1030 cm^{-1}), and protein phosphorylation (approximately 970 cm^{-1}).

Additionally, the high region, from 3700 to 2800 cm^{-1}, can also provide valuable information for biological analysis. This region includes absorption peaks for water ($-$OH stretching at approximately 3275 cm^{-1}), proteins (symmetric $-$NH stretching at approximately 3132 cm^{-1}), fatty acids and lipids (=C$-$H asymmetric stretching at approximately 3005 cm^{-1}, CH_3 asymmetric stretching at approximately 2970 cm^{-1}, CH_2 asymmetric stretching at approximately 2942 cm^{-1}, and CH_2 symmetric stretching at approximately 2855 cm^{-1}). These peaks can be used to obtain complementary information on the molecular composition of biological samples and provide a more comprehensive understanding of their chemical structure [17].

3.4.3 Smoothing

Smoothing removes random noise while preserving useful spectral information. This is performed by applying spectral filters, where the Savitzky-Golay (SG) algorithm is the most used one. This technique is a moving average-like filter, where the coefficients are derived by an unweighted linear least-square fit using a polynomial order [53].The mathematical description of the SG process is given by [54].

$$s_i^* = \frac{\sum_{j=-m}^{j=m} C_j s_{i+j}}{N} \tag{5}$$

where s_i is a point of the spectrum $S = (s_1, s_2, ..., s_N)$, treated with a set of m convolution coefficients, C_j.

The major drawback of SG smoothing is that the polynomial order and window size used in the polynomial fitting can have a significant impact on the resulting data. It is important to choose a polynomial order that matches the spectral shape features, such as a second-order polynomial for vibrational spectroscopy data, and the window size should be an odd number that's neither too small, as this can leave noise in the data, nor too large, as this can change the spectral shape [17]. Usually, window sizes from 3 to 21 may be evaluated [16].

3.4.4 Light scattering correction

Light scattering can occur when a material being analyzed contains particles of different sizes, particularly those smaller than the employed wavelength. These particles cause a systematic shift in the absorbance or spectral intensity. Light scattering correction can by accomplished by using Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC) or second derivative [15, 17].

While the second derivative can be coupled to the SG smoothing, the SNV can be calculated by:

$$s_i^* = \frac{s_i - \overline{S}}{\mathrm{SD}_S} \tag{6}$$

where s_i is a point of the spectrum $S = (s_1, s_2, ..., s_N)$, $\overline{S}$ and SD_S are the mean and standard deviation, respectively, of S.

The MSC represents each spectrum in terms of the average spectrum:

$$A(\tilde{\nu}) = a + \overline{x}(\tilde{\nu}) \cdot b + e(\tilde{\nu}) \tag{7}$$

where $A(\tilde{\nu})$ is the absorbance in a particular wavenumber, a_i and b_i are parameters calculated by a least squares regression, related to the baseline and the thickness, respectively, the reference spectrum $\overline{x}(\tilde{\nu})$ can be the mean of all collected spectra or of a established material, such as Matrigel (Corning Inc., USA), and $e(\tilde{\nu})$ is the residual term, which can be calculated by a PCA model. Thus, the corrected spectrum is given by [55]:

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - a}{b} \tag{8}$$

3.4.5 Baseline correction

Baseline correction removes absorption interference, such as fluorescence interference. There are several techniques, such as rubber-band, polynomial and automatic weighted least squares corrections. However, if spectral differentiation or the Extended MSC (EMSC) is applied for the light scattering correction, the baseline will already be corrected [17]. EMSC is one of the most used technique for baseline correction. It adds a polynomial to the previous MSC algorithm, since the baseline usually cannot be represented by a straight line:

$$A(\tilde{\nu}) = a + \overline{x}(\tilde{\nu}) \cdot b + d_1\tilde{\nu} + d_2\tilde{\nu}^2 + ... + d_n\tilde{\nu}^n + e(\tilde{\nu}) \tag{9}$$

where d_1 to d_n are the coefficients of the polynomial of order n. Therefore, the spectra are corrected by the EMSC according to:

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - a + d_1\tilde{\nu} + d_2\tilde{\nu}^2 + ... + d_n\tilde{\nu}^n}{b} \tag{10}$$

where the polynomial coefficients can be estimated using an ordinary or weighted least squares regression.

3.4.6 Spectral differentiation

Applying first and second derivatives can effectively correct for baseline distortions and light scattering, and can be coupled with SG smoothing. These techniques can also enhance the detection of smaller spectral differences between samples, making them particularly useful for identifying distinctive spectral features in complex samples with overlapping bands. Although, it is important to carefully choose the order of the derivative function to avoid introducing excessive noise.

Derivatives transform the spectral scale to mathematical coefficients, rather than absorbance, which means that spectral intensity cannot be directly correlated with chemical concentrations. Furthermore, identifying spectral biomarkers requires careful consideration, as the derivative function shifts the spectral band positions in ixd wavenumbers, where i is the derivative order and d is the data spacing resolution.

3.4.7 Substrate and environment contributions removal

Substrate contributions can mask the target signal, preventing the properly analysis. These contribution may be originated from components such as glass slide, paraffin and water vapor. Glass contributions are decreased in the high wavenumber region, enabling a suitable analysis at 3800 to 2500 cm^{-1}. Hence, a simple truncation can solve the problem [17].

The analysis of samples embedded in paraffin presents many advantages, such as the refractive index match of the biological tissue, which decreases significantly the

Mie scattering; it is the gold standard in tissue preservation and storage; chemical de-waxing usually warms and chemically degrade the samples, mainly protein content, while most commonly not removing the paraffin completely. However, paraffin has two strong signal bands in 1500 to 1300 cm^{-1} that can mask the target signal. Thereby, the spectra of a paraffin embedded sample can be truncated, excluding the paraffin region and also the relevant sample signal, or a digital de-waxing can be coupled to the EMSC [56].

The proceeding of the digital de-waxing is to take a hyperspectral image of pure paraffin and build a PCA model. This model and the average paraffin spectrum are added to the EMSC model, which is solved in the same way, using least squares regression. It is important to properly preprocess the paraffin spectra, e.g. applying quality test and smoothing. Analogously, the water vapor contribution may also be added to the EMSC model, where a hyperspectral image of the pure slide containing water vapor variation has to be acquired, e.g. turning off the purge system after the acquisition starts to potentialize the variation [56].

3.4.8 Normalization

Normalization is used to correct different sample thickness and concentration. Amide I, vector and min-max normalization are the most employed ones. The first can be used when the amide I band is not a distinguishing feature, and is simply given by dividing the spectrum intensities by the Amide I peak intensity [15, 17]:

$$s_i^* = \frac{s_i}{P} \tag{11}$$

where s_i is a point of the spectrum $S = (s_1, s_2, ..., s_N)$, and P is the Amide I peak. Note that the Amide I peak can be replaced by any other desired peaky, if suitable. Yet, wavenumber shift may occur and change the band position across the spectra, negatively affecting this kind of normalization.

Vector normalization is defined as:

$$s_i^* = \frac{s_i}{\sqrt{s_1^2 + s_1^2 + ... + s_N^2}} \tag{12}$$

where the denominator is called the “norm” of the spectrum.

Min-max normalization is usually applied to a 0 to 1 range, where this can be calculated by:

$$s_i^* = \frac{s_i - s_{min}}{s_{max} - s_{min}} \tag{13}$$

where s_{min} and s_{max} are the minimum and maximum intensities of the spectrum, respectively [15].

3.4.9 Outlier detection

Outlier are samples where the spectral signal differs significantly from the spectral signal of most of the acquired samples, even after all the previous preprocessing steps. This may be due chemical structure or concentration differences, or by measurement error. There are several types of outlier detection algorithm which may be applied to spectral data, such as the Jack-knife, Z-score and K-mode clustering, although the Hotelling's T^2 vs Q residuals is one of the most popular and visually intuitive [17]. Its usage is the same from the quality test step.

3.4.10 Dimensionality reduction

Feature selection and extraction techniques may be employed before data modeling to obtain dimensionality reduction, once spectral data contain a large number of features. The key difference between these two techniques is that the feature selection looks for the subset of features which efficiently define the data. It selects important and relevant features without transforming them. On the other hand, feature extraction creates new features that depend on the original ones. It calculates more significant features by transforming them using algebraic algorithms and optimization criteria [57].

Several feature selection methods have been applied to different fields of study using FTIR, such as Random Forest (RF), penalized regression through the least Absolute Shrinkage and Selection Operator or Ridge Regression [58, 59], Genetic Algorithm [60], and Recursive Feature Elimination [61].

The most frequently employed feature extraction algorithm for spectral data is the Principal Component Analysis (PCA), for both dimensionality reduction before modeling and exploratory data analysis. The main goal of PCA is to transform the data into a more relevant Principal Component (PC) coordinate space. These obtained PCs can be defined as variance-scaled vectors in the variable space, and are obtained calculating the eigenvectors and eigenvalues of the covariance matrix obtained from the X data. First PC relates to the greatest variance in the data. Scree plots exhibits the accumulated explained variance percentage by PCs, which is used to determine the optimal number of PCs. Each PC is also called a loading, and the coefficients in the linear combination representing the PC indicate the contributions of each wavenumber in the original variable space. The values of the new coordinate system are called PC scores, e.g. a 467 points biofingerprint truncated spectra after modeled by a PCA model with 10 PCs, will present 10 score points. Mathematically, PCA decompose a data matrix X of m x n size by [15]:

$$X = Y(U_K)^T + E \tag{14}$$

where Y is the score matrix, U_K is the loadings matrix, and E is the residual term.

3.5 Machine Learning Modeling

Classical inference statistics always had a strong preference for low-dimensional parametric models [62], not being able to handle the increasing volume of generated data and its high dimensionality [63]. Thus, Machine Learning (ML) models are required to perform such tasks. There are three major categories of ML: unsupervised, semi-supervised, and supervised.

Unsupervised ML enables the learning process without available labeled data. The agent learns significant features to predict the unidentified structure or relationships in the input data. In semi-supervised ML models, the learning process uses semi-labeled datasets. This minimizes the amount of labeled data required when dealing with difficulties of obtaining large amount of labeling, such as in natural language processing task. However, irrelevant input features may provide incorrect decisions. Supervised learning deals with labeled data. In this technique, the datasets have a collection of inputs and target outputs. Using prior knowledge data is simpler than other techniques in the form of high-performance learning, but it takes a high labor cost to label all the data and the decision boundaries of a class can be tenuous [64]. Although there are a large number of models in the literature, the main algorithms for spectra analysis are discussed in the following, where only the K-Means clustering is a unsupervised method, while the subsequent models are described regarding their supervised forms.

3.5.1 K-Means

K-Means is one of the most used clustering techniques. Clustering algorithms are grouped in two major categories: hierarchical clustering and partitional clustering. K-Means is a partitional type, where a single partition of the dataset is produced, instead of agglomerative or divisive methods used in hierarchical clustering.

The standard K-Means algorithm finds the minimum squared error between the data points and the mean of a cluster, assigning each point to the closest cluster center. The minimization process for each cluster can be expressed by [65]:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} ||x_i - \mu_k||^2 \tag{15}$$

where x_i, the value of point i of the dataset, is partitioned into k clusters C, using the clusters centroids μ.

In this standard K-Means, the distance calculation uses the Euclidean distance metric, but other metrics may be employed. The number of clusters k is manually defined by the user in standard K-Means algorithms. Several iterations are demanded for a given k with different initial cluster centroids to optimize the cluster result, where the iterations number is also a hyperparameter defined by the user.

K-means naturally results in crisp clusters, struggling to detect overlapping classes of data. Time and space complexities are strongly dependent on the size of the input data, as the distances are calculated several times [65].

3.5.2 Linear discriminant analysis

Linear Discriminant Analysis (LDA) calculates discriminant projections vectors to achieve a data dimensionality reduction. These projections form the maximum intraclass separation and minimum interclass distance. Given two labeled data x_1 and x_2, the linear projection w that maximizes the interclass distance is obtained by the following [66]:

- Calculate the classes center m
- Calculate the intraclass and interclass scatter matrices

$$S_{inter} = (m_1 - m_2)(m_1 - m_2)^T \tag{16}$$

$$S_{intra} = S_1 + S_2 \tag{17}$$

where S_1 and S_2 are given by

$$S_i = \sum (x - m_i)(x - m_i)^T \tag{18}$$

- Compute the LDA projection by maximizing J

$$J = \frac{w^T S_{inter} w}{w^T S_{intra} w} \tag{19}$$

- To solve the J optimization problem, i.e., the minimum w for $w^T S_{intra} x$ and $w^T S_{inter} w = c$, where c is a non-zero value, the Lagrangian can be constructed as:

$$L(w, \lambda_{lda}) = w^T S_{intra} w - \lambda_{lda}(w^T S_{inter} w - c) \tag{20}$$

$$\frac{\partial L(w, \lambda_{lda})}{\partial w} = S_{intra} w \ - \lambda_{lda} \ S_{inter} w \tag{21}$$

- The optimal w can be obtained as

$$S_{intra} w = \lambda_{lda} S_{inter} w \tag{22}$$

LDA assumes that the data is gaussian shaped, i.e., a normal distribution. Spectral data usually does not present normal distribution, although LDA can handle modeling this kind of data, and according to the central limit theorem, increasing the dataset size can help to overcome this issue. The number of samples should be significantly larger than the number of features. In addition, LDA is negatively affected when the attribute variance is not equally distributed, which is usually the case of complex biological media [17].

3.5.3 Partial least squares discriminant analysis

The Partial Least Squares (PLS) seeks to maximize the covariance, finding the direction of the space of the predictors that explains the greatest variance of class space. PLS in its classical form is based on nonlinear iterative partial least squares

(NIPALS) algorithm. Given a sample data X, its label matrix Y, and a number of latent variables, NIPALS steps are [66]:

- Calculate the sample weights and normalize it

$$w = X^T y \tag{23}$$

$$w = \frac{w}{||w||} \tag{24}$$

- Obtain the data scores

$$t = Xw \tag{25}$$

- Calculate the weight of the labels and normalize it

$$c_y = \frac{Y^T t}{(t^T t)} \tag{26}$$

$$c_y = \frac{c_y}{||c_y||} \tag{27}$$

- Get the vector u

$$u = Y c_y \tag{28}$$

- Iterate steps 1 to 4 until t converges, i.e., current t equal to last t
- Compute the loading vector of X and Y, respectively

$$p = \frac{X^T t}{t^T t} \tag{29}$$

$$q = \frac{Y^T t}{t^T t} \tag{30}$$

- Calculate the coefficient b

$$b = \frac{u^T t}{(t^T t)} \tag{31}$$

- Update data and label matrices as

$$X = X - tq^T \tag{32}$$

$$Y = Y - tq^T \tag{33}$$

- The vectors t, p, u, b and q can be saved and the next component can be obtained restarting the first step

The PLS model looks for the multidimensional direction in the data space that explains the maximum variance direction in the label space:

$$X = TP^T + E \tag{34}$$

$$Y + UQ^T + F \tag{35}$$

$$U = TD + H \tag{36}$$

where T and U are the score matrices; P, Y and Q are the loadings; and E, F and H are the residuals.

The discriminant analysis (DA) term for PLS refers to the use of a threshold after the decomposition to enable a classification. As it is a binary classifier, the threshold is usually set to 0.5. PLS-DA is negatively affected by imbalanced classes and the number of latent variables requires a grid search optimization.

3.5.4 K-nearest neighbors

The K-Nearest Neighbors (KNN) algorithm calculates the distance between a test data and the training data (labeled) [67]. Euclidean Distance is frequently used, but others may also be applied, as they are special cases of a more general family of distance functions, L_k [67]:

$$d_k(p,q) = \sqrt[k]{\sum_{i=1}^{d} |p_i - q_i|^k} \tag{37}$$

where p and q are a set of points. When $k = 1$, Manhattan distance is calculated; $k = 2$ is the Euclidean Distance; and $k = 3$ is the Maximum Component.

The nearest K data define the classification of the tested data by a voting system. Voting can be uniform, where all the data has the same weight, or weights can be assigned to the data, such as the inverse of the distance, increasing the influence of closer data. k values usually range from 3 to 50, and the optimal value should be grid searched.

For a classification with a large number of data, search optimization algorithms are used, such as Voronoi diagrams and k-d tree, partitioning the dimensional space. The effectiveness of these optimizations and the high training time are highly affected by the number of features, and dimensionality reduction techniques, such as PCA, may be applied before modeling.

KNN can be modeled into almost all type of data and distribution, although imbalanced classes may add a larger class bias and lead to an overfitting issue. When a new sample has to be classified, KNN must re-run all the model training, calculating the distance metrics again, thus being considered a lazy model [17].

3.5.5 Support vector machines

Support Vector Machines (SVM) look for a hyperplane that presents the best separation between two classes, based on the points closest to it, called support vectors [67]. If a point is erroneously separated, i.e., it is not on the separation side of its correct class, then its distance to the hyperplane is given as an error. The model error is calculated through the sum of all errors, where a cost constant is added as a penalty for incorrect classification. The separating plane of linear SVM can be written as:

$$w \cdot x - b = 0 \tag{38}$$

where w is a vector of coefficients and x the input variables. Thus, the constraints for class 1 and class -1 for each x_i point are, respectively:

$$w \cdot x - b \geq 1 \tag{39}$$

$$w \cdot x - b \leq -1 \tag{40}$$

The problem optimization for classes y_i of x_i is:

$$\mathit{max}\,||w||\,, \tag{41}$$

$$\text{where } y_i(w \cdot x_i - b) \geq 1, \tag{42}$$

$$\text{for all } 1 \leq i \leq n \tag{43}$$

For application in non-linear models, SVM use the concept of kernel, where a function is applied to the predictors, in order to increase dimensionality and make separation possible. Most used kernels are the polynomial and the radial basis function (RBF or gaussian), respectively defined by:

$$K_{poly}(x_1, x_2) = (a + {x_1}^T x_2)^b \tag{44}$$

$$K_{RBF}(x_1, x_2) = \mathit{exp}(-\gamma(d_{12})^2) \tag{45}$$

where b is the polynomial order, a is a constant term, d_{12} the euclidean distance between x_1 and x_2, and γ is the inverse of the radius of influence of samples selected by the model as support vectors [67].

The kernel type and parameters optimization is a laborious step, yet RBF usually presents the best adaptation to several data distributions. As SVM are binary classifiers, multiclass problems can be solved using “One versus Rest” or “One versus One” approaches. SVM tend to struggle with the increasing data complexity and size. Overlapped classes also compromises the model performance [17].

3.5.6 Random forest

The Random Forest (RF) algorithm is an ensemble learning method using decision trees. A decision tree is generated by dividing the data into nodes, which are divided into subsequent nodes through binary choices of a predictor, for example, if the intensity in a wavenumber is greater or lower than a certain value, until it arrives at a classification, given by final nodes, called leaves. This whole structure is also called Classification and Regression Trees [67].

To evaluate each predicate and how they will contribute to partitioning the set S of the samples, it is used the information entropy or Gini impurity. Given f_i as the fraction of S, the information entropy is defined as

$$H(S) = -\sum_{i=1}^{m} f_i log_2 f_i \tag{46}$$

The potential split is evaluated by how much it decreases the system entropy. Considering a predicate splitting S in two subsets, the information gain function is given by

$$G(S) = H(S) - \sum_{j=1}^{2} \frac{|S_i|}{|S|} H(S_i) \tag{47}$$

If Gini impurity is used, then it is based on another quantity $(f_i(1 - f_i))$:

$$G(f) = \sum_{i=1}^{m} f_i(1 - f_i) \tag{48}$$

The number of knots can be limited through the concept of pruning, in order to reduce overfitting. When N trees are built, where N is defined by the user, usually with initial tests of 20 to 100, a RF is obtained. The final classification is given by a voting system for each of the trees, which can follow a uniform or weighted voting. This strategy of combining different classifiers into one is called ensemble learning. To avoid high correlation between trees, the bagging or bootstrap approach is used to build the best possible tree in small random subsets of predictors.

RF is robust to outliers and presents lower overfitting than most of the machine learning algorithms, as it performs feature selection and generates uncorrelated decision trees. The subsampling method also makes it a good model for dealing with data with large feature quantity. Sparse data usually decreases the model performance [67].

3.5.7 Extreme gradient boosting

Extreme Gradient Boosting (XGB) is a gradient tree boosting implementation. Boosting is a process to weight the classifiers based on how hard the classification is. The gradient descent algorithm is used to minimize the classification errors. In this way, each tree learns from the previous, instead of random non-related trees like in the RF algorithm. To learn the set of functions used in the model, the following objective function is minimized [68].

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{49}$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \left|\left|w\right|\right|^2 \tag{50}$$

The first term in L is related to the loss function, where l calculates the residuals of the predicted $\hat{y}_i$ and the true y_i values. The second part, Ω, is a regularization term, where the number of leaves T is pruned by the penalty γ, and the leaf weights w is regularized by the λ term. The objective function has functions as parameters and cannot be optimized by traditional methods, hence it is trained in an additive manner, where the f_t that improves the model the most is added:

$$L(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{51}$$

XGB shows a fast model training in comparison to most machine learning algorithms. As each tree learns from the previous, it is more prone to overfitting than RF ensemble learning, however, it usually learns better from the features and presents a higher performance. To overcome the overfitting issue, the regularization term helps to generalize the model. Preprocessing steps to remove data noise may also play an important role for the modeling [68].

3.5.8 Deep learning

Deep Learning (DL), a subset of ML, is inspired by the information processing patterns found in the human brain. DL does not require any human-designed rules to operate; rather, it uses a large amount of data to map the given input to specific labels [64, 69]. It presents a very good performance at understanding complex high-dimensional data, which broadens its scope to several domains, such as science, business and government. DL has outperformed other ML techniques in several tasks, specially in image and speech recognition, and with complex medical data [20].

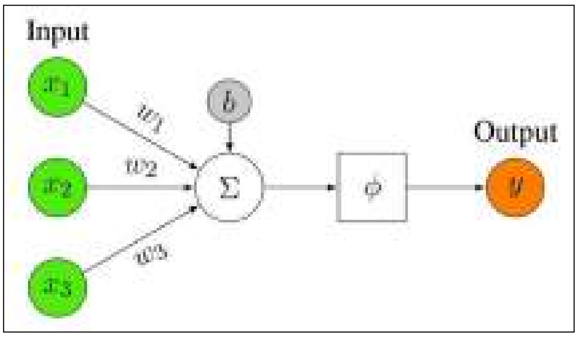
There are plentiful types of DL approaches and model architectures. Next, it is described the basis of every DL approach, the Feedfoward Neural Network (FFNN), and the most common approach for image recognition, the Convolutional Neural Network (CNN).

3.5.8.1 Feedfoward neural network

The basic entity of any neural network is a model of a neuron, where an input (x) and a bias (b) are weighted (w) and then summarized together. The bias term is usually omitted from the equations. The sum of these terms forms the argument of an activation function, ϕ, leading to the output of the neuron (Figure 6), as per the following equation [70]:

$$y = \phi(z) = \phi(w^T x + b) \tag{52}$$

Figure 6 – Representation of a neuron. Inputs x, weights w, bias b, activation function ϕ, and output y.



Source: Emmert-Streib et al., 2020 [70].

Although a liner activation function may be employed, such as in the output neuron of regression tasks, where the output is continuous and non-limited, the activation functions used in a deep learning architecture usually perform non-linear transformation. Some of the most used activation includes ReLU (Rectified Linear Unit), LeakyReLU, Sigmoid, tanh, and Softmax, respectively define by [64, 70]:

$$\phi(x)_{\text{ReLU}} = \text{max}(0, z) \tag{53}$$

$$\phi(x)_{\text{Sigmoid}} = \frac{1}{1 + e^{-z}} \tag{54}$$

$$\phi(x)_{\text{Softmax}} = \frac{e^z}{\sum_i^n e^z} \tag{55}$$

$$\phi(x)_{\text{Tanh}} = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{56}$$

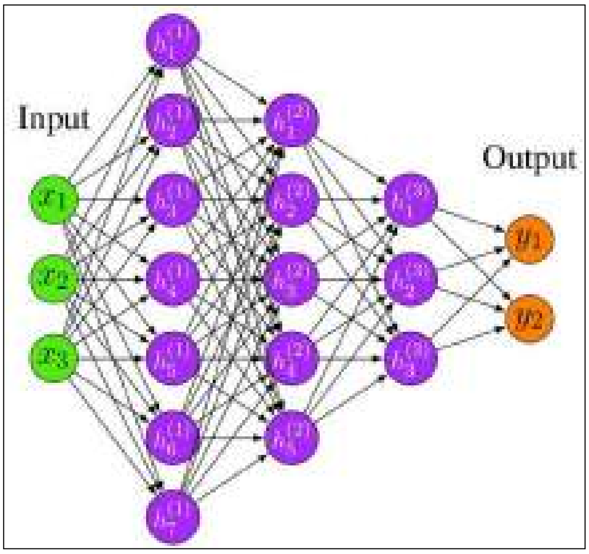
$$\phi(x)_{\text{LeakyReLU}} = \begin{cases} z, & z > 0 \\ mz, & z \leq 0 \end{cases} \tag{57}$$

where m is a small value, such as 1e-3.

These functions need the ability to differentiate, allowing the learning process by applying gradients, a partial derivatives method better explained later on. ReLU is the most used activation function. Even so, it may lead to the Dying ReLU issue, where a larger gradient is generated and the ReLU function will update the sum of the weights in a way that the neuron will not activate once more. The LeakyReLU is usually employed to solve scenarios with this issue. Sigmoid, Tanh and Softmax are mostly employed within the final classification neurons layer, where Softmax specially deals with multi-class tasks [64].

To build a neural network (NN), several neurons have to be connected by each other. The feedforward structure is the most basic one, where the layers between the input and the output are called hidden layers. When more than two hidden layers are stacked, the architecture is commonly considered as a Deep Feedforward Neural Network (D-FFNN, Figure 7).

Figure 7 – Representation of a Deep Feedforward Neural Network (D-FFNN). Input neurons x, hidden neurons h, output neurons y.



Source: Emmert-Streib et al., 2020 [70].

A FFNN needs a learning rule to optimize its parameters. Hence, an optimization algorithm finds the parameters that minimize the error of the training data, given by a cost function. Approximating an unknown function f^* can be written as:

$$y = f^*(x) \approx f(x, w) \approx \phi(x^T w) \tag{58}$$

where f is a function dependent of parameters θ, and ϕ the non-linear activation of one layer. If several hidden layers are considered, ϕ can be written as [70]:

$$\phi = \phi^{(n)}(...\phi^{(2)}(\phi^{(1)}(x))...) \tag{59}$$

The learning process to minimize the training error can be accomplished by the Gradient Descent or a gradient-based learning algorithm. It operates by iteratively updating the network parameters throughout each training epoch. The algorithm computes the gradient (slope) of the objective function via first-order derivative of the network parameters. Subsequently, the parameter is adjusted in the opposite direction of the gradient to effectively decrease the error. This process is accomplished through backpropagation, where the gradient at each neuron is propagated backward to all neurons in the preceding layer. The equation of this operation is [64]:

$$w_{ij^t} = w_{ij^{t-1}} - \Delta w_{ij^t} = w_{ij^{t-1}} - \left(\alpha \frac{\partial E}{\partial w_{ij}}\right) \tag{60}$$

where w denotes the weight i in the epoch t, α is the learning rate, defined by the user, and E is the error of the prediction in comparison to the target.

