Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

Characterization and classification of deep endometriosis lesions in magnetic resonance imaging

Supervisor:
Dr. Nico Curti

Co-supervisor:
Dr. Sara Peluso

Submitted by:
Elettra Lucchesi

Academic Year 2023/2024

Abstract

Machine learning is increasingly being explored as a tool for improving the diagnosis of deep endometriosis. This study investigates its potential for distinguishing between active and fibrotic lesions using radiomic and clinical data extracted from 3D MRI scans. The goal is to develop a classification model that could assist in the diagnostic process by analyzing lesion characteristics automatically.
The dataset consists of 3D MRI scans from 61 patients, in which active and fibrotic lesions were segmented. From these images, radiomic features were extracted either from the 3D volume of each segmented lesion or from their corresponding 2D slices. In parallel, clinical data of the patient, were linked to each lesion based on the patient in whom it was found.
This approach allowed for the construction of six different datasets: 3D Radiomic, containing radiomic features extracted from the full 3D volume of each lesion; 3D Clinical, which includes the clinical data of the patient associated with each specific lesion; and 3D Radiomic Clinical, which integrates both radiomic and clinical features. The same structure was applied to the 2D-based datasets, resulting in 2D Radiomic, 2D Clinical, and 2D Radiomic Clinical, where the features were extracted from the corresponding 2D slices of the 3D lesion volume.

A first analysis was conducted using dimensionality reduction techniques such as Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and Pairwise Controlled Manifold Approximation Projection (PaCMAP). However, no clear separation between active and fibrotic lesions was observed, suggesting that the selected features may not contain enough distinctive information to naturally differentiate the two lesion types.
The classification task was then approached using automated machine learning (AutoML) tools, specifically Tree-based Pipeline Optimization Tool (TPOT), to optimize classification pipelines for each of the six different datasets.
The best performance was achieved with the 3D Radiomic Clinical dataset, which combines both radiomic and clinical features, reaching a balanced accuracy of 0.65 ± 0.19 and AUC of 0.60 ± 0.20. These results were based on 10-fold cross-validation, and the performance variability across different splits is reflected in the error values.
All other datasets showed poor predictive performance, with balanced accuracy values close to 0.5.
A significant misclassification of fibrotic lesions was also observed, likely due to the imbalanced dataset (25 fibrotic lesions vs. 64 active lesions). A limitation of the study was the relatively small sample size (61 patients, 89 total lesions), which made it more

challenging to achieve robust results and harder to characterize lesion differences.

These findings highlight the challenge of differentiating between active and fibrotic lesions in deep endometriosis using a combination of radiomic and clinical features. Future research should explore additional radiomic features and alternative modeling approaches to improve classification performance and enhance the diagnostic potential of these methods.

Contents

Introduction

Endometriosis is a chronic gynecological disorder affecting around 10% of women worldwide. It is characterized by the growth of endometrial-like tissue outside the uterus, leading to inflammation, fibrosis, and lesion formation. Among its different forms, deep infiltrating endometriosis (DIE) is the most severe, as lesions penetrate deeper than 5 mm into surrounding tissues.

MRI is a key imaging tool for diagnosing DIE, and in particular, T2-weighted sequences allow for a clear visualization of lesion characteristics. However, traditional MRI analysis relies on clinicians' expertise, which can lead to variability in diagnosis. This has led to growing interest in computational techniques such as radiomics and machine learning to enhance diagnostic accuracy.

Machine learning (ML) enables feature extraction and classification, offering a promising approach to differentiating lesion types in DIE. Radiomics, which extracts imaging features related to texture, shape, and intensity, provides additional insights into tissue characteristics that may not be visually apparent. Integrating radiomic data with ML models could improve the distinction between endometriotic lesions. In particular, they can be used to characterize and distinguish active and fibrotic endometriotic lesions, with active lesions being associated with inflammation, while fibrotic lesions are denser and less metabolically active.

This study aims to classify active and fibrotic DIE lesions using 3D MRI scans from 61 patients with endometriotic lesions, developing a machine learning based diagnostic tools for endometriosis to offer a standardized method for lesion classification. To achieve this, the study first focuses on extracting radiomic features from MRI images that are then combined with patient clinical data, which includes patient history, symptoms, and surgical information, to explore whether integrating both sources of information improves classification accuracy.
Machine learning pipelines are implemented to automate the process of feature selection, model training, and classification and are then tested and compared to identify the most effective approach for lesion classification. The performance of these models is assessed using cross-validation and classification metrics such as balanced accuracy, precision, re-

call, and F1-score.

The first chapter 1 introduces the study, explaining the clinical challenges of DIE, the role of MRI, and the potential of machine learning in lesion classification. The second chapter 2 explains the machine learning tools available and the and ML applications in medical imaging. The third chapter 3 details the dataset and methodology, including MRI images preprocessing, feature extraction, machine learning models, and evaluation techniques.
In the fourth chapter 4 the results of the study are described, comparing different classification models and feature sets, with an analysis of model performance and feature importance. Lastly, conclusion are drawn in the 5 chapter.

Chapter 1

Introduction

1.1 Deep Infiltrating Endometriosis

Endometriosis is a chronic gynecological disorder that affects approximately 10% [1] of women worldwide.
This condition is characterized by the presence of tissue similar to the lining of uterus, known as endometrium, that grows outside the uterine cavity. The endometrium consists of glands embedded in a supportive tissue called stroma, and its structure changes with age and throughout menstrual cycle due to its sensitivity to the hormones estrogen and progesterone.
In endometriosis, these glands and stroma (1.1) grow in area where they are not normally located such as the ovaries, fallopian tubes, or the pelvic regions. These ectopic tissues (tissues located in abnormal positions), continue to respond to hormonal changes, leading to inflammation, pain, and scarring, which result in various symptoms and complications.

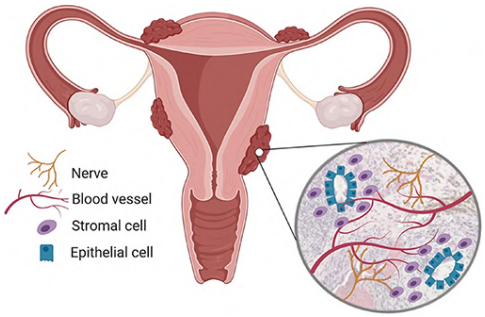


Figure 1.1: *Representation of an endometriotic lesion, highlighting the presence of endometrial stromal cells, epithelial glands, blood vessels, and nerve fibers. F rom: Hogg, Chloe and Horne, Andrew W. and Greaves, Erin, (2020), Endometriosis-Associated Macrophages: Origin, Phenotype, and Function*

Endometriosis is particularly challenging as abnormal growths of endometrial tissue differ at a molecular level from the normal endometrial tissue that lines the inside of

the uterus. This difference complicates the development of effective treatments, as the ectopic tissue behave differently, even though both types respond to hormonal changes. As a result, managing symptoms and finding therapies that work for all individuals affected by the condition remain significant challenges.

Endometriosis can be categorized into several types based on the location and extent of the disease. The Endometriosis Foundation of America suggests classifications according to the anatomical location of endometriosis inside the pelvic and abdominal cavities: superficial peritoneal endometriosis (SPE), ovarian endometriomas (OMA), and deep infiltrating endometriosis (DIE).
SPE is the most common form and involves lesions on the surface of the organs within the pelvic and abdominal cavities. OMA are cysts filled with menstrual blood that develop in the ovaries [2].
DIE is the most advanced and challenging form of the disease, with an incidence of about $1 - 2\%$ [3] in the global female population. It is associated with very severe pain in over 95% of cases, which is often accompanied by fertility issues. While the condition is relatively rare, its impact is debilitating for those affected. [4]
In figure below, the three types of endometriosis lesions are illustrated, along with their typical locations within the body, providing a clear representation of where each lesion is most commonly found.

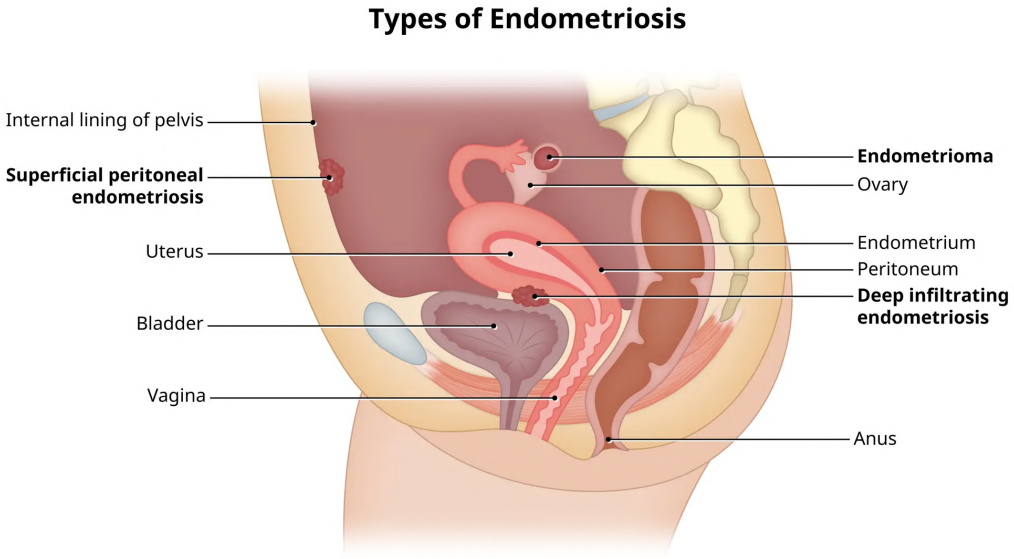


Figure 1.2: *Image showing the typical locations of the three types of endometriosis lesions: SPE, OMA, and DIE, as classified by the Endometriosis Foundation of America (EFA). From: IHH Healthcare Singapore, - Gleneagles Hospital - https://www.gleneagles.com.sg/conditions-diseases/endometriosis/symptoms-causes*

DIE occurs when endometrial tissue grows deeper than 5 millimeters into the pelvic organs. DIE is typically found in the posterior compartment of the pelvis, affecting

areas such as the vagina, rectum, uterosacral ligaments, and ureters, although it can also involve the anterior compartment, such as the bladder.
A sagittal view of the pelvic organs, showing their relative positions within the pelvic cavity, is provided in Figure 1.3 to help illustrate these anatomical locations.

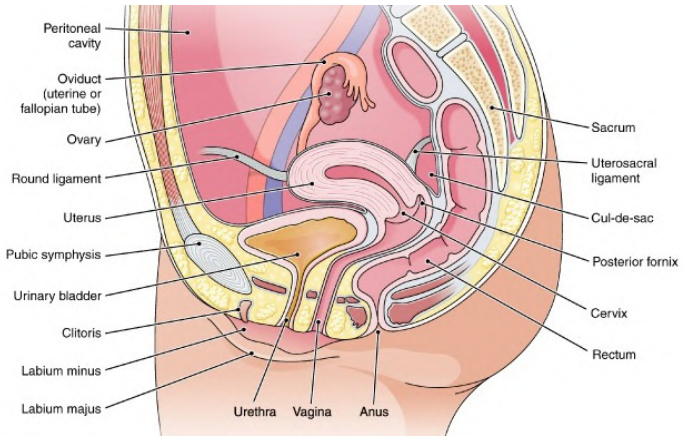


Figure 1.3: *Sagittal view of the pelvic organs, showing their relative positions within the pelvic cavity. From: Cohen BJ., (2007), Medical Terminology: An Illustrated Guide, Lippincott Williams & Wilkins*

Recently DIE has also been described as "adenomyosis externa". Adonomyosis is a condition where endometrial tissue grows within the muscular wall of the uterus. However, in "adenomyosis externa", this endometrial tissue grows outside the uterus, forming a single nodule that is typically larger than 1 cm in diameter. Unlike typical DIE lesions, which tend to be more widespread, adenomyiosis externa presents as a more compact, isolated mass. This makes it harder to treat surgically, as the lesions are often deeply embedded in the pelvic organs.
In severe cases, DIE may result in "adenomyosis externa", a condition in which the pelvic organs become stuck to each other due to scar tissue or adhesions, caused by ongoing inflammation from endometriosis. These adhesions make the pelvic organs less flexible and can cause significant pain, especially during ovulation and menstrual periods [5]. In the worst cases, adhesions can become so widespread that they severely restrict movement of the pelvic organs.

Deep endometriosis shows two main types of lesions, each with different characteristics based on their activity, stage of the disease and response to treatment. Active lesions, often referred to as "red"lesions, are typically marked by their strong blood supply or by being blisters filled with a reddish fluid, showing that the disease is actively inflaming the area. These lesions, observed in the early active stage of endometriosis, appear as small raised spots or fluid-filled bubbles with visible blood vessels on their surface. They are attached to the surrounding tissue, and may contain enlarged glands surrounded by blood vessels, indicating that the disease is still actively growing.

Depending on the course of the disease or the course of treatment, some of these active lesions may develop or disappear over time. It has been observed that hormone therapies that reduce inflammation and slow growth are particularly effective in treating red lesions [6].
Fibrotic lesions, like the "blue-black plaques", appear when the disease has been present for a long time and has led to scarring or tissue hardening. These lesions are characterized by the accumulation of scar tissues, which makes the affected area rigid and often retracted. The blue-black plaques are typically the result of the chronic nature of the disease, where inflammation has slowed down, and the body has begun to repair the tissue. For this reason, these types of lesions tend to be more stable and generally do not shrink or disappear. Unlike active lesions, they do not respond well to hormonal therapy. In some cases, fibrotic lesions may require surgical intervention to remove deeply affected tissue.

The distinction between active and fibrotic endometriosis lesions is crucial for understanding disease progression and determining the appropriate treatment. Active lesions are potentially responsive to pharmacological treatments and may benefit from medical therapy, while fibrotic lesions are stable and do not respond to hormonal therapy. This differentiation is crucial for pre-operative evaluations as well as for guiding therapeutic choices, ensuring that treatment plans are customized for the specific type of lesion.

1.1.1 Symptoms

The chronic and inflammatory nature of DIE can significantly impact the quality of life for affected women. Symptoms such as severe pelvic pain, bowel dysfunction, and sexual discomfort often result in physical, emotional, and social burdens. DIE is also a leading cause of infertility, affecting 30-50% of women diagnosed with endometriosis. DIE is often characterized by severe, stabbing, and persistent pain, with patients also experiencing abdominal pressure or a sensation of heaviness that further decreases their quality of life [5]. For this condition symptoms can vary significantly based on the lesions' location. In over 95% of cases, DIE is associated with severe pelvic pain, often leading to significant discomfort during menstruation (dysmenorrhea), intercourse (dyspareunia), and bowel movements. [4]
Depending on the affected organs, DIE can lead to a range of specific symptoms:

- Bowel involvement: endometriosis affecting the bowel may lead to symptoms such as abdominal bloating, diarrhea, constipation, and blood in the stool. In severe cases, adhesions can cause the bowel to fuse with the vagina, leading to intense pain during sex or bowel movements. Dyschezia (painful defecation) is a common symptom when gastrointestinal structures are involved, particularly during menstruation [7].

- Bladder involvement: lesions affecting the bladder can cause urinary symptoms such as frequent urination, urgency, painful urination (dysuria), and hematuria (blood in the urine), particularly during menstruation [5].
- Uterosacral ligaments: lesions in the uterosacral ligaments often lead to deep dyspareunia, causing significant discomfort during intercourse.
- Rectovaginal area: in cases where the rectovaginal septum is affected, women may experience lower abdominal pain, vaginal discomfort, and painful bowel movements, sometimes accompanied by menstrual blood in the stool or diarrhea.
- Rare cases: though uncommon, endometriosis may also involve other areas, such as the pleura, leading to chest pain, dyspnea, or even hemoptysis (coughing up blood) during menstruation.

Despite the severity of the symptoms, early diagnosis and appropriate treatment can significantly improve a patient's condition and fertility prospects.

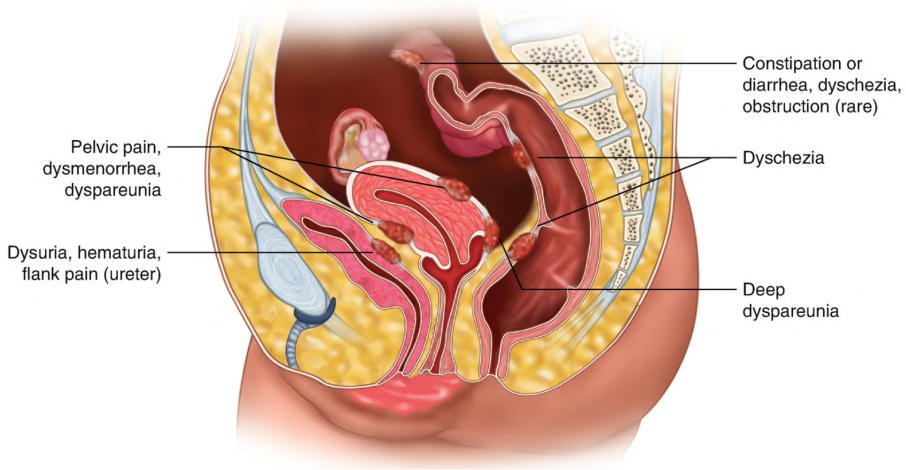


Figure 1.4: *Sagittal view of the pelvis, displaying typical endometriosis symptoms in relation to the location of DIE lesions; these symptoms often overlap. From: Salari, Salomeh and Coyne, Kathryn and Flyckt, Rebecca, (2022), Deep Infiltrating Endometriosis: Diagnosis and Fertility-Sparing Management in the ART Patient*

1.1.2 Diagnosis and treatment

Despite its significant impact on women's health, DIE remains challenging to diagnose due to its variety of symptoms.
Deep endometriosis should be suspected in all women presenting with chronic and debilitating pelvic pain, particularly dysmenorrhea, deep dyspareunia, dyschezia and cyclic rectal bleeding [4]. Although no single symptom is specific to DIE, certain symptoms tend to correlate with the location of the lesions. However, since lesion size and position

do not always reflect symptom severity, clinical awareness remains crucial for rapid diagnosis and effective treatment.
Late diagnosis is often the result of symptom misinterpretation, the masking effect of hormonal contraceptives, and the normalization of pelvic pain as part of the menstrual cycle [2].
A detailed medical history and physical examination are the first steps in diagnosing DIE. A gynecological exam may detect abnormalities in certain areas of the pelvis, but it is not always reliable since many lesions cannot be easily felt, making additional imaging tests necessary for confirmation.
Laparoscopy is the gold standard for the surgical diagnosis of endometriosis, providing direct visualization of endometrial tissue outside the uterus and enabling potential treatment of lesions. However, since laparoscopy is an invasive procedure, imaging techniques have become essential for identifying the locations of DIE lesions, suggesting that the role of diagnostic laparoscopy may need to be reconsidered, with imaging methods playing a more central role in the diagnostic process [8].
Transvaginal ultrasound (TVUS) is the first-line imaging-tool for diagnosis DIE [4], with reported sensitivity and specificity exceeding 85% [5]. However, its accuracy is highly operator-dependent, and its effectiveness varies based on lesion size and location [7].
In contrast, Magnetic Resonance Imaging (MRI) provides a more operator-independent approach and is valuable for detecting lesions in locations difficult to assess via ultrasound, such as the sigmoid colon and upper abdomen. MRI offers a detailed three-dimensional view of pelvic structures, allowing for the preoperative assessment of lesion size, lateral extension, and depth of invasion, which are essential for surgical planning.
The decision to perform surgery is mainly based on clinical evaluation and MRI offers a useful tool to support this decision [4].

Given the complex nature of DIE, a multidisciplinary approach is essential for accurate diagnosis and effective management. Collaboration between gynecologists, radiologists, surgeons, and fertility specialists ensures a comprehensive evaluation and tailored treatment planning, preventing unnecessary surgeries, reducing surgical complications, and improving fertility outcomes [2].

The treatment for DIE generally involves surgery to remove the endometriosis lesions and alleviate pain. This approach helps restore normal pelvic anatomy and can significantly improve patients' quality of life. Hormonal treatments after surgery are often recommended to reduce the risk of recurrence and to manage ongoing symptoms.
For patients facing fertility issues, the treatment plan should consider the possibility of pregnancy. In some situations, assisted reproduction techniques, such as in vitro fertilization (IVF), may be recommended, particularly when factors like previous surgeries or low ovarian reserve affect fertility. In general, addressing endometriosis surgically can improve the chances of pregnancy, whether naturally or through IVF [7] [9].

1.2 Nuclear Magnetic Resonance

Magnetic Resonance (MR) techniques represent a class of non-invasive methodologies widely used across various scientific and technological fields to investigate the properties of matter.
These techniques are based on the phenomenon of Nuclear Magnetic Resonance (NMR), which occurs when atomic nuclei, characterized by a non-zero spin quantum number I, interact with external magnetic fields through their intrinsic magnetic moment μ.
The magnetic moment μ of a nucleus is directly proportional to its spin I, as expressed by the relationship:

$$\mu = \gamma_n \frac{h}{2\pi} I \tag{1.1}$$

where γ_n is the gyromagnetic ratio of the n nucleus, h Planck's constant.
When a static and homogeneous magnetic field B_0 is applied, the interaction between μ and B_0, known as Zeeman interaction, induces a splitting of the nuclear energy levels into $2I + 1$ discrete levels.
This splitting causes the nuclei to redistribute among these energy levels according to the Boltzmann distribution, resulting in a population imbalance that, macroscopically, gives rise to a detectable bulk magnetization M.
For nuclei with spin $I = \frac{1}{2}$, such as hydrogen nuclei, the bulk magnetization can be represented as a vector in Cartesian coordinates, which at thermal equilibrium aligns along the direction of the applied magnetic field B_0, conventionally defined along the z-axis.
To detect this magnetization, it is necessary to perturb the system by transferring energy under the resonance condition. This means that the nuclear spins ensemble is exposed to an electromagnetic wave whose energy matches the energy difference between the Zeeman-split levels.
Thereby, to satisfy the resonance condition, a secondary oscillating magnetic field B_1 is applied orthogonally to B_0 at the specific frequency required for resonance. This radiofrequency (RF) pulse tilts the bulk magnetization vector away from its equilibrium position along the z-axis, resulting in the development of a transverse magnetization component in the xy-plane.[10]
After excitation, the system gradually returns to thermal equilibrium, with the magnetization vector regaining its equilibrium position. This vector has two components — longitudinal and transverse — which relax independently, each driven by distinct physical mechanisms and occurring over different timescales:

- Longitudinal Relaxation: the recovery of the longitudinal magnetization component (M_z) involves energy transfer between the nuclear spins and their surrounding lattice, a process known as spin lattice relaxation. Over time, M_z returns to its equilibrium value along the z-axis, with a rate of recovery characterized by the longitudinal relaxation time $T1$, which depends on the material's specific physical and chemical properties.

- Transverse Relaxation: the decay of the transverse magnetization component (M_{xy}) is driven by dephasing within the spin ensemble, an entropic process caused by interactions between individual spins and their local environments, referred to as spin-spin relaxation. This decay is described by the transverse relaxation time $T2$, which quantifies the rate of dephasing.

1.2.1 Magnetic resonance imaging in medicine

Medical imaging has significantly improved modern medicine by providing a variety of tools designed for different clinical purposes.
From methods like computed tomography (CT) and ultrasound to advanced nuclear medicine techniques such as positron emission tomography (PET) and single photon emission computed tomography (SPECT), these tools allow clinicians to visualize anatomical structures and assess functional or metabolic processes, with CT and ultrasound focusing on anatomical visualization and PET and SPECT assessing metabolic and functional activity, respectively. They are particularly helpful in cancer diagnosis, staging, and treatment planning, as well as in monitoring conditions in fields like heart diseases and brain disorders.

Among the various imaging techniques, MRI is particularly relevant due to its wide range of capabilities and non-invasive nature.
MRI uses strong magnetic fields and radio waves to generate images of internal structures, including organs, bones, muscles and blood vessels, without exposing patients to ionizing radiation.
The ability to provide detailed anatomical information, combined with insights into metabolic, functional and molecular properties of tissues, as well as its capacity to offer both qualitative and quantitative data, enhances its applicability in a wide range of clinical applications, including the detection, diagnosis, staging, and grading of various diseases [10].

The technology of an MRI clinical system relies on the principles of Nuclear Magnetic Resonance, which exploit the magnetic properties of atomic nuclei, specifically hydrogen atoms in the human body, to generate high-resolution images.

An MRI clinical system consist of several integrated components that work in together to produce these images. At its core lies the magnet, which generates the strong and uniform static magnetic field (B_0) around the patient. This field aligns the hydrogen protons within tissues, creating the conditions necessary for imaging. Clinical MRI scanners typically operates at field strengths of $1.5\,T$ or $3\,T$, with research system reaching $7\,T$ or higher. The strength of the magnetic field is a crucial parameter, as it directly influences image resolution and the signal-to-noise ratio (SNR): stronger field results in superior image quality, enabling the visualization of finer anatomical details.
Superconducting magnets are the most common type of magnet used in clinical MRI

system, requiring cooling to cryogenic temperatures with liquid helium to maintain their superconducting properties.
These magnets are housed in a Faraday cage to shield the system from external electromagnetic interference.
To maintain the homogeneity of the static magnetic field B_0, shim coils are used to generate localized corrective magnetic fields that compensate for inhomogeneities in B_0, which may arise due to imperfections in the magnet itself or the introduction of the patient into the scanner.

For spatial information encoding of the MRI signal, gradient coils are employed. These coils generate rapidly changing magnetic fields along the three axes (x, y, z), which are superimposed on the static B_0 field, enabling the localization of signals in three dimensions.
The role of radiofrequency coils is equally crucial: these coils transmit RF pulses (B_1) that excite the protons, temporarily displacing them from their alignment with B_0. As the protons relax back to equilibrium, they emit RF signals that are detected by the same or separate RF coils.
In some cases, contrast agents, typically containing gadolinium, are administered to patients to enhance differentiation between healthy and pathological tissues. These agents are able to accelerate the relaxation processes of protons, producing brighter and more detailed images [11].

The intrinsic properties of tissues, such as $T1$ and $T2$ relaxation times and proton density (PD), play a crucial role in determining the signal intensity in an MRI image, influencing how bright or dark tissues appear. However, by using different pulse sequences, it is possible to manipulate these intrinsic parameters to highlight specific tissue characteristics. Through the careful design of radiofrequency (RF) pulses and gradient waveforms, MRI systems can be modified to emphasize one property over another. Commonly used pulse sequences, such as $T1$-weighted ($T1$w), $T2$-weighted ($T2$w), Proton Density-weighted (PDw), and advanced variations like Fluid-Attenuated Inversion Recovery (FLAIR), allow for the visualization of different tissue properties or abnormalities thus the sequence selection depends on the diagnostic requirements.
The two main sequences employed are $T1$-weighted and $T2$-weighted.
$T1$-weighted is mainly used to highlight anatomical structures, producing images where tissues with high fat content appear bright (hyperintense), while fluid-filled areas, such as the brain's ventricles, appear dark (hypointense). $T1$-weighted imaging provides clear, detailed images of anatomy, making it ideal for assessing tissue composition and structural integrity.
Instead, T2-weighted imaging relies on the transverse relaxation time ($T2$) of tissues, which indicates how quickly protons lose phase coherence after an RF pulse.
Tissues with high water content, such as areas affected by edema or inflammation, maintain their transverse magnetization longer, resulting in a bright (hyperintense) signal on

$T2$-weighted images. In contrast, tissues with lower water content, like fat and bone, exhibit shorter $T2$ relaxation times and appear dark (hypointense).
This sensitivity to water content makes $T2$-weighted imaging particularly effective for visualizing pathological processes and, those conditions that increase water content within tissues - inflammation, cysts, edema, or tumor - are easily identifiable on $T1$-weighted scans.

An important extension of $T2$-weighted imaging is the Fluid-Attenuated Inversion Recovery (FLAIR) sequence, which is able to suppress the strong signal due to fluids that might otherwise obscure other signals. This suppression enhances the visibility of hyperintense lesions near fluid-filled spaces, such as edema, tumors, and other abnormalities, allowing for a more accurate diagnosis [10] [12].

Figure 1.5: *MR images of the pelvis: On the left, the $T1$-weighted image and on the right, the $T2$-weighted image. In both sequences, veins and arteries appear dark. Fat is bright in $T1$-weighted and also bright in $T2$-weighted , though less intense. The urinary bladder appears dark in $T1$-weighted and bright in $T2$-weighted, while ligaments are darker in $T1$-weighted. Bone is dark in both sequences, and the bone marrow shows intermediate brightness in both $T1$-weighted and $T2$-weighted. Muscles have an intermediate brightness in $T1$-weighted and appear more intermediate dark in $T2$-weighted. From: https://mrimaster.com/t1-vs-t2-mri/*

1.3 MRI in deep endometriosis

MRI is an important tool for the diagnosis of deep endometriosis due to its ability to produce high-resolution images of soft tissues, and assess the extent and location of the type of lesion.
The European Society of Urogenital Radiology (ESUR) has established recommendations for the optimal MRI protocol and diagnostic criteria for pelvic endometriosis, providing standardized guidelines to improve diagnostic accuracy.
MRI helps the identification of key pathological features associated with DIE by detecting signal intensity and morphological abnormalities that correspond to hemorrhagic lesions, fibrosis, or tissue masses. Hemorrhagic lesions typically appear as hyperintense foci (localized regions of high signal intensity) on $T1$-weighted images, reflecting the presence of blood caused by the cyclic bleeding of ectopic endometrial tissue. Fibrotic

areas, on the other hand, exhibit signal intensity similar to pelvic muscles on both $T1$-weighted and $T2$-weighted sequences and may or may not exhibit an enhancement after the administration of gadolinium-based contrast agents.
Morphologic abnormalities vary by anatomical compartment. For instance, in the posterior compartment, the torus uterinus and uterosacral ligaments (USLs) are the most frequently affected structures, presenting with mass thickening or nodular formations with stellate or irregular margins.

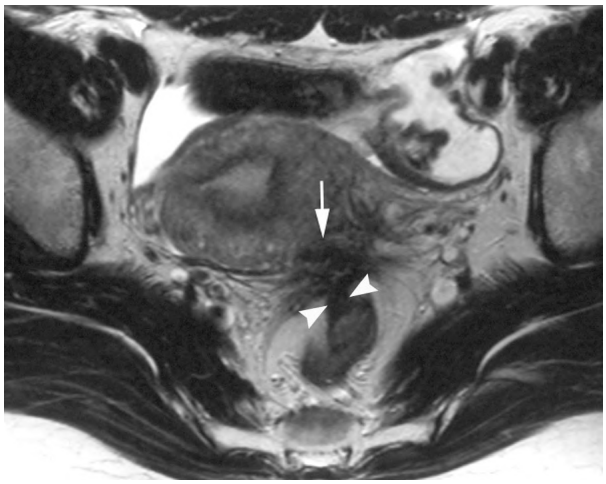


Figure 1.6: *Transverse MRI $T2$-weighted image showing irregular tissue area (related to fibrosis), with signal intensity close to that of pelvic muscle in the torus uterinus and USLs (arrow). Thickening of anterior rectal wall, which forms obtuse angle with normal wall, suggests rectal wall involvement (arrowheads). From Bazot M, (2004), Deep pelvic endometriosis: MR imaging for diagnosis and prediction of extension of disease*

Vaginal involvement is typically associated with the obliteration of the hypointense signal of the posterior vaginal wall on $T2$-weighted images, accompanied by thickening or mass formation. Similarly, rectosigmoid endometriosis is characterized by anterior displacement of the rectum toward the torus uterinus, thickening of the anterior rectal wall, and occasional hyperintense foci on $T2$-weighted or fat-suppressed $T1$-weighted images, often better delineated with contrast enhancement.

The obliteration of the pouch of Douglas and parametrium involvement are also key MRI indicators, with the latter presenting as low-signal-intensity areas in T2-weighted images.
For the anterior compartment, bladder endometriosis is identified by a nodule or hypointense mass near the viscouterine pouch. MRI also helps assess intestinal involvement in DIE, which commonly affects the sigmoid colon and rectum. Lesions in these areas may appear with or without adhesions to the posterior wall of the uterus, often leading to structural distortion.
Finally, endometriotic lesions in the round ligament are associated with fibrotic thicken-

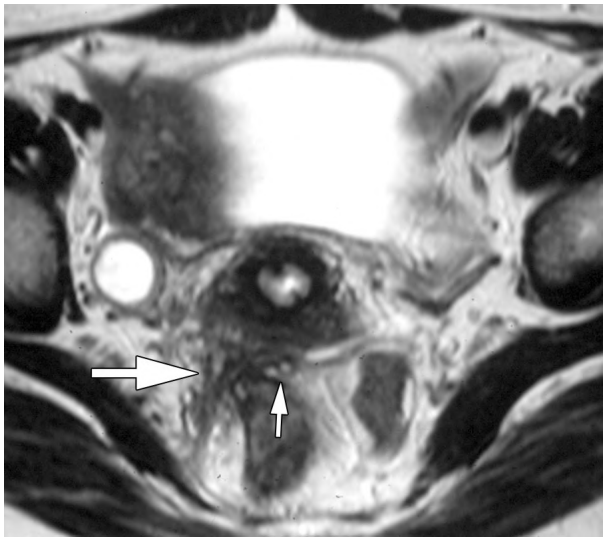


Figure 1.7: *Transverse T2-weighted MR image demonstrates irregular solid area with signal intensity close to that of pelvic muscle (thick arrow) at patient's right; area contains foci of high signal intensity (thin arrow) between lateral part of rectovaginal septum and rectal wall. From Bazot M, (2004), Deep pelvic endometriosis: MR imaging for diagnosis and prediction of extension of disease*

ing exceeding 1 cm, with either regular or irregular margins.

