POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Symptomatic and Asymptomatic Carotid Plaque Classification: An Integrated Approach Using Radiomic and Deep Learning Features

Tesi di Laurea Magistrale in
Computer Engineering - Ingegneria Informatica

Authors: **Stefano Baroni**, **Alessia Menozzi**

Student IDs: 10939098, 10684261
Advisor: Prof. Anna Corti
Co-advisors: Prof. Valentina Corino
Academic Year: 2023-24

Abstract

Ischemich stroke is one of the leading causes of vascular morbidity and mortality worldwide, with approximately 12 million people affected every year. 10-15% of all ischemic strokes are due to carotid atherosclerotic plaque rupture. Thus, the early identification of vulnerable carotid plaques, namely plaques at risk of rupture, is crucial for preventing acute cerebrovascular events and stroke and for guiding effective clinical intervention. This study is based on a dataset of 129 patients, of which 53 symptomatic (i.e., subjected to transient ischemic attack or stroke), who underwent computer tomography imaging, prior to carotid endarterectomy procedure. Accordingly, the aim of the present study is the identification of image-based plaque features that effectively stratify symptomatic and asymptomatic patients, thus characterizing plaque vulnerability. Radiomics is applied to extract quantitative features from CT images. Furthermore, deep learning-derived features are extracted from several pre-trained convolutional neural networks including ResNet50, InceptionV3, Inception-ResNetV2, and VGG19. Radiomic, deep and combined features are then tested in a machine learning framework. To address data imbalance in the training set, SMOTE oversampling is applied, along with cross-validation. Two types of slices, full and cropped around the plaque, as well as with four feature selection techniques and six machine learning classifiers are used to identify the best-performing combinations. Finally ensemble models incorporating all the best models for radiomics and deep networks are tested to enhance classification accuracy. The analysis is conducted first on a single slice per carotid, the one presenting the largest plaque area (2D approach). Then, multiple slices, presenting with a plaque area $> 30\%$ of the largest slice, were included in the analysis (2.5D approach). The best classification performance was obtained with the 2.5D approach on VGG19 combined with radiomics, with a ROC AUC of 0.889 and a balanced accuracy of 0.898. To the best of our knowledge, this is the first study that uses deep learning networks applied to CTA scans and tries a combination of radiomic and deep features. Moreover, the obtained excellent stratification results demonstrated the potentials of the proposed approach. In future, further validation on larger, multicentric and prospective dataset will be pursued to confirm the findings.

Keywords: Carotid artery plaques, Atherosclerosis, Radiomics, Deep Learning, Machine Learning, Artificial Intelligence, CTA imaging, Feature Combination, Symptomatic and Asymptomatic Classification

Abstract in lingua italiana

L'ictus ischemico è una delle principali cause di morbilità e mortalità vascolare a livello mondiale, con circa 12 milioni di persone colpite ogni anno. Il 10-15% di questi eventi è dovuto alla rottura di placche aterosclerotiche carotidee. Pertanto, l'identificazione precoce delle placche carotidee vulnerabili, ovvero quelle a rischio di rottura, è cruciale per prevenire eventi cerebrovascolari e ictus, nonché per guidare interventi clinici efficaci. Questo studio si basa su un dataset di 129 pazienti, di cui 53 sintomatici (ossia soggetti a attacco ischemico transitorio o ictus), sottoposti a tomografia computerizzata prima della procedura di endoarterectomia carotidea. Di conseguenza, l'obiettivo del presente studio è identificare caratteristiche delle placche basate su immagini che permettono di stratificare efficacemente pazienti sintomatici e asintomatici, caratterizzando così la vulnerabilità della placca. Viene applicata la radiomica per estrarre features quantitative dalle immagini CT. Inoltre, vengono estratte deep learning-derived features da diverse reti neurali convoluzionali pre-addestrate, tra cui ResNet50, InceptionV3, Inception-ResNetV2 e VGG19. Le features radiomiche, deep e combinate vengono analizzate all'interno di un framework di machine learning. A fronte dello sbilanciamento dei dati nel set di addestramento, è stato applicato l'oversampling tramite SMOTE, insieme alla cross-validation. Vengono utilizzati due tipi di slice, una intera e una ritagliata intorno alla placca, così come quattro tecniche di feature selection e sei classificatori di machine learning per identificare le combinazioni migliori. Infine vengono testati modelli ensemble che incorporano tutti i migliori modelli per radiomica e reti neurali. L'analisi viene condotta inizialmente su una singola slice per carotide, rappresentante l'area di placca più estesa (approccio 2D). Successivamente, vengono incluse nell'analisi multiple slice con un'area di placca $>$ 30% di quella maggiore (approccio 2.5D). Le migliori performance di classificazione sono state ottenute con l'approccio 2.5D su VGG19 combinato alla radiomica, con ROC AUC di 0.889 e balanced accuracy di 0.898. Stando alle ricerche analizzate, questo è il primo studio che utilizza reti di deep learning su scansioni CTA e prova combinazioni di features radiomiche e deep. Inoltre, gli ottimi risultati di stratificazione ottenuti dimostrano il potenziale dell'approccio proposto. In futuro, verrà perseguita un'ulteriore validazione su un dataset più ampio, multicentrico e prospettico per confermare i risultati ottenuti.

Parole chiave: Placche Carotidee, Aterosclerosi, Radiomica, Deep Learning, Machine Learning, Intelligenza Artificiale, Immagini CTA, Combinazione di features, Classificazione di sintomatici e asintomatici

Contents

List of Figures

List of Tables

List of Abbreviations

ACA Anterior Cerebral Artery

AIC AKaike Information Criterion

ANOVA Analysis of Variance

AUC Area Under Curve

BMI Body Mass Index

CAD Computer-Aided Diagnosis

CART Classification and Regression Trees

CAS Carotid Artery Stenting

CCA Common Carotid Artery

CEA Carotid Endarterectomy

CNN Convolutional Neural Network

CT Computed Tomography

CTA Computed Tomography Angiography

CV Cross Validation

DL Deep Learning

DSA Digital Subtraction Angiography

DT Decision Tree

DUS Doppler Ultrasound

ECA External Carotid Artery

ESC European Society of Cardiology

ESVS European Society for Vascular Surgery

FN False Negative

FP False Positive

FPR False Positive Rate

GLCM Gray Level Co-occurrence Matrix

GLDM Gray Level Dependence Matrix

GLRLM Gray Level Run Length Matrix

GLSZM Gray Level Size Zone Matrix

GMP Global Max Pooling

HH High-High

HL High-Low

HRMRI High-Resolution Magnetic Resonance

ICA Internal Carotid Artery

ICC Intraclass Correlation

INC Inception

INCRES InceptionResnet

IPH Intraplaque Hemorrhage

KNN K Nearest Neighbour

LASSO Least Absolute Shrinkage and Selection Operator

LBP Local Binary Pattern

LCCA Left Common Carotid Artery

LH Low-High

LL Low-Low

LR Logistic Regression

MCA Middle Cerebral Artery

ML Machine Learning

MLP MultiLayer Perceptron

MRA Magnetic Resonance Angiography

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

mRMR Minimum Redundancy Maximum Relevance

MV Majority Voting

NGTDM Neighboring Gray Tone Difference Matrix

NN Neural Network

PET Positron Emission Tomography

PVAT Perivascular Adipose Tissue

RBF Radial Basis Function

RCCA Right Common Carotid Artery

RES Resnet

RF Random Forest

ROI Region of Interest

ROC Receiver Operating Characteristic

SMOTE Synthetic Minority Over-Sampling Technique

SVM Support Vector Machine

TIA Transient Ischemic Attack

TN True Negative

TP True Positive

TPR True Positive Rate

US Ultrasound

VOI Volume of Interest

VGG Visual Geometry Group

XGBoost eXtreme Gradient Boosting

1 | Introduction

1.1. Clinical Problem

1.1.1. Anatomy of Carotid Arteries

The carotid arteries are crucial vessels responsible for supplying oxygen-rich blood to the head and brain. Blood flow to these arteries begins as oxygenated blood exits the heart through the ascending aorta. The first major branch is the brachiocephalic artery, which gives rise to the right common carotid artery (RCCA) and the right subclavian artery. The left common carotid artery (LCCA) branches directly from the aortic arch as the second major branch and then together with RCCA ascends through the neck. The carotid bifurcation, where the common carotid artery (CCA) divides into the internal carotid artery (ICA) and external carotid artery (ECA), particularly at the carotid bulb, is a common site for the development of atherosclerotic plaques. *Figure 1.1* provides a detailed visualization of it.

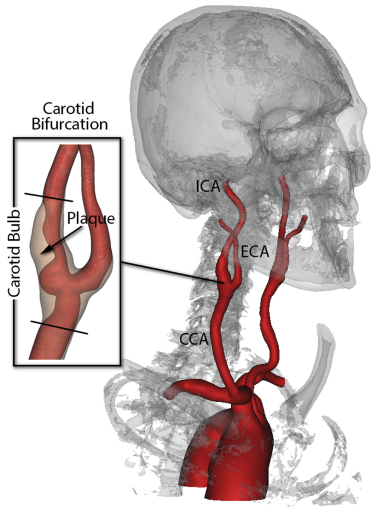


Figure 1.1: Carotid bifurcation illustration

The ECAs supply blood to the face, scalp, and neck, while the ICAs enter the skull and play a crucial role in cerebral circulation. Inside the skull, the ICAs contribute to the circle

of Willis, ensuring stable blood flow to the brain even if a major artery is compromised. The first branch of the ICA, the ophthalmic artery, supplies the eyes. The ICA then divides into the middle cerebral artery (MCA) and anterior cerebral artery (ACA). The MCA supplies areas involved in upper limb, facial motor, and sensory functions, along with language centers, while the ACA supplies regions controlling lower limb functions.

At a histological level, carotid arteries, like all arteries in the human body, are composed of three distinct layers: the tunica intima, tunica media, and tunica adventitia. The tunica intima, the innermost layer, is a thin layer of endothelial cells supported by elastic and collagen fibers. This layer plays a vital role in maintaining vascular health by producing substances such as nitric oxide, which acts as a vasodilator to regulate blood flow. It is within the tunica intima that plaques containing low-density lipoprotein cholesterol can form, leading to atherosclerosis. The tunica media, the middle layer, is composed primarily of smooth muscle cells, responsible for controlling the diameter of the artery and regulating blood flow. This layer also contains collagen and elastin fibers, which provide strength and elasticity. The tunica adventitia, the outermost layer, consists of connective tissue, primarily collagen fibers, and serves to anchor the artery to surrounding structures. It also contains small blood vessels known as vasa vasorum, which nourish the outer regions of the arterial wall [6].

1.1.2. Epidemiology

Stroke

Stroke is one of the leading causes of death and disability in all the world. It is estimated that each year, approximately 12 million people around the world experience their first stroke, with 6.5 million dying as its consequence. For those who survive, its effects can be severe, impacting physical mobility, speech, eating, emotional regulation, and cognitive functions, all these leading to significant care needs. In addition to this, stroke represents a substantial economic burden on a global scale, placing substantial demand on healthcare and social welfare systems [32]. Numerous studies have been conducted to analyze these costs, for instance Katan and Luft (2018) estimated stroke-related costs to account for approximately 3 to 4% of total healthcare expenditures in Western countries [27]. The World stroke organization stated that the economic burden of stroke represents around 0.66% of the world's GDP (Gross Domestic Product), with the total costs projected to surpass US$1 trillion by 2030 [32].

Atherosclerosis

Atherosclerosis plays a critical role in the development of cerebrovascular diseases, more specifically atherosclerosis in carotid plaques is estimated to be the cause of the 10–15% of stroke and transient ischemic attack (TIA) cases [16]. Atherosclerosis is a chronic and progressive inflammatory condition characterized by the accumulation of lipids and fibrous tissue within the arterial wall [30]. Carotid atheromatous plaque is generally associated with the restriction or disruption of blood flow [41]. Its rupture typically occurs at the carotid bifurcation or within the ICA, resulting in the formation of a thrombus — a blood clot that can partially or fully obstruct blood flow (as shown in *Figure 1.2*). Furthermore, fragments of the thrombus or atherosclerotic plaque may detach and form emboli, which can travel through the bloodstream, end up lodging in smaller arteries in the brain and potentially cause a blockage that leads to an ischemic stroke.

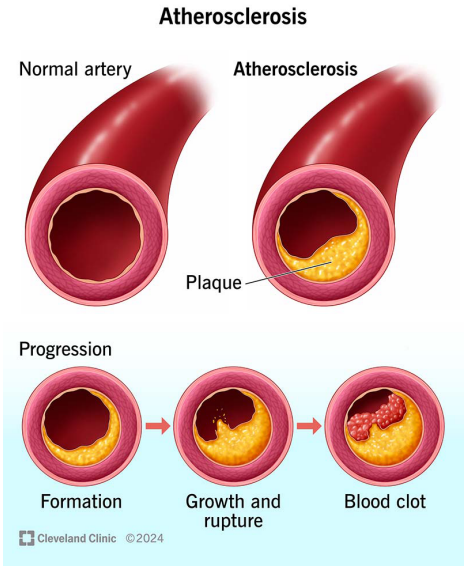


Figure 1.2: Illustration of atherosclerosis progression

Vessel stenosis (i.e., the narrowing of blood vessel) has always been the primary parameter for the classification and stratification of the disease and is adopted by the main guidelines, for example European Society of Cardiology (ESC) and European Society for Vascular Surgery (ESVS), to decide for surgical intervention and clinical treatment [16]. As an example, according to the ESC guidelines, surgery is suggested in patients with symptomatic carotid stenosis of 70-99% [1]. Histopathological studies, however, have demonstrated that certain morphological features of carotid plaque like the fibrous cap and a lipid-rich necrotic core are linked to higher rates of plaque rupture, even in cases of moderate stenosis [12]. These features are potential markers of vulnerability to plaque

rupture and are key targets in the development of new imaging techniques able to capture them as precisely as possible [42].

1.1.3. Vulnerable Plaques

Over 20 years ago, the concept of "vulnerable plaque" was introduced to describe an atheroma that is particularly prone to rupture, thrombosis, and subsequently cause a cardiac ischemic event [30]. Numerous studies ([28], [22], [16], [33]) have explored the link between specific morphological characteristics of atherosclerotic plaques and their susceptibility to rupture. These studies have identified the features listed below as the most significant factors contributing to plaque vulnerability.

A primary characteristic of vulnerable plaques is the presence of a thin fibrous cap [22]. This cap serves as a protective barrier, separating the contents of the plaque from the bloodstream. When the fibrous cap becomes thin or weakened, the plaque is more likely to rupture, exposing thrombogenic material to the circulating blood and increasing the risk of clot formation. Carr et al. in 1996 found that cap thinning was identified in 95% of plaques associated with symptoms, which are TIA or ischemic strokes, and in 48% of those without [28]. Another hallmark of plaque vulnerability is a large lipid core [22], accounting for at least 25% of the plaque area [42], which reflects the accumulation of cholesterol deposits within the arterial wall. Intraplaque hemorrhage (IPH), or bleeding within the plaque, also plays a vital role in destabilizing plaques [22]. IPH can increase plaque volume, elevate internal pressure, and further weaken the fibrous cap, making the plaque more susceptible to rupture. Active inflammation within the plaque is another significant factor that accelerates vulnerability [22]. Inflammatory cells, including macrophages and T-lymphocytes, release enzymes that degrade the extracellular matrix and compromise the structural integrity of the fibrous cap. Neovascularization, or the formation of new, fragile blood vessels within the plaque, is also linked to increased vulnerability [22]. These small vessels are prone to rupture, leading to intraplaque hemorrhage and further destabilization of the plaque. Plaques with an irregular or ulcerated surface are additionally considered high-risk for causing embolic events [22]. Such surface irregularities can encourage clot formation, and these clots may detach and travel to the brain. Lastly, the composition (e.g., high lipid content and calcification) and volume of the plaque contribute to its instability [22]. Larger plaques, often containing more necrotic material, experience greater mechanical stress and are more susceptible to rupture. As plaques grow, they may induce arterial remodeling, where the artery expands to maintain blood flow. However, if this compensatory dilation is insufficient, the artery can narrow, raising the likelihood of ischemic events [21]. On the other side, plaques that are highly

calcified tend to be more stable and less likely to rupture, although extensive calcification can sometimes interfere with the accurate assessment of other plaque components during imaging.

1.2. Imaging Techniques

Advances in imaging have enabled a more detailed characterization of the plaque, allowing to manage patients not only on the base of stenosis degree but also on morphological characteristics. Both invasive and non-invasive imaging techniques play a crucial role. Invasive techniques, such as digital subtraction angiography (DSA), have long been considered the gold standard for evaluating carotid stenosis. DSA works by inserting a catheter into the femoral artery, which is advanced to the carotid artery under imaging guidance. Once in place, a contrast medium is injected, and a series of X-rays are taken to visualize the arterial lumen in great detail. The contrast enhances the visibility of the blood vessels, allowing precise identification of stenosis or occlusion. DSA, while accurate, is invasive, costly, and carries risks like hematomas and, in rare cases, neurological complications. All this led to shift the focus towards non-invasive alternatives [41].

Ultrasound (US) is a widely used non-invasive technique for evaluating carotid artery disease, particularly Doppler ultrasound (DUS), which measures blood flow speed and direction to assess stenosis (with stenosed areas typically showing increased velocities). DUS functions by emitting high-frequency sound waves from a handheld transducer placed on the patient's neck. In addition to Doppler, B-mode US is used to generate two-dimensional images of the carotid artery and surrounding structures. It allows for the visualization of plaque morphology and the intima-media thickness. B-mode imaging distinguishes between stable, calcified plaques, which appear more echogenic (brighter), and vulnerable plaques, which are more echolucent (darker) due to their lipid content or intraplaque hemorrhage. US is a low-cost, low-risk tool, and is well-tolerated by patients, making it the initial imaging modality of choice in evaluating carotid artery stenosis but has some limitations, especially in assessing plaques that are heavily calcified or in distinguishing between complex plaque compositions [21, 28, 30, 41].

Computed Tomography Angiography (CTA) has become one of the leading and most accurate methods for assessing carotid artery disease. The procedure begins with the intravenous injection of an iodinated contrast agent, followed by rapid acquisition of images as the multidetector CT scanner rotates around the patient. These images are then reconstructed into three-dimensional models of the carotid arteries, providing high-resolution views of the vascular lumen, plaque morphology, and surrounding structures.

Given its excellent spatial resolution, CTA is particularly useful for the quantification of calcification and the detection of fibrous tissue and surface irregularities [30, 41]. One of the main challenges for CTA is differentiating between soft plaque components, such as lipid cores and IPH, especially in the presence of extensive calcification. Beam-hardening artifacts caused by calcified plaques can also obscure the view of the arterial lumen, making it more difficult to assess the degree of stenosis accurately [21].

Magnetic Resonance Angiography (MRA) and High-Resolution Magnetic Resonance Imaging (HRMRI) are highly valuable for providing detailed information on both the arterial lumen and the composition of atherosclerotic plaques. MRA uses strong magnetic fields and radio frequency pulses to generate images, and it can be performed with or without the use of contrast agents . HRMRI is particularly effective in characterizing the components of atherosclerotic plaques, offering detailed views of soft tissue structures, like IPH. Additionally, it can assess the thickness of the fibrous cap and detect the presence of lipid-rich necrotic cores, both of which are indicators of a plaque's risk of rupture. MRI is also effective in evaluating remodeling of the artery and can detect inflammation through the use of specialized imaging protocols, making it a comprehensive tool for plaque assessment. Its limitations compared to CTA include lower effectiveness in detecting calcifications. MRI is also less widely available and can be more expensive, which may limit its use in some clinical settings [21, 30, 41].

Finally, Positron Emission Tomography (PET), particularly when combined with CT, is an emerging technique for assessing plaque vulnerability. PET uses radiotracers, such as 18F-fluorodeoxyglucose, which accumulate in metabolically active tissues, including inflamed atherosclerotic plaques and help in highlighting areas of active inflammation. Although PET is excellent for detecting metabolic activity within plaques, it is limited by lower spatial resolution compared to CTA or MRI and cannot differentiate between specific plaque components like lipid cores or fibrous caps. Moreover, the use of radiation and the high cost of PET imaging restrict its frequent use in routine clinical practice [21, 30].

To conclude, it's hard to state what is the best imaging method: DSA provides the most accurate assessment of stenosis but is invasive, while non-invasive options offer safer alternatives. US is ideal for initial screening, CTA excels in detecting calcifications and structural details, MRI provides unparalleled detail in assessing plaque vulnerability through soft tissue characterization, and PET highlights metabolic processes within plaques, identifying areas of active inflammation.

The ESC guidelines recommend relying on DUS just as an initial screening tool due to

the ease of use and ability to provide real-time information. However, for more detailed imaging, MRA and CTA are preferred as they offer higher spatial resolution and greater accuracy in assessing plaque morphology and vascular structures. These advanced imaging techniques provide clearer visualization of the arterial anatomy and are essential for precise evaluation, especially when planning surgical interventions [1].

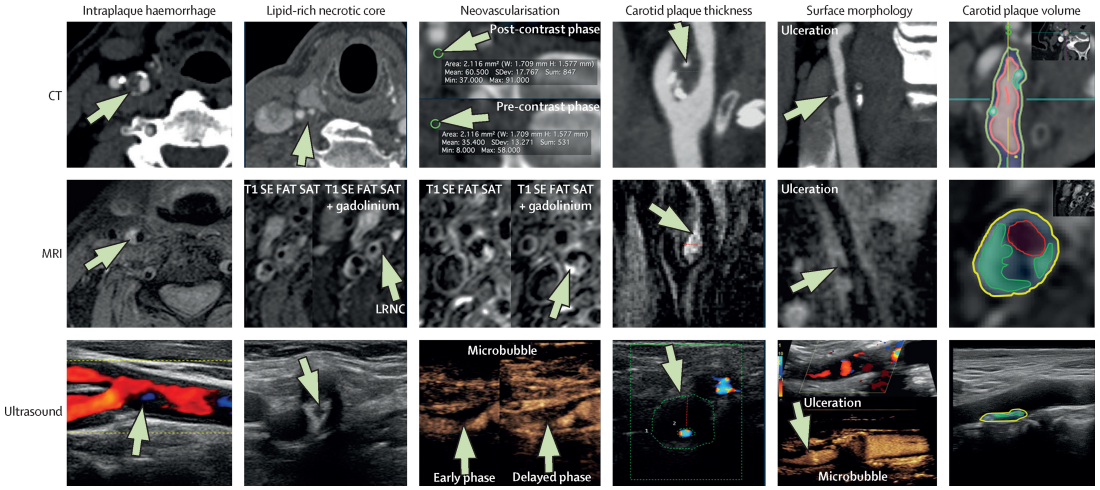


Figure 1.3: Multimodal imaging of vulnerable carotid plaque features in CT, MRI, and US

1.3. Risk prevention

Carotid atherosclerotic disease has drawn significant clinical attention due to its role as a major, potentially preventable cause of stroke [42]. According to ESC definitions, carotid stenosis is defined as 'symptomatic' if associated with symptoms as ipsilateral retinal or cerebral ischemia in the preceding 6 months and 'asymptomatic' if no prior symptoms can be identified or when symptoms occurred >6 months ago [1]. Carotid atheromatous plaque usually remains asymptomatic for many years, making early detection challenging. It is estimated that between 5% and 10% of individuals over the age of 65 have asymptomatic carotid artery atherosclerosis with at least 50% artery stenosis [22].

Currently, symptomatic extracranial internal carotid artery occlusive disease is primarily treated through carotid endarterectomy (CEA), a surgical procedure that carries significant risks for the patient. The primary criterion for this surgery is the severity of luminal stenosis in the ICA, along with the presence of relevant symptoms. However, stenosis severity alone is not an accurate predictor of stroke risk in asymptomatic patients. Thus, identifying alternative approaches to distinguish symptomatic from asymptomatic patients becomes essential. Furthermore, CEA indications for asymptomatic patients

remain controversial and choosing the most suitable treatment continues to pose a substantial clinical challenge.

In clinical practice, it's always been doctors that manually identify carotid plaques based on neck images, which is time-consuming and labor-intensive. Research from coronary artery studies and, more recently, findings in carotid atherosclerosis indicate that evaluating atherosclerosis pathology via non-invasive plaque imaging could identify patients who might benefit from early intervention. Such imaging can be performed to characterize carotid plaques, support risk stratification also for asymptomatic patients, and provide appropriate preventive therapies. Evidence suggests that making treatment decisions based on plaque characteristics can be cost-effective [42].

In addition to imaging analysis, various other risk predictors are traditionally used to identify individuals at high risk, often assessing key factors such as age, gender, cholesterol levels, blood pressure, body mass index, family medical history, smoking status, and diabetes [41]. In recent years, also artificial intelligence-based carotid plaque diagnosis has demonstrated significant potential for these applications, and it is this approach the one explored in this study.

1.4. Plaque classification state of art

In recent years, several studies have explored the use of machine learning (ML) algorithms for the classification of carotid plaques, focusing on distinguishing between symptomatic and asymptomatic plaques, assessing the degree of stenosis or predicting the likelihood of future cerebrovascular events based on features such as clinical risk factors (e.g., age, cholesterol levels, BMI), morphological features derived from imaging modalities (e.g., US, CTA and MRI), plaque composition and tissue characteristics (e.g., fibrous cap thickness, IPH, lipid core size) and plaque surface characteristics (e.g., surface irregularity). In this review, the focus is exclusively on features of the same type used in the study, specifically radiomic and deep learning-derived ones.

1.4.1. Radiomics approach

Extensive research was conducted on carotid plaques using the radiomics approach, however, this section is specifically focused on studies focused on classifying symptomatic and asymptomatic, with an emphasis on those that extract features from CT images. Table 1.1 provides a summary of the analyzed studies, focusing on the characteristics of the dataset, the aim of the study, the type of the features involved and their selection, the

ML trained model and the performance achieved with the best model found. Scicolone et al. in 2024 [35] summarized the radiomics studies based on different kind of images, five of the eight studies here listed are present in that review under the CT image section, three more were added to this list.

Acharya et al. in 2013, were the firsts to perform this classification by combining Local Binary Pattern (LBP) texture features with 2D radiomic wavelet features. The dataset only consists of 20 patients, 11 of them symptomatic. After the use of t-test for feature selection, the SVM algorithm with RBF kernel is the model that resulted in the best performance (0.88 of accuracy) [2].

Xia et al. in 2023 developed ML models for predicting the risk of TIA in patients with mild carotid artery stenosis. The aim of this study is to compare the performances of three different models: clinical-only model (features coming from routine blood test data, demographic data, stenosis of the carotid artery, and medication data), a radiomic-only model, and a combined model of the two. The dataset used for the study consists of 179 patients where only 34 are symptomatic and with a stenosis percentage of 30 to 50%. After performing feature selection, five classification algorithms were used to train models. The best one was found to be Random Forest (RF), constructed using radiomics and clinical feature information, which exhibited the highest accuracy on both training (0.988) and testing sets (0.863), with corresponding AUC values of 0.983 and 0.879 respectively, by far outperforming the clinical model, while the radiomic model resulted in train accuracy of 0.973, AUC of 0.982 and test accuracy of 0.787, AUC of 0.746 [45].

In the same year, Shi et al. performed the same kind of analysis on a cohort of 167 patients with 70 symptomatic cases, but combining wavelet radiomic features with clinical and imaging CTA features. Various steps of feature selection were performed to obtain 2 conventional features and 6 radiomic ones, then a Multiple Logistic Regression (LR) model was chosen. This led to lower results for the combined compared to the study mentioned before (test AUC= 0.832 and test acc = 0.761 [36]) but still showed that combining radiomics to conventional features improves the performances of the model.

Le et al. considered 41 patients and 82 carotids, of which 41 culprit and 41 non-culprit with a stenosis in the range 29-99% for the culprit patients and in the range 3-88% for the non culprit ones. They extracted from CTA images 93 original radiomic features and, after applying intraclass Spearman correlation to increase robustness and univariate LR for relevancy, only 10 were given to several ML algorithms (Decision Tree (DT), RF, LASSO, ElasticNet, Neural Network (NN) and XGBoost). Both 2D and 3D approach were tested and the best performance was achieved by ElasticNet which only with 3D

radiomic features achieved a mean AUC on the train set of 0.670 and when paired with calcium score increased to 0.730 [20].

Dong et al. realized a study on 120 patients, with only 34 symptomatic case, and a degree of stenosis greater or equal to 50%. They were able to obtain a train AUC of 0.858, compared to 0.706 using conventional plaques features, highlighting again the improvements of adding radiomics features to ML models [14].