Although the backpropagation algorithm can solve the parameters optimization, other gradient-based algorithms are employed for a better computational efficiency, such as the Stochastic Gradient Descent (SGD) and the Adaptive Moment Estimation (Adam). SGD approach involves presenting the input vector for a few examples, calculating their corresponding outputs and errors, computing the average gradient based on these examples, and adjusting the weights accordingly. This process is then repeated for numerous small sets of examples from the training set until the average of the objective function no longer decreases. The term "stochastic" arises from the fact that each small set of examples provides a noisy estimate of the average gradient across all examples [20].

Adam optimization calculates an adaptive learning rate for each parameter, estimating not only the first moment (the mean) of the gradient like other gradient-based algorithms, but also the second raw moment (the uncentered variance), through the calculation of an exponential moving average of these gradients. The decay rate of these moving averages are controlled by the parameters β_1 and β_2, which are manually selected by the user. Similar to the basic backpropagation, the equation of the weights update using Adam can be represented by [71]:

$$w_{ij^t} = w_{ij^{t-1}} - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{61}$$

where α is the learning rate, ϵ is a small constant for numerical stability, such as 1e-8, and $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second raw moment estimate, respectively, given by:

$$\hat{m}_t = \frac{\beta_1 m_{t-1} + (1 - \beta_1) g_t}{1 - \beta_1^t} \tag{62}$$

$$\hat{v}_t = \frac{\beta_2 v_{t-1} + (1 - \beta_2) g_t^2}{1 - \beta_2^t} \tag{63}$$

where g_t is the gradient applied to the cost function.

A loss function, part of the cost function, is applied to quantify how much the model prediction probability (p) differs from the target class (y). There are several loss functions, such as the cross-entropy (CE), commonly applied for classification tasks, based on the softmax activation, and the mean squared error (MSE), widely applied to regression problems. Their equations are described as [64]:

$$CE = -\sum_{i=1}^{n} y_i \text{log}(p_i) \tag{64}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2 \tag{65}$$

Class weights may be applied to loss functions when dealing with imbalanced datasets, i.e., when there is considerably more input classes from one class than the others. In this way, the loss will greater punish the predictions related to the class with low amount of examples, indicating to the optimization algorithms to better learn this prediction.

To properly evaluate a model, several metrics are employed both during and after the training phase, thus providing enough information to choose the best model. The most used metrics for classification tasks are related to True Positive (TP) and True Negative (TN) values, successfully classified positive and negatives instances, respectively, and False Positive (FP) and False Negative (FN) values, misclassified positive and negative instances, respectively. The metrics can be formulated as [64]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{66}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{67}$$

$$Specificity = \frac{TN}{TN + FP} \tag{68}$$

$$Precision = \frac{TP}{TP + FP} \tag{69}$$

$$F1_{score} = 2 \cdot \left(\frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \right) \tag{70}$$

Regression metrics may include the previous described MSE, and Mean Absolute Error (MAE) and Root-Mean-Square Error ($RMSE$). The last two are defined by [72]:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - p_i| \tag{71}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - p_i)^2} \tag{72}$$

To deal with overfitting issues, regularization approaches may be employed to NN, such as dropout and batch normalization. Dropout assists the generalization of the model, where neurons and their connections are randomly dropped by each training epoch, acting like a feature selection method. Batch normalization guarantees the performance of the output activation by applying a Standard Normal Variate (SNV) normalization, acting like a preprocessing tool inside each layer of the NN. This stabilizes the model by helping to prevent vanishing gradients and consuming less time for converge, besides a minor influence on regularization [64].

3.5.8.2 Convolutional neural network

Convolutional Neural Networks (CNN) were inspired by cells in animal visual cortex that detects light in receptive fields. The first application was to classify images of handwritten digits, presenting representations of the original image to identify visual patterns [73].

The input of a CNN is frequently a three-dimensional (3D) matrix of size m x n x d (width x height x depth or channels), usually with $m = n$ for computational efficiency reasons. This 3D matrix is basically the organization of a 2D image with three color channels (RGB – Red, Green and Blue). Yet, a CNN can be build and trained for any dimensional size, as long as computing power is available. Kernels are the basis of connections in a CNN. They share the same bias (b) and weights (w) to generate k feature maps (h). The convolution layer calculates a dot product ($*$) between its inputs and weights, as per the following [64]:

$$h^k = f(w^k * x + b^k) \tag{73}$$

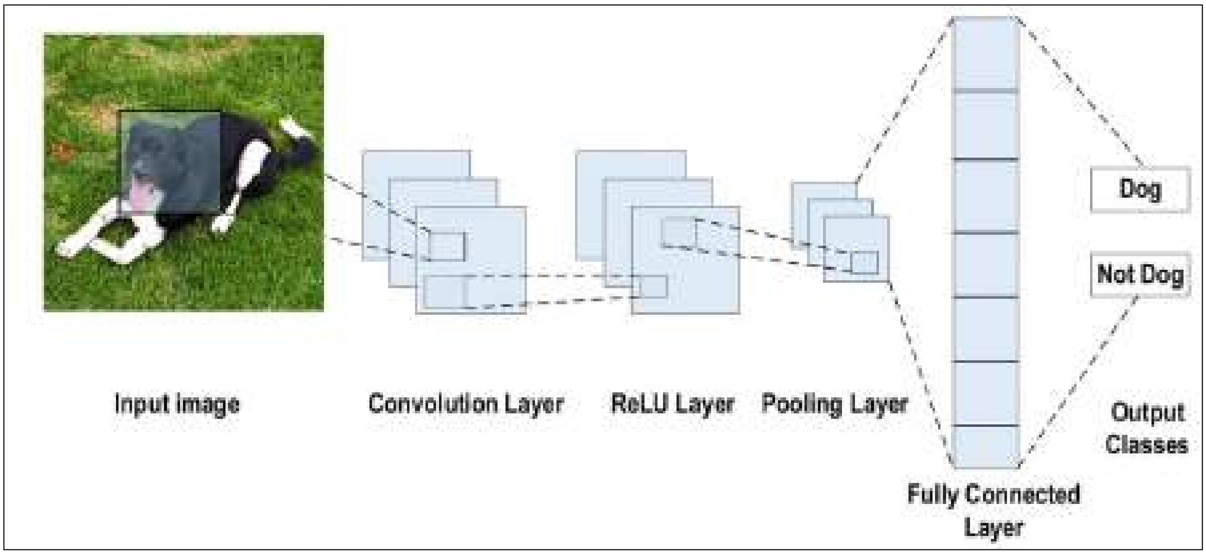
The convolution process relies on the definition of spatial attributes, where the most common one are the size of kernels (N), also called the receptive field or window size, stride (S) and zero-padding (P). N is the size of the kernel matrix, given by width x height, usually N x N, for a 2D kernel. The convolution is applied only to the inputs inside the specified window, and then the kernel take a S amount of steps, in pixels, before applying the next convolution. Therefore, when $S > 1$, a downsampling task is performed to the feature maps. P is the quantity of zeroed rows and columns to be added to the image. This is useful for preserving the original image dimension. The number of output feature maps (h_{out}) of a convolutional layer is defined by the selected number of kernels, i.e., output depth $d_{out} = k$; while the output image width (m_{out}) and height (n_{out}) can be calculate from the input image width (m_{in}) and height (n_{in}) by [70]:

$$m_{out} = \frac{m_{in} - N + 2P}{S + 1} \tag{74}$$

$$n_{out} = \frac{n_{in} - N + 2P}{S + 1} \tag{75}$$

CNN are usually composed of different types of layers, besides the convolution layer, such as activation, pooling and fully connected layers (Figure 8). Fully connected (or dense) layers, are the same as in FFNN. When no dense layers are added at the end, the designated term is a Fully Convolutional Network (FCN), instead of CNN [74].

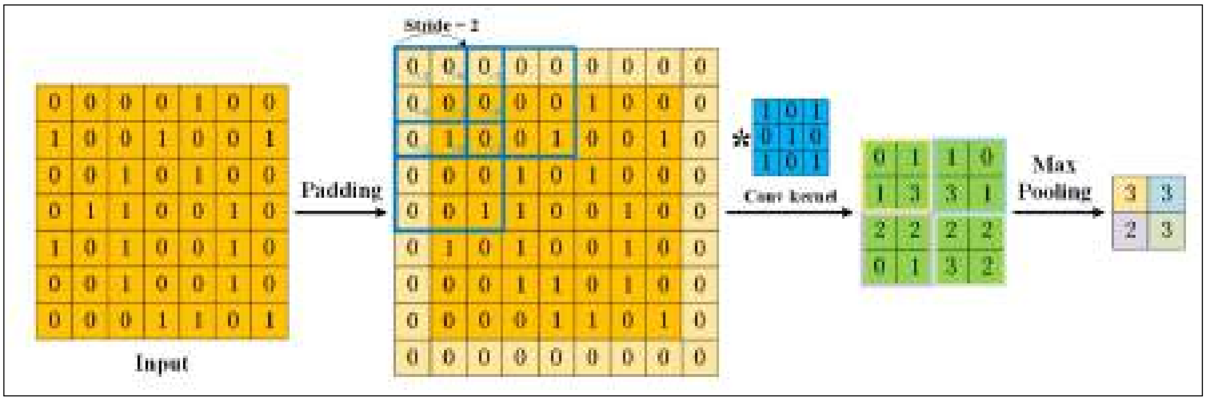
Figure 8 – Example architecture of a Convolutional Neural Networks (CNN) for a colored image classification.



Source: Adapted from Alzubaidi et al., 2021 [64].

Pooling layers are mainly applied to perform the downsampling of the feature maps with lower computational cost than convolutions, besides adding spatial invariance into the network, which can assist to improve the model generalization. Pooling operations use the same parameters of window size, stride and zero-padding as for the convolutional operations. In the same manner as the kernels, only the inputs inside the specified window are considered before taking the stride for the next inputs. There are several types of pooling, where two of the most basic and common are the average and max pooling. The first one calculates the mean value of the inputs in the window, while the later extracts the maximum of the inputs [70]. Figure 9 exhibits the procedure of a CNN regarding the convolution and pooling operations.

Figure 9 – Procedure of a 2D CNN.



Source: Li et al., 2022 [74].

Activation layers perform the previous describe non-linear transformations of the activation functions. ReLU speeds up the compute time in comparison to Sigmoid and Tanh, besides inducing sparsity in the hidden layers. Still, its discontinuity at zero may negatively affect the backpropagation performance. As in FFNN, Dying ReLU is also a problem in CNN, resulted from the zero gradient. It can also be solve by using LeakyReLU, since it allows a small non-zero gradient when the unit is not active [73].

The same regularization approaches, loss functions and metrics described for FFNN can be applied to CNN.

4 MATERIAL AND METHODS

This thesis is split in three studies. The first one to define an optimal number of co-added scans; the second, a one-dimensional (1D) deep learning approach to classify cancer and the subtype, and to assess the biochemical content impact; the third, a two-dimensional (2D) deep learning to classify cancer, subtype, and biomarkers levels expression.

4.1 Scan Study

4.1.1 Dataset

SK-BR-3 cells (ATCC number: HTB-30), a HER 2 subtype, ER/PR negative, HER2 positive, and BT-474 (ATCC number: HTB-20), a luminal B subtype ER/PR/HER2 positive [13], were cultured in DMEM (Gibco, Life technologies, MD, USA) supplemented with 10% of fetal bovine serum (Gibco, Life technologies, MD, USA) and 50 μg/mL of gentamicin (Gibco, Life technologies, MD, USA). For in vivo studies, eight-week-old female Balb/c nude mice were subcutaneously injected with 1x10^6 BT-474 or SK-BR-3 cells and tumor growth was followed for 4 weeks. Balb/c nude mice were bred at the animal facility of Nuclear and Energy Research Institute, and all experiments complied with the relevant laws and were approved by local animal ethics committees (protocol number: 203/17). When tumors volume reached approximately 0.5 cm^3, biopsies were collected and processed by formalin fixation and paraffin embedding (FFPE) method. Twenty sections of 5 μm, ten for each cell line, were then obtained using a microtome and fixed in low-e microscope slides (MirrIR, Kevley Technologies, USA).

Spectral images acquisition was accomplished using a Cary Series 600 system (Agilent Technologies, USA), composed by a Cary 660 FTIR spectrometer and a Cary 620 FTIR microscope. This system has a focal plane array (FPA) detector of 32x32 elements and 5.5 μm spatial resolution, providing 1024 spectra per acquisition. The system was set to operate between 3950 and 900 cm^{-1}, with 4 cm^{-1} spectral resolution in transflection mode, due to the use low-e slides.

In each histological section, the FTIR image was acquired by varying twice the number of co-added scans of the background, where each one was acquired before a batch of varying sample scans six times, totalizing 12 acquisition per section. Background scans were set to 128 and 256, while sample scans were set to 4, 8, 16, 32, 64 and 128. The groups were labeled as “bx_y”, where x and y are the number of scans for the background and sample, respectively. Furthermore, adjacent paraffin regions images were acquired, varying scans from 4 to 128 as sample scans.

For a reproducibly purpose, the same histological section region was settled during the collection of all scans. Hence, single acquisitions were performed instead of mosaics (grouping several single acquisitions in one measure), resulting in 12288 sample spectra acquired for each section. In addition, scans of the same section were collected in sequence, within the same spatial position, even for different background scans.

4.1.2 Data preprocessing and analysis

Data preprocessing was applied according to the protocols [13,16,17]. The pipeline was defined depending on the analysis, as describe in the next paragraph. General steps included fingerprint truncation; Savitzky-Golay (SG) filtering for smoothing with window size of 7 and for obtain the second derivative; extended multiplicative signal correction (EMSC) [55], with a zero degree polynomy, as second derivative does not need baseline correction, and digital de-waxing [56]; and outlier detection (quality test) using Hotelling’s T^2 vs Q residuals with a fixed removal threshold of 95% confidence interval.

The analysis was divided in two: a spectral analysis to better visualize the spectra distribution and variations; and a classification analysis to check how each group performs within the machine learning approaches. For the spectral analysis, each scan group was preprocessed by three different pipelines, which originated three main groups division: RAW, PP and OUT. The Table 1 describes the applied preprocessing steps in each one.

Table 1 – Preprocessing steps applied in each group: RAW, PP, OUT.

Step	RAW	PP	OUT
Outlier	✗	✗	✓
Fingerprint	✓	✓	✓
Derivative	✓	✓	✓
Smoothing	✗	✓	✓
EMSC	✗	✓	✓

Cell lines were not considered as a grouping feature for spectral analysis, therefore only scan number and preprocessing steps were compared in this part. The groups were analyzed using mean ± standard deviation (SD) plots and principal component analysis (PCA) scores plots. The PCA scores and all the other analysis which used the fingerprint preprocessing refer to the biofingerprint region (1800 to 900 cm^{-1}), excepts the spectra for mean + SD plots, where the dead region (2000 to 1800 cm^{-1}) was added, hence showing spectra truncated from 2000 to 900 cm^{-1}.

For the classification analysis, PP and OUT groups division were used, besides the scan group division. In addition, another main division was also applied in this part, splitting the previous groups according to the cell line. In this way, the classification could be compared by scan number, preprocessing steps, and machine learning technique.

Six machine learning classifiers were modeled using default hyperparameters:

- *Linear Discriminant Analysis (LDA)*: number of components = 1; singular value decomposition solver with tolerance (significance threshold) = 1e-4.
- *Partial Least Squares Discriminant Analysis (PLS-DA)*: number of components = 10; nonlinear iterative solver with tolerance (convergence criteria) = 1e-6; maximum number of iterations = 500; discriminant threshold = 0.5.
- *K-Nearest Neighbors (KNN)*: number of neighbors = 5; Euclidean distance metric.
- *Support Vector Machine (SVM)*: radial basis function kernel with gamma coefficient = number features times feature variance; cost parameter (regularization) = 1; tolerance (convergence criteria) = 1e-3.
- *Random Forest (RF)*: number of trees = 100; split metric = Gini impurity; maximum features = 21 (square root of the number of features); minimum impurity

decrease = 0; no pruning.

- *Extreme Gradient Boosting (XGB)*: gbtree booster; learning rate = 0.3; maximum depth = 6; L2 regularization = 1; no L1 regularization; minimum loss decrease = 0;

Model training was performed by a stratified 5-fold cross-validation (CV), varying the fold split seed 10 times, resulting in 50 trainings per model. Preprocessing steps were applied after each fold split, where test data were only transformed according to the train fitting, preventing information leakage between train and test data. Models were assessed independently, where intra-groups test accuracies, i.e., different scans performances from a same model and same PP or OUT group, were evaluated using Friedman + Nemenyi test [75] with a significance level of 5%.

All the study was accomplished using in house algorithms in Python, except by Friedman + Nemenyi test, written in R language.

4.2 One-dimensional Deep Learning

4.2.1 Dataset

The BR804b (Biomax, Inc, USA) breast cancer microarray was ordered with formalin fixation and paraffin embedding (FFPE) histological sections of 8 μm, fixed in calcium fluoride (CaF_2) slides (Crystran, UK). A total of 60 cores of 1.5 mm were imaged, one Cancer and one Adjacent Tissue (AT) core for each of the 30 unique patients. Cores were already classified by Biomax regarding their type, and receptors and Ki67 expression levels. Molecular subtypes were classified based on St. Gallen International Expert Consensus guidelines [8, 9], thus resulting in the distribution (where N denotes the samples number): Type – Cancer (CA; N=30) and Adjacent Tissue (AT; N=30); Subtype – Luminal A (LA; N=8), Luminal B (LB; N=8), HER2 (N=7) and Triple-negative Breast Cancer (TNBC; N=7).

Image acquisition was performed using a Cary Series 600 system (Agilent Technologies, USA), composed by a Cary 660 FTIR spectrometer and a Cary 620 FTIR microscope. It was used the focal plane array (FPA) detector to acquire 320x320 pixels hyperspectral images with 5.5 μm spatial resolution for each core, resulting in 6,144,000 raw spectra. Tissue and borderline paraffin were gathered. The system was set to operate between 3950 and 900 cm^{-1}, with 4 cm^{-1} spectral resolution in transmission mode. Background images were acquired with 256 co-added scans, while 64 scans were set to the sample images acquisition.

A single 320x320 image of environment spectra was collected using a clean slide to obtain water vapor (H2O) variation. The air purge of the FTIR microscope acrylic box was turned off and the box left open when this acquisition started to increase the H2O variation.

4.2.2 Data preprocessing

Preprocessing steps were applied individually to each image. Tissue and paraffin regions were selected using a two-step K-means clustering, while pure slide spectra were excluded. Firstly, for the tissue identification, raw spectra were truncated at the Amide I and II region (1700 to 1500 cm^{-1}) and modeled with $k = 2$. Secondly, for the paraffin identification, raw spectra were truncated at the highest paraffin intensity band (1480 to 1450 cm^{-1}) and also modeled with $k = 2$, but spectra previously clustered as tissue were set to zero intensity. An area integration check was performed to guarantee tissue and paraffin as cluster 1 in each K-means.

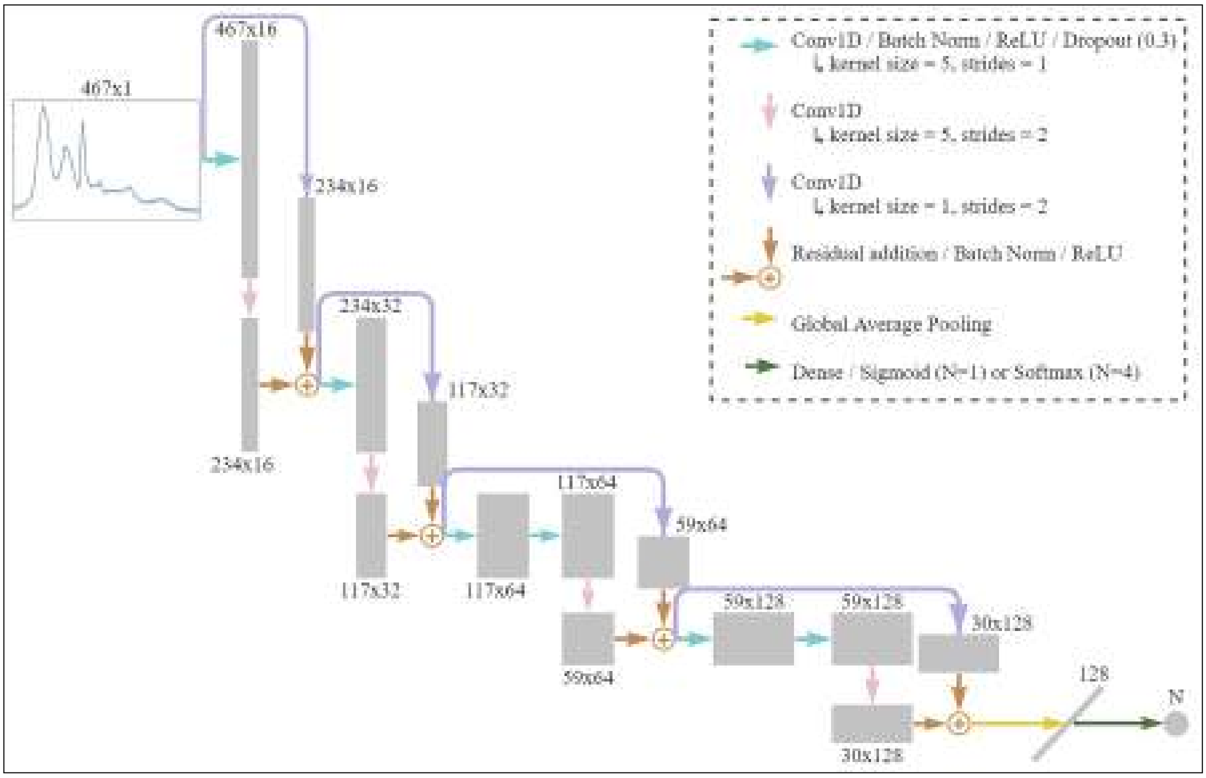
A biofingerprint truncation was applied from 1800 to 900 cm^{-1}, resulting in 467 points. This number of point is due to the spectral resolution and zero padding applied by the equipment in the interferogram. Outlier removal was applied using the Hotelling's T^2 vs Q residuals method, with 10 Principal Components (PC) and a 95% confidence interval fixed threshold removal. Data were smoothed using Savitzky-Golay (SG) filtering with window size of 11 and polynomial order of 2. These steps were also applied to the H2O spectra.

Extended Multiplicative Signal Correction (EMSC) [55] was applied coupled with paraffin removal [56], from 1500 to 1350 cm^{-1}, and water vapor (H2O) removal, from 1800 to 1300 cm^{-1}. The EMSC model was built by a polynomial baseline of order 4; PCA from paraffin and H2O using the number of PCs that corresponded to 99% of explained variance; and global mean spectra from tissue, paraffin, and H2O as references. The model was solved by least squares estimation. Spectra were normalized by min-max method and another outlier removal was performed using Hotelling's T^2 vs Q residuals method. Preprocessed spectra were saved in a HDF5 file. Spectra were labeled as binary encoding for the type classification, while the subtype was handled by a one-hot encoding.

4.2.3 Deep learning

A novel 1D convolutional neural network (CNN) called CaReNet-V1 was developed based on VGG (Visual Geometry Group) [76], ResNet (Residual Neural Network) [77, 78] and reported 1D models for spectroscopy analysis [79–82]. Figure 10 presents the CaReNet-V1 architecture.

Figure 10 – CaReNet-V1 architecture.


467x1
467x16
234x16
234x16
234x32
117x32
117x32
117x64
117x64
59x64
59x64
59x128
59x128
30x128
30x128
128
N
Conv1D / Batch Norm / ReLU / Dropout (0.3)
↳ kernel size = 5, strides = 1
Conv1D
↳ kernel size = 5, strides = 2
Conv1D
↳ kernel size = 1, strides = 2
Residual addition / Batch Norm / ReLU
Global Average Pooling
Dense / Sigmoid (N=1) or Softmax (N=4)

All convolutional layers were created using HeNormal kernel initialization [83] and zero padding. The type classification model was created with a single final neuron, sigmoid activation and binary cross-entropy loss, while subtype one was built using four final neurons with softmax activation and categorical cross-entropy loss. Models were trained using Adam optimizer [71], with learning rate of 1e-3, beta 1 of 0.9 and beta 2 of 0.999. A reduce learning rate on plateau callback [84] was employed to monitor the testing loss, with patience of 4, reduce factor of 0.5 and minimum learning rate of 1e-4.

A total of 4 patients were held-out for the test set, one for each subtype class and two for each type class. The 26 remaining patients were addressed for a balanced stratified 4-fold cross-validation, resulting in 21 patients for the train set and 5 for the development (dev) set of each fold, except for the last one, containing 20/6 (train/dev). The balance was executed by randomly undersampling the training spectra until reaching the same quantity per class. A batch size of 250 spectra was defined, along 50 training epochs. The training spectra order was randomly shuffled for each epoch using a data generator.

Performance was evaluated by assessing each spectrum prediction and the sample prediction, where the most predicted class among all sample spectra was determined as the final classification. A 0.5 threshold was applied for type classification and the maximum argument was chosen as the subtype prediction. The evaluation was accomplished by accuracy, specificity and sensitivity metrics, and by a 1D adaptation of Gradient-weighted Class Activation Mapping (Grad-CAM) [85]. Final Grad-CAM was calculated using the best fold model from dev set and averaging the Grad-CAM of each sample, grouping by classes. A min-max normalization was applied to generate the Grad-CAM heatmap. All the study was performed by in house algorithms in Python, mainly Tensorflow and Keras libraries, and using a GeForce GTX 1080 GPU with 8 GB of memory.

4.3 Two-dimensional Deep Learning

4.3.1 Dataset

A total of 60 cores from the BR804b (Biomax, Inc, USA) breast cancer microarray was imaged. Histological sections of 8 μm were formalin fixed and paraffin embedded (FFPE) in calcium fluoride (CaF_2) slides (Crystran, UK). The company also provided the Ground Truth (GT) labeling of receptors and Ki67 expression levels using immunohistochemistry (IHC). Molecular subtypes were classified in accordance with St. Gallen International Expert Consensus guidelines [8, 9]. Table 2 presents the distribution of the dataset acquired.

Table 2 – Dataset distribution regarding each label. Type with AT (adjacent tissue) and CA (cancer) classes; Subtype with LA (luminal A), LB (luminal B), HER2 and TNBC (triple-negative breast cancer); ER (estrogen receptor), PR (progesterone receptor), and HER2 receptors expression levels; Ki67 percentage levels. HER2 1$+$ and 2$+$, and other Ki67 levels were not considered due to small quantity (only one core).