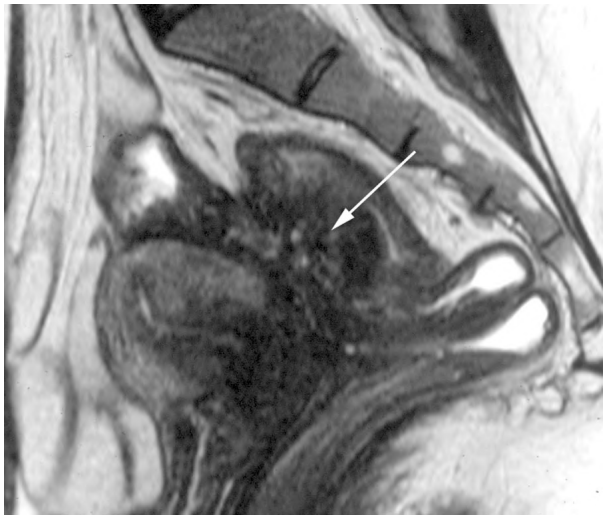


Figure 1.8: *Sagittal T2-weighted MR image showing fibromuscular lesions of endometriosis (arrow) that contain hyperintense foci in intestinal wall and extend into posterior wall of uterus. From Bazot M, (2004), Deep pelvic endometriosis: MR imaging for diagnosis and prediction of extension of disease Caption*

These MRI features are crucial for accurately diagnosing DE, guiding clinical decision-making and optimizing patient management [13] [8].

Chapter 2

Machine Learning

2.1 Introduction to machine learning

Machine learning (ML) is a branch of artificial intelligence (AI) which uses algorithms that learn from data to make predictions.
ML algorithms are able to detect patterns in data and use mathematical models to approximate real-world problems.
Depending on its purpose, ML can be used to describe past events by analyzing data to explain what happened, predict future outcomes by using data to estimate what will happen, or recommend the best actions to take based on the data.
Machine learning methods are generally categorized into three main types [14][15][16]:

- Supervised learning: type of ML where the goal is to learn a mathematical mapping between an input space and an output space. Since the relationship between these spaces is unknown, the algorithm is trained on a dataset containing input-output pairs, allowing it to estimate the correct output for new, unseen inputs. This process is known as prediction.
  Supervised learning is divided into classification and regression, depending on the nature of the output space: if the outputs are discrete values, the task is classification, and the model is a classifier, while if the outputs are continuous variables, the task is regression, and the model is a regressor.
  The most common models for classification include Decision Trees, Linear Classifiers, and Support Vector Machines, while for regression, they include Logistic Regression, Linear Regression, and Polynomial Regression.
  The strength of supervised learning lies in its ability to generalize from known examples, meaning it can make accurate predictions for inputs it has never seen before. This is achieved through an automated learning process in which the algorithm, guided by labeled data, refines its ability to produce the desired outputs without human intervention.

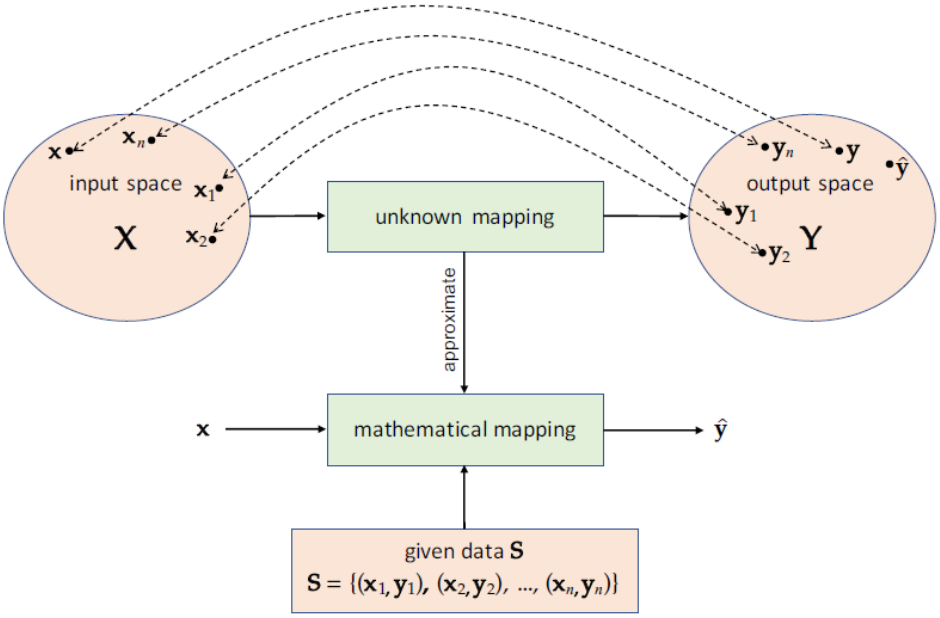


Figure 2.1: *Supervised learning: the data consists of a set S containing elements from the input space $(x_1, x_2, ..., x_n)$ and their corresponding values from the output space $(y_1, y_2, ..., y_n)$. The goal is to learn a mathematical mapping using S that can estimate (predict) the corresponding the output value for any given input element. Zollanvari A., (2023), Machine Learning with Python: Theory and Implementation*

- Unsupervised learning: type of ML where the algorithm is trained only on input data without corresponding output labels. The goal is to extract meaningful structures or patterns from the data defining an output space based on the specific task.
  One of the most common applications is clustering, where the algorithm groups similar observations, creating a partition of the dataset based on shared characteristics. This approach is widely used in image segmentation, a technique in digital image processing that divides an image into multiple regions based on pixel properties, such as color or shape. In medical imaging, for example, segmentation is essential for identifying and labeling regions corresponding to tumors or other anatomical structures.
  Unsupervised learning also includes a variety of dimensionality reduction techniques, which seek to project data into a lower dimensional space while preserving essential information. This is particularly useful when dealing with dataset that have many features, making analysis and visualization more efficient.
  Key models for unsupervised learning include K-means clustering, Hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Principal Component Analysis.

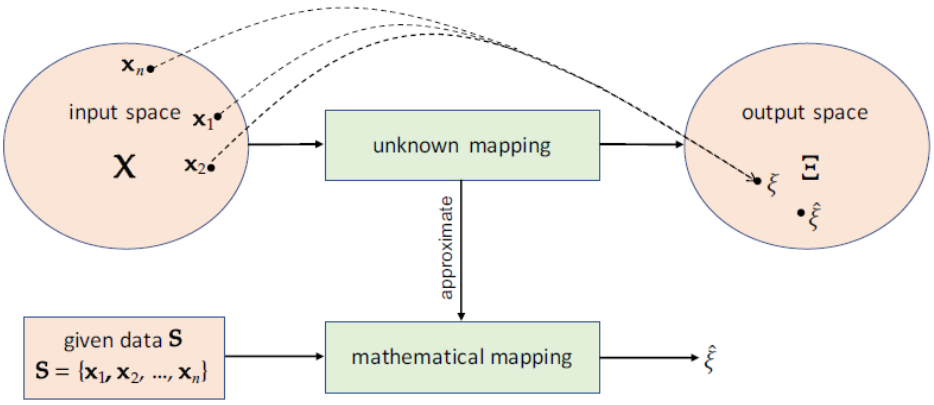


Figure 2.2: *Unsupervised learning: the data consists of a set S containing elements from the input space* $(x_1, x_2, ..., x_n)$. *The goal is to learn a mathematical mapping using S that can project S to an elements of the output space. Zollanvari A., (2023), Machine Learning with Python: Theory and Implementation*

- Semi-supervised learning: it connects supervised and unsupervised learning by using both labeled and unlabeled data. In this case, part of the dataset consists of input-output pairs, similar to supervised learning, while the remaining data is unlabeled, requiring the model to detect patterns without direct instruction. This method is particularly useful when labeled data is limited or difficult to acquire, but a large amount of unlabeled data is available.
  One important technique in this area is self-training, where a model first learn from the labeled data and then progressively refines its predictions by assigning labels to the unlabeled data.
  This approach minimizes the need for manual labeling while still enabling the model to use a larger dataset.

2.1.1 Data splitting and generalization

A fundamental step in ML is splitting the dataset into a training set and a test set, where the training set is used to fit the model, while the test set evaluates the model's ability to generalize to unseen data. This process is essential to assess the generalization performance of the model, ensuring that it does not merely memorize the training data but can also make accurate predictions on new inputs, simulating in this way real-world conditions, where the model is applied to new cases.
One common issue that arises during training is overfitting, which occurs when a model becomes too complex relative to the amount of information available in the dataset. An overfitted model performs exceptionally well on the training set but fails to generalize, leading to poor performance in new data. This happens because the model captures noise and specific pattern in the training data.
On the other hand, underfitting occurs when the model is too simple to capture the

underlying structure of the data. An underfitted model fails to learn relevant patterns even within the training data, resulting in poor performance across both training and test sets.
There is a trade-off between model complexity and generalization, where increasing complexity improves training performance but, beyond a certain threshold, begins to degrade test performance. The goal is to find an optimal balance, often referred to as the sweet spot, where the model is complex enough to learn meaningful patterns but not so complex that it memorizes noise [14].

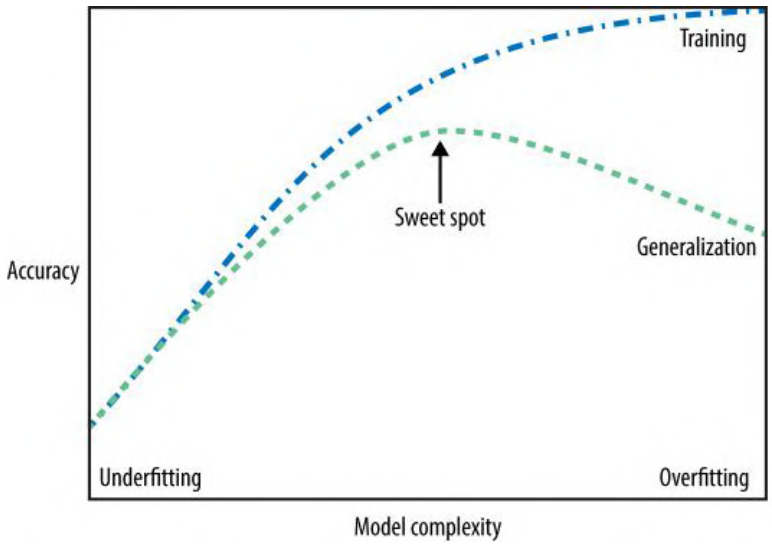


Figure 2.3: *Trade-off of model complexity against training and test accuracy. From Müller, A.C. and Guido, S., (2016), Introduction to Machine Learning with Python: A Guide for Data Scientists*

2.1.2 Preprocessing and feature engineering

ML datasets consist of samples (e.g., patients in a medical study) and their corresponding features (e.g., age, lesion size). However, raw data is rarely suitable for immediate use in model training. Before feeding a dataset into a machine learning algorithm, it must undergo preprocessing to ensure that the data is clean, standardized, and properly formatted for analysis. This includes steps such as handling missing values, normalizing numerical features, and encoding categorical variables to improve consistency and comparability.
Beyond standardization, not all features contribute equally to a model's predictive accuracy. Some may be redundant, irrelevant, or even introduce noise, negatively impacting performance.
To improve model efficiency and generalization, dimensionality reduction techniques are applied to retain the most informative aspects of the data while eliminating unnecessary complexity. These techniques are generally classified into two key steps:

- Feature Extraction, which consists of transforming raw or high-dimensional data into a structured set of features. This step is particularly crucial when dealing with unstructured data, as it enables the quantification of meaningful variables.
- Feature Selection,which follows feature extraction and involves identifying the most informative features while discarding those that do not contribute significantly to the model's performance. This process may also involve combining features to enhance their predictive power.

These preprocessing and feature engineering steps are crucial for enhancing model performance, reducing computational costs, and mitigating overfitting, ensuring that the data used for training is both optimized and meaningful.

Feature scaling

One of the most common preprocessing techniques is feature scaling. It ensures that different features are comparable in scale, and well suited for the model's learning process, preventing it from being biased toward features with larger magnitudes. This process can involve:

- Standardization: using method like *StandardScaler*, where features are transformed to have zero mean ($\mu = 0$) and unit variance ($\sigma = 1$).
- Normalization: Rescaling values within a fixed range ($[0, 1]$ or $[-1, 1]$).

To ensure consistency and avoid cross-contamination of information, it is essential to apply the same process using the same indicators: the mean and standard deviation calculated from the training set must be used to standardize the test set. [14].

Feature selection

Feature selection identifies the most informative features in a dataset, reducing overfitting, improving interpretability and enhancing computational efficiency. Selection criteria can be based on priori knowledge or data-driven methods which exploit statistical and algorithmic techniques that rank features by statistic importance in predicting the target variable or contribution to the model's performance. As it can be observe in figure 2.4 below, feature selection techniques are typically classified into three categories.

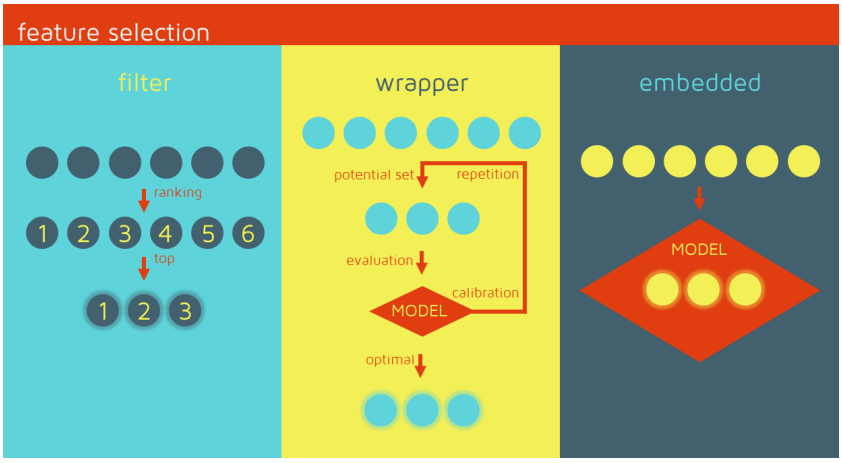


Figure 2.4: *Feature selection: selects a subset of relevant features while keeping the original feature space intact. From ETS Asset Management Factory, (2019), Machine learning, what is the difference between feature extraction and feature selection?*

1. Wrappers
   Wrapper methods iteratively evaluate different feature subset by training a model and measuring its performance using evaluation rules along with evaluation metrics (accuracy, ROC-AUC curve). The goal is to find the subset of features the yields the best results. Common wrapper techniques include:

   - Sequential Feature Search (SFS): iteratively adds features one by one, selecting those that improve model performance.
   - Sequential Backward Search (SBS): starts with all features and removes them gradually to find the optimal subset based on the evaluation metric score.

2. Filter Methods
   Filter methods rank features based on their statistical properties, independently of any learning algorithm. Since these methods do not involve model training, they are computationally efficient but may not account for interactions between features.
   A common filter-based selection technique is Mutual Information (MI). MI measures the dependence between two variables. Given two discrete random variables x and y with a joint probability distribution $p(x, y)$, MI is defined as:

$$I(X, Y) = \sum_{x,y} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \tag{2.1}$$

   where:
   $p(x)$ and $p(y)$ are the marginal probabilities of x and y
   $p(x, y)$ is the joint probability of x and y occuring together

MI values range from 0 to positive infinity, where zero implies that there is no relationship between the feature and the target variable, while as high is MI as informative is the feature for the prediction task [17].

3. Embedded Methods
   Embedded methods incorporate feature selection directly into the learning algorithm itself. In some cases, once the model is trained, it can provide an intrinsic measure of feature importance, depending on the type of algorithm used. These methods are often more computationally efficient than wrapper methods and tend to generalize well. Embedded techniques can be further categorize as iterative or non iterative, where an important iterative method is Recursive Feature Elimination (RFE) that recursively removes the least important features based on model-specific criteria. The process continues until an optimal subset of features is identified, often determined via cross-validation.

Feature extraction

Feature extraction, unlike feature selection, transforms the original feature space into a new lower-dimensional feature space, while preserving as much information as possible. This precess is particularly useful for high-dimensional datasets, where many features may be redundant or highly correlated.
Feature extraction includes techniques as Principal component Analysis (PCA), Linear Discriminant Analysis (LDA), Uniform Manifold Approximation and Projection (UMAP) and Pairwise Controlled Manifold Approximation Projection (PaCMAP).

Principal component Analysis (PCA)

PCA is an unsupervised dimensionality reduction techniques used to transform a high-dimensional dataset into a smaller set of linearly uncorrelated variables, called principal components (PCs).
Unlike feature selection methods that retain a subset of the original features, PCA creates new features by forming linear combinations of the existing ones. The primary objective is to remove redundant variables and focus on combinations of variables that preserve most of the informative content of the dataset. This is particularly useful in high-dimensional datasets where many features are highly correlated, leading to redundancy and increased computational complexity.
The principal components are ordered based on the amount of variance they explain:

- The first principal component (PC1) captures the highest variance in the dataset.

- The second principal component (PC2) captures the next highest variance while remaining orthogonal (uncorrelated) to PC1.
- Each subsequent component explains progressively less variance than the previous one.

To ensure that all features contribute equally, PCA requires data standardization, especially when the features have different units or scales (e.g., "age" in years vs. "lesion size" in millimeters).
Once standardized, the covariance matrix is computed to quantify how different features vary together, as it captures the relationships between the two. If the dataset contains d features, the covariance matrix is a $d \times d$ symmetric matrix where each element represents the covariance between two features. The covariance between two features x^i and x^j is given by the following equation:

$$c_{ij} = cov(x^i, x^j) = \frac{1}{n} \sum_{k=1} (x_k^i - \mu_i)(x_k^j - \mu_j) \tag{2.2}$$

where:
c_{ij} covariance value between feature i and feature j
n total number of samples in the dataset
x_k^i value of the i-th feature for the k-th sample
x_k^j value of the j-th feature for the k-th sample
μ_i mean of the i-th feature across all n samples
μ_j mean of the j-th feature across all samples

The sign of the covariance provides insight into these relationships: a positive covariance ($c_{ij} > 0$) indicates that as one feature increases, the other tends to increase as well, suggesting a direct relationship between the two variables. Conversely, a negative covariance ($c_{ij} < 0$) signifies an inverse relationship, meaning that when one feature increases, the other is more likely to decrease. If the covariance is zero, it implies that the two features are uncorrelated, so variations in one do not influence the other.
To determine the principal components, PCA computes the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors define the directions of the new feature space, while the eigenvalues indicate how much variance each principal component captures. The eigenvectors corresponding to the largest eigenvalues represent the directions of maximum variance, forming the new axes along which the data is projected.
By selecting the top k eigenvectors, where k is much smaller than the original number of dimensions d, PCA reduces the dataset's dimensionality while preserving as much variance as possible.
The transformed dataset retains its most significant patterns, making it useful for machine learning tasks such as clustering, classification, and visualization. Reducing dimensionality not only improves computational efficiency but also enhances generalization by

removing irrelevant or redundant features. When projected into two or three dimensions, the data can also be visualized more easily, helping to identify underlying structures and patterns [18].

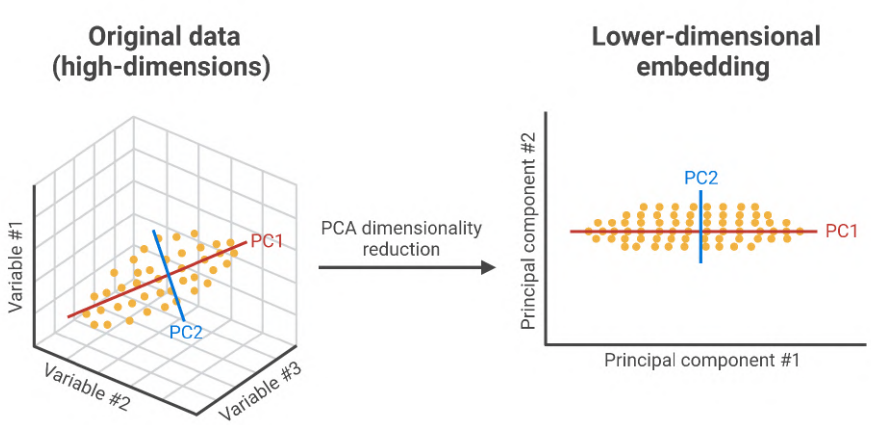


Figure 2.5: *PCA transformation: from original high dimensional feature space to low dimensional features space. From: Mina Nashed, https://www.biorender.com/template/principal-component-analysis-pca-transformation*

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction technique used for classification tasks. Unlike Principal Component Analysis (PCA), which is an unsupervised method that focuses on maximizing variance, LDA aims to find the best projection that maximizes the distance between different classes while minimizing the spread of samples within the same class. This means that while PCA looks for directions where the data is most spread out, LDA looks for directions where different classes are best separated.
The algorithm works by reducing the dimensionality of the original high-dimensional space by projecting the data onto a lower-dimensional subspace, where classification becomes more effective. This is done by identifying a new set of axes that maximize the separation between classes.
LDA achieves this by following two key principles: it maximizes the distance between class means, ensuring that the classes are well separated, while simultaneously minimizing the variance within each class, keeping data points belonging to the same class together. This transformation enhances class separability.
For binary classification, this process results in a one dimensional projection (a single line), while for multi-class classification, LDA can find up to $C - 1$ dimensions, with C is the number of classes.
LDA algorithm constructs two scatter matrices that measure how data varies within

each class (within class scatter matrix S_W (2.3)) and between different classes measures how much the class means differ from the overall mean (between class scatter matrix S_B (2.4).

$$S_W = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (x_i - \mu_c)(x_i - \mu_c)^T \tag{2.3}$$

where:
C total number of c classes
N_c total number of samples within each c class
μ_c mean vector for c
x_i sample of class c

Each term $(x_i - \mu_c)(x_i - \mu_c)^T$ represents the deviation of a sample from its class mean. The sum accumulates this variance across all classes and high S_W value means that the data points within a class are spread out, making classification harder.

$$S_B = \sum_{c=1}^{C} N_c(\mu_c - \mu)(\mu_c - \mu)^T \tag{2.4}$$

where:
μ overall mean of all samples
N_c weights the contribution of each class based on the number of samples

Each term $(\mu_c - \mu)(\mu_c - \mu)^T$ measures how far a class centroid is from the overall mean: for high S_B values the class centroids are well separated, a desirable results for classification.

Once the matrices are initialized, the LDA algorithm determines the optimal direction for projecting the data by computing eigenvectors and eigenvalues. The top eigenvectors, corresponding to the largest eigenvalues, are selected to construct the projection matrix. Using this matrix, the original dataset is mapped onto a new subspace where class separation is maximized, making classification more effective.
By transforming the data into a space where distinctions between classes is more defined, LDA enhances the performance of classification models, establishing itself as a valuable tool in machine learning.

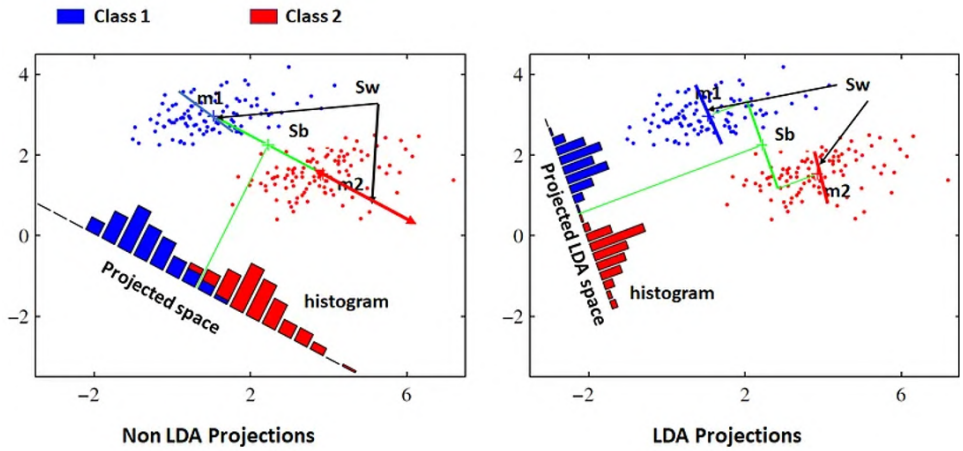


Figure 2.6: *Each color represents a different class, with m_1 and m_2 the mean of Class 1 and Class 2, respectively. The left plot illustrates data projections onto the line connecting m_1 and m_2, resulting in significant overlap between the two classes. In contrast, the right plot shows the data projected using LDA, effectively reducing class overlap and improving separation. From Christopher M. Bishop, (2011), Pattern Recognition and Machine Learning*

Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a non linear dimensionality reduction techniques that preserves local structure while maintaining some aspects of the global structure. It is particularly effective for visualizing high dimensional data in two or three dimensions trying to retain meaningful patterns present in the original space. Unlike other techniques which relies on linear projections, UMAP build a graph based representation of the high-dimensional data and optimizes a corresponding low-dimensional graph to approximate its structure.
The algorithm consists of two primary phases: graph construction in the high dimensional space and graph optimization in the low dimensional space.
The first step is based on the construction of a weighted nearest-neighbor graph in the original high-dimensional space, where for each point x_i the algorithm tries to find its k nearest neighbors using a defined distance metric. To account for variations in local density, a parameter ρ_i is computed for each point, ensuring that neighborhoods are properly defined even in regions of varying density.
After this step a scaling parameter σ_i is determined by solving:

$$log_2(k) = \sum_{j=1}^{k} w(x_i, x_j) = \sum_{j=1}^{k} exp\left(-\frac{max(0, Distance_{i,j} - \rho_i)}{\sigma_i}\right) \tag{2.5}$$

where:
$x_{i,j}$ data point in high-dimensional space
k nearest neighbors of x_i

ρ_i the minimum positive distance from point x_i to any of its k k-nearest neighbors
$w(x_i, x_j)$ weight function, used to defined to quantify the connection strength between points
$Distance_{i,j}$ distance between observations i and j in the original high-dimensional space.
σ_i parameter that normalizes distances between each point i and its neighbors j to preserve relative proximities.

The result is a weighted nearest-neighbor graph, where edges between points represent probabilistic relationships derived from their distances in the high-dimensional space.
The final connection strength between two points x_i and x_j is adjusted to ensure that their relationship is represented in both directions. This step helps create a stable graph that accurately reflects the data's underlying patterns. In UMAP, this means that if point x_i is a neighbor of point x_j, but x_j was not initially a neighbor of x_i, the algorithm ensures the connection is mutual. This mutual reinforcement between points strengthens the graph, making it a more accurate representation of the high-dimensional relationships.
Once the high-dimensional graph has been constructed, UMAP seeks to create a corresponding low-dimensional graph that preserves the relationships encoded in the original graph. This is achieved through an iterative optimization process that adjusts the positions of the points in the reduced space.
It initializes the points in the low-dimensional space and their positions are then refined through a series of updates that apply attractive force, which pulls connected points closer together, and repulsive forces, which pushes apart points that are not connected in the original high-dimensional graph, preventing unrelated points from collapsing together in the low-dimensional representation.
In this way, the relative distances and cluster separation observed in the high-dimensional space are preserved [19].

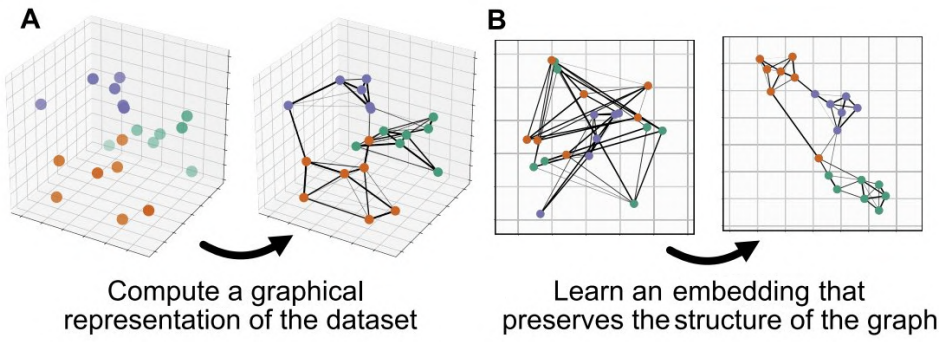


Figure 2.7: *UMAP: (A) the first stage of the UMAP algorithm is to compute a probabilistic graphical representation of the data. (B) The second step optimizes the embeddings to maintain the structure of the graphical representation. From Tim Sainburg, Leland McInnes, Timothy Q Gentner, (2020), Parametric UMAP embeddings for representation and semi-supervised learning*

Pairwise Controlled Manifold Approximation Projection (PaCMAP)

Pairwise Controlled Manifold Approximation Projection (PaCMAP) is a dimensionality reduction algorithm able to preserve both global and local structure in a way more effective with respect to methods like UMAP. UMAP in fact enhances local structure that can in some cases lead to distorted embedding, while PaCMAP controls the relationship between different types of point pairs to ensure an actual representation of high-dimensional data.
In order to accomplish this, PaCMAP dynamically adjusts the importance of different pairwise relationships during the optimization process, where the algorithm initially define a global layout and then gradually refine local structures. The result is a robust and computationally efficient method that adapts well to various data distributions.
PaCMAP is based on the idea that high-dimensional data contain various levels of different meaningful relationships, and it processes the dataset to keep close points together in the low-dimensional embedding, preserve the separation of distant points to maintain global structure, and use moderately distant points to help define the overall shape of the dataset.
It structures the data as a graph, where edges represent relationships between points in the high-dimensional space, and can be divided into three categories: neighbor pairs (close points), mid-near pairs (moderately distant points), and further pairs (widely separated points), each of them with a different role in shaping the embedding.
PaCMAP optimizes a loss function. In machine learning, a loss function quantifies how well a model's predictions match the true values. It acts as an objective function that the algorithm seeks to minimize during training. A well-designed loss function ensures that the learned representation or predictions are as close as possible to the ground truth.
Here the loss function is designed to minimize distances between similar points while maximizing separation between dissimilar ones to preserve high dimensional relationships in the low dimensional one. The total loss function in PaCMAP consists of three components:

$$\begin{aligned} Loss_{PaCMAP} = w_{NB} \sum_{neighbors(i,j)} \frac{\tilde{d}_{ij}}{10 + \tilde{d}_{ij}} + w_{MN} \sum_{mid-near(i,k)} \frac{\tilde{d}_{ik}}{10000 + \tilde{d}_{ik}} + \\ + w_{FP} \sum_{further(i,l)} \frac{1}{1 + \tilde{d}_{il}} \end{aligned} \tag{2.6}$$

where:
(i, j) neighbor pairs
(i, k) mid-near pairs
(i, l) further pairs

w_{NB} weight for neighbor pairs, w_{MN} weight for mid-near pairs, w_{FP} weight for further pairs
$\tilde{d}_{ab}$ distances in the low-dimensional space: $\tilde{d}_{ab} = ||y_a - y_b||^2 + 1$ with y_a and y_b low-dimensional representations of points a and b

The first term is the loss function of neighbor pairs and it ensures that nearby points remain close in the low dimensional space: when $\tilde{d}_{ij}$ is small (i.e., points are close), the loss decreases, while as $\tilde{d}_{ij}$ increases the loss approaches 1, preventing distant neighbors from being treated as equally important as close ones.
The second term applies to mid-near pairs (moderately distant points). Since the denominator in this term is larger than in the neighbor loss, its impact is lower. This means that mid-near points influence the embedding only slightly, helping to define the overall structure without dominating the optimization process.
The third term controls further pairs (distant points) by penalizing small distances between them. If two far-apart points are mapped too closely in the low-dimensional space, the loss is high, encouraging them to remain well-separated. Conversely, when these points are already far apart, the loss is minimal, meaning no additional penalty is applied.

To refine the embedding effectively, PaCMAP dynamically adjusts the weights of these terms in three optimization phases:

1. Global phase: the weight of mid-near pairs (w_{MN}) is high, emphasizing global structure and preserving large-scale relationships.

2. Transition phase: the weight of neighbor pairs (w_{NB}) increases while w_{MN} decreases, shifting the focus toward local neighborhood refinement.

3. Local refinement phase: the mid-near term is removed ($w_{MN} = 0$), while neighbor and further pair weights refines cluster boundaries and prevent overlap.

By using adaptive weights and a three-stage optimization process, PaCMAP effectively preserves both local and global structure, making it a powerful tool for high-dimensional data visualization [19].

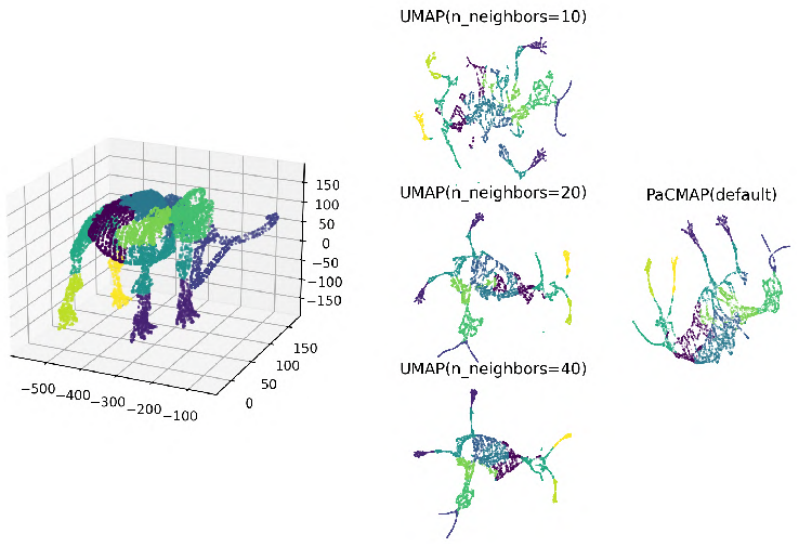


Figure 2.8: *Application of UMAP and PaCMAP to the Mammoth dataset. In judging these figures, one should consider preservation of both local structure (e.g., mammoth toes) and global structure (overall shape of the mammoth). From Yingfan Wang, Haiyang Huang, C. Rudin, Yaron Shaposhnik, (2020), Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization*

2.1.3 Model evaluation

Assessing the performance of a predictive model is a critical step in machine learning, as it provides insight into how well the model generalizes to unseen data. A model that performs well on training data but fails to make accurate predictions on new samples is unlikely to be useful in real-world applications. Therefore, model evaluation is not only about measuring performances but also about ensuring generalization, and so the model ability to maintain accuracy when applied to previously unseen inputs.
Model evaluation plays also a key role in model selection, where the most appropriate model must be chosen among several candidates. By applying evaluation techniques, it is possible to compare models objectively and select the one that best balances predictive accuracy and robustness. To achieve this, model evaluation relies on both evaluation metrics, which quantify performance, and evaluation rules, which define how these metrics are estimated.