Cilla et al. classified plaques, on a dataset of 30 patients with more than 70% stenosis, into vulnerable and non-vulnerable to rupture, with the aim of identifying surgical candidates in a timely manner. They tested LR and SVM as binary classifiers and reached an accuracy of 0.880 and an F1 score of 0.867 with the latter. After performing feature selection through Spearman rank correlation coefficient, univariate analysis and backward elimination using AIC, they found that volume and entropy features are the most significantly associated with the two plaque groups [11].

Zhang et al. published in 2022 a research where they developed a model called radiomics nomogram obtained by the combination of a clinical factor model and a radiomics signature model. The dataset is composed of 110 patients for training and 38 patients for testing with a general stenosis between 15-70%. The training dataset is composed of 46 patients with IPH and 64 without it while the testing dataset contains 18 patients with IPH and 20 without it. The radiomics signature model was obtained after performing features selection with different techniques such as: ICCs, ANOVA test and LASSO. The nomogram model proved to be more efficient in detecting IPH in symptomatic patients, with an AUC score on the test set of 0.811 [48].

Finally, Chen et al. created a model that combines radiomic features extracted from the plaques with perivascular adipose tissue's (PVAT) radiomic features and traditional CTA features and is able to distinguish between symptomatic and asymptomatic patients. The dataset consisted of 60 asymptomatic patients and 84 asymptomatic one with a degree of stenosis greater than 30%. They were able to reach an AUC of 0.840 [7].

Authors & Country	Number of Patients	Purpose	Feature Type	Feature Selection	ML model	Best model performance
Acharya et al., 2013 [2]	20 patients (11 sym and 9 asym), 400 carotid artery images. Stenosis >50%	Propose a non-invasive CAD technique to classify the plaque into the two classes	Wavelet transform and LBP	Student t-test p_value (0.01), 9 kept	SVM with different kernels	SVM RBF kernel (best) test acc= 0.88
Xia et al., 2023 [45]	179 patients (34 symp, 145 asymp) with 30-50% stenosis.	Predict the risk of TIA in patients with mild carotid stenosis through a combined model	129 3D radiomic features, clinical features	RF feature importance, LR (9 radiomics kept, 2 clinical). 3 kept for the combined model	Testing of 5 ML alg: RF, XGBoost, LR, SVM, KNN	Rad model: RF (train AUC= 0.982 test AUC= 0.746), Comb model: RF (train AUC = 0.983, test AUC = 0.879)
Shi et al., 2023 [36]	167 patients (70 sym, 97 asym). 91 patients with 0-49.9% of stenosis and 76 patients with 50-99.9% of stenosis	Develop a combined model to assess plaque vulnerability	Wavelet transform features, imaging and clinical features	ICC, linear correlation value, mutual information alg (6 radiomics + 2 conventional kept)	Multiple LR	Rad model: train AUC= 0.84 test AUC= 0.767, Comb model: train AUC = 0.856, test AUC = 0.832
Le et al., 2021 [20]	41 patients, 82 carotids, 41 sym and 41 asym. Stenosis of 29-99% for culprit and 3-88% for non culprit.	Test feature robustness and to identify culprit and non-culprit arteries using calcium score and radiomics features	93 radiomic features	Robustness through Spearman correlation and relevancy for univaritate LR (10 kept)	Testing of 6 ML alg: DT, RF, LASSO, ElasticNet, NN, XgBooost	ElasticNet radiomic-only: train mean AUC= 0.67 radiomic + calcium: train mean AUC = 0.73, acc= 0.69
Dong et al., 2022 China [14]	120 patients with >= 50% stenosis, 148 plaques (34 sym)	Identify symptomatic patients with CAS through radiomics analysis	8 plaque features + 2,107 radiomics parameters	20 kept	Tested: SVM, XGBoost, LR	train AUC = 0.858 (conventional model AUC = 0.706)
Cilla et al., 2022 [11]	30 patients with > 70% internal carotid stenosis	Discriminate vulnerable from non-vulnerable plaques	203 radiomic features	Spearman rank correlation coefficient, univariate analysis, backward elimination using AIC (2 kept)	Tested LR, SVM with CART tree analisis	SVM with rbf kernel acc = 0.880, F-score = 0.867

Continued on next page

Authors	Number of Patients	Purpose	Feature Type	Feature Selection	ML model	Best model performance
Zhang et al., 2022 [48]	46 patients with IPH and 64 without IPH for training. 18 patients with IPH and 20 without IPH for testing. Stenosis >15% and <70%	Develop and validate CT-based radiomics nomogram incorporating clinical factors and a radiomics signature for the detection of IPH	1409 radiomics features extracted from CTA images	ICCs > 0.75, ANOVA and LASSO ending with 8 features	Multiple LR	Radiomics signature: 0.725 AUC Radiomics nomogram: 0.811 AUC
Chen et al., 2023 [7]	144 patients (60 sym, 84 asym) with >30% stenosis and PVAT	Identify high-risk carotid plaques and compare the diagnostic value between radiomics models and traditional CTA model	10 radiological features and radiomic features extracted with wavelet and Laplacian of Gaussian	Mann-Whitney U-test and LASSO	SVM	Plaque + PVAT + Traditional CTA model obtained test AUC = 0.840

Table 1.1: Summary of Radiomics studies on carotid plaques with CTA images

1.4.2. Deep Learning approach

All the methods discussed above primarily rely on manually designed features to train classification models. However, these hand-crafted features are limited to describing low-level image characteristics, which may not fully capture the complexity of plaque characterization and might struggle to differentiate between various types of carotid plaques in gray-scale images. Deep learning (DL) methods, by automatically extracting features, have the potential to uncover new, high-level features of plaques that may not be detectable through traditional approaches.

All the applications of deep features identified in the literature involve the use of deep neural networks with an end-to-end approach, directly classifying carotid plaques based on learned features. However, no studies to date have applied these methods specifically to CT images; therefore, the review includes studies that utilize other imaging modalities (mostly US), while still narrowing the focus to studies aimed at the desired classification of symptomatic plaques.

In 2021, Ganitidis et al. applied DL methods to B-mode US images using a Convolutional Neural Networks (CNN) for feature extraction, followed by classification into symptomatic

and asymptomatic plaques using fully connected layers, achieving an AUC of 0.730. The dataset used was composed of 16 symptomatic patients and 58 asymptomatic. Interesting about their work was they way they addressed the highly imbalanced distribution of patients. They applied an ensemble learning scheme based on a sub-sampling approach along with a two-phase, cost-sensitive strategy of learning, that uses once the original and then the resampled data set [40].

In the same year, Saba et al. built a 13-layers CNN taking as input manually delineated plaques US images and reached an accuracy of 0.897 and an AUC of 0.910, demonstrating the model's effectiveness compared to the previous ML methods. The study was conducted on 346 plaques of which 150 are asymptomatic and 196 are symptomatic [34].

In 2022, Wei et al. conducted a study on 333 patients where the 117 symptomatic patients are identified by atherosclerotic events such as TIA or ischemic stroke. They developed an object-specific four-path network (OSFP-Net), which integrates images of carotid plaques from both transverse and longitudinal sections of the bilateral carotid arteries. Each path of the OSFP-Net consists of two components: a feature extraction subnetwork (FE) and a feature downsampling subnetwork (FD). To accommodate images of arbitrary size and extract richer information, FE and FD use two pooling strategies: spatial pyramid pooling and multilevel strip pooling. The study compares the performance of OSFP-Net with several state-of-the-art DL models, including ResNet50, DenseNet121, and EfficientNet-b7, showing the superiority of OSFP-Net (AUC= 0.990 , accuracy=0.974, sensitivity=0.962, specificity=0.976) [43].

Another interesting study was conducted in 2023 by Gui et al. on 104 patients with 74 symptomatic cases and 30 asymptomatic, with the aim to compare the performances of several radiomics-based ML methods and an end-to-end DL approach for plaques classification, using HRMRI images. They tested two DL networks, 3D-DenseNet and 3D-SE-DenseNet, with the addition of the SE module able to automatically determine the importance of each feature channel through learning. The result showed that the 3D-SE-DenseNet-121 model outperformed the best radiomics-based ML model found which was MLP (AUC of 0.930 vs 0.880) [10].

Zhang et al. proposed a self-supervised learning approach. They built a fusion network (DBResNet) which consists of two parallel branches for feature extraction, one with the aim of classifying image block order and the other of predicting image rotation angles, followed by one fusion layer that fuses the feature presentations learned by the two tasks and then the classification layer. The study, based on 844 patients and with only 40% of labeled data, achieved comparable performances to ResNet101 method on 100% la-

beled training images, showing the potential of this network to overcome the problem of unlabeled images [49].

The research conducted by Saba et al. used a mixed dataset obtained from two different datasets for a total of 246 symptomatic and 260 asymptomatic scans. They showed how a transfer learning approach (VGG19) can outperform a relatively simple network trained from scratch. Moreover it proved that the features extracted from a deep network are more useful than features extracted through more classical approach, such as Histogram of Oriented Gradients, and then classified with traditional ML classifiers. The study achieved a final results of 0.946 AUC in the case of transfer learning network [15].

Finally the study of He et al. of 2024 aimed at identifying and classify carotid plaque. The dataset is composed of 665 patients without plaques and 510 patients with plaques of which 156 are stable plaques and 354 are unstable plaques. They obtained an AUC of 0.854 by building a custom deep network which combines the features extracted by two different ResNet50 networks, that are then classified by a fully connected layer [19].

Authors	Image Type	Number of Patients	Purpose	Modeling Method	Best Model Performance
Ganitidis et al., 2021 [40]	US	74 patients (58 asym, 16 sym)	Leverage feature extraction using CNN for identifying symptomatic and asymptomatic plaques	Six CNN layers with two fully connected layers	AUC = 0.73
Saba et al., 2021 [34]	US	346 plaques (196 sym, 150 asym) with > 50% stenosis	Develop an automated carotid plaque classification system on a supercomputer	Atheromatic 2.0: A 13-layer CNN, with 5 convolution layers, pooling layers, dense layers, and a softmax output layer	AUC = 0.910, acc= 0.897
Wei et al., 2022 [43]	US	333 patients (1332 images) where 117 had atherosclerotic events and 216 without.	Propose a DL approach for the identification of symptomatic and asymptomatic plaques	Object-Specific Four-Path Network (OSFP-Net) which integrates US carotid plaques in both transverse and longitudinal sections	AUC = 0.990, f1 = 0.959
Gui et al., 2023 [10]	HRMRI	104 patients with stenosis between 30% and 70% (74 sym, 30 asym)	Compare the performance in classifying symptomatic plaques using radiomics-based ML approach and a DL approach	3D DenseNet and 3D SE-DenseNet against several ML alg (KNN, LR, SVM, DT, RF, XGBoost, AdaBoost, LightGBM, CatBoost, MLP)	AUC = 0.93, acc= 0.931, f1 = 0.861
Zhang et al., 2024 [49]	US	844 patients, 1270 plaque images	Develop a dual-branch residual network (DBResNet) to improve classification of carotid plaques	The DBResNet consists of a feature extraction layer, feature fusion layer, and fully connected layer	acc = 0.812, f1 = 0.801
Saba et al., 2021 [15]	US	260 asymptomatic scans and 246 symptomatic scans obtained as a combination of two datasets (London and Lisbon)	Classify the carotid plaque (vulnerable or stable) taken from multicenter studies using automated algorithms	CNN consisting on 7 layers applied on augmented data, compared to VGG19 transfer learning model and 4 different ML models (KNN, SVM, DT, RF)	DL mean AUC= 0.938, TL mean AUC= 0.946, ML mean AUC=0.889
He et al., 2024 [19]	Doppler US	665 patients without plaques and 510 patients with plaques of which 156 stable and 354 unstable.	Develop an automated algorithm to identify the presence and stability of carotid plaques using DL	BCNN-ResNet: uses two ResNet50 networks and one single input. The two networks perform feature extraction separately that are then combined. Finally fully connected layer for classification	internal testing dataset: AUC= 0.896 external testing dataset: AUC=0.854

Table 1.2: Summary of DL models for carotid plaque classification

1.5. Thesis Objective and Structure

The objective of the dissertation is to test a non-explored approach in the existing body of research for the classification of symptomatic and asymptomatic carotid plaques. This approach consists in extracting features from various pre-trained deep neural networks from two types of CT images: one cropped around the plaque and one including the full scan. These features are then given to different feature selection algorithms and fed into various ML classification algorithms to produce the best binary prediction. This deep feature-based approach is then compared with a traditional radiomics approach and combinations of the two are experimented. The study was first conducted in 2D, analyzing only the slice with the largest plaque area. Subsequently, the analysis is expanded to a method (termed '2.5D'), where several slices per patient are included to assess potential improvements. At the conclusion of the study, it will be possible to determine which neural network architecture performs best and which model could, with further validation, be considered for implementation in clinical settings to improve diagnostic capabilities for carotid plaque classification.

This chapter provided a clinical introduction of atherosclerosis, carotid plaques and their risks associated, followed by a review of the current state of the art and the objectives of this study. Chapter Two supports the study with a theoretical overview of all the techniques used. Chapter Three details the materials involved, the workflow of the analysis and the methods. Chapter Four presents the obtained results, leading into the discussions and potential future improvements in Chapter Five. Chapter Six ends the study presenting the final conclusions.

2 | Theoretical Context

This section provides a theoretical context for all the concepts and methods included in the study. It begins with a broad overview of the two types of features utilized: radiomic and DL. Following this, the discussion transitions to the algorithms employed for feature selection and classification, to then conclude with an explanation of the metrics used to evaluate the results.

2.1. Radiomics

Radiomics is a rapidly developing field of research dedicated to extracting quantitative features from medical images, thereby transforming these digital images into minable, high-dimensional data that reveal biological insights not readily visible to the human eye and offer unique biological information that can deepen the understanding of disease processes and provide clinical decision support. While the primary focus of radiomics research has been oncology, its applications are expanding to include cardiovascular disease, neurological disorders, and other fields, where it shows promise for improving diagnosis, prognosis, and treatment planning [25].
The radiomics analysis process comprises various standard steps, as depicted in *Figure 2.1.*

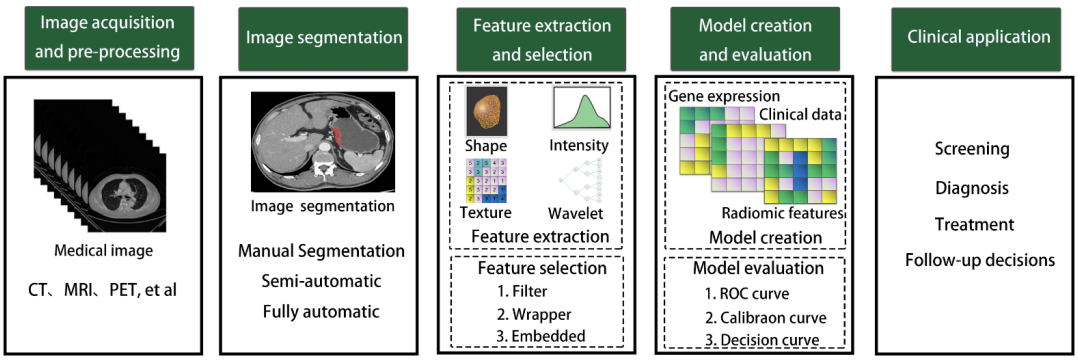


Figure 2.1: Steps of radiomics analysis

The initial stage involves image acquisition, from various modalities such as CT, MRI, PET, and US. One of the biggest challenges in radiomics is to reduce the influence of the imaging protocol and one of the ways to achieve this is post-image acquisition pre-processing.

The next step is to accurately segment Regions of Interest (ROIs) for 2D images or Volumes of Interest (VOIs) for 3D. Segmentation methods include manual, semi-automatic, and fully automatic approaches. Since manual segmentation is time-intensive and susceptible to bias, many studies reported that employing automatic or semi-automatic segmentation techniques can mitigate human errors and enhance efficiency. There are numerous software programs and segmentation algorithms available for performing automatic segmentation, an example is 3D Slicer. Additionally, DL is emerging as a powerful paradigm, achieving remarkable results across various applications in the medical field. [44].

Moving to the third stage, automatic identification, computation, and extraction of radiomic features is allowed by various open-source software tools (e.g., PyRadiomics, ITK-SNAP, 3D Slicer with Radiomics Extension). Features in detail are explained in *section 3.2.2*. Then, selecting a subset of these features relevant to the disease process or research question becomes crucial (feature selection methods will be explained in *section 2.3*).

The selected features can be utilized in various model types, including diagnostic, prognostic, and predictive models. ML has become a popular approach for making radiomics-based predictions. At its core, ML involves training models on sets of input-output data, allowing the system to identify complex relationships autonomously, with minimal human intervention. Through repeated cycles of training and validation, ML models enhance their predictive capability, providing valuable insights for clinical applications.

Finally, the clinical application stage applies the trained model to support clinical decisions, including screening, diagnosis, treatment planning, and follow-up care. This step translates radiomic data into actionable insights, helping healthcare providers make informed decisions based on quantitative imaging biomarkers.

Radiomics features fall into three categories:

1. First-order statistical: they relate to the properties of individual voxels without reference to their spatial distribution and are expressed through commonly used and basic metrics. They reflect the symmetry, homogeneity, and local intensity distribution variations of the measured voxels. The features extracted by PyRadiomics are listed in *Table 2.1* .

First Order Statistics Features	
Energy	Total Energy
Entropy	Minimum
10th percentile	90th percentile
Maximum	Mean
Median	Interquartile Range
Range	Mean Absolute Deviation
Robust Mean Absolute Deviation	Root Mean Squared
Standard Deviation	Skewness
Kurtosis	Variance
Uniformity	

Table 2.1: The PyRadiomics 19 First Order Statistics Features

2. Shape-based: they describe the shape of the traced ROI and its geometric properties including simple descriptions of size, such as dimensions but also more complex topological characterizations such as roundness, compactness, speculation, or convexity [25]. These features are independent from the gray level intensity distribution in the ROI and are only calculated on the non-derived image and mask [17] (*Table 2.2*).

Shape 2D Features	
Mesh Surface	Pixel Surface
Perimeter	Perimeter to Surface ratio
Sphericity	Spherical Disproportion
Maximum 2D diameter	Major Axis Length
Minor Axis Length	Elongation

Table 2.2: The 10 PyRadiomics features including descriptors of the two-dimensional size and shape of the ROI

3. Textural features: they are statistical interrelationships of the voxel intensities within the ROI, they can discern spatial variations intensities by evaluating the spatial dispersion among voxels. They can be categorized into several groups:

   (a) The Gray Level Co-occurrence Matrix (GLCM) represents the probability of a voxel value occurring at a specific direction and distance *(Table 2.3)*;

   (b) The Gray Level Run Length Matrix (GLRLM) describes the length of consecutive voxels with the same gray level in a specified direction *(Table 2.4)*;

(c) The Gray Level Size Zone Matrix (GLSZM) segments an image into regions with contiguous voxel values *(Table 2.5)*;

(d) The Neighboring Gray Tone Difference Matrix (NGTDM) quantifies the gray value of a voxel by considering the difference between its average gray value and the gray value within a certain distance of the neighborhood *(Table 2.6)*;

(e) The Gray Level Dependence Matrix (GLDM) calculates the difference between adjacent voxels based on their values *(Table 2.7)*.

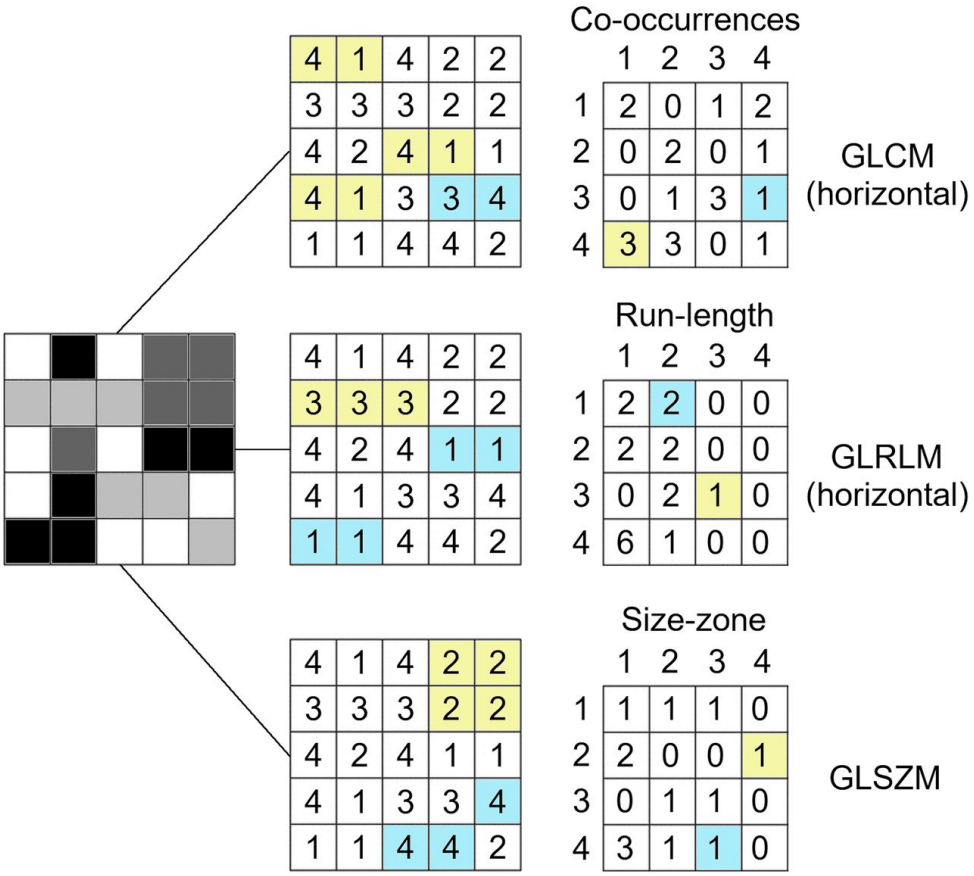


Figure 2.2: Diagram illustrating the calculation of three different texture feature matrices: GLCM represents the co-occurrence of pairs of gray levels, GLRLM measures the length of horizontal sequences of pixels with the same gray level, and GLSZM identifies the size of regions with uniform pixel intensity.

GLCM features	
Autocorrelation	Contrast
Joint Average	Correlation
Cluster Prominence	Cluster Tendency
Cluster Shade	Difference Average
Difference Entropy	Difference Variance
Joint Energy	Joint Entropy
Informational Measure of Correlation 1	Informational Measure of Correlation 2
Inverse Difference Moment	Maximal Correlation Coefficient
Inverse Difference Moment Normalized	Inverse Difference
Inverse Difference Normalized	Inverse Variance
Maximum Probability	Sum Average
Sum Entropy	Sum of Squares

Table 2.3: The 24 GLCM PyRadiomics features

GLRLM features	
Short Run Emphasis	Long Run Emphasis
Gray Level Non-Uniformity	Gray Level Non-Uniformity Normalized
Run Length Non-Uniformity	Run Length Non-Uniformity Normalized
Run Percentage	Gray Level Variance
Run Variance	Run Entropy
Low Gray Level Run Emphasis	High Gray Level Run Emphasis
Short Run Low Gray Level Emphasis	Short Run High Gray Level Emphasis
Long Run Low Gray Level Emphasis	Long Run High Gray Level Emphasis

Table 2.4: The 16 GLRLM PyRadiomics features

GLSZM features	
Small Area Emphasis	Large Area Emphasis
Gray Level Non-Uniformity	Gray Level Non-Uniformity Normalized
Size-Zone Non-Uniformity	Size-Zone Non-Uniformity Normalized
Zone Percentage	Gray Level Variance
Zone Variance	Zone Entropy
Low Gray Level Zone Emphasis	High Gray Level Zone Emphasis
Small Area Low Gray Level Emphasis	Small Area High Gray Level Emphasis
Large Area Low Gray Level Emphasis	Large Area High Gray Level Emphasis

Table 2.5: The 16 GLSZM PyRadiomics features

NGTDM features	
Coarseness	Contrast
Busnyness	Complexity
Strenght	

Table 2.6: The 5 NGTDM PyRadiomics features

GLDM features	
Small Dependence Emphasis	Large Dependence Emphasis
Gray Level Non-Uniformity	Dependence Non-Uniformity
Dependence Non-Uniformity Normalized	Gray Level Variance
Dependence Variance	Dependence Entropy
Low Gray Level Emphasis	High Gray Level Emphasis
Small Dependence Low Gray Level Emphasis	Small Dependence High Gray Level Emphasis
Large Dependence Low Gray Level Emphasis	Large Dependence High Gray Level Emphasis

Table 2.7: The 14 GLDM PyRadiomics features

In addition to extracting features from the original image, it is common practice to apply specific filters to generate additional feature-rich images. Filters such as Gaussian, Laplacian of Gaussian and wavelet transforms highlight different aspects of the image structure, extracting complementary features from different frequency domains and resolutions.

Wavelet-domain images are transformed by applying a pair of quadrature mirror filters, a high-pass and a low-pass filter, in each of the three dimensions of the image, allowing to take into account the spectral dimension of the data. Although the high-pass filter highlights the changes in gray level and thus emphasizes image details, the low-pass filter smooths the image in terms of gray-level, removing image details [24].

The image is decomposed into 4 sub-bands based on spatial frequencies, labeled as LL (low frequency component), LH, HL, and HH. The LL sub-band represents the low-frequency component, which captures the overall structure and smooth variations in the image, while the other three capture high-frequency details in the horizontal, vertical, and diagonal directions, respectively. Once this decomposition is completed, radiomic features are extracted individually from each sub-band.

This approach, leveraging multi-scale and multi-directional information, significantly increases the number of features available for analysis, as each sub-band provides a unique perspective on the texture and structure within the image.

2.2. Artificial neural networks and deep learning

Artificial neural networks are a fundamental application of artificial intelligence and DL. They are designed to emulate the learning process of the human brain and its ability to recognize patterns. The underlying idea is that, by being trained on a sufficient amount of specific cases, the network is able to generalize and recognize patterns on unseen data. This ability has become incredibly useful in different fields such as features extraction, image recognition, classification and prediction.
Generally a neural network is composed by an input layer, an output layer and a variable number of hidden layers in between. DL is a specialized subset of ML, that consists of using neural networks with a high amount of hidden layers. The layer's depth allows the network to learn more intricate patterns within high dimensional data, increasing flexibility and accuracy.

2.2.1. Neural network structure

As the name suggests, an artificial neural network is composed by a different number of artificial neurons, called perceptrons. In a perceptron, each input x_i (where $i = 1, 2, \ldots, m$) is associated with a weight w_i that determines the input's influence on the final output. Additionally the perceptron contains a bias term b which acts as threshold and helps the model to generalize better. Finally, the *activation function* φ introduces non-linearity in the output.

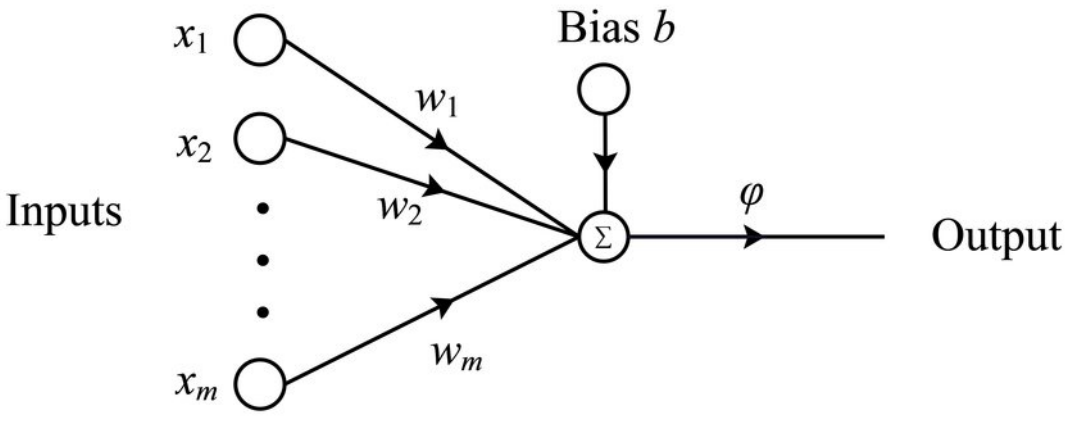


Figure 2.3: Schematic representation of a perceptron

Mathematically:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

where:

- y: The output of the perceptron.
- φ: Activation function (e.g., step, sigmoid, ReLU).
- x_i: Input features, where $i = 1, 2, \ldots, m$.
- w_i: Weights associated with each input x_i.
- b: Bias term.