Label	**Class**	**Quantity**
Type	AT	30
	CA	30
Subtype	LA	8
	LB	8
	HER2	7
	TNBC	7
ER	$-$	11
	$+$	3
	$++$	3
	$+++$	13
PR	$-$	15
	$+$	3
	$++$	2
	$+++$	10
HER2	0	16
	3$+$	11
Ki67	5%	12
	10%	6
	20%	4
	30%	3

Hyperspectral images mosaics of 320x320 were acquired using a Cary Series 600 system (Agilent Technologies, USA) with a focal plane array (FPA) detector of 32x32 and spatial resolution of 5.5 μm, resulting in a total of 6,144,000 raw spectra for the 60 cores. The full equipment range was collected, from 3950 to 900 cm^{-1}, with spectral resolution of 4 cm^{-1}, transmission mode, and 256 and 64 background and sample co-added scans, respectively.

During the image acquisition, the mosaic region was positioned to cover each core, while also collecting the paraffin around the tissue. In addition, it was acquired a single mosaic using a clean slide and turning off the air purge of the microscope acrylic box to obtain spectra with water vapor (H2O) variation.

4.3.2 Data preprocessing

Images were preprocessed individually. Tissue, paraffin and possible pure slide regions were selected using a two K-means clustering in sequence. The first one clustered the raw spectra truncated at the Amide I and II region (1700 to 1500 cm^{-1}) into two clusters: tissue and paraffin + pure slide. Raw spectra were then truncated at the highest paraffin intensity band (1480 to 1450 cm^{-1}) and tissue previously clustered were set to zero for the second K-mean, grouping paraffin and zeroed tissue + pure slide. Spectra were preprocessed by the following steps:

Spectra were truncated in the biofingerprint region (1800 to 900 cm^{-1}), decreasing the size to 467 points. This number of point is due to the spectral resolution and zero padding applied by the equipment in the interferogram. Outlier removal was performed by the Hotelling's T^2 vs Q residuals approach, with 10 Principal Components (PC) and removing spectra above the 95% confidence interval threshold. Spectra were smoothed adopting Savitzky-Golay method with window size of 11 and polynomial order of 2.

Extended Multiplicative Signal Correction (EMSC) [55] with digital de-waxing [56] and H2O removal was employed. PC quantity was selected until 99% of explained variance. Global mean spectrum, calculated from all samples, was used as reference. Baseline correction was accomplished by polynomial of order 4. EMSC model was solved by least squares estimation.

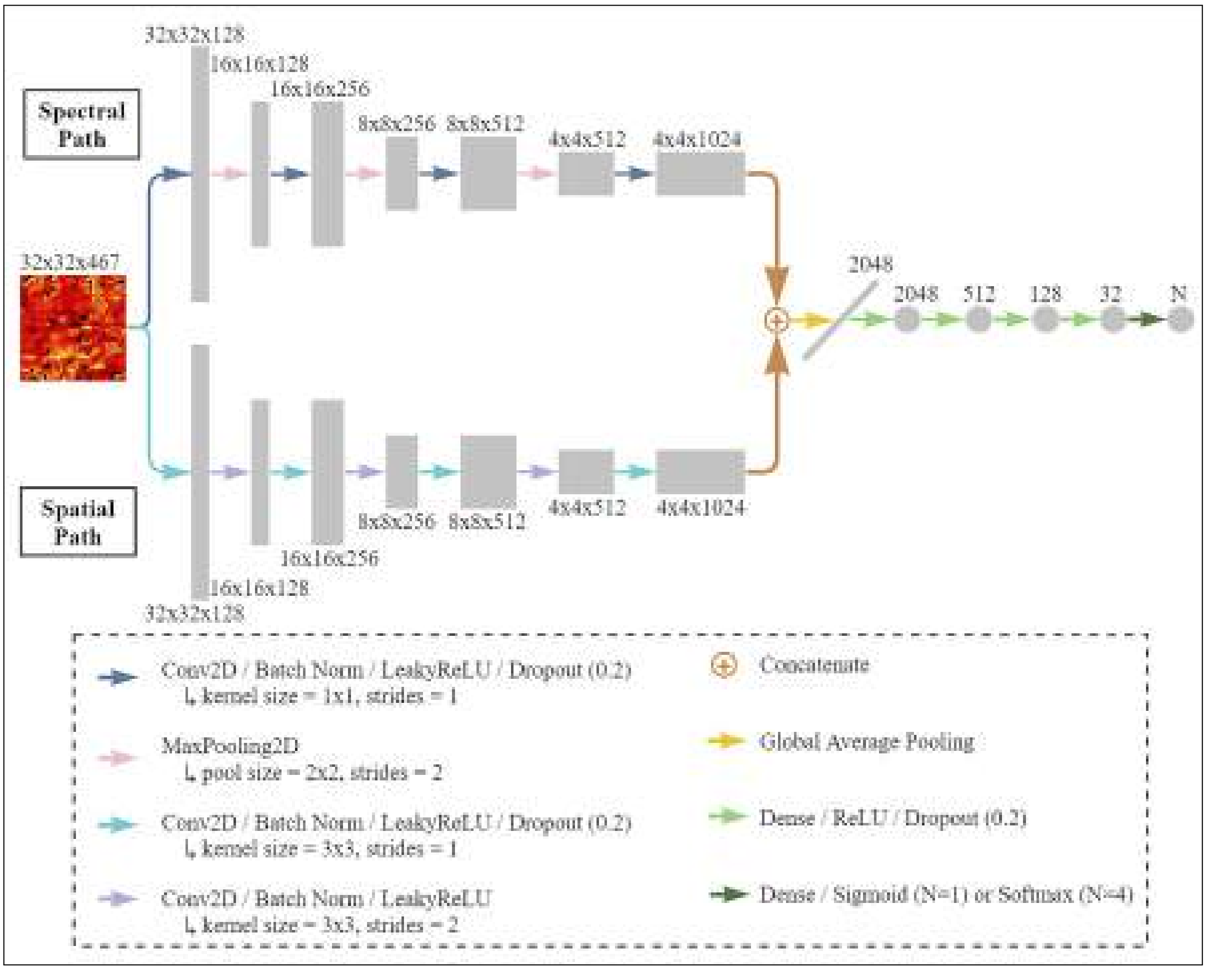
Corrected spectra were normalized by the min-max method, and a second outlier removal was applied. The 2D mosaics were reconstructed with preprocessed tissue spectra and zeroing paraffin and pure slide spectra. Final mosaics of 320x320x467 were divided into 6000 patches of 32x32x467. Patches with half or more of the pixels' quantity zeroed were excluded.

Patches were labeled by a binary encoding for the type and HER2 level ($1+$ and $2+$ were not considered due to quantity limitations) classification; by a one-hot encoding for the subtype classification; and by an ordinal one-hot-like encoding [86] for the receptor levels. The percentage of Ki67 was min-max scaled to a 0 to 1 expression fraction for a regression.

4.3.3 Deep learning

A 2D convolutional neural network (CNN) called CaReNet-V2 was developed inspired on hyperspectral images classification [87–89], VGG (Visual Geometry Group) [76], and generators of generative adversarial networks (GAN) [90, 91]. Figure 11 presents the CaReNet-V2 architecture. The model has two channels path: one to target spectral feature extraction; and another for spatial extraction.

Figure 11 – CaReNet-V2 architecture.


Spectral Path
Spatial Path
32x32x467
32x32x128
16x16x128
16x16x256
8x8x256
8x8x512
4x4x512
4x4x1024
2048
2048
512
128
32
N
Conv2D / Batch Norm / LeakyReLU / Dropout (0.2)
↳ kernel size = 1x1, strides = 1
MaxPooling2D
↳ pool size = 2x2, strides = 2
Conv2D / Batch Norm / LeakyReLU / Dropout (0.2)
↳ kernel size = 3x3, strides = 1
Conv2D / Batch Norm / LeakyReLU
↳ kernel size = 3x3, strides = 2
Concatenate
Global Average Pooling
Dense / ReLU / Dropout (0.2)
Dense / Sigmoid (N=1) or Softmax (N=4)

Convolutional layers were created using HeNormal kernel initialization [83]. Zero padding was applied to both convolutional and pooling layers. A total of six models were created, one per label, where the final dense layer, activation and losses were dependent to the encoding: binary (type and HER2) – single neuron, sigmoid activation, and binary cross-entropy loss; one-hot encoded (subtype) – four neurons, softmax, and categorical cross-entropy; ordinal (ER and PR) – four neurons, sigmoid, and square error [86]; regression (Ki67) – one neuron, linear, and mean squared error.

Adam [71] was set as the optimization algorithm, with learning rate of 1e-3, beta 1 of 0.9 and beta 2 of 0.999. A cosine decay schedule with restarts [92] was applied with initial learning rate of 1e-3, first decay step with the length of the training set, epochs multiplier in the decay cycle (t_mul) of 1.5, initial learning rate multiplier (m_mul) of 1.0, and minimum learning rate (alpha) of 1e-5. Class weights were calculated and applied to the losses to correct the learning process with imbalanced classes.

Four patients were held-out for the test set, resulting in one patient of each class for multi-class and regression models, and two of each for binary models. The 26 remaining patients, or 21 remaining for Ki67 regression, were split in train and development (dev) sets by a stratified 4-fold cross-validation. Type, subtype and HER2 were split with unique patients for train and dev, while the others, due to quantity limitations, were split by patches, presenting a same patient in both train and dev sets.

Models were trained by a batch size of ten patches by 300 epochs. Training patches were randomly shuffled and augmented by each epoch using a data generator. Data augmentation involved different random rotation (90°, 180° and 270°) and flip (horizontally and vertically) transformations every epoch, without duplicating data or increasing the dataset size.

Performance was evaluated by each patch prediction and by the sample prediction with a voting system. Final classification predictions were standardized in relation to the encoding: a fixed threshold of 0.5 for binary approaches (type and HER2); the maximum argument for one-hot encoded (subtype); the maximum argument above 0.5 for ordinal encoding (ER and PR). Then, for the sample voting system, the most predicted class among all sample patches was chosen as the final classification, and Ki67 final regression was defined by the mean of all patches. Classification evaluation was accomplished by accuracy, specificity and sensitivity metrics, while the regression was assessed using the mean absolute error (MAE), mean squared error (MSE) and root-mean-square error (RMSE).

Gradient-weighted Class Activation Mapping (Grad-CAM) [85] was employed to analyze spatial activation. The best dev set model from each group was selected to calculated the Grad-CAM using the last convolutional layer of the spatial path channel.

The channel importance, i.e., the contribution by wavenumber to the classification, was analyzed by summing the kernels values from the first convolutional layer of the spectral path. Spectral and spatial paths relative contributions were calculated by their respective sum of GAP feature values and first dense weights multiplication. All the study was performed by in house algorithms in Python, mainly Tensorflow and Keras libraries, and using a GeForce GTX 1080 GPU with 8 GB of memory.

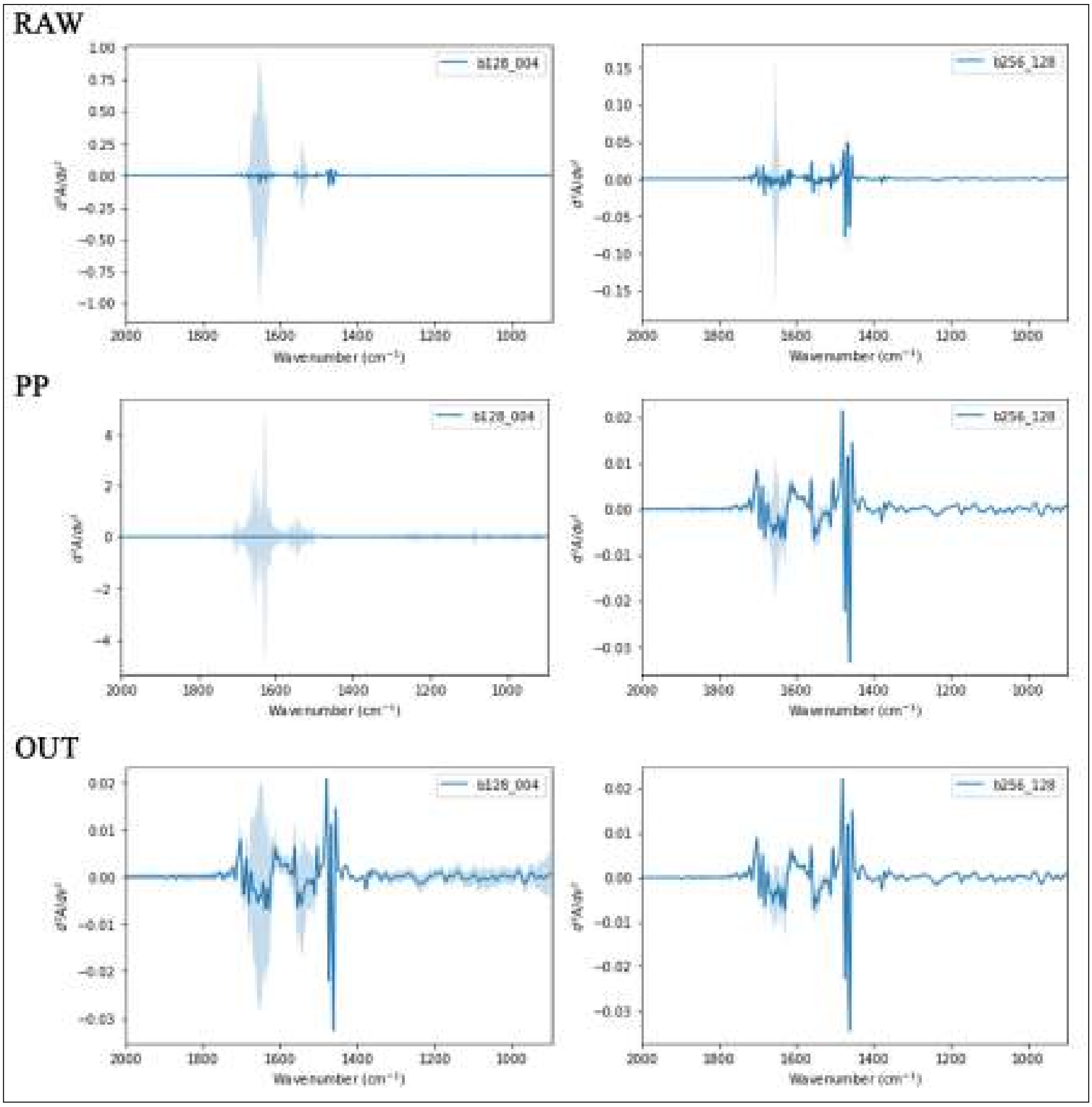
5 RESULTS AND DISCUSSION

5.1 Scan Study

5.1.1 Spectral analysis

The Figure 12 shows a representative spectra comparison of lowest and highest scans for both background and sample (b128_004 and b256_128). The RAW plots demonstrate greater SD of b128_004 group in comparison to b256_128, as evidenced by its wider range of shades in relation to its mean, especially at Amide I and II region (1700 to 1500 cm^{-1}) [13, 93]. The PP group presented a decrease in b256_128 SD, enabling a better visualization of the mean spectrum shape, while b128_004 did not present visual improvements due to the high SD. These findings suggest more outliers in b128_004, as the preprocessing techniques so far were not able to improve its quality. This tendency also occurs in OUT group plots, as the implementation of outlier removal method decreased b128_004 SD and made possible to distinguish its mean spectrum shape, which became like to b256_128. When compared to its PP, the OUT b256_128 group presented a slight decrease of its SD and a similar mean spectrum, indicating less impact of outliers, since there was less improvement than b128_004 after the outlier removal.

Figure 12 – Spectra comparison of b128_004 and b256_004 for groups RAW, PP, and OUT. Mean spectrum (solid line) and standard deviation (shades).

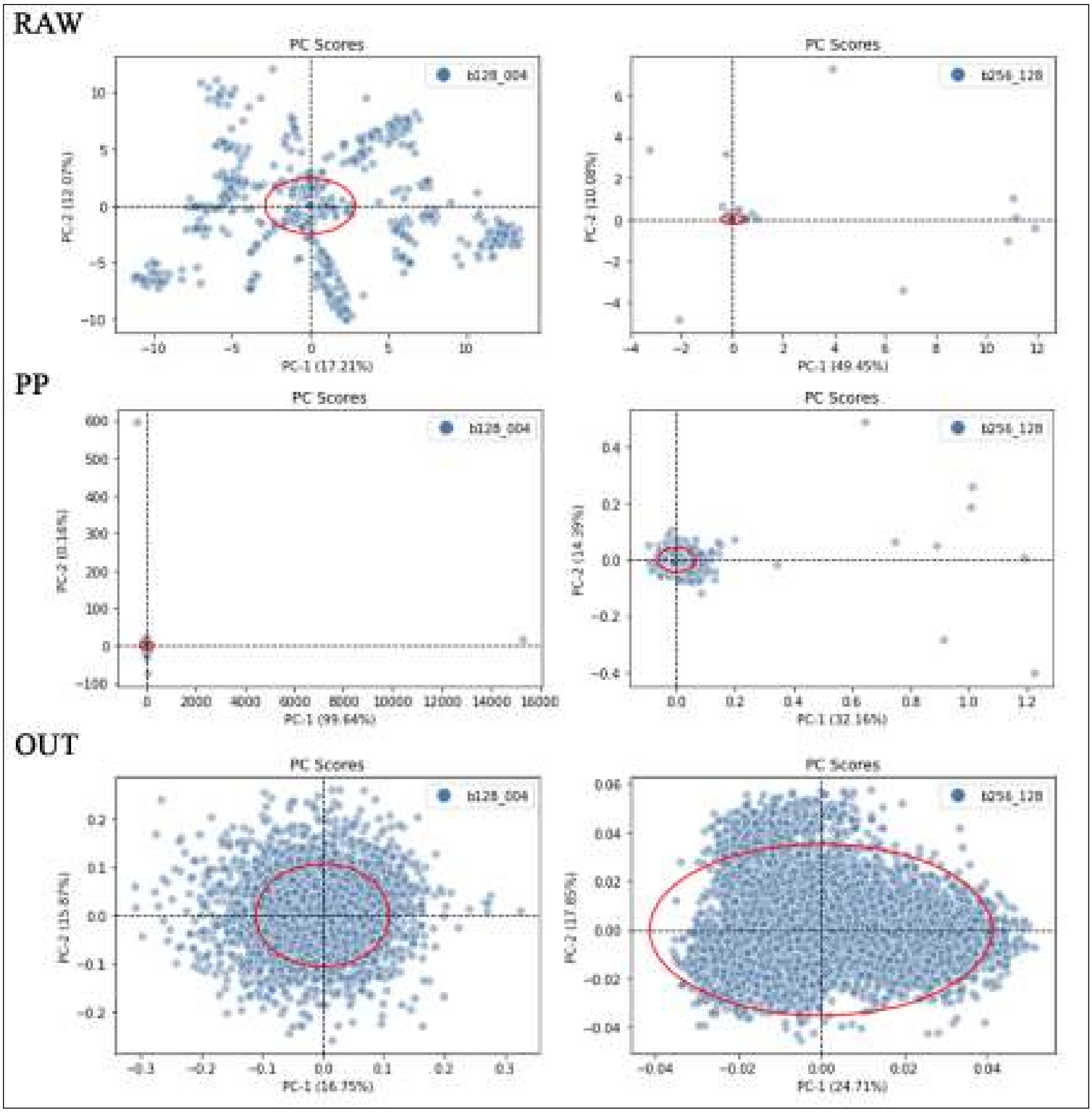


Comparing all groups (please check Appendices A to F), it is possible to verify the same changes, SD decrease and better definition of the mean spectrum shape, progressively from lowest to highest sample scan (004 to 128). The OUT plots (Appendices E and F) evidence a SD decrease not only in Amide I and II region, but also within 1350 to 900 cm^{-1}, which are mainly related to amide III (1350 to 1200 cm^{-1}), DNA and RNA (1235 to 1080 cm^{-1}), and carbohydrates (1200 to 900 cm^{-1}) content. Lower SD in this region is expected as amide III is not reported with intense variation in breast cancer [93], and as BT-474 and SK-BR-3 present similar DNA features [94]. Besides, tumors from xenograft cells exhibit a less heterogenous tissue than a regular human tissue [95], thus decreasing features variation in this analysis.

The dead region (2000 to 1800 cm^{-1}) does not contain tissue information, thus the SD of this region is only affected by environmental fluctuations [15]. These fluctuations are mainly related to water vapor content, as its absorption region is characterized from 2072 to 1205 cm^{-1} [96]. The dead region did not exhibit SD in any group, thus indicating that there was no water vapor contribution to the different SD range in the biofingerprint region (1800 to 900 cm^{-1}) among the groups. This fact is mainly related to the fast single scan acquisitions in a sequential way. A 128 scan took an average of 2 minutes to be acquired, where halving the scan almost halves the acquisition time, hence obtaining scans variations in similar ambient conditions.

PCA plots (Figure 13) corroborates with spectra comparison, where b128_004 presented sparse scores distribution in RAW and PP groups, which can also be evidenced by the scale range, and as observed in b256_128 the distribution of same groups was concentric with few sparse points. Both OUT plots show clustered scores, however b256_128 scale range is lower than b128_004. As in spectra plots, when comparing all PCA plots (Appendices G to L), it is possible to visualize the detailed changes along the increase of scan number.

Figure 13 – Principal Component Analysis (PCA) of b128_004 and b256_004 for groups RAW, PP, and OUT. Dots denote PC-scores and red line the Hotelling's T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.

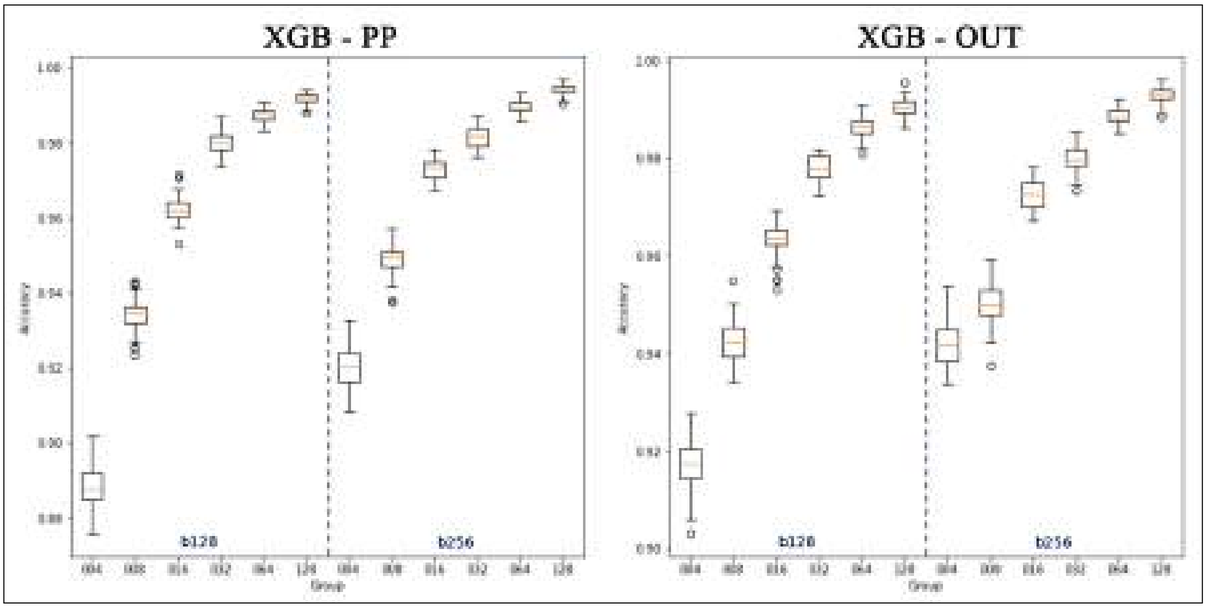


Although spectra and PCA plotting make possible to visualize the changes as sample scan number is increased from 004 to 128, it is hard to perceive major changes among highest scans, such as 064 and 128, especially for OUT groups, where the preprocessing techniques, mostly the outlier removal, approximate lower and higher scan samples. It is even harder to distinguish between b128 and b256 for same sample scans. In this way, a classification quantitative analysis can give an additional perspective of these different scans and how they perform when modeled into machine learning techniques, which is the main objective for many cancer evaluations.

5.1.2 Classification analysis

The Figure 14 presents the best model test accuracy boxplot, XGB, where the highest mean accuracy of 0.995 was achieved by OUT b256_128 group. It is possible to see an accuracy augmentation tendency in relation to higher sample scans, as well as for higher background scans, but in a less evident way. Lower sample scans also demonstrate higher accuracy SD, indicating a model generalization difficulty along the CV, possibly due to noisy data varying during the folds. All models' boxplots can be seen in Appendices M and N.

Figure 14 – Best model (XGB, Extreme Gradient Boost) test accuracy boxplot by groups. Dashed blue line splits different background scans (b128 and b256).



RAW groups were excluded from this analysis as EMSC application is necessary. Most models are negatively affected when using non-normalized data, and all of them would use paraffin bands variation to misclassify the data. The fixed threshold for outlier removal in OUT groups assisted to not result in biased data, such as removing more data from lower scan numbers than from higher ones. On average, 10% of the data were removed by the outlier algorithm.

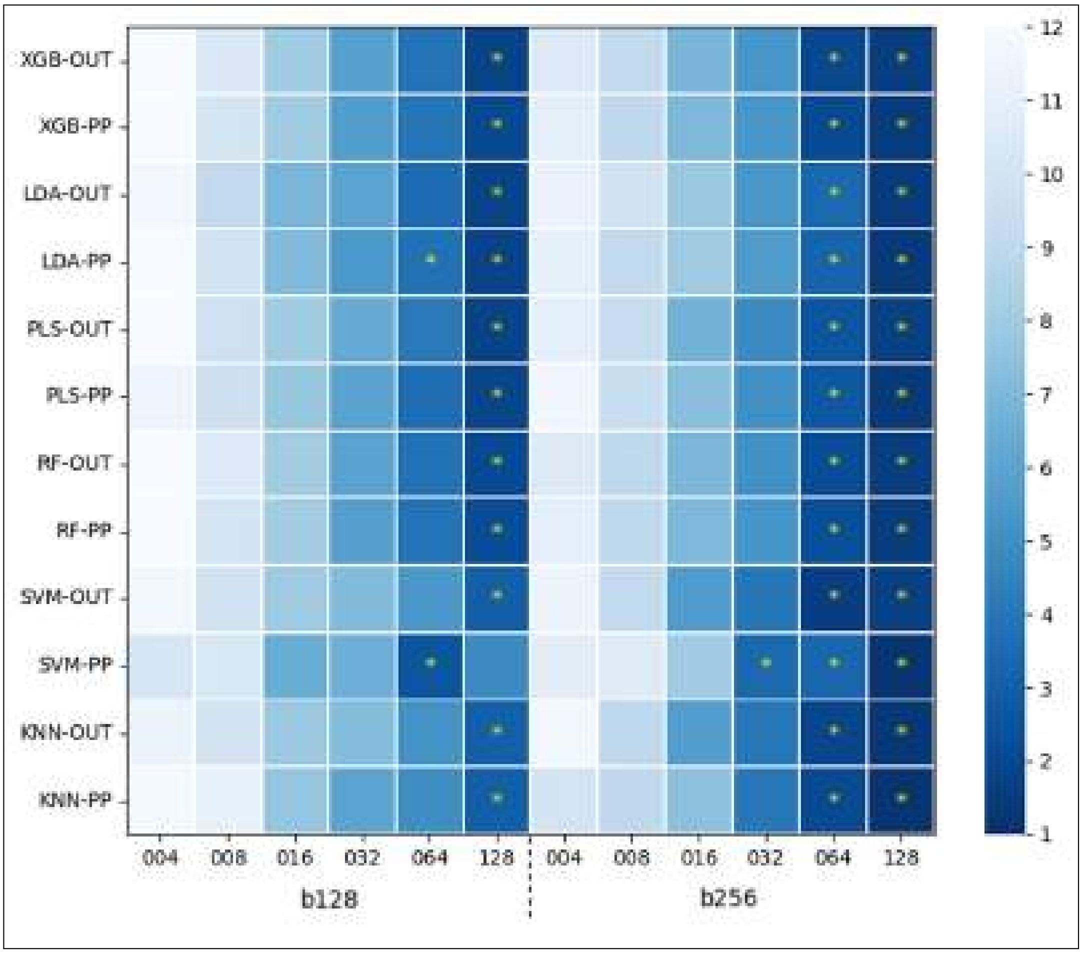
Tree-based models, XGB and RF, presented similar PP and OUT groups metrics, endorsing their robustness to outliers [97]. Even though XGB tends to be more prone of being affected by outlier than RF, since each individual tree learns from the previous, its techniques, mainly the regularization [68], enabled to perform similar with PP and OUT. The others models exhibited better accuracies for OUT groups than for PP, especially the SVM model, once outliers affect its hyperplane separation and make it a fragile model for this case [98].