One of the most used evaluation techniques is cross-validation.
Cross validation is an extension of the train-test split method, which divides the dataset into two subsets: one for training and the other for evaluation. While simple, this approach can lead to unreliable estimates of model performance due to its dependence on a single train-test partition. Cross-validation accounts for this problem by repeatedly splitting the data into multiple training and validation sets thereby providing a more stable and representative estimate of model's generalization ability.

By evaluating the model across different train-test splits, cross-validation can reveal whether a model is too dependent on the training data, which would lead to poor performance on new samples.
The basic form of cross-validation is k-fold cross -validation, where the dataset is divided into k equally sized subsets, called folds. The model is trained and evaluated k times, each time using $k-1$ folds for training and the remaining fold for testing. The process is repeated until each fold has been used as the test set exactly once, ensuring that every data point is used for both training and testing, leading to a more robust performance estimate. At the end of the procedure, multiple evaluation scores are obtained, which can be aggregated (e.g., by computing the mean and standard deviation) to assess the model's stability and sensitivity to data variations [14].
Different cross-validation techniques exist to handle specific characteristics of the dataset:

- Stratified k-Fold cross-validation: ensures that each fold maintains the same proportion of classes as in the original dataset, which is particularly useful for imbalanced classification problems.
- Group k-Fold cross-validation: used when the dataset contains grouped observations (e.g., multiple samples from the same subject in medical applications). To prevent data leakage, all samples from the same group must be assigned to either the training or test set, but never both. This is particularly important when the goal is to develop models that can be applied to entirely new groups, such as new patients in a medical study.

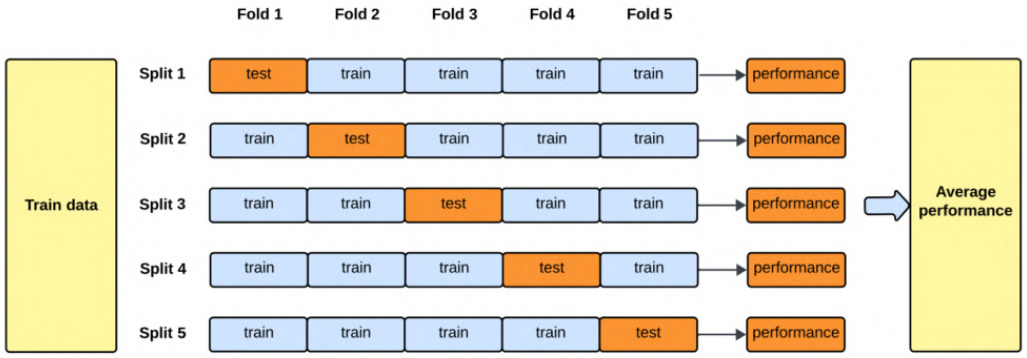


Figure 2.9: *Standard k-fold cross validation. From Haden Pelletier, (2023), Towards Data Science: Cross Validation with Time Series Data by author*

Evaluating a classification model requires the use of various metrics that quantify its ability to distinguish between different classes. These metrics help asses how well

the model generalizes to unseen data and whether it makes reliable predictions.
The foundation of many classification metrics is the confusion matrix, which provides a structured way to analyze the model's performance in terms of correctly and incorrectly classified instances.
For a binary problem this matrix is a 2×2 matrix where rows represent predicted labels, so the output give by the model, and the columns represent the actual label, the true class values.
Along the diagonal elements of the matrix there are correct classifications, while off-diagonal elements represent misclassifications. Given a dataset with two classes, positive (P) and negative (N), the confusion matrix contains the following elements:

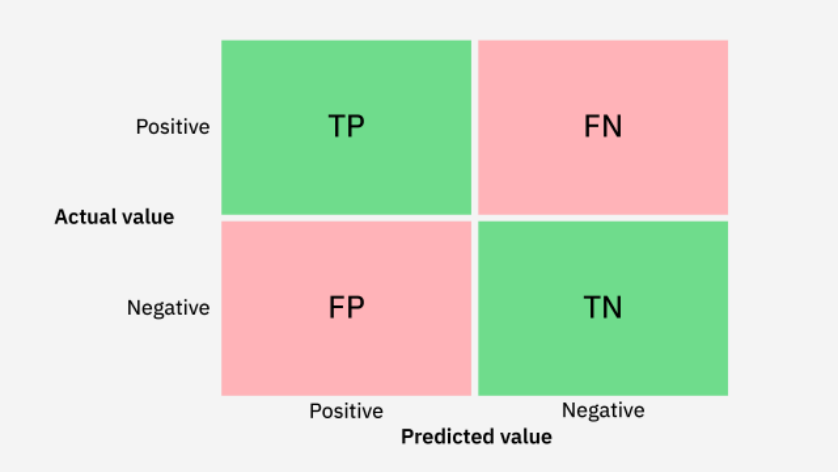


Figure 2.10: *Confusion Matrix: top-left cell represents the True Positives (TP), indicating the number of correctly predicted positive cases. The bottom-left cell corresponds to False Positives (FP), where negative cases are incorrectly classified as positive—also. The top-right cell shows the False Negatives (FN), where actual positive cases are mistakenly predicted as negative. The bottom-right cell contains the True Negatives (TN), representing the correctly identified negative instances.From Jacob Murel Ph.D., Eda Kavlakoglu, (2024), Che cos'è una matrice di confusione*

Each cell corresponds to a different evaluation metrics used to analyze the classifier's performance, that are:

- True Positive Rate (TPR) / Sensitivity / Recall: measures how many actual positive instances are correctly classified by the model. In medical applications, this is referred to as sensitivity:

$$t\hat{p}r = \frac{n_{TP}}{n_P} = \frac{n_{TP}}{n_{TP} + n_{FP}} \tag{2.7}$$

- False Positive Rate (FPR): measures the proportion of negative instances that are incorrectly classified as positive:

$$f\hat{p}r = \frac{n_{FP}}{n_N} = \frac{n_{FP}}{n_{TN} + n_{FP}} \tag{2.8}$$

- True Negative Rate (TNR) / Specificity: measures how many actual negative instances are correctly classified:

$$t\hat{n}r = \frac{n_{TN}}{n_N} = \frac{n_{TN}}{n_{TN} + n_{FP}} \tag{2.9}$$

- False Negative Rate (FNR): measures the proportion of positive instances that the model fails to classify correctly:

$$f\hat{n}r = \frac{n_{FN}}{n_P} = \frac{n_{FN}}{n_{TP} + n_{FN}} \tag{2.10}$$

Other metrics can be derived to evaluate the model's performance:

- Precision (Positive Predictive Value, PPV): represents the proportion of instances predicted as positive that are actually positive:

$$p\hat{p}v = \frac{n_{TP}}{n_{TP} + n_{FP}} \tag{2.11}$$

- False Discovery Rate (FDR): measures the proportion of instances predicted as positive that actually belong to the negative class:

$$f\hat{d}r = 1 - p\hat{p}v = \frac{n_{FP}}{n_{TP} + n_{FP}} \tag{2.12}$$

- F1-Score: provides a single metric that balances precision and recall, defined as their harmonic mean. This is particularly useful when dealing with imbalanced datasets:

$$f_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \frac{2recall \times precision}{recall + precision} \tag{2.13}$$

- Accuracy: represents the proportion of correctly classified instances across all samples. However, accuracy can be misleading when dealing with imbalanced datasets, as a model may achieve high accuracy simply by predicting the majority class:

$$acc = \frac{n_{TP} + n_{TN}}{n_{FP} + n_{FN} + n_{TP} + n_{TN}} \tag{2.14}$$

- Balanced Accuracy: version of accuracy that considers both class distributions. It is calculated as the average of sensitivity (recall for the positive class) and specificity (recall for the negative class):

$$balancedaccuracy = \frac{sensitivity + specificity}{2} \tag{2.15}$$

- Error Rate: Represents the proportion of misclassified instances, complementary to accuracy:

$$ErrorRate = 1 - accuracy \tag{2.16}$$

In many classification models, predictions are made based on a probability score, and a threshold is applied to decide whether an instance belongs to a certain class. For example, a model might classify an instance as positive if its predicted probability is greater than 0.5. However, this threshold can be adjusted, affecting the balance between true positives and false positives.
To analyze how a model performs across all possible thresholds, we use the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the TPR against the FPR for various threshold values. A well-performing model achieves a curve that is closer to the top-left corner, indicating a high recall with a low false positive rate.

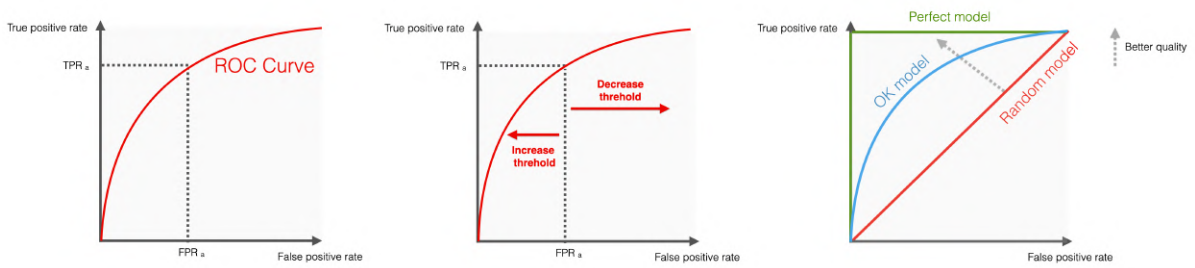


Figure 2.11: *ROC curve: relationship between FPR (x axis) and TPR (y axis) at different threshold values. As the curve apporaches a step function as the model makes precise classification. From Evidently AI Team, (2025), How to explain the ROC curve and ROC AUC score?*

To summarize the ROC curve into a single value the Area Under the Curve (AUC) is computed, which quantifies the model's ability to distinguish between classes: an AUC value of 1 represents a perfect classifier, while as it approaches 0.5 as the prediction made by the model is randomic.

2.1.4 Pipelines

Machine learning models are built through a sequence of steps, from data preprocessing to model training and evaluation. Managing these steps individually can be inefficient, especially when multiple models or data processing techniques need to be tested. To simplify this process, machine learning pipeline provide a structured way to connect these steps into a single automated workflow.
A pipeline organizes the different phases of machine learning into a predefined sequence, where each step takes input from the previous one and passes its output to the next. This approach ensures that data transformations, model selection and evaluation follow a consistent procedure, reducing the risk of errors and making it easier to experiment

with different configurations.
Pipelines also improve reproducibility, allowing models to be trained and tested in the same way each time, regardless of changes in the dataset or parameters.
Another important key advantage of pipelines is their modularity: each component, whether a data preprocessing step, a feature selection method, or a machine learning model, can be adjusted or replaced independently without disrupting the entire workflow. This flexibility allows for systematic testing of different techniques, helping to identify the most effective combination for a given problem.
Pipelines also make it easier to maintain and update models over time. Once a pipeline is set up, new data can be automatically processed and analyzed, reducing the need for manual intervention [20].

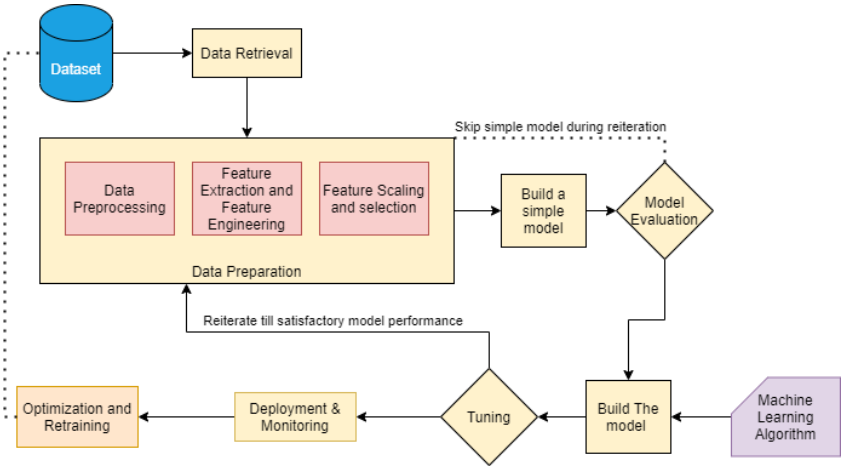


Figure 2.12: *High-level workflow of a standard machine learning pipeline. From Prakshaal Jain, (2021),Standard Machine Learning Pipeline*

2.2 Machine learning in medicine

Machine learning is becoming increasingly important in medicine, offering new ways to improve disease diagnosis, treatment planning and overall healthcare management. By analyzing large amounts of data, machine learning models can identify patterns, make predictions and assist clinicians in making more accurate decisions.
One of the main application of ML in medicine is disease diagnosis: doctors rely on their experience and medical tests to diagnose illnesses, but ML can process vast amount of data and detect subtle patterns that might go unnoticed. This is particularly useful in fields like radiology, where models trained on X-rays, MRIs and CT scans can recognize tumors or other conditions with high accuracy.
Another key area is disease prediction, where through the analysis of patients data ML helps doctors anticipate disease progression, complications, and hospital re-admissions,

allowing them to take preventive measures. In oncology predictive models help determine which treatments are most likely to be effective for individual patients, reducing side effects and improving success rates.
Also by analyzing a patient's medical history, genetic data, and lifestyle, machine learning models can identify the most effective treatment options, tailoring the treatment to the needs of each patient rather than following a standard approach [21].

In endometriosis ML has emerged as a promising tool for improving the detection, prediction and understanding of endometriosis, with applications in diagnostic model development, patient outcome prediction and research optimization.
Several studies [22] have exploited machine learning models to identify endometriosis using clinical symptoms, imaging data and biochemical markers, where many of these models employ logistic regression, decision trees, support vector machines (SVMs). These models incorporate variables such as age, history of pelvic pain, dysmenorrhea, infertility, and prior pelvic surgeries, allowing clinicians to assess risk levels more effectively.
As a predictive tool ML can give information on treatment responses and disease progression, estimating for example the success rates of fertility treatments in endometriosis patients, predict the likelihood of deep endometriosis (DE) versus other pelvic pain disorders, and assess the risk of post-surgical complications.
Despite challenges such as symptom variability, data limitations, and the lack of standardized diagnostic criteria, which complicate model development and reduce prediction accuracy, machine learning remains a promising tool for improving endometriosis detection and treatment. Continued research and better data quality can enhance its accuracy, leading to earlier diagnosis and more effective, patient-specific treatments [22].

Chapter 3

Materials and Methods

This study focuses on the implementation of a machine learning algorithm to classify fibrotic and active lesions in cases of deep endometriosis.
The following dataset and tools were used in this study:

- 3D MRI scans: 64 patients' 3D MRI scans in NIfTI format, acquired using T2-weighted fast spin echo (FSE) sequences. Corresponding segmentations of the lesions were also available.
- Clinical data: dataset including various clinical parameters, categorized into general patient characteristics, preoperative symptoms, surgical procedure details and locations, and postoperative symptoms.
  - The general characteristics (GC) included medical history data, such as weight, height, body mass index (BMI), and smoking status. Additionally, it considered whether the patient had prior pregnancies, had undergone hormonal therapy (including progestin-only pills (POP), combined oral contraceptives (COC), intrauterine systems (IUS), or gonadotropin-releasing hormone (GnRH) agonists), or had a history of abdominal surgery. The presence of dysmenorrhea was also recorded. The surgical indication was classified as pain, infertility, or organ dysfunction.
  - Preoperative symptoms included amenorrhea, chronic pelvic pain (CPP), dysuria, rectal bleeding, and bowel dysfunction, such as stipsi or diarrhea.
  - Surgical details and the location of the intervention were also considered, specifying where it was performed.
  - Postoperative symptoms focused on the presence or persistence of dysmenorrhea or amenorrhea.
- Software and libraries: Python 3.11, including libraries such as Pyradiomics 3.0.1 and Scikit-learn 1.5.2.

Dataset prepocessing

From the initial 64 MRI scans, three were excluded due to differences in acquisition protocols, as they were not focused on the pelvic region. The remaining 61 MRI scans, along with their respective lesion segmentations, were considered for analysis.
Anonymized Digital Imaging and Communications in Medicine (DICOM) files were retrieved from the hospital's Radiological Information System (RIS) and Picture Archiving and Communications System (PACS). Manual segmentation was performed, where lesion contours were manually outlined across all image slices to ensure accurate three-dimensional reconstruction using 3D Slicer version 5.6.1. The DICOM files were loaded into the software, and axial T2-weighted acquisitions (excluding post-contrast images when contrast was administered) were selected, with a slice thickness of 3 mm.
Then, the DICOM files were converted into the Neuroimaging Informatics Technology Initiative (NIfTI).

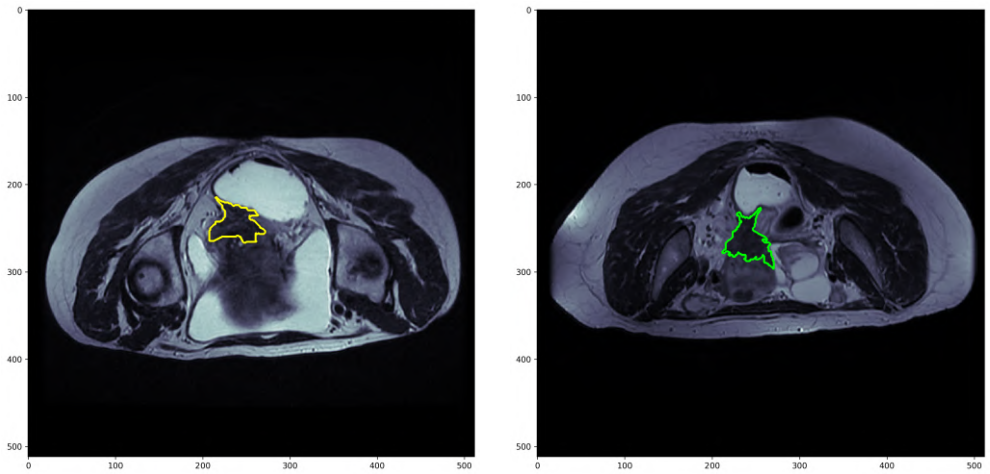


Figure 3.1: *MRI scans for patient 18 (slice 27) with fibrotic lesion (left), and for patient 4 (slice 27) with active lesion*

This study examines both 3D MRI scans and their corresponding 2D slices, focusing on the slices where the lesion was present. The inclusion of 2D images analysis was introduced to enhance lesion characterization in the axial plane. As illustrated in Figure 3.2, the number of samples increased significantly when considering individual slices. As can be observed the number of active lesions is much higher than the number of fibrotic ones, leading to imbalanced dataset.

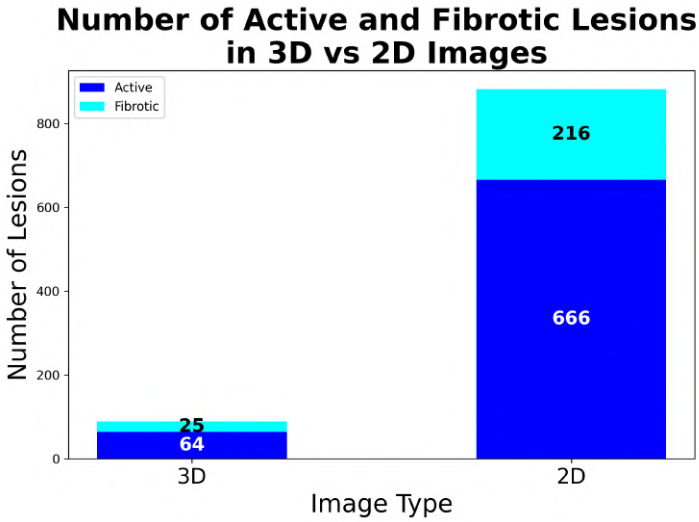


Figure 3.2: *Number of fibrotic and active lesions for 3D images vs their corresponding 2D slices*

Dataset construction

Radiomics aims to extract quantitative metrics, the radiomic features, within medical images, capturing morphological and textural characteristics that may be difficult to recognize or quantify by the human eye. In this study, extracted radiomic features include first-order statistics, texture-based features, and shape descriptors. These features were computed from both original and transformed images, where specific filtering techniques were applied to improve lesion analysis, such as wavelet decomposition and the Laplacian of Gaussian (LoG) filter. Wavelet filtering was used to divide the image into multiple frequency components, while the LoG filter, applied with sigma values of 0.5, 1.0, 1.5, and 2.0, emphasized local intensity variations, highlighting edges and texture patterns within the lesion.
The extracted first-order features describe the statistical distribution of pixel intensities within the region of interest (ROI), including metrics such as mean intensity, standard deviation, entropy, skewness, and kurtosis. These provide insight into the overall signal variability, but do not account for spatial relationships between pixels.
To capture other tissue characteristics, texture-based features were derived from different gray-level matrices, which analyze the spatial relationships between pixels or voxels [41]:

- Gray-Level Co-occurrence Matrix (GLCM): measures how often pairs of pixel intensities levels occur together in a specific pattern, showing variations in texture.
- Gray-Level Run-Length Matrix (GLRLM): measures the distribution of consecutive pixels with the same intensity along a given direction.
- Gray-Level Size Zone Matrix (GLSZM): evaluates the number and size of homogeneous pixel clusters, providing information on regional intensity variations.

- Gray-Level Distance Zone Matrix (GLDZM): similar to GLSZM, but also considers the distance between areas of similar intensities, providing a more detailed evaluation of lesion homogeneity.
- Neighborhood Gray-Tone Difference Matrix (NGTDM): captures the contrast between pixel intensities and their local neighborhood.

Radiomic features were extracted from both 3D MRI scans and 2D slices using the pyradiomics 3.0.1 library. For the 2D dataset, only the slices containing the lesion mask were selected. In cases where multiple regions of the same lesion appeared within the same slice, only the largest region was considered to maintain consistency.

In addition to radiomic features, clinical features were also incorporated. This led to the cration of six datasets:

1. 3D MRI datasets:
    - Radiomic features only
    - Clinical features only
    - Combined radiomic and clinical features
2. 2D slice datasets:
    - Radiomic features only
    - Clinical features only
    - Combined radiomic and clinical features

Lesions were categorized into two classes: active lesions (class 0) and fibrotic lesions (class 1).
The radiomic datasets contained a total of 1228 features for the 3D extraction and 842 features for the 2D extraction. The clinical dataset comprised 59 features.

Data Analysis

The first step of the analysis involved the evaluation of the mutual information between features and class labels to asses their relevance for distinguish between lesion types.

Instead, to visualize the dataset and determine whether the extracted features provided class separation, dimensionality reduction techniques were applied. In particular, as supervised approach, LDA was employed, also providing a measure of balanced accuracy, while as unsupervised approach PCA, UMAP and PaCMAP were selected. For all these techniques, scatterplots were generated. Additionally, for LDA, a Kernel Density

Estimation (KDE) plot was implemented to visualize the distribution of the two classes along the identified direction. Unlike traditional histograms, KDE estimates the probability density function (PDF) of the data, offering a smoothed representation that is less sensitive to bin size. Instead, KDE distributes the influence of each data point across a continuous range, providing a more accurate and representative view of the underlying data structure.

As a preliminary step, the datasets were scaled to provide a common scale for all features before applying these techniques.

Pipeline implementation

In order to perform classification Tree-based Pipeline Optimization Tool (TPOT) [39] was used. TPOT is an automated machine learning tool based used to optimize model selection and hyperparameter tuning. It employs a tree based approach where each pipeline consists of a sequence of machine learning operations structured in a tree format: the root node represents the final predictive model (e.g. a classifier or regressor), while the intermediate nodes correspond to preprocessing steps such as feature selection or dimensionality reduction. The leaf nodes contains hyperparameter values that define specific configurations of the algorithms.

TPOT identifies the best model by generating an initial population of random pipelines. In this study, 100 pipelines were created in the first generation. Each pipeline consists of different steps, including feature selection, preprocessing, model selection, and hyperparameter tuning, all implemented using the scikit-learn library.

Scikit-learn [40] is an open-source Python library that provides a wide range of tools for data preprocessing, model training, and evaluation, including classification, regression, and clustering algorithms, but also for feature scaling, encoding, selection.

Once TPOT generates the initial pipelines, the optimization process begins. In this study, 200 generations were used to refine the pipelines. This process involves cross-validation to evaluate model performance, followed by mutation and crossover operations to create new candidate pipelines
For the 3D dataset, Stratified 10-fold cross-validation was applied to ensure a balanced representation of classes across training and validation sets. For the 2D dataset, Stratified Group k-Fold cross-validation (k=10) was used to guarantee that all slices from the same patient were assigned exclusively to either the training or test set.
At the end of the process, only the best performing pipelines are carried for the next generation, where they are iteratively refined.

The TPOT drawback is its high computational cost. The optimization process can

take hours to days to complete, depending on factors such as dataset complexity and hardware capabilities. In this study, pipeline optimization required significant computational time, particularly due to the high-dimensional nature of the dataset.

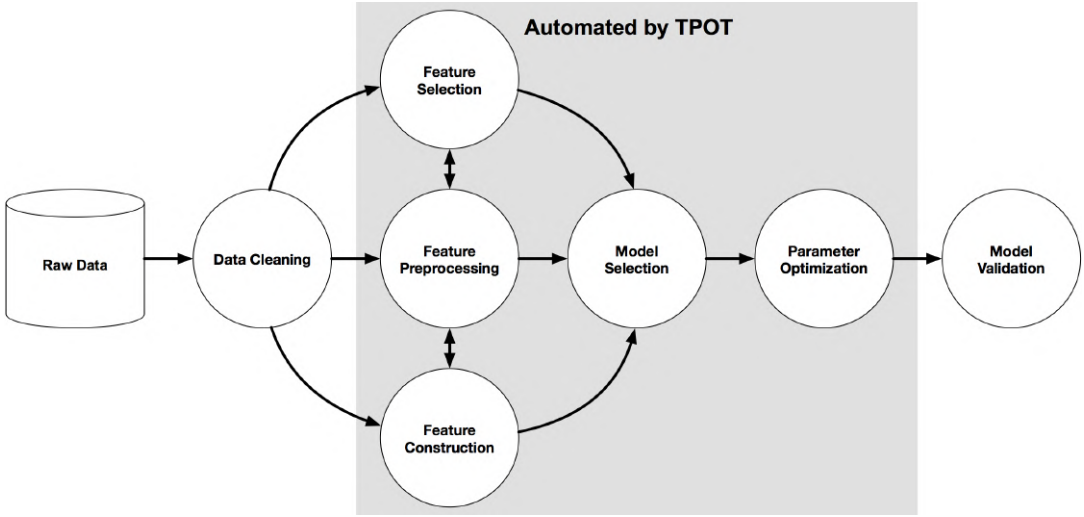


Figure 3.3: *This diagram illustrates the automated steps performed by TPOT. TPOT handles feature selection, preprocessing, construction, model selection, and parameter optimization, allowing to focus on data analysis and interpretation. Form https://epistasislab.github.io/tpot/*

The pipelines produced by TPOT varied significantly, incorporating different combinations of feature selection, feature extraction, and classification methods, each discussed below.

Preprocessing methods

- MinMaxScaler class of scikit-learn: scales features to a predefined range (typically $[0, 1]$).
- RobustScaler class of scikit-learn: reduces sensitivity to outliers by centering the data on the median value and scaling it based on the interquartile range (IQR).
- Kernel approximation: transforms input features into a higher-dimensional space. TPOT utilized two kernel approximation methods, based on scikit-learn classes:
  - Radial Basis Function (RBF) Sampler: approximate the Radial Basis Function (RBF) kernel, which is a Gaussian kernel, using random Fourier features.
  - Nystroem: it is an approximation technique that computes the kernel matrix only for a subset of seleted points. The smaller kernel is used then to approximate the entire kernel matrix.

Feature selection and engineering

- SelectPercentile class of scikit-learn: this method selects only the features with higher score based on a statistical test. TPOT selected ANOVA F-values to evaluate the relationship between each feature and the target class.

- FastICA class of scikit-learn: this algorithm computes indipendent component analysis and finds the features that maximize the statistical indipendence.

- Polynomial Features class of scikit-learn: it generates new features by computing polynomial combinations of existing ones.

Classification models

- Linear SVC [31]: machine learning algorithm that finds an optimal hyperplane that separates classes while maximizing the distance between the decision boundary and the nearest data points from each class, also known as margins. The main parameter is C that controls the margin's width: for high C values hard margins are defined, which try to find a decision boundary that perfectly separates the data into two classes without any misclassification, while for low C values, the so called soft margins allow for some misclassification.

- Decision Tree Classifier: it is a supervised learning algorithm with a tree structure composed of nodes and branches. It is a hierarchical model that recursively splits data into subgroups based on feature values. The tree consists of decision nodes, that are the points where the dataset is split based on condition on features, and leaf nodes, the terminal nodes that assign class labels based on the majority of samples reaching that node.
  The tree is constructed by initially choosing the best feature to split on, and then the dataset is iteratively split until reaching pure leaf nodes or a stopping criterion (e.g., maximum depth).
  If a leaf node contains mixed class samples, a majority vote determines the predicted class.

- Random Forest Classifier [33]: it is an ensemble method that constructs multiple decision trees and aggregates their predictions. It introduces randomness by training each tree on a random subset of the data, and selecting a random subset of features at each split, reducing correlation between trees.
  For classification, the final prediction is determined by majority voting across all the trees or by out-of-bag (OOB) error estimation, which assesses model performance on unseen excluded samples.

- Extra Trees Classifier: ExtraTreesClassifier is an ensemble learning methods also based on decision trees. It function similarly to Random Forest but in this case the feature splits are chosen randomly rather than based on optimal thresholds, producing more diversified trees.

- MLP Classifier [34]: the multilayer perceptron (MLP) is a neural network model able to learn non-linear relationships. It consist of an input layer, where each neuron represents an input feature, hidden layers, that consist of interconnected neurons that perform computations on the input data, applying weighted sums and non-linear functions, and the output layer, that generates the final prediction for classes. MLPs learn through backpropagation, adjusting weights to minimize the error between predicted and actual labels.

- SGD Classifier [35]: Stochastic Gradient Descent (SGD) is an optimization technique that randomly selects a single training sample to compute the gradient, and then updates the model parameters accordingly. The primary goal of SGD is to minimize a specified loss function. A loss function is an error function that quantifies the difference between the predicted outputs of a machine learning algorithm and the actual target values. In this case, TPOT chose the hinge loss, which is a type of loss function that penalizes both misclassified samples and correctly classified samples that are too close to the decision boundary. The hinge loss evaluates how closely a model's predictions align with the actual labels.

- Gradient Boosting Classifier [36]: Gradient Boosting is an ensemble learning method that builds a sequence of models, each correcting errors made by the previous one. Models are trained to minimize residual errors using a gradient descent approach, while the new models focus on samples that are hard to classify, improving in this way the accuracy.

- Naive Bayes Classifier (Multinomial and Bernoulli) [38]: Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem, which describes how to update the probability of a hypothesis based on new evidence. It assumes that all features are conditionally independent given the class label, which, despite being a strong assumption, often works well in practice, especially with high-dimensional data. In particular TPOT chose two particular types of Naive Bayes classifier: multinomial Naive Bayes, which assumes that features follow a multinomial distribution and computes class probabilities based on the frequency of each feature, and Bernouli Naive Bayes, typically used for datasets with binary features.

Also regularization techniques are incorporated into TPOT pipelines. These techniques change the model learning behavior during the training phase, in order to reduce

the overfitting. This is done by assigning a penalty term to the model, that increases with the model complexity.
In particular L1 and L2 regularizations are used. The first one adds penalty on the sum of the absolute values of the model's weights; this implies that weights that do not significantly contribute to the model will be set to zero, which can lead to automatic feature selection. The L2 regularization instead penalizes features with large coefficient values, in order to distribute equally feature importance.

In some pipelines, classifiers were used as stacking estimators, meaning their outputs were appended as new features rather than being used directly for classification.

The final pipelines are evaluated based on the balanced accuracy score, ROC-AUC, a classification report that includes precision, recall, and F1-score, and a normalized confusion matrix plot, with normalization based on true labels.

Chapter 4

Results and Discussion

This section presents the results.
Throughout this chapter, the term 3D Radiomic refers to the dataset containing radiomic features extracted from the 3D volume of each segmented lesion. Similarly, 3D Clinical includes the clinical information of the patient associated with each specific lesion, meaning that for every segmented lesion, the dataset contains the clinical data of the patient in whom that lesion was identified. The 3D Radiomic Clinical dataset combines both types of features.

The same structure applies to 2D datasets. 2D Radiomic consists of radiomic features extracted from individual slices of each lesion volume. 2D Clinical includes the clinical information of the patient related to each specific lesion slice, ensuring that for every analyzed lesion slice, the dataset contains the relevant patient's clinical data. Finally, 2D Radiomic Clinical integrates both radiomic and clinical features.