The layers of a neural network are the aggregation of multiple neurons, they can be of different types depending on the task that needs to be executed.

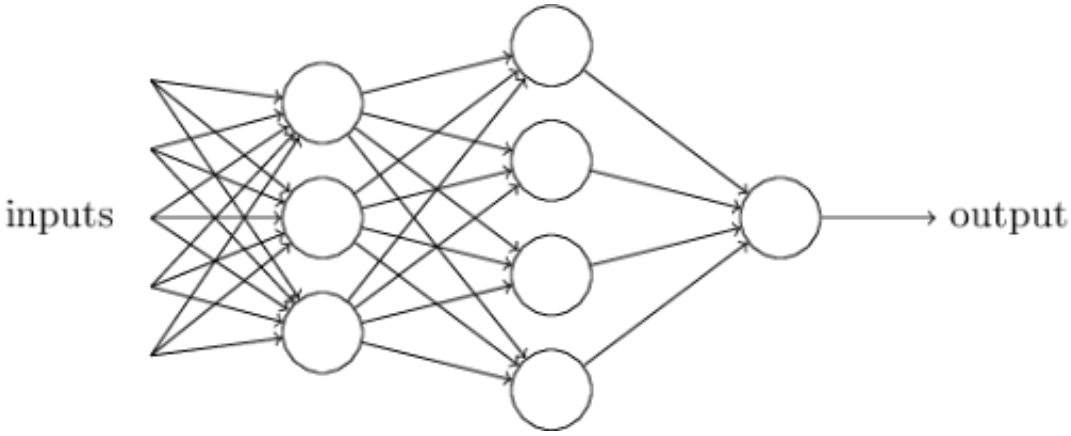


Figure 2.4: An example of neural network with two hidden layers

Convolutional layers, which are fundamental components of the networks used in this study, are an example. These layers apply a set of filters (also called *kernels*) across the input data to detect patterns such as edges and textures. The advantage of convolutional layers lies in their ability to maintain spatial relationships between data by capturing local patterns but also reducing the number of parameters if compared to fully connected layers.

The core of the learning process of a neural network is back-propagation, a technique that allows the network to adjust its weights to improve the performance. The goal is to minimize a loss function that measures the error between the network's predictions and the actual values. An example of loss function is the Mean Squared Error (MSE) computed as:

$$L = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

where:

- L: The loss function value.
- m: The number of training examples in the dataset.
- y_i: The true value for the i-th training example.
- $\hat{y}_i$: The predicted value for the i-th training example.

The back-propagation algorithm works by calculating the partial derivatives (*gradient*) of the loss function with respect to each weight in the network. By applying this procedures iteratively it is possible to adjust all the weights, increasing the accuracy of the network [3].

In this study different pre-trained networks are employed, which consist of neural networks with custom weights obtained through training on a big dataset, to extract deep features from patient's images that will later be used for classification. These datasets are:

1. *ImageNet* is a massive dataset created by a group of researchers [13] that contains millions of labeled images across thousand of categories, ranging from animals and objects to scenes and everyday items. Each image is assigned with one or more labels that specify its content. This large-scale labeled dataset allows neural networks to generalize well to various visual tasks such as edge detection and object recognition.
2. *RadImageNet* is a specialized medical imaging dataset containing over 1.3 million labeled images across a variety of radiological categories, such as CT, MRI, US and X-ray images. The dataset is publicly available and can be used for different applications [26].

2.2.2. Networks descriptions

Here is a description of the structure of the four networks involved in the study.

ResNet50

ResNet, short for *ResidualNetwork*, is a deep CNN architecture that uses residual connections which allow the network to be much deeper without degrading performance [18]. ResNet50 is one of the variants of ResNet containing 50 layers of various components such as convolutional layers, batch normalization layers, pooling layers and residual blocks. The residual blocks are the turning point in ResNet's design since they solve the vanishing gradient problem that occurs when gradients diminish as they propagate back through

the network.

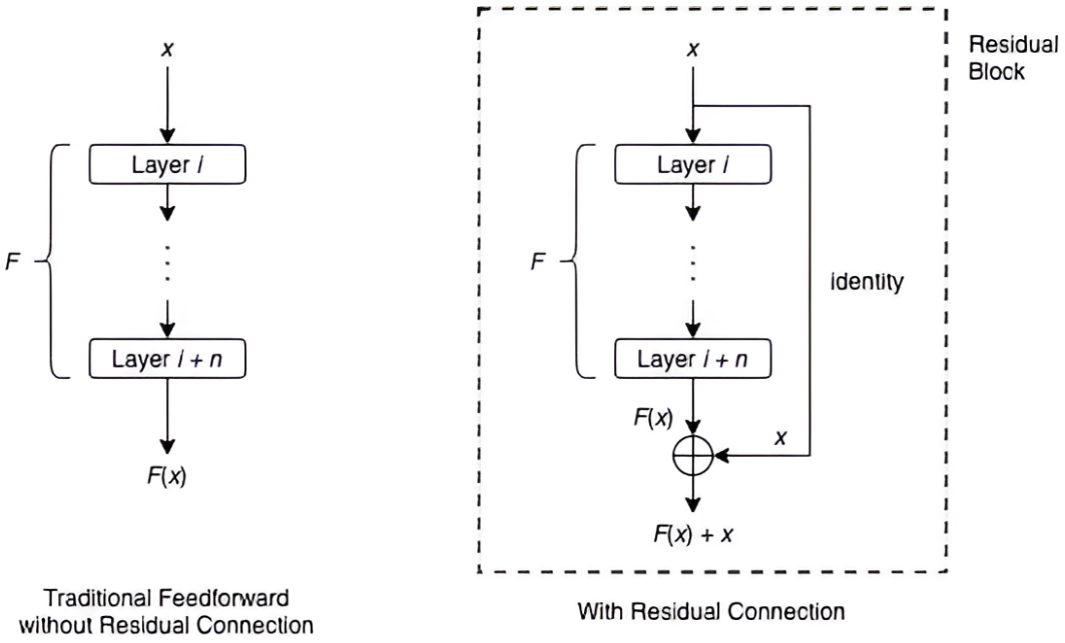


Figure 2.5: A traditional feedforward block on the left and a residual block on the right

As shown above, the residual block contains an additional connection called *skip connection* allowing the network to learn residuals and not the entire transformation from scratch. The generated output is of the type:

$$Output = F(x) + x$$

where *F(x)* represents the transformation of the input throughout the different layers and x represents the input [26].

In this study, ResNet50 is tested in the version trained on RadImageNet.

InceptionV3

InceptionV3 is a deep CNN architecture developed by Google that focuses on efficient computation and multi-scale feature extraction [38]. Unlike traditional networks that apply a single convolution at each layer, InceptionV3 uses inception modules that process information at multiple scales simultaneously, allowing the network to capture fine details.

Firstly the inception module performs dimensionality reduction to minimize the computational load. Secondly the module performs multi-scale feature extraction by applying multiple convolutional filters of different size in parallel. Finally the outputs are concate-

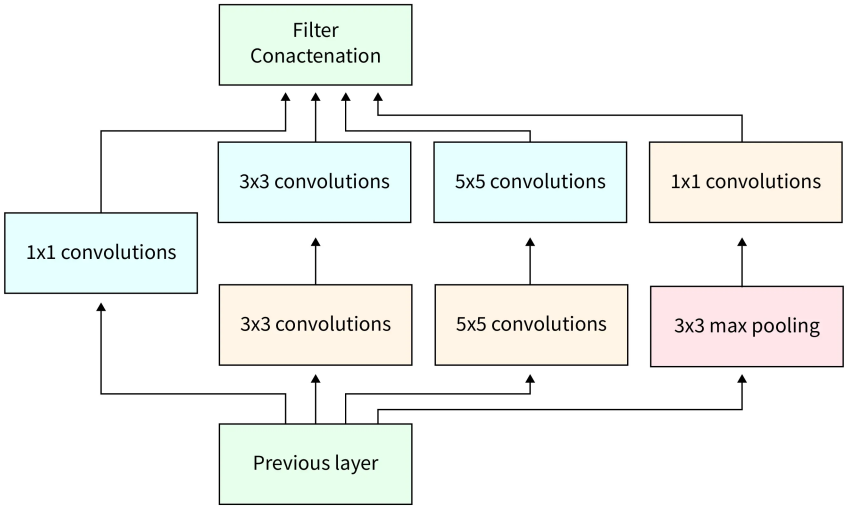


Figure 2.6: The inception module of InceptionV3

nated allowing the network to combine the information captured at different scales into a single feature map [26].

This model is used in the version trained on RadImageNet.

InceptionResnetV2

InceptionResNetV2 is a hybrid architecture that combines the multi-scale feature extraction of the Inception modules with the residual connections of ResNet. The goal of the architecture is to capture complex features across different scales while also addressing the problem of the vanishing gradients [39].

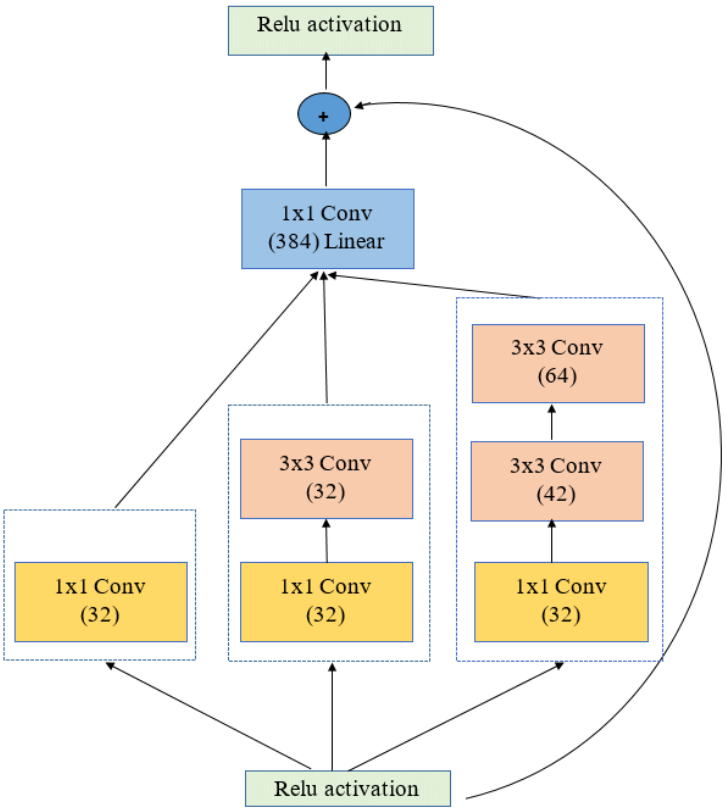


Figure 2.7: The InceptionResNet module obtained as a combination of residual blocks and Inception module

This model is again used in the version trained on RadImageNet.

VGG19

VGG19 is a deep CNN architecture developed by the Visual Geometry Group (VGG) at the University of Oxford. It is formed by 19 layers where 16 are convolutional layers with a filter dimension of 3x3 with a stride of 1 [37].

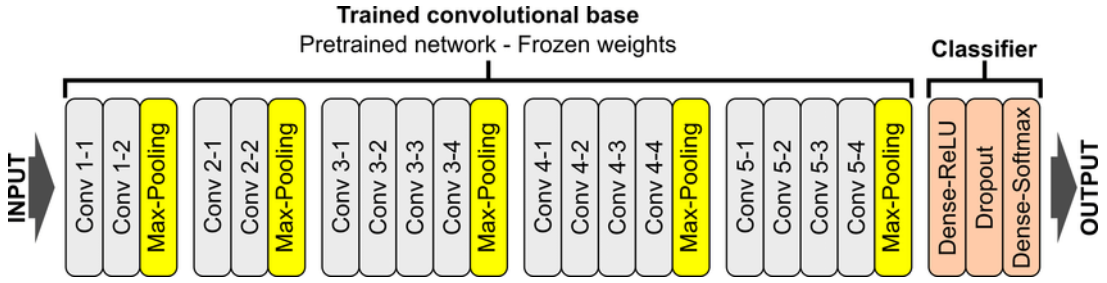


Figure 2.8: Schema of the architecture of VGG

Differently from the other networks, VGG19 model in this research is pre-trained on Imagenet.

2.2.3. Feature extraction

Prior to feature extraction from the pre-trained neural networks utilized in this study, the fully connected layers of the networks are removed to obtain a vector of feature maps represented as follows:

$$F = (C, H, W)$$

where:

- C: The number of feature maps.
- H: The height dimension of each feature map.
- W: The width dimension of each feature map.

Once obtained the feature maps vector, Global Max Pooling (GMP) is applied as a first dimensionality reduction technique, keeping only the most relevant features. GMP is defined as:

$$\text{GMP}(F)_c = \max_{0 \le i < H} \max_{0 \le j < W} F_{c,i,j} \quad \text{for } c = 1, 2, \ldots, C$$

where:

- $\text{GMP}(F)_c$: The result of the global max pooling for the c-th feature map, which is a single value.
- $\max_{0 \le i < H} \max_{0 \le j < W} F_{c,i,j}$: The maximum value taken across all spatial positions (i, j) within the c-th feature map.

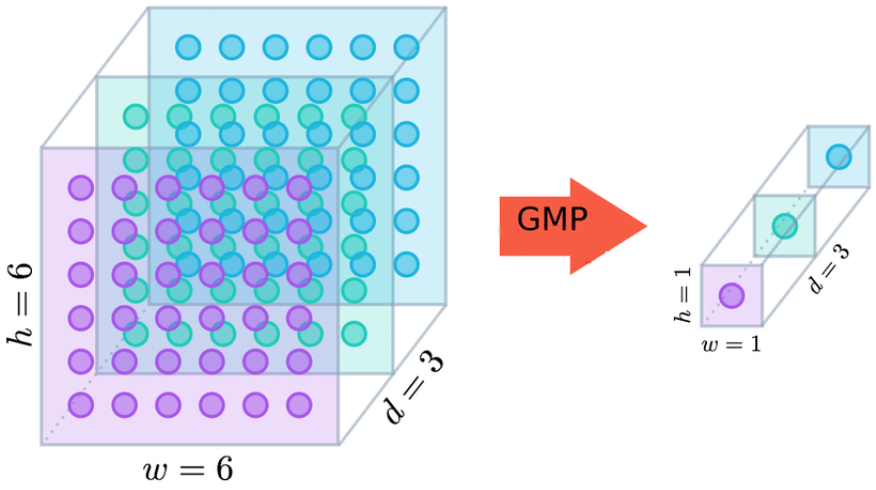


Figure 2.9: Graphical representation of GMP where C = 3, H = 6 and W = 6

2.3. Feature selection

Feature selection is a dimensionality reduction technique used to identify the features that are relevant to ML tasks. By removing redundant features that provide duplicate information, as well as irrelevant features that don't add value to the learning process, feature selection can enhance the performance of ML algorithms, accelerate training, reduce overfitting, and simplify models.

For a dataset d, feature selection involves selecting a subset of features, $f' = \{f'_i | i = 1, 2, 3, ..., m\}$, from the original feature set $f = \{f_i | i = 1, 2, 3, ..., n\}$ by satisfying conditions $m < n$ and $\arg\max(T) = f'$, where the goal is to maximize T, a target function such as classification accuracy or cluster quality [31].

Feature selection techniques are classified as supervised or unsupervised, depending on the availability of labeled data:

1. *Supervised Feature Selection* applies to labeled data, evaluating the relationships between features and the target variable using an evaluation criterion or classifier. In this context, irrelevant features are those with weak or no association with the target variable.

2. *Unsupervised Feature Selection* is used on unlabeled data. Here selection relies on an assessment of clustering tendency, which determines if the data inherently contains meaningful clusters. These algorithms (i.e., PCA) choose features that enhance this tendency, removing noisy features that detract from cluster formation.

Several studies have classified feature selection methods into three main categories:

- *Filter Methods*: rather than selecting features directly, they rank the entire feature set based on an evaluation function (e.g., distance, information entropy, accuracy, correlation, or consistency). Feature selection is then performed by the user based on these rankings. These methods rely on statistical and mathematical criteria rather than a classifier to determine feature relevance. Filters are further divided into univariate and multivariate approaches. Univariate filters assess each feature independently, while multivariate filters consider feature inter-dependencies.
- *Wrapper Methods*: they generate feature subsets using a search strategy (e.g., randomized search), evaluate each subset with a classifier as a "black box," and apply a stopping criterion (such as a maximum iteration count). Wrappers consider feature dependencies and generally generalize better than filters, but are computationally intensive. Recursive Feature Elimination is a classic example.
- *Embedded Methods*: Feature selection in these methods occurs within the classifier training process itself, yielding both a refined feature set and a trained model. They are less computationally intensive and less prone to overfitting than wrapper methods. Additionally, embedded methods can account for feature dependencies. Tree-based algorithms, such as CART, exemplify embedded methods by selecting features based on their contribution to classification during the training process [31].

2.3.1. Feature selection algorithms employed

In this study, various feature selection algorithms were employed in different steps. In the table below these selectors are listed and classified into categories, followed by a theoretical description.

Algorithm	Method	Label	Type/Search Strategy
Pearson Correlation	Filter	Supervised/ Unsupervised	Univariate/Multivariate
ANOVA F-test p-value	Filter	Supervised	Univariate
mRMR	Filter	Supervised	Univariate, Multivariate
RF	Embedded	Supervised	Tree-based method
LR	Embedded	Supervised	Regularization Method
LASSO	Embedded	Supervised	Regularization Method

Table 2.8: Classification of Feature Selection Algorithms used

The *Pearson correlation coefficient* is a statistical measure that quantifies the strength and direction of the linear relationship between two quantitative variables. Denoted by r, this coefficient is calculated by evaluating the covariance between the variables, normalized

by the product of their standard deviations. The value of r ranges between -1 and 1, where $r = 1$ indicates a perfect positive linear correlation and $r = -1$ indicates a perfect negative linear correlation [4]

The formula for calculating the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where:

- x_i and y_i represent the observed values of the two variables,
- $\overline{x}$ and $\overline{y}$ are the arithmetic means of variables x and y, respectively,
- n is the total number of observations in the sample.

The Pearson correlation is considered a filter method because it evaluates each feature independently of a classifier. It can be used both as an unlabeled method, relying on the correlation coefficient to measure the linear relationship between features, or as a labeled method, when the correlation is between each feature and the target variable.

The *ANOVA F-test p-value* is derived from the One Way ANOVA (Analysis of Variance) F-test, a parametric test used in feature selection to evaluate the significance of differences in means across two or more independent classes. This test assumes that the data are normally distributed, that the variances are equal across groups, and that the observations are independent. It is used to determine whether a specific feature has a statistically significant difference in mean values among multiple classes, making it a *supervised*, *filter*, and *univariate* method. The test compares the means of the groups and calculates a p-value, indicating the probability of observing the data if the null hypothesis is true. The null hypothesis posits that there is no significant difference in the means of the groups. If the p-value is below a chosen significance threshold, the null hypothesis can be rejected, suggesting that the feature significantly distinguishes among the classes and is therefore important to retain.

The formula for the F-statistic F in the ANOVA F-test is:

$$F = \frac{\text{between-group variance}}{\text{within-group variance}} = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2}{\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}$$

where:

- k is the number of groups,
- n_i is the sample size of the i-th group,
- $\overline{x}_i$ is the mean of the i-th group,
- $\overline{x}$ is the overall mean of all observations,
- N is the total number of observations.

The *Minimum Redundancy Maximum Relevance* (mRMR) method is a feature selection approach that identifies features which are both highly relevant to the target variable, measuring target-feature mutual information and so contributing valuable predictive information, and are minimally redundant with each other, calculating the mutual information between them. The goal of mRMR is to maximize relevance while minimizing redundancy, as expressed in the following objective function:

$$\text{mRMR} = \max \left(\frac{1}{|S|} \sum_{x_i \in S} I(x_i; y) - \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \right)$$

where:

- $I(x_i; y)$ represents the mutual information between feature x_i and the target y, measuring relevance,
- $I(x_i; x_j)$ represents the mutual information between pairs of features x_i and x_j, measuring redundancy,
- S is the set of selected features [46].

The *RF* feature selection method leverages the importance scores generated by a RF model to identify the most informative features. RF builds multiple decision trees on subsets of data and aggregates their results. It is an embedded method since during the learning process it calculates importance scores for each feature based on how much each feature contributes to reducing impurity, specifically Gini impurity in the chosen implementation. For a node with K classes, the Gini impurity is defined as:

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k is the probability of selecting an observation belonging to class k at that node.

The importance of a feature is calculated based on the average reduction in Gini impurity across all splits where the feature is used [47]. The formula for calculating the importance of a feature x_i is:

$$\text{Importance}(x_i) = \frac{1}{T} \sum_{t=1}^{T} \sum_{s \in S_t(x_i)} \Delta\text{Gini}(s)$$

where:

- T is the total number of trees in the RF,
- $S_t(x_i)$ is the set of all splits based on feature x_i in tree t,
- $\Delta\text{Gini}(s)$ is the reduction in Gini impurity achieved by split s.

The *LR* feature selection method identifies important features by analyzing the absolute values of the coefficients associated with each feature after training a LR model. LR calculates a coefficient for each feature, indicating its impact on predicting the probability of belonging to a specific class. Higher absolute coefficient values signify a stronger influence on the target variable, while they are irrelevant the more they get towards 0.

The probability of belonging to class 1 given the feature vector X is calculated as:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

where:

- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for features $x_1, x_2, \ldots, x_n$,
- e is the base of the natural logarithm.

After training, the importance score for each feature x_i is given by the absolute value of its coefficient $|\beta_i|$ [9].

LASSO (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that allows both feature selection and regularization by incorporating L1 regularization to the MSE Loss function. The objective function can be written as:

$$\text{Loss} = \sum_{i=1}^{N} \left(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where:

- y_i is the actual value for observation i,
- x_{ij} is the j-th feature of observation i,
- β_j is the coefficient associated with feature x_j,
- λ is the regularization parameter controlling the strength of the L1 penalty.

The L1 penalty term, $\lambda \sum_{j=1}^{p} |\beta_j|$, encourages sparsity in the model by shrinking some coefficients exactly to zero as λ increases. This means that as the penalty strength grows, LASSO will set the coefficients of less important features to zero, thereby selecting only the most relevant features and reducing the complexity of the model [29].

2.4. Feature classification

Feature classification is another fundamental process in ML that entails training a classifier to assign labels to a set of input features. This process involves learning the relationship between the selected features and the corresponding output labels from the training data. Once trained, the classifier can then be used to make predictions on new, unseen data points. The primary objective is to accurately categorize these data points, allowing for effective decision-making based on the learned patterns in the dataset.

In the context of a dataset d, feature classification involves applying a function $C : f' \rightarrow y$, where $f' = \{f'_i | i = 1, 2, 3, \ldots, m\}$ is the subset of selected features, and y represents the predicted class labels. This study focuses specifically on binary classification, which involves categorizing data points into one of two distinct classes. The objective is to optimize the performance of the classifier by maximizing a performance metric P, ensuring that the model generalizes well to unseen data. In the next two sections it's given a theoretical description of the classification algorithms employed and of the metrics taken into account when evaluating the results.

2.4.1. Classification Algorithms employed

Random Forest

The RF algorithm operates by constructing an ensemble of decision trees, each trained on different subsets of the data. This approach, known as bootstrap aggregation or bagging, involves generating multiple samples of the dataset by randomly selecting instances with replacement, allowing each tree to specialize in different aspects of the data. In addition to sampling data points, RF also selects a random subset of features for each split in the trees, promoting further diversity among the trees. This dual randomness—sampling both instances and features—helps capture a broader representation of the data patterns, ultimately producing a more generalizable model and avoid overfitting.

A key element in building each decision tree is determining how to split the data at each node. For this, RF uses the Gini index (explained in *section 3.3.4*), a measure of "impurity" that indicates how well a potential split separates the classes. A node with mixed classes has high impurity, whereas a node with a single class has zero impurity. The algorithm iterates through possible splits, choosing the one that most reduces impurity, thus leading to purer nodes.

RF is particularly effective in handling complex, high-dimensional datasets, where relationships among features may not be linear or immediately apparent [3].

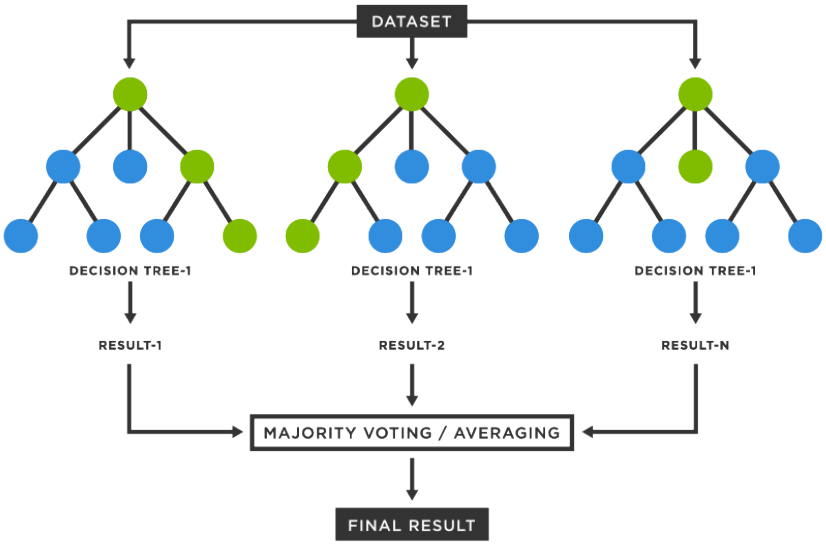


Figure 2.10: Basic functioning of the RF algorithm

Logistic Regression

The LR algorithm is a popular classification technique used to model the probability of a binary outcome based on one or more predictor variables. LR takes a more direct approach since it fits a single linear model that predicts the likelihood of a given class. This prediction is achieved through the logistic function (or *sigmoid*) which has been explained in *section 3.3.4*.

In LR, the parameters β define a decision boundary, which separates the classes based on the predictor variables and the goal is to find the values of β that best fit the observed data.

In order to do this it uses the Maximum Likelihood Estimation, that tries to find the parameters β that maximize the probability of correctly classifying the observed outcomes. The likelihood function is expressed as:

$$L(\beta) = \prod_{i=1}^{N} P(y_i|x_i, \beta)$$

where $P(y_i|x_i, \beta)$ represents the predicted probability of the observed class y_i given the predictor values x_i and the coefficients β.

For computational simplicity the log-likelihood is instead maximized:

$$\log L(\beta) = \sum_{i=1}^{N} [y_i \log(P(y_i|x_i, \beta)) + (1 - y_i) \log(1 - P(y_i|x_i, \beta))]$$

In this expression:

- y_i is the actual class label (0 or 1) for each observation,
- $P(y_i|x_i, \beta)$ is the probability that the model assigns to the observed class.

Maximizing this log-likelihood function allows to find the coefficients β that make the decision boundary most effective at separating the classes.

Logistic Regression is particularly effective for datasets where relationships among features and outcomes are linear or approximately linear [3].

Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm that aims at finding the optimal hyperplane that maximizes the margin between classes. This margin maximization helps improve the model's generalizability. For data that is not linearly separable, SVM uses kernel functions to project data into a higher-dimensional space, allowing it to construct nonlinear boundaries.

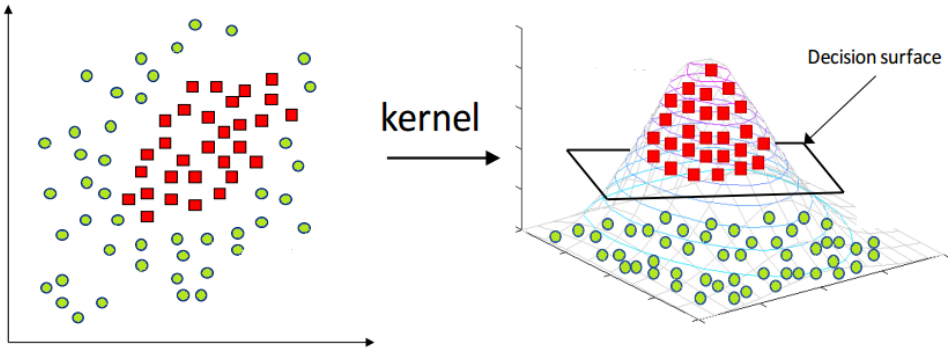


Figure 2.11: An example of how the points are projected in a high dimensional space to find a decision boundary

Several types of kernels can be applied based on the characteristics of the data. The linear kernel is useful when data is approximately linearly separable, while the polynomial kernel introduces polynomial combinations of features, enabling the model to capture more complex, curved decision boundaries. The radial basis function (RBF) or Gaussian kernel is particularly popular for handling nonlinear relationships, as it clusters nearby points together to create smooth decision boundaries in higher-dimensional space. Additionally, the sigmoid kernel, inspired by neural networks, acts similarly to an activation function but is less commonly applied in SVM.