SVM and KNN models presented the lowest accuracies, that may be related to their difficulties when dealing with large features number [98, 99], while tree-based models apply subsampling techniques, and LDA and PLS employ dimensionality reduction. Using feature selection and extraction approaches, such as feeding the models with n PCs instead of the whole spectra, could help to overcome this issue, although it was not tested in this study for the sake of comparison. Additionally, to standardize the comparison, default hyperparameters were chosen for all the models to avoid better results due to improved optimization in one model in comparison to the others. Still, feasible hyperparameters were used in relation to the data, making a general evaluation of the models.

PCA noise reduction [13] was also tested, using n PCs until 99% of explained variance was obtained, but no accuracy improvement was presented by the models, corroborating with the clustering analysis performed by Sacharz et al., 2020 [19], thus this result was omitted for simplicity.

The Friedman test indicated significant difference in all the models, for both PP and OUT, hence the Nemenyi test could be performed, where its results are shown in Figure 15. It is possible to see a pattern of better critical values (closer to one) when increasing the number of sample scans for both backgrounds. All the models presented no significant statistical difference between b256_128, b256_64 and b128_128, except for SVM-PP model, which had its shortcomings already discussed and exhibited difference for b128_128. Therefore, these three scan groups demonstrated similar prediction performance, especially when applying the outlier removal algorithm.

Figure 15 – Critical values heatmap for the Nemenyi test. Yellow stars denote the best tied accuracies within each model (scans where the critical distance presented no significant statistical difference).



While it is plausible in terms of accuracy to choose 128 sample scans with any of the backgrounds, b256 or b128, the b256_64 choice brings the additional of time optimization. In real clinical tasks, larger areas of the biopsies have to be evaluated, requiring mosaic acquisition to cover a region in order of centimeters. In this way, halving the sample scan, almost halves the time acquisition, improving the clinical applicability of the technique. As the background can be collected in a single scan, even if sample mosaics are performed, using b256 scans instead of b128 does not imply in a significant impact to acquisition time.

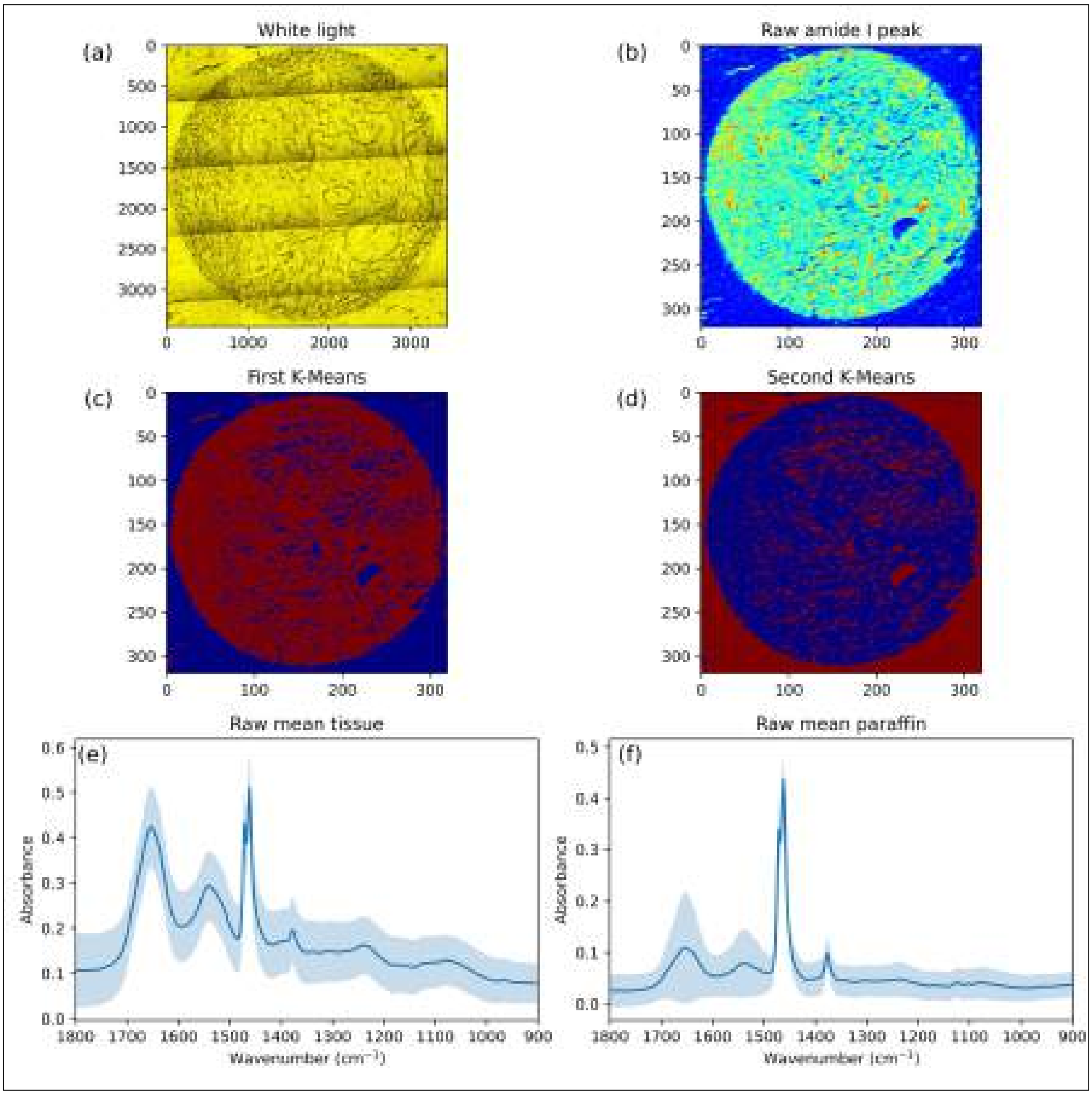
Luminal B and HER2 subtypes were chosen for this study to focus on ER/PR classification, as up to 80% of breast cancer are ER positives and up to 65% are PR positives, besides presenting the best treatment outcome when they are both positive and diagnosed in early stage [100]. Hence, a better evaluation between them may provide important prognostic and therapeutic information. In addition, this scan evaluation can be used as a basis for other studies using machine learning with cancer samples.

5.2 One-dimensional Deep Learning

5.2.1 Data preprocessing

Figure 16 shows a representative image of the clustering process. The Amide I peak image in Figure 16(b) is only for visual comparison, where it is possible to evidence its relation to the white light image in Figure 16(a). Despite that, a single peak may not a good approach to cluster the spectra, since relying on more information is more reliable.

Figure 16 – Representative figure of the clustering process. (a) White light image acquired by Cary microscope in transmission mode; (b) Amide I intensity peak plot from the raw spectra; (c) First K-Means clustering, evidencing tissue regions in red; (d) Second K-Means clustering, evidencing paraffin regions in red; (e) Raw mean spectrum (solid line) and standard deviation (light shadow) from tissue regions in (c); (f) Raw mean spectrum (solid line) and standard deviation (light shadow) from paraffin regions in (d). Spatial scale of images (a) to (d) in pixels.



Amide bands are indicators of biological tissue presence, as other regions such as paraffin and pure slide does not present these bands [15]. Therefore, the first K-Means clustering using Amide bands was able to select tissue related spectra as Cluster 1 (red), leaving paraffin and any pure slide spectra as Cluster 0 (blue), as showed in Figure 16(c). The raw spectrum in Figure 16(e), identified as Cluster 1 by the first

K-Means, exhibits the pattern of an usual biological tissue spectrum, being possible to evidence the characteristic Amide I and II bands (1700 to 1500 cm^{-1}), as well as the paraffin bands, since the tissue is embedded in paraffin, with peaks in $\sim$1462 cm^{-1} and $\sim$1373 cm^{-1} [13].

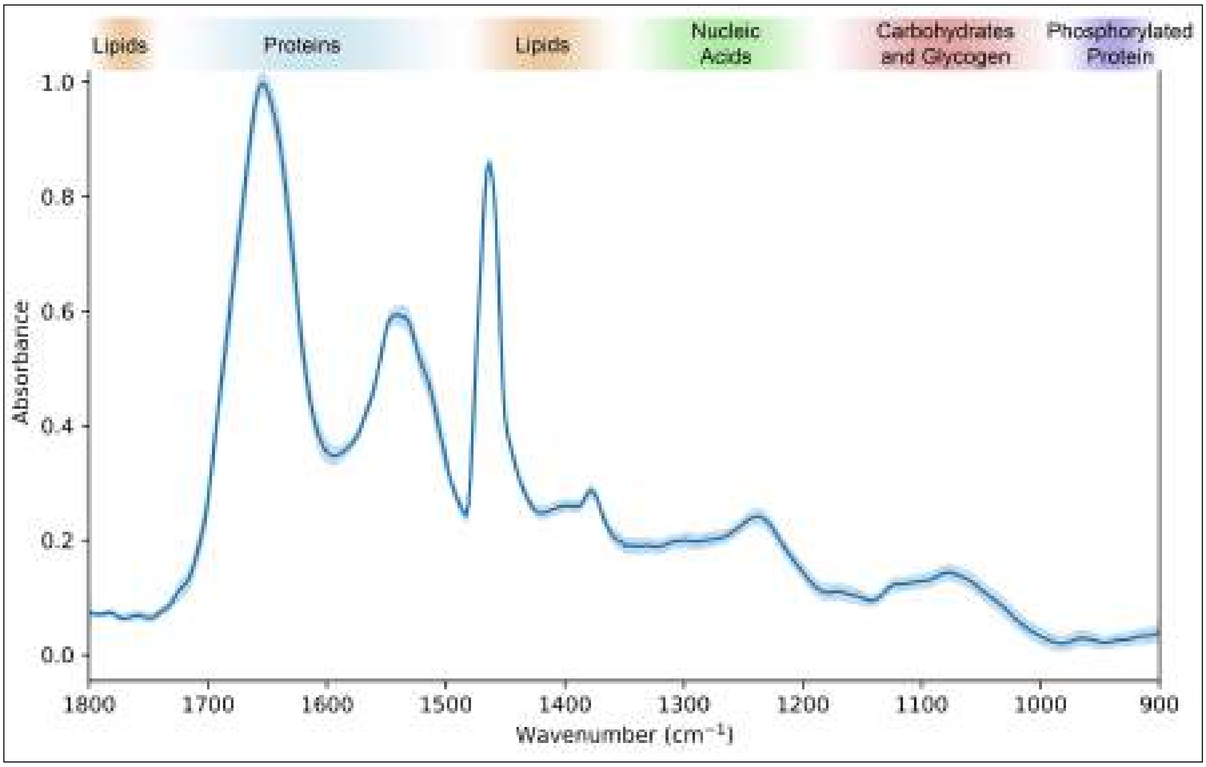
If it was guaranteed no pure slide spectra in the acquisition, only the first K-Means would be enough, however, some regions may present the absence of tissue and paraffin. In this way, a second K-Means is necessary, being able to identify paraffin spectra as Cluster 1 (red) as in Figure 16(d). As tissue spectra identified in the first clustering were set to zero, tissue and pure slide spectra should be similar with low or absent signal at the highest paraffin intensity band (1480 to 1450 cm^{-1}), being both selected as Cluster 2 (blue).

The raw spectrum in Figure 16(f), identified as Cluster 1 by the second K-Means, shows the selected paraffin. It is possible to evidence an intensity variation in the Amide I and II bands, which are not related to pure paraffin. This may be due to thinner tissue regions from the histological sectioning, hence not presenting enough Amide intensities to be clustered as tissue, but as paraffin.

Small contribution of tissue regions in the paraffin spectra, does not affect non-paraffin regions of the final tissue during the EMSC modeling, since the paraffin model masks the region from 1500 to 1350 cm^{-1}. Although this contribution my affect the variance of the masked paraffin region, the PCA model in the EMSC should take more into account the paraffin variance. Analogously to paraffin, H2O was masked from 1300 cm^{-1} onward and, therefore, only this region accounted for the EMSC model. In addition, outlier removal techniques assist to overcome these issues.

Figure 17 presents a final spectrum, using the sample from Figure 16, with the biochemical regions identification [1, 24]. The smaller standard deviation, covered by the spectrum midline in some regions, is a result of the preprocessing steps, making the tissue spectra comparable.

Figure 17 – Representative final tissue spectrum with the corresponding biochemical regions. Mean spectrum (solid line) and standard deviation (light shadow).



5.2.2 Deep learning

Table 3 shows the dev and test sets performances for each class. The type classification presented higher values than the subtype, besides closer values of dev and test, while subtype demonstrated lower test values in comparison with dev values. This indicates the type classification as the easiest prediction to be learned, as it was expected, since cancer and AT are the most different samples analyzed.

Table 3 – CaReNet-V1 dev and test performance. Results grouped by classes of each model (type and subtype). Mean values ± standard deviation.

Set	Label	Class	Accuracy	Specificity	Sensitivity
Dev	Type	CA	0.95 ± 0.02	0.92 ± 0.03	0.97 ± 0.02
	Subtype	LA	0.89 ± 0.04	0.83 ± 0.05	0.82 ± 0.08
		LB	0.87 ± 0.05	0.83 ± 0.09	0.79 ± 0.05
		HER2	0.93 ± 0.03	0.91 ± 0.02	0.92 ± 0.04
		TNBC	0.92 ± 0.02	0.87 ± 0.05	0.89 ± 0.07
Test	Type	CA	0.89 ± 0.03	0.89 ± 0.02	0.93 ± 0.03
	Subtype	LA	0.74 ± 0.08	0.45 ± 0.09	0.62 ± 0.06
		LB	0.68 ± 0.09	0.61 ± 0.10	0.42 ± 0.11
		HER2	0.83 ± 0.05	0.89 ± 0.02	0.78 ± 0.06
		TNBC	0.86 ± 0.03	0.85 ± 0.04	0.92 ± 0.03

LA and LB metrics were the lowest, however, dev results exhibit that the model was able to learn how to extract features of the spectra to predict the classes. Nevertheless, test results show specificity and sensitivity near 0.5, indicating the model had more difficulties with this class. Even though several spectra predictions were wrong, the final test patient could be right, as it is the most predicted class, where Table 4 demonstrates this performance.

Table 4 – CaReNet-V1 performance for each test patient and each of the four models from the folds in relation to the GT (Ground Truth). Light blue indicates correct predictions, while light red are wrong predictions.

Label	GT	Predicted class – Model fold:			
	Class	1	2	3	4
Type	AT	AT	AT	AT	AT
	AT	AT	CA	AT	AT
	CA	CA	CA	CA	CA
	CA	CA	CA	CA	CA
Subtype	LA	LB	LA	LA	LA
	LB	LA	LA	LB	LB
	HER2	HER2	HER2	HER2	LA
	TNBC	TNBC	TNBC	TNBC	TNBC

The model efficiently classified the type, as verified in the single spectra performance, with only one AT being classified as CA. This negative impact can be reduced when considering that it is better to have a false positive than a false negative for a cancer evaluation.

The model struggled to differentiate LA and LB samples. Luminal subtypes are the most similar, as they can even present the same receptors expression, being different only regarding the Ki67 level [9]. HER2 and TNBC were correctly classified in all cases, except by one HER2 as LA.

The concern of distinguishing LA and LB is lower than finding HER2 and TNBC cases, once they are more aggressive manifestations of the breast cancer than the luminal ones [100]. Perfect TNBC prediction is especially important as this subtype exhibits the worst prognosis due to lack of drug targets and high risk of brain metastasis [101]. Therefore, these findings can be considered an important step towards an automated laboratory screening technique, and may help to prioritize patients' analysis by the health professionals.

Grad-CAM results are depicted in Figures 18 and 19. This is a localization approach to identify important regions of the image that influenced the classification decision [85]. Once CaReNet-V1 is a 1D model, the localization is related to the wavenumber instead of the spatial information. Thus, a 1D Grad-CAM can be applied as a feature importance tool.

Figure 18 – Gradient-weighted Class Activation Mapping (Grad-CAM) of the Type classification model. Darker blue areas identify wavenumber regions that contributed the most for the Cancer (CA) activation.

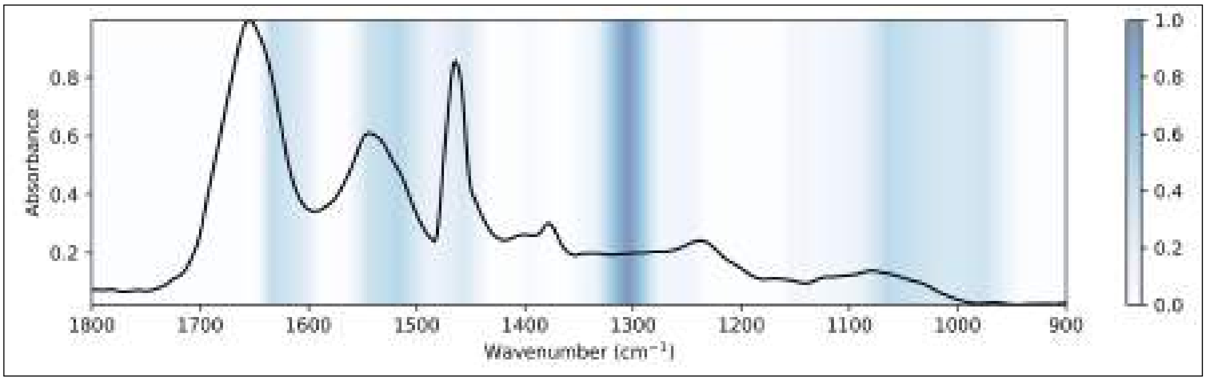
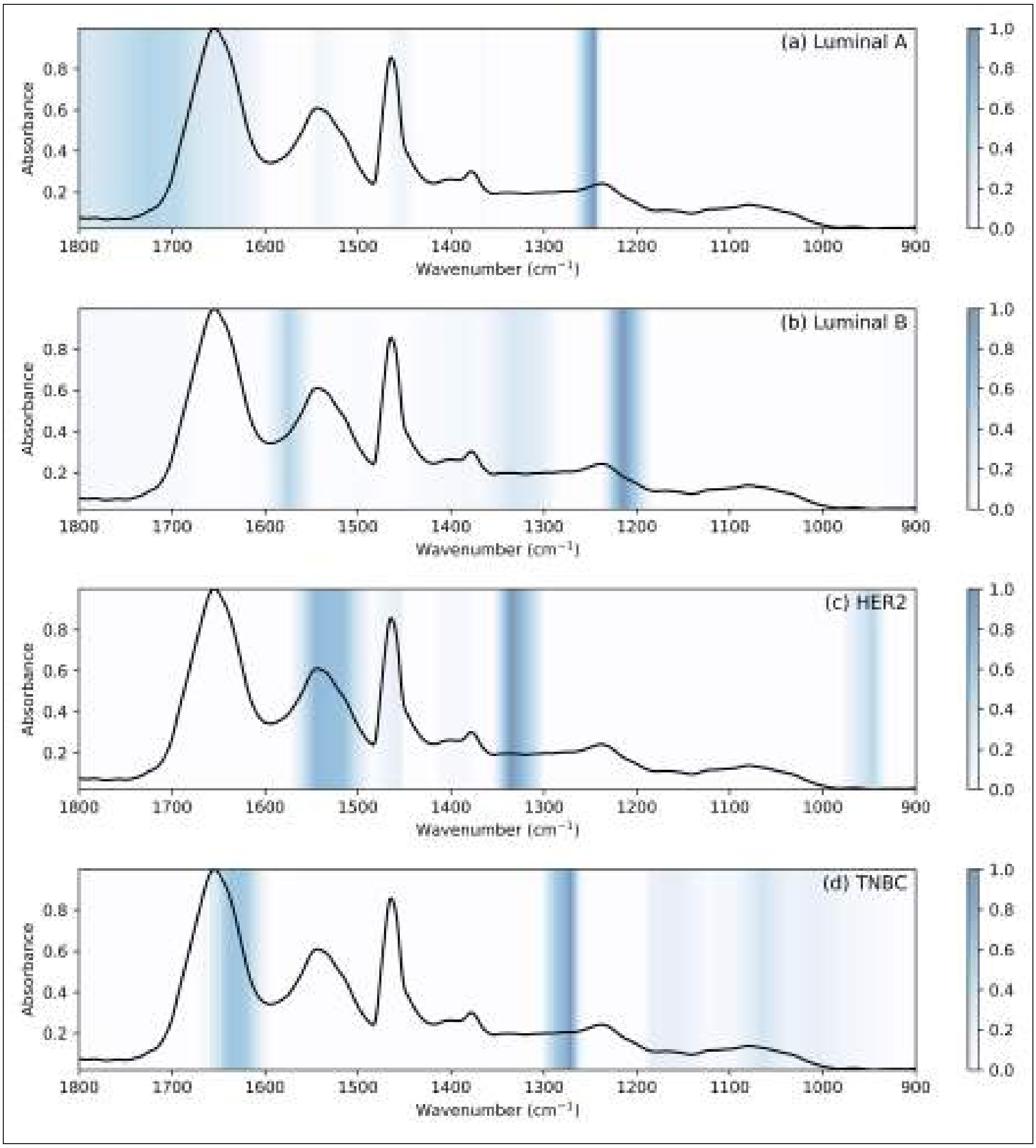


Figure 19 – Gradient-weighted Class Activation Mapping (Grad-CAM) of the Subtype classification model. Darker blue areas identify wavenumber regions that contributed the most for the activation of (a) Luminal A (LA), (b) Luminal b (LB), (c) HER2, and (d) Triple-Negative Breast Cancer (TNBC).



Type classification was performed by one output (one neuron with sigmoid activation), thus only one Grad-CAM can show the activation or not of the CA class. On the other hand, subtype with four output probabilities leads to four Grad-CAM. These results can be related to the molecular footprints of vibrational spectroscopy [1, 2], as per the following, to understand the composition impacts to the model.

Cancer type activation was mainly related to four regions. The first one in 1640-1600 cm^{-1} is mainly assigned to adenine vibrations in DNA and one part of Amide I. Second region of 1550-1510 cm^{-1} is strongly related to Amide II band, where there is C$-$N stretching and C$-$N$-$H bending vibrations weakly coupled to the C$=$O stretching mode, besides the nucleobases C$=$N. Studies have reported amide I and II to differentiate cancerous and healthy tissue [102]. The third, in 1320-1280 cm^{-1} band, there is the leading contribution of Amide III, essentially related to C$-$N stretching, N$-$H in plane bending and CH_2 wagging vibrations; and collagen associated vibrations. The most intense activation contribution in 1310-1300 cm^{-1} is totally related to Amide III. At last, the 1080-980 cm^{-1} region is pertinent to glycogen (1050-1020 cm^{-1}) and symmetric PO_2^- stretching in RNA and DNA (1100-1040). Glucose expressions have been linked to cancer cells during the neoplastic process, while DNA and RNA oscillation is directly associated with cancer diagnosis [103].

Regions that contributed the most for LA classification (Figure 19 (a)) are around 1750-1680 and 1260-1240 cm^{-1}. The first one is mostly related to C$=$O stretch from bases of nucleic acids in 1717-1681 cm^{-1}, and from lipids and fatty acids in 1750-1725 cm^{-1}. It is stated that tumor progression and cancer cell survival favored by fatty acids overproduction [104]. In the second region, there is the contribution of Amide III, phosphate, and collagen I and IV. Amide III region is observed in 1340-1240 cm^{-1} due to C$-$N stretching of proteins, indicating mainly α-helix conformation. In 1245-1240 cm^{-1} there are several PO_2^- asymmetric stretching originated from the phosphodiester groups of nucleic acids, which suggest nucleic acids increase in malignant tissues [1]. Presence of PO_2^- stretching vibrations may be associated to DNA damage caused by reactive oxygen species [103].

LB classification contribution (Figure 19 (b)) from 1590-1570 cm^{-1} relates to C$=$N adenine and to phenyl ring C$-$C stretch. The accelerated metabolism of DNA/RNA leads to oscillatory deformations of adenine [103]. The 1225-1200 cm^{-1} region is associated with PO_2^-: asymetric phosphate I vibrations at 1217-1207 cm^{-1}; asymmetric phosphate vibrations of nucleic acids when highly hydrogen-bonded; and phosphate II asymmetric vibration in B-form DNA.

Three regions appear with high impact for HER 2 classification (Figure 19 (c)): 1550-1510, 1350-1300, and 980-950 cm^{-1}. The first region is all covered by the amide

II band, with N$-$H bending vibration coupled to C$-$N stretching. Guanine C$=$N is also present in 1534-1526 cm^{-1}. In 1317-1307 cm^{-1} there are amide III band components of proteins, and in 1340-1317 cm^{-1} collagen related assignments. Last region is mainly attributed to deoxyribose C$-$O and symmetric stretching mode of dianionic phosphate monoesters in phosphorylated proteins and nucleic acids. Increased intensities in this region are correlated to cells in malignant tissue [2].

TNBC classification (Figure 19 (d)) showed the influence in 1660-1610 cm^{-1}, which covers a large area of the amide I band, besides adenine vibration in DNA in 1609-1601 cm^{-1}. In 1300-1270 cm^{-1}, there are mainly amide III and collagen vibrational modes, and also a CH_2 wagging vibration of phospholipids acyl chains. Phospholipids expression is used to distinguish subtypes membrane remodeling, where TNBC and HER2 demonstrates the greater difference and potentially reflects their greater ability to grow [104]. The presented vibrational modes of the collagen are distinctly stronger for breast carcinoma [1]. A slight contribution in 1070-1050 cm^{-1} is mainly associated to phosphate and oligosaccharides, such as P$-$O$-$C antisymmetric stretching and C$-$OH stretching.