4.1 Feature Selection and Extraction

4.1.1 Mutual Information

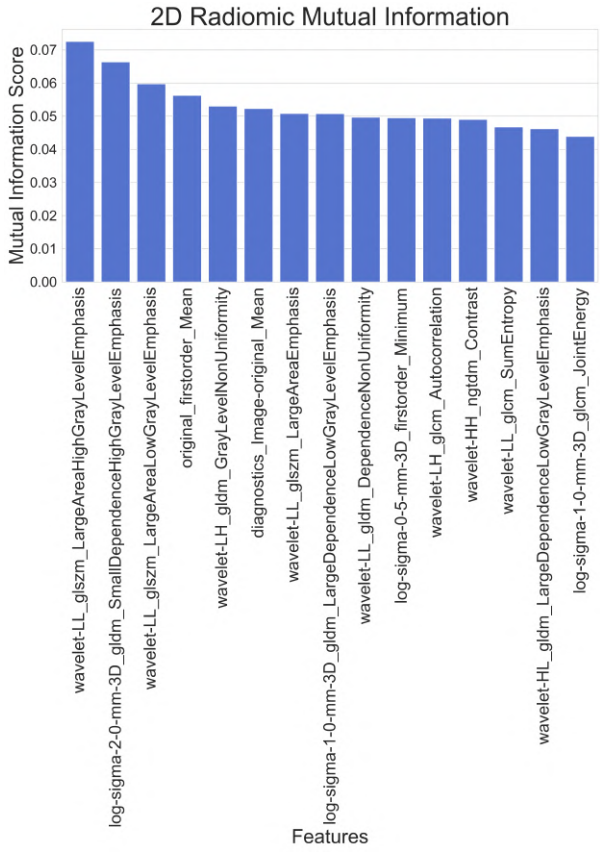
Mutual information was computed across all datasets, both 2D and 3D, to evaluate the dependency between features and class labels. The results are summarized in bar plots, Figure 4.2, displaying the top 15 features ranked by mutual information score.

In the 2D Radiomic Clinical dataset the most informative feature is GC:BMI (General Characteristic: Body Mass Index), with a mutual information score of approximately 0.25, significantly higher than the dataset average of 0.1. This trend is consistent with the 2D Clinical dataset, where BMI also emerges as the most relevant feature.
In contrast, radiomic features appear only from the sixth position, with mutual infor-

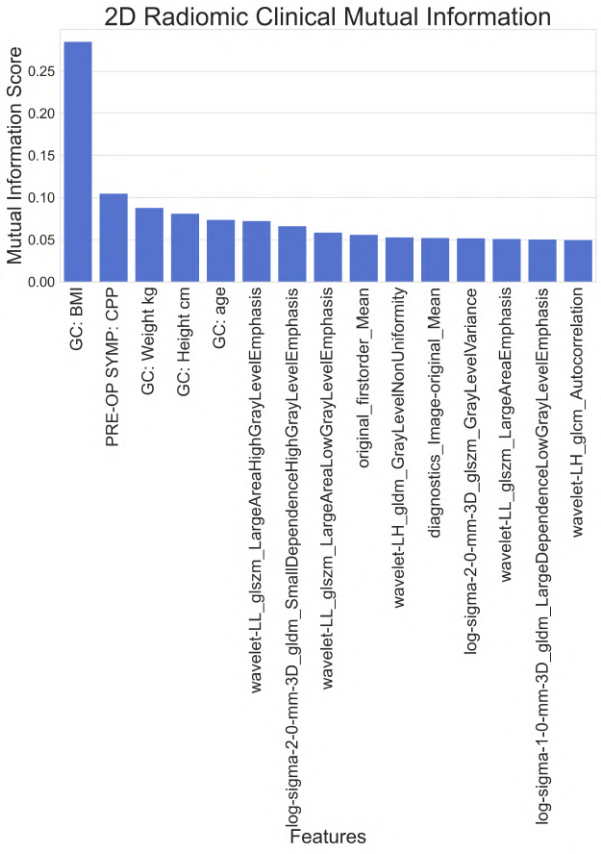
mation values ranging between 0.05 and 0.06, indicating that clinical features contribute more prominently to the classification task in this dataset.

The most informative radiomic feature in the 2D radiomic dataset is wavelet-LL-glszm_LargeAreaHighGreyLevelEmphasis, with a mutual information score of 0.07. This feature is also present in the combined clinical-radiomic 2D dataset on sixth position. This feature combines wavelet-transformed imaging and texture analysis to describe the intensity and spatial distribution of grey levels within a lesion. In particular, wavelet-LL transformation involves applying a low-pass filter along all three spatial axes (x, y, z), preserving only low-frequency components while removing high-frequency details such as sharp edges and fine textures.
The GLSZM method is used for texture analysis, quantifying how groups of neighboring pixels with the same grey level are distributed within a lesion. The LargeAreaHighGreyLevelEmphasis measure, extracted from the GLSZM, quantifies how much the image contains large, high-intensity homogeneous zones (regions where a significant number of connected voxels share a high grey-level intensity) [42]. The relevance of this feature suggests that heterogeneity in lesion texture and the presence of high-intensity may be important discriminative factors in the classification task.

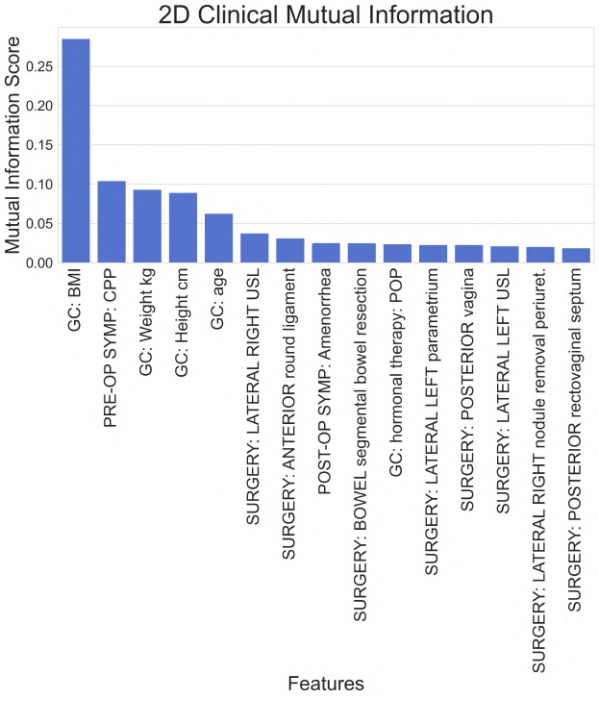
In the 3D Radiomic Clinical dataset clinical features are no longer among the top-ranked features (Figure 4.2 (f)). The bar plot is defined only by radiomic features, with log-sigma-1.5mm-3D-first-order-mean as the most informative feature, achieving 0.2. This feature is derived from a LoG filter applied with a standard deviation of 1.5 mm, where the first-order mean represents the average voxel intensity within the lesion after filtering. This feature can provide discriminative information for lesions classification.
The BMI, which played a crucial role in the 2D dataset, does not appear among the top 15 features in the 3D clinical dataset. Instead, the most informative clinical feature in this setting is "SURGERY: OVARY endometriomas: monolat", which refers to a patient's history of unilateral ovarian surgery for endometriomas. However, its contribution remains relatively low, with a mutual information score of 0.08, as it is possible to observe in Figure 4.2 (d).


2D Radiomic Mutual Information
Mutual Information Score
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
wavelet-LL_glszm_LargeAreaHighGrayLevelEmphasis
log-sigma-2-0-mm-3D_gldm_SmallDependenceHighGrayLevelEmphasis
wavelet-LL_glszm_LargeAreaLowGrayLevelEmphasis
original_firstorder_Mean
wavelet-LH_gldm_GrayLevelNonUniformity
diagnostics_Image-original_Mean
wavelet-LL_glszm_LargeAreaEmphasis
log-sigma-1-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis
wavelet-LL_gldm_DependenceNonUniformity
log-sigma-0-5-mm-3D_firstorder_Minimum
wavelet-LH_glcm_Autocorrelation
wavelet-HH_ngtdm_Contrast
wavelet-LL_glcm_SumEntropy
wavelet-HL_gldm_LargeDependenceLowGrayLevelEmphasis
log-sigma-1-0-mm-3D_glcm_JointEnergy
Features

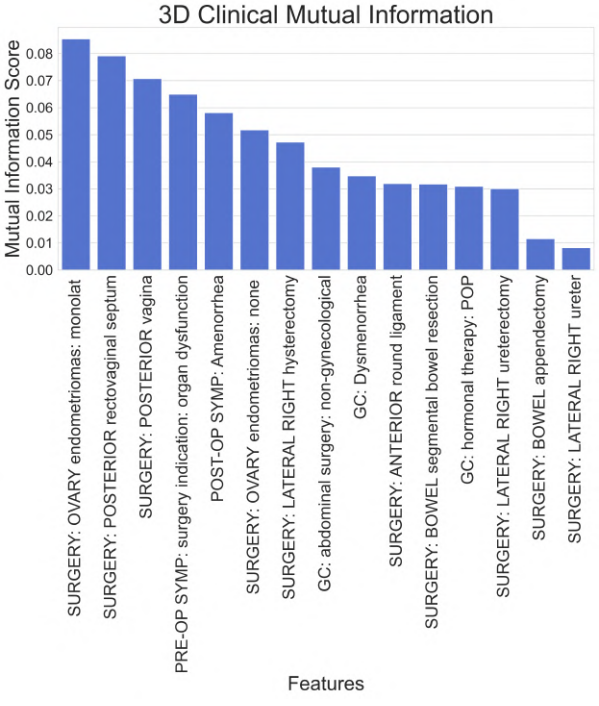
(a)


2D Radiomic Clinical Mutual Information
Mutual Information Score
0.00
0.05
0.10
0.15
0.20
0.25
GC: BMI
PRE-OP SYMP: CPP
GC: Weight kg
GC: Height cm
GC: age
wavelet-LL_glszm_LargeAreaHighGrayLevelEmphasis
log-sigma-2-0-mm-3D_gldm_SmallDependenceHighGrayLevelEmphasis
wavelet-LL_glszm_LargeAreaLowGrayLevelEmphasis
original_firstorder_Mean
wavelet-LH_gldm_GrayLevelNonUniformity
diagnostics_Image-original_Mean
log-sigma-2-0-mm-3D_glszm_GrayLevelVariance
wavelet-LL_glszm_LargeAreaEmphasis
log-sigma-1-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis
wavelet-LH_glcm_Autocorrelation
Features

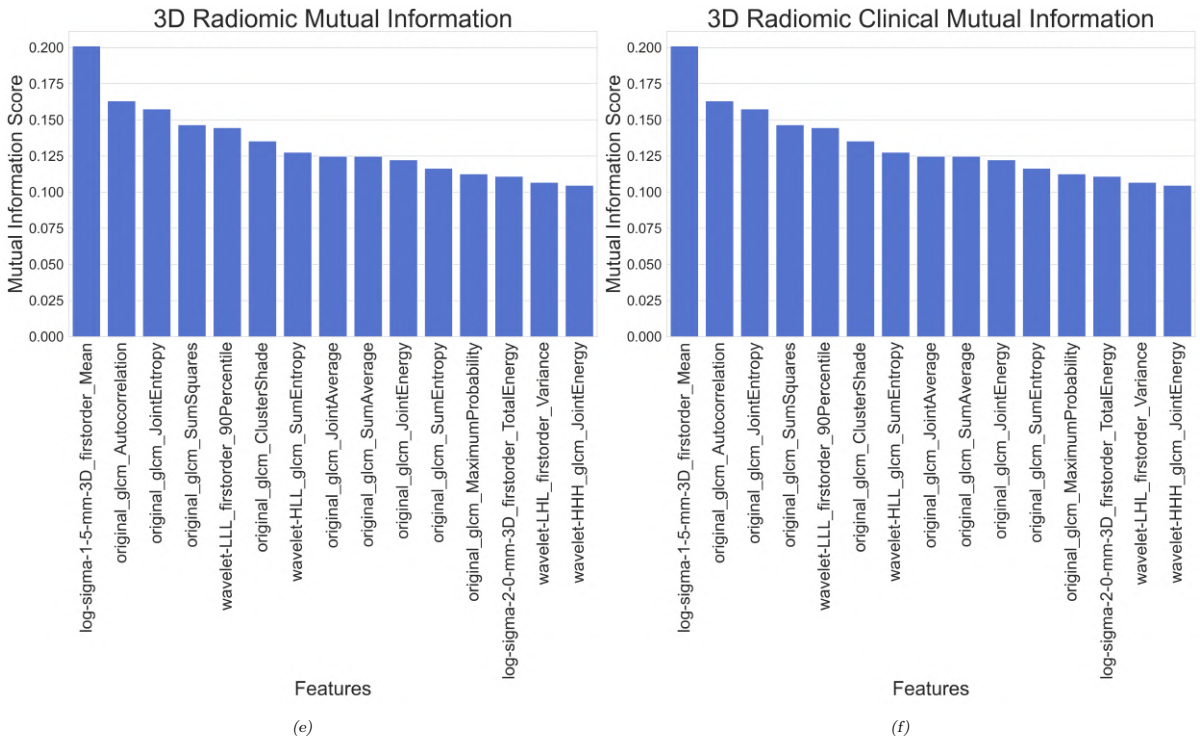
(b)


2D Clinical Mutual Information
Mutual Information Score
0.00
0.05
0.10
0.15
0.20
0.25
GC: BMI
PRE-OP SYMP: CPP
GC: Weight kg
GC: Height cm
GC: age
SURGERY: LATERAL RIGHT USL
SURGERY: ANTERIOR round ligament
POST-OP SYMP: Amenorrhea
SURGERY: BOWEL segmental bowel resection
GC: hormonal therapy: POP
SURGERY: LATERAL LEFT parametrium
SURGERY: POSTERIOR vagina
SURGERY: LATERAL LEFT USL
SURGERY: LATERAL RIGHT nodule removal periuret.
SURGERY: POSTERIOR rectovaginal septum
Features

(c)


3D Clinical Mutual Information
Mutual Information Score
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
SURGERY: OVARY endometriomas: monolat
SURGERY: POSTERIOR rectovaginal septum
SURGERY: POSTERIOR vagina
PRE-OP SYMP: surgery indication: organ dysfunction
POST-OP SYMP: Amenorrhea
SURGERY: OVARY endometriomas: none
SURGERY: LATERAL RIGHT hysterectomy
GC: abdominal surgery: non-gynecological
GC: Dysmenorrhea
SURGERY: ANTERIOR round ligament
SURGERY: BOWEL segmental bowel resection
GC: hormonal therapy: POP
SURGERY: LATERAL RIGHT ureterectomy
SURGERY: BOWEL appendectomy
SURGERY: LATERAL RIGHT ureter
Features

(d)



(e) *(f)*

Figure 4.2: *Barplots of mutual information score of each feature. The first 15 feature with the higher score are shown. (a) 2D Radiomic dataset, (b) 2D Radiomic Clinical, (c) 2D Clinical (d) 3D Clinical (e) 3D Radiomic (f) 3D Radiomic Clinical dataset*

4.1.2 Linear Discriminant Analysis

LDA is the first technique of feature extraction used in this study for data visualization. For each split generated by the cross-validation method applied (stratified K fold or stratified group K fold), the LDA model was trained using the training set and then evaluated on the test set.

After testing, the Kernel Density Estimation (KDE) plot was computed to visualize class distributions, feature importance values were analyzed, and the balanced accuracy score was recorded.

From the KDE plots, illustrated in Figure 4.3, it is possible to observe a clear overfitting. In the training set, LDA fits the data almost perfectly, effectively separating the two distributions corresponding to the active lesions and the fibrotic lesions. However, when the model is applied to previously unseen data, its performance decreases significantly leading to overlapped distributions.

For the sake of discussion, only the 2D Radiomic Clinical and 3D Radiomic Clinical datasets are shown, while plots for the other datasets can be found in the appendix **??**.

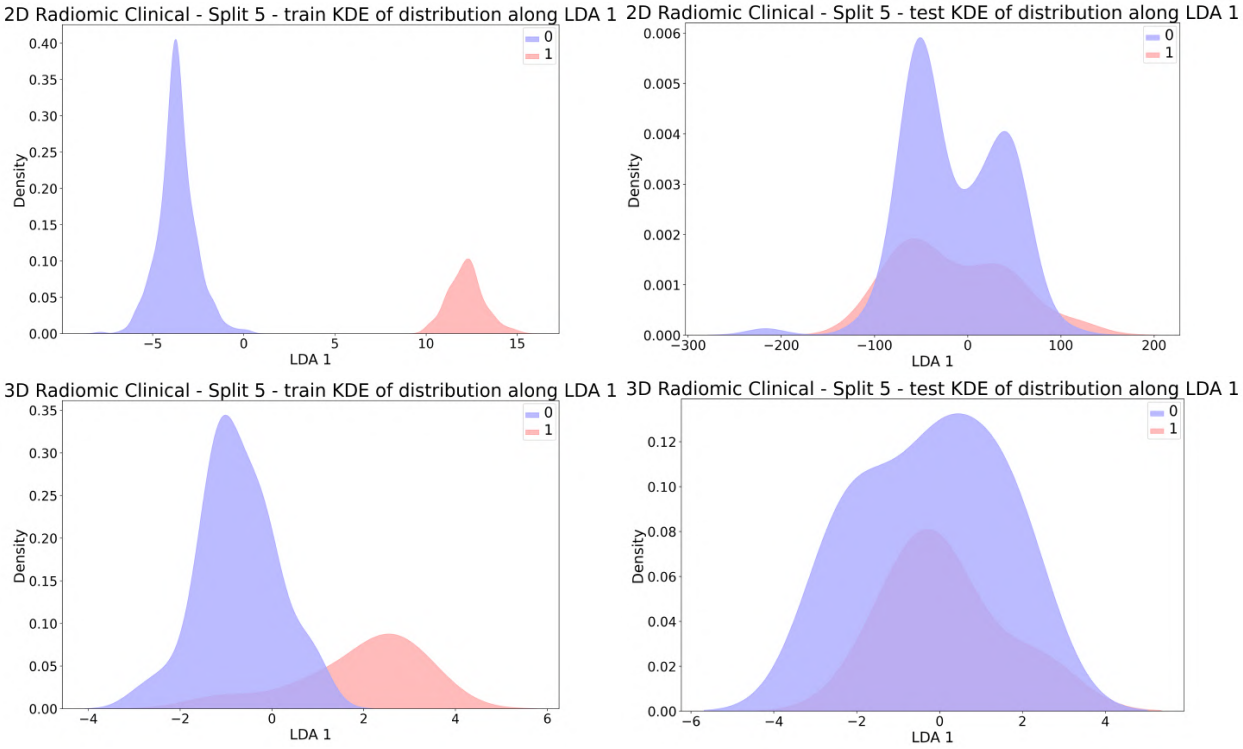


Figure 4.3: *Distributions of class 0, active lesion, and class 1, fibrotic lesion, along LDA 1 for 2D and 3D Clinical radiomic datasets. (a)-(b) 2D Radiomic Clinical KDE plot on train/test set of split 5, (c)-(d) 3D Radiomic Clinical KDE plot on train/test set of split 5*

Further evidence of overfitting is provided by the values shown in Table 4.1, which presents the average balanced accuracy with its associated error. A noticeable difference between training and test accuracy confirms that the model generalizes poorly, achieving high performance on training data but failing to maintain the same level of accuracy on unseen data.

Table of mean LDA balanced accuracy on 2D and 3D datasets		
	Balanced accuracy	
	Train	Test
2D Radiomic	0.998 ± 0.002	0.521 ± 0.049
2D Clinical	0.771 ± 0.034	0.469 ± 0.029
2D Radiomic Clinical	1.000 ± 0.000	0.481 ± 0.017
3D Radiomic	0.903 ± 0.020	0.495 ± 0.051
3D Clinical	0.796 ± 0.018	0.393 ± 0.116
3D Radiomic Clinical	0.879 ± 0.032	0.522 ± 0.110

Table 4.1: *Table of mean balanced accuracy of stratified 5 fold splits on 2D and 3D dataset resulting from the LDA models*

To quantify the contribution of each feature in LDA, the coefficients of importance

are analyzed and a comprehensive summary plot, for each dataset, is shown in figure 4.4. Figure 4.4 displays the average contribution of each feature across all cross-validation splits, in particular for the 10 most important features, ranked based on how frequently they were selected across different splits and their mean coefficient value.
In these plots, the bar length represents the absolute importance of the feature, while the direction (positive or negative) indicates whether the feature is more informative for one class or the other.

In 2D Radiomic dataset, numerous features contribute across all dataset splits, presenting high mean coefficients. Conversely, in the 2D Clinical dataset, feature importance is considerably lower. This pattern is reflected in the 2D Radiomic Clinical dataset, where radiomic features prevail over the clinical ones.

In the 3D dataset that includes both radiomic and clinical features (3D Radiomic-Clinical), a more balanced contribution from both feature types is observed. In particular, for negative class, the most relevant feature is GC: Hormonal Therapy: COC, which indicates that the patient has undergone combined oral contraceptive (COC) therapy in previous years.
The second most important contribution is the radiomic feature wavelet-HHL-firstorder-Kurtosis. This feature is derived from a wavelet transformation, where wavelet-HHL indicates the specific type of filtering applied. In this case, a high-pass filter was used along the x and y axes, to retain high-frequency details, while a low-pass filter was applied along the z axis. As a result, the transformed image emphasizes sharp intensity variations in the xy plane while smoothing intensity variations along the Z-axis.
The firstorder-kurtosis component is instead a first-order statistical measure that describes the shape of the intensity distribution within the region of interest (ROI). Specifically, kurtosis quantifies how much the pixel intensity distribution deviates from a normal distribution [42].

Despite the inclusion of both clinical and radiomic features, the LDA coefficients for 3D datasets remain relatively low. This suggests that LDA does not lead to a well-defined separation between active and fibrotic lesions.

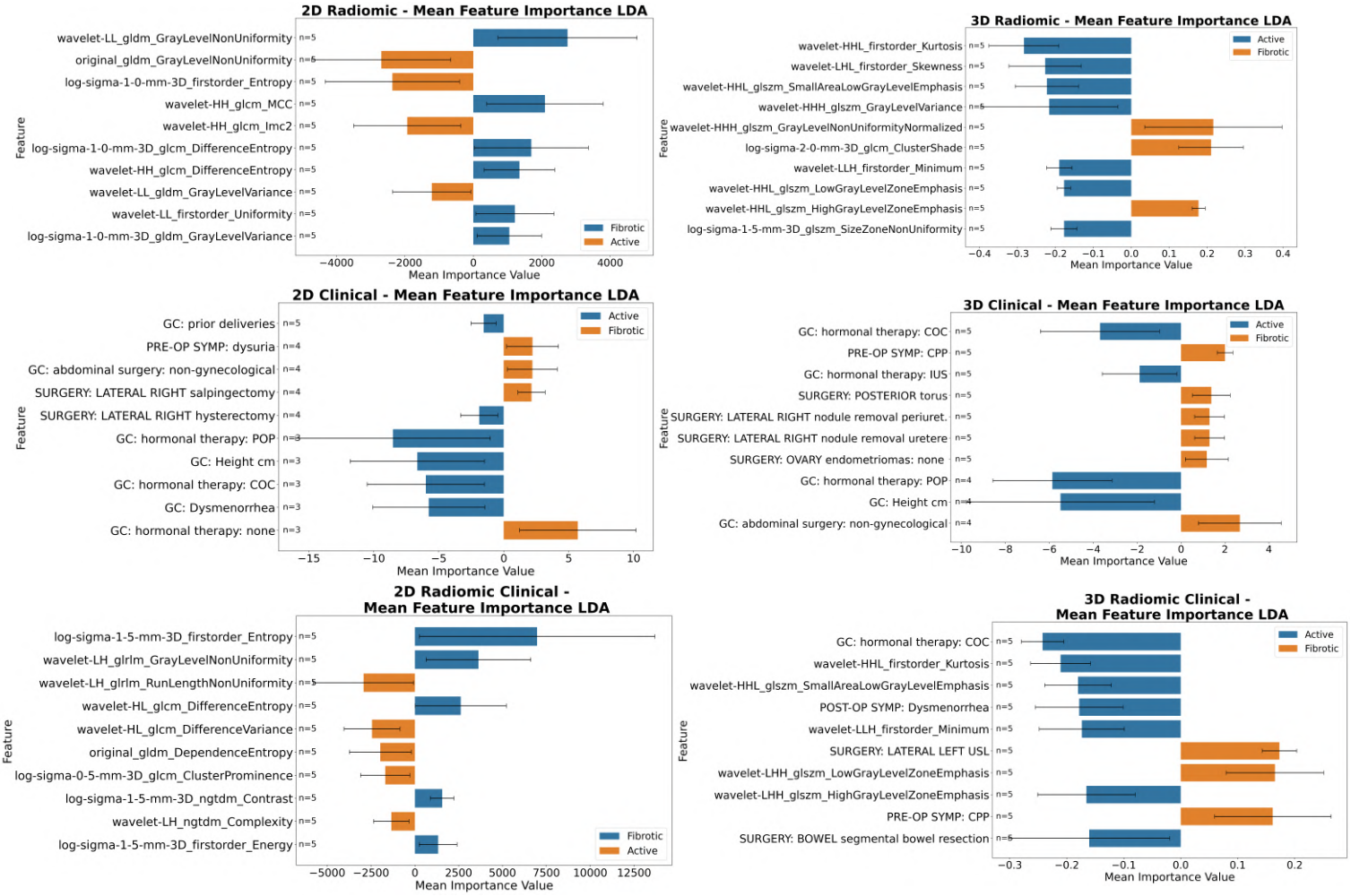


Figure 4.4: *Average contribution of the top 10 features across all cross-validation splits. Features are ranked by their frequency of selection and mean coefficient value.*

4.1.3 Principal Component Analysis

Another analysis was performed using PCA to reduce the dataset to two principal axes, maximizing the variance between the data points.
The results are visualized in scatter plots, in Figure 4.5, where the x-axis and y-axis represent the first and second principal components (PC1 and PC2), respectively. These scatter plots illustrate the relationship between data points (3D image or 2D slice of lesions) and the new variables generated by PCA.

As can be observed in Figure 4.5, in all datasets except for 2D and 3D clinical datasets, the data points tend to align along the diagonal, indicating a linear relationship between the first and the second principal components. This suggests that a preferred direction

of variance exists, but there is no explicit separation between the two classes.

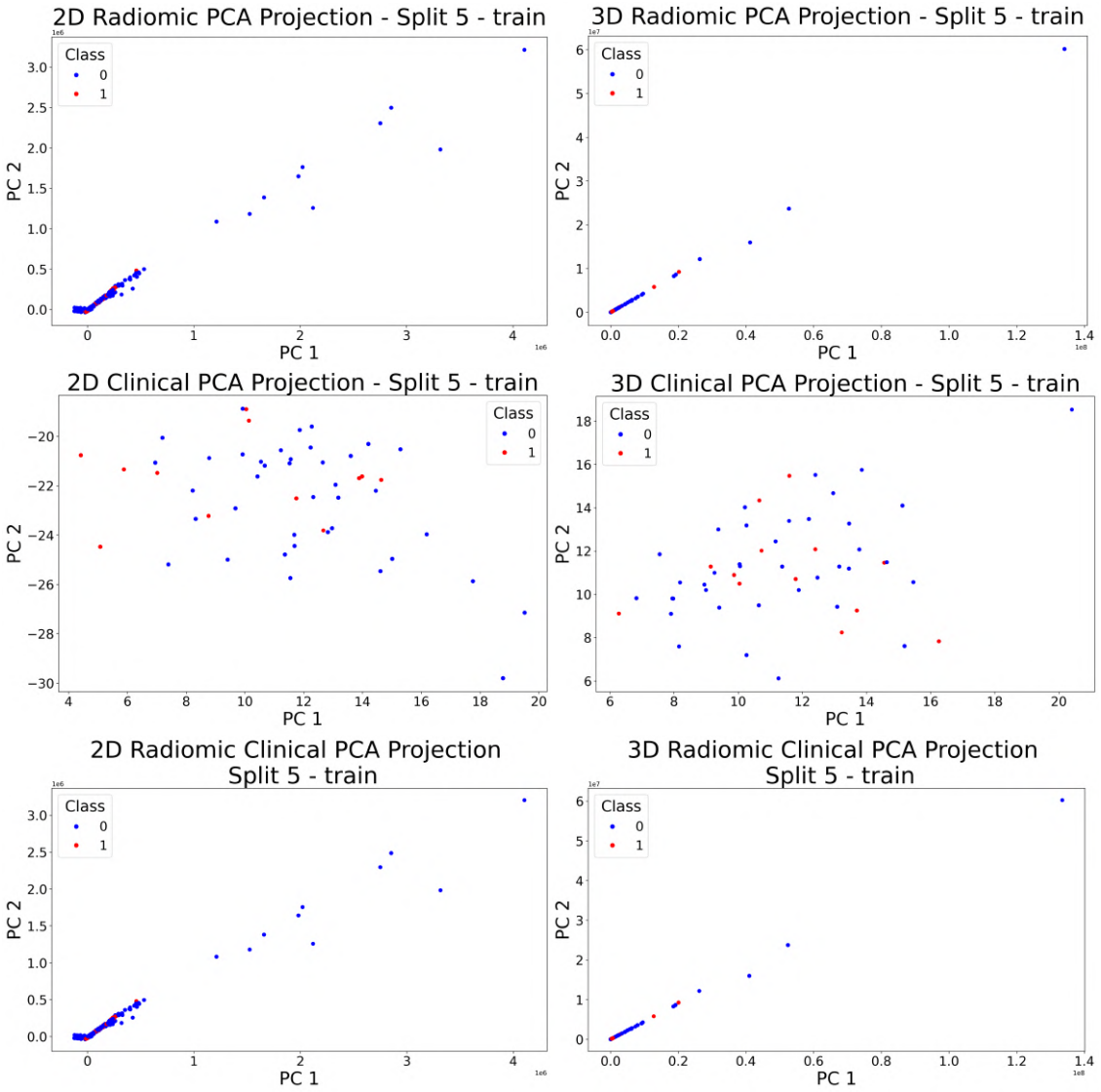


Figure 4.5: *Scatter plot of relationship between data points (3D images or 2D slices) and principal component. Active class in blue, fibrotic class in red.*

To futher investigate feature importance in PCA, the mean contribution of each feature to the principal components was analyzed, and the results are shown in Figure 4.6. The importance of a feature is determined by the magnitude of its corresponding value in the eigenvectors: features with higher absolute values in the principal components contribute more to the overall variance. In 2D Radiomic Clinical dataset, only radiomic features contribute significantly to the principal components, even if with a very low

mean importance. In all datasets, except for clinical datasets, the mean importance values remain low, typically of the order of 10^{-2}, while for clinical datasets the mean importance is higher, reaching the order of 10^{-1}.
Similar to the results of LDA, in both the 2D and 3D Radiomic Clinical datasets, clinical features do not show high enough importance, indicating that only radiomic features play a more important role in the definition of the principal components.