The RBF kernel function is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

where:

- $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors for two data points,
- γ is a parameter that controls the width of the RBF kernel.

Using the kernel trick, SVM can incorporate this RBF transformation without explicitly

calculating the coordinates in the higher-dimensional space. The decision function in the dual form becomes:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

where:

- α_i are the Lagrange multipliers,
- y_i represents the class label,
- $K(\mathbf{x}_i, \mathbf{x})$ is the RBF kernel function,
- b is the bias term.

The optimization process in SVM aims to find the values of α_i that will maximize the margin between the classes [3].

Ensemble

An ensemble classifier combines multiple models to improve predictive performance. In this study, the ensemble classifier is constructed by combining three different models: RF, SVM with RBF kernel, and LR. Each of these models have the possibility contributes unique strengths: RF captures complex feature interactions through multiple decision trees, SVM with RBF kernel effectively handles nonlinear relationships by mapping data into higher-dimensional space, and LR provides a straightforward, interpretable model for linear boundaries.

To combine the predictions of these individual models, a technique called soft voting was applied. In soft voting, each model in the ensemble outputs a probability for each class rather than a direct class prediction (hard voting). These probabilities indicate each model's confidence in an instance belonging to a specific class. The ensemble then calculates the average of the predicted probabilities across the models, assigning the final class label based on the highest averaged probability.

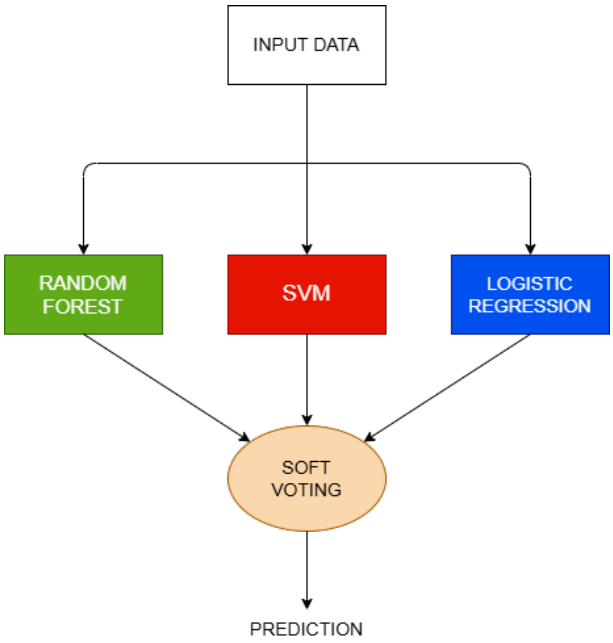


Figure 2.12: Representation of the ensemble involved in this study

eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is an advanced and highly efficient machine learning algorithm within the gradient boosting framework, known for its strong performance in classification tasks. XGBoost builds a series of decision trees sequentially, where each tree is trained to correct the residual errors of the predictions from the previous trees. This iterative process allows XGBoost to progressively improve its predictions by focusing on harder-to-classify instances. At each step t, XGBoost minimizes a specific loss function L using gradient descent, which updates the model by adding a new tree f_t that best reduces the errors of the previous predictions. The objective function is defined as:

$$\text{Objective} = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} \Omega(f_t),$$

where:

- $L(y_i, \hat{y}_i^{(t)})$ measures the difference between the true values y_i and predicted values $\hat{y}_i^{(t)}$,
- $\Omega(f_t)$ is a regularization term that controls the complexity of each tree f_t to prevent overfitting.

By adjusting predictions based on both the gradient and second-order gradient (or Hes-

sian) of the loss function, XGBoost accurately corrects the model's previous errors [8].

MultiLayer Perceptron

Multilayer Perceptron (MLP) is a type of neural network commonly used for classification and regression tasks. It composed of multiple layers of interconnected neurons arranged in an input layer, one or more hidden layers, and an output layer. Each neuron in a layer connects to every neuron in the subsequent layer, creating a fully connected network. MLP learns to map input features to target outputs by adjusting weights associated with each connection based on the back-propagation algorithm. This process involves calculating the loss function L, which measures the error between the predicted output $\hat{y}$ and the actual target y, and then propagating this error backward through the network.

The loss function used for classification in MLP is cross-entropy, defined as:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}),$$

where:

- N is the number of samples,
- C is the number of classes,
- $y_{i,c}$ is a binary indicator if class label c is the correct classification for sample i,
- $\hat{y}_{i,c}$ is the predicted probability for sample i belonging to class c.

This cross-entropy loss penalizes incorrect classifications by assigning a high loss to predictions far from the true class probability, thus guiding the model towards higher accuracy [3].

2.4.2. Metrics

The *Precision* (or Specificity) metric measures the accuracy of positive predictions, defined as the ratio of true positives (TPs) to the total predicted positives (the sum of TPs and false positives (FPs) [5]. It is formulated as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

The *Recall* (or Sensitivity) metric indicates the ability of the model to identify all actual

positives in the dataset. It is defined as the ratio of TP to the sum of TPs and false negatives (FNs):

$$\text{Recall} = \frac{TP}{TP + FN}$$

Using Precision and Recall, *F1-score* can be calculated. It is the harmonic mean of Precision and Recall, offering a single metric that balances both. It is given by:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Another essential metric is the *Accuracy*, which measures the proportion of correct predictions (TPs and TNs) out of all predictions. The formula for accuracy is:

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

To better understand the model's classification ability, the ROC AUC (Receiver Operating Characteristic - Area Under Curve) was examined. The ROC AUC score represents the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative instance.

The ROC curve itself is a plot that represents the relationship between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. Each point on the curve corresponds to a specific threshold, with the curve starting from the origin (0,0) — indicating no positive predictions — and ending at (1,1), where all predictions are classified as positive. A model with perfect discrimination would have an ROC curve that passes through the point (0,1), indicating a TPR of 1 and a FPR of 0, which would result in an AUC of 1 [5].

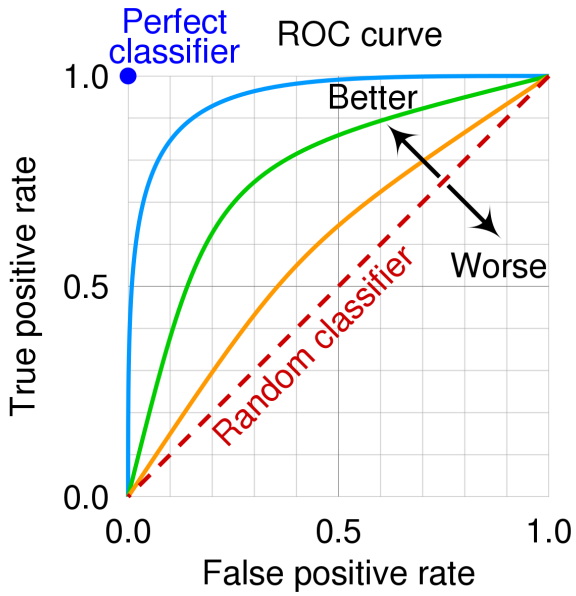


Figure 2.13: ROC curve showing different classifier performances. Better-performing models have curves closer to the top-left corner, while the red dashed line indicates a random classifier (AUC = 0.5).

The Confusion Matrix provides detailed insights into the types of errors the model makes, displaying counts of TPs, FPs, TNs and FNs. This matrix is particularly useful for understanding the distribution of predictions and misclassifications across classes.

		Predicted Values
Actual Values	TP	FP
	FN	TN

Table 2.9: Confusion Matrix showing TPs, FPs, FNs, and TNs

Finally, the Balanced Accuracy metric, which was used as main metric for result evaluation, addresses the issue of class imbalance by averaging the recall (or TPR) for each class. This approach ensures that each class contributes equally to the final metric, regardless of its frequency in the dataset. Balanced accuracy can be calculated as follows:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

In this formula, the balanced accuracy considers both the recall of the positive class and the recall of the negative class, providing a fairer evaluation when the classes are not evenly represented [5].

3 | Materials and Methods

3.1. Patient and image Dataset

This dataset includes carotid plaques images of 129 patients who underwent elective CEA at the Vascular Surgery Operative Unit of Fondazione IRCSS Ca 'Granda, Ospedale Maggiore Policlinico, Milan, and had pre-operative echo-color Doppler and CTA scans available. Echo-color Doppler analysis was performed using an Affinity 50 ultrasound scanner with an 8MHz linear probe (Philips Ultrasound, Bothell, WA) to evaluate the degree of stenosis pre-operatively. Specifically, the stenosis percentage was classified following the guidelines of the ESVS, based on the North American Symptomatic Carotid Endarterectomy Trial (NASCET) criteria, which considers peak systolic velocity, end-diastolic velocity, and their ratios in both the ICA and CCA. The study received approval from the IRCCS Fondazione Policlinico Ethical Committee in accordance with institutional ethical guidelines, and all patients provided informed consent.

Of the 129 patients, 53 were symptomatic, presenting with either TIA or ischemic strokes. *Table 3.1* outlines the main clinical characteristics of these patients.

The type of image used in this study is CTA. These images were acquired using a GE Light Speed VCT 64-slice 3T scanner (GE Healthcare, Little Chalfont, UK)*Figure 3.1.* The key acquisition parameters included a slice thickness of 0.625 mm, a reconstruction matrix of 512x512 pixels, and a final resolution of 0.39 mm x 0.39 mm x 0.625 mm. The scans obtained were saved in DICOM format, while the ROIs in NRRD format.

Characteristic	Set (129)	Asymptomatic (76)	Symptomatic (53)
Age (mean, SD)	73.6 (7.87)	72.76 (8.09)	74.81 (7.46)
Male	82 (63.6%)	43	37
Female	47 (36.4%)	33	14
Diabetes	27 (21%)	13	14
Hypercholesterolemia	57 (44%)	37 [1]	20 [1]
Smoking	21 (16%)	10	11
Hypertension	99 (77%)	58 [2]	41 [1]
Hypertension treatment	104 (81%)	58 [2]	41 [3]
Statin	103 (80%)	59 [3]	44 [3]
Antiplatelet treatment	120 (93%)	70 [1]	47 [2]
Anticoagulants	12 (9%)	3 [2]	9 [3]
Obesity	5 (4%)	3	2 [1]
Stenosis (mean, SD)	77% (10%)	78% (8%)	76% (13%)

Table 3.1: Clinical characteristics of the 129 patients, grouped into asymptomatic and symptomatic categories. Missing data is indicated in square brackets.

Figure 3.1: GE Light Speed VCT scanner

3.1.1. Image segmentation and visualization

Arterial lumen and external wall surfaces were manually segmented from CTA images using MIMICS software and exported as stereolitographic files. Plaque proximal and distal locations were provided by expert radiologists. To generate binary masks of the carotid plaques representing the ROI, the lumen and external artery surfaces were semi-automatically cut in correspondence of the proximal and distal plaque locations and processed with 3D Slicer.

3D Slicer is an open-source software platform specifically designed for medical image analysis. It supports a wide range of imaging data types, including DICOM and NRRD, and offers comprehensive tools for image segmentation, registration, and visualization. It allows for the simultaneous visualization of three main sections of the plaque—axial, coronal, and sagittal planes—as well as a 3D reconstruction of the plaque. 3D Slicer enables the overlay of the ROI, which appears in green in the images, providing a clear delineation of the plaque boundaries (as shown in *Image 3.2*). For image extraction purposes, the axial section was selected.

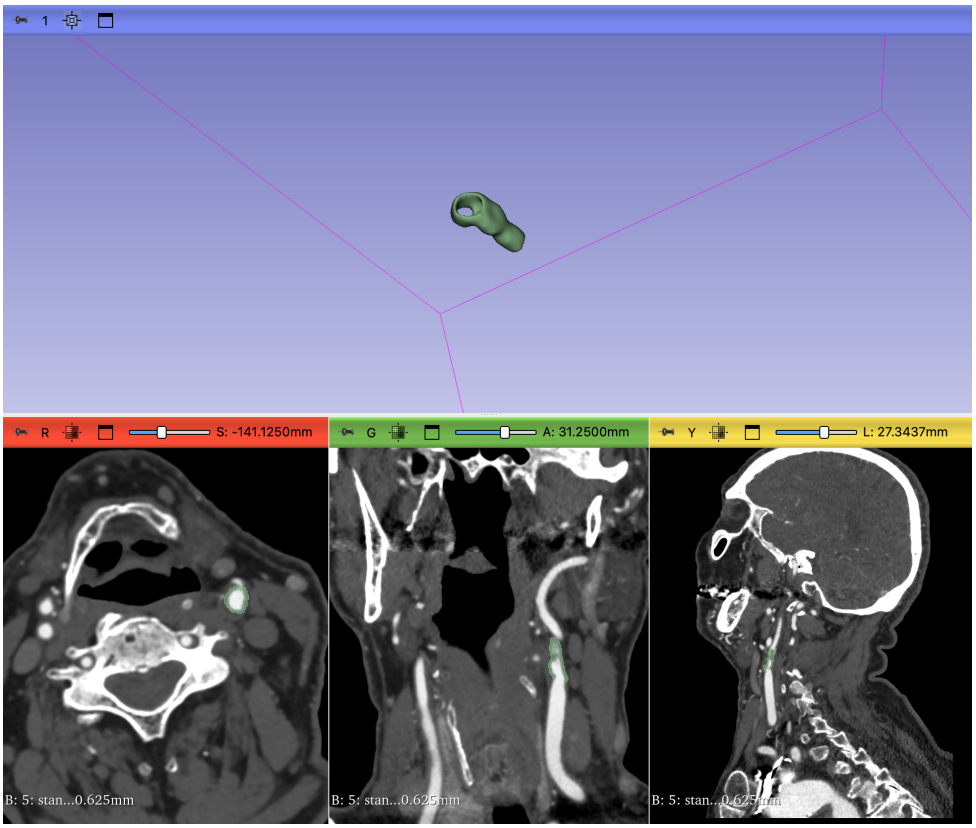


Figure 3.2: Visualization of carotid plaque in 3D Slicer showing three planes—axial (left), coronal (middle), and sagittal (right)—along with a 3D reconstruction of the plaque (top). The green overlay indicates the ROI

In this study, it was necessary to overlay the full plaque images with the NRRD mask (ROI) images. However, in some cases, the images did not align perfectly. To overcome this issue, 3D Slicer was used to convert the DICOM images into NRRD format, ensuring that all images were in a compatible format and could be accurately overlapped.

3.2. Feature Extraction

Feature extraction plays a huge role when it comes to ML and DL. It transforms raw data into a set of meaningful and easily understandable features. These features are then used as input for the algorithms. The extraction process for radiomic analysis is focused on pre-set feature definitions (i.e., handcrafted features), as opposed to the ones based on DL which autonomously identify complex patterns and hierarchical representations from the images.

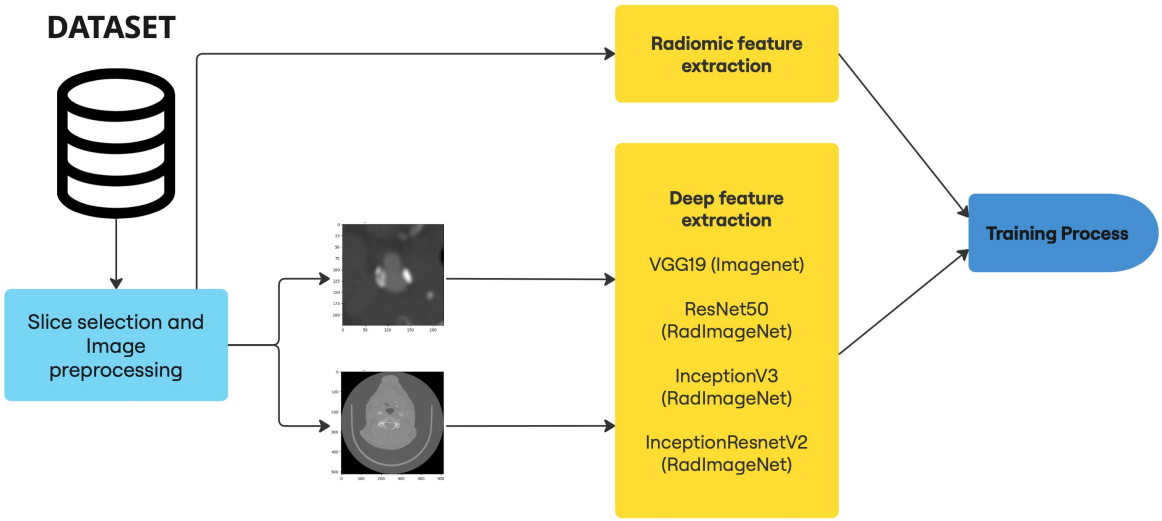


Figure 3.3: Schema representing the primary steps to feature extraction

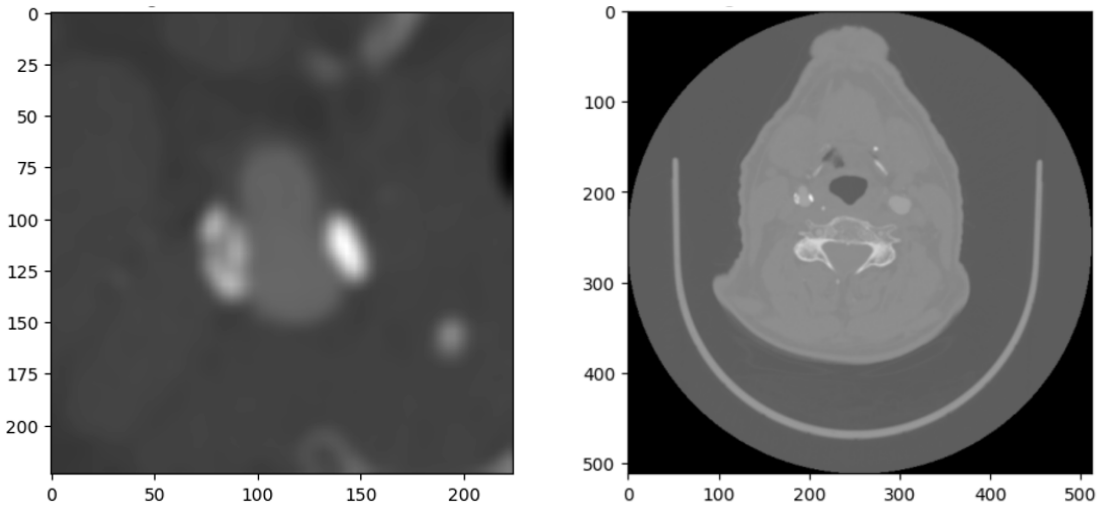
3.2.1. Slice selection and Image preprocessing

In the case of radiomics features no image preprocessing was performed. Since the analysis is on 2D images and every slice is associated with a set of radiomic features, all the slices and the ROIs were passed one by one to the feature extractor.

In order to extract the deep features instead, some preprocessing operations were performed. Two types of slices were created:

1. *Cropped* slices have been obtained by cropping the region around the plaque.

2. *Full* slices consist of the full original slice containing all the scan.



Cropped image

Full image

Figure 3.4: Examples of the two types of slices

The two types of slices (shown in *Figure 3.4*) can be considered as a sort of different dataset on which the analysis was conducted separately. The obtained slices are then both resized to a dimension of 224 x 224, which is the dimension input accepted by all the networks.

In the case of the network trained on ImageNet (VGG19), slices are normalized using the specific function of the network *preprocess_input*(). This function adjusts the input slices to match the format and range expected by the specific model.

In the case of network trained on RadImageNet (ResNet50, InceptionV3, InceptionResNetV2), min-max normalization was performed to obtain an input scaled between [0,1]. The min-max normalization formula scales a value x to a new range based on the minimum and maximum values in the data:

$$x' = \frac{x - \min}{\max - \min}$$

where:

- x: The original value.
- x': The normalized value within the range $[0, 1]$.
- min: The minimum value in the original image.

- max: The maximum value in the original image.

In addition, two different approaches were investigated for slice selection:

1. 2D approach only considers one single slice per patient, that is the one with the largest ROI area,
2. 2.5D approach takes all the slices of the patient whose ROI area is at least 30% of the largest ROI, resulting in different number of slices for every patient (being 13 the minimum and 97 the max number of slices considered, with an average of 32). *Table C.1* in *Appendix C* reports the number of slices kept for every patient.

The analysis was performed separately for the two approaches.

In order to create the right input for the networks, firstly the single channel (being the images black and white) was repeated in order to create a vector of dimensions (224, 224, 3). Then the batch dimension was added at the beginning of the vector, this is (1, 224, 224, 3) for 2D since only one slice is used, and (n, 224,224,3) for 2.5D where n is the number of slices that at least contains one piece of the plaque, hence different for every patient as a consequence of the different plaque configuration. The selection of the 30% of the area was performed before classification for noise removal purposes, to select only the most informative slices.

3.2.2. Radiomic feature extraction

Radiomic features were extracted from CT images using the PyRadiomics library, a widely-used Python package for radiomics analysis. PyRadiomics performs the automated extraction of a wide range of quantitative features from different medical images. Most of the features extracted are in compliance with feature definitions as described by the Imaging Biomarker Standardization Initiative, which consists of a standardized set of 169 features. This initiative was created to facilitate the validation and calibration of various radiomics software tools and addresses the inconsistencies in extraction methods employed across different studies and clinical settings [44].
The feature extraction process involved using both an image file and a corresponding ROI mask, which were converted to NRRD format.

A total of 474 radiomic features were extracted, of which 102 coming from the original image and 372 from the wavelet-transformed one (90 first-order statistical, 9 shape-based and 375 textural features).

3.2.3. Deep feature extraction

The features, extracted using the GMP method (as described in *Section 2.2.3*), were saved in a file containing one row for every slice of each patient and as many columns as the number of features. Given the differences in the network structure, different amount of features were obtained:

- 512 for VGG19
- 1536 for InceptionResNetV2
- 2048 for ResNet50
- 2048 for InceptionV3

3.3. Training Process

Figure 3.5 delineates the training workflow and the main steps.

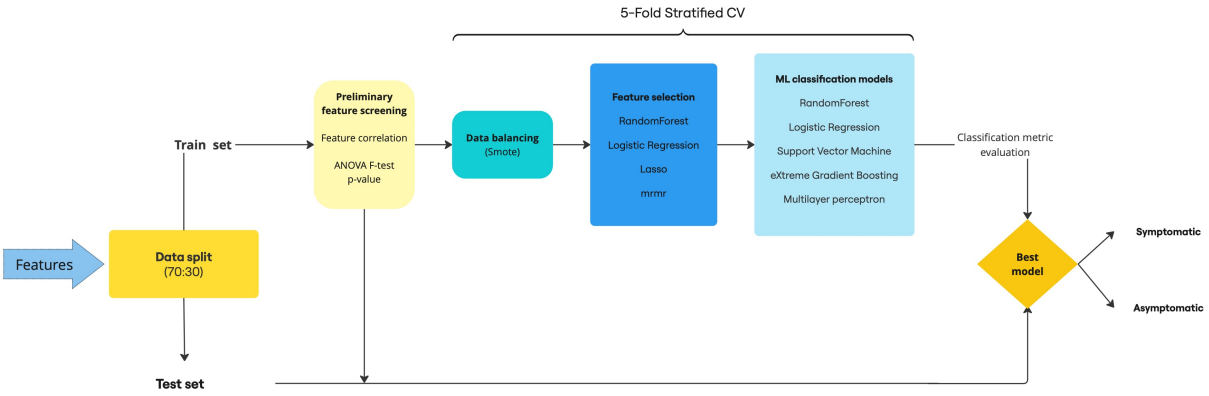


Figure 3.5: Schema representing the steps of the training process

The training process began by splitting the dataset consisting on the extracted features into training and testing sets (90 vs 39 patients), where 30% of the patients were reserved for testing, according to the date of surgery. This choice reflects a practice commonly adopted in clinical studies in which the aim is to develop a model that can generalize to future cases based on prior data, simulating a retrospective scenario where past patient information informs future diagnostic predictions.

3.3.1. Preliminary Feature Screening

Multiple feature selection steps were applied to reduce dataset dimensionality, especially given the high number of initial deep features. The functioning of all these methods is explained in *Section 2.3.1.* The first step was a feature correlation analysis using Pearson correlation, setting a threshold of 0.9 to eliminate highly redundant features. Next, a ANOVA F-test p-value filter with a threshold of 0.05 was applied to further refine the features based on their statistical relevance to the symptomatic and asymptomatic classes.

3.3.2. Cross Validation

In order to provide a robust estimate of model performance and reduce the risk of overfitting, especially given the limitations in the dimensionality of the dataset, stratified 5-fold cross validation (CV) was applied on the training set. The central idea behind k-fold CV is to partition the dataset into k equally sized subsets, known as "folds," and iteratively train and validate the model across different portions of the data. In each iteration, one fold is used as a validation set while the model is trained on the remaining k-1 folds. Stratified CV is used when dealing with imbalanced datasets, as it maintains the original class distribution within each fold.
All the subsequent steps described were performed iteratively on every train-validation fold.

3.3.3. Data Balancing

To address the imbalance of our dataset (76 asymptomatic vs 53 symptomatic), the Synthetic Minority Over-sampling Technique (SMOTE) was applied to all the 5 training sets to balance the class distribution. This is how SMOTE algorithm works: for each minority instance, it identifies its k (5 by default) nearest neighbors within the same class, then it generates synthetic points by randomly selecting positions along the line segments between the instance and its neighbors. These synthetic samples are added to the dataset, increasing the minority class's representation [23].

3.3.4. Feature selection

After the initial filtering steps, additional feature selection techniques were explored to identify the best features combination to give as inputs to the classifier. The feature selection methods tested are:

- RF, which uses Gini impurity and 100 estimators,

- LR, which has 2000 as maximum iterations parameter,
- LASSO,
- mRMR

All these selectors were tested in the CV loop with *num_features* parameter varying from 2 to 30 (for LASSO a range of 30 *alpha* values different for every network).

3.3.5. Machine learning classification models

The selected features were then passed to various ML classification algorithms (theoretically explained in *Section 2.4.1*). These are:

- RF, which uses Gini impurity and 100 estimators,
- LR, which has 2000 as maximum iterations parameter,
- SVM, with RBF kernel and enables probability estimates,
- Ensemble, combining RF, LR, and SVM using soft voting,
- MLP, initialized with three hidden layers of sizes 128, 64, and 32 neurons, a maximum of 1000 iterations, logistic activation, adaptive learning rate, and early stopping enabled,
- XGBoost, which utilizes default parameters.

3.3.6. Best model selection

At the end of CV process, aimed at assessing model performance across different configuration, the optimal combination of classifier, feature selector, and number of features (or α for LASSO) as hyper-parameter of the selector, could be identified as best model. This was done by calculating the mean and standard deviation of the balanced accuracy and ROC AUC for each configuration across the five validation folds. Then a grid search was implemented to systematically examine all combinations of classifiers, selectors, and hyper-parameters, focusing on finding the configuration with the highest mean balanced accuracy across folds. Grid search is a method that exhaustively explores a predefined parameter grid, in this case using the mean balanced accuracy as the primary criterion for ranking each configuration. By comparing these mean values, grid search allowed to identify the model yielding the best overall balanced accuracy, which was prioritized to account for class imbalance and minimization of the number of incorrect predictions. If multiple configurations achieved similar mean balanced accuracy values, models with

lower standard deviation of balanced accuracy or highest ROC AUC score were preferred, as this last metric provides insight into the model's ability to distinguish between classes at various thresholds.

3.4. Testing

Once the best model was determined for every network, the selected classifier was retrained on the entire training set (90 patients) to fully leverage the available data. For feature selection in this final model, only the features chosen in at least three out of the five folds in the initial CV cycle were retained. This approach ensured that only the most stable and relevant features, identified repeatedly across multiple folds, were used in the final model for testing. At these point the final classification metrics were obtained.