In this way, the 1D deep learning prediction coupled with a Grad-CAM analysis of micro-FTIR images can assist to understand the breast cancer composition that distinguish the cancer and subtypes in a label-free approach. Although wavenumber shifts may occur according to the sample, the band (region) analysis offers more reliable evaluation. This analysis may be employed not only for diagnosis purposes, but also for treatment efficacy assessment and development of new therapeutic methods.

Models created with a 1D approach have the advantage of consuming less memory, hence being easier to be trained. CaReNet-V1 architecture, using 467x1 input shape, resulted in 277,236 parameters. In contrast, albeit one spectrum prediction is fast, as there are several spectra in one mosaic, it takes longer to predict one whole patient. One-dimensional adaptations of traditional CNN, such as VGG, and models based on the 1D spectroscopy analysis reported in [79–82] were tested, yet the learning process was only possible when using residuals approaches as in ResNets, where CaReNet-V1 model presented best results.

Using single spectra increases the dataset size, where one mosaic of 320x320 turns into 102.400 single spectra, facilitating the training with even a couple of mosaics.

Even so, training with several patients is recommended to enable a generalist model. In addition, the single spectra approach raises the issue that not all spectra may be representative of the class, as a breast cancer biopsy usually presents a very heterogeneous tissue. This can directly decrease the 1D model performance, since all spectra is considered as being from the same class, while still correctly predicting the final patient classification due to the incorporate voting system.

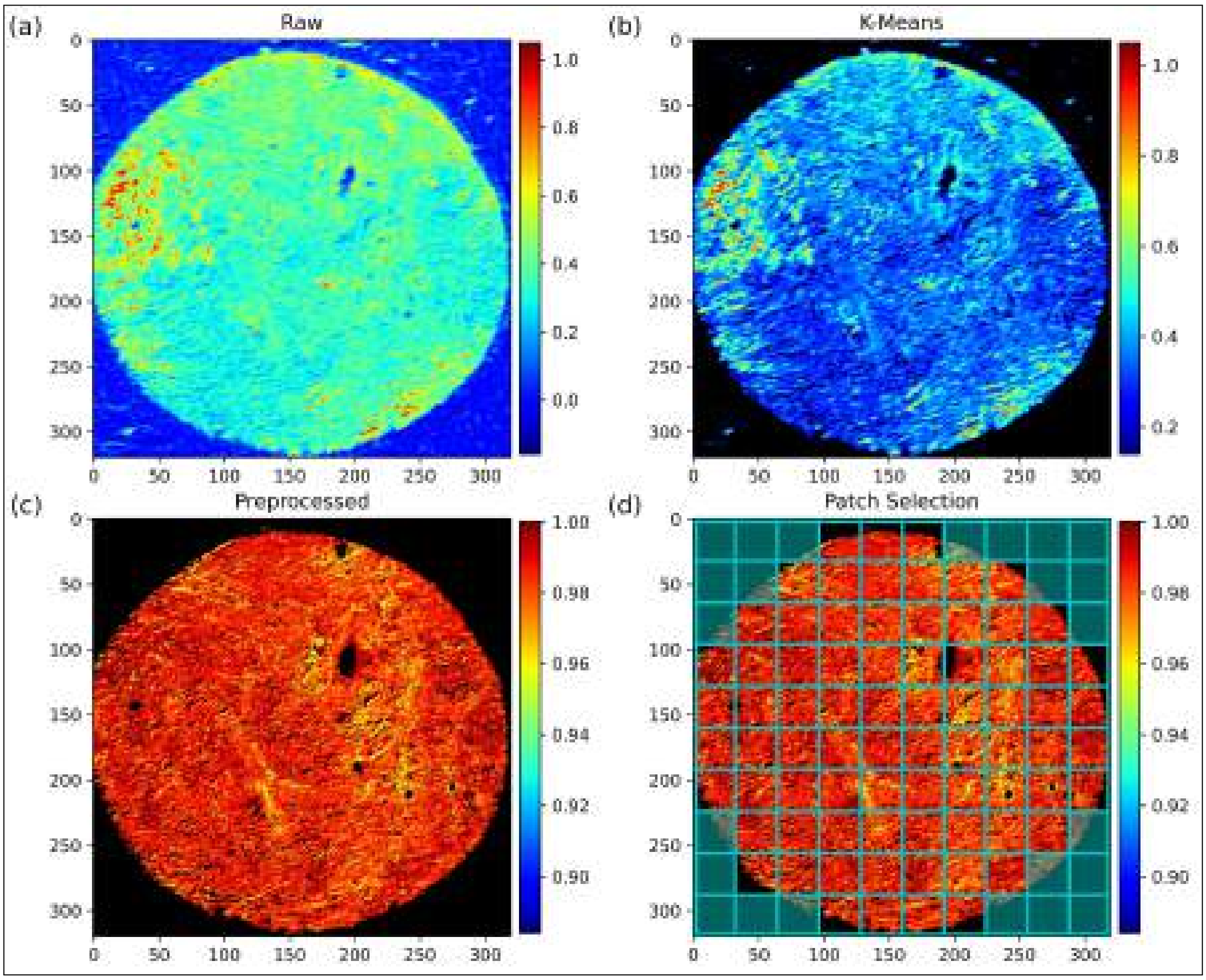
The second derivative was tested as input instead of the regular absorbance, and also a double channel input with absorbance and its second derivative, but in all cases using only absorbance resulted in better outcomes. Hence, second derivative results were omitted from this study for simplicity. Analogously, a cosine decay restarts schedule [92] was tested, but no improvement was observed. While it enables longer training, such as 500 epochs, to increase the probability of the optimizer to reach the global minimum loss, the reduce learning rate on plateau callback granted similar performance with only 50 epochs.

Next studies should consider adding more breast cancer biopsies. Increasing the dataset size may help not only by giving more training and dev examples, but mainly by augmenting the test set quantity, hence aiding to achieve a better real world performance evaluation of the breast cancer subtypes. In addition, individual receptor expressions should be assessed along biopsies augmentation, once some tests with the current dataset have demonstrated poor performance of CaReNet-V1 for these predictions.

5.3 Two-dimensional Deep Learning

Figure 20 depicts a representative image of the preprocessing process. The amide I peak demonstrates the impact of each process, as amide bands are indicators of biological tissue [15]. Paraffin blue border regions in Figure 20 (a) were clustered and zeroed by the K-means process in Figure 20 (b). Some residual tissue regions still appear after the clustering, evidenced by the blue chunks in the black zeroed paraffin area. This may be due to thin tissue residues, where the amide band intensities are present, but it is not thick enough to have a similar spectrum from the rest, thus being identified as outliers and eliminated after the remaining preprocessing steps in Figure 20 (c). Besides the clean black border, it is possible to visualize the more defined "holes" inside the core, where borderline outliers were excluded.

Figure 20 – Representative figure of the preprocessing process using the Amide I intensity peak image for better visualization. (a) Raw spectra; (b) Tissue raw spectra after K-means clustering. Paraffin as black (zeroed); (c) Preprocessed spectra; (d) Patches selection. Painted cyan squares are excluded patches. Spatial scale of images in pixels.



EMSC and normalization improved the scale presentation in Figure 20 (c), where values close to the maximum (1) are due to the fact that the amide I peak is usually the one with the highest intensity of the spectrum. The plot was normalized itself, hence improving the variation visualization. Patches with more than 50% of the pixels as zeroes were automatically removed, which were mainly the paraffin border patches, as shown in Figure 20 (d).

Test sets classifications performance are shown in Table 5. Type model presented the best metrics, as expected due to the disparity between malignant and benign tissues. Even though they were both above 0.9, the higher sensitivity than specificity, of 0.95 and 0.91, respectively, is preferred for cancer diagnosis, once the sensitivity is the probability of correct identifying a truly present cancer [105]. Therefore, it is better not to predict a false negative, while it is acceptable a certain level of false positives.

Table 5 – CaReNet-V2 performance for test patches classification. Results grouped by classes of each model. Mean values ± standard deviation.

Label	Class	Accuracy	Specificity	Sensitivity
Type	CA	0.91 ± 0.02	0.91 ± 0.03	0.95 ± 0.02
Subtype	LA	0.85 ± 0.05	0.80 ± 0.06	0.89 ± 0.04
	LB	0.84 ± 0.04	0.81 ± 0.05	0.77 ± 0.05
	HER2	0.90 ± 0.03	0.87 ± 0.04	0.91 ± 0.05
	TNBC	0.92 ± 0.02	0.90 ± 0.03	0.89 ± 0.03
ER	−	0.65 ± 0.06	0.62 ± 0.07	0.69 ± 0.10
	+	0.57 ± 0.12	0.62 ± 0.08	0.53 ± 0.09
	++	0.59 ± 0.09	0.66 ± 0.10	0.53 ± 0.09
	+++	0.66 ± 0.07	0.61 ± 0.05	0.71 ± 0.06
PR	−	0.63 ± 0.05	0.58 ± 0.08	0.70 ± 0.08
	+	0.54 ± 0.09	0.60 ± 0.08	0.51 ± 0.11
	++	0.51 ± 0.07	0.56 ± 0.10	0.50 ± 0.12
	+++	0.62 ± 0.05	0.68 ± 0.10	0.68 ± 0.07
HER2	3+	0.82 ± 0.04	0.84 ± 0.05	0.77 ± 0.06

Subtypes demonstrated metrics above 0.77, indicating a nice performance where the models learned how to extract features from the samples for this classification. HER2 and TNBC were better classified than LA and LB, with metrics around 0.9. This may be due the similarity between LA and LB samples, since their expression levels of receptors may be the same, demanding the Ki67 level to distinguish them [9].

ER and PR labels presented the lowest metrics, especially regarding borderline classes (+ and ++). The classification of the expression levels of each receptor is a finer prediction than the subtypes, since it is necessary to differentiate the same characteristics, at different levels, instead of a macro grouping of the characteristics as in the subtypes. Furthermore, the built dataset contains considerably fewer borderline classes samples (2 or 3) in comparison to others (10 or more samples), making generalization a difficult task, even with loss punishment by class weights.

HER2 models achieved metrics close to those of subtype, higher than ER and PR ones. However, a binary classification is easier than multi-class, especially in this case where the borderline classes (1+ and 2+) were not considered. Nevertheless, the models showed the ability of learning the HER2 evaluation.

Ki67 regression test performance is exhibited in Table 6. A mean difference between GT and predicted of 2.3% is a good overall prediction for the current dataset, 5 and 10% will remain as low level, and 20 and 30% as high level, since the cutoff is usually given by 15% [106]. MSE and RMSE measure the variance and standard deviation of the residuals, respectively, amplifying high errors more than lower ones. Therefore, the fact that they did not scale too far from the MAE indicates low impact from outliers.

Table 6 – CaReNet-V2 performance for test patches Ki67 regression. MAE (Mean Absolute Error), MSE (Mean Squared Error) and RMSE (Root-Mean-Square Error) according to the predictions of the models on the min-max fraction scale (0-1 range) and rescaled to percentage (5-30% range). Mean values ± standard deviation.

Scale	**MAE**	**MSE**	**RMSE**
Min-Max	0.094 ± 0.015	0.021 ± 0.003	0.145 ± 0.009
Rescaled (%)	2.3 ± 0.4	13.1 ± 1.6	3.6 ± 0.2

Missing some patches predictions does not imply in a wrong final test patient prediction, as the voting system may overcome these mistakes. Table 7 exhibits the final test patient classification results after the voting system.

Table 7 – CaReNet-V2 classification performance for each test patient and each of the four models from the folds. Light blue corresponds to correct predictions in comparison to the GT (Ground Truth), whilst light red are wrong predictions.

Label	**GT**	**Predicted class – Model fold:**			
	Class	**1**	**2**	**3**	**4**
Type	AT	AT	AT	AT	AT
	AT	AT	AT	AT	AT
	CA	CA	CA	CA	CA
	CA	CA	CA	CA	CA
Subtype	LA	LA	LA	LA	LA
	LB	LB	LA	LA	LB
	HER2	HER2	HER2	HER2	HER2
	TNBC	TNBC	TNBC	TNBC	TNBC
ER	−	−	−	−	−
	+	−	+	+	+
	++	++	++	++	+++
	+++	+++	+++	+++	+++
PR	−	−	−	−	−
	+	+	+	−	+
	++	−	++	++	+++
	+++	+++	+++	+++	+++
HER2	0	0	0	0	0
	0	0	0	0	0
	3+	3+	0	3+	3+
	3+	3+	3+	3+	0

The models from the four folds were able to correctly classify the Type (Cancer vs AT) for all test patients, withstanding the highest test patches metrics among all predictions. Type classification was expected to show the best performance, since it is a binary classification of the most different tissues: malignant and non-malignant. Still, a perfect cancer identification is an excellent characteristic for a biopsy screening tool.

Subtypes were classified correctly, except for two LB misclassified as LA. This also expresses the similarity between LA and LB, corroborating the findings in single patches metrics. It is important to discern luminal subtypes from HER2 and TNBC due to their higher incidence and better prognosis. LA represents 30 to 40% of breast cancers and LB 20 to 30%, while HER2 and TNBC ranges from 12 to 20% each [100, 107].

Luminal subtypes demonstrates better treatment outcome and survival rate [101, 108, 109], whilst TNBC leads the worst, mainly because it develops resistance to its often treatment, chemotherapy, and has a high risk of evolving brain metastasis [101]. Luminal is also less recurrent, where LA evolves slowly within time and LB presents a peak incidence of recurrence in the first 5 years. On the other hand, HER2 and TNBC manifest a peak of recurrence in one or two year [100, 110]. Hence, considering a biopsy screening technique, it is less critical to misclassify between LA and LA than a wrong HER2 or TNBC prediction.

ER and PR levels were satisfactory predicted, showing difficulties with borderline levels ($+$ and $++$). This may be related to the lack of borderline samples, not providing enough examples for the models to learn their characteristics. Loss punishment by classes weights and the acquired knowledge from $-$ and $+++$ samples may have assisted in the learning of these classes' prediction, however more samples are still necessary. If only $-$ and $+++$ samples were considered, the models would be classifying mainly LA/LB vs HER2/TNBC, since there few examples of LA and LB with negative ER or PR. Performance could be improved if ER and PR expressions were grouped as negative ($-$ and $+$) or positive ($++$ and $+++$) classification [111], once binary classifications are usually easier to be modeled. Despite this information being the most critical for the prognosis, these receptors play a substantial role in the assessment, and further details should be considered whenever possible.

Positive hormone receptor expression status is a favorable prognostic factor and a predictor of response to endocrine therapy [100]. Patients with both ER and PR positivity usually experience better outcomes than single positivity, especially single PR one, once believed a rare phenomenon, now reported as exhibiting a behavior as aggressive as HER2 and TNBC [112, 113]. ER and PR positive and HER negative is the most prevalent with 60 to 70% of all breast cancers, where antiestrogen target therapy is associated with improvement in overall survival in both early and advanced phases, in addition to good responses to adjuvant chemotherapy [114]. Thereby, the proper analysis of these receptors can change the whole treatment strategy for a wide range of patients.

Instead of using four levels, receptors can be analyzed by the expression percentage [115, 116]. It would be of great value to predict this percentage, using a regression process analogously to the Ki67 one, accomplished in this study. Nevertheless, this would require a whole new and larger dataset, with representative expressions from all the possible range assessed by gold standard techniques. This kind of samples should be evaluated in future studies.

HER2 levels were properly predicted, except for two 3+ classified as 0. Associated with the single patches performance similar to the subtypes, which have four classes, this may indicate HER2 as a harder prediction, since it was a binary classification with larger samples number per class. Indeed, preliminary tests were performed using the 1+ and 2+ samples, in which the models were not able to learn this classification using only train/dev sets, as there were not enough samples for a test set, thus being omitted from this study. Even so, the ability to predict 0 or 3+ indicates that the model may learn the four levels classification if more samples examples is available.

Adequate HER2 assessment directly affects the prognosis. Tumors related to HER2-overexpression are regarded as aggressive neoplasms, associated with chemoresistance and poor survival rates. The most promising treatments are the use of tyrosine kinase inhibitors and immunotherapy with monoclonal antibodies [117]. These target therapies are usually employed in an adjuvant setting, and although had improved the prognosis, the high number of deaths from HER-positive breast cancers and researches for newer therapies persists [118, 119].

Table 8 displays Ki67 test patient predictions after the voting system. This is the only regression approach, hence not presenting the highlight for correct or not. Even though, the GT and predictions can be compared in terms of absolute error. Lower Ki67 levels presented better predictions than higher, with four-folds MAE of 1.5% for the lowest GT expression (5%), gradually increasing to 3.7% for the highest (30%).

Table 8 – CaReNet-V2 Ki67 regression performance rescaled to percentage for each test patient and each of the four models from the folds. MAE (Mean Absolute Error) calculated from the four models' predictions with respect to the GT (Ground Truth).

Label	GT	Predicted % – Model fold:				MAE
	%	1	2	3	4	%
Ki67	5	4.2	3.0	6.8	6.2	1.5
	10	7.1	8.2	12.2	11.5	2.1
	20	18.3	17.2	22.6	17.1	2.5
	30	25.5	25.9	27.5	26.5	3.7

Using the MSE as the loss function helps dealing with outliers predictions. Decreasing outliers is important once the Ki67 is macro-divided in low or high expression if it is below or above 15%, respectively. Thus, near misses can still be in the same category, although outliers will probably lead to an incorrect one. Considering this cutoff point, all test patients were predicted within the correct low/high range, even though there was no sample with a borderline GT of 15% to better evaluate this occurrence. MSE usually does not deal well with imbalanced datasets [120], however the usage of class weights assisted to overcome this issue, as a large distribution difference was present on the dataset

Ki67 modeling involved four target values from two macro-levels, hence it could be more appropriated to deal with it as a classification approach. A binary low/high classification is useful, but it does not provide as much information as all the percentages, especially when dealing with borderline expressions. Even a categorical multi-class approach does not represent all possible real-life Ki67 levels. Therefore, the regression method was chosen to verify how the model would perform with an approach that could account all Ki67 expression levels, which can range from barely 0 to almost 100% [121]. Even so, it is required a much larger dataset with several samples in this range to properly evaluate this process.

Besides the usage to distinguish LA from HER2 negative LB, Ki67 expression is important to evaluate treatment responsiveness, endocrine or chemotherapy resistance, residual risk, and a dynamic biomarker during therapy [106]. High Ki67 is associated with poor survival, however the cutoff may vary between studies. It is reported variations on the cutoff of 10 to 20% [122,123]. Other assessments may also be indicated, such as relating a cutoff of 40% to a higher risk of recurrence and death for resected TNBC [124]. Therefore, a complete Ki67 regression is an advantageous analysis in comparison to binary or multi-class classifications.

Figure 21 depicts the Grad-CAM analysis. It is possible to visualize a well-distributed high intensity heatmap all over the tissue region, indicating the spatial contribution of a large area. Zeroed black pixels have a low classification contribution, as their weights are multiplied by zero. The spatial path of the model is responsible for spatial feature extraction by evaluating the 3x3 kernel, i.e., spatial evaluations of 9 spectra per step. Spectral path convolutions only assess individual pixels, where the downsampling is executed by pooling layers. Hence, the Grad-CAM of this path does not provide useful information, presenting meaningless heatmaps.

Figure 21 – Representative image of the Gradient-weighted Class Activation Mapping (Grad-CAM) for Type classification.

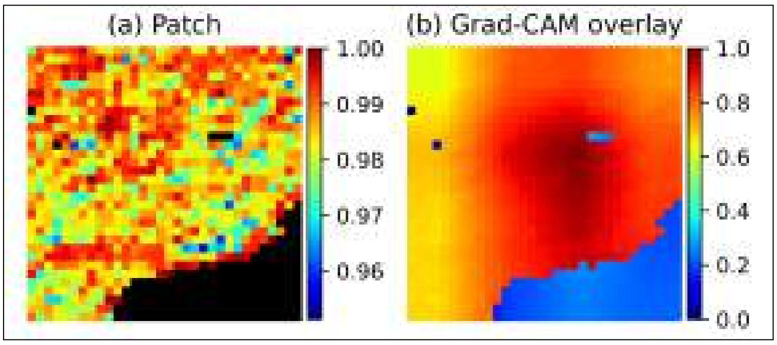


Table 9 lists the most influential channels for each label prediction. These channels, or wavenumber bands, are assigned to biochemical informations [1, 2].

Table 9 – Channel importance for each label. Top three wavenumber bands and their main assignments [1, 2].

Label	#1		#2		#3	
	Band (cm^{-1})	**Assignments**	**Band (cm^{-1})**	**Assignments**	**Band (cm^{-1})**	**Assignments**
Type	1658-1650	Amide I	1240-1236	Phosphodiester PO_2^-	1553-1539	Amide II
Subtype	1597-1588	Adenine Phenyl ring	1753-1742	Lipids Fatty Acids	1073-1065	Nucleic Acids Phosphate
ER	1055-1050	RNA DNA	1612-1603	Adenine	1287-1281	Collagen
PR	1206-1192	Collagen	1047-1036	Carbohydrates RNA	1647-1641	Amide I
HER2	972-968	Nucleic acids	1506-1493	Amide II Phenyl rings	1026-1017	Glycogen
Ki67	1088-1085	Phosphate PO_2^-	1242-1234	Phosphodiester PO_2^-	1649-1639	Amide I

Amide I and II, listed in type, PR, HER2 and Ki67, have been used to differentiate cancerous and normal tissues [102], specially the 1655 cm^{-1} in type importance range, which is related to α-helix amide I and is reported to have its intensity decreased for malignant tissues [1]. The 1240 cm^{-1} in type and Ki67 is regarded to asymmetric non-hydrogen-bonded phosphate stretching modes from phosphodiester vibration, which suggest an increase in the nucleic acids in cancerous tissue.

Adenine bands, as in subtype and ER, are reported to be higher in patients with cancer. This is due to the higher accelerated metabolism of the cells, which entails in oscillatory deformations of the C$-$H peak of adenine [103]. Overproduction of fatty acids, listed in subtype importance, facilitates the tumor evolution and the survival of cancerous cells [103].

Differences in DNA and RNA vibration frequency, as in the band shown in #1 ER importance, is an important evaluation to discriminate between normal and cancer spectra [103]. PR importance presented the collagen (amide III) influence, where the 1204 cm^{-1} is associated with higher intensities for breast carcinoma tissue [1]. Still in PR, it is noted an absence of carbohydrates peaks in breast cancer spectra, which may be related to the higher glucose metabolism in cancer cells [103].

Nucleic acids, as in HER2 #1 importance band, are augmented in cancer tissue due to their increase in the relative content [2]. The 1026 to 1017 cm^{-1} HER2 #2 band, may be linked to the higher metabolism during the neoplastic process [104]. The 1086 cm^{-1} listed in Ki67 band #1 due to symmetric stretching modes is reported to be increased in nucleic acids in malignant tissues [1].

Wavenumber shifts of the intensities may take place, where the band-region evaluation provides more reliable information. The channel importance analysis can help to understand the impact of the biochemical composition of the breast cancer to the predictions in a label-free approach. However, a fully 1D model may provide more information and facilitate interpretation, once the whole model extracts intensity features from single spectra, totally related to wavenumber variations and without any influence from neighboring spectra. In contrast, for 2D and 3D models it is necessary the evaluation of each channel individually, which becomes a difficult task since the models accounts the spatial relation of the spectra and their extracted features are mixed. Thus, only the first conv layer was considered for the current channel importance analysis, which is not directly related to the final model prediction, once the features pass through several other conv and dense layers.

The relative GAP importance analysis revealed that spectral path features accounted for 58 to 79% of the total contribution, while the rest was related to spatial features. This indicates the greater importance from extracting individual spectral features, with deeper assessment of the local biochemical information. Spatial path also calculates spectral features, although it takes more into account the spatial relationship of the spectra. Even so, it also plays an important role by analyzing heterogeneous samples such as breast cancer tissue.

Models created with a 2D approach demands more computing memory than 1D to be trained. CaReNet-V2 architecture with a 32x32x467 image patch as input and four neurons as output classification resulted in 15,914,660 parameters. On the other hand, predicting a patch implies 1024 spectra processed together, speeding up the overall prediction time for a large mosaic.

Other models reported on the literature were tested, such as the well-established VGG [76], and the state of the art models EfficientNet [125] and ConvNeXt [126]. Still, they were not able to learn how to extract features and properly classify the samples. Adding residuals convolutions increased the training time and did not improve the performance.

A 32x32x467 image can be compared to a deep hidden layer of a standard RGB (Red, Gree, Blue) classification model, where after several downsampling layers and progressive augmentation of the filters number to better extract the information, features maps images may approximate the size of the hyperspectral patches in this study [76, 127]. In this way, for both spectral and spatial paths, the first convolutional layer with 128 filters assists to extract the most important information from the 467 wavenumber channels, decreasing the channels size at first, and then the next layers work on understanding these characteristics.

Patches approach was necessary to not overload the GPU memory, although it was also a benefit due to increasing the dataset size by 100x. Combined with the fact that hyperspectral imaging provides more information than RGB, it was possible to train the models with only 30 breast cancer patient samples. In addition, it enables the prediction of any mosaic size as long as it is possible to build 32x32 patches. Patches augmentation during the data generator aids to achieve a better generalization, as the model doesn't get used to predicting the same image position; while the voting system assists for a better prediction, since it is not dependent on a single patch evaluation, thus final prediction may overcome possible mistakes.

All spectral path layers containing 1x1 kernels causes each spectrum features to be extracted individually, with no influences of neighbors' spectra. The max pooling layers select the most discriminating features, without reprocessing them with convolution kernels. It would be possible to keep extracting features with 1x1 kernels and apply a single Global Max Pooling (GMP) at the end of the spectral path, instead of the GAP, but that would make the model more complex and remove the most discriminating feature emphasizing. GAP before dense layers is preferred to map the complete extent of extracted features instead of selecting the most discriminative ones as in GMP, besides acting as a structural regularizer and helping to prevent overfitting [128].

Loss punishment based on class weights supports the development of imbalanced datasets without using oversampling or downsampling approaches. Nevertheless, the training process with balanced classes, avoiding class weights or other balancing methods, usually present better results [129]. Thereupon, a larger and balanced dataset may considerably improve CaReNet-V2 performance. Increasing dataset size would also provide more test samples, aiding to achieve a better real world performance evaluation.

Other tested approaches did not exhibit satisfactory results and were omitted from this study for simplicity. Even so, it is worth describing them to guide future research: second derivative led to dummy classifiers; PCA instead of patches, downsampling the input to 320x320x10, did not supply enough information for the feature extraction; reduce learning rate on plateau callback [84] was not able to leave local minima; one-hot rather than ordinal encoding for ER and PR levels displayed lower metrics, possibly due to the models not learning the ranking relationship between the levels.

A microarray sample with representative subtypes distribution was chosen for this study as subtypes stratifies patients for treatment, guiding the systemic therapy in preoperative, postoperative or both scenarios [107, 130]. Yet, a complete pathology report screening tool should address other evaluation methods, such as breast cancer histology (50 to 75% of patients present invasive ductal carcinoma while 5 to 15% shows invasive lobular carcinoma) [107], TNM (Tumor, Node, Metastasis) staging [131], and grade system [132], once these methods may also assist to a better understanding of the breast cancer and should be considered in future studies with larger samples number.