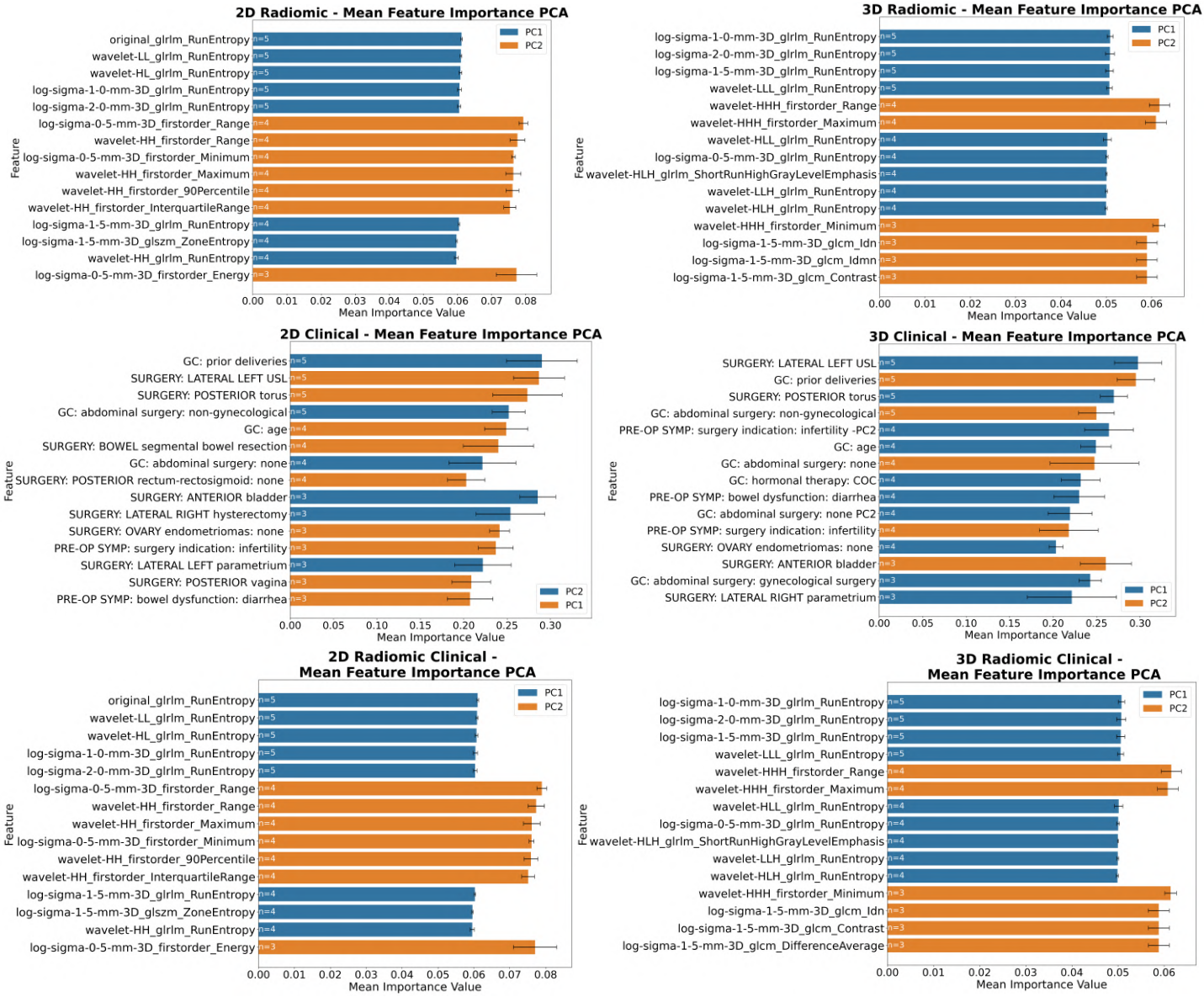
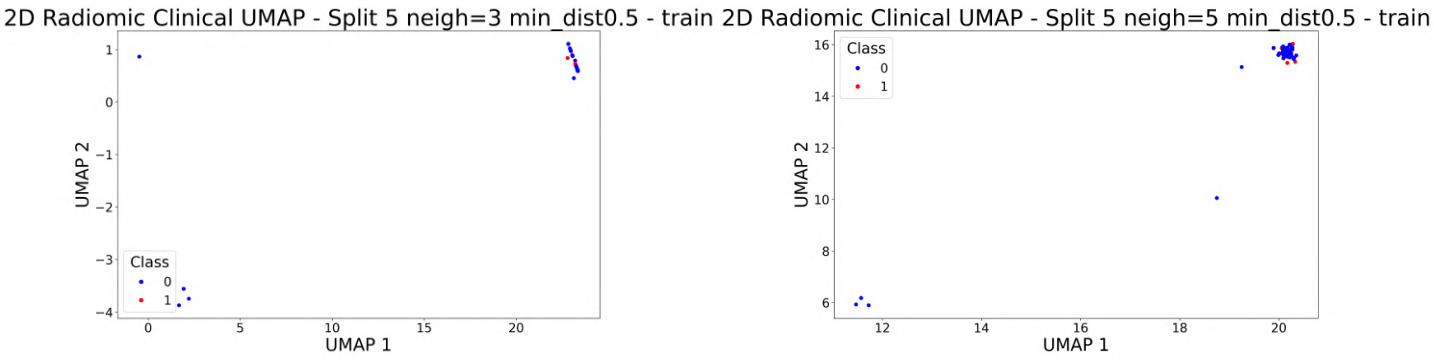


Figure 4.6: *Bar plots illustrating the mean contribution of each feature - across the cross validation splits - to the principal components*

4.1.4 Uniform Manifold Approximation and Projection

After applying PCA and LDA, UMAP was used to further investigate whether the combination of radiomic and clinical features allows for a clear separation between active and fibrotic lesions. The UMAP algorithm tries to preserve both local and global relationships in the dataset by constructing a high-dimensional graph and optimizing its projection into a two-dimensional space.
UMAP arranges data points in a lower-dimensional space based on two parameters, the number of neighbors (n_neighbors), which represents number of points in the low dimensional space to consider around a central point, and minimum distance (min_dist), which defines the minimum distance between points in low-dimensional space, where lower values result in more compact clusters.
To investigate these effects, UMAP was first applied with n_neighbors of 3 and 5. As shown in Figure 4.8, increasing the number of neighbors leads to a more continuous distribution of points, reducing the formation of small, well-separated clusters. This is expected, as lower values of n_neighbors prioritize local structure and small-scale patterns, while higher values focus on global relationships, showing the overall structure.
However, even at low values of n_neighbors, no clear separation between active and fibrotic lesions arises, suggesting that neither radiomic nor clinical features provide strong class separability in this representation.
For the sake of discussion, only the 2D and 3D Radiomic Clinical are shown in Figure 4.8, while the other datasets can be found in the appendix **??**.



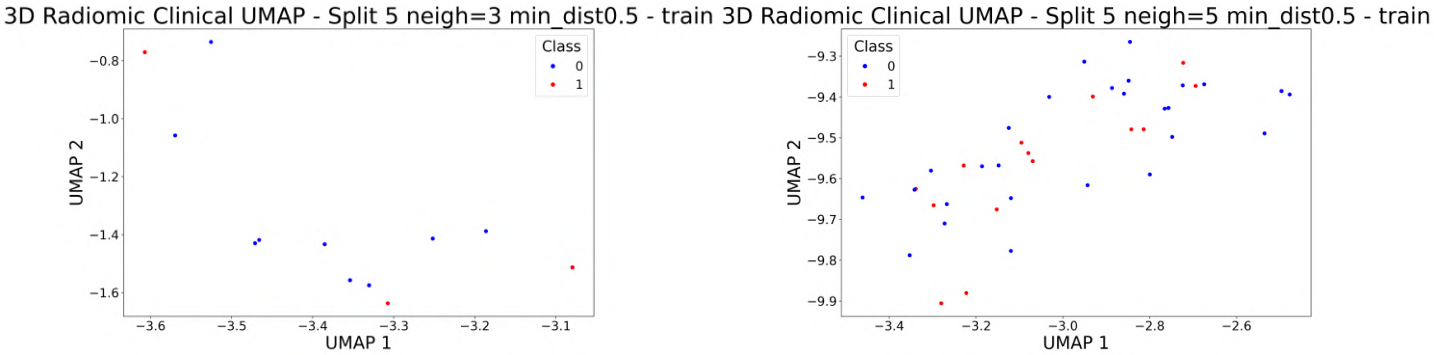
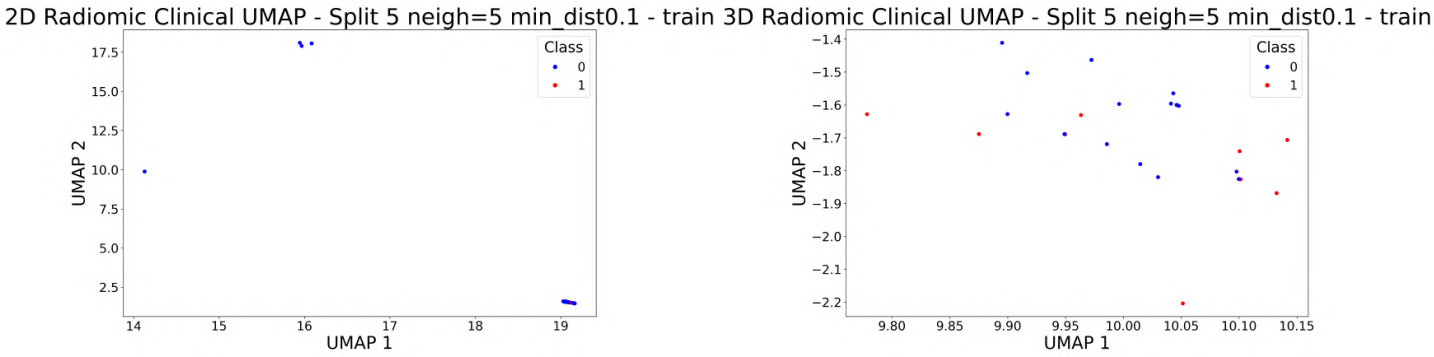


Figure 4.8: *UMAP scatter plots of 2D and 3D Radiomic Clinical datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

Next, the min_dist parameter was varied (0.1, 0.5, and 1) to observe how it affects the spread of points, and the results of 2D and 3D Radiomic Clinical dataset are shown in Figure 4.10 (see appendix for other dataset).
In the 2D Radiomic dataset, UMAP produced localized groups of points, indicating that certain radiomic features tend to cluster together. However, these clusters did not correspond to a clear separation of lesion types. In contrast, the 3D Radiomic-Clinical dataset exhibits a more dispersed structure, especially as min_dist increases. This suggests that clinical features introduce additional variability, leading to a more continuous distribution rather than distinct groupings.

Also UMAP does not reveal any natural separation between lesion types. This aligns with the findings from PCA and LDA.



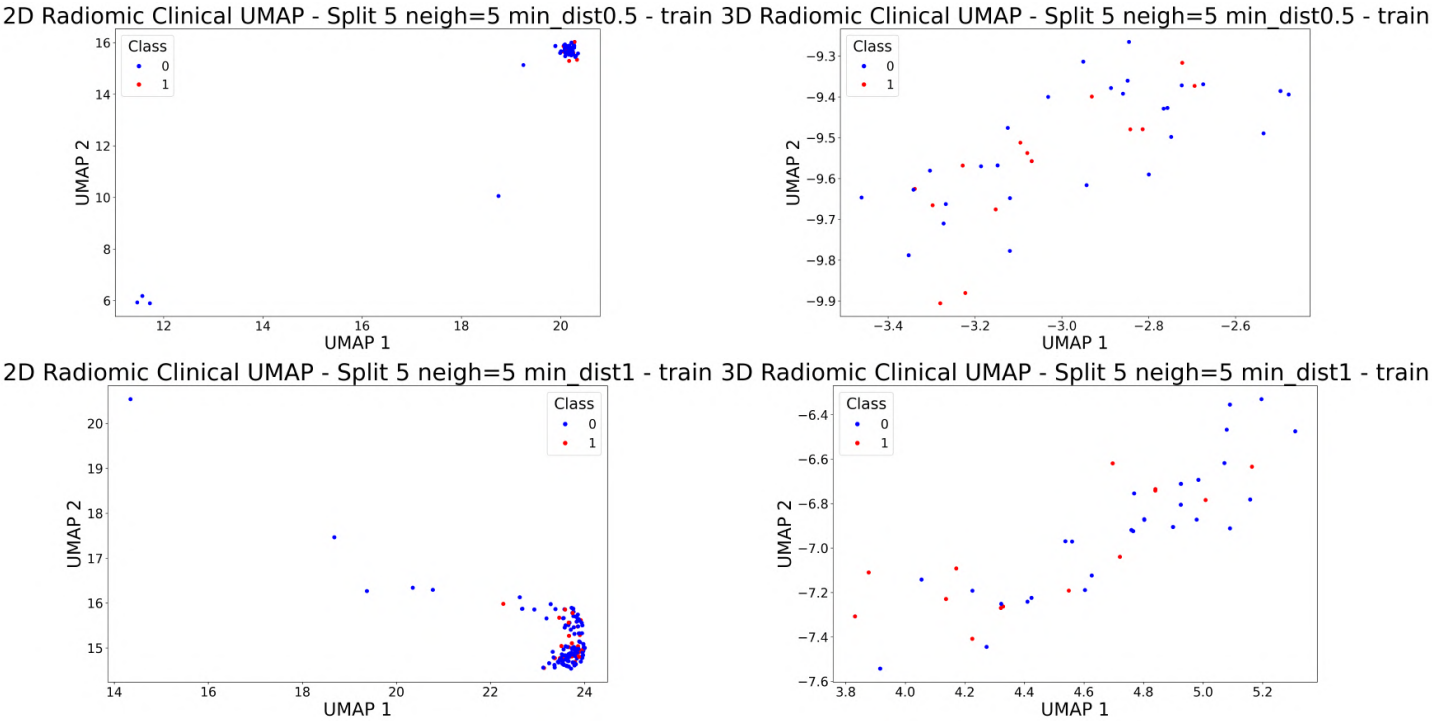


Figure 4.10: *UMAP scatter plots of 2D and 3D Radiomic Clinical datasets, at different value of min_dist. Active class in blue, fibrotic class in red.*

4.1.5 Pairwise Controlled Manifold Approximation Projection

PaCMAP was also employed as a visualization technique to generate a low dimensional representation of the dataset, in order to assess whether active and fibrotic lesions can be effectively separated in another reduced-dimensional space.
To visualize the results given by PaCMAP, two-dimensional scatterplots were produced. In Figure 4.11 the 2D and 3D Radiomic Clinical datasets behavior (see appendix **??** for other datasets) with a number of three and five neighbors is illustrated. Despite the fact that differences in point arrangment can be observed when the parameter changes, no clear separation between active and fibrotic lesions arises.
As observed for LDA, PCA and UMAP, also PaCMAP has not identified a projection where the lesion types are well distinguishable.

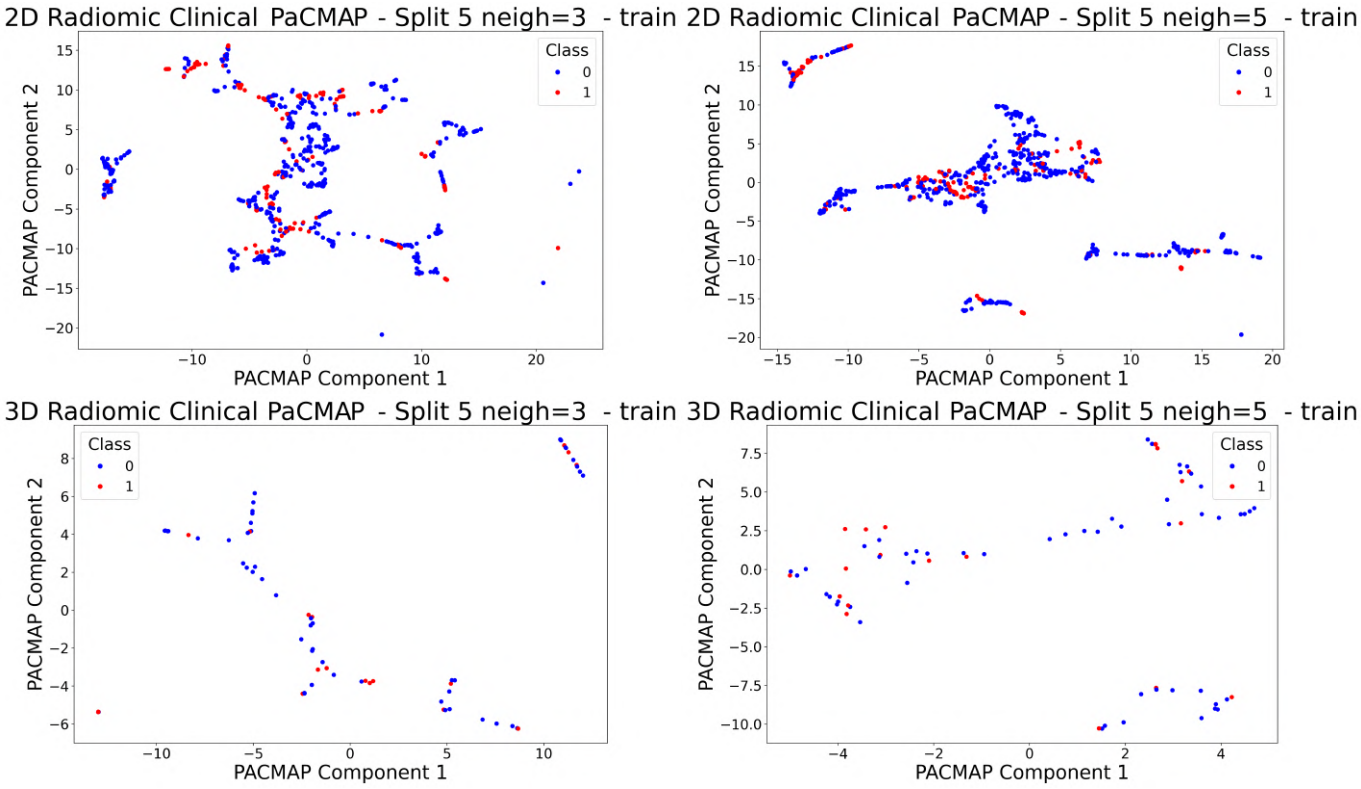


Figure 4.11: *PacMap scatterplots for 2D and 3D Radiomic Clinical datasets with three (left) and five (right) number of neighbors. Active class in blue, fibrotic class in red.*

For all feature extraction methods (LDA, PCA, UMAP, PaCMAP) no clear separation between active and fibrotic lesions was observed, suggesting that the selected features may not contain sufficient discriminative information to naturally separate the two classes within the reduced space, further emphasizing the need for more advanced classification strategies.

4.2 Pipeline analysis

By exploiting TPOT tools, a pipeline is automatically produced for each dataset.

4.2.1 3D Radiomic Pipeline

The TPOT pipeline for the 3D Radiomic dataset begins by transforming and refining the feature space, then selects the most important features, and ultimately uses a decision tree to classify various lesion types.

```
make_pipeline(RBFSampler(gamma=0.5),
    RFE(estimator=ExtraTreesClassifier(criterion="entropy",
        max_features=0.05, n_estimators=100),
        step=0.7500000000000001),
    Normalizer(norm="l1"),
    DecisionTreeClassifier(criterion="entropy", max_depth=6,
        min_samples_leaf=14, min_samples_split=9))
```

The pipeline starts with the RBFSampler, which applies a random Fourier feature transformation using the Radial Basis Function (RBF) kernel, that maps data into higher dimensional space. Its gamma parameter controls the spread of the transformation and affects how points are separated into the new feature space.

Next, the pipeline applies Recursive Feature Elimination (RFE) to select the most relevant features. It does this by repeatedly training an ExtraTreesClassifier, evaluating feature importance, and removing the least useful features.
The ExtraTreesClassifier employs an ensemble approach, utilizing entropy to evaluate the quality of splits. It decreases the number of features considered at each split to 5% of the total, ensuring that only the most informative features are retained for

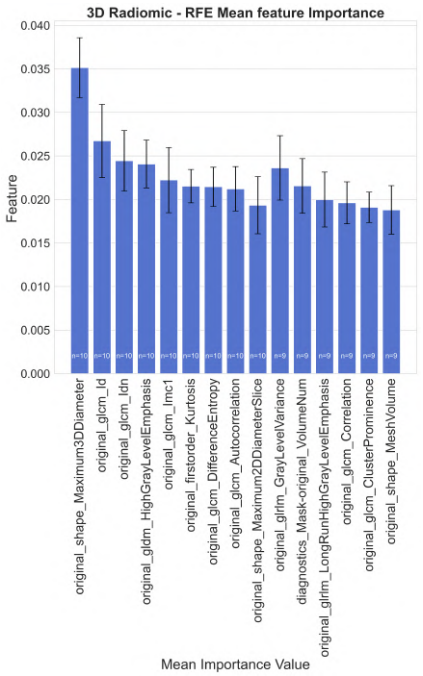


Figure 4.12: *Mean feature importance across 10 Fold Split Stratified Cross-Validation, ranked based on their mean importance and number of times they were selected across the splits*

classification.
Along with the ExtraTreesClassifier, RFE uses a step size of 0.75, meaning that in each iteration, 75% of the least important features are removed, making the selection process faster. After the application of the RFE the number of features is reduced from 1228 to 94. The top 15 most important features are shown in Figure 4.12, with the original_shape_Maximum3D_Diameter being the most relevant feature. In particular, original_shape indicates that the feature is extracted from the original, untransformed 3D lesion volume (not derived from filtered or processed images), while Maximum3D_Diameter quantifies the largest Euclidean distance between any two points on the lesion's surface within the segmented 3D volume.

After feature selection, the dataset is normalized using L1 normalization so that its absolute sum equals one.
The processed features are then classified using a Decision Tree Classifier. The tree is built using the entropy criterion, meaning it selects splits that maximize the information gain. The tree is limited to a maximum depth of 6, preventing it from becoming too complex and reducing the risk of overfitting. Additionally, it requires at least 14 samples in a leaf node and at least 9 samples to split a node.

At the final stage of the pipeline, the number of remaining features is approximately 13 and their mean importance is illustrated in Figure 4.13, where features are ranked according to their average importance and the frequency with which they are selected across cross-validation splits. The only feature consistently selected across all splits is original_shape_Maximum3DDiameter, the same feature with the highest importance also in the previous plot.

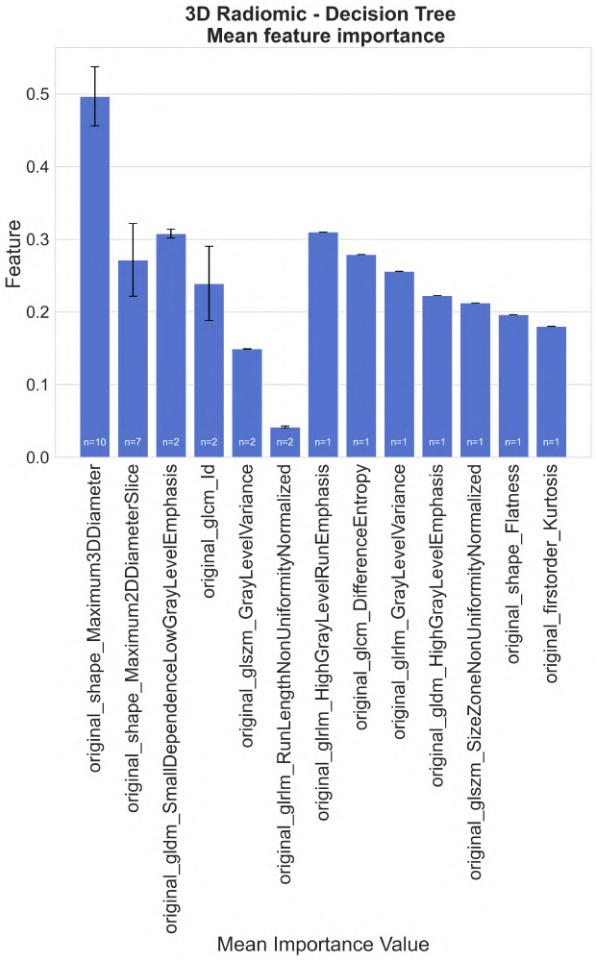


Figure 4.13: *Mean feature importance of the 3D Radiomic dataset across 10 Fold Split Stratified Cross-Validation. The feature are ranked based on their mean importance and number of times their were selected across the splits*

The performance of the pipeline was

evaluated using multiple classification metrics, with results reported as mean values across a 10-fold cross-validation (Table 4.2). The balanced accuracy of 0.56 ± 0.22 is only slightly higher than the 0.5 threshold for random classification, suggesting that the model struggles to achieve reliable discrimination between the two classes.
The test AUC is 0.66 ± 0.20, which is higher than the balanced accuracy score but still indicates moderate discriminative ability. This suggests that while the model can capture some degree of class separability, its overall predictive performance remains limited. In particular, the model does not perform equally well for both classes, leading to imbalanced classification outcomes.

The AUROC curves shown in Figure 4.14, illustrate the model's behavior across different cross-validation splits, showing a deviation from the diagonal (random classifier), which indicates a certain level of predictive power, though not strong enough for reliable classification.

Table of mean balanced accuracy and AUC 3D Radiomic dataset			
Train Balanced Accuracy	Train AUC	Test Balanced Accuracy	Test AUC
0.78 ± 0.04	0.90 ± 0.01	0.56 ± 0.22	0.66 ± 0.20

Table 4.2: *Table of mean balanced accuracy and AUC for test and train of 3D Radiomic dataset across 10 Fold Split Stratified Cross-Validation*

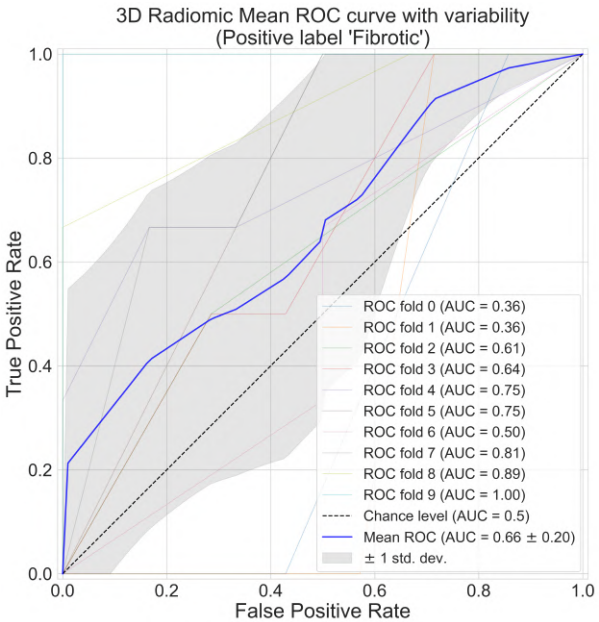


Figure 4.14: *Mean ROC-AUC curve of 3D Radiomic dataset across 10 Fold Split Stratified Cross-Validation*

A more detailed evaluation is provided in Table 4.3, where additional metrics such

as precision, recall and F1-score, are reported and the normalized confusion matrix is displayed in Figure 4.15. Along the diagonal, we observe a mean value of 0.80 for true positive predictions (active lesions) and 0.32 for true negative predictions (fibrotic lesions), which correspond to the recall values in Table 4.3. The off-diagonal values, 0.2 (false positives) and 0.68 (false negatives), indicate that the model performs better in identifying active lesions but struggles significantly with fibrotic lesions, leading to a high false-negative rate.

Table of mean classification report 3D Radiomic			
	precision	recall	F1-score
0	0.75 ± 0.12	0.80 ± 0.21	0.77 ± 0.15
1	0.38 ± 0.4	0.32 ± 0.33	0.33 ± 0.33

Table 4.3: *Table of mean classification report across 10 Fold Split Stratified Cross-Validation of 3D Radiomic dataset*

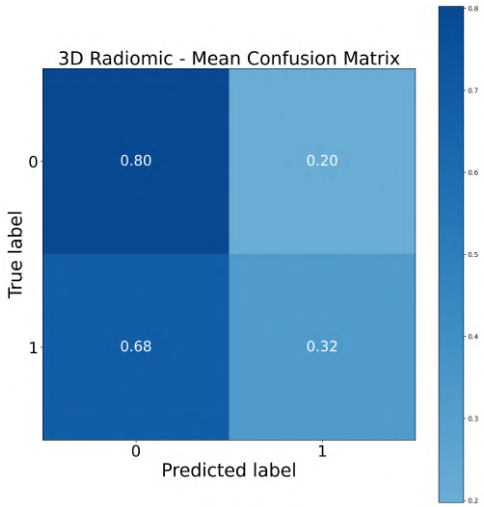


Figure 4.15: *Normalized mean confusion matrix of 3D Radiomic dataset across 10 Fold Split Stratified Cross-Validation*

These findings suggest that while the model shows some predictive capability, it is far from optimal for clinical application. The high misclassification rate for fibrotic lesions is particularly concerning, as it implies poor sensitivity in detecting these cases, which could result in inaccurate patient diagnosis.

4.2.2 3D Clinical

For 3D clinical dataset the generated pipeline includes a MLPClassifier used as a stacking estimator, then a scaling of the data through a Robust Scaler, which ensures the same scale for all features and reduces sensitivity to outliers, and finally applies an ensemble

method for classification, in particular Gradient Boosting Classifier (GBC).

```
exported_pipeline = make_pipeline(
    StackingEstimator(estimator=MLPClassifier(alpha=0.01,
        learning_rate_init=0.001)),
    RobustScaler(),
    GradientBoostingClassifier(learning_rate=1.0, max_depth=3,
        max_features=0.45, min_samples_leaf=3, min_samples_split=5,
        n_estimators=100, subsample=0.45))
```

The pipeline starts with a MLPClassifier estimator for stacking estimator, where it is used as a feature engineering tool: the outputs of this neural network are appended to the original feature set.

The MLP introduces new, abstract feature representations by learning complex relationships between input variables.
In Figure 4.16 the weights between the input layer and the first hidden layer of MLP neural network are shown, sorted by mean feature importance across the ten splits. A barplot of the weights reveals which inputs have the most influence in the first layer of processing. In Figure 4.16, the weights between the input layer and the first hidden layer of the MLP neural network are sorted by mean feature weight across ten splits and the feature with the highest importance is surgery of the anterior round ligaments, indicating that the patient has undergone a surgical procedure in that area.

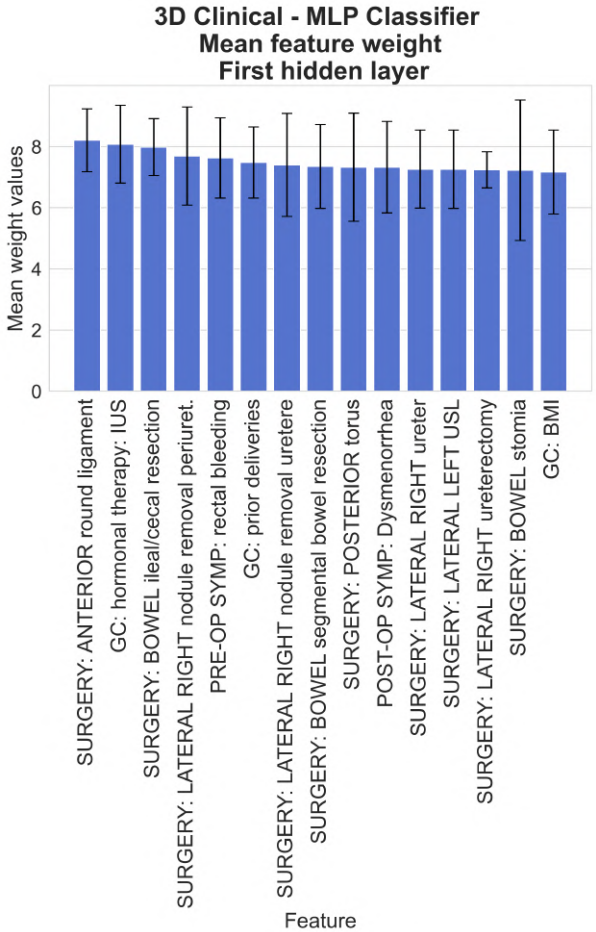


Figure 4.16: *Features weight between the input layer and the first hidden layer of the MLP neural network, sorted by mean value across ten splits*

Then the Robust Scaler is applied.

The Gradient Boosting Classifier is selected as the final classifier.

This model was selected by TPOT, which optimized its hyperparameters, resulting in a high learning rate of 1.0. While a high learning rate accelerates convergence, it also increases the risk of overfitting, as each individual tree has a greater influence on the final prediction. The model consists of 100 decision trees, the default setting, with a depth of each decision tree restricted to three levels.
Additionally, for each split within a tree, only 45% of the features are considered.
The classifier requires a minimum of three samples per leaf node and at least five samples per split, preventing too much specific splits that could fit noise rather than meaningful patterns.
After applying the Gradient Boosting Classifier, 47 out of 58 features were assigned a nonzero importance value, indicating their contribution to the classification process. The top 15 most important features, ranked by their mean importance across the ten splits, are presented in Figure 4.17, the first positions are occupied from general patients characteristcs as height, weight and BMI.

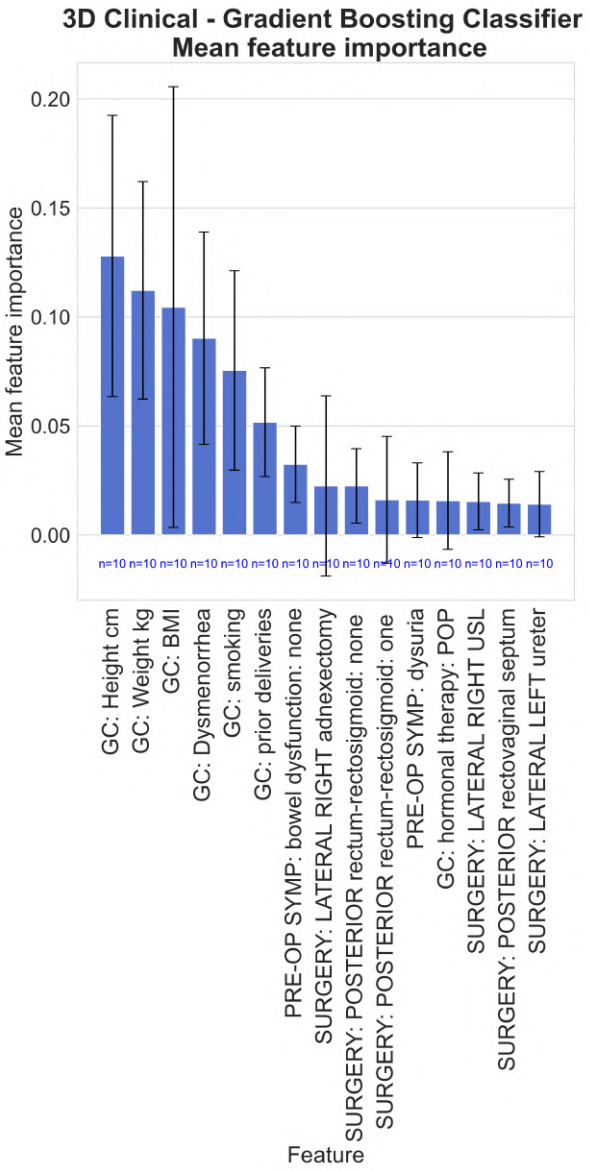


Figure 4.17: *Mean Feature Importance of the 3D Clinical Dataset across 10 Fold Split Stratified Cross-Validation. The feature importance values are derived from a Gradient Boosting Classifier and sorted by their mean importance, and on the number of times their importance was greater than zero across the splits.*

The pipeline is evaluated, and the metric score is presented in Table 4.4, calculated as the average values from a 10-fold cross-validation.
As it is possible to observe in Table 4.4 the pipeline reveals poor classification performance. The balanced accuracy is 0.45 ± 0.25, indicating that the model performs a

random classification. Similarly, the AUC score on the test set is about 0.44 ± 0.20, and still does not suggest a meaningful discriminative capability. These two results indicate that the model is misclassifying more often than it is correctly distinguishing between classes.
Figure 4.18 presents the mean AUROC curve across the 10-fold splits, alongside individual ROC curves for each fold. The observed deviation from the diagonal (random classifier) is minimal, further confirming the model's lack of discriminative power.