3.5. Classification Approaches

3.5.1. 2D

Following the 2D approach, classifiers were trained on a dataset consisting of single slices, specifically selecting the slice with the largest area for each patient. It is in this case that the best model was found through iteration of all the possible combinations of classifiers, selector and features number as parameter. This approach provides a foundational model that was later utilized for the 2.5D approach and for the combination and ensemble methods.

3.5.2. 2.5D

Once the best model for each network and for radiomics in the 2D case had been determined, it was then used in the 2.5D approach, which can be seen as a sort of data augmentation, since it takes into account not only one slice but several ones. Predictions were obtained for every slice and, since a single prediction is needed for every patient, three aggregation methods were tested. The method yielding to the highest balanced accuracy for one validation set (consisting of 27 patients taken from the 90 patients of the initial train), was then applied to the test set. This resulted in a different aggregation method for every network. These methods are:

- *Majority Voting* (MV) involves classifying each slice individually, where each slice casts a "vote". The final patient prediction is determined by the majority of these votes.

- *Mean* calculates the average predicted probability of the positive class across all slices. If the mean probability surpasses the threshold (0.5), the patient is classified as positive; if not, the classification is negative.
- *Max* identifies the slice with the highest probability across both classes and bases the prediction on this slice alone.

All the combination and ensembles described below were performed in both 2D and 2.5D approaches.

3.5.3. Combination of radiomic and deep features

Combining two different types of features is widely adopted in the state of the art ([2], [45], [36], [20], [7]) . Specifically, this methodology was applied here by integrating DL features with radiomic features. The goal was to enhance the predictive capability of the model and achieve a more comprehensive and accurate data representation, as demonstrated in the analyzed studies. In the feature combination process, the best models found through grid search was considered and each model was combined, corresponding to a different deep network, paired with the same radiomic features. The combination process results in a CSV file with both deep and radiomic features.

Once this combined feature set was obtained, a single validation step was used to select the best classifier among those previously tested. Before classification, the features underwent a selection process where their importance was calculated. Only features with an importance (determined by the Gini importance mechanism of RF) equal to or above 50% of the most important feature were retained. With the selected features and the best classifier chosen, classification was then re-executed on the test set. In the 2.5D approach, the combined features were the same as those obtained in the 2D case, and were therefore filtered accordingly. The classifier applied was the one trained on the 2D feature combination, using the aggregation mode identified as optimal in single network case.

3.5.4. Ensemble

Ensemble of 4 Deep Networks

An ensemble of the best model for every deep network (both for 2D and 2.5D approach) was created aimed at assessing the potential improvement in performance, as the ensemble approach enables leveraging the strengths of each model variant and mitigating their individual weaknesses. These top models provided predictions for each patient, which

were aggregated through a hard voting mechanism, resulting in a final prediction per patient. In hard voting, each model casts a "vote" for a prediction, and the majority vote determines the final output, allowing the ensemble to leverage consensus among the models for increased robustness. Given that ties may occur, in this case the final prediction was assigned as class 1 (symptomatic) for clinical preventive reasons.

Ensemble 4 Deep Networks and Radiomics

Additionally, a five-model ensemble was tested in which the fifth model is the radiomic model. The process remains the same, with predictions subject to hard voting.

4 | Results

4.1. Feature selection and dimensionality reduction

Given that the number of significant features was quite high compared to the small sample size available, few steps of preliminary dimensionality reduction have been performed. This section details the steps along with the resulting dimensionalities at each stage.

Type of Features	Initial # of Features	Pearson Correlation	ANOVA F-test p-value
Radiomics	474	111	39

Table 4.1: Dimensionality reductions results for radiomics

From the removal of features with high Pearson correlation (threshold of 0.9) , 111 features remain, of which only 100 are wavelet features, with approximately 25 features from each band. Among the wavelet features, 14 are first-order wavelet features, 6 are shape-based wavelet features, and 91 are texture-based wavelet features, including 25 from GLCM, 10 from GLDM, 11 from GLRLM, 25 from GLSZM, and 13 from NGTDM. The remaining 11 features are original, all texture-based, with 9 from GLCM and 2 from NGTDM.

Then, setting a threshold of 0.05 for the ANOVA F-test p-value, 39 features are retained, only one of which is original. Between the 38 wavelet features, there is a balance of 9 or 10 features for each band. Only one feature is first order, all the others are texture-based (25 of GLCM 25, 10 of GLDM, 11 of GLRLM, 25 of GLSZM and 13 of NGTDM).

The names of the features just mentioned are listed in *Table A.1* at Appendix A.

Type of Network	Initial # of Features	Pearson Correlation	ANOVA F-test p-value
VGG19	512	512	6
ResNet50	2048	915	142
InceptionV3	2048	1225	191
InceptionResNetV2	1536	852	162

Table 4.2: Dimensionality reductions results for cropped image

Type of Network	Initial # of Features	Pearson Correlation	ANOVA F-test p-value
VGG19	512	483	67
ResNet50	2048	1862	268
InceptionV3	2048	1600	380
InceptionResNetV2	1536	1151	158

Table 4.3: Dimensionality reductions results for full image

For the cropped image, VGG shows no reduction in features for Pearson Correlation, while ResNet and Inception demonstrate significant reductions of 55.32% and 40.29%, respectively, and InceptionResNet shows a 44.61% reduction. Regarding the p-value, VGG experiences a substantial reduction of 98.83%, indicating that most features are not statistically significant, followed by ResNet (93.08%), Inception (90.68%), and InceptionResNet (89.45%).

For the full image, VGG has a small reduction of 5.66% with Pearson Correlation, while ResNet shows a slightly higher reduction of 9.07%. Inception and InceptionResNet exhibit more substantial reductions of 21.85% and 25.0%, respectively, with InceptionResNet reducing the most. For the p-value, VGG again shows a large reduction of 86.91%, similar to ResNet's 87.02%, with Inception (81.43%) and InceptionResNet (89.68%) having high reductions, the latter showing the greatest decrease in the p-value.

The Pearson Correlation threshold, set at 0.9, demonstrates that all networks have a significant number of features with high correlation, as shown by the substantial reduction in features across all models, especially for ResNet, Inception, and InceptionResNet. Differently VGG retains all of its features in the cropped image case and almost all in the full image. On the other hand, the p-value threshold, set at 0.05, removes many features, with large reductions observed in all networks. This suggests that a significant portion of features are not statistically significant, especially for networks like VGG and ResNet, which exhibit the highest reductions. In contrast, InceptionResNet shows the most dras-

tic feature reductions for both correlation and p-value, indicating that it eliminates the biggest number of features due to both low correlation and lack of statistical significance.

4.2. 2D Classification results

Heatmaps of the average balanced accuracy on the validations set obtained by CV are shown. For each combination of selector (x-axis) and classifier (y-axis) the best result is provided. The one showing the highest mean is the one applied to the test set, in case of two combinations with very similar mean, the one with the lowest standard deviation is chosen.

4.2.1. Radiomics

The chosen combination is SVM - mRMR showing a balanced accuracy across the CV folds of 0.732 ± 0.120, ROC AUC 0.726 ± 0.135 and F1 score of 0.715 ± 0.118.

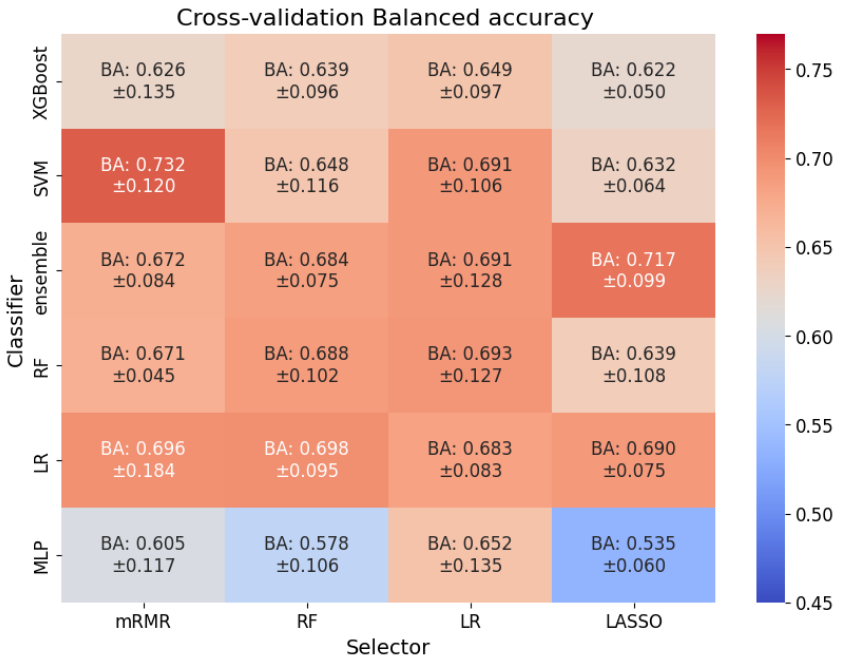


Figure 4.1: Validation results for radiomics in 2D approach

SVM stands out as the best classifier, consistently outperforming the others across all metrics. It achieves the highest balanced accuracy (0.732), ROC AUC (0.726), and F1 score (0.715). RF also performs well, with solid results across all metrics. On the other hand, XGBoost appears to be the weakest model, with the lowest values across all metrics (balanced accuracy: 0.649, ROC AUC: 0.656, F1 score: 0.608).

Table 4.4: Test metrics on the best model for radiomics fo the 2D approach

Classifier	Selector	Features number	ROC AUC	Balanced accuracy	F1
SVM	mRMR (4)	4	0.620	0.532	0.300

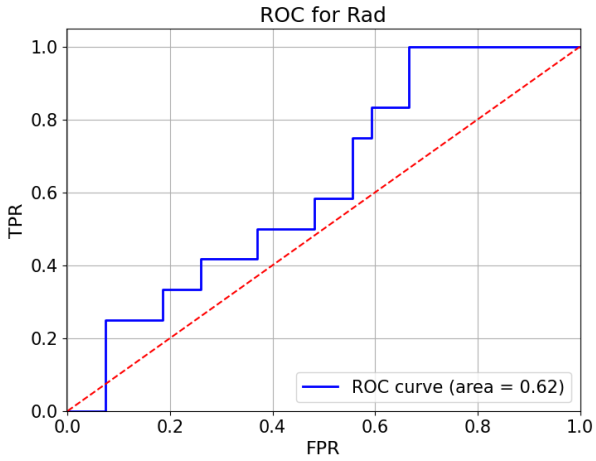


Figure 4.2: ROC curve test for radiomics

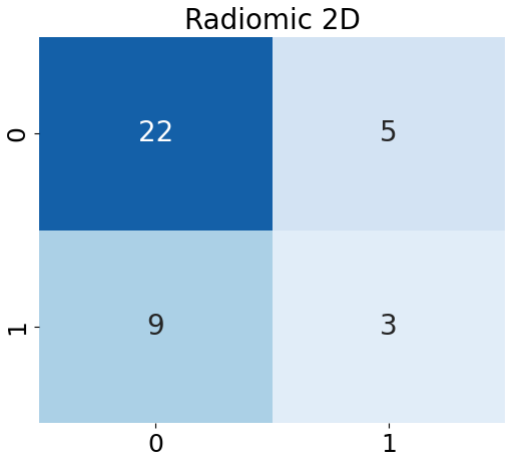


Figure 4.3: Confusion matrix of test set for radiomics

Table 4.4 shows the performance metrics on the test set while *Figure 4.2* and *Figure 4.3* show the ROC curve and the confusion matrix respectively. All the analyzed metrics worsen, possibly due to different data distributions or the lack of generalization of the chosen model. All the chosen features are texture-based and 3 of the 4 are of HL band, these are:

- wavelet-HL_glcm_ldn
- wavelet-HL_glrlm_GrayLevelNonUniformity
- wavelet-HL_gldm_DependenceVariance
- wavelet-LH_gldm_DependenceVariance

4.2.2. Deep learning features from cropped slices

In this section a comprehensive evaluation of different ML classifiers applied to various neural network architectures for cropped images is shown. The analysis begins with a series of heatmaps, each illustrating the mean balanced accuracy and its standard deviation for combinations of classifiers and feature selectors (on the best *num_features* found) within a given network. These visualizations offer insights into the variability and robustness of each classifier-selector combination, helping to identify configurations that perform consistently well. Following the heatmaps, three detailed tables provide quantitative comparisons of the top-performing classifiers for each network across three performance metrics: balanced accuracy, ROC AUC, and F1 score. Finally, a set of histograms summarizes these metrics, comparing the best classifier for each neural network in order to select the best network across the four.

From *Figure 4.4*, it can be seen that the chosen model for VGG consists in LR classifier and LASSO selector. *Figure 4.5* highlights for ResNet that LR classifier with mRMR selector provides the highest balanced accuracy. *Figure 4.6* then shows that for Inception the best performing model is LR with RF selector and finally *Figure 4.7* demonstrates that the best performing model for InceptionResnet is RF with LR as selector. Visually, it is clear that VGG is the worst efficient network while InceptionResnet's classifiers, followed by Inception's, are the ones with the highest balanced accuracy.

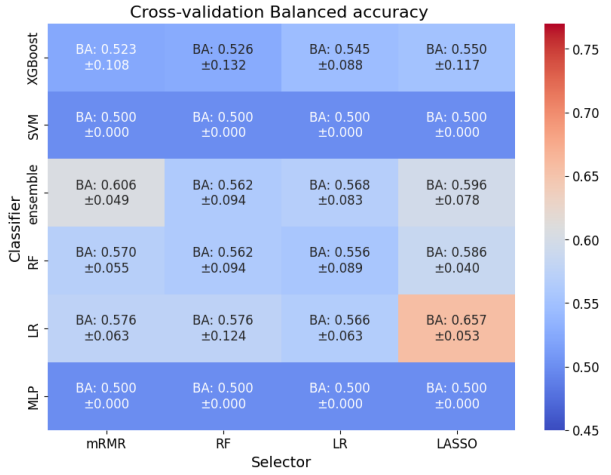


Figure 4.4: Validation results VGG in 2D (Cropped)

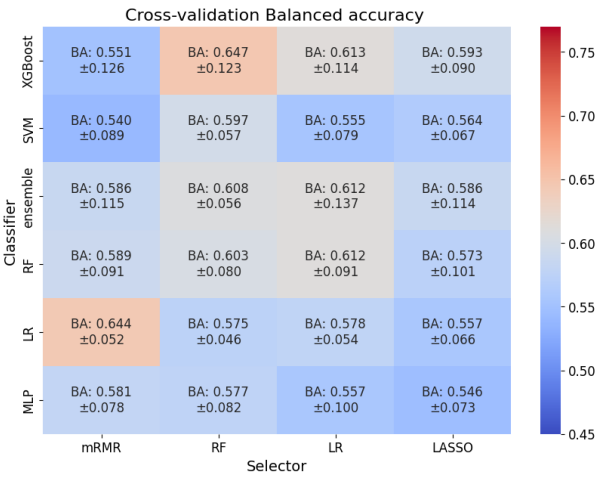


Figure 4.5: Validation results Res in 2D (Cropped)

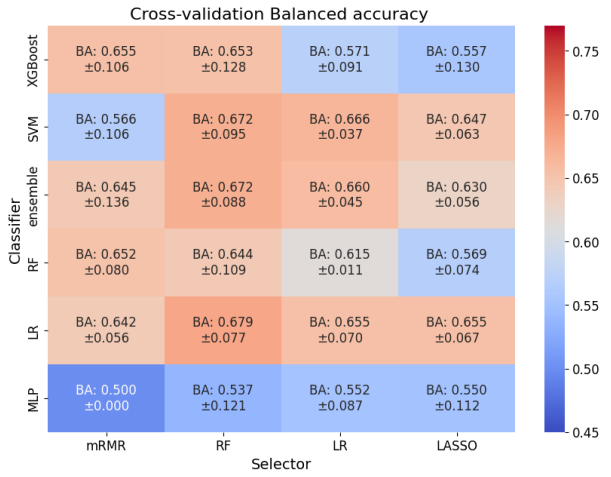


Figure 4.6: Validation results Inc in 2D (Cropped)

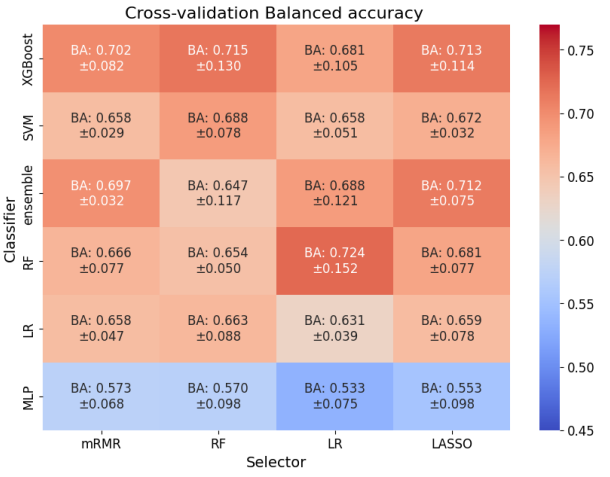


Figure 4.7: Validation results IncRes in 2D (Cropped)

Comparing the general performance of ML classifiers on all the networks, MLP consistently underperforms compared to other models, particularly on InceptionResnet. RF demonstrates the strongest overall performance, with high scores in balanced accuracy, ROC AUC, and F1, consistently excelling across all datasets.

Finally, *Figure 4.8* present an histogram that summarizes the mean and standard deviation scores of the three metrics across the best-performing classifiers for each neural network. The diagram provides an aggregated view of each network's strengths across the three metrics, with InceptionResNet standing out due to its superiority for all the mean metrics, reinforcing its overall performance advantage. Notably, ResNet shows relatively low variability in all metrics, suggesting it may offer reliable performance, but it was not

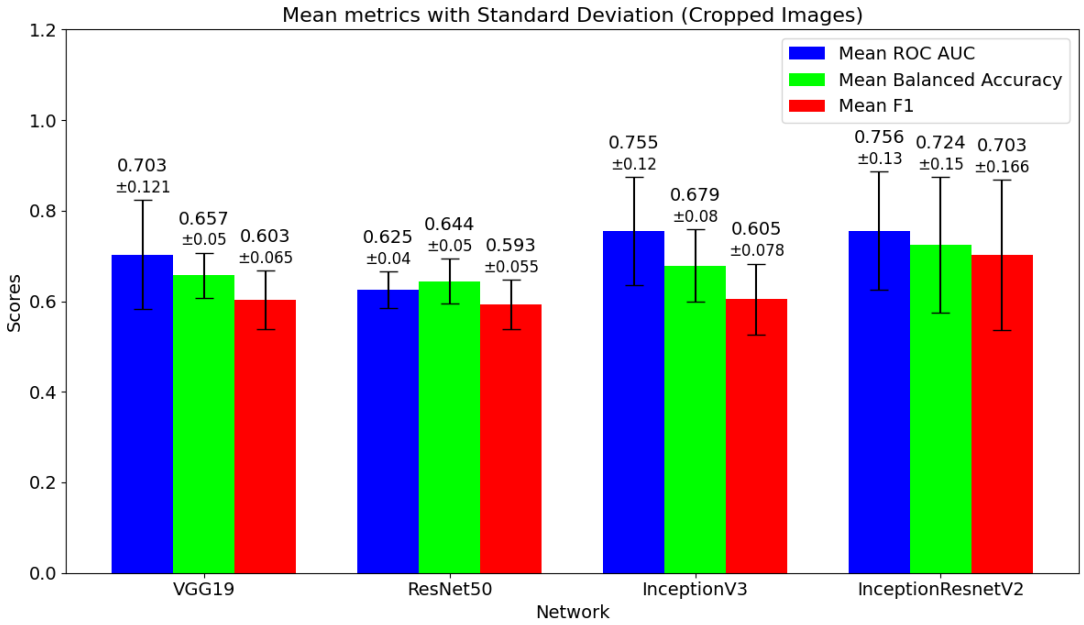


Figure 4.8: 2D mean and std diagram (Cropped slice)

chosen due to the low performances.

InceptionResNet has been selected as the optimal network, with the RF and LR as best classifier-selector combination. Thus, this model was applied to the test set. The final configuration used 26 features, with 19 of these being consistently selected across at least three folds, ensuring that only the most relevant and stable features contributed to the final predictions. Test performances on the other networks are reported in *Table B.1 Appendix B.*

Table 4.5: 2D test metrics on the best model for the best network for cropped slices

Network	Classifier	Selector	Features number	ROC AUC	Balanced accuracy	F1
IncRes	RF	LR (26)	19	0.807	0.787	0.690

The ROC curve illustrates high discrimination capability, reflecting the model's accuracy in distinguishing between classes. Additionally, the confusion matrix provides further insights, highlighting the model's ability to correctly classify both positive and negative cases.

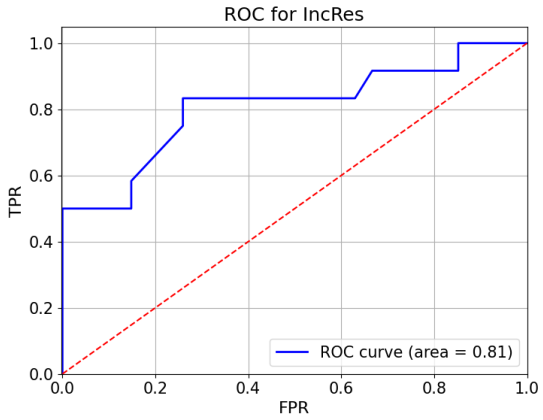


Figure 4.9: 2D ROC curve test for IncRes (Cropped slice)

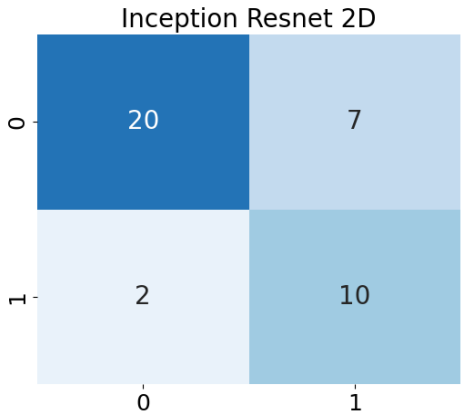


Figure 4.10: 2D Confusion Matrix on test set for IncRes (Cropped slice)

4.2.3. Deep learning features from full slices

In this section, the evaluation of different classifier - selector combinations on full slices is presented. This analysis follows the same structured approach, with heatmaps, detailed tables, and histogram comparisons, to examine the balanced accuracy, ROC AUC, and F1 scores across neural network architectures.

Figure 4.11, shows that, for the VGG network, the RF classifier clearly stands out as the best classifier, achieving high balanced accuracy with both the mRMR and RF selectors. Even though the std for mRMR is lower than for RF, RF is higher both in F1 score (0.718 vs 0.709) and in ROC AUC (0.746 vs 0.712). *Figure 4.12* shows that the optimal configuration for ResNet is the SVM classifier paired with the LASSO selector. *Figure 4.13* indicates that for Inception, the highest balanced accuracy is achieved with LR selector and with both XGBoost and RF as classifier (0.739 and 0.726). RF was chosen since it has lower std but also higher ROC AUC (0.770 vs 0.751) and higher F1 score (0.716 vs 0.703). Lastly, *Figure 4.14* demonstrates that for InceptionResNet, XGBoost with LR selector, emerges as the most efficient configuration. Observing the overall heatmaps colours, it is evident that VGG and Inception exhibit stronger generalization capabilities compared to the other two networks.

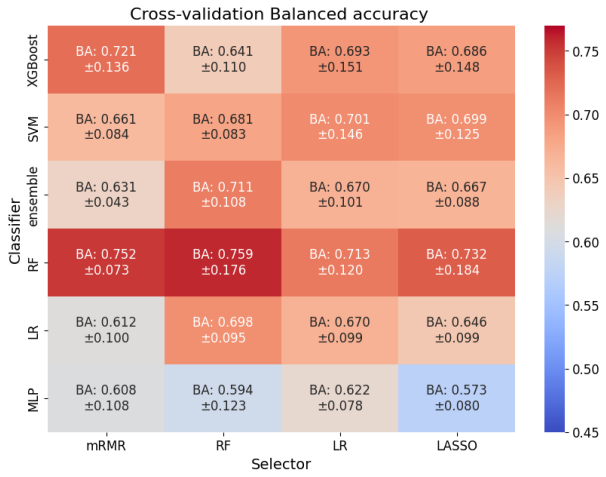


Figure 4.11: Validation results VGG in 2D (Full)

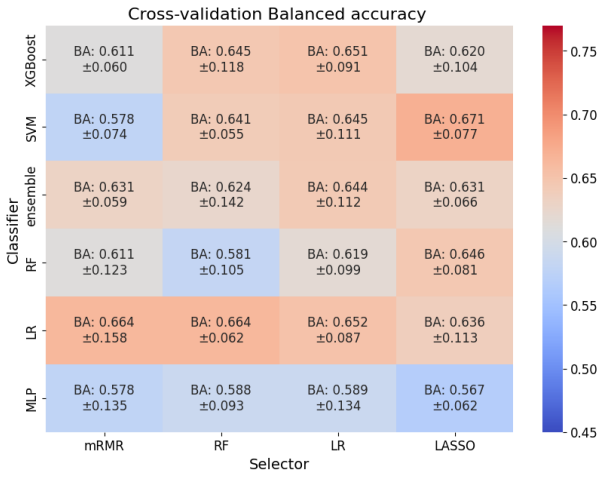


Figure 4.12: Validation results Res in 2D (Full)

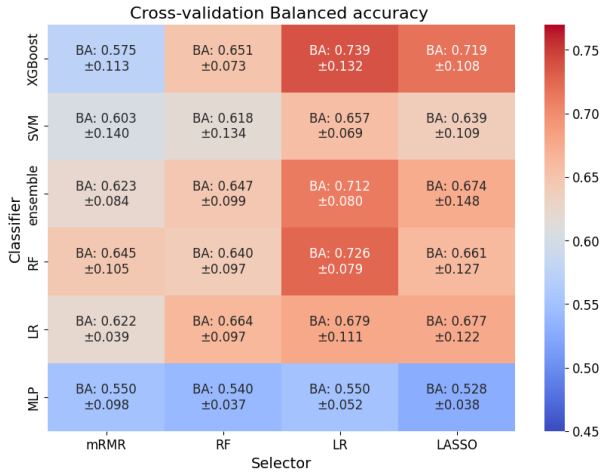


Figure 4.13: Validation results Inc in 2D (Full)

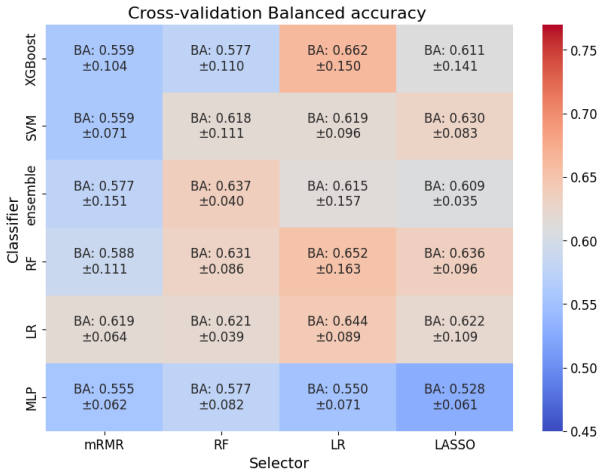


Figure 4.14: Validation results IncRes in 2D (Full)

Generally across all the networks it is evident that RF is the most consistent model, performing well across all metrics. MLP, followed by ensemble and LR, generally show the weakest performance.

Finally *Figure 4.15* provides a visual comparison of the top-performing models across the four networks. Since balanced accuracy was chosen to be the primary evaluation metric, VGG, with RF as classifier and selector, was selected as the preferred configuration for the full slice approach.