6 CONCLUSIONS

The analysis of co-added scans number for FTIR spectroscopy images demonstrated great impact on the acquired spectra, where higher sample scans decreased the standard deviation and the outlier impact. The 256 background and 064 sample scans group (b256_064) showed up as the best cost-benefit for machine learning classification tasks, presenting the best classifications together with b256_128 and b128_128, but with approximately half the acquisition time, thus a better clinical translational potential.

The clustering method using two K-Means was able to identify tissue and paraffin spectra, enabling a fully automated data organization and preprocessing for both 1D and 2D approaches. Individual spectra processing (1D) and patches selection (2D) augmented the dataset size, enabling to develop and train the models with only 30 breast cancer patients, where a larger dataset should improve the metrics and better generalize the models.

CaReNet-V1 efficiently classified breast cancer tissue (CA) against adjacent tissue (AT), with only one test patient false positive. Subtypes were correctly classified, except for four LA/LB wrong predictions. The 1D model coupled with a Gradient-weighted Class Activation Mapping (Grad-CAM) enabled a detailed evaluation of the feature importance, directly correlated to the biochemical composition of the samples, which may assist to better understand the subtypes composition, diagnosis and therapeutic approaches.

CaReNet-V2 showed better metrics than the 1D approach, with perfect AT vs CA and only two LA/LB mistaken for test patients' classification. Furthermore, it provided the biomarkers (estrogen receptor – ER, progesterone receptor – PR, HER2, and Ki67) levels prediction with good overall metrics, but demonstrated lower performance for borderline classes.

The 2D model added the spatial relation to the feature extraction, which is an important task to be executed when dealing with high heterogeneous tissue such as breast cancer. On the other hand, the 1D model can provide a more detailed and easier to interpret feature importance evaluation, directly correlated with the biochemical content of the tissues.

These findings indicate the novel deep learning approach using FTIR hyperspectral images as a potential technique for breast cancer evaluation, providing additional information to the pathology report and also standing out as a biopsy screening analysis technique, helping to prioritize patients. A larger dataset and other classifications should be assessed in following studies, such as TNM (Tumor, Node, Metastasis), staging and grade system, thus providing a more detailed automated report.

REFERENCES

[1] Movasaghi Z, Rehman S, ur Rehman DI. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. Applied Spectroscopy Reviews. 2008 feb;43(2):134-79.

[2] Malek K, Wood BR, Bambery KR. FTIR Imaging of Tissues: Techniques and Methods of Analysis; 2014. p. 419-73.

[3] World Health Organization. Cancer;. Available from: `https://www.who.int/health-topics/cancer`.

[4] Hanf V, Kreienberg R. WHO Report on Cancer Setting Prioritie, Inversting Wisely and Providing Care for All. Geneva; 2003.

[5] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021 may;71(3):209-49.

[6] Brasil. Estimativa 2020 incidencia de cancer no brasil. Rio de Janeiro; 2019.

[7] Hennigs A, Riedel F, Gondos A, Sinn P, Schirmacher P, Marmé F, et al. Prognosis of breast cancer molecular subtypes in routine clinical care: A large prospective cohort study. BMC Cancer. 2016 dec;16(1):734.

[8] Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thürlimann B, Senn HJ. Strategies for subtypes-dealing with the diversity of breast cancer: Highlights of the St Gallen international expert consensus on the primary therapy of early breast cancer 2011. Annals of Oncology. 2011 aug;22(8):1736-47.

[9] Kos Z, Dabbs DJ. Biomarker assessment and molecular testing for prognostication in breast cancer. Histopathology. 2016 jan;68(1):70-85.

[10] Russnes HG, Lingjærde OC, Børresen-Dale AL, Caldas C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. American Journal of Pathology. 2017 oct;187(10):2152-62.

[11] Kalmodia S, Parameswaran S, Yang W, Barrow CJ, Krishnakumar S. Attenuated total reflectance Fourier Transform Infrared spectroscopy: an analytical technique to understand therapeutic responses at the molecular level. Scientific Reports. 2015 dec;5(1):16649.

[12] Kumar S, Srinivasan A, Nikolajeff F. Role of Infrared Spectroscopy and Imaging in Cancer Diagnosis. Current Medicinal Chemistry. 2017 mar;25(9):1055-72.

[13] Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nature Protocols. 2014 aug;9(8):1771-91.

[14] Su KY, Lee WL. Fourier Transform Infrared Spectroscopy as a Cancer Screening and Diagnostic Tool: A Review and Prospects. Cancers. 2020 jan;12(1):115.

[15] Gautam R, Vanga S, Ariese F, Umapathy S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. EPJ Techniques and Instrumentation. 2015 dec;2(1):8.

[16] Morais CLM, Paraskevaidi M, Cui L, Fullwood NJ, Isabelle M, Lima KMG, et al. Standardization of complex biologically derived spectrochemical datasets. Nature Protocols. 2019 may;14(5):1546-77.

[17] Morais CLM, Lima KMG, Singh M, Martin FL. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. Nature Protocols. 2020 jul;15(7):2143-62.

[18] Tahtouh M, Despland P, Shimmon R, Kalman JR, Reedy BJ. The application of infrared chemical imaging to the detection and enhancement of latent fingerprints: Method optimization and further findings. Journal of Forensic Sciences. 2007 sep;52(5):1089-96.

[19] Sacharz J, Perez-Guaita D, Kansiz M, Nazeer SS, Wesełucha-Birczyńska A, Petratos S, et al. Empirical study on the effects of acquisition parameters for FTIR hyperspectral imaging of brain tissue. Analytical Methods. 2020;12(35):4334-42.

[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 may;521(7553):436-44.

[21] Giambagli L, Buffoni L, Carletti T, Nocentini W, Fanelli D. Machine learning in spectral domain. Nature Communications. 2021 dec;12(1):1330.

[22] Wang L, Ren B, He H, Yan S, Lyu D, Xu M, et al. Deep learning for biospectroscopy and biospectral imaging: State-of-the-art and perspectives. Analytical Chemistry. 2021 mar;93(8):3653-65.

[23] Yang J, Xu J, Zhang X, Wu C, Lin T, Ying Y. Deep learning for vibrational spectral analysis: Recent progress and a practical guide. Analytica Chimica Acta. 2019 nov;1081:6-17.

[24] Tomas RC, Sayat AJ, Atienza AN, Danganan JL, Ramos MR, Fellizar A, et al. Detection of breast cancer by ATR-FTIR spectroscopy using artificial neural networks. PLoS ONE. 2022 jan;17(1 January):e0262489.

[25] Du Y, Xie F, Yin L, Yang Y, Yang H, Wu G, et al. Breast cancer early detection by using Fourier-transform infrared spectroscopy combined with different classification algorithms. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 2022 dec;283:121715.

[26] Berisha S, Lotfollahi M, Jahanipour J, Gurcan I, Walsh M, Bhargava R, et al. Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks. Analyst. 2019;144(5):1642-53.

[27] Tortora GJ, Derrickson B. Physiology, principles of anatomy &. 14th ed. Hoboken: John Wiley & Sons; 2021. Available from: https://books.google.com/books/about/Principles_of_Anatomy_and_Physiology.html?id=0c_noQEACAAJ.

[28] McGuire KP. Breast Anatomy and Physiology. In: Breast Disease: Diagnosis and Pathology, Volume 1: Second Edition. vol. 1. 2nd ed. Cham: Springer; 2019. p. 1-9.

[29] Tortora GJ. Anatomy and Physiology: From Science to Life, Illustrated Notebook. 2nd ed. Hoboken: John Wiley & Sons; 2006.

[30] Martini FH. Fundamentals of Anatomy and Physiology [Paperback]. 4th ed. Boston: Cengage Learning; 1989. Available from: http://www.amazon.co.uk/Fundamentals-Anatomy-Physiology-Frederic-Martini/dp/0133326020.

[31] Bland KI, Copeland EM, Klimberg VS. Anatomy of the breast, axilla, chest wall, and related metastatic sites. In: The Breast: Comprehensive Management of Benign and Malignant Diseases. 5th ed. Philadelphia: Elsevier; 2018. p. 20-36.e2.

[32] Pandya S, Moore RG. Breast development and anatomy. Clinical Obstetrics and Gynecology. 2011;54(1):91-5.

[33] Valerie C Scanlon TS. Textbook Essentials of Anatomy and Physiology. 7th ed. Philadelphia: F. A. Davis Company; 2012. Available from: https://books.google.com/books?id=MdlVBgAAQBAJ.

[34] Khonsary S. Guyton and Hall: Textbook of Medical Physiology. vol. 8. 13th ed. Philadelphia: Elsevier; 2017.

[35] Bazira PJ, Ellis H, Mahadevan V. Anatomy and physiology of the breast. In: Surgery (United Kingdom). vol. 40. 2nd ed. Cham: Springer; 2022. p. 79-83.

[36] Eliyatkin N, Yalcin E, Zengel B, Aktaş S, Vardar E. Molecular Classification of Breast Carcinoma: From Traditional, Old-Fashioned Way to A New Age, and A New Way. Journal of Breast Health. 2015;11(2):59-66.

[37] Tuzlali S. Pathology of Breast Cancer. In: Breast Disease: Diagnosis and Pathology, Volume 1: Second Edition. vol. 1. 1st ed. Cham: Springer; 2019. p. 201-20.

[38] Strah KM, Love SM. The in situ carcinomas of the breast. In: Journal of the American Medical Women's Association. vol. 47. 5th ed. Philadelphia: Elsevier; 1992. p. 165-8.

[39] Lakhani S, Ellis I, Schnitt S, Tan P, van de Vijver M. IARC WHO Classification of Tumours of the Breast, Volume 4. 4th ed. Lyon: International Agency for Research on Cancer; 2012.

[40] Cserni G. Histological type and typing of breast carcinomas and the WHO classification changes over time. Pathologica. 2020;112(1):25-41.

[41] Lamb CA, Fabris VT, Lanari C. Progesterone and breast. Best Practice and Research: Clinical Obstetrics and Gynaecology. 2020 nov;69(x):85-94.

[42] Dean-Colomb W, Esteva FJ. Her2-positive breast cancer: Herceptin and beyond. European Journal of Cancer. 2008 dec;44(18):2806-12.

[43] Nielsen TO, Leung SCY, Rimm DL, Dodson A, Acs B, Badve S, et al. Assessment of Ki67 in Breast Cancer: Updated Recommendations From the International Ki67 in Breast Cancer Working Group. JNCI: Journal of the National Cancer Institute. 2021 jul;113(7):808-19.

[44] Penault-Llorca F, Radosevic-Robin N. Ki67 assessment in breast cancer: an update. Pathology. 2017 feb;49(2):166-71.

[45] Escórcio-Dourado CS, Alves-Ribeiro FA, Lima-Dourado JC, dos Santos AR, de Oliveira Pereira R, Tavares CB, et al. Human Epidermal Growth Factor Receptor-2 gene polymorphism and breast cancer risk in women from the Northeastern region of Brazil. Clinics. 2020;75:e2360.

[46] Kadamkulam Syriac A, Nandu NS, Leone JP. Central Nervous System Metastases from Triple-Negative Breast Cancer: Current Treatments and Future Prospective. Breast Cancer: Targets and Therapy. 2022 jan;Volume 14:1-13.

[47] Smith BC. Fundamentals of fourier transform infrared spectroscopy, second edition. 2nd ed. Boca Raton: CRC Press; 2011.

[48] Mayerhöfer TG, Pahlow S, Hübner U, Popp J. Removing interference-based effects from the infrared transflectance spectra of thin films on metallic substrates: a fast and wave optics conform solution. The Analyst. 2018;143(13):3164-75.

[49] Schneider M, Demoulin P, Sussmann R, Notholt J. Fourier transform infrared spectrometry. vol. 10. 2nd ed. Hoboken: John Wiley & Sons; 2013.

[50] Xu X, Xu S, Jin L, Song E. Characteristic analysis of Otsu threshold and its applications. Pattern Recognition Letters. 2011 may;32(7):956-61.

[51] Rainer RJ, Mayr M, Himmelbauer J, Nikzad-Langerodi R. Opening the black-box of Neighbor Embeddings with Hotelling's <math altimg="si7.svg" display="inline" id="d1e317"> <msup> <mrow> <mi>T</mi> </mrow> <mrow> <mn>2</mn> </mrow> </msup> </math> statistic and <math altimg="si8.svg" display="inline" id="d1e327". Chemometrics and Intelligent Laboratory Systems. 2023 jul;238:104840.

[52] Miller CE. Chemometrics in Process Analytical Technology (PAT). In: Bakeev KA, editor. Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries. Jersey City: John Wiley & Sons; 2010. p. 353-438.

[53] Zhao AX, Tang XJ, Zhang ZH, Liu JH. The parameters optimization selection of Savitzky-Golay filter and its application in smoothing pretreatment for FTIR spectra. In: 2014 9th IEEE Conference on Industrial Electronics and Applications. IEEE; 2014. p. 516-21.

[54] Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry. 1964 jul;36(8):1627-39.

[55] Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems. 2012 aug;117:92-9.

[56] De Lima FA, Gobinet C, Sockalingum G, Garcia SB, Manfait M, Untereiner V, et al. Digital de-waxing on FTIR images. Analyst. 2017;142(8):1358-70.

[57] Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends. 2020 may;1(2):56-70.

[58] Mota LFM, Pegolo S, Baba T, Peñagaricano F, Morota G, Bittante G, et al. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. Journal of Dairy Science. 2021 jul;104(7):8107-21.

[59] Pfeiffer P, Ronai B, Vorlaufer G, Dörr N, Filzmoser P. Weighted LASSO variable selection for the analysis of FTIR spectra applied to the prediction of engine oil degradation. Chemometrics and Intelligent Laboratory Systems. 2022 sep;228:104617.

[60] Khanmohammadi M, Ghasemi K, Garmarudi AB. Genetic algorithm spectral feature selection coupled with quadratic discriminant analysis for ATR-FTIR spectrometric diagnosis of basal cell carcinoma via blood sample analysis. RSC Adv. 2014;4(78):41484-90.

[61] de Magalhães CR, Carrilho R, Schrama D, Cerqueira M, Rosa da Costa AM, Rodrigues PM. Mid-infrared spectroscopic screening of metabolic alterations in stress-exposed gilthead seabream (Sparus aurata). Scientific Reports. 2020 oct;10(1):16343.

[62] Efron B, Hastie T. Computer Age Statistical Inference. Cambridge: Cambridge University Press; 2016.

[63] Bzdok D, Yeo BTT. Inference in the age of big data: Future perspectives on neuroscience. NeuroImage. 2017 jul;155:549-64.

[64] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data. 2021 mar;8(1):53.

[65] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences. 2023 apr;622:178-210.

[66] Tang L, Peng S, Bi Y, Shan P, Hu X. A New Method Combining LDA and PLS for Dimension Reduction. PLoS ONE. 2014 may;9(5):e96944.

[67] Skiena SS. The Data Science Design Manual. Texts in Computer Science. Cham: Springer International Publishing; 2017.

[68] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 mar;13-17-Augu:785-94.

[69] Iliadis LS, Kurkova V, Hammer B. Brain-inspired computing and machine learning. Neural Computing and Applications. 2020 jun;32(11):6641-3.

[70] Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An Introductory Review of Deep Learning for Prediction Models With Big Data. Frontiers in Artificial Intelligence. 2020 feb;3.

[71] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 dec.

[72] Goodfellow IJ, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016. Available from: http://www.deeplearningbook.org.

[73] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognition. 2018 may;77:354-77.

[74] Li Z, Liu F, Yang W, Peng S, Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems. 2022 dec;33(12):6999-7019.

[75] Demšar J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. 2006;7:1-30.

[76] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings; 2015. .

[77] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2016-Decem; 2016. p. 770-8.

[78] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2016 mar;9908 LNCS:630-45.

[79] Korb E, Bağcıoğlu M, Garner-Spitzer E, Wiedermann U, Ehling-Schulz M, Schabussova I. Machine learning-empowered ftir spectroscopy serum analysis

stratifies healthy, allergic, and sit-treated mice and humans. Biomolecules. 2020 jul;10(7):1-17.

[80] Zhang L, Ding X, Hou R. Classification Modeling Method for Near-Infrared Spectroscopy of Tobacco Based on Multimodal Convolution Neural Networks. Journal of Analytical Methods in Chemistry. 2020 feb;2020:1-13.

[81] Chen X, Chai Q, Lin N, Li X, Wang W. 1D convolutional neural network for the discrimination of aristolochic acids and their analogues based on near-infrared spectroscopy. Analytical Methods. 2019;11(40):5118-25.

[82] Paoletti ME, Haut JM, Plaza J, Plaza A. A new deep convolutional neural network for fast hyperspectral image classification. ISPRS Journal of Photogrammetry and Remote Sensing. 2018 nov;145:120-47.

[83] He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 feb.

[84] Smith LN, Topin N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. 2017 aug.

[85] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2020 oct;128(2):336-59.

[86] Jianlin Cheng, Zheng Wang, Pollastri G. A neural network approach to ordinal regression. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE; 2008. p. 1279-84.

[87] Xu Y, Li Z, Li W, Du Q, Liu C, Fang Z, et al. Dual-Channel Residual Network for Hyperspectral Image Classification With Noisy Labels. IEEE Transactions on Geoscience and Remote Sensing. 2022;60:1-11.

[88] Yu S, Jia S, Xu C. Convolutional neural networks for hyperspectral image classification. Neurocomputing. 2017 jan;219:88-98.

[89] Song H, Yang W, Dai S, Du L, Sun Y. Using dual-channel CNN to classify hyperspectral image based on spatial-spectral information. Mathematical Biosciences and Engineering. 2020;17(4):3450-77.

[90] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine. 2018 jan;35(1):53-65.

[91] Iqbal T, Ali H. Generative Adversarial Network for Medical Images (MI-GAN). Journal of Medical Systems. 2018 nov;42(11):231.

[92] Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts. 2016 aug.

[93] Balan V, Mihai CT, Cojocaru FD, Uritu CM, Dodi G, Botezat D, et al. Vibrational spectroscopy fingerprinting in medicine: From molecular to clinical practice. Materials. 2019 sep;12(18):2884.

[94] Meksiarun P, Aoki PHB, Van Nest SJ, Sobral-Filho RG, Lum JJ, Brolo AG, et al. Breast cancer subtype specific biochemical responses to radiation. Analyst. 2018;143(16):3850-8.

[95] Murayama T, Gotoh N. Patient-derived xenograft models of breast cancer and their application. Cells. 2019 jun;8(6):621.

[96] Bruun SW, Kohler A, Adt I, Sockalingum GD, Manfait M, Martens H. Correcting Attenuated Total Reflection—Fourier Transform Infrared Spectra for Water Vapor and Carbon Dioxide. Applied Spectroscopy. 2006 sep;60(9):1029-39.

[97] Hodge VJ, Austin J. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review. 2004 oct;22(2):85-126.

[98] Kanamori T, Fujiwara S, Takeda A. Breakdown point of robust support vector machines. Entropy. 2017 sep;19(2).

[99] Khan MMR, Arif RB, Siddique AB, Oishe MR. Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository. 4th International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2018. 2018 sep:124-9.

[100] Fragomeni SM, Sciallis A, Jeruss JS. Molecular Subtypes and Local-Regional Control of Breast Cancer. Surgical Oncology Clinics of North America. 2018 jan;27(1):95-120.

[101] Lv Y, Ma X, Du Y, Feng J. Understanding Patterns of Brain Metastasis in Triple-Negative Breast Cancer and Exploring Potential Therapeutic Targets. OncoTargets and Therapy. 2021 jan;Volume 14:589-607.

[102] Kar S, Katti DR, Katti KS. Fourier transform infrared spectroscopy based spectral biomarkers of metastasized breast cancer progression. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 2019 feb;208:85-96.

[103] Kołodziej M, Kaznowska E, Paszek S, Cebulski J, Barnaś E, Cholewa M, et al. Characterisation of breast cancer molecular signature and treatment assessment with vibrational spectroscopy and chemometric approach. PLOS ONE. 2022 mar;17(3):e0264347.

[104] Cappelletti V, Iorio E, Miodini P, Silvestri M, Dugo M, Daidone MG. Metabolic Footprints and Molecular Subtypes in Breast Cancer. Disease Markers. 2017;2017:1-19.

[105] Herman CR, Gill HK, Eng J, Fajardo LL. Screening for Preclinical Disease: Test and Disease Characteristics. American Journal of Roentgenology. 2002 oct;179(4):825-31.

[106] Nahed AS, Shaimaa MY. Ki-67 as a prognostic marker according to breast cancer molecular subtype. Cancer Biology & Medicine. 2016;13(4):496.

[107] Waks AG, Winer EP. Breast Cancer Treatment. JAMA. 2019 jan;321(3):288.

[108] Hwang KT, Kim J, Jung J, Chang JH, Chai YJ, Oh SW, et al. Impact of Breast Cancer Subtypes on Prognosis of Women with Operable Invasive Breast Cancer: A Population-based Study Using SEER Database. Clinical Cancer Research. 2019 mar;25(6):1970-9.

[109] Howlader N, Cronin KA, Kurian AW, Andridge R. Differences in Breast Cancer Survival by Molecular Subtypes in the United States. Cancer Epidemiology, Biomarkers & Prevention. 2018 jun;27(6):619-26.

[110] Tsoutsou PG, Vozenin MC, Durham AD, Bourhis J. How could breast cancer molecular features contribute to locoregional treatment decision making? Critical Reviews in Oncology/Hematology. 2017 feb;110:43-8.

[111] Yang F, Li J, Zhang H, Zhang S, Ye J, Cheng Y, et al. Correlation between Androgen Receptor Expression in Luminal B (HER–2 Negative) Breast Cancer and Disease Outcomes. Journal of Personalized Medicine. 2022 dec;12(12):1988.

[112] Zhao H, Gong Y. The Prognosis of Single Hormone Receptor-Positive Breast Cancer Stratified by HER2 Status. Frontiers in Oncology. 2021 may;11.

[113] Fan Y, Ding X, Xu B, Ma F, Yuan P, Wang J, et al. Prognostic Significance of Single Progesterone Receptor Positivity. Medicine. 2015 nov;94(46):e2066.

[114] Andrahennadi S, Sami A, Manna M, Pauls M, Ahmed S. Current Landscape of Targeted Therapy in Hormone Receptor-Positive and HER2-Negative Breast Cancer. Current Oncology. 2021 may;28(3):1803-22.

[115] Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. Journal of Clinical Oncology. 2020 apr;38(12):1346-66.

[116] Yi M, Huo L, Koenig KB, Mittendorf EA, Meric-Bernstam F, Kuerer HM, et al. Which threshold for ER positivity? a retrospective study based on 9639 patients. Annals of Oncology. 2014 may;25(5):1004-11.

[117] English DP, Roque DM, Santin AD. HER2 Expression Beyond Breast Cancer: Therapeutic Implications for Gynecologic Malignancies. Molecular Diagnosis & Therapy. 2013 apr;17(2):85-99.

[118] Figueroa-Magalhães MC, Jelovac D, Connolly RM, Wolff AC. Treatment of HER2-positive breast cancer. The Breast. 2014 apr;23(2):128-36.

[119] Loibl S, Gianni L. HER2-positive breast cancer. The Lancet. 2017 jun;389(10087):2415-29.

[120] Wang S, Liu W, Wu J, Cao L, Meng Q, Kennedy PJ. Training deep neural networks on imbalanced data sets. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE; 2016. p. 4368-74.

[121] Liang Q, Ma D, Gao RF, Yu KD. Effect of Ki-67 Expression Levels and Histological Grade on Breast Cancer Early Relapse in Patients with Different Immunohistochemical-based Subtypes. Scientific Reports. 2020 may;10(1):7648.

[122] Choi SB, Park JM, Ahn JH, Go J, Kim J, Park HS, et al. Ki-67 and breast cancer prognosis: does it matter if Ki-67 level is examined using preoperative biopsy or postoperative specimen? Breast Cancer Research and Treatment. 2022 apr;192(2):343-52.

[123] Skjervold AH, Pettersen HS, Valla M, Opdahl S, Bofin AM. Visual and digital assessment of Ki-67 in breast cancer tissue - a comparison of methods. Diagnostic Pathology. 2022 dec;17(1):45.

[124] Wu Q, Ma G, Deng Y, Luo W, Zhao Y, Li W, et al. Prognostic Value of Ki-67 in Patients With Resected Triple-Negative Breast Cancer: A Meta-Analysis. Frontiers in Oncology. 2019 oct;9.

[125] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019 may.

[126] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. 2022 jan.

[127] Ahmed WS, Karim AaA. The Impact of Filter Size and Number of Filters on Classification Accuracy in CNN. In: 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE; 2020. p. 88-93.

[128] Lin M, Chen Q, Yan S. Network In Network. 2013 dec.

[129] Dablain D, Jacobson KN, Bellinger C, Roberts M, Chawla N. Understanding CNN Fragility When Learning With Imbalanced Data. 2022 oct.

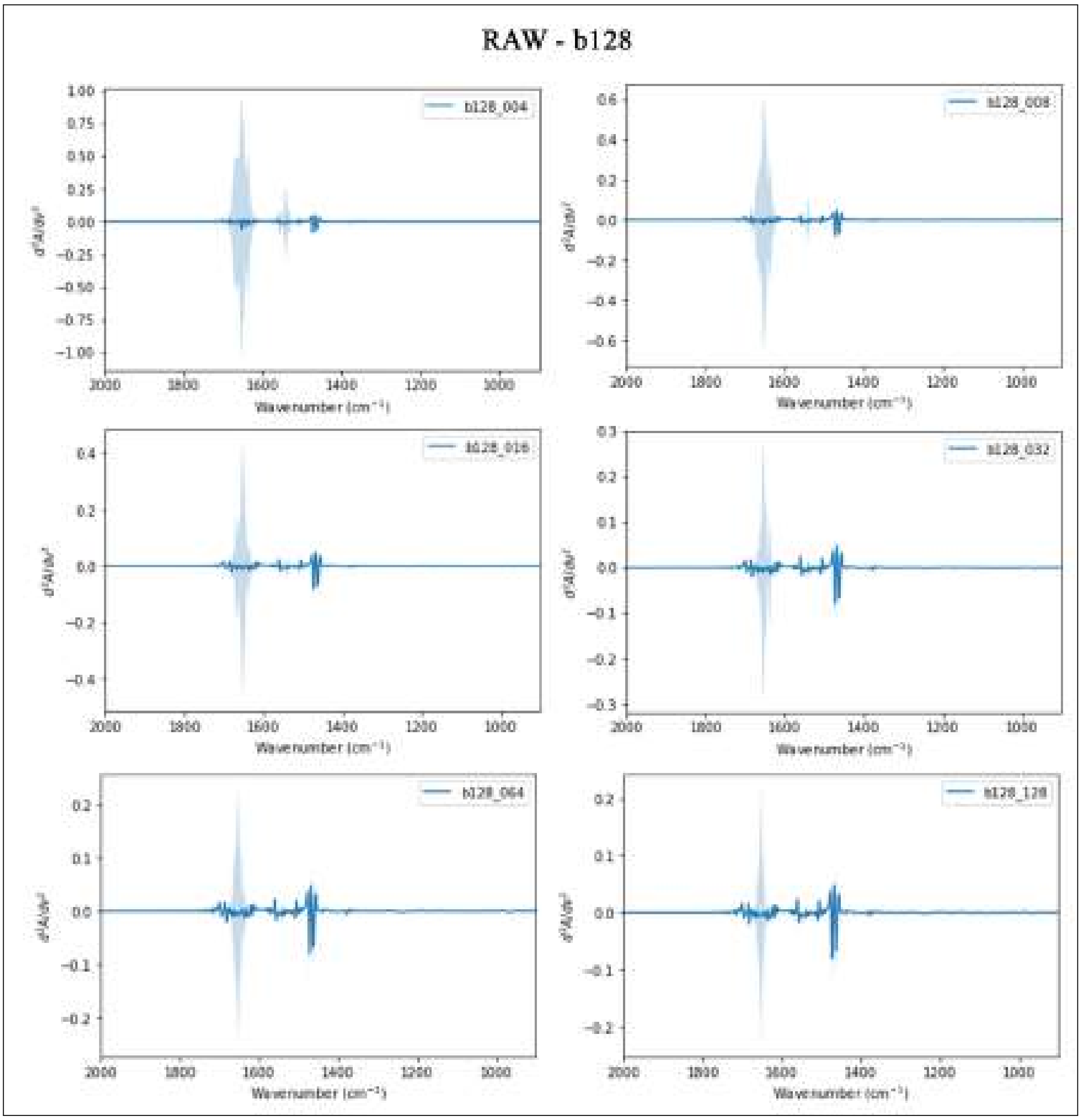
[130] Wolf DM, Yau C, Wulfkuhle J, Brown-Swigart L, Gallagher RI, Lee PRE, et al. Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. Cancer Cell. 2022 jun;40(6):609-23.e6.