Table of mean balanced accuracy and AUC of 3D Clinical dataset			
Train Balanced Accuracy	Train AUC	Test Balanced Accuracy	Test AUC
0.56 ± 0.05	0.56 ± 0.05	0.45 ± 0.25	0.44 ± 0.20

Table 4.4: *Table of mean balanced accuracy and AUC for test and train of 3D Clinical dataset across 10 Fold Split Stratified Cross-Validation*

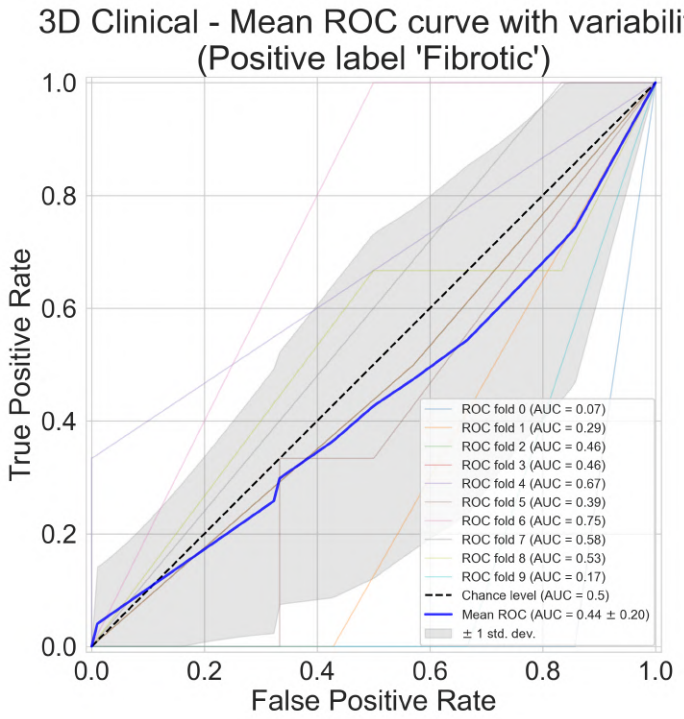


Figure 4.18: *Mean ROC-AUC curve of 3D Clinical dataset across 10 Fold Split Stratified Cross-Validation*

A more detailed analysis, including precision, recall and F1-score, is summarized in Table 4.5 and the normalized confusion matrix (Figure 4.19) provides further insights into the classification errors. The diagonal elements, representing correctly classified cases, show a mean of 0.59 (true negative, active lesions) and 0.32 (true positive, fibrotic lesions). However, the off-diagonal values are high, with 0.41 false positives (active

lesions misclassified as fibrotic) and 0.68 false negatives (fibrotic lesions misclassified as active). This suggests that the model struggles significantly with correctly identifying fibrotic lesions.
It is likely due to imbalanced feature classes or insufficient discriminative information in the extracted features.

Table of mean classification report 3D Clinical			
	precision	recall	F1-score
0	0.68 ± 0.21	0.59 ± 0.29	0.60 ± 0.23
1	0.24 ± 0.28	0.32 ± 0.37	0.26 ± 0.29

Table 4.5: *Table of mean classification report across 10 Fold Split Stratified Cross-Validation of 3D Clinical dataset*

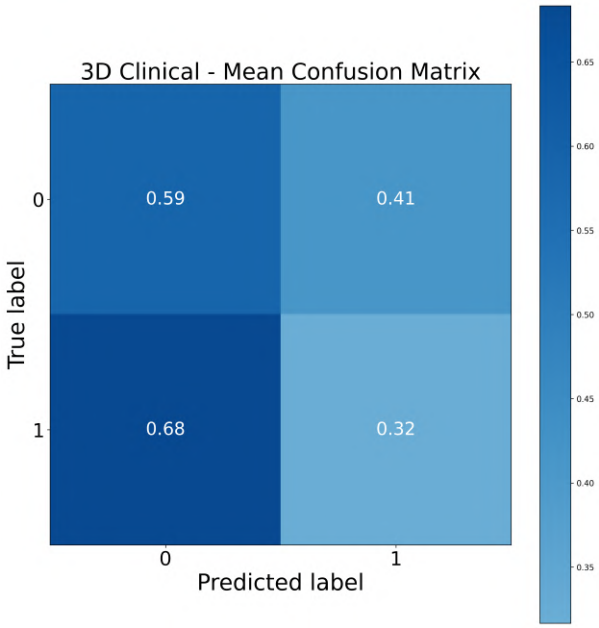


Figure 4.19: *Normalized mean confusion matrix of 3D Clinical dataset across 10 Fold Split Stratified Cross-Validation*

In general, these results indicate that the current pipeline does not provide a reliable classification between active and fibrotic lesions.

4.2.3 3D Radiomic Clinical

The optimized pipeline selected by TPOT for processing the 3D Radiomic Clinical dataset transforms and refine the feature space through a series of steps, and the apply a LinearSVC classification model to define the final classes.

```
exported_pipeline = make_pipeline(
    StackingEstimator(estimator=SGDClassifier(alpha=0.001, eta0=0.1,
        fit_intercept=False, l1_ratio=0.0, learning_rate="invscaling",
        loss="hinge", penalty="elasticnet", power_t=0.1)),
    ZeroCount(),
    SelectPercentile(score_func=f_classif, percentile=62),
    VarianceThreshold(threshold=0.05),
    LinearSVC(C=25.0, dual=False, loss="squared_hinge",
    penalty="l1", tol=0.01)
)
```

The pipeline starts with a StackingEstimator, which applies a Stochastic Gradient Descent (SGD) classifier as an intermediate transformation step. This classifier uses a hinge loss function, and employs L2 regularization. By doing so, it increases the feature space with linear projections of the original features, creating a new representation of the data before feature selection.

Elastic Net regularization, the hyperparameter of SDG, is a linear regression technique that combines L1 and L2 penalties, but in this case the l1_ratio is set to 0.0, and so only the L2 regularization is applied. Also, by setting the invscaling, the learning rate (invscaling) decreases progressively during training, preventing abrupt updates.

Since this classifier is used within a StackingEstimator, its outputs are not used for direct classification. Instead, they serve as transformed features that enrich the dataset. This transformation add one feature to the dataset, expanding it from 1283 to 1284 fea-

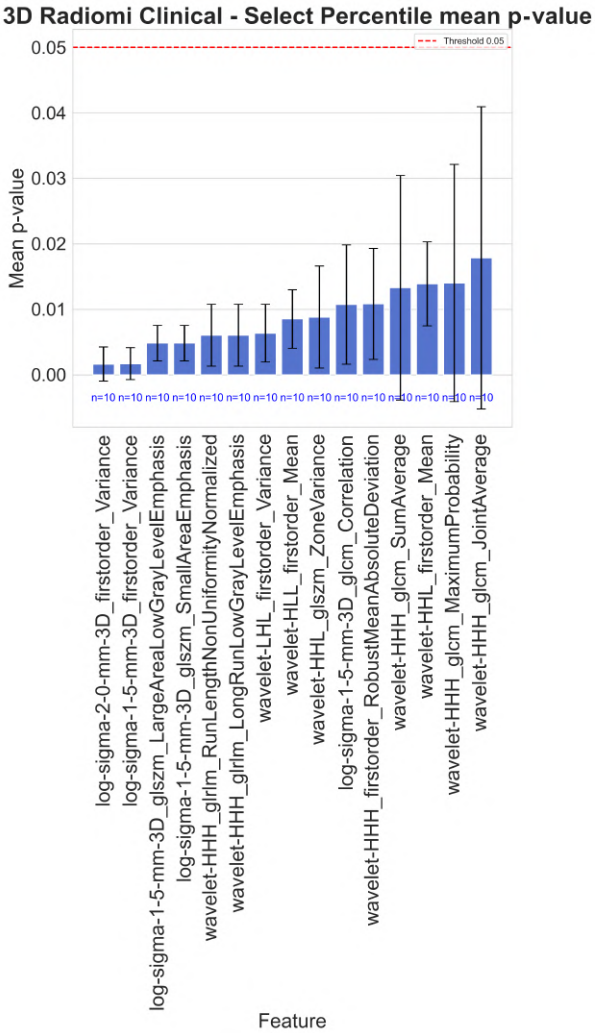


Figure 4.20: *Top 15 mean p-values selected by SelectPercentile across all 10 fold splits, ordered per frequency and lower pvalue.*

tures.

The successive step applies ZeroCount, which calculates how often each feature is equal to zero across the dataset for each sample, and helping the next steps in the identification of redundant or uninformative features.

To reduce the number of features and keep the most informative ones, the SelectPercentile method is applied. This step uses the ANOVA F-test (f_classif) to evaluate how well each feature distinguishes between the target classes. Features with higher F-values are considered more relevant, while those with low statistical significance are removed.
As a result the number of features is reduced from 1284 to 797. The top 15 features with the lowest p-values (indicating higher statistical importance) are shown in Figure 4.20. The values represent the mean feature importance across 10 cross-validation splits, with error bars indicating the associated standard deviation. Additionally, the plot displays the frequency with which each feature was selected across the 10 folds, ranking them based on both selection frequency and mean p-value (lower p-values correspond to higher importance).
It can be observed that only radiomic features appear in the ranking, suggesting that clinical features have lower statistical significance in distinguishing the target classes. The most important feature is log-sigma-2-0-mm-3D_firstorder_Variance, which quantifies the degree of intensity variation within a lesion after applying a LoG filter with a 2.0 mm smoothing scale. The sigma value defines the standard deviation of the Gaussian smoothing applied before the Laplacian operator is used. A sigma of 2.0 mm means that the filter selectively enhances structures that are approximately 2.0 mm in size, while smaller details are suppressed. The 3D designation indicates that the LoG filter was applied to the entire three-dimensional lesion volume, while variance measures how widely intensity values are spread around the mean: a high variance means that the lesion contains voxels with widely varying intensity values.

A VarianceThreshold filter (0.05) further removes near-constant features, so that only variables with meaningful variation contribute to the model. The dataset is further reduced reaching a value of 397 features, that are the features passed to the final classifier.

The classification step is performed using LinearSVC, SVM model that incorporates L1 regularization. The classifier employs a squared hinge loss function with an high regularization parameter (C = 25.0), creating a strict decision boundary.
The final number of selected features is approximately 350. Figure 4.21 illustrates the top 10 most important features after the selection process, ranked by their frequency of selection and absolute importance. The sign of the coefficients indicates whether a feature contributes more to the negative class (active lesions) or the positive class (fibrotic lesions).

The most important feature for active lesion classification is wavelet-LHL-first_order_variance, which measures intensity variance after applying a wavelet transformation (LHL decomposition), highlighting textural variations.
For fibrotic lesion classification, the most important feature is log-sigma-0-5-mm-3D_first _order_variance, which captures intensity variations after applying a Laplacian of Gaussian filter with a standard deviation of 0.5 mm.

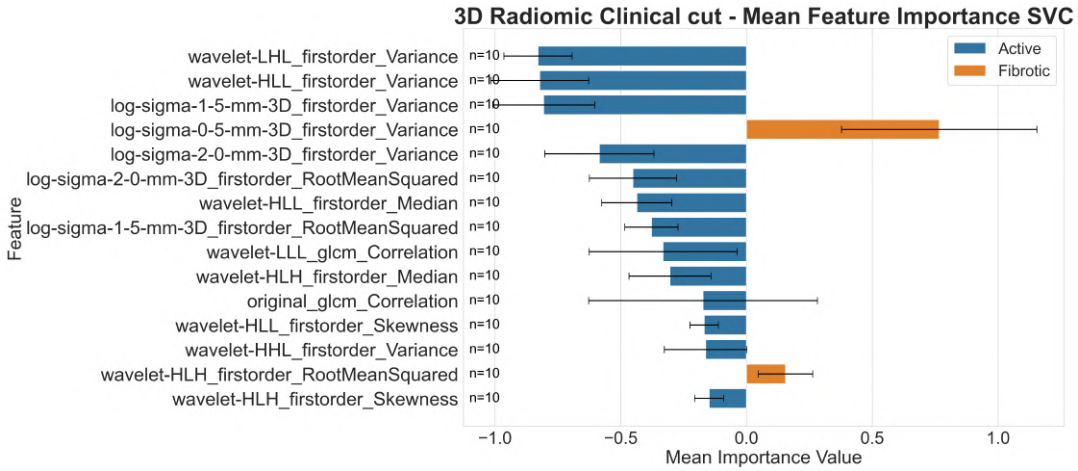


Figure 4.21: *Mean feature importance of the top 10 features after SVC classification, averaged across 10-fold cross-validation. Features are ranked by frequency of selection and mean importance. Bars pointing towards positive values indicate association with the positive class (fibrotic lesions), while bars towards negative values indicate association with the negative class (active lesions).*

The pipeline's performance is assessed using balanced accuracy and AUC and the results are summarized in Table 4.6.
For the balanced accuracy the model achieves a value of 0.65 ± 0.19 for the test set. This suggests that the model struggles to effectively differentiate between classes.
The test AUC of 0.60 ± 0.20 and the mean ROC curve across the 10 cross validation folds (Figure 4.22, displaying a deviation from the diagonal trend), confirm the model's limited performance, showing that it does slightly better than random guessing but lacks strong predictive capability.
It is important to notice that the training set achieves a balanced accuracy and AUC of 1.0, indicating important overfitting. This means that while the model performs perfectly on training data, it fails to generalize well to unseen test data.

Table of mean balanced accuracy and AUC - 3D Radiomic Clinical			
Train Balanced Accuracy	Train AUC	Test Balanced Accuracy	Test AUC
1.0 ± 0.0	1.0 ± 0.0	0.65 ± 0.19	0.60 ± 0.20

Table 4.6: *Table of mean balanced accuracy and AUC for test and train of 3D Radiomic Clinical dataset across 10 Fold Split Stratified Cross-Validation*

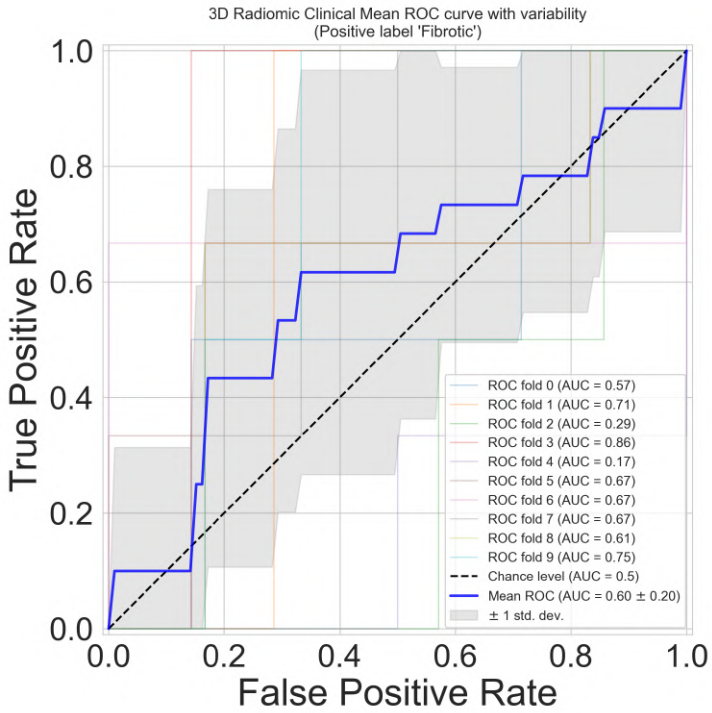


Figure 4.22: *Mean ROC-AUC curve of 3D Radiomic Clinical dataset across 10 Fold Split Stratified Cross-Validation*

As with previous pipelines, the classification report for this pipeline is also computed and reported in Table 4.7, while the confusion matrix is shown in Figure 4.23.
In the confusion matrix, the mean recall values are of the order of 0.77 and 0.53; while the off-diagonal values (misclassifications) are 0.23 (active lesion misclassified as fibrotic) and 0.47 (fibrotic misclassified as active).
This suggests that the model is better at identifying active lesions than fibrotic lesions, leading to higher false negatives for fibrotic cases.

Table of mean classification report
3D Radiomic Clinical dataset

	precision	recall	F1-score
0	0.82 ± 0.13	0.77 ± 0.11	0.78 ± 0.09
1	0.45 ± 0.31	0.53 ± 0.34	0.47 ± 0.29

Table 4.7: *Table of mean classification report across 10 Fold Split Stratified Cross-Validation of 3D Radiomic Clinical dataset*

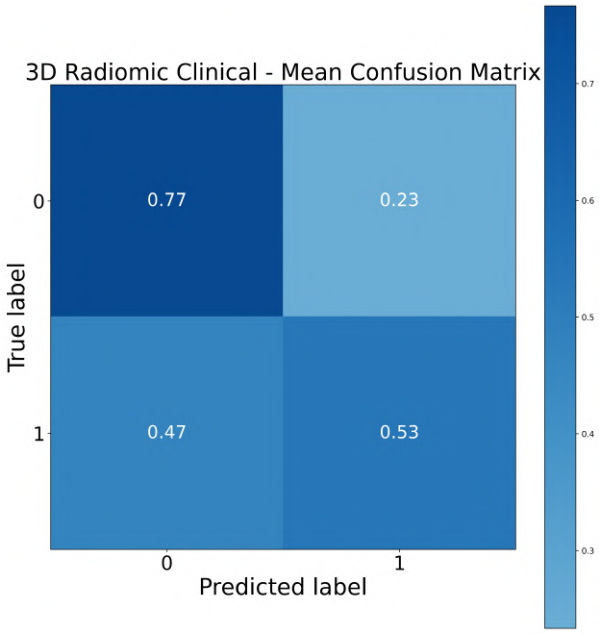


Figure 4.23: *Normalized mean confusion matrix of 3D Radiomic Clinical dataset across 10 Fold Split Stratified Cross-Validation*

4.2.4 2D Radiomic

The pipeline selected from TPOT for 2D Radiomic dataset consists of a polynomial feature expansion, kernel approximations, feature scaling, and ensemble-based transformations, followed by a probabilistic classification model.

```
exported_pipeline = make_pipeline( PolynomialFeatures(degree=2,
        include_bias=False, interaction_only=False),
    Nystroem(gamma=0.6000000000000001, kernel="linear",
        n_components=3),
    RobustScaler(),
    StackingEstimator(estimator=RandomForestClassifier(bootstrap=False,
        criterion="entropy", max_features=0.55, min_samples_leaf=18,
        min_samples_split=4, n_estimators=100)),
    ZeroCount(),
    BernoulliNB(alpha=1.0, fit_prior=False)
)
```

The polynomial feature transformation is applied as first step. It generates quadratic interaction terms between features, that can allow the model to capture non-linear dependencies. However, these terms significantly increase the dimensionality of the dataset

and can lead to redundancy. This is evident from the significant increase in feature number that follows the polynomial feature application: from 842 to 355745.

To address this issue, a Nystroem kernel approximation is applied, projecting the data into a lower-dimensional space via a linear kernel, thereby reducing the number of features to three.
The three new features are visualized in a 3D scatter plot, Figure 4.24, in particular for split 7 of the 10 fold split cross validation. These points, indicating the active and fibrotic classes, remain very close to each other, indicating a poor class separability in the transformed feature space.

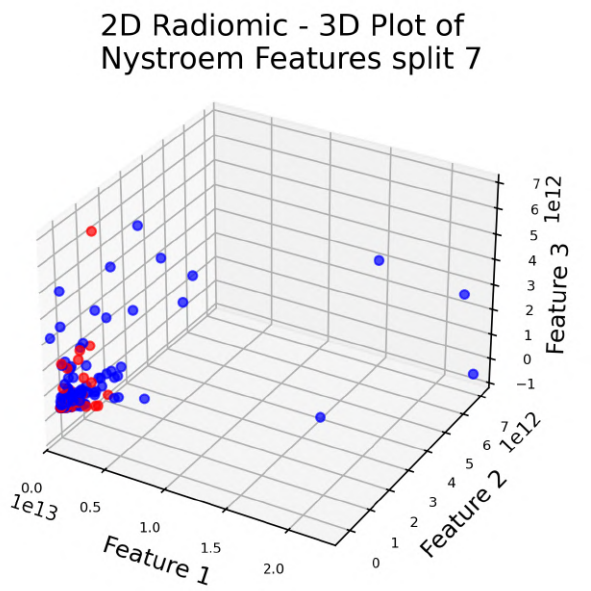


Figure 4.24: *3D feature plot using Nystroem kernel features for split 7 of the 10 fold cross-validation. Each axis represents one of the three features, with blue points indicating the active class and red points indicating the fibrotic class.*

After feature transformation, the data undergo robust scaling, which scales feature distributions based on the median and interquartile range.

Next, the Stacking Estimator applies a Random Forest Classifier (RFC), which generates new transformed features that are used as input for subsequent pipeline steps.
The RFC used here has 100 trees, with each terminal node requiring a minimum of 18 samples. At each split, only 55% of the total features are considered, and a node will only be further split if it contains at least 4 samples. This approach trains the Random Forest classifier on the input data and generates new features based on its learned patterns. As

a result, the dataset is expanded from three to six features, followed by the application of a ZeroCount transformation.

The final classification step is performed using Bernoulli Naïve Bayes (BernoulliNB). This model assumes that each feature follows a Bernoulli distribution, meaning it considers whether a feature is present (1) or absent (0) rather than its exact numerical value. The key hyperparameters are alpha=1.0, which corresponds to the Laplace smoothing parameter and prevents zero probabilities that could arise when a feature is absent in all training samples of a class, and fit_prior=False, meaning the model does not assume a prior class distribution to estimates class probabilities, relying only on data.
In Figure 4.25 a log-odds plot is used to visualize the probability estimates from a Naive Bayes classifier. In particular, the log-odds values for each feature indicate how much the feature contributes to distinguishing between the two types of lesions, and they are computed as mean values across 10 fold Stratified Group cross-validation.
In the plot, the feature with a log-odds value of 2.4 suggests that it strongly favors one class over the other, while other features have log-odds around -0.3, meaning they slightly favor the opposite class. This suggests that the feature with a log-odds of 2.4 is likely one of the most important predictors in the model.

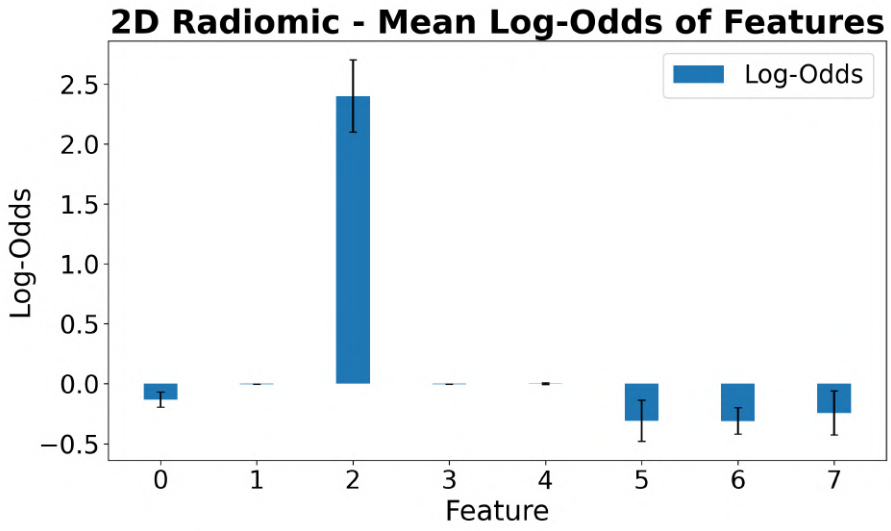


Figure 4.25: *Log-odds plot illustrating the probability estimates from a Naive Bayes classifier. The log-odds values are averaged across 10 Fold Stratified Group cross-validation.*

The performance of the pipeline is evaluated based on balanced accuracy and AUC scores, averaged over 10-fold cross-validation. The results, presented in Table 4.8, show values of balanced accuracy and AUC very close to 0.5 suggesting that the model has weak discriminative power.
The mean ROC curve (Figure 4.26) deviates slightly from the diagonal, indicating that the model learns some useful information, but not enough for reliable classification.

Table of mean balanced accuracy and AUC of 2D Radiomic dataset			
Train Balanced Accuracy	Train AUC	Test Balanced Accuracy	Test AUC
0.59 ± 0.01	0.62 ± 0.02	0.54 ± 0.14	0.56 ± 0.16

Table 4.8: *Table of mean balanced accuracy and AUC for test and train of 2D Radiomic dataset across 10 Fold Split Stratified Group Cross-Validation*

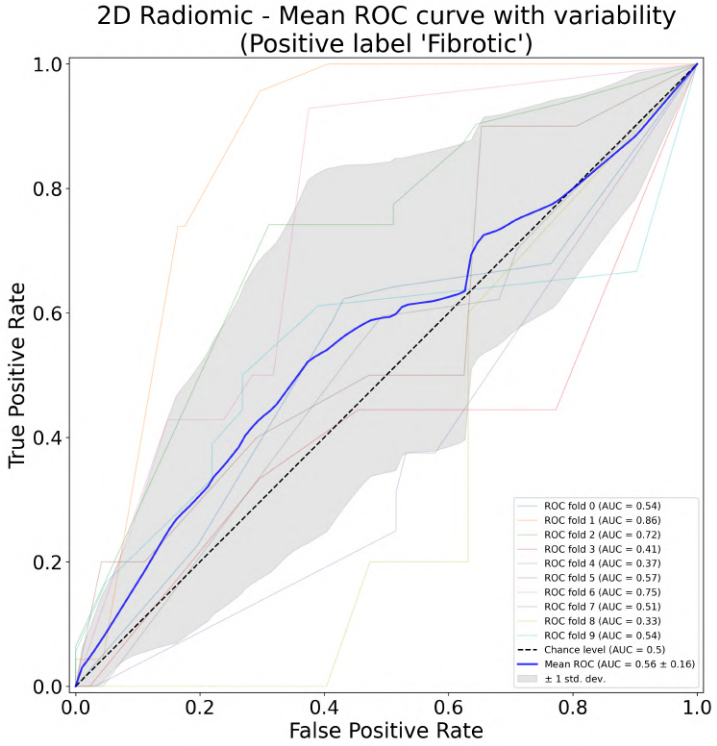


Figure 4.26: *Mean ROC-AUC curve of 2D Radiomic dataset across 10 Fold Split Stratified Group Cross-Validation*

From Table 4.9 and Figure 4.27 it can be said that the model correctly identifies active lesions more often than fibrotic ones, but misclassifies a substantial number of fibrotic lesions, leading to a high false negative rate for class 1. This means that fibrotic lesions are harder to identify, suggesting that the features extracted from the dataset may not effectively capture the distinguishing characteristics of fibrotic tissues.

Table of mean classification report 2D Radiomic			
	precision	recall	F1-score
0	0.79 ± 0.12	0.62 ± 0.14	0.69 ± 0.11
1	0.31 ± 0.21	0.46 ± 0.23	0.35 ± 0.22

Table 4.9: *Table of mean classification report across 10 Fold Split Stratified Group Cross-Validation for 2D Radiomic dataset*

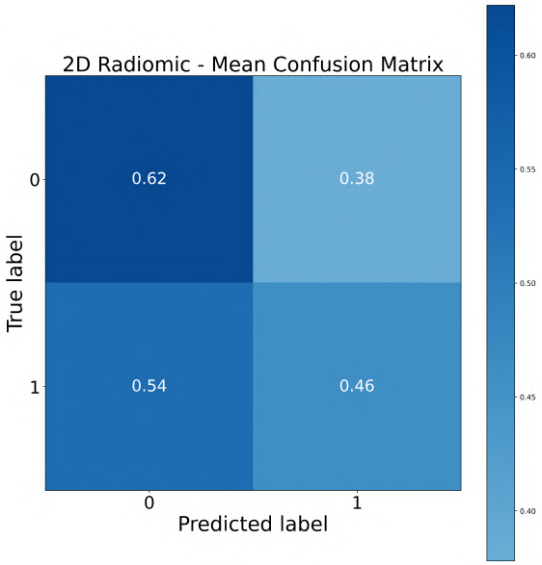


Figure 4.27: *Normalized mean confusion matrix of 2D Radiomic across 10 Fold Split Stratified Cross-Validation*

4.2.5 2D Clinical

The pipeline for 2D Clinical dataset integrates different steps as feature normalization, feature extraction, thresholding, and probabilistic classification.

```
exported_pipeline = make_pipeline(
    make_union( make_pipeline(
                MinMaxScaler(),
                StackingEstimator(estimator=MLPClassifier(
                    alpha=0.001, learning_rate_init=0.1)),
                Binarizer(threshold=0.7000000000000001)),
            StackingEstimator(estimator=MLPClassifier(alpha=0.001,
            learning_rate_init=0.1))),
    MultinomialNB(alpha=0.1, fit_prior=False)
)
```

The first step in data preprocessing is MinMax scaling, which normalizes all feature values within the range $[0, 1]$. After normalization, the pipeline utilizes a MLP classifier to extract relevant features. The Figure 4.28 illustrates the mean weights of the connections between the input layer and the first hidden layer. For the 2D Clinical dataset, the bar plot displays the feature weights sorted by their average values across ten data splits.
In this first hidden layer, the feature with the highest mean weight corresponds to a clinical variable indicating the absence of ovarian endometriomas in the patient. The second most relevant feature is associated with patients who have undergone bowel surgery,

specifically segmental bowel resection.

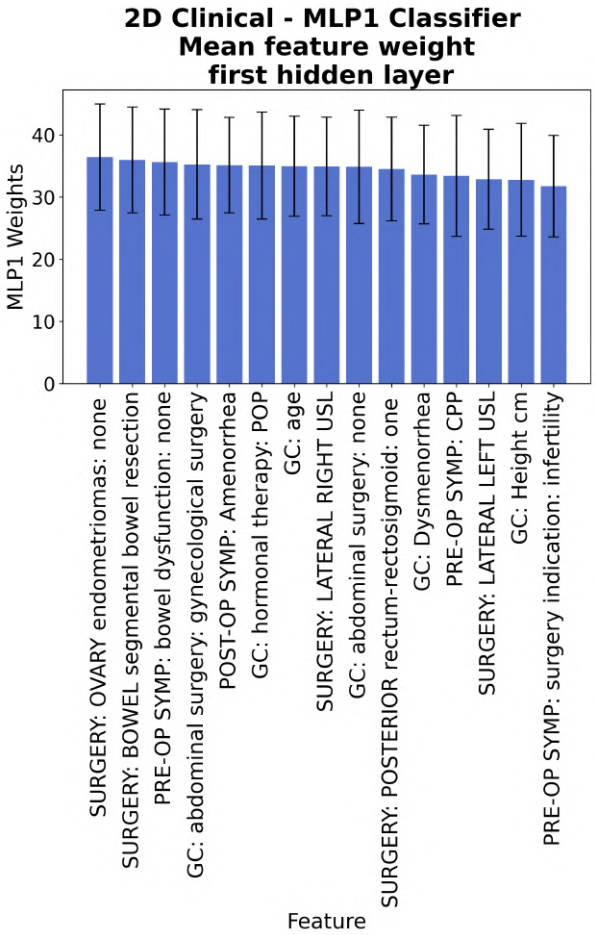


Figure 4.28: *Weights between the input layer and the first hidden layer of the first MLP neural network applied, sorted by mean feature weight across ten splits, for 2D Clinical dataset*

Next, the Binarization transformation applies a threshold of 0.7 to the feature values, converting to 1 any feature value above 0.7 and to 0 the others.

A second transformation using the MLP classifier is then applied, introducing additional features. The mean weights assigned to the connections between the input layer and the first hidden layer of this second MLP classifier are shown in Figure 23.
Compared to the previous MLP, the most important features are general patient characteristics such as weight, age, and height, similarly to the 3D clinical case.

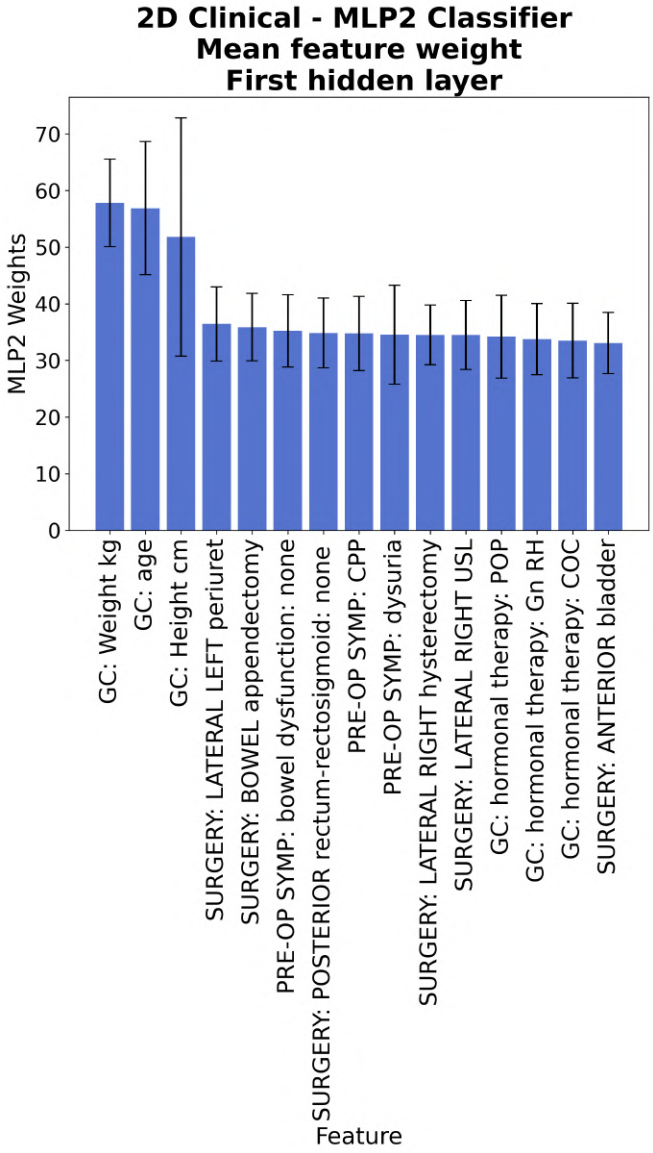


Figure 4.29: *Weights between the input layer and the first hidden layer of the second MLPClassifier applied, sorted by mean feature weight across ten splits, for 2D Clinical dataset*

The feature set is then passed to the Multinomial Naive Bayes classifier, implemented with alpha=0.1 Laplace smoothing, to prevent zero probabilities from occurring.