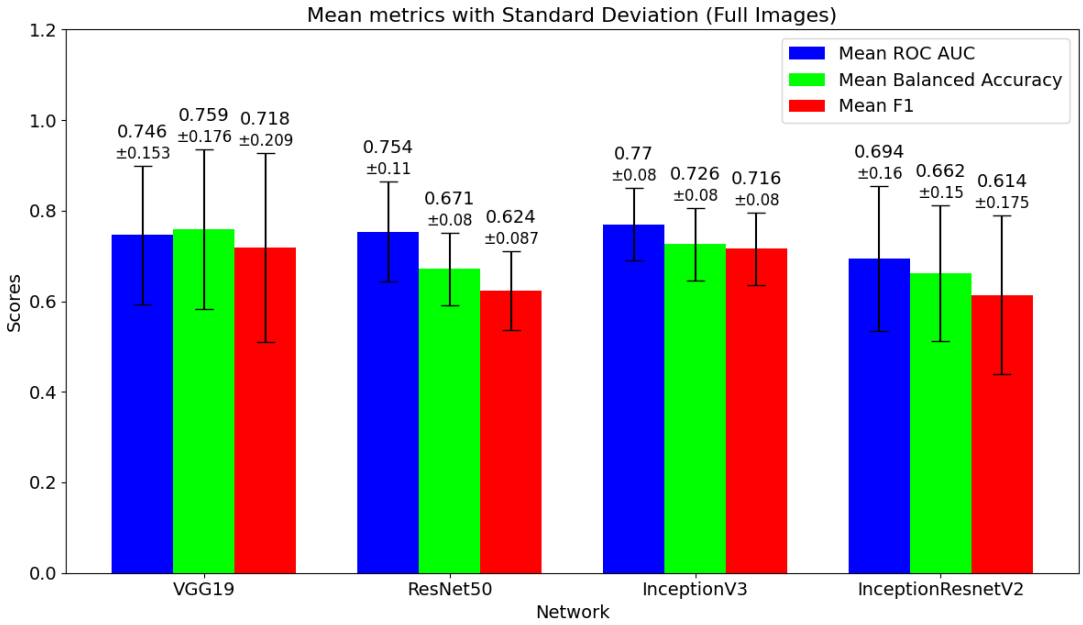


Figure 4.15: 2D mean and std diagram diagram (Full slice)

Table 4.6: 2D test metrics on the best model for the best network for full slices

Network	Classifier	Selector	Features number	ROC AUC	Balanced accuracy	F1
VGG	RF	RF (23)	20	0.877	0.824	0.741

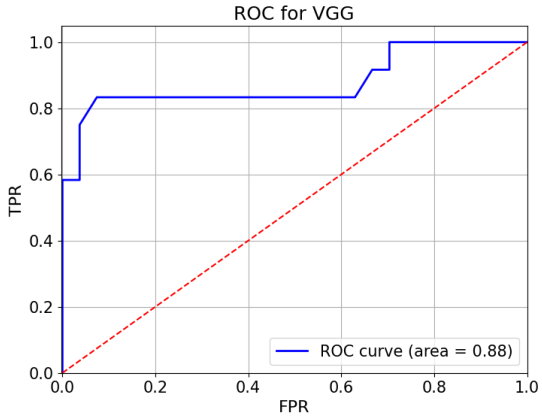


Figure 4.16: 2D ROC curve test for VGG (Full slice)

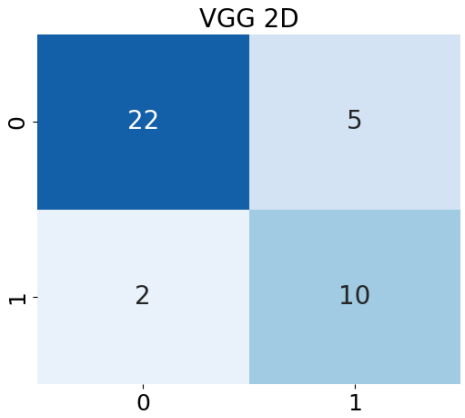


Figure 4.17: 2D Confusion matrix on test set for VGG (Full slice)

Table B.2 reports the performance of the final models for each network.

4.2.4. Combination Radiomic and Deep features

This section presents the results obtained by combining the optimal features identified by the radiomic model with those from the optimal models in both the cropped and full slice cases. Initially, a histogram illustrates the selected features (both deep and radiomic), reduced from the initial number by looking at their importance (as explained in *Section 3.5.3*) that are used for classification. A final validation step is conducted on a single fold to determine the new best-performing classifier on the selected features. *Tables 4.7* and *Tables 4.8* show the performances of the classifiers on the validation set. The ones with higher metrics are used to classify the combined features.

Table 4.7: Validation metrics for classifiers in the combined approach for IncRes

Classifier	**Balanced accuracy**	**ROC AUC**	**F1**
XGBoost	0.742	0.811	0.720
SVM	0.508	0.650	0.435
Ensemble	0.625	0.728	0.583
RF	0.583	0.703	0.522
LR	0.583	0.683	0.522
MLP	0.500	0.661	0

Table 4.8: Validation metrics for classifiers in the combined approach for VGG

Classifier	**Balanced accuracy**	**ROC AUC**	**F1**
XGBoost	0.658	0.728	0.609
SVM	0.683	0.739	0.600
Ensemble	0.800	0.761	0.762
RF	0.683	0.803	0.600
LR	0.733	0.750	0.696
MLP	0.583	0.744	0.286

In the histograms, the radiomic features are labeled with their original names, while the remaining features represent the deep features extracted by InceptionResnet in *Figure 4.18* and by VGG in *Figure 4.19*. The features whose columns are above the red dashed line are used for the classification.

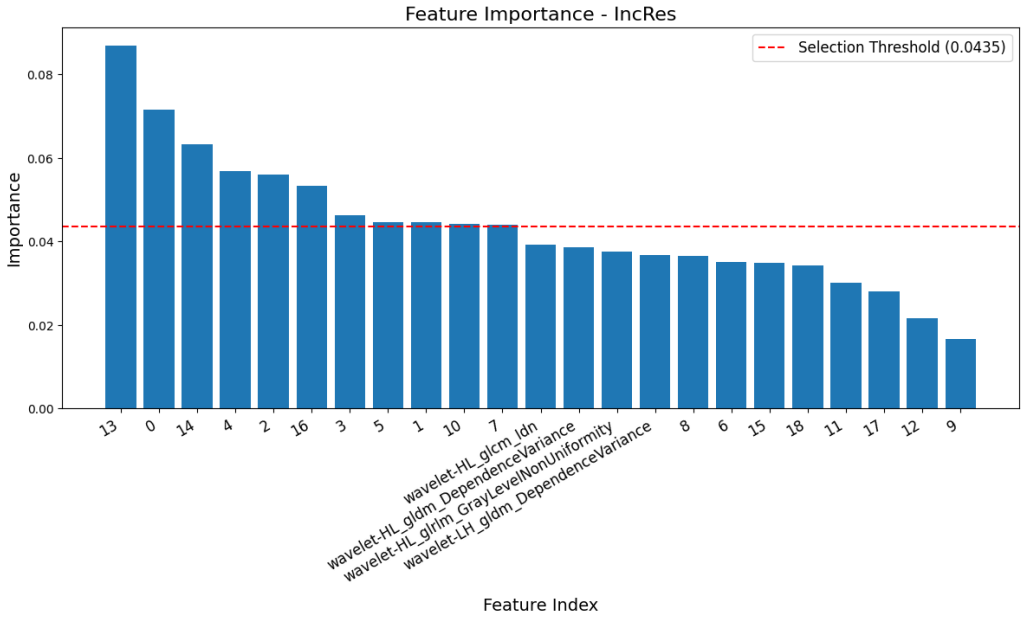


Figure 4.18: Importance diagram for IncRes (Cropped slice)

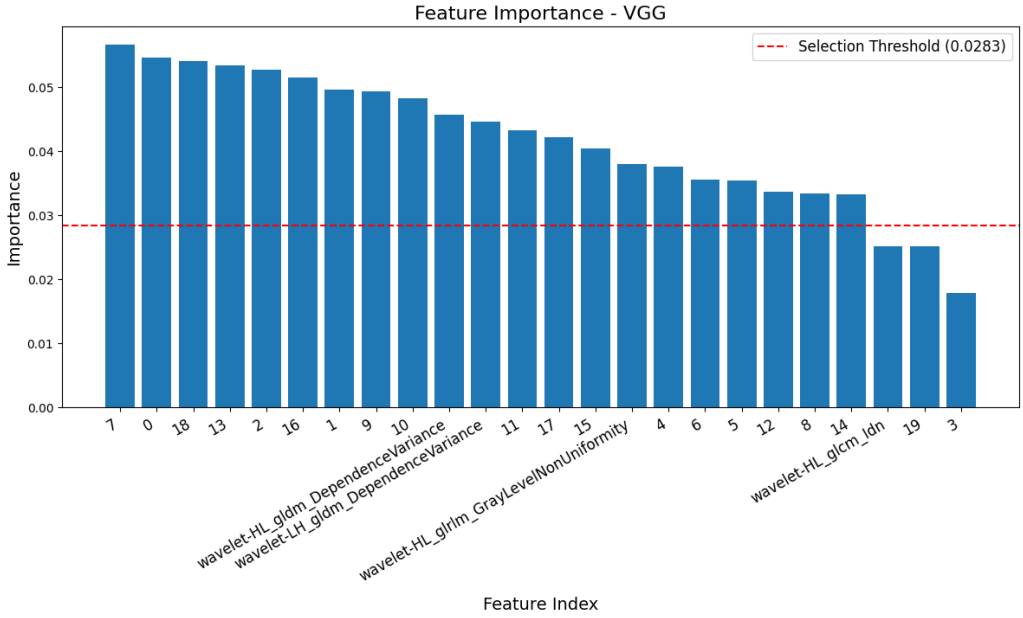


Figure 4.19: Importance diagram for VGG (Full slice)

It can be seen from *Figure 4.18* that no radiomic feature is kept by the importance feature selection for the InceptionResnet model. On the other side VGG (*Figure 4.19*) only discards one of the four features.

Finally, the test set metrics are presented for this combined feature approach while metrics for all the networks are in *Table B.3*.

Table 4.9: 2D test metrics on combination

Image Type	Classifier	Network	Features number	ROC AUC	Balanced accuracy	F1
Cropped	XGBoost	INCRES	11	0.698	0.583	0.444
Full	ensemble	VGG	21	0.907	0.787	0.690

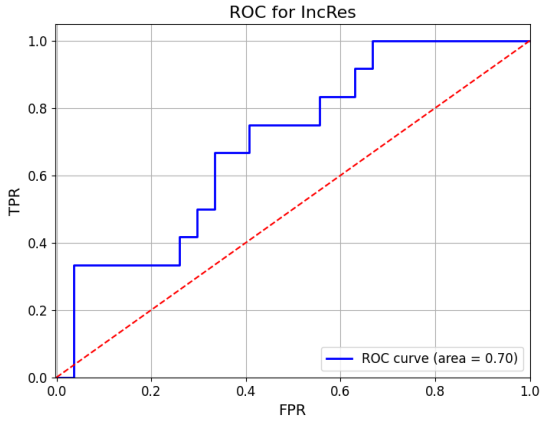


Figure 4.20: 2D ROC curve for IncRes (Cropped slice)

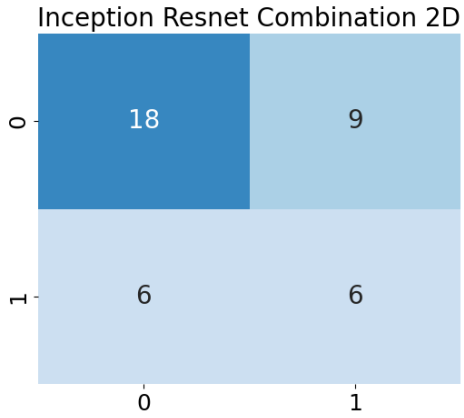


Figure 4.21: 2D Confusion matrix on test set for IncRes (Cropped slice)

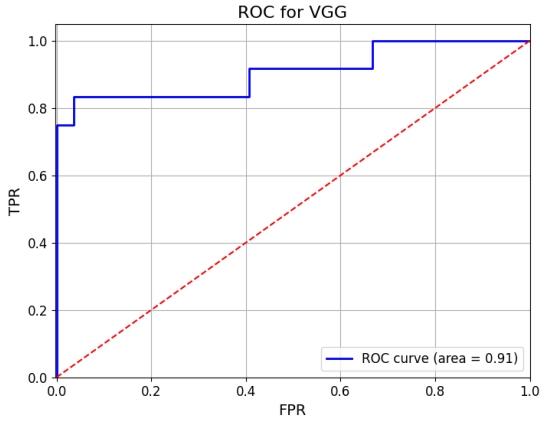


Figure 4.22: 2D ROC curve for VGG (Full slice)

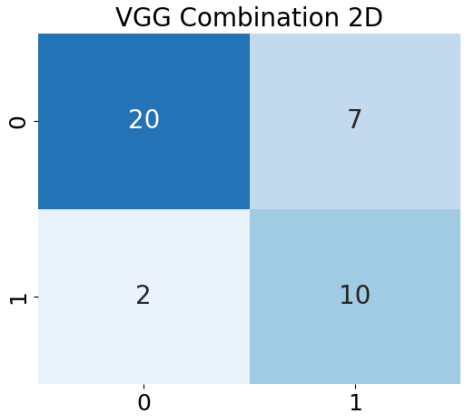


Figure 4.23: 2D Confusion matrix on test set for VGG (Full slice)

4.2.5. Ensemble

The top-performing models from each of the four neural networks are combined into an ensemble to provide a refined, final classification. Additionally, an extended ensemble of

five models is evaluated, incorporating the radiomic model to further enhance classification performance. *Table 4.18* summarizes the test set metrics for both ensemble configurations, along with the respective confusion matrices.

Ensemble Type	Image Type	Balanced accuracy	F1-score	Confusion Matrix
4 networks	Cropped	0.856	0.818	*Figure 4.24*
4 networks + radiomics	Cropped	0.708	0.588	*Figure 4.26*
4 networks	Full	0.727	0.621	*Figure 4.25*
4 networks + radiomics	Full	0.796	0.727	*Figure 4.27*

Table 4.10: Ensemble performance metrics

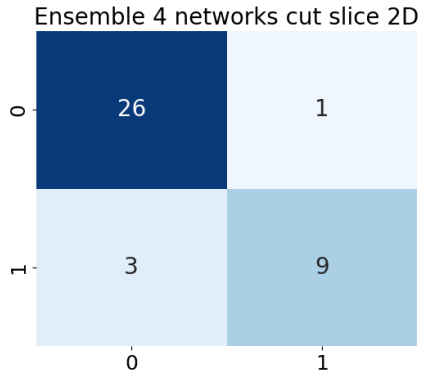


Figure 4.24: 2D Confusion matrix for 4 networks ensemble (Cropped slice)

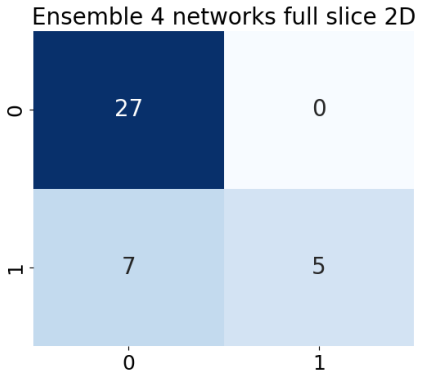


Figure 4.25: 2D Confusion matrix for 4 networks ensemble (Full slice)

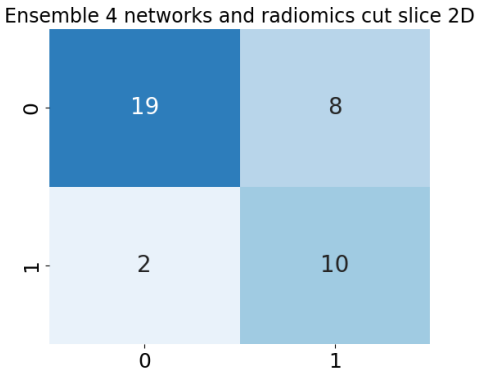


Figure 4.26: 2D Confusion matrix for networks and radiomics ensemble (Cropped slice)

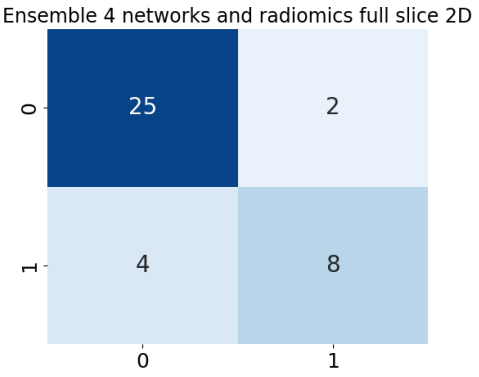


Figure 4.27: 2D Confusion matrix for networks and radiomics ensemble (Full slice)

4.3. 2.5D Classification results

The 2.5D approach builds on the 2D model by applying the same trained classifier to multiple slices that have an area of at least 30% of the largest plaque area. This method expands the dataset meaningfully without requiring additional feature selection or classifier-selector optimization. Considering that multiple predictions are obtained for each patient, three methods (*Majority Voting, Mean* and *Max*) are tested to combine them into a single prediction (as explained in *Section 3.5.2*). To determine the most appropriate method to apply to the test set, the different methods are evaluated on a validation set consisting of 27 patients (30% of the original train set) and balanced accuracy is used as main evaluation metric. A different mode is found for every network and for radiomics and it is reported in the tables presenting the test metrics.
Results are reported for radiomics and for the best networks found in the 2D analysis for the two types of images. Results for all the other networks can be found in *Appendix B*.

4.3.1. Radiomics

Table 4.11 shows the validation metrics obtained on a single fold in order to choose the prediction mode for the radiomic, while *Table 4.12* shows the metrics on the test set using the trained model of the 2D approach with the prediction mode chosen. The ROC AUC metric is not shown since it's equal in the 3 cases and thus not useful for the choice of the best prediction mode.

Table 4.11: Validation metrics for mode prediction in the radiomic case

Mode	Balanced accuracy	F1
MV	0.467	0
Mean	0.525	0.333
Max	0.467	0

Table 4.12: Test metrics on radiomic features 2.5

Mode	Features number	ROC AUC	Balanced accuracy	F1
Mean	4	0.654	0.542	0.154

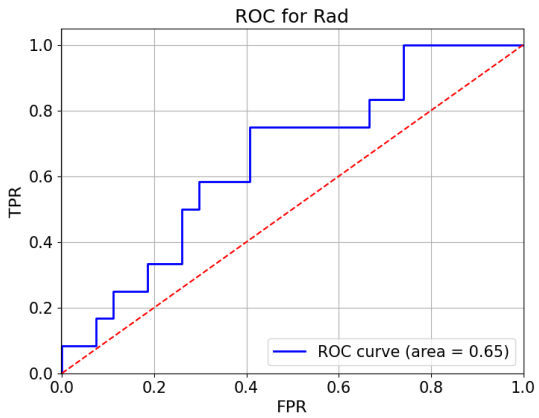


Figure 4.28: 2.5D ROC curve test for radiomics

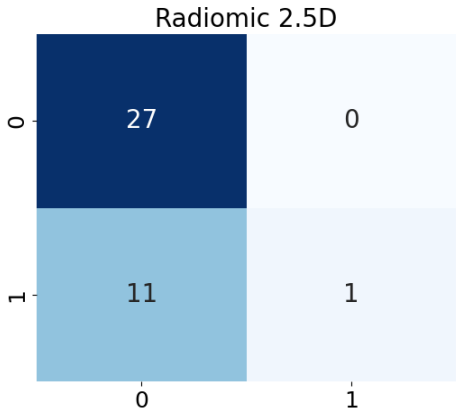


Figure 4.29: 2.5D Confusion matrix of test set for radiomics

4.3.2. Deep learning features from Cropped slices

Table 4.13: Validation metrics for mode prediction in the IncRes case for Cropped slices

Mode	Balanced accuracy	F1
MV	0.683	0.6
Mean	0.767	0.727
Max	0.883	0.869

Table 4.14: 2.5D test metrics on best deep network on Cropped slices

Network	Mode	Features number	ROC AUC	Balanced accuracy	F1
Incres	Max	19	0.809	0.745	0.643

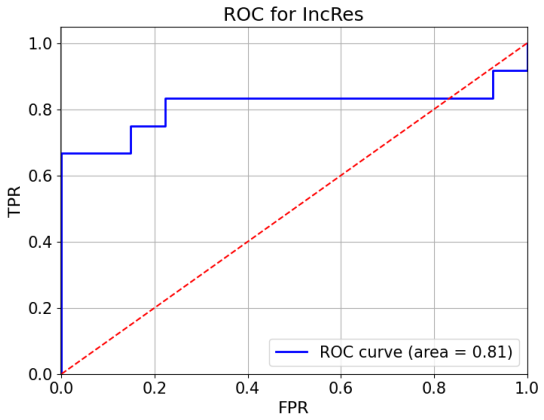


Figure 4.30: 2.5D ROC curve test for IncRes (Cropped slices)

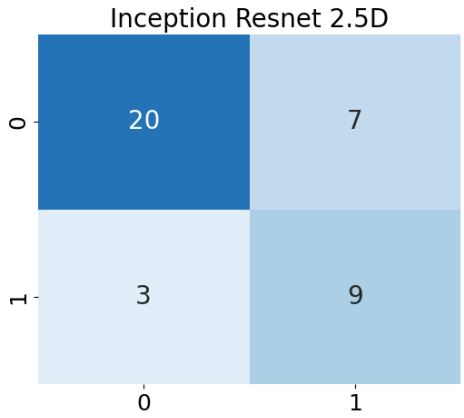


Figure 4.31: 2.5D Confusion Matrix on test set for IncRes (Cropped slices)

4.3.3. Deep learning features from full slices

Table 4.15: Validation metrics for the prediction mode in the VGG case for full slices

Mode	Balanced accuracy	F1
MV	0.8	0.762
Mean	0.8	0.762
Max	0.842	0.818

Table 4.16: 2.5D Test metrics on best model for the best network on full slices

Network	Mode	Features number	ROC AUC	Balanced accuracy	F1
VGG	Max	20	0.849	0.880	0.833

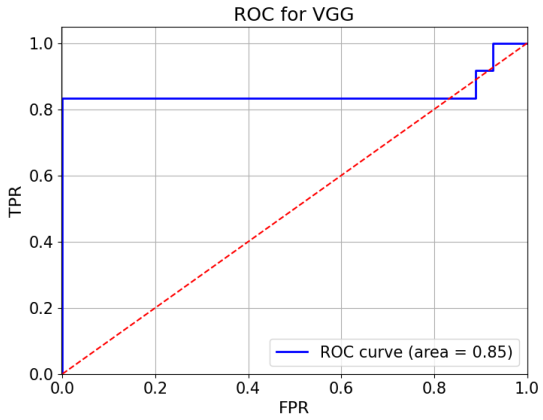


Figure 4.32: 2.5D ROC curve test for VGG (Full slices)

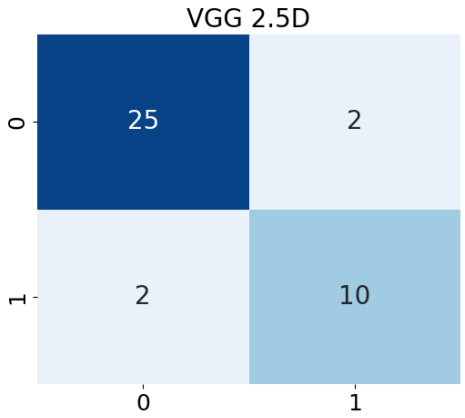


Figure 4.33: 2.5D Confusion Matrix on test set for VGG (Full slices)

4.3.4. Combination Radiomic and Deep features

This section explores the impact of combining radiomic and deep features in the 2.5D setup. The best combined model and the features obtained in the 2D analysis are employed with this multi-slice approach. The prediction mode is the one chosen by the single networks (as displayed in *Tables 4.14 and 4.16*).

Table 4.17: 2.5D test metrics on combinations

Image type	Network	Mode	Features number	AUC	Balanced accuracy	F1
Cropped	IncRes	Max	11	0.741	0.616	0.514
Full	VGG	Max	21	0.889	0.898	0.870

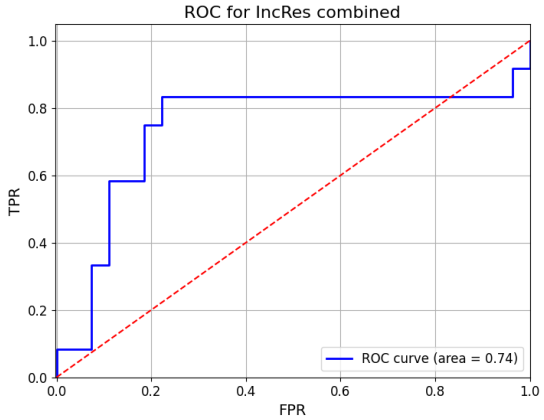


Figure 4.34: 2.5D ROC curve for IncRes combined with radiomics (Cropped slices)

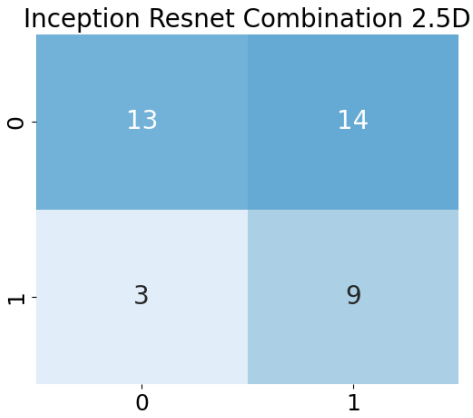


Figure 4.35: 2.5D Confusion matrix on test set for IncRes combined with radiomics (Cropped slices)

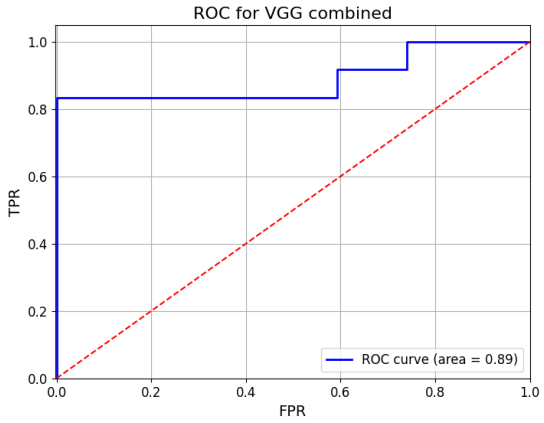


Figure 4.36: 2.5D ROC curve for VGG combined with radiomics (Full slices)

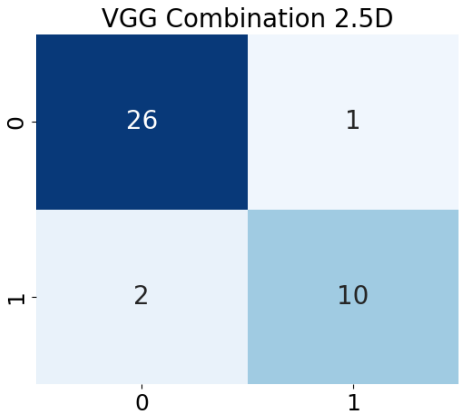


Figure 4.37: 2.5D Confusion matrix on test set for VGG combined with radiomics (Full slices)

4.3.5. Ensemble

Ensemble Type	Image Type	Balanced accuracy	F1	Confusion Matrix
4 networks	Cropped	0.727	0.621	*Figure 4.24*
4 networks + radiomics	Cropped	0.824	0.741	*Figure 4.26*
4 networks	Full	0.838	0.783	*Figure 4.25*
4 networks + radiomics	Full	0.856	0.818	*Figure 4.27*

Table 4.18: Ensemble performance metrics

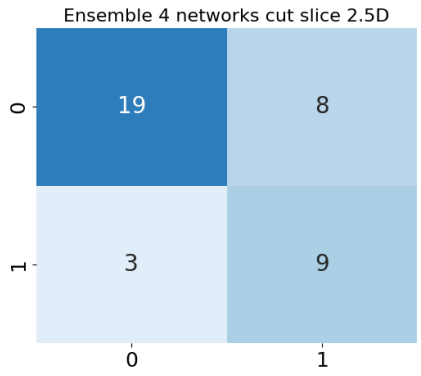


Figure 4.38: 2.5D Confusion matrix on test set for 4 networks ensemble (Cropped Slices)

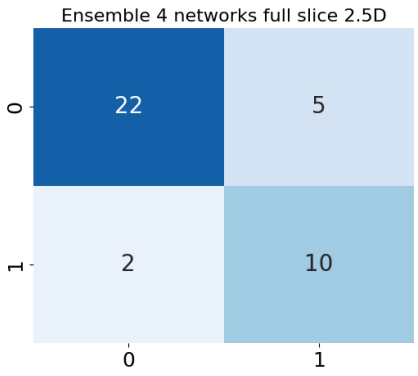


Figure 4.39: 2.5D Confusion matrix on test set for 4 networks ensemble (Full Slices)

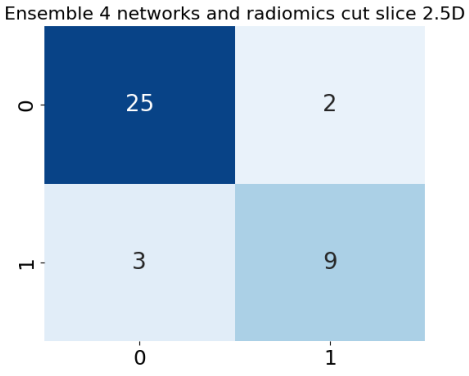


Figure 4.40: 2.5D Confusion matrix on test set for networks + radiomics ensemble (Cropped Slices)

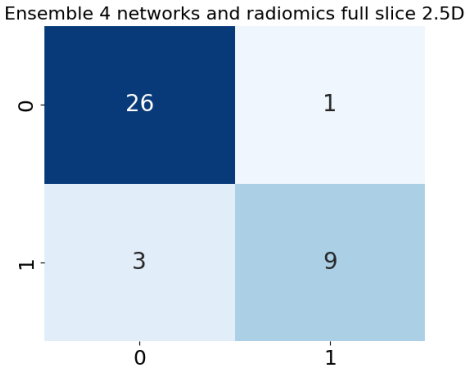


Figure 4.41: 2.5D Confusion matrix on test set for networks + radiomics ensemble (Full Slices)

5 | Discussion, limitations and future developments

An extensive analysis was conducted in this study to classify carotid plaques, leveraging both radiomic and deep learning-based approaches. The investigation evaluated various models, including 2D and 2.5D methodologies, combinations of radiomic and deep features, and ensemble techniques. This comprehensive approach allowed for the determination of the most effective strategy for distinguishing between symptomatic and asymptomatic plaques. This chapter provides a comprehensive overview of the findings of this study and their implications. The first section discusses and compares the results obtained, highlighting the advantages of our proposed approach. Then a critical evaluation of the developed method relatively to the state-of-the-art is provided. Finally, current limitations of the study are outlined and potential directions for future research are proposed.