[131] Hortobagyi GN, Edge SB, Giuliano A. New and Important Changes in the TNM Staging System for Breast Cancer. American Society of Clinical Oncology Educational Book. 2018 may;(38):457-67.

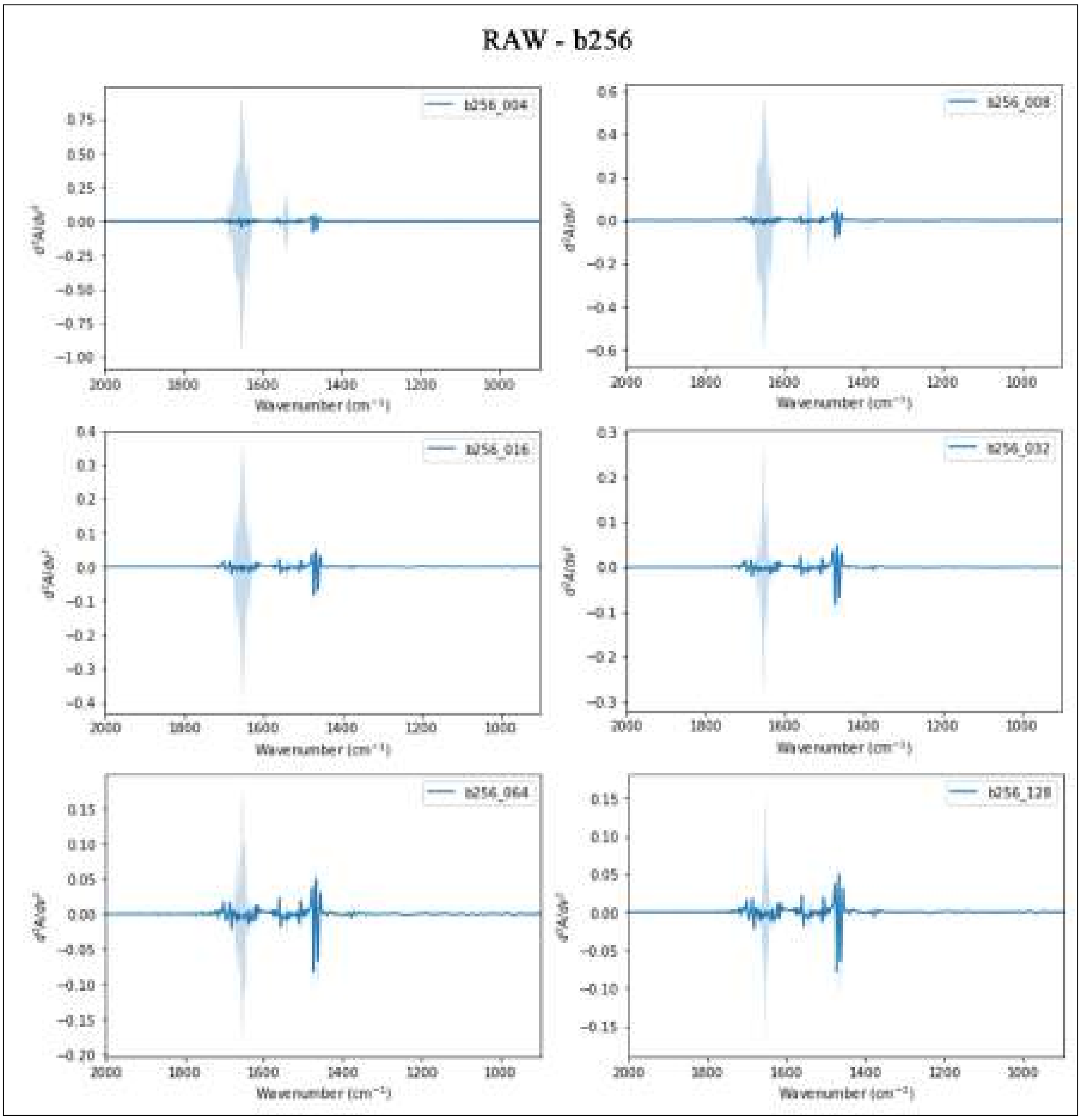
[132] van Dooijeweert C, Baas IO, Deckers IAG, Siesling S, van Diest PJ, van der Wall E. The increasing importance of histologic grading in tailoring adjuvant systemic therapy in 30,843 breast cancer patients. Breast Cancer Research and Treatment. 2021 jun;187(2):577-86.

APPENDICES

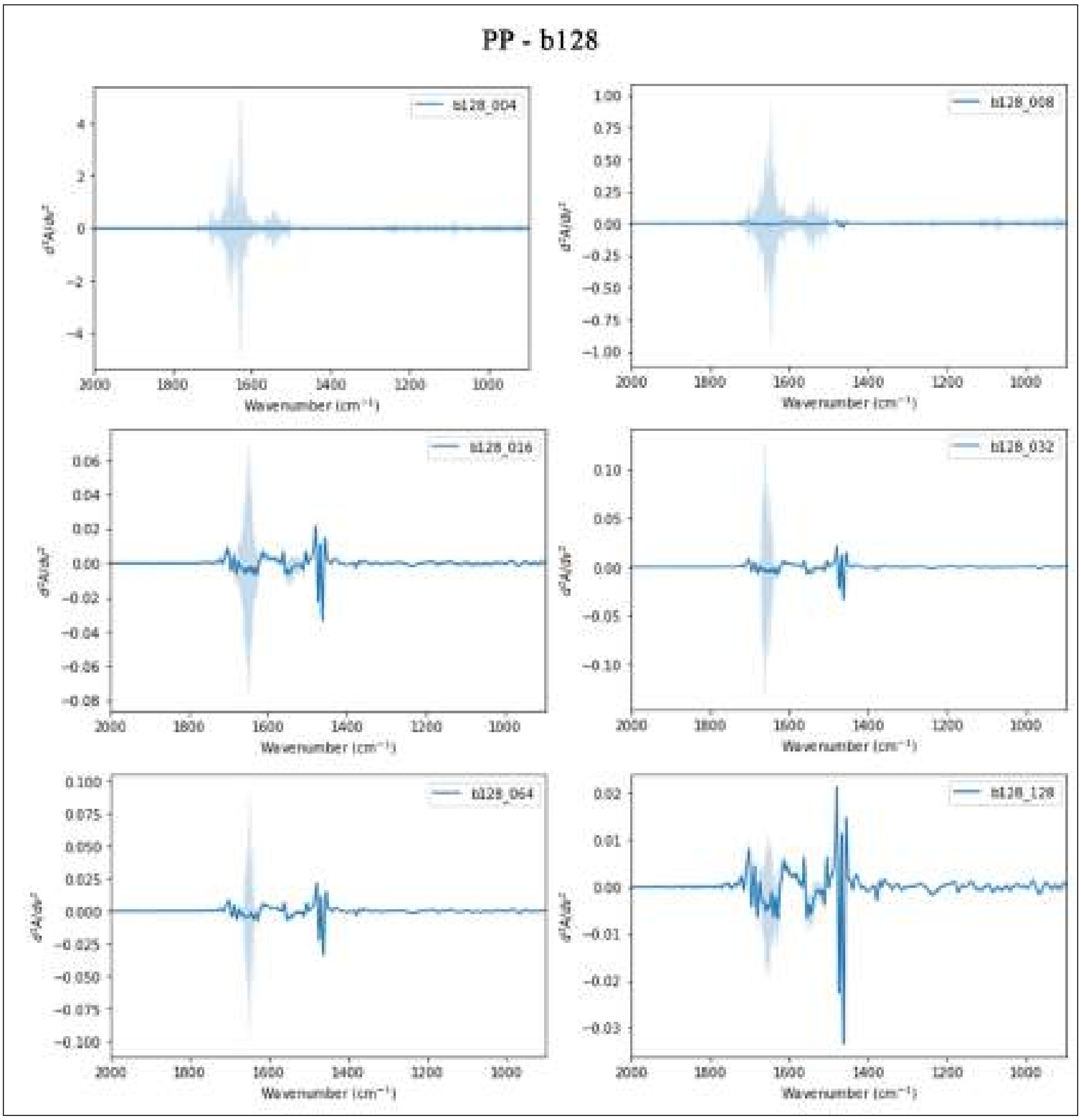
APPENDIX A – Spectra comparison of RAW b128 group. Mean spectrum (solid line) and standard deviation (shades).



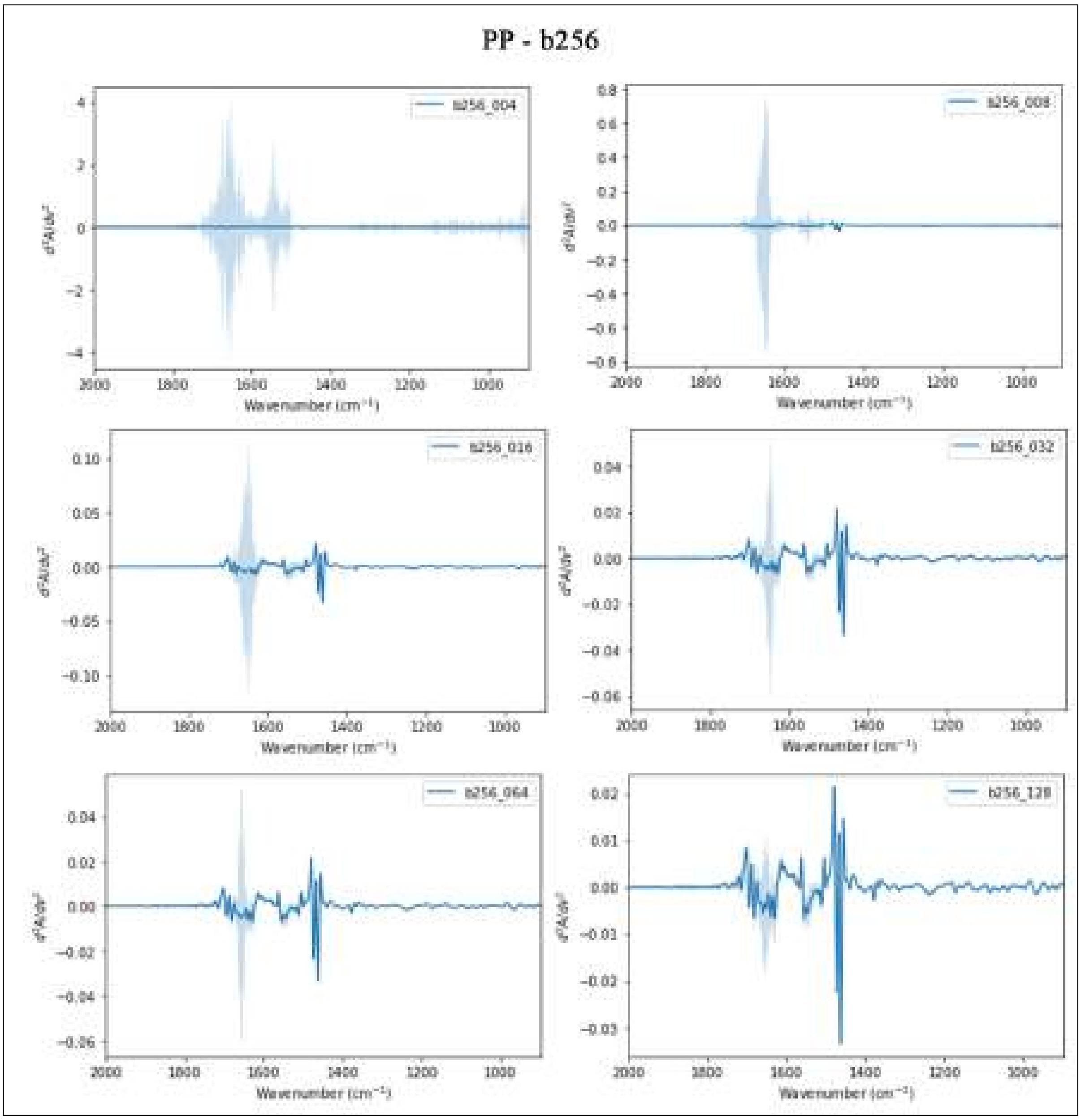
APPENDIX B – Spectra comparison of RAW b256 group. Mean spectrum (solid line) and standard deviation (shades).



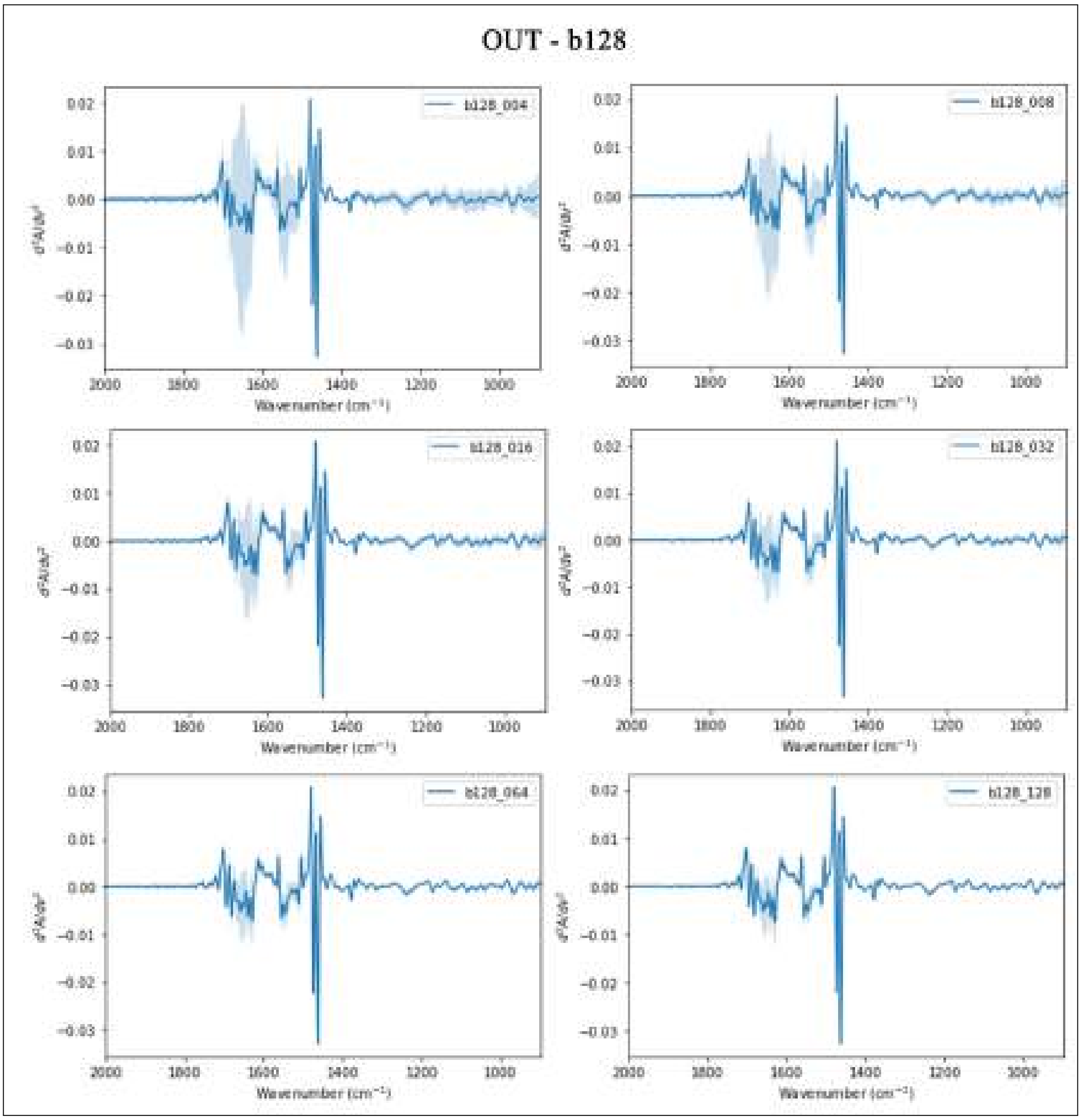
APPENDIX C – Spectra comparison of PP b128 group. Mean spectrum (solid line) and standard deviation (shades).



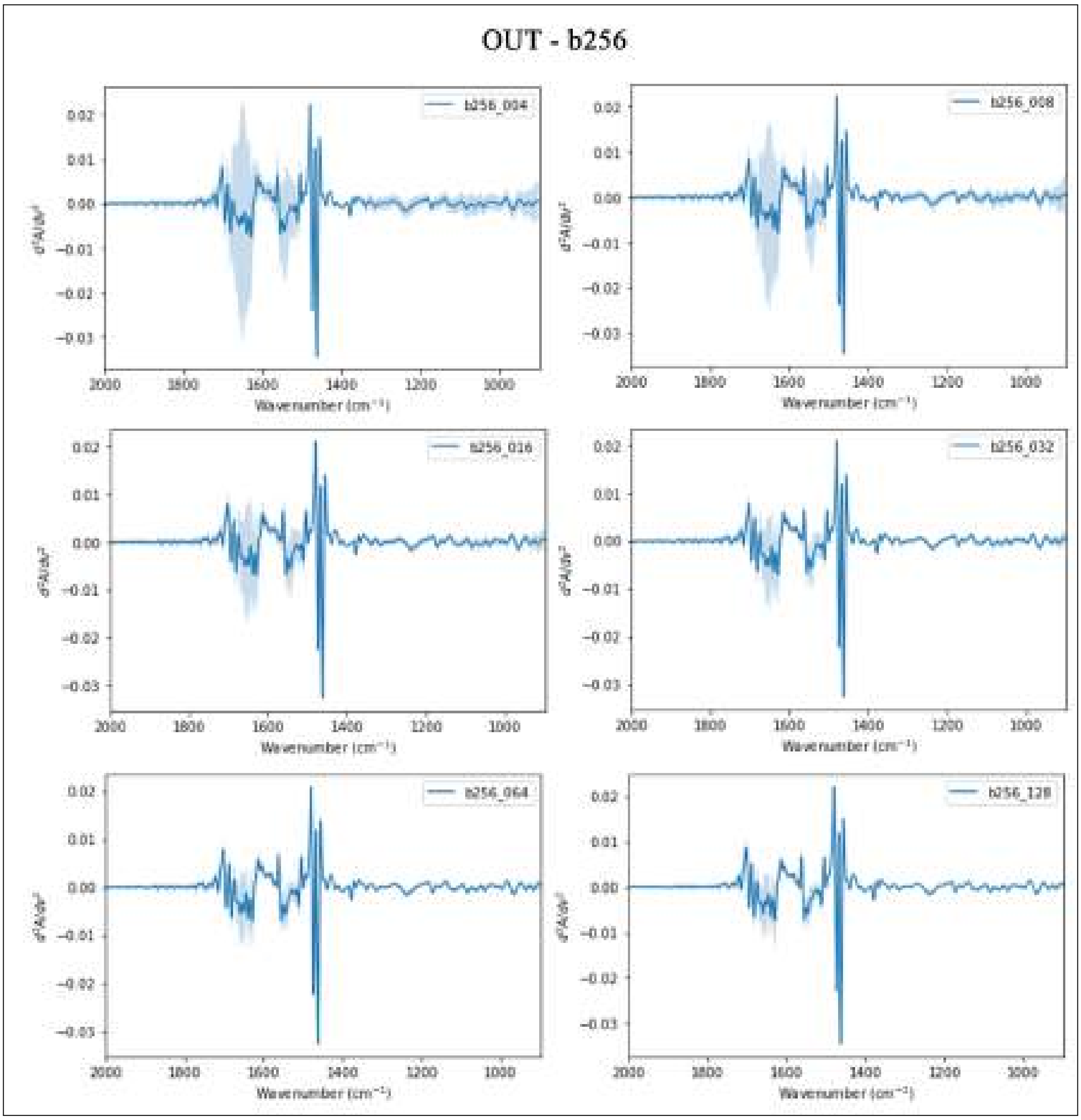
APPENDIX D – Spectra comparison of PP b256 group. Mean spectrum (solid line) and standard deviation (shades).



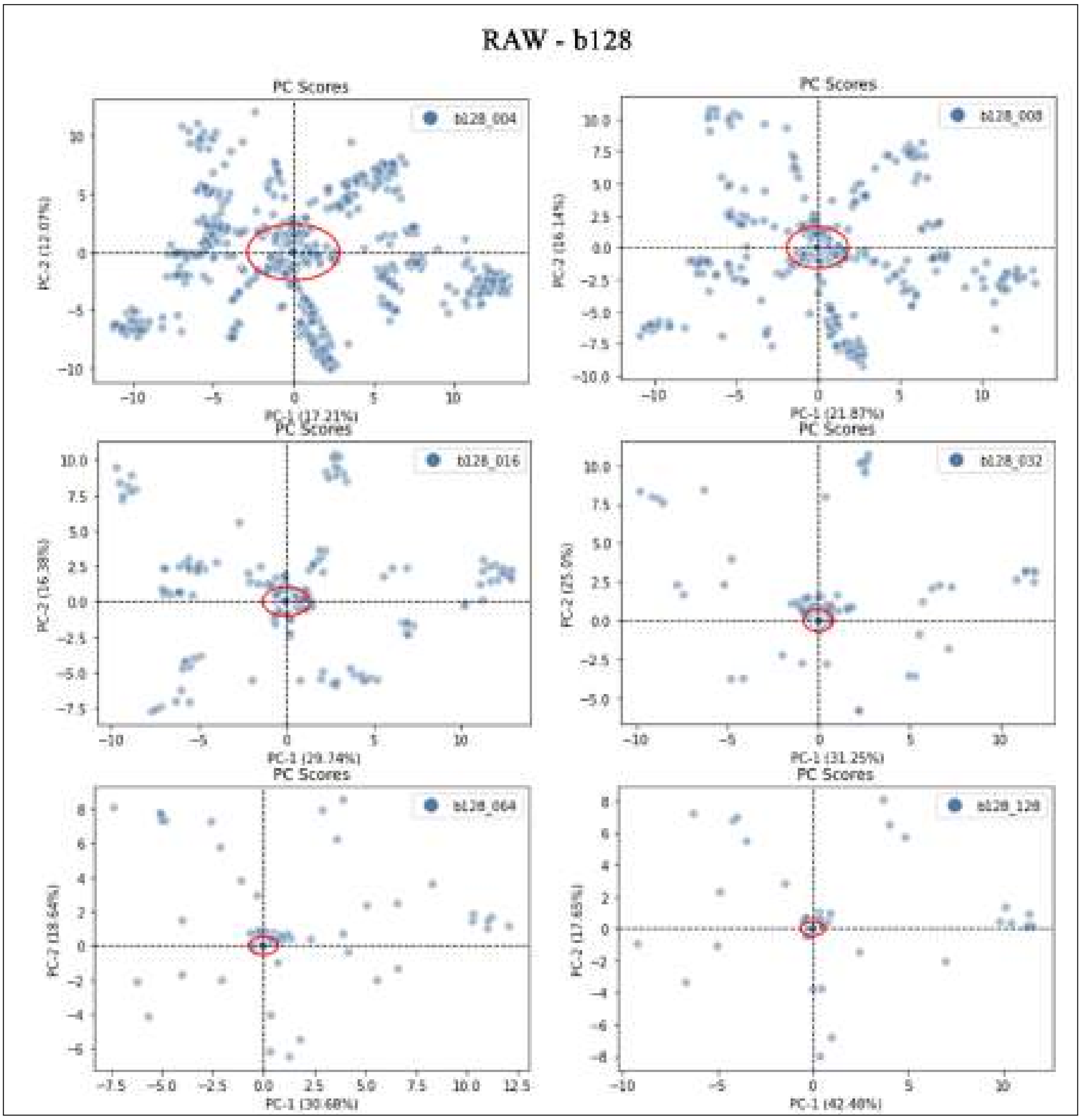
APPENDIX E – Spectra comparison of OUT b128 group. Mean spectrum (solid line) and standard deviation (shades).