To visualize the classifier's probability estimates, a log-odds plot is presented in Figure 4.30. Similarly to the 2D Radiomic dataset, log-odds values for each feature are computed as the average across a 10 fold Stratified Group cross-validation. The features are ranked based on their mean log-odds values, and the top 15 are displayed.
The most influential feature for the first class is a clinical variable indicating that the patient has had prior deliveries. For the second class, the most relevant feature identifies patients who have undergone left lateral ultrasound-guided surgery. Additionally, a feature introduced by the MLP classifier is among the top 15, showing positive log-odds values.

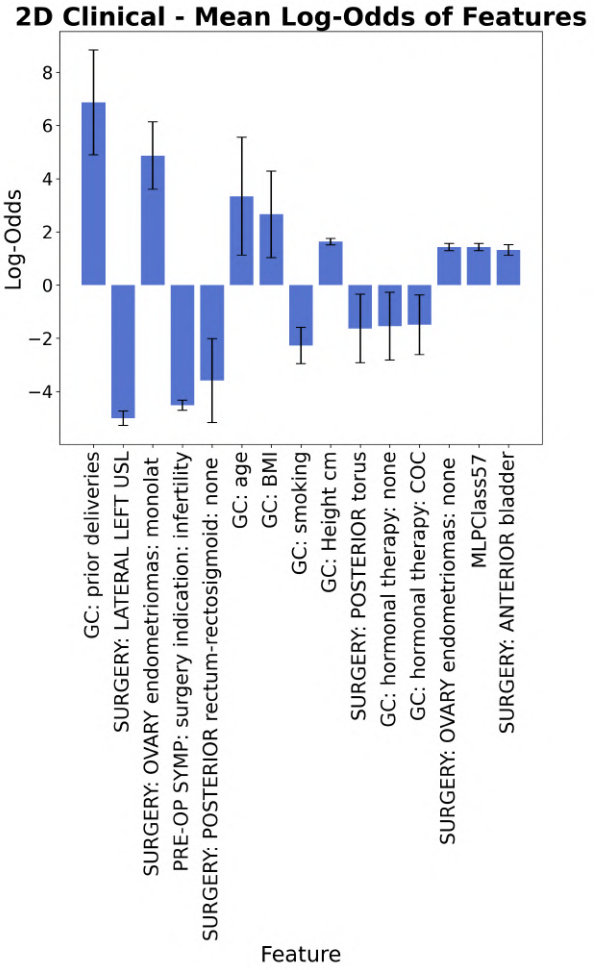


Figure 4.30: *Log-odds plot illustrating the probability estimates from a MultinomialNB classifier. The log-odds values are averaged across 10 Fold Stratified Group cross-validation.*

The performance of the pipeline is evaluated using the metrics reported in Table 4.10, computed as the mean values across a 10-fold cross-validation.
The balanced accuracy score is approximately 0.52 ± 0.19 indicating poor performance. The test AUC is slightly higher, reaching 0.59 ± 0.26. While this suggests a weak ability to discriminate between classes, the large standard deviation indicates high variability across splits.
The ROC-AUC shown in Figure 4.31, displays the mean ROC curve across all ten splits along with individual curves for each split. As for the previous datsaset also this curve has a slight but not significant deviation from the diagonal, and so a low predictive capability.

Table of mean balanced accuracy and AUC of 2D Clinical dataset			
Train Balanced Accuracy	Train AUC	Test Balanced Accuracy	Test AUC
0.78 ± 0.02	0.84 ± 0.02	0.52 ± 0.19	0.59 ± 0.26

Table 4.10: *Table of mean balanced accuracy and AUC for test and train of 2D Clinical dataset across 10 Fold Split Stratified Group Cross-Validation*

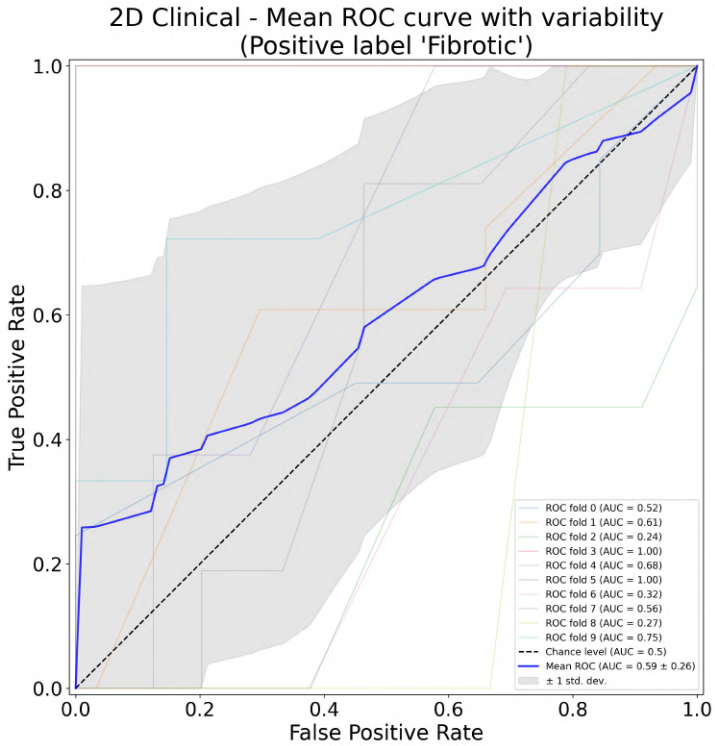


Figure 4.31: *Mean ROC-AUC curve of 2D Clinical dataset across 10 Fold Split Stratified Group Cross-Validation*

A more detailed performance analysis is provided in Table 4.11, which includes metrics such as precision, recall and F1-score. Additionally, the normalized confusion matrix in Figure 4.32 reveals that the mean value of correctly classified instances is 0.69 for class 0 and 0.36 for class 1. Misclassification is more frequent for class 1, with a value of 0.64. These results suggest that the model performs better at detecting active lesions but struggles significantly in identifying fibrotic cases.

Table of mean classification report 2D Clinical			
	precision	recall	F1-score
0	0.76 ± 0.22	0.69 ± 0.24	0.71 ± 0.21
1	0.17 ± 0.19	0.36 ± 0.43	0.21 ± 0.25

Table 4.11: *Table of mean classification report across 10 Fold Split Stratified Group Cross-Validation for 2D Clinical dataset*

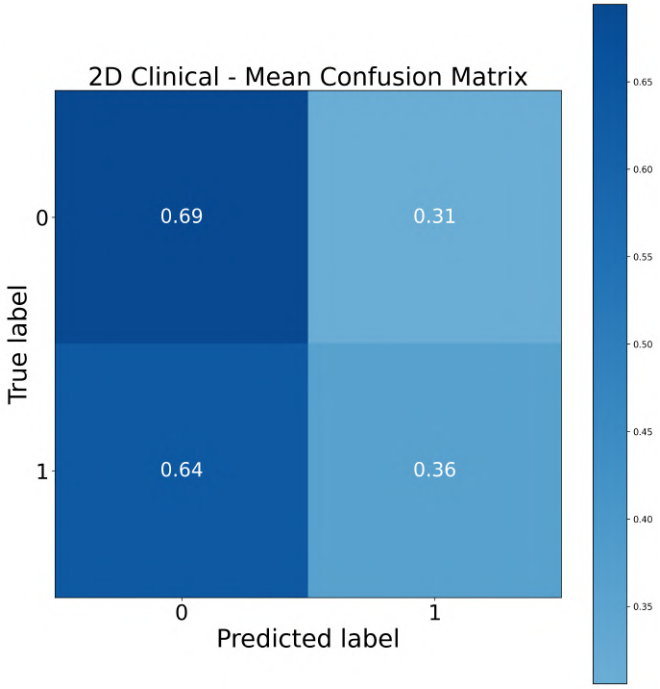


Figure 4.32: *Normalized mean confusion matrix of 2D Clinical across 10 Fold Split Stratified Cross-Validation*

The high misclassification rate indicates that the current feature set and modeling approach may not be well-suited for distinguishing between these conditions, requiring further refinement.

4.2.6 2D Radiomic Clinical

The proposed machine learning pipeline for the 2D Radiomic Clinical dataset consists of a Stacking Estimator using SGDClassifier, dimensionality reduction via FastICA, and a final classification step also employing SGDClassifier.

```
exported_pipeline = make_pipeline(
    StackingEstimator(estimator=SGDClassifier(alpha=0.0, eta0=0.01,
        fit_intercept=False, l1_ratio=1.0, learning_rate="constant",
        loss="perceptron", penalty="elasticnet", power_t=50.0)),
    FastICA(tol=0.9500000000000001),
    SGDClassifier(alpha=0.01, eta0=1.0, fit_intercept=False,
        l1_ratio=0.25, learning_rate="invscaling",
        loss="squared_hinge", penalty="elasticnet", power_t=0.5)
)
```

The first step applies a SGD classifier with a perceptron loss, which means that the model is trying to separate data with a hyperplane. Elastic net regularization is also employed, combining both L1 (LASSO) and L2 (ridge) penalties, in order to try to improve generalization. The model tries to select the most relevant features while reducing model complexity.
In particular, in Figure 4.33, the importance of each feature for the SGD classifier is visualized as a bar plot.
The importance score assigned to each feature corresponds to the weight attributed by the SGD classifier. These values can be interpreted as an indication of the feature's influence on classification, where higher absolute values represent stronger contributions toward either the positive or negative class.
The values shown represent the mean importance scores across 10 cross-validation splits. The presence of positive and negative values indicates the classifier's preference towards one class or the other. The plot highlights the 15 most important features, ranked based on their absolute mean importance. It can be observed that no clinical feature appears among the top-ranked variables, suggesting that the model mainly relies on radiomic features for classification.

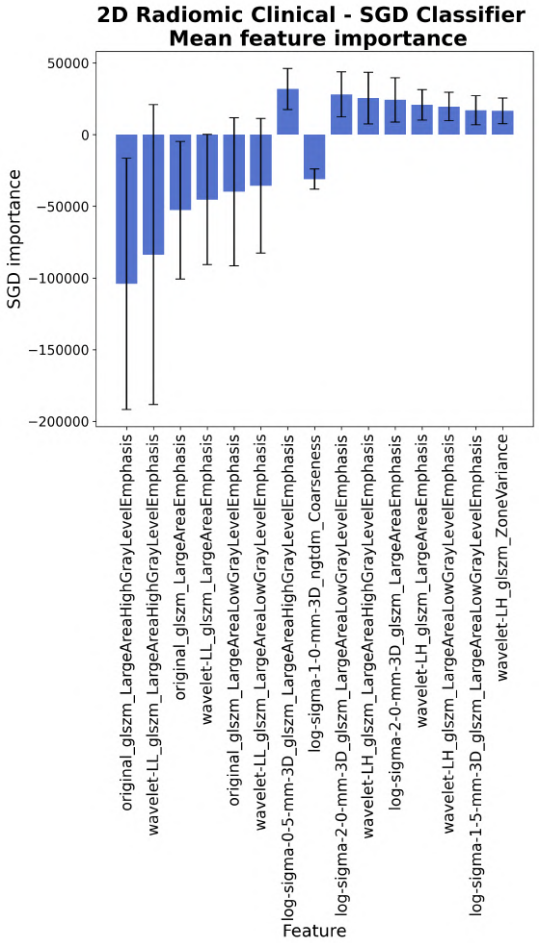


Figure 4.33: *The figure shows a bar plot of the importance assigned to each feature by the SGD classifier, specifically the top 15 in order of highest value. The values are averaged across the 10 Folds of cross-validation.*

Following feature selection, the pipeline applies FastICA, a technique designed to extract statistically independent components from the dataset to reduce feature redundancy. The consequence of this transformation is the loss of the original feature names, as the method generates new independent components. After applying FastICA, the total number of features is reduced from 897 to 778.
The transformed dataset is then passed to the final SGD classifier, which employs a squared hinge loss function. Once again, elastic net regularization is used, but with a different balance between L1 and L2 penalties. Additionally, the learning rate decreases over time, promoting more stable model optimization.

The pipeline performance is evaluated and shown in Table 4.12. The balanced accuracy and AUC have similar values, around 0.44, suggesting that the model shows insufficient predictive capability. This fact can be also observed in the ROC-AUC plot, Figure 4.34, where the mean curve approaches the diagonal trend.
A critical issue is highlighted by the training evaluation results, which show an extremely high balanced accuracy (0.99 ± 0.003) and AUC (0.99 ± 0.001). This discrepancy between training and test performance indicates severe overfitting.

Table of mean balanced accuracy and AUC of 2D Radiomic Clinical dataset			
Train Balanced Accuracy	Train AUC	Test Balanced Accuracy	Test AUC
0.99 ± 0.003	0.99 ± 0.001	0.44 ± 0.16	0.44 ± 0.21

Table 4.12: *Table of mean balanced accuracy and AUC for test and train of 2D Radiomic Clinical dataset across 10 Fold Split Stratified Group Cross-Validation*

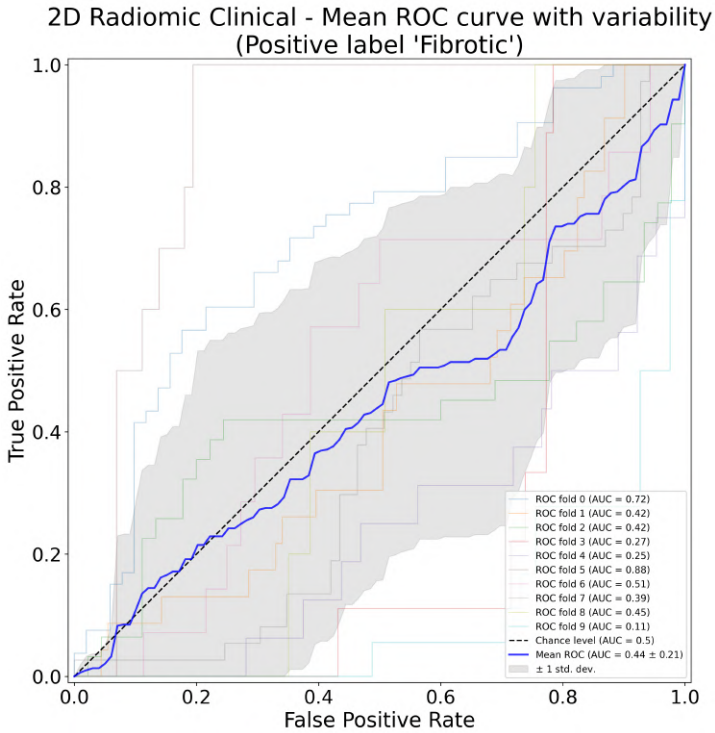


Figure 4.34: *Mean ROC-AUC curve of 2D Radiomic Clinical dataset across 10 Fold Split Stratified Group Cross-Validation*

In Table 4.13 are shown precision, recall, and F1-score values for the pipeline, and Figure 4.35 shows the normalized confusion matrix. The normalized mean number of correctly classified instances is 0.47 for active lesions and 0.41 for fibrotic cases. The non diagonal values result higher than the recall values, indicating that the model struggles

in the classification of both active and fibrotic lesions.

Table of mean classification report 2D Radiomic Clinical			
	precision	recall	F1-score
0	0.76 ± 0.15	0.47 ± 0.19	0.55 ± 0.16
1	0.21 ± 0.20	0.41 ± 0.31	0.25 ± 0.20

Table 4.13: *Table of mean classification report across 10 Fold Split Stratified Group Cross-Validation for 2D Radiomic Clinical dataset*

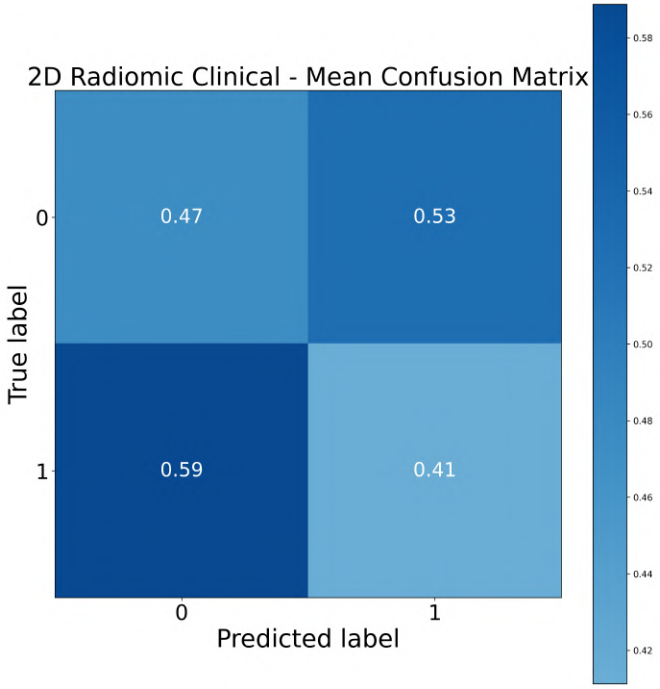


Figure 4.35: *Normalized mean confusion matrix of 2D Radiomic Clinical across 10 Fold Split Stratified Cross-Validation*

Chapter 5

Conclusions

This study investigated the classification of active and fibrotic lesions in deep endometriosis using machine learning models trained on radiomic and clinical features extracted from 3D MRI scans of 61 patients, as well as from the corresponding 2D slices. The comparative analysis across different datasets and feature sets provided key insights into the potential and limitations of these approaches.

Dimensionality reduction techniques such as LDA, PCA, UMAP and PaCMAP, failed to reveal a clear separation between lesion classes, suggesting that the extracted features do not have strong discriminative power.

Across all datasets, TPOT pipelines struggled to achieve robust discrimination between active and fibrotic lesions.
From a comparison of the TPOT pipelines between 3D and 2D datasets, they present almost the same low discriminative performance, with balanced accuracy and AUC values that remain around the random classification range of 0.5.
The highest performance is reached by the 3D Radiomic Clinical dataset with values of 0.65 ± 0.19 for balanced accuracy and of 0.60 ± 0.20 for AUC. The high error value must be taken into account, indicating a high variability of the model performance across the different cross-validation splits.

The balanced accuracy for 2D Radiomic Clinical (0.44 ± 0.16) is noticeably lower than its 3D counterpart, suggesting that including both radiomic and clinical features of the 2D lesion slices does not improve classification performance. However, when considering radiomic features alone, the balanced accuracy of 2D Radiomic (0.54 ± 0.14) is roughly the same as that of 3D Radiomic (0.56 ± 0.22), indicating that 2D radiomic information captures similar patterns to its 3D equivalent.
In contrast, the 2D Clinical dataset shows a slight improvement in balanced accuracy (0.52 ± 0.19) compared to 3D Clinical (0.45 ± 0.25), although with a high error margin.

Feature importance rankings varied between 2D and 3D datasets, but in general, models relied more on radiomic features than on clinical ones. Additionally, a comparison of training and test scores revealed overfitting in most models, as performance on training data was significantly better than on test data. Misclassification was particularly high for fibrotic lesions, likely due to the class imbalance in the dataset (64 active vs. 25 fibrotic lesions).

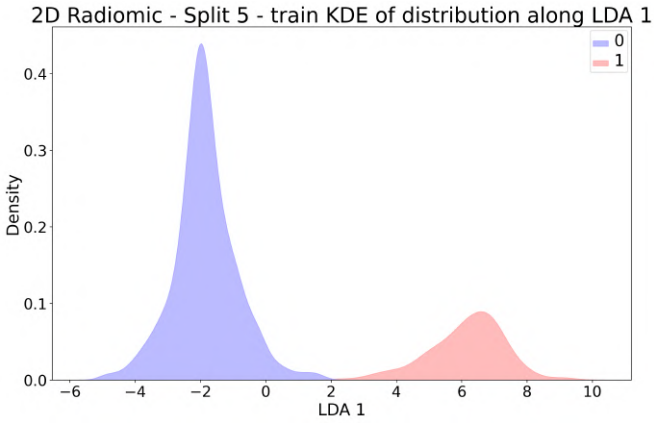
An important limitation of this study is the small sample size, which complicates the classification task. The dataset consists of 61 patients, with a total of 64 active lesions and 25 fibrotic lesions. This imbalance and the overall limited number of samples reduce the model's ability to generalize, increasing variability in performance.

Overall, these findings show that analyzing both 2D and 3D representations produced comparable classification results. However, the small sample size and class imbalance may have played a significant role in limiting the model's ability to differentiate between lesion types effectively.
In conclusion, this study highlights the challenges of classifying active and fibrotic lesions in deep endometriosis using radiomic and clinical features. Future work should explore additional radiomic features and alternative modeling approaches to improve classification performance and enhance the diagnostic potential of these methods.

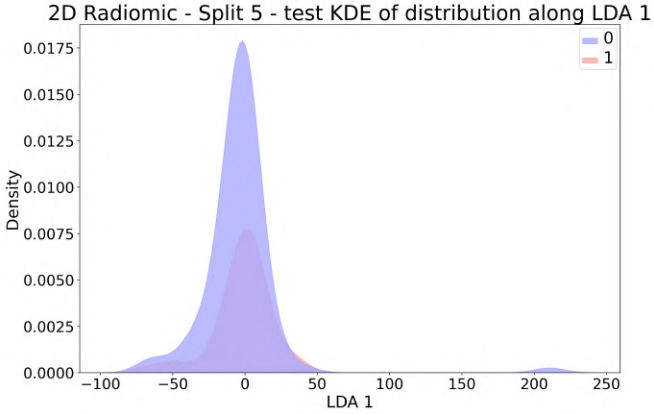
Appendix A

Appendix

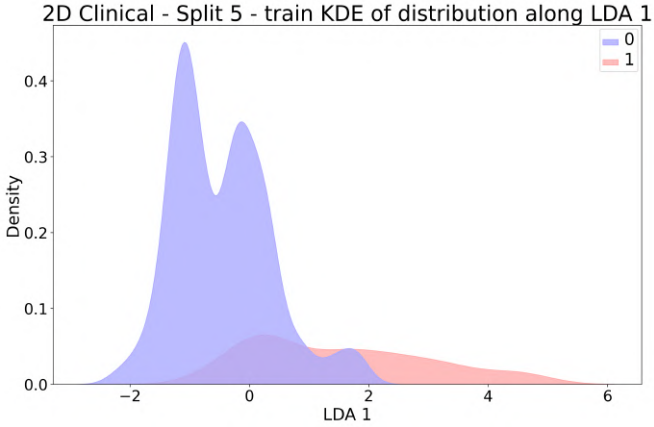
A.1 LDA KDE plot



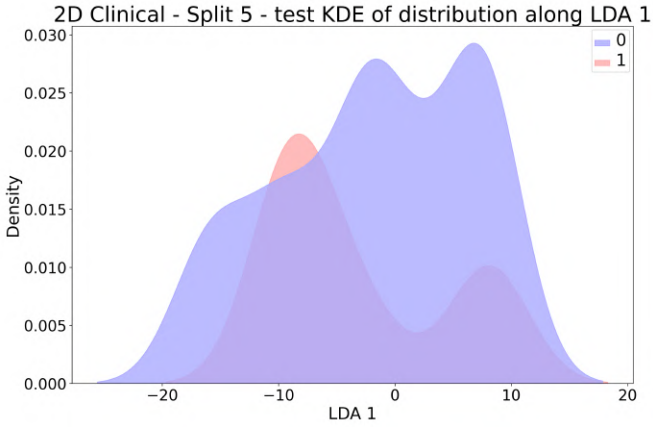
(a)



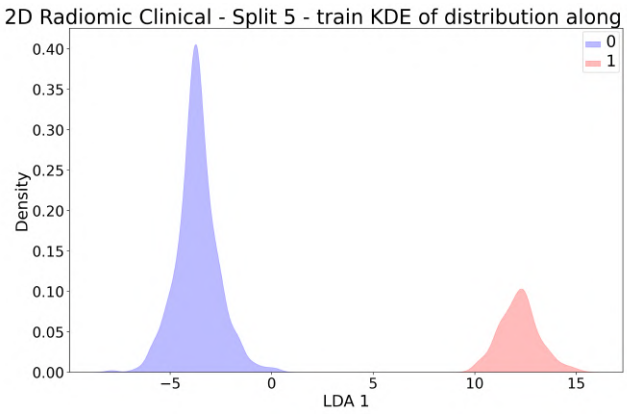
(b)



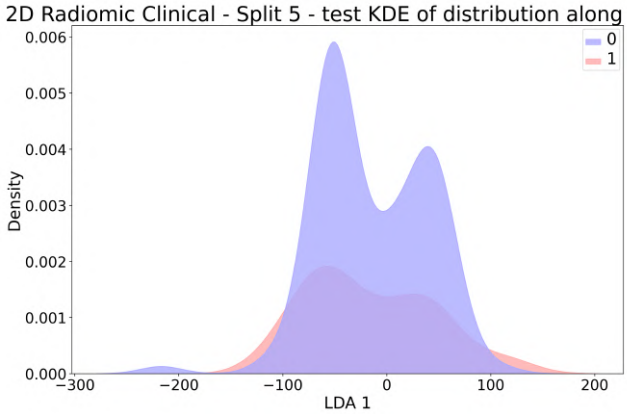
(c)



(d)

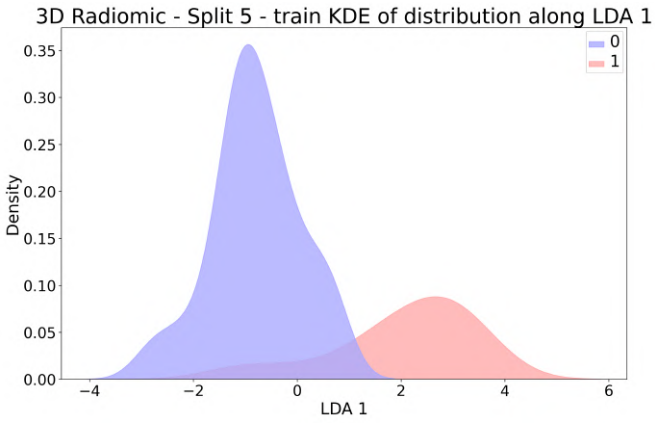


(e)

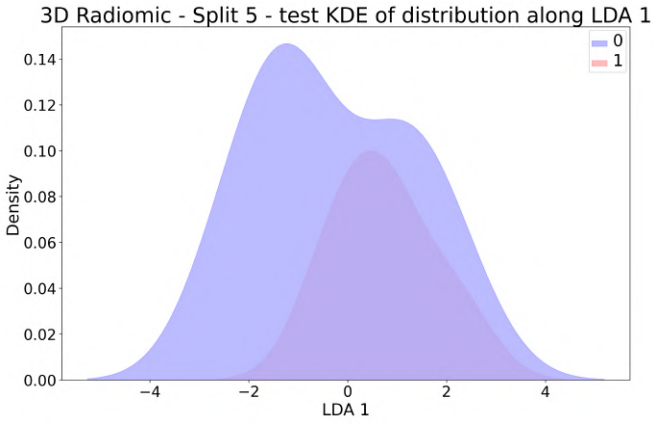


(f)

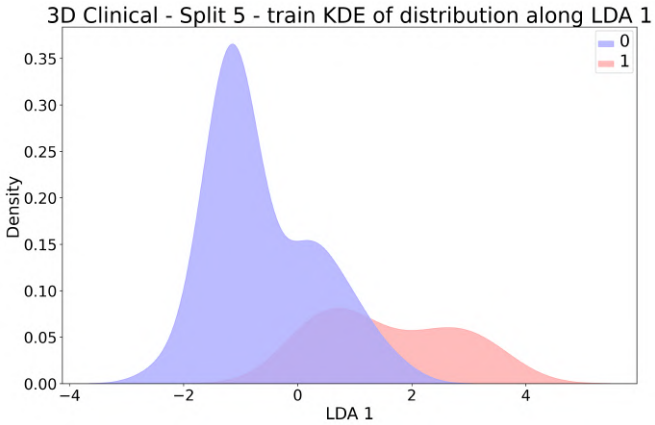
Figure A.2: *Distributions of class 0, active lesion, and class 1, fibrotic lesion, along LDA 1 for 2D datasets. (a)-(b) 2D Radiomic KDE plot on train/test set of split 5, (c)-(d) 2D Clinical KDE plot on train/test set of split 5, (e)-(f) 2D Radiomic Clinical KDE plot on train/test set of split 5*



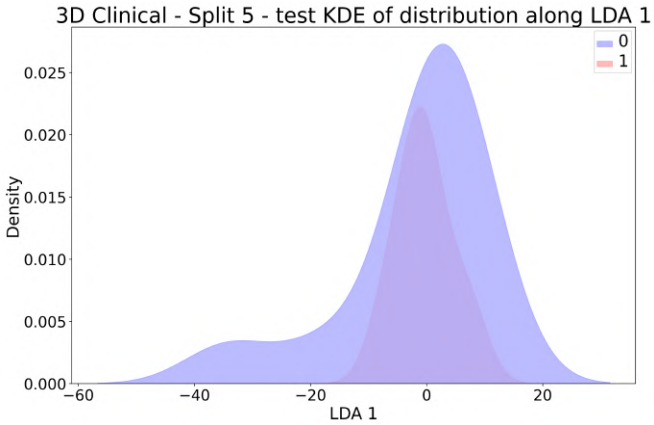
(a)



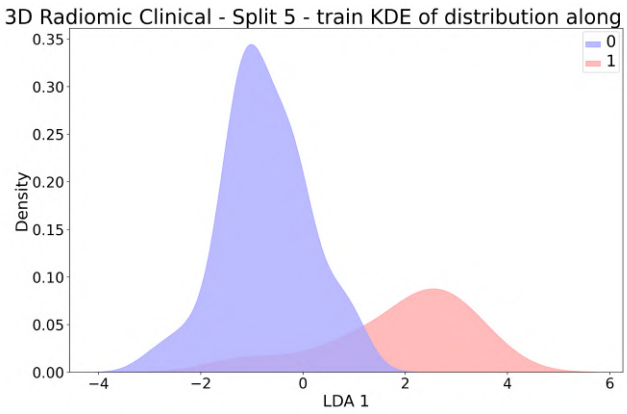
(b)



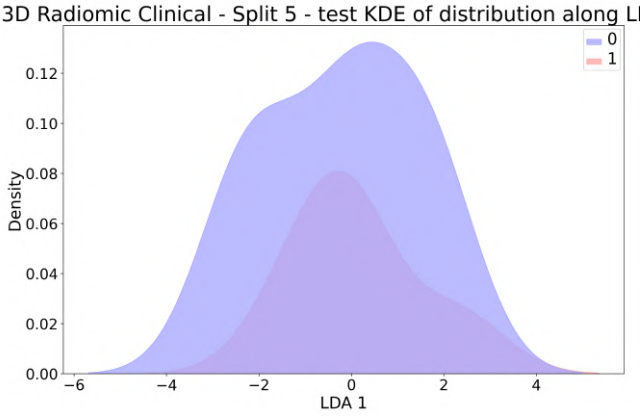
(c)



(d)



(e)



(f)

Figure A.4: *Distributions of active lesions (class 0), and fibrotic lesions (class 1), along LDA 1 for 3D datasets. (a)-(b) 3D Radiomic KDE plot on train/test set of split 5, (c)-(d) 3D Clinical KDE plot on train/test set of split 5, (e)-(f) 3D Radiomic Clinical KDE plot on train/test set of split 5*

A.2 UMAP Scatterplots

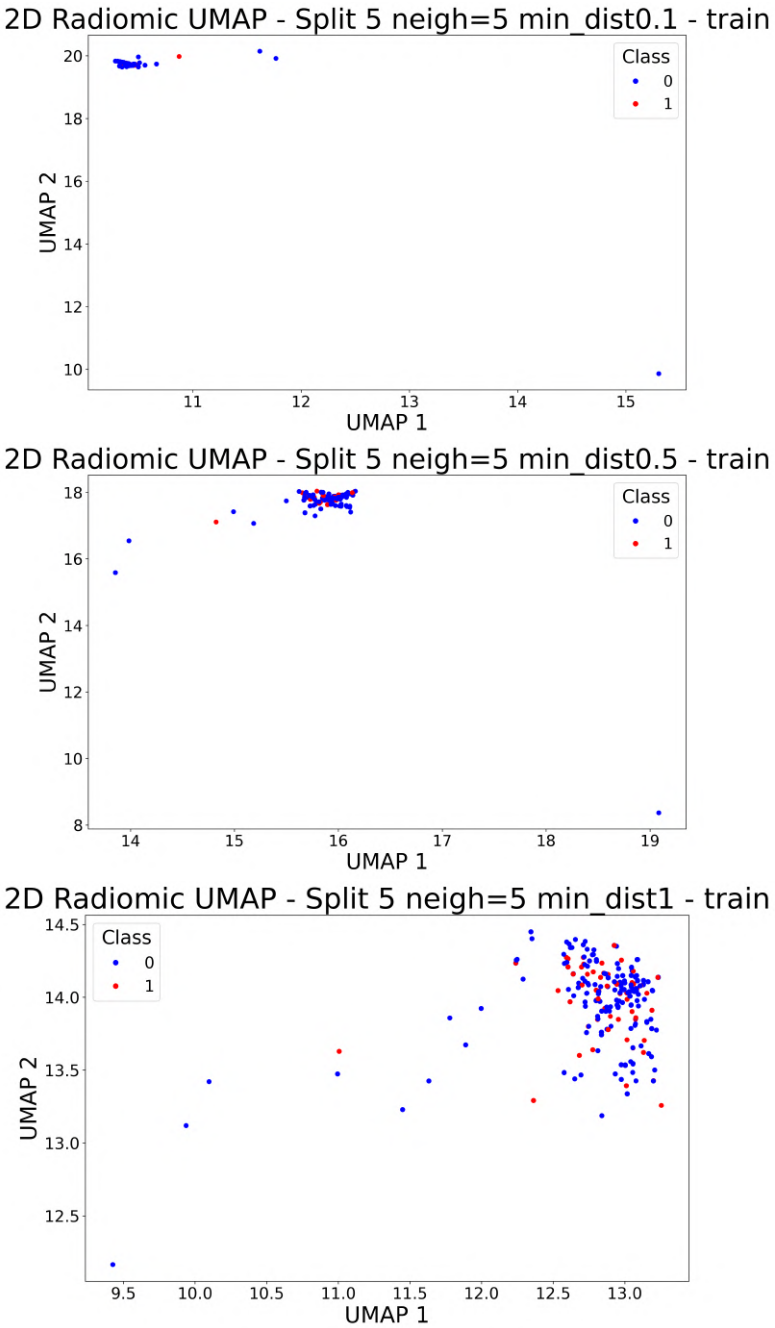


Figure A.5: *UMAP scatter plots of 2D Radiomic datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