5.1. Results discussion

The analysis begins with an in-depth exploration of the 2D approach, considering only the slice with the largest plaque area. From the preliminary feature selection in the radiomics analysis, it is evident that wavelet features are more significant than original ones, highlighting the efficiency of applying filters prior to feature extraction. During validation, the radiomics model achieved promising results, with a mean balanced accuracy of 0.732 $\pm$ 0.120. After feature selection, four features were retained, three of which belonged to the HL wavelet band and were texture-based. However, the performance of the model on the test set dropped, with a balanced accuracy of 0.532, with 5 asymptomatic and 9 symptomatic patients being misclassified. Such inaccuracies carry serious clinical implications, emphasizing the model's limitations in a real-world setting. The discrepancy between validation and test results might be attributed to differences in data distribution, the limited size of the dataset, or the presence of patterns in the test set that were not captured during training.

The 2D deep analysis of cropped images identified InceptionResNet as the best-performing network, achieving a balanced accuracy of 0.724 ± 0.077, a ROC AUC of 0.756 ± 0.131, and an F1 score of 0.703 ± 0.166 across CV folds. On the test set, the network outperformed radiomics, with a balanced accuracy of 0.787, a ROC AUC of 0.807, and an F1 score of 0.690. Despite this improvement, the model misclassified 7 asymptomatic and 2 symptomatic patients. While these errors are fewer than those of the radiomics model, they still have clinical relevance. Moving to full slices, the VGG network demonstrated stronger performance, achieving a balanced accuracy of 0.759 ± 0.176, a ROC AUC of 0.746 ± 0.153, and an F1 score of 0.718 ± 0.209 during validation, surpassing all the cropped image metrics except ROC AUC. On the test set, the VGG network reached a balanced accuracy of 0.824, a ROC AUC of 0.877, and an F1 score of 0.741. This model reduced FPs to 5 compared to the 7 of InceptionResNet on cropped images, while also maintaining only 2 FNs.

When combining radiomic features with deep learning models, no substantial improvements were observed. For full images, the combination with VGG showed a slight increase in ROC AUC from 0.877 to 0.907, but this came at the cost of a decrease in balanced accuracy due to an increase in FPs (7 cases). On the other hand, combining all the deep learning networks for cropped images led to a significant improvement, achieving a balanced accuracy of 0.856 on the test set. This enhancement likely comes from the complementary strengths of InceptionResNet and ResNet, with the latter achieving a balanced accuracy of 0.819 on the test set. However, adding radiomics to this ensemble caused a sharp decline, reducing balanced accuracy to 0.708. For full images, combining the four networks decreased performance compared to the single VGG model. Surprisingly, while ResNet alone achieved a balanced accuracy of 0.843, the ensemble dropped to 0.727. Interestingly, adding radiomics to this ensemble improved balanced accuracy to 0.796, with 2 FPs and 4 FNs.

Across all these results, full images resulted in better classification metrics. The only exception to this is the ensemble of the four networks on cropped images, which achieved a balanced accuracy of 0.856 with 3 FNs and 1 FP. This number of FPs can be explained by the unbalanced predictions of the single networks; in fact two out of these four networks predict almost, if not all, patients as 0 (see *Table B.1*). Additionally, the fact that cropped images underperform full ones in all the other cases leads the decision of VGG on full images as the best 2D model. This network achieved a balanced accuracy of 0.824 and has 2 FNs and 5 FPs. This network, when combined with radiomics, achieved a ROC AUC of 0.907 compared to 0.877, but at the expense of a lower balanced accuracy of 0.787.

The 2.5D approach yielded negligible improvements in radiomics with the balanced accu-

racy increasing slightly from 0.532 of the 2D to 0.542 of the 2.5D, ROC AUC from 0.620 to 0.654 and the F1 score dropping from 0.300 to 0.154 respectively. Furthermore, the model misclassified 11 symptomatic patients while correctly classifying all asymptomatic patients. Differently, a clear trend did not emerge when applying the 2.5D approach to the best-performing deep learning models. For InceptionResNet on cropped images, the balanced accuracy decreased from 0.787 to 0.745, and the F1 score dropped from 0.690 to 0.643. While the ROC AUC remained nearly unchanged (0.807 vs. 0.809), the model misclassified one additional symptomatic patient. Conversely, the 2.5D approach improved the VGG model's performance on full slices. The balanced accuracy increased from 0.824 to 0.880, with the number of FPs decreasing from 5 to 2. However, the ROC AUC dropped slightly from 0.877 to 0.849, while FNs remained at 2 cases.

Combining radiomics with the 2.5D VGG model proved to be the most effective strategy. This combination improved the ROC AUC from 0.849 of VGG and 0.654 of radiomics to 0.889, the F1 score from 0.833 and 0.542 respectively to 0.870, and the balanced accuracy from 0.880 and 0.154 to 0.898. This result, achieved with 21 features (including 3 radiomic ones), represents the best overall performance observed in this study. For cropped images, combining all deep learning networks with radiomics improved performance, resulting in a balanced accuracy of 0.824. However, for full slices, ensembles did not outperform the single VGG model, despite the high performance of individual networks such as ResNet (0.866 of balanced accuracy) and InceptionResNet (0.806 of balanced accuracy).

The 2.5D analysis showed that full images are more effective in terms of extracting significant features for the performed binary classification. For these images, the results show that both individual networks, combinations, and ensembles benefit from the multi-slice approach. By selecting among the various predictions generated by each slice of the patient, especially using the approach based on the maximum predicted probability (*Max* approach), some improvements are achieved. In contrast, for cropped images, while there is some improvement, it is not as pronounced. Therefore, a clear enhancement cannot be claimed for the cropped images in the same way as for the full images.

The combined VGG-radiomics applied to the 2.5D analysis on the full image provided the best results, with a ROC AUC of 0.889, balanced accuracy of 0.898 and F1 score of 0.870, with the incorrect prediction of 2 symptomatic patients and 1 asymptomatic patient.

The results of this study clearly show that deep learning approach outperformed radiomics. Despite its limited standalone impact, radiomics still plays a complementary role, contributing to the improvement in the performance when combined with deep learning in the 2.5D approach.

When comparing the 2D and 2.5D approaches, the two methodologies showed relatively similar performance. The best 2.5D model achieved a small but noteworthy improvement, classifying 2 FNs as the 2D but 2 FPs instead of 5. However, this gain must be balanced against the increased effort required for 2.5D, which involves analyzing multiple slices instead of a single one. This additional step, while manageable, may not always justify the modest improvement in performance, especially in clinical settings where efficiency is a priority. The simplicity of the 2D approach offers the significant advantage of reducing the segmentation effort for radiologists, while still maintaining strong classification performance. This underscores the value of the largest plaque slice, which appears to contain nearly all the critical information necessary to distinguish between symptomatic and asymptomatic cases. The 2.5D approach, even though it requires the segmentation of multiple slices, remains practical, since it is limited to a small number of slices —those with the largest plaque areas—which are relatively easy for a radiologist to identify. This keeps the workload manageable and avoids the need to segment the entire plaque volume. Moreover, this multi-slice approach allows for the inclusion of an arbitrary number of slices from the scan, enabling a more flexible and adaptable analysis. The added effort for 2.5D is rewarded with an increase in performance, providing a more comprehensive analysis without dramatically increasing complexity. Another observation to be done is that, for both VGG and InceptionResNet, the prediction modality based on the maximum predicted probability (across the slices) emerged as the best approach (*Max* approach). This modality identifies the slice with the highest prediction probability and bases the prediction on that slice alone. This aligns with the finding that the most critical diagnostic information for the ML model is concentrated in one slice, which can differ from the one having the largest plaque area. This is likely because plaque vulnerability is influenced not only by its size but also by its composition and morphological characteristics, as well as biological and inflammatory factors. Therefore, the slice with the largest plaque area may not necessarily be the most "informative" or comprehensive for diagnostic purposes.

This study has demonstrated that using full CT scan images is superior to using cropped images. This allows the model to take into account the entire context of the scan, preserving important surrounding information that might be relevant for classification and avoiding the risk of losing important contextual data that may be discarded during the cropping process. The model can potentially detect subtle patterns that are not confined to a specific region, which could improve its ability to capture more comprehensive information about the structure. This approach also makes the model more adaptable to real-world clinical scenarios where full images are often available.

5.2. Comparative Analysis

Although the radiomics results obtained in this study are not comparable to the current state-of-the-art, it is important to highlight the novel contributions made. This study explored 2D and 2.5D approaches for radiomics analysis, methodologies that have been scarcely reported in the existing literature. In fact, only two studies were identified that extracts features using a 2D approach.

The first was conducted by Acharya et al. [2] and it used a dataset where patients were selected based on the presence of either stenosis >50% or plaque alterations like irregular surface, IPH, and presence of ulceration. However, the study does not specify how many patients had stenosis versus plaque alterations, which could introduce a potential bias in distinguishing symptomatic and asymptomatic cases. In contrast, in our study, all patients undergone endarterectomy and had a significant stenosis percentage (77% ± 10%), irrespective of whether they were symptomatic or asymptomatic, ensuring a more homogeneous cohort and reducing the likelihood of selection bias. Another notable difference lies in the handling of data for classification: the cited study included only 20 patients but generated 200 symptomatic and 200 asymptomatic images, treating each slice as an independent sample. This approach, while common in image-level analysis, does not take into account that multiple slices from the same patient might share similarities, potentially altering the classification performance. In our study, while multiple slices were extracted per patient, classification was always performed at the patient level. Specifically, in the 2.5D approach, predictions from multiple slices were aggregated to make a patient-level decision, thereby aligning the classification task with clinical needs and increasing its robustness. Furthermore, the referenced study achieved an accuracy of 0.880; however, the performance evaluation was based on CV performed on individual plaques, without providing details on how the validation sets were constructed. This raises concerns that slices from the same patient might have appeared in both training and validation sets, simplifying the classification task. In our study, CV was performed at the patient level, ensuring that slices from the same patient were never present in both training and validation sets. We also included an independent test set to evaluate the generalizability of our results. It is also worth noting that while the referenced study combined wavelet features with LBP features, our research focused solely on radiomic wavelet features.

The second one, conducted by Le et al. [20], tested both a single-slice and a multi-slice approach. Unlike our methodology, their single-slice analysis used slices extracted specifically at the carotid bifurcation, whereas we focused on the slice with the largest plaque area, targeting the region of greatest clinical significance. Their multi-slice approach was

instead based on a traditional 3D methodology, where 3D radiomic features were extracted from the volume. In contrast, our approach treats slices as 2D, aggregating predictions from multiple slices to achieve patient-level classifications. A critical difference between the datasets lies in the stenosis levels of culprit and non-culprit arteries. Their dataset had a significant disparity, with culprit arteries averaging 72% stenosis and non-culprit arteries only 40%. This introduces a potential bias while, as already said, our dataset included only arteries with significant stenosis, minimizing the bias. Notably, their results for the 2D analysis were not explicitly reported, likely because the performance was inferior to the multi-slice approach. Additionally, they did not validate their method on an external test set, relying instead on a 5-fold CV of the entire dataset. In terms of dataset size, their study included 82 arteries, making it comparable to our cohort of 90 patients when considering only the CV results. In their best-performing configuration, radiomic features alone achieved a ROC AUC of 0.67 in the multi-slice analysis. When combined with calcium scoring, the performance improved to a ROC AUC of 0.73 and reached an accuracy of 0.69. In contrast, our radiomic-only approach achieved, on 2D slices, a ROC AUC of 0.726 and balanced accuracy of 0.732, relying solely on radiomic wavelet features without incorporating external biomarkers like calcium scoring.

Zhang et al. [48] conducted a study using a 3D radiomics approach that achieved comparable performance on the training dataset to our results. Their study focused on classifying plaques based on the presence or absence of IPH, which is not directly equivalent to distinguishing symptomatic from asymptomatic plaques. Their dataset consisted of 106 patients in the training set, and 38 patients in the external test set, with stenosis ranging from 32% to 78% for plaques with IPH and 16% to 56% for plaques without IPH, possibly introducing a small bias due to the IPH stenosis being higher than the other. While the total number of patients is similar to ours (90 patients in the training set and 39 in the test set), their external test set was sourced from a different institution. Zhang et al.'s 3D approach using a radiomic signature model achieved a train AUC of 0.717, increasing to 0.743 when combined with the clinical factor of stenosis, which is comparable to ours of 0.726. While comparing the results obtained on the test set they achieved an AUC of 0.725 for the radiomic model only, which is a small improvement compared to our 0.654 obtained in the 2.5D case, that increased to 0.811 when combined to the stenosis factor.

Moreover, unlike the studies conducted by Xia et al. [45], Shi et al. [36], and Chen et al. [7], which integrate multimodal data such as patient demographics, medical history, or additional plaque-specific metrics like composition and calcification levels, our study exclusively focused on radiomic features. While this choice limits achieving state-of-the-art results, it provides a clearer view of radiomics' standalone predictive power. According

to current research, no other studies have applied a framework comparable to ours while exploring both 2D and 2.5D approaches. This combination of rigor and innovation makes our work an important progress in the field and can encourage future studies on radiomics in carotid plaque analysis.

For what concerns the deep approach, to the best of our knowledge, no studies have been found that apply deep learning specifically to CTA images for the classification of carotid plaques. This highlights the novelty of our work, as it demonstrates the potential of deep networks in a domain where their application has been largely unexplored. While deep learning has been employed in previous studies, these approaches have primarily utilized imaging modalities such as US and MRI. Furthermore, the results obtained in our work were compared with end-to-end approaches, where the network automatically learns to extract highly specialized features specifically tailored for the classification task. In contrast, our approach involves extracting deep features from pre-trained networks, which are then passed to ML models for classification. These features are not task-driven but rather serve as a general-purpose representation of the data. The use of pre-trained models also significantly reduces computational demands compared to more complex, custom-built models. This makes the approach scalable, quick and applicable to a wide range of clinical environments. Additionally, in light of the analyzed literature, no previous studies have utilized the combination of radiomic and deep features and the ensemble modeling for carotid plaque classification. In our case, certain scenarios demonstrated that this integration improved predictions and outperformed individual networks.

Gui et al. [10](2021) demonstrated, as in our study, notable improvements in classification performance when using deep learning features instead of radiomic ones on HRMRI images and a dataset of 104 patients. Their research reported increases in ROC AUC, accuracy, and F1-score, from 0.861 to 0.930, from 0.819 to 0.931 and from 0.753 to 0.861 respectively. In comparison, the improvements observed in our study when transitioning from radiomics to the deep learning approach are substantially larger. For the 2D method, we achieved an increase from 0.620 to 0.877 in ROC AUC, from 0.532 to 0.824 in balanced accuracy, and from 0.300 to 0.741 in F1-score. Similarly, for the 2.5D method, the improvements were from 0.654 to 0.849, from 0.542 to 0.880, and from 0.154 to 0.833, respectively. These results demonstrate that, despite the methodological differences, our approach provides competitive outcomes with reduced complexity. Saba et al. [15] conducted a study using a larger dataset of 346 plaques and US images, testing both with and without data augmentation. With augmentation they achieved an accuracy of 0.897 and a ROC AUC of 0.910. Without augmentation, their performance dropped to 0.862 for accuracy and 0.860 for ROC AUC. Despite not applying augmentation techniques, our results are

superior to their non-augmented case. Gui's study on HRMRI images using a deep end-to-end approach achieved even higher results, with a ROC AUC of 0.930 and an accuracy of 0.931 [10]. It is important to highlight that their study employed a comprehensive 3D analysis, utilizing a 3D deep neural network for feature extraction and classification. In contrast, our approach, while simpler, achieved comparable performance. Specifically, our best model achieved an AUC of 0.889 and a balanced accuracy of 0.898. Also, these results relied heavily on data augmentation (performed 60 times), given their dataset consisted of only 104 patients. The differences in imaging modalities and the extensive augmentation applied in Gui's and Saba's studies may explain their improved performance compared to ours, suggesting that future efforts to incorporate data augmentation in our work could further enhance results. When comparing to studies with smaller datasets, such as Ganitidis et al. (2021) [40], who classified US images from 74 patients using a simple end-to-end CNN, the advantage of our approach becomes evident. Their results, with a ROC AUC of 0.730 and a balanced accuracy of 0.725, are notably lower than our best results (ROC AUC of 0.889, balanced accuracy of 0.898). This comparison underscores how small datasets can limit model performance. On the other hand, studies with larger datasets show improved outcomes. For example, Saba et al. [34], in another work, employed a pre-trained VGG19 network in an end-to-end manner on a dataset of 500 plaques. They achieved a mean ROC AUC of 0.946 and an accuracy of 0.9455 across 10-fold CV. Similarly, He et al.(2024) [19] used a large dataset of 510 plaques and an end-to-end network for US images, achieving a ROC AUC of 0.854 on the test set, which is outperformed by our results.

CTA images, employed in this study, offer several advantages over the US and MRI traditionally used in the literature. CTA provides high-resolution, detailed images of vascular structures, allowing for clearer delineation of carotid plaques and surrounding tissues. Compared to US, CTA images are less impacted by operator variability and imaging depth limitations, resulting in more consistent and reproducible data. Additionally, CTA offers superior contrast resolution for visualizing blood vessels and plaque compositions, especially when dealing with calcified or complex plaques, where MRI may be less effective. CTA is also generally more cost-effective and efficient in clinical settings, making it a practical and reliable choice for routine carotid plaque detection.

5.3. Limitations and future developments

The study presented several limitations that must be acknowledged. One of the primary constraints was the size of the dataset, with the best model being trained on only 90 patients. This limited the model's robustness and generalizability, as such a small dataset may not fully represent the diversity of clinical cases. Expanding the dataset in future studies would be highly beneficial, as it would enable a more comprehensive evaluation of the model's performance across a broader range of cases. Larger cohorts would also support more rigorous validation of the proposed methods, allowing for a better assessment of their clinical efficacy and making them more applicable in real-world scenarios.

Another significant challenge encountered was that radiomic features were unable to capture the complex patterns present in the medical images, which resulted in lower performance. Future research could explore additional techniques to enhance the performance of the radiomic model. One potential avenue is to conduct a 3D radiomic analysis, which could provide a more comprehensive representation of the plaque by including volumetric information. However, it is important to consider that for clinicians, segmenting VOIs requires significantly more time and effort compared to segmenting individual slices. Additionally, performing integration of other features like the ones related to plaque composition, morphological characteristics, or clinical data, could further improve the robustness and accuracy of the model. Lastly, the application of various image transformations other than wavelet ones could also be explored to extract more diverse and meaningful features. Examples are Laplacian of Gaussian, Gabor filters, or Fourier transformations.

The dependence on manually defined ROIs posed an additional challenge. Defining ROIs manually is a time-consuming task that requires substantial domain expertise, which limits the scalability of the model. A potential solution to this issue would be to incorporate the models into an application that allows clinicians to quickly select relevant slices, thus eliminating the need for full segmentation. This would significantly speed up the process and make it more feasible for routine clinical use. Moreover, for the results presented, slices that contain at least 30% of the maximum area of the ROI were selected. However, in future works, more informative slices could be selected directly by clinicians, potentially optimizing the model's performance by including only the most relevant sections of the scan.

Another limitation of this study is the lack of interpretability, particularly in understanding how deep learning models extract features from the images. This is a common issue with deep models, which are often seen as black boxes. To address this, future work could implement interpretability mechanisms, such as heatmaps or attention maps, which would

visually highlight the specific areas and features the model relies on. This would improve the transparency of the model and enhance its clinical usability by providing clinicians with insights into the decision-making process.

Investigating a 3D approach could offer other significant improvements. A 3D model could capture volumetric information from the plaque images, thus providing a more comprehensive and potentially more discriminative dataset for classification. This approach would be particularly useful in medical imaging, where spatial relationships in all three dimensions are often critical for accurate diagnosis. Furthermore, while this study focused on four deep learning architectures, future research could expand this by testing a wider range of networks. Other models, such as those trained on RadImageNet, like VGG, could offer new insights and potentially improve performance.

Lastly, the ensemble approach demonstrated good overall results, particularly in minimizing the number of asymptomatic misclassified cases. This outcome highlights the potential of ensemble methods in enhancing classification accuracy, but it also underscores the need for future studies to address the observed imbalance between asymptomatic and symptomatic cases for improved clinical relevance. The misclassification of symptomatic cases could limit the generalizability of the model in clinical settings where correctly identifying these cases is crucial for appropriate patient management. Further analysis could be conducted to investigate different methods to aggregate information and explore why they perform better in certain scenarios. Exploring advanced ensemble strategies, such as weighted ensembles or those leveraging case-specific thresholds, may help address this class imbalance.

6 | Conclusions

This study presents innovative approaches for assessing symptomatic and asymptomatic carotid plaques through the application of radiomics and deep learning to CTA images, offering significant contributions in an under-explored area of the scientific literature. The research started with the investigation of a 2D approach, which works by focusing on the slice with the largest plaque area. This approach demonstrated that a single slice can often capture the majority of diagnostic information, offering an efficient methodology for real-world applications. It was able to balance simplicity with effectiveness, potentially reducing the effort required for segmentation while still delivering clinically relevant results. While radiomics alone did not produce a model capable of competitive performance, significant improvements were demonstrated with the use of deep learning features extracted from pre-trained networks, which leverage robust, general-purpose representations developed on large and diverse datasets. Building on the 2D approach, the study introduced the 2.5D method using the slices with a plaque area of at least the 30% of the area of the largest plaque. By aggregating predictions from multiple slices, this technique provides a more comprehensive understanding of plaque characteristics without the complexity of full 3D analysis. This approach generally enhanced the 2D, and proved especially effective, reaching a balanced accuracy of 0.9 for the combination of radiomic and deep learning features, a technique still under-investigated. Two versions of the same image were involved, one full and one cropped, with the former consistently demonstrating superior performance compared to the latter. The ability to retain contextual information in full images proved crucial, allowing models to identify patterns extending beyond the ROI. Moreover, the use of these images can remove the need for pre-processing steps like segmentation, reducing the workload for clinicians and making it more adaptable to real-world clinical settings. Finally ensemble methods proved to be another innovative technique, showing promising results as well as limitations, particularly in misclassifying symptomatic cases, which could be further investigated to refine their application and improve clinical reliability. Despite the small dataset size, this research demonstrated the ability to effectively distinguish between symptomatic and asymptomatic plaques. These findings establish a strong foundation for future research, offering a framework for the de-

velopment of clinically applicable tools able to enhance the accuracy and reliability of the classification. This system has the potential to provide critical insights into plaque vulnerability, enable more accurate risk assessments and offer valuable support management of carotid artery disease.

Bibliography

[1] B. e. a. Aboyans, Ricco. 2017 esc guidelines on the diagnosis and treatment of peripheral arterial diseases, in collaboration with the european society for vascular surgery (esvs):. *European Heart Journal*, 39, 2017.

[2] M. e. a. Acharya, Vinitha Sree. Computed tomography carotid wall plaque characterization using a combination of discrete wavelet transform and texture features: A pilot study. *Part H: Journal of Engineering in Medicine*, 2013.

[3] A. Alnuaimi and T. Albaldawi. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97, 2024.

[4] J. J. Berman. Chapter 4 - understanding your data. In *Data Simplification*, pages 135–187. Morgan Kaufmann, 2016.

[5] G. Canbek, S. Sagiroglu, T. T. Temizel, and N. Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826, 2017.

[6] D. J. Charlick M. *Anatomy, Head and Neck: Internal Carotid Arteries*. Treasure Island (FL): StatPearls, 2024.

[7] C. Chen, W. Tang, Y. Chen, W. Xu, N. Yu, C. Liu, Z. Li, Z. Tang, and X. Zhang. Computed tomography angiography-based radiomics model to identify high-risk carotid plaques. *Quantitative Imaging in Medicine and Surgery*, 13, 2023. URL `https://qims.amegroups.org/article/view/116470`.

[8] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, 2016.

[9] Q. Cheng, P. Varshney, and M. Arora. Logistic regression for feature selection and soft classification of remote sensing data. *Geoscience and Remote Sensing Letters, IEEE*, 3, 2006.

[10] X. Z. J. Z. G. N. Chengzhi Gui, Chen Cao and D. Ming. Radiomics and artificial neural networks modelling for identification of high-risk carotid plaques. *Front Cardiovascular Medicine*, 2023.

[11] M. G. L. J. e. a. Cilla, S. Ct angiography-based radiomics as a tool for carotid plaque characterization: a pilot study. *La radiologia medica*, 127:743–753, 2022.

[12] A. C. S. K. K. B. G. A. K. G. Dakis K, Nana P. Carotid plaque vulnerability diagnosis by cta versus mra: A systematic review. *Diagnostics (Basel)*, 2023.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[14] L. H. S. J. L. J. L. Q. S. X. Z. F. C. X. L. G. Dong Z, Zhou C. Radiomics versus conventional assessment to identify symptomatic participants at carotid computed tomography angiography. *Cerebrovascular diseases*, 51:647–654, 2022.

[15] L. S. et al. A multicenter study on carotid ultrasound plaque tissue characterization and classification using six deep artificial intelligence models: A stroke application. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[16] S. M. A. W. F. C. S. S. G. Geiger MA, Flumignan RLG. Carotid plaque composition and the importance of non-invasive in imaging stroke prevention. *Front Cardiovasc Med.*, 2022.

[17] A. P. e. a. Griethuysen, Fedorov. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 2017. URL `https://doi.org/10.1158/0008-5472.CAN-17-0339`.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

[19] W. Y. C. W. D. L. W. Y. Y. W. L. X. Z. Y. H. Y. S. E. He L, Yang Z. A deep learning algorithm to identify carotid plaques and assess their stability. *Frontiers Artificial Intelligence*, 2024.

[20] T. e. a. Le, Rundo. Assessing robustness of carotid artery ct angiography radiomics in the identification of culprit lesions in cerebrovascular events. *scientific reports*, 11, 2021.

[21] B. C. M. M. P. E. R. P. P. B. A. N. L. M. C. C. M. W. Luca Saba, Michele Anzidei.

Imaging of the carotid artery vulnerable plaque. *CardioVascular and Interventional Radiology*, 2014.

[22] H. R. J. C. Y. T. S. H. D. S. B. A. W. L. H. B. M. W. Luca Saba, Tobias Saam. Imaging biomarkers of vulnerable carotid plaques for stroke risk prediction and their potential clinical implications. *Lancet Neurol*, 2019.

[23] C. Maklin. Synthetic minority over-sampling technique (smote), 2022. URL `https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c`.

[24] L. e. a. Mayerhoefer, Materka. Introduction to radiomics. *Journal of Nuclear Medicine*, 2020.