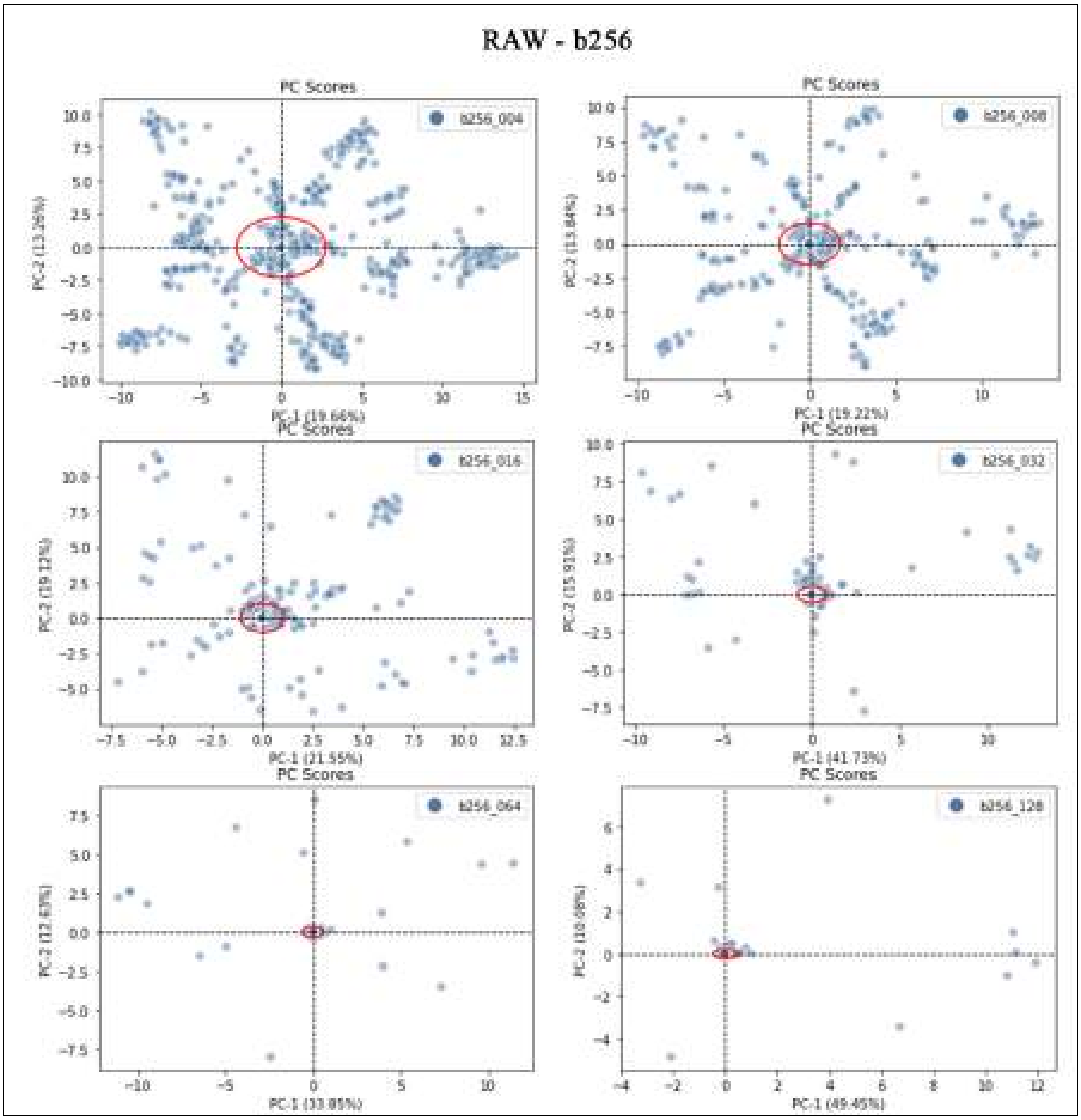
APPENDIX F – Spectra comparison of OUT b256 group. Mean spectrum (solid line) and standard deviation (shades).



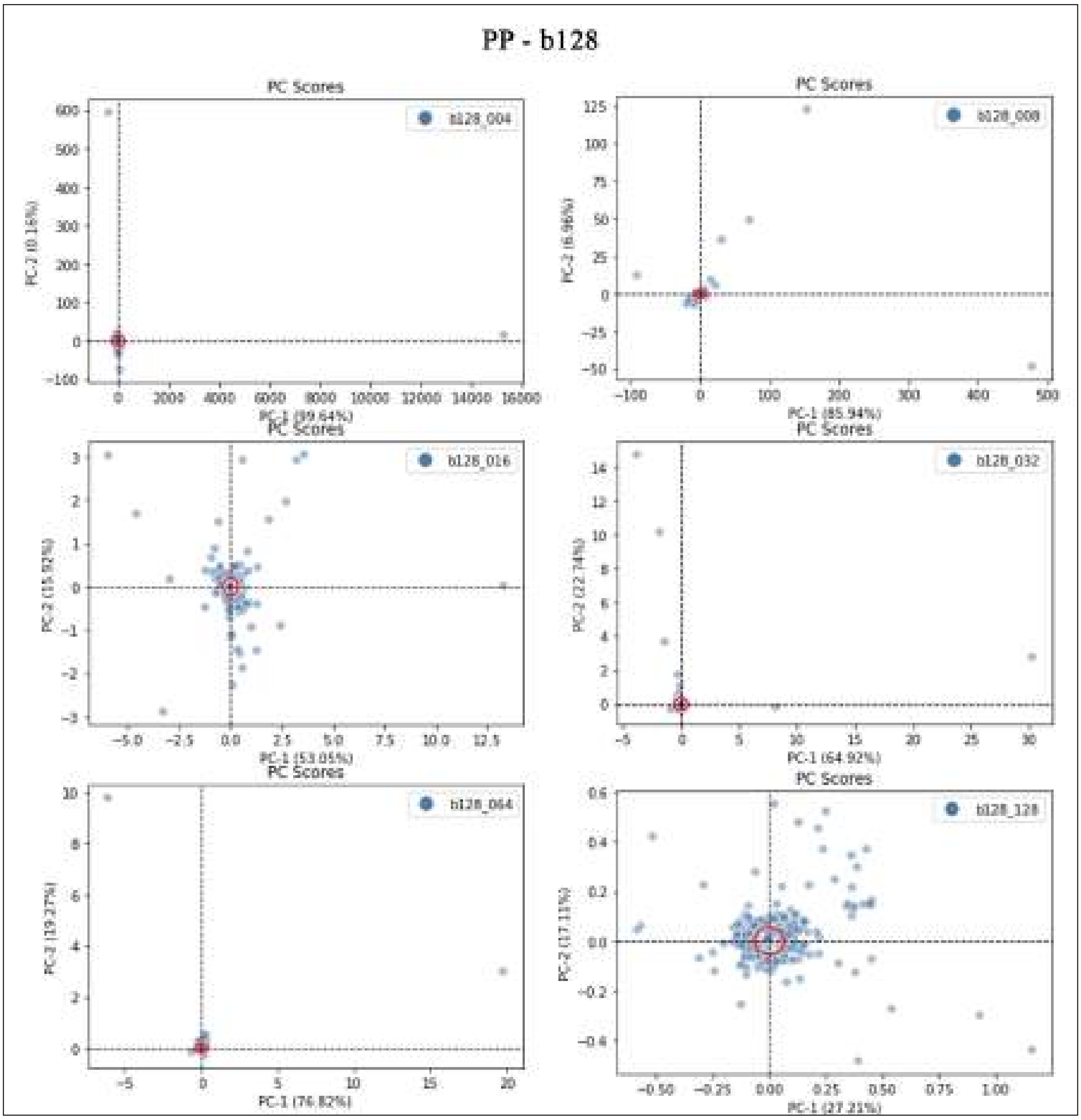
APPENDIX G – Principal Component Analysis (PCA) of RAW b128 for group. Dots denote PC-scores and red line the Hotelling's T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.



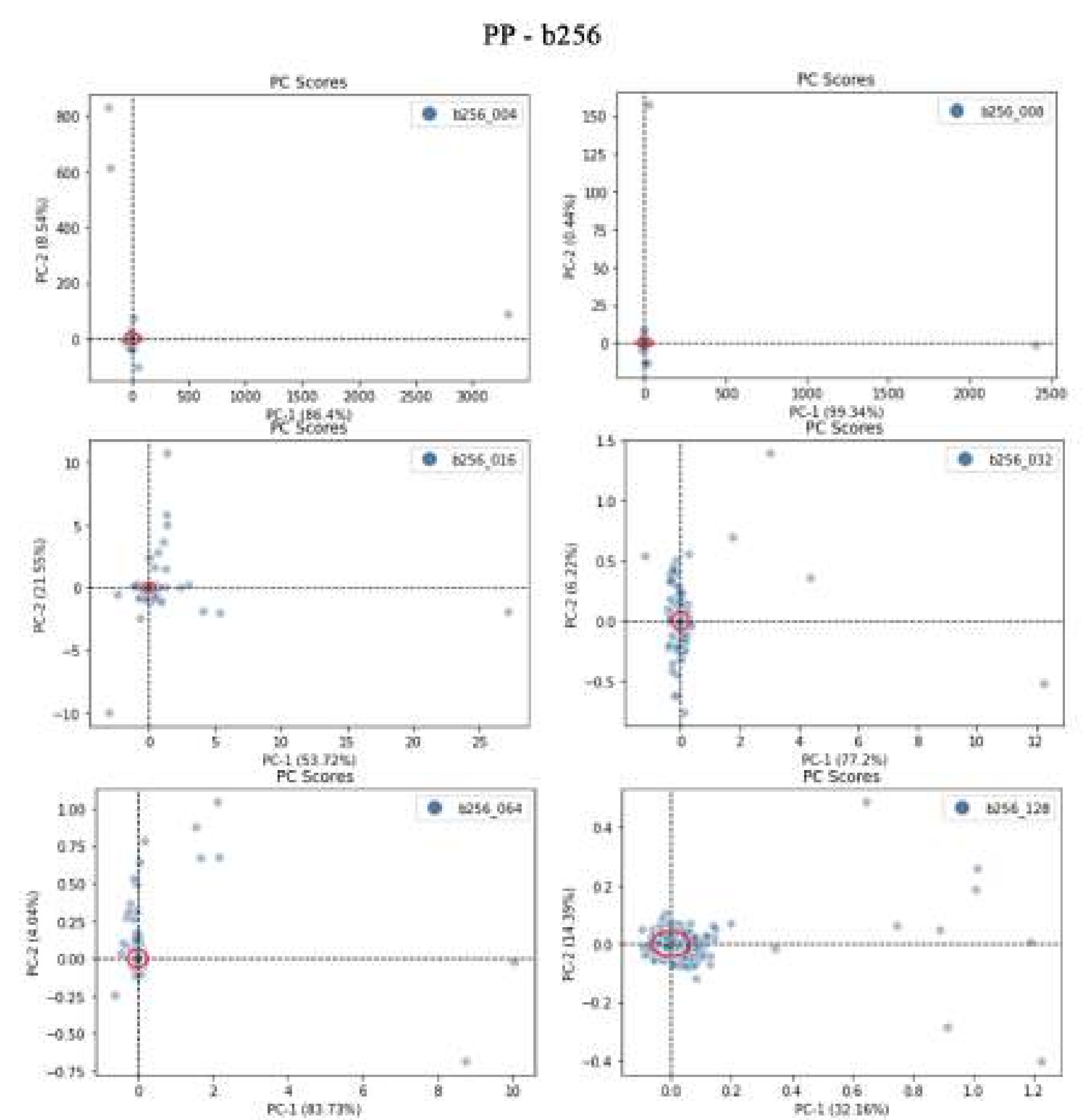
APPENDIX H – Principal Component Analysis (PCA) of RAW b256 for group. Dots denote PC-scores and red line the Hotelling's T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.



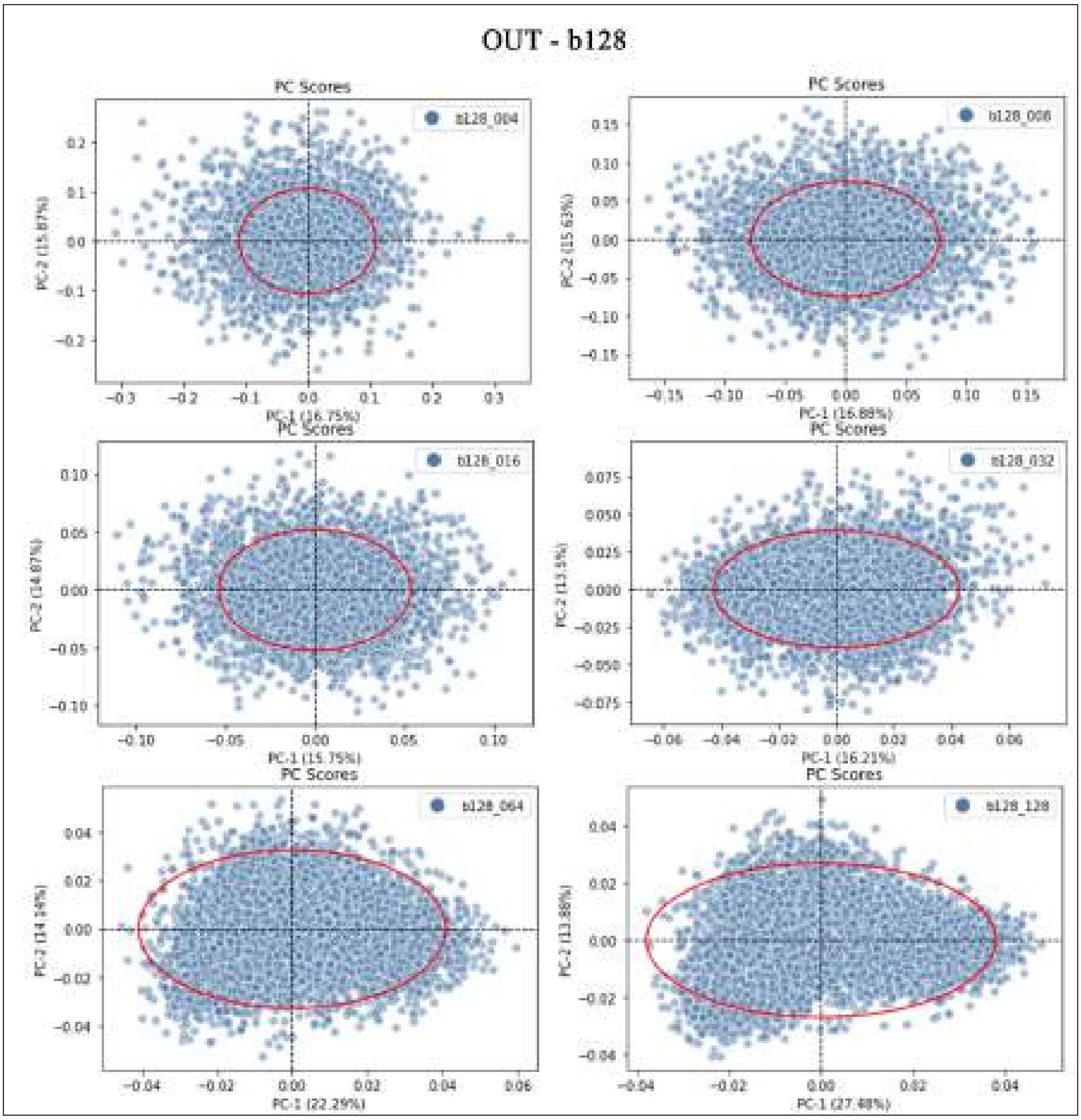
APPENDIX I – Principal Component Analysis (PCA) of PP b128 for group. Dots denote PC-scores and red line the Hotelling's T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.



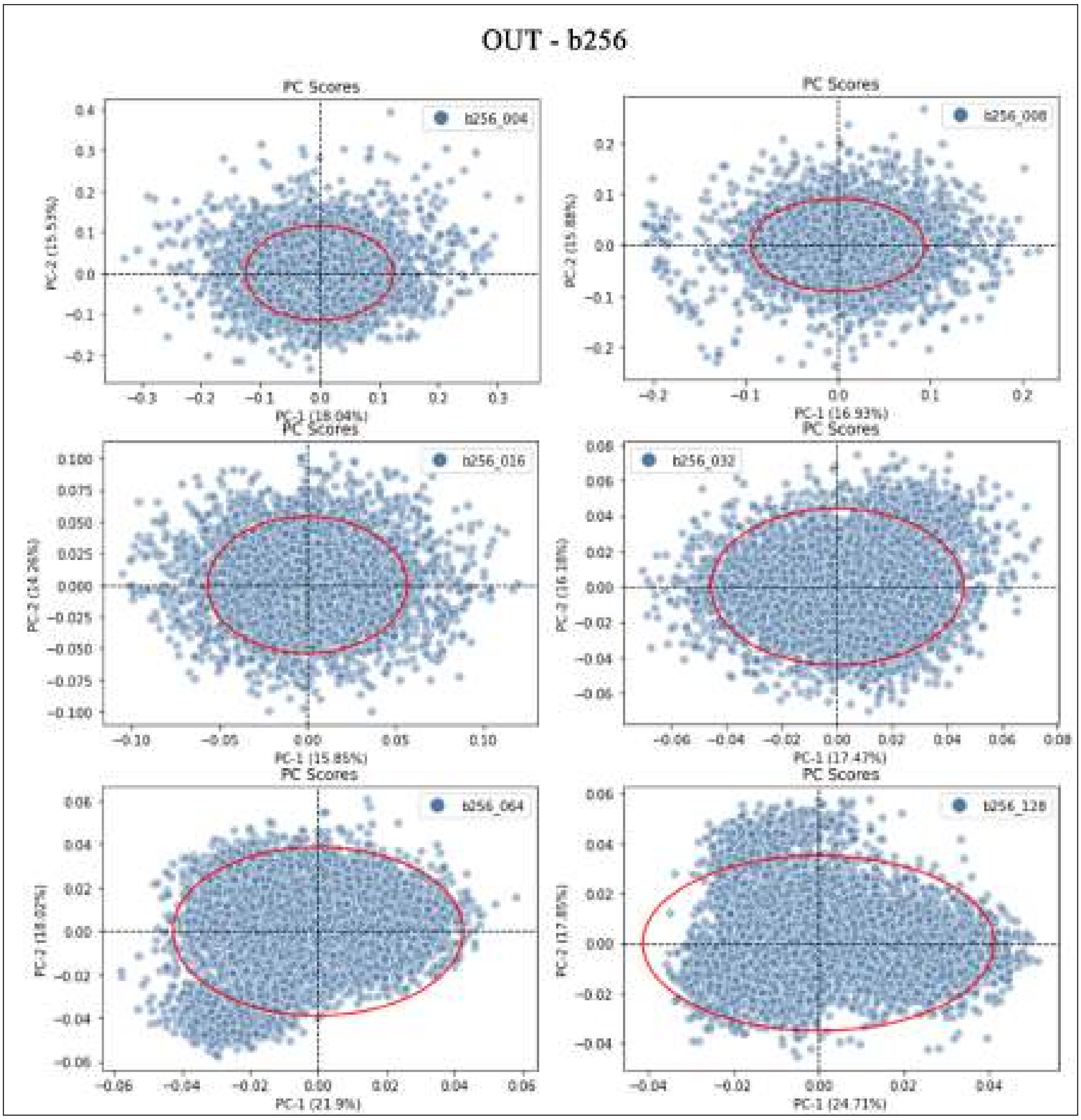
APPENDIX J – Principal Component Analysis (PCA) of PP b256 for group. Dots denote PC-scores and red line the Hotelling’s T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.



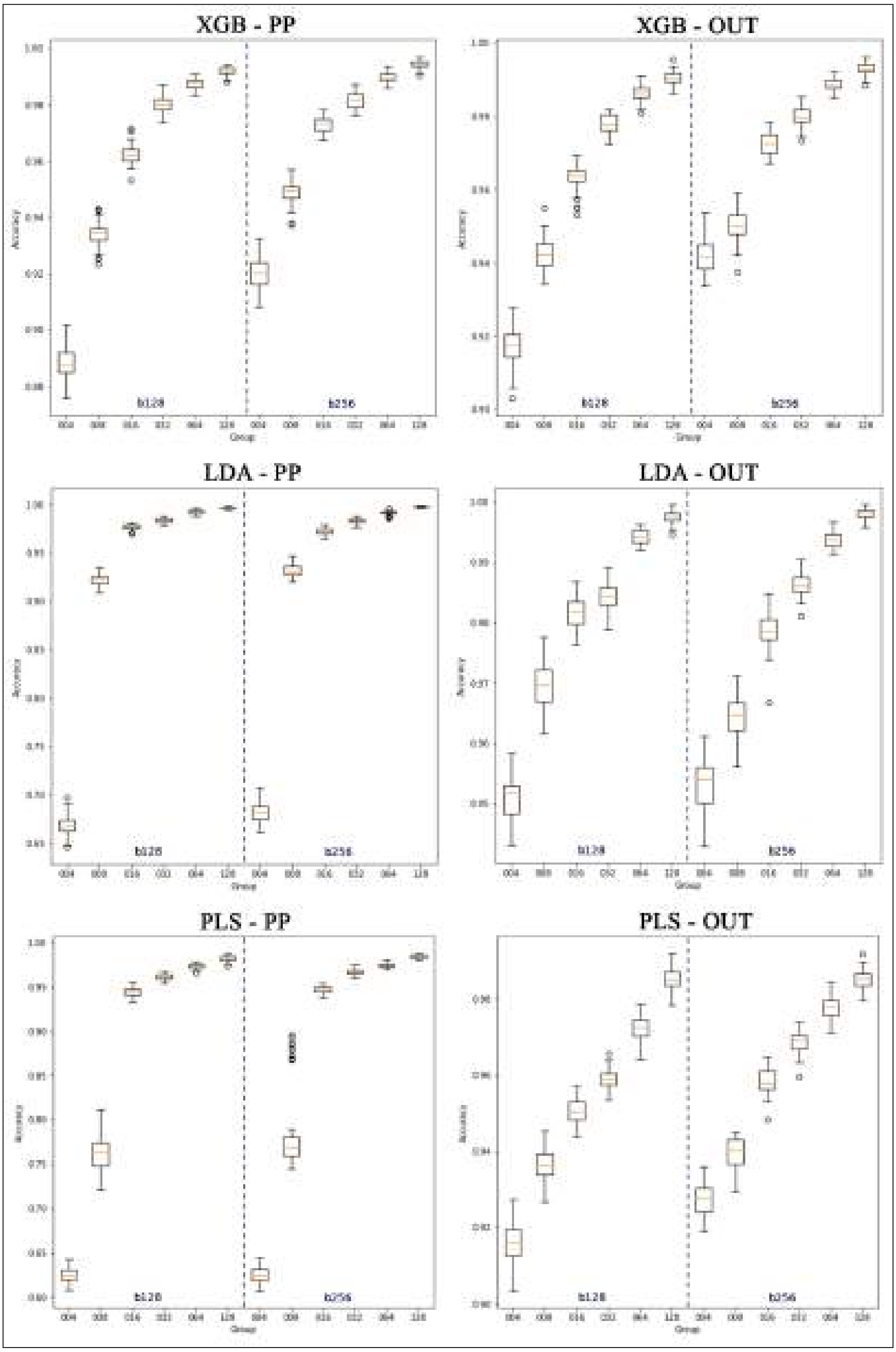
APPENDIX K – Principal Component Analysis (PCA) of OUT b128 for group. Dots denote PC-scores and red line the Hotelling's T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.



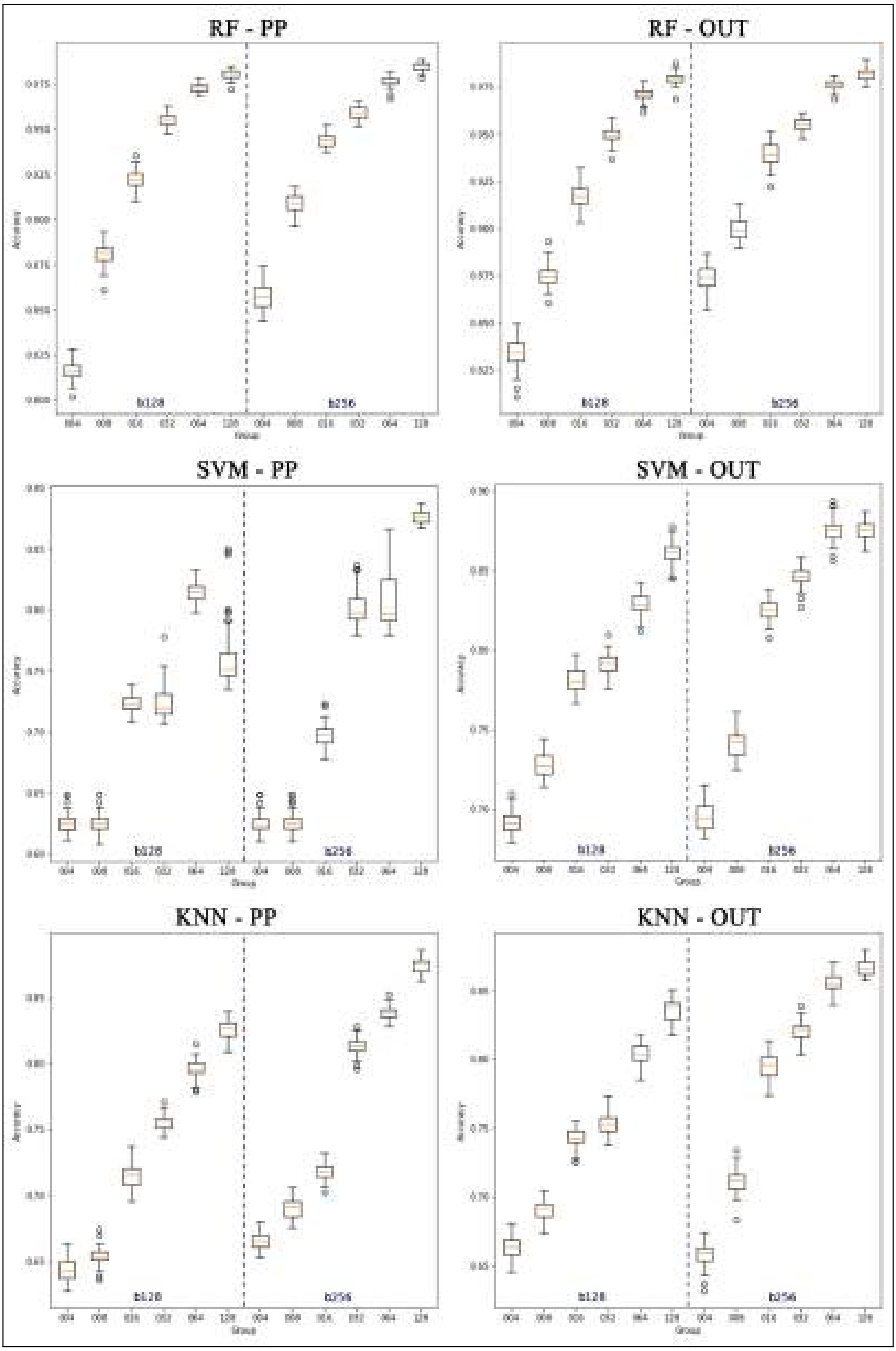
APPENDIX L – Principal Component Analysis (PCA) of OUT b256 for group. Dots denote PC-scores and red line the Hotelling's T^2 95% confidence ellipse. PC-1 and -2 cumulative variance between parentheses.



APPENDIX M – Test accuracy boxplot by groups of the models XGB (Extreme Gradient Boost), LDA (Linear Discriminant Analysis), and Partial Least Squares Discriminant Analysis (PLS-DA). Dashed blue line splits different background scans (b128 and b256).


XGB - PP
1.00
0.98
0.96
0.94
0.92
0.90
0.88
Accuracy
b128
b256
004
008
016
032
064
128
Group
XGB - OUT
1.00
0.98
0.96
0.94
0.92
0.90
LDA - PP
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
LDA - OUT
1.00
0.99
0.98
0.97
0.96
0.95
PLS - PP
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
PLS - OUT
0.98
0.96
0.94
0.92
0.90

APPENDIX N – Test accuracy boxplot by groups of the models RF (Random Forest), SVM (Support Vector Machine), and KNN (K-Nearest Neighbors). Dashed blue line splits different background scans (b128 and b256).



INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Diretoria de Pesquisa, Desenvolvimento e Ensino
Av. Prof. Lineu Prestes, 2242 – Cidade Universitária CEP: 05508-000
Fone/Fax(0XX11) 3133-8908
SÃO PAULO – São Paulo – Brasil
http://www.ipen.br