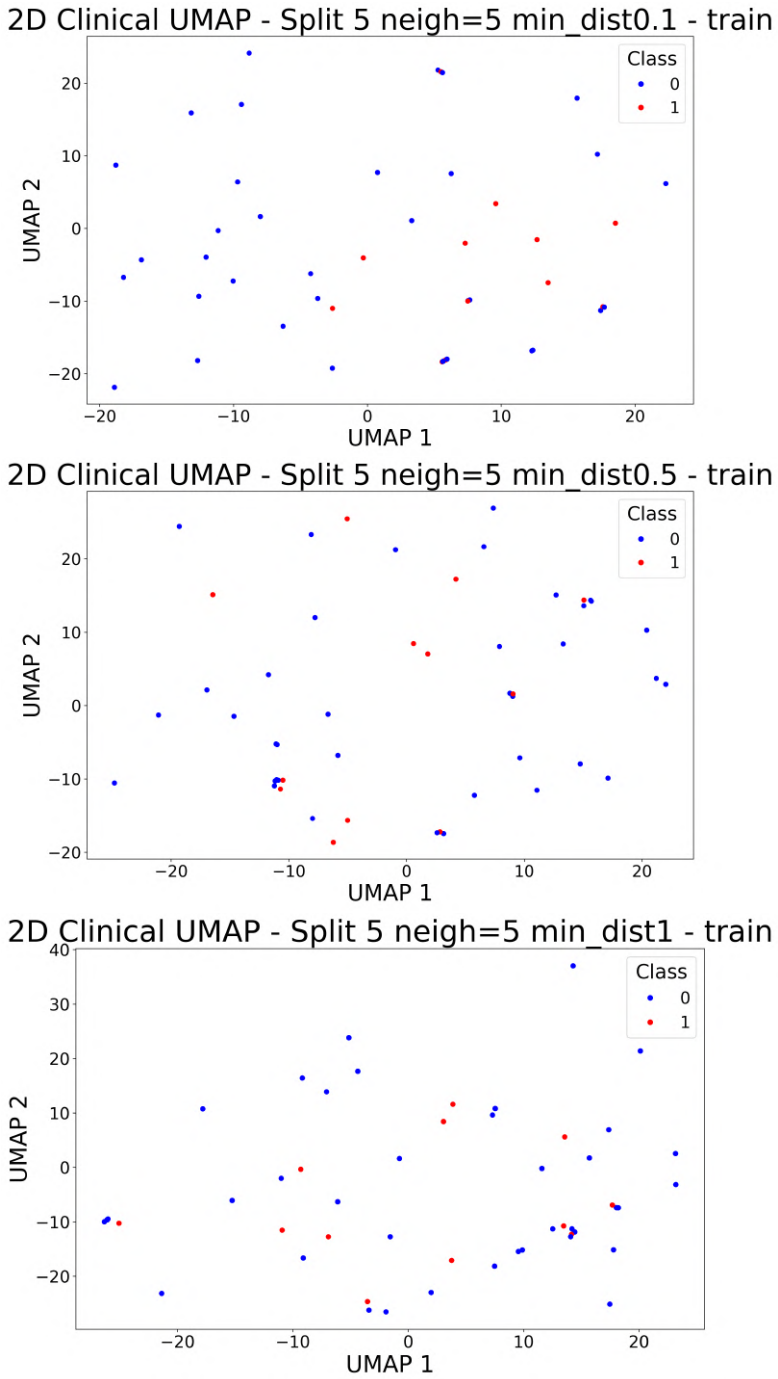


Figure A.6: *UMAP scatter plots of 2D Clinical datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

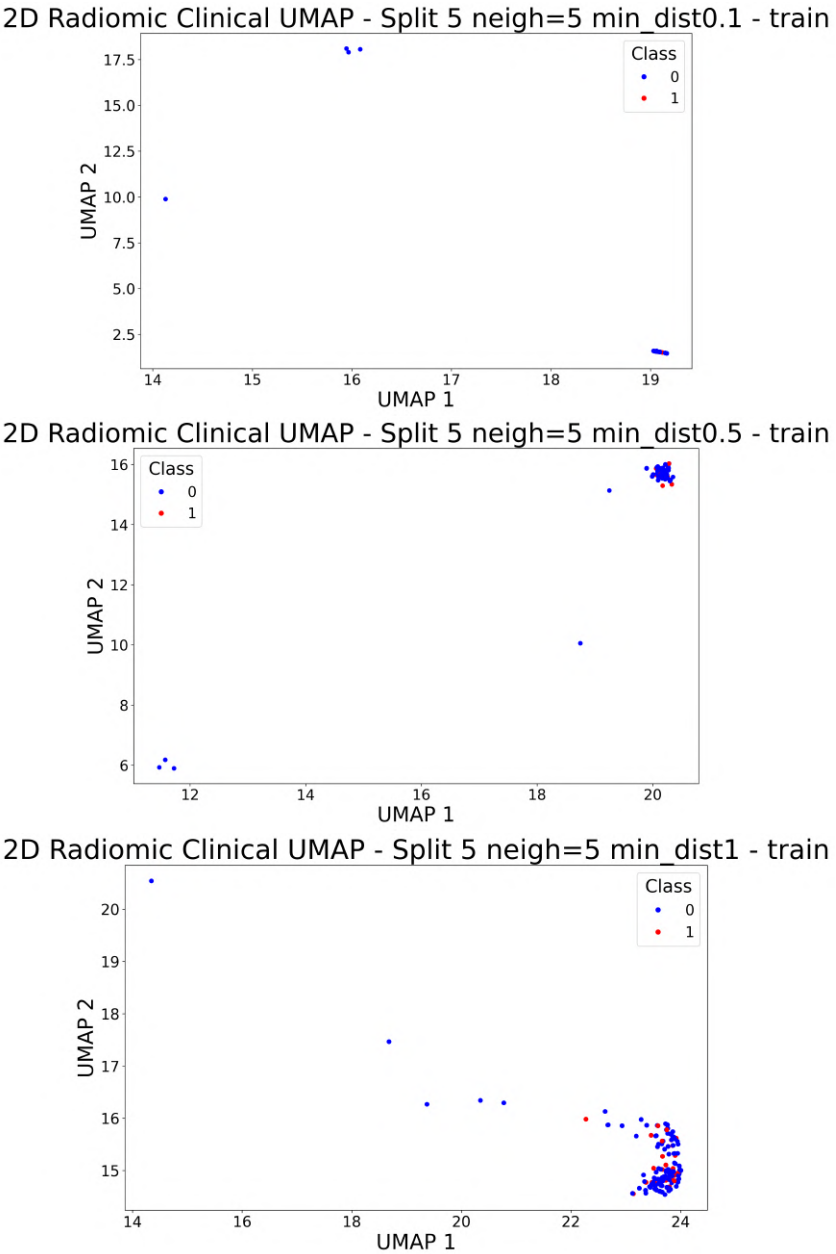


Figure A.7: *UMAP scatter plots of 2D Radiomic Clinical datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

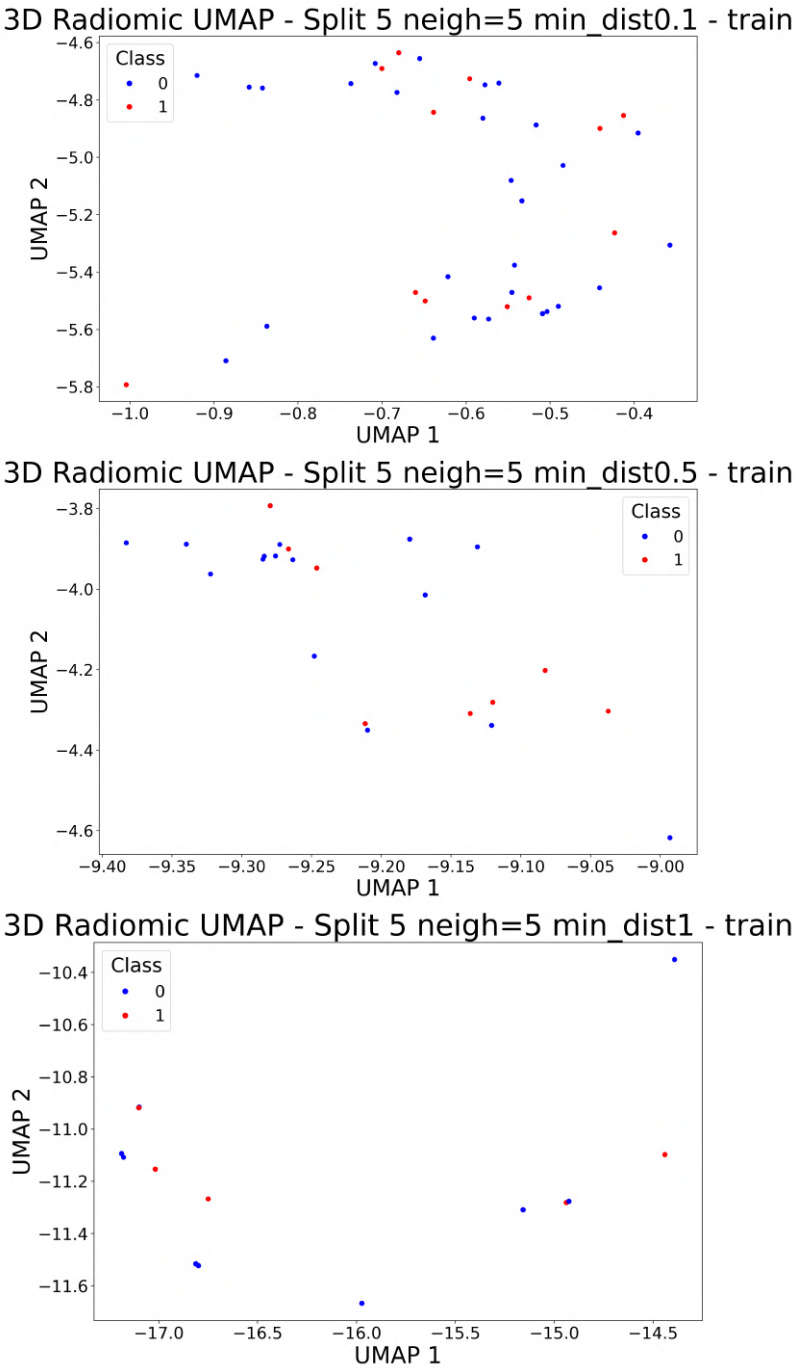


Figure A.8: *UMAP scatter plots of 3D Radiomic datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

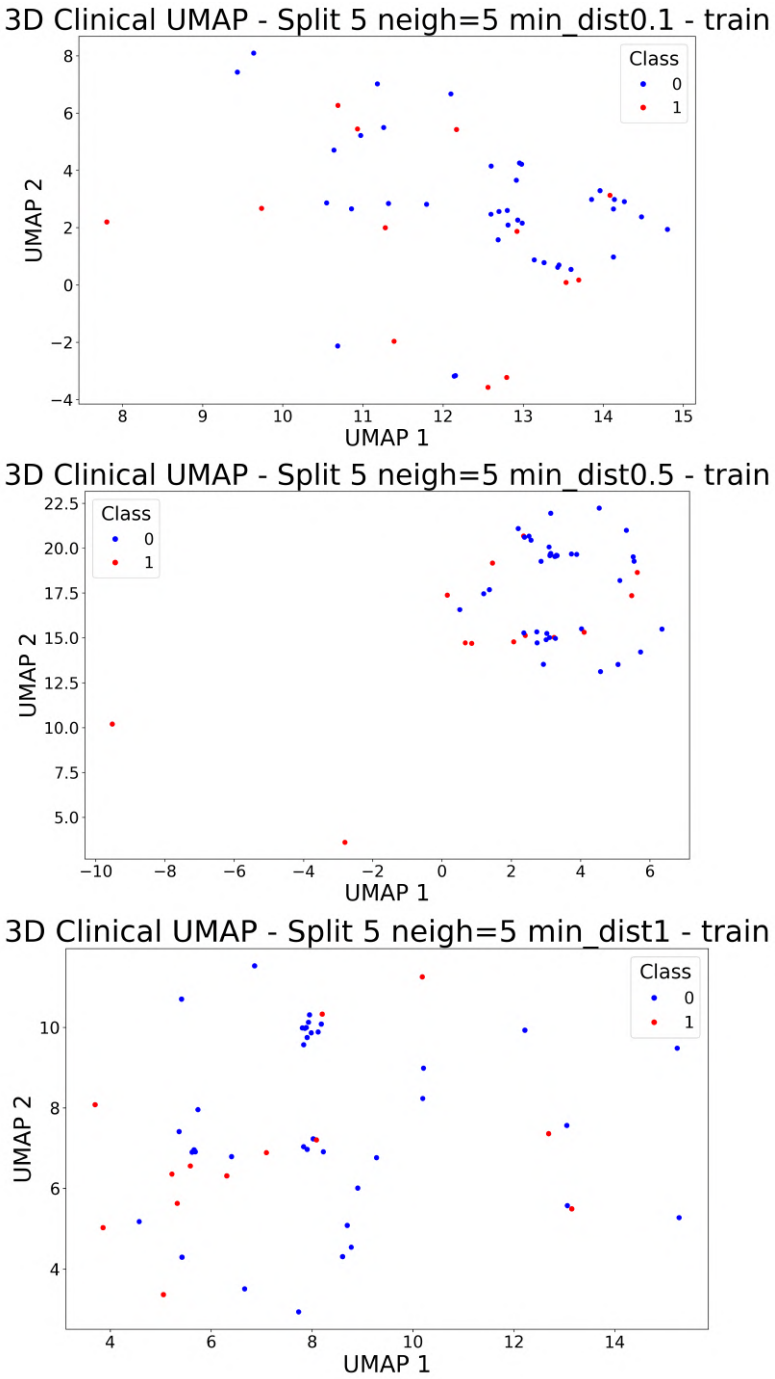


Figure A.9: *UMAP scatter plots of 3D Clinical datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

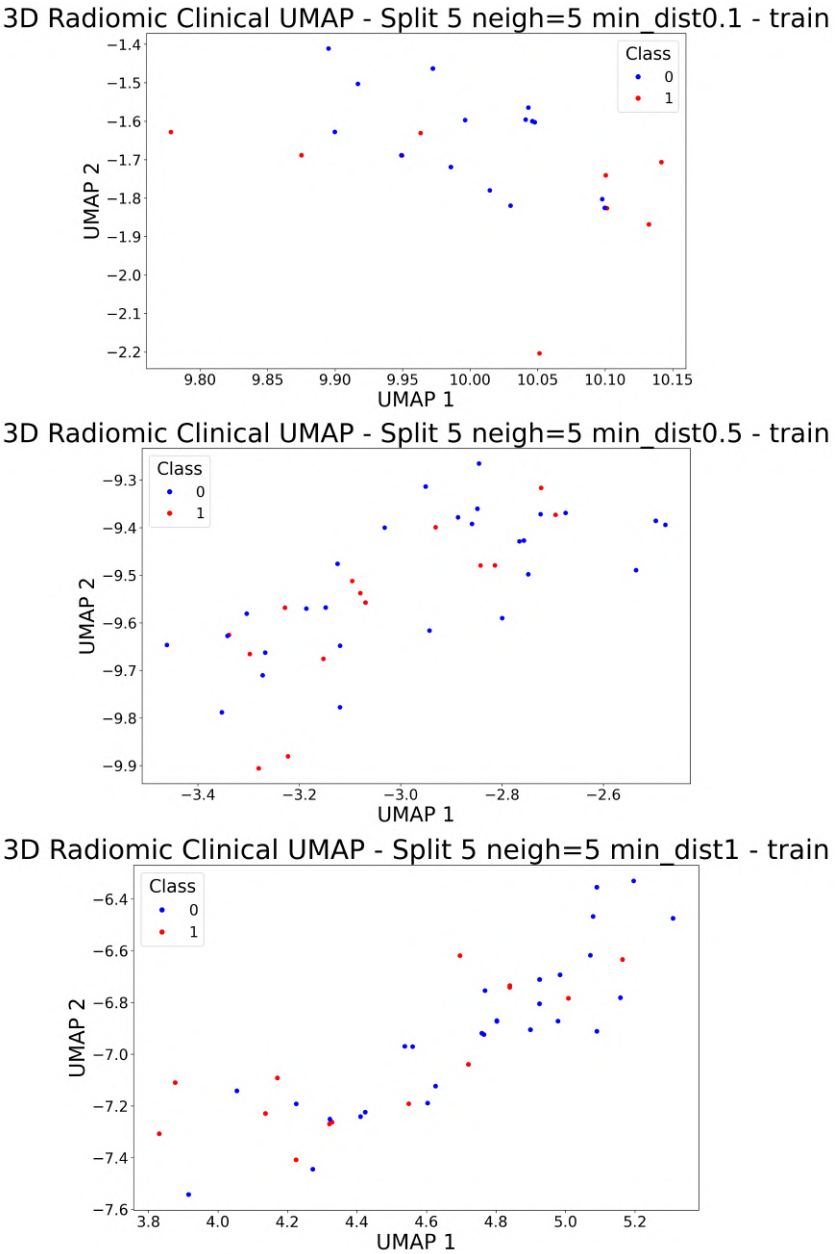


Figure A.10: *UMAP scatter plots of 3D Radiomic Clinical datasets, at different value of number of neighbors. Active class in blue, fibrotic class in red.*

A.3 PaCMAP scatterplots

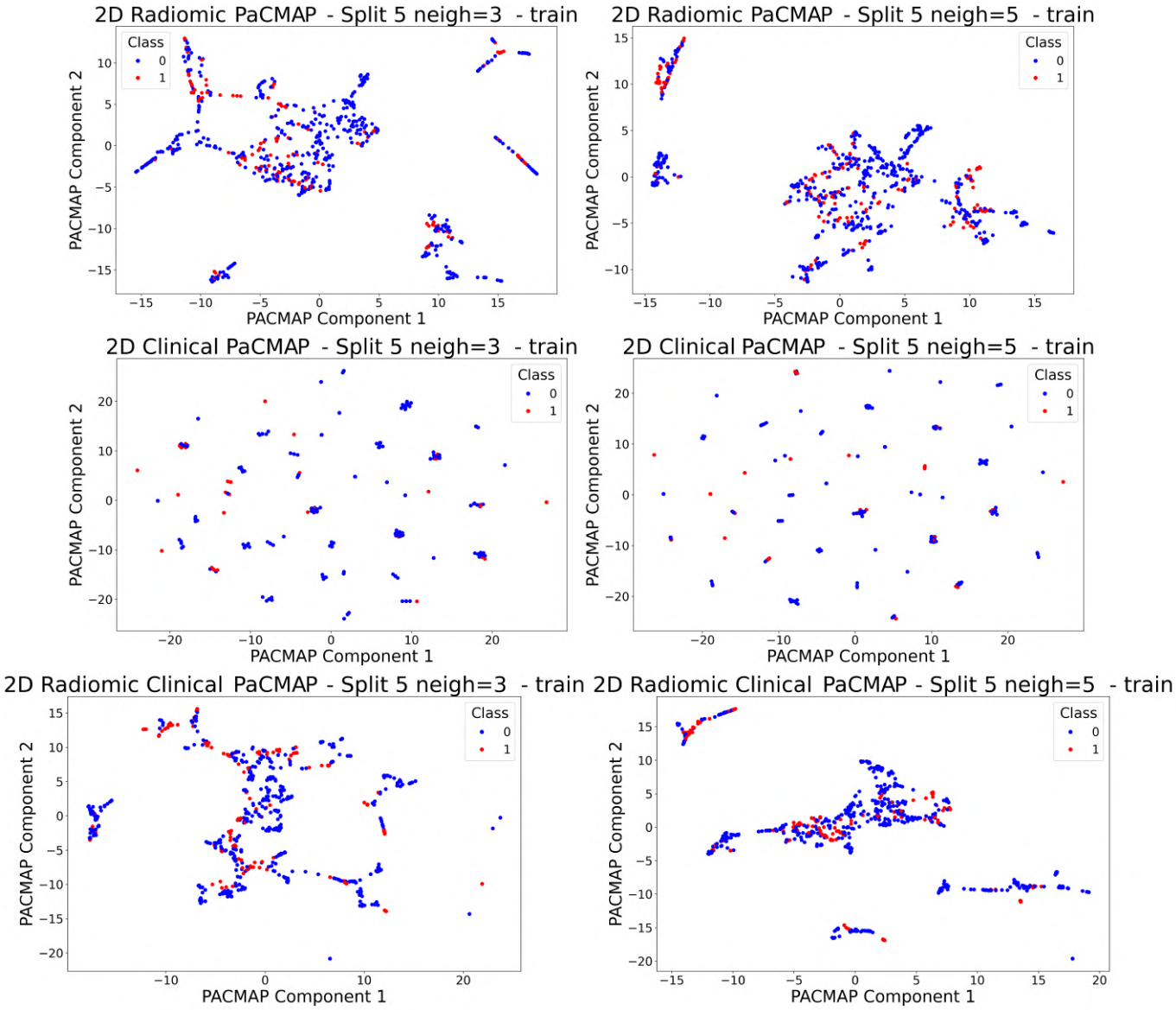


Figure A.11: *PaCMAP scatterplots for 2D datasets with three (left) and five (right) number of neighbors. Active class in blue, fibrotic class in red.*

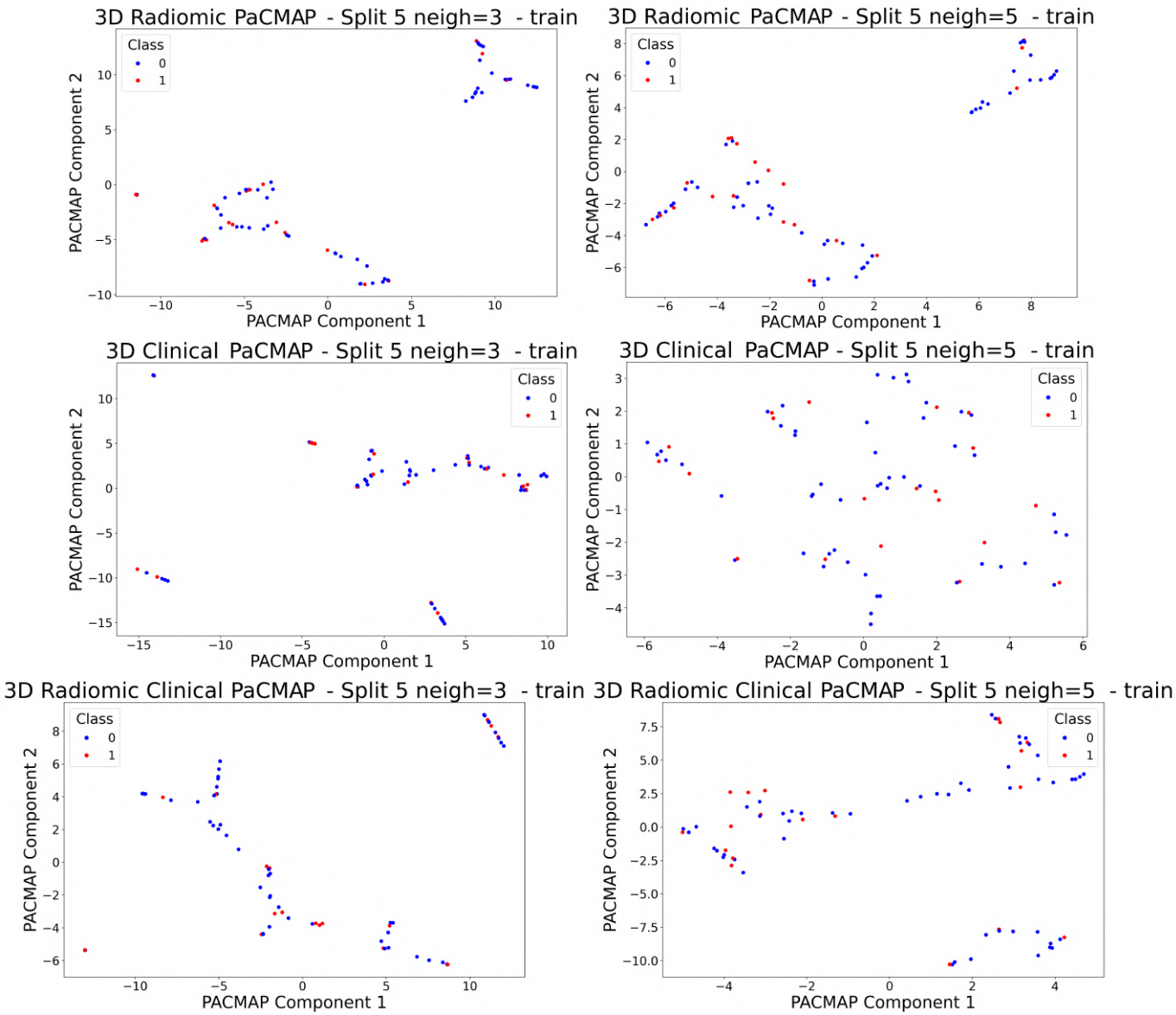


Figure A.12: *PaCMAP scatterplots for 3D datasets with three (left) and five (right) number of neighbors. Active class in blue, fibrotic class in red.*

Bibliography

[1] World Health Organization. *Endometriosis*. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/endometriosis.

[2] Dr. Pedro Barri Soldevila. *The latest in advanced managment of deep infiltrating endometriosis*. 2023. URL: https://myacare.com/blog/the-latest-in-advanced-management-of-deep-infiltrating-endometriosis.

[3] Ludovica Imperiale et al. "Three Types of Endometriosis: Pathogenesis, Diagnosis and Treatment. State of the Art". In: *Journal of Clinical Medicine* 12 (Jan. 2023), p. 994. DOI: 10.3390/jcm12030994.

[4] Philippe Koninckx et al. "Deep endometriosis: Definition, diagnosis, and treatment". In: *Fertility and sterility* 98 (Sept. 2012), pp. 564–71. DOI: 10.1016/j.fertnstert.2012.07.1061.

[5] Dr. Shiva Harikrishnan. *Deep Infiltrating Endometriosis*. 2022. URL: https://www.drshivahk.com/deep-infiltrating-endometriosis/.

[6] Brosens IA Wiegerinck MA Van Dop PA. "The staging of peritoneal endometriosis by the type of active lesion in addition to the revised American Fertility Society classification". In: *Fertil Steril* (1993). DOI: 10.1016/s0015-0282(16)56161-5.

[7] The Leeds Teaching Hospitals NHS Trust. *Endometriosis*. Version 1.0. 2023.

[8] Daraï E Bazot M. "Diagnosis of deep endometriosis: clinical examination, ultrasonography, magnetic resonance imaging, and other techniques". In: *Fertil Steril* (2017). DOI: 10.1016/j.fertnstert.

[9] Salomeh Salari, Kathryn Coyne, and Rebecca Flyckt. "Deep Infiltrating Endometriosis: Diagnosis and Fertility-Sparing Management in the ART Patient". In: *Reproductive Surgery: Current Techniques to Optimize Fertility*. Ed. by Steven R. Lindheim and John C. Petrozza. Springer International Publishing, 2022. DOI: 10.1007/978-3-031-05240-8_20. URL: https://doi.org/10.1007/978-3-031-05240-8_20.

[10] Jin ZY Xian JF Chen M. "Magnetic resonance imaging in clinical medicine: current status and potential future developments in China". In: *Chin Med J (Engl)* (2015). DOI: 10.4103/0366-6999.151637.

[11] Claudia Testa. *MRI Lesson III*. Physics in Neuroscience and Medicine, University of Bologna. 2022.

[12] mrimaster.com. *T1 vs T2 MRI*. URL: https://mrimaster.com/t1-vs-t2-mri/.

[13] Marc Bazot et al. "Deep Pelvic Endometriosis: MR Imaging for Diagnosis and Prediction of Extension of Disease1". In: *Radiology* 232 (Sept. 2004), pp. 379–89. DOI: 10.1148/radiol.2322030762.

[14] A.C. Müller and S. Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016. ISBN: 9781449369897.

[15] Sara Brown. *Machine learning, explained*. 2021. URL: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[16] Amin Zollanvari. *Machine learning with Python: Theory and implementation*. Publisher Copyright: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023. All rights reserved. Springer International Publishing, 2023. ISBN: 9783031333415. DOI: 10.1007/978-3-031-33342-2.

[17] Prof. Mai Vu. *Lecture 1: Entropy and mutual information*. EE194 – Network Information Theory, Tufts University. 2022.

[18] IBM. *What is principal component analysis (PCA)?* 2023. URL: https://www.ibm.com.

[19] Yingfan Wang et al. *Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization*. 2021. arXiv: 2012.04456 [cs.LG]. URL: https://arxiv.org/abs/2012.04456.

[20] IBM. *What is a machine learning pipeline?* URL: https://www.ibm.com.

[21] EIT Health. *Machine learning in healthcare: Uses, benefits and pioneers in the field*. 2024. URL: https://eithealth.eu/news-article/machine-learning-in-healthcare-uses-benefits-and-pioneers-in-the-field/.

[22] Brintha Sivajohan et al. "Clinical use of artificial intelligence in endometriosis: a scoping review". In: *npj Digital Medicine* 5 (Aug. 2022), p. 109. DOI: 10.1038/s41746-022-00638-1.

[23] Nguyen T. Thalluri AL Knox S. "MRI findings in deep infiltrating endometriosis: A pictorial essay". In: *J Med Imaging Radiat Oncol.* (2017). DOI: 10.1111/1754-9485.12680..

[24] A. Zollanvari. *Machine Learning with Python: Theory and Implementation*. Springer International Publishing, 2023. ISBN: 9783031333422. URL: https://books.google.it/books?id=f_zKEAAAQBAJ.

[25] Asset Management Factory. *What is the difference between feature extraction and feature selection?* 2019. URL: https://quantdare.com/what-is-the-difference-between-feature-extraction-and-feature-selection/.

[26] EIT Health. *Learning Model Building in Scikit-learn*. 2025. URL: https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/.

[27] Dagang Wei. *Essential Math for Machine Learning: Kernel Density Estimation*. 2024. URL: https://medium.com/@weidagang/essential-math-for-machine-learning-kernel-density-estimation-d014df073770.

[28] Travis Tang. *Automate Your Machine Learning Training Process with TPOT*. 2021. URL: https://medium.com/analytics-vidhya/automate-your-machine-learning-training-process-97e63c584716.

[29] Jason H. Moore et al. "Genetic Programming as an Innovation Engine for Automated Machine Learning: The Tree-Based Pipeline Optimization Tool (TPOT)". In: *Handbook of Evolutionary Machine Learning*. Ed. by Wolfgang Banzhaf, Penousal Machado, and Mengjie Zhang. Singapore: Springer Nature Singapore, 2024, pp. 439–455.

[30] GeeksforGeeks. *Radial Basis Function Kernel – Machine Learning*. 2024. URL: https://www.geeksforgeeks.org/radial-basis-function-kernel-machine-learning/.

[31] Akhil Soni. *Linear SVM Classification*. 2023. URL: https://medium.com/@akhil0435/linear-svm-classification-40dde297c931.

[32] Scikit-learn. *scikit-learn Machine Learning in Python*. 2023. URL: https://scikit-learn.org/stable/index.html#.

[33] IBM. *Che cos'è la foresta casuale?* URL: https://www.ibm.com/it-it/topics/random-forest.

[34] Sejal Jaiswal. *Multilayer Perceptrons in Machine Learning: A Comprehensive Guide*. 2024. URL: https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning?dc_referrer=https%3A%2F%2Fwww.google.com%2F.

[35] GeeksforGeeks. *ML — Stochastic Gradient Descent (SGD)*. 2024. URL: https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/.

[36] GeeksforGeeks. *Gradient Boosting in ML*. 2023. URL: https://www.geeksforgeeks.org/ml-gradient-boosting/.

[37] GeeksforGeeks. *Multinomial Naive Bayes*. 2025. URL: https://www.geeksforgeeks.org/multinomial-naive-bayes/?ref=ml_lbp.

[38] IBM. *Cosa sono i classificatori Naïve Bayes?* URL: https://www.ibm.com/it-it/think/topics/naive-bayes.

[39] Randal S. Olson et al. “Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science”. In: *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. GECCO ’16. Denver, Colorado, USA: ACM, 2016, pp. 485–492. ISBN: 978-1-4503-4206-3. DOI: 10.1145/2908812.2908918. URL: http://doi.acm.org/10.1145/2908812.2908918.

[40] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[41] Marius Mayerhoefer et al. “Introduction to Radiomics”. In: *Journal of Nuclear Medicine* 61 (Feb. 2020), jnumed.118.222893. DOI: 10.2967/jnumed.118.222893.

[42] pyradiomics. *Radiomic Features*. URL: https://pyradiomics.readthedocs.io/en/latest/features.html#.