[25] C. McCague, S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, and R. Woitek. Introduction to radiomics for a clinical audience. *Clinical Radiology*, 78:83–98, 2023.

[26] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 2022. URL `https://doi.org/10.1148/ryai.210315`.

[27] A. L. Mira Katan. Global burden of stroke. *Seminar in Neurology*, 2018.

[28] D. M. S. A. Mughal, Khan. Symptomatic and asymptomatic carotid artery plaque. *Expert Rev Cardiovasc Ther*, 2011.

[29] R. Muthukrishnan and R. Rohini. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016.

[30] T. E. R. P. G. M. A. F. C. G. S. G. Naim C, Douziech M. Vulnerable atherosclerotic carotid plaque evaluation by ultrasound, computed tomography angiography, and magnetic resonance imaging: an overview. *Canadian Association of Radiologists*, 2014.

[31] M. Okur. A comprehensive review of feature selection and feature selection stability in machine learning. *GAZI UNIVERSITY JOURNAL OF SCIENCE*, 36, 2022.

[32] W. S. Organization. Impact of stroke. https://www.world-stroke.org.

[33] C. e. a. Saba, Agarwal. Review of imaging biomarkers for the vulnerable carotid plaque. *JVS-Vascular Science*, 2021.

[34] S. S. G. S. e. a. Saba, L. Ultrasound-based internal carotid artery plaque characterization using deep learning paradigm on a supercomputer: a cardiovascular disease/stroke risk assessment system. *The International Journal of Cardiovascular Imaging*, 27:1511–1528, 2021.

[35] R. Scicolone, S. Vacca, F. Pisu, J. C. Benson, V. Nardi, G. Lanzino, J. S. Suri, and L. Saba. Radiomics and artificial intelligence: General notions and applications in the carotid vulnerable plaque. *European Journal of Radiology*, 176, 2024.

[36] H. e. a. Shi, Sun. Radiomics signatures of carotid plaque on computed tomography angiography. *Clinical Neuroradiology*, 33, 2023.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL `https://arxiv.org/abs/1409.1556`.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, 2015. URL `https://arxiv.org/abs/1512.00567`.

[39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016. URL `https://arxiv.org/abs/1602.07261`.

[40] K. D. N. M. S. G. K. S. N. Theofanis Ganitidis, Maria Athanasiou. Stratification of carotid atheromatous plaque using interpretable deep learning methods on b-mode ultrasound images. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society*, 2021.

[41] K. D. N. M. S. G. K. S. N. Theofanis Ganitidis, Maria Athanasiou. Stratification of carotid atheromatous plaque using interpretable deep learning methods on b-mode ultrasound images. *43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society*, 2022.

[42] G. J. U-King-Im JM, Young V. Carotid-artery imaging in the diagnosis and management of patients at risk of stroke. *The Lancet Neurology*, 2009.

[43] X. W. Z. Y. X. C. A. F. M. D. Wei Ma, Yujiao Xia. Object-specific four-path network for stroke risk stratification of carotid arteries in ultrasound images. *Computational and Mathematical Methods in Medicine*, 2022.

[44] Q. J. Wenchao Zhang, Yu Guo. Radiomics and its feature selection: A review. *Symmetry 2023*, 15, 2023.

[45] Z. e. a. Xia, Yuan. Predicting transient ischemic attack risk in patients with mild carotid stenosis using machine learning and ct radiomics. *Frontiers in Neurology*, 14, 2023.

[46] S. Xie, Y. Zhang, D. Lv, X. Chen, J. Lu, and J. Liu. A new improved maximal relevance and minimal redundancy method based on feature subset. *The Journal of Supercomputing*, 2023.

[47] R. Yao, J. Li, M. Hui, L. Bai, and Q. Wu. Feature selection based on random forest for partial discharges characteristic set. *IEEE Access*, 8, 2020.

[48] G. L. K. B. e. a. Zhang, S. Radiomics assessment of carotid intraplaque hemorrhage: detecting the vulnerable patients. *Insights Imaging*, 13, 2022. URL `https://doi.org/10.1186/s13244-022-01324-2`.

[49] W. F. C. X. W. X. Y. J. Y. Z. Z. R. Zhang Y, Gan H. A self-supervised fusion network for carotid plaque ultrasound image classification. *Math Biosci Eng*, 2024.

A | Appendix A

A.1. Radiomics preliminary feature selection

The following table shows the 111 radiomic features kept after the Pearson Correlation selection. In bold instead the 39 features kept after applying ANOVA F-test p-value.

original_shape2D_Elongation	original_shape2D_MaximumDiameter
original_shape2D_MinorAxisLength	original_shape2D_Perimeter
original_shape2D_PerimeterSurfaceRatio	original_shape2D_Sphericity
original_gldm_SmallDependenceLowGrayLevelEmphasis	original_glrlm_LongRunLowGrayLevelEmphasis
original_glszm_LargeAreaHighGrayLevelEmphasis	**original_glszm_SmallAreaEmphasis**
original_glszm_SmallAreaLowGrayLevelEmphasis	wavelet-LH_firstorder_Kurtosis
wavelet-LH_firstorder_Mean	wavelet-LH_firstorder_Median
wavelet-LH_firstorder_Skewness	**wavelet-LH_glcm_Imc1**
wavelet-LH_glcm_Correlation	wavelet-LH_glcm_Idn
wavelet-LH_glcm_InverseVariance	wavelet-LH_glcm_MCC
wavelet-LH_gldm_DependenceVariance	wavelet-LH_gldm_SmallDependenceLowGrayLevelEmphasis
wavelet-LH_glrlm_LongRunHighGrayLevelEmphasis	wavelet-LH_glszm_LargeAreaHighGrayLevelEmphasis
wavelet-LH_glszm_LargeAreaLowGrayLevelEmphasis	**wavelet-LH_glszm_SizeZoneNonUniformity**
wavelet-LH_glszm_SmallAreaEmphasis	wavelet-LH_glszm_SmallAreaLowGrayLevelEmphasis
wavelet-LH_glszm_ZoneEntropy	wavelet-LH_glszm_ZoneVariance
wavelet-LH_ngtdm_Complexity	**wavelet-LH_ngtdm_Contrast**
wavelet-LH_ngtdm_Strength	**wavelet-HL_firstorder_Mean**
wavelet-HL_firstorder_Median	wavelet-HL_firstorder_Skewness
wavelet-HL_glcm_Correlation	**wavelet-HL_glcm_Idn**
wavelet-HL_glcm_Imc1	wavelet-HL_glcm_Imc2
wavelet-HL_glcm_InverseVariance	wavelet-HL_glcm_MCC
wavelet-HL_gldm_DependenceVariance	**wavelet-HL_glrlm_GrayLevelNonUniformity**
wavelet-HL_glrlm_LongRunHighGrayLevelEmphasis	**wavelet-HL_glrlm_RunEntropy**
wavelet-HL_glszm_GrayLevelNonUniformity	wavelet-HL_glszm_LargeAreaHighGrayLevelEmphasis
wavelet-HL_glszm_LargeAreaLowGrayLevelEmphasis	wavelet-HL_glszm_LowGrayLevelZoneEmphasis
wavelet-HL_glszm_SmallAreaEmphasis	wavelet-HL_glszm_SmallAreaLowGrayLevelEmphasis

wavelet-HL_glszm_ZoneEntropy	wavelet-HL_glszm_ZoneVariance
wavelet-HL_ngtdm_Busyness	**wavelet-HL_ngtdm_Complexity**
wavelet-HL_ngtdm_Contrast	wavelet-HL_ngtdm_Strength
wavelet-HH_firstorder_Kurtosis	wavelet-HH_firstorder_Mean
wavelet-HH_firstorder_Median	wavelet-HH_firstorder_Skewness
wavelet-HH_glcm_ClusterShade	**wavelet-HH_glcm_Correlation**
wavelet-HH_glcm_InverseVariance	wavelet-HH_glcm_MCC
wavelet-HH_gldm_DependenceVariance	wavelet-HH_glrlm_GrayLevelNonUniformity
wavelet-HH_glrlm_LongRunHighGrayLevelEmphasis	**wavelet-HH_glrlm_RunEntropy**
wavelet-HH_glrlm_RunLengthNonUniformity	**wavelet-HH_glrlm_RunVariance**
wavelet-HH_glrlm_ShortRunLowGrayLevelEmphasis	**wavelet-HH_glszm_GrayLevelNonUniformity**
wavelet-HH_glszm_GrayLevelNonUniformityNormalized	wavelet-HH_glszm_LargeAreaLowGrayLevelEmphasis
wavelet-HH_glszm_LowGrayLevelZoneEmphasis	**wavelet-HH_glszm_SizeZoneNonUniformity**
wavelet-HH_glszm_SmallAreaEmphasis	wavelet-HH_glszm_SmallAreaLowGrayLevelEmphasis
wavelet-HH_glszm_ZoneEntropy	**wavelet-HH_glszm_ZonePercentage**
wavelet-HH_glszm_ZoneVariance	wavelet-HH_ngtdm_Coarseness
wavelet-HH_ngtdm_Complexity	wavelet-HH_ngtdm_Contrast
wavelet-HH_ngtdm_Strength	wavelet-LL_firstorder_10Percentile
wavelet-LL_firstorder_Minimum	wavelet-LL_firstorder_Skewness
wavelet-LL_glcm_Correlation	wavelet-LL_glcm_Idn
wavelet-LL_glcm_Imc1	wavelet-LL_glcm_Imc2
wavelet-LL_glcm_MCC	wavelet-LL_glrlm_GrayLevelNonUniformity
wavelet-LL_glrlm_LongRunLowGrayLevelEmphasis	**wavelet-LL_glszm_GrayLevelNonUniformity**
wavelet-LL_glszm_LargeAreaHighGrayLevelEmphasis	wavelet-LL_glszm_LargeAreaLowGrayLevelEmphasis
wavelet-LL_glszm_SizeZoneNonUniformity	**wavelet-LL_glszm_SmallAreaEmphasis**
wavelet-LL_glszm_SmallAreaLowGrayLevelEmphasis	wavelet-LL_glszm_ZoneEntropy
wavelet-LL_glszm_ZonePercentage	wavelet-LL_glszm_ZoneVariance
wavelet-LL_ngtdm_Busyness	wavelet-LL_ngtdm_Coarseness
wavelet-LL_ngtdm_Complexity	**wavelet-LL_ngtdm_Contrast**
wavelet-LL_ngtdm_Strength	

Table A.1: Table of the radiomic features kept by Pearson correlation, in bold the ones kept by p_value

B | Appendix B

B.1. 2D Test results

B.1.1. Single deep network

Network	Classifier	Selector	Features number	ROC AUC	Bal acc	F1	Confusion matrix
VGG	LR	LASSO (6)	6	0.858	0.500	0	$\begin{bmatrix} 27 & 0 \\ 12 & 0 \end{bmatrix}$
RES	LR	mRMR (4)	1	0.853	0.819	0.750	$\begin{bmatrix} 24 & 3 \\ 3 & 9 \end{bmatrix}$
INC	LR	RF (4)	1	0.784	0.667	0.500	$\begin{bmatrix} 27 & 0 \\ 8 & 4 \end{bmatrix}$
INCRES	RF	LR (26)	19	0.807	0.787	0.690	$\begin{bmatrix} 20 & 7 \\ 2 & 10 \end{bmatrix}$

Table B.1: Test metrics on the best model for all networks (Cropped image)

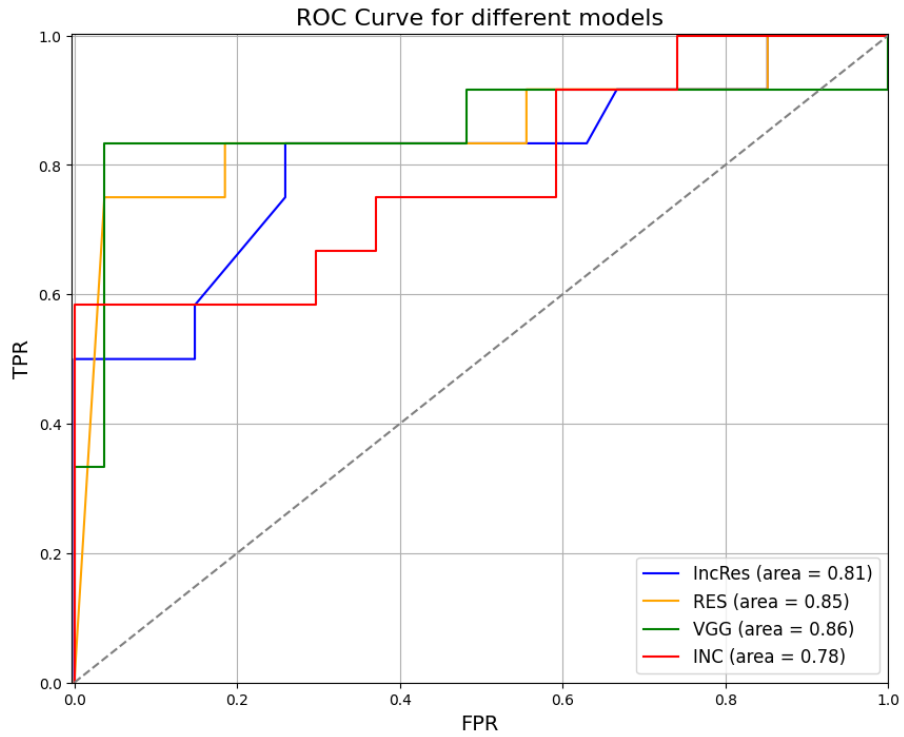


Figure B.1: ROC curves of all networks (Cropped image)

Network	Classifier	Selector	Features number	ROC AUC	Bal acc	F1	Confusion matrix
VGG	RF	RF (23)	20	0.877	0.824	0.741	$\begin{bmatrix} 22 & 5 \\ 2 & 10 \end{bmatrix}$
RES	SVM	LASSO (0.015)	12	0.910	0.843	0.769	$\begin{bmatrix} 23 & 4 \\ 2 & 10 \end{bmatrix}$
INC	RF	LR (4)	3	0.670	0.676	0.571	$\begin{bmatrix} 14 & 13 \\ 2 & 10 \end{bmatrix}$
INCRES	XGBoost	LR (24)	21	0.744	0.648	0.641	$\begin{bmatrix} 17 & 10 \\ 4 & 8 \end{bmatrix}$

Table B.2: Test metrics on the best model for all networks (Full image)

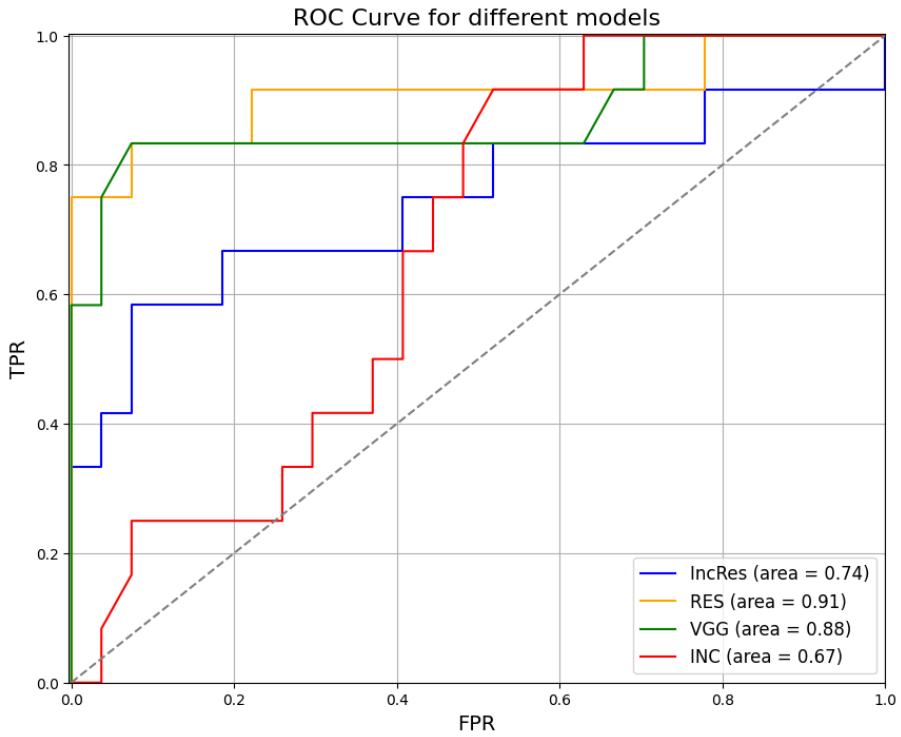


Figure B.2: ROC curves of all networks (Full image)

B.1.2. Combination deep and radiomic features

Img Type	Network	Classifier	Features number	ROC	Bal Acc	F1	Confusion Matrix
Cropped	VGG	ensemble	10	0.670	0.611	0.500	$\begin{bmatrix} 15 & 12 \\ 3 & 9 \end{bmatrix}$
Cropped	RES	LR	5	0.775	0.667	0.552	$\begin{bmatrix} 18 & 9 \\ 4 & 8 \end{bmatrix}$
Cropped	INC	ensemble	5	0.661	0.639	0.5	$\begin{bmatrix} 21 & 6 \\ 6 & 6 \end{bmatrix}$
Cropped	INCRES	XGBoost	11	0.698	0.583	0.444	$\begin{bmatrix} 18 & 9 \\ 6 & 6 \end{bmatrix}$
Full	VGG	ensemble	21	0.907	0.787	0.690	$\begin{bmatrix} 20 & 7 \\ 2 & 10 \end{bmatrix}$
Full	RES	ensemble	15	0.864	0.759	0.667	$\begin{bmatrix} 23 & 4 \\ 4 & 8 \end{bmatrix}$
Full	INC	ensemble	7	0.707	0.588	0.467	$\begin{bmatrix} 16 & 11 \\ 5 & 7 \end{bmatrix}$
Full	INCRES	RF	23	0.836	0.671	0.563	$\begin{bmatrix} 16 & 11 \\ 3 & 9 \end{bmatrix}$

Table B.3: Test metrics on combination

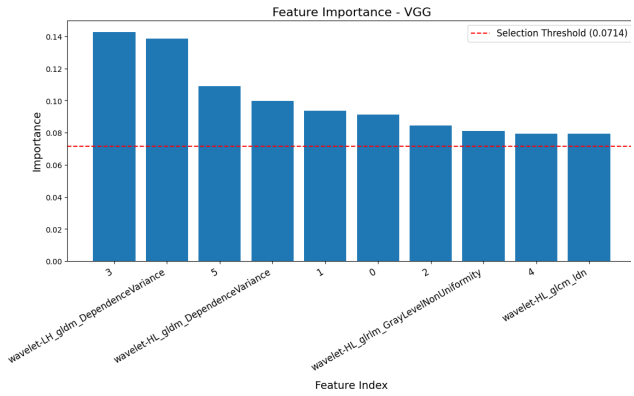


Figure B.3: Importance diagram for VGG (Cropped)

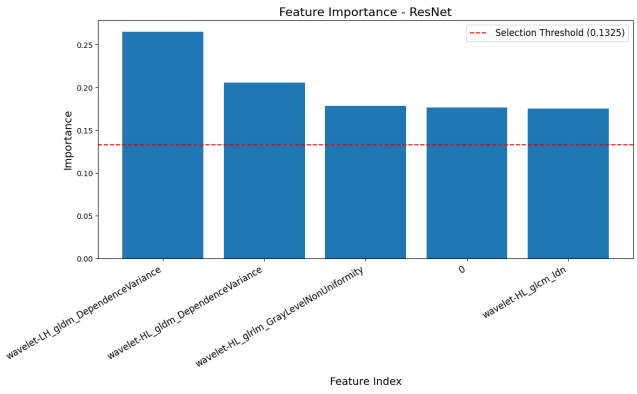


Figure B.4: Importance diagram for RES (Cropped)

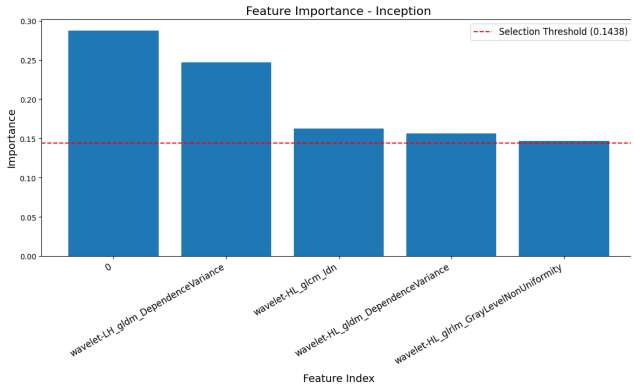


Figure B.5: Importance diagram for INC (Cropped)

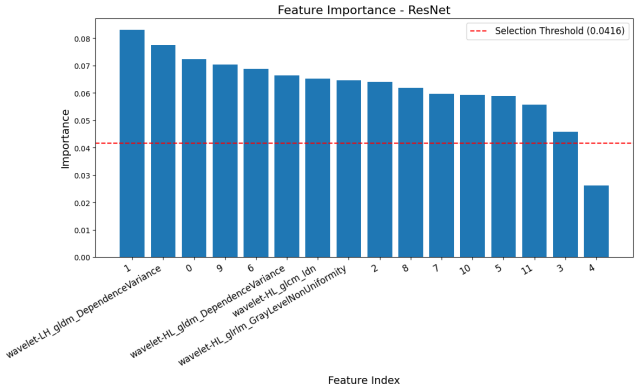


Figure B.6: Importance diagram for RES (Full)

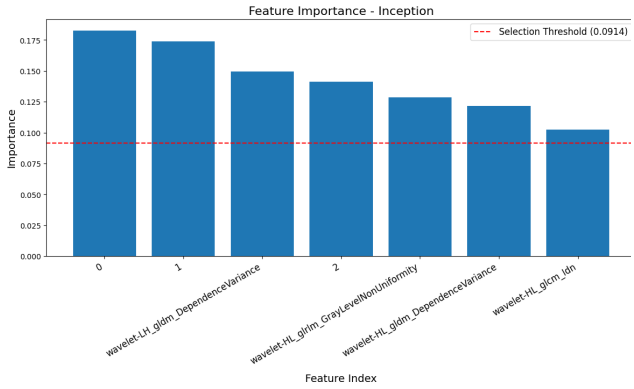


Figure B.7: Importance diagram for INC (Full)

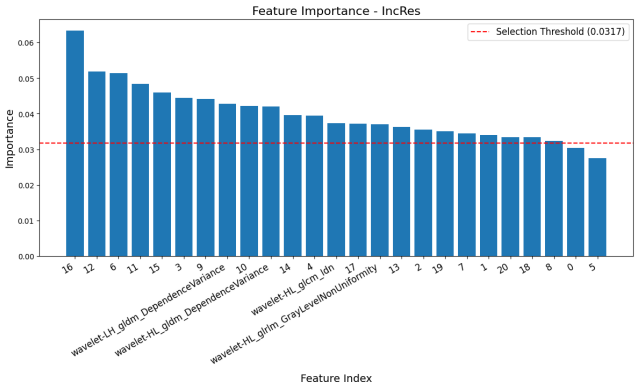


Figure B.8: Importance diagram for IncRes (Full)

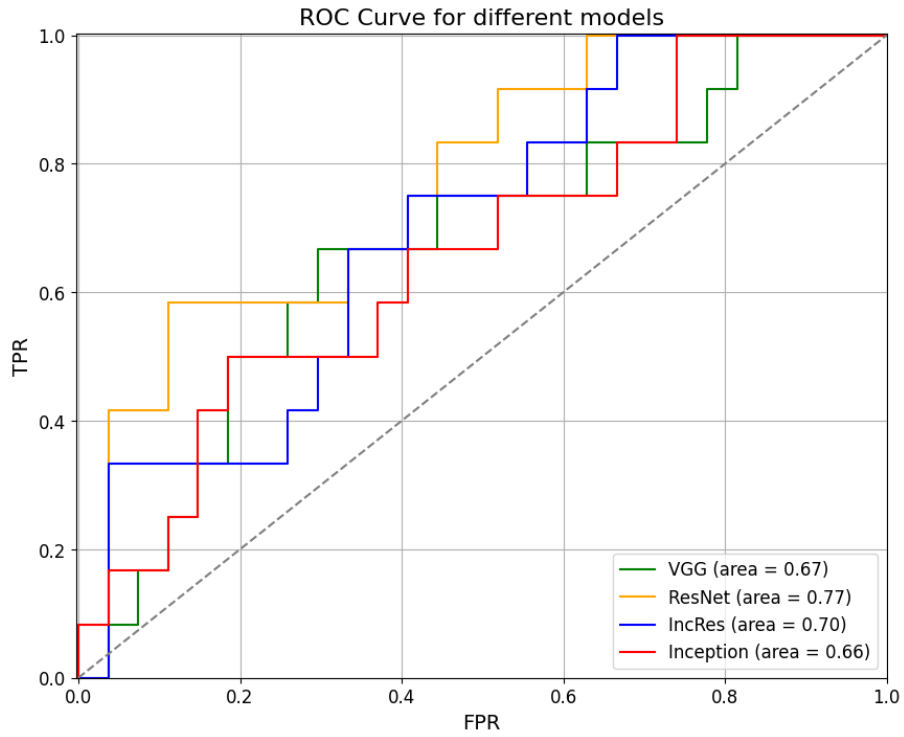


Figure B.9: ROC curve of all networks combined (Cropped image)

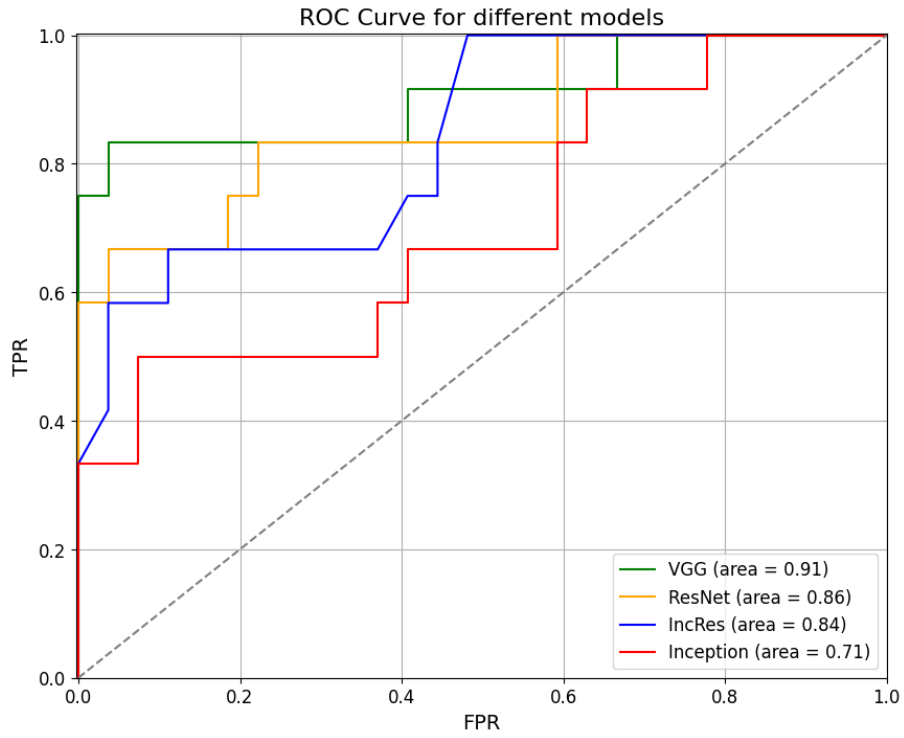


Figure B.10: ROC curve of all networks combined (Full image)

B.2. 2.5D Test results

B.2.1. Single deep network

Img Type	Network	Classifier	Features number	ROC	Bal Acc	F1	Confusion Matrix
Cropped	VGG	Mean	6	0.478	0.500	0.471	$\begin{bmatrix} 0 & 27 \\ 0 & 12 \end{bmatrix}$
Cropped	Res	Max	1	0.798	0.819	0.750	$\begin{bmatrix} 24 & 3 \\ 3 & 9 \end{bmatrix}$
Cropped	Inc	Mean	1	0.725	0.667	0.500	$\begin{bmatrix} 27 & 0 \\ 8 & 4 \end{bmatrix}$
Cropped	IncRes	Max	19	0.809	0.745	0.643	$\begin{bmatrix} 20 & 7 \\ 3 & 9 \end{bmatrix}$
Full	VGG	Max	20	0.849	0.880	0.833	$\begin{bmatrix} 25 & 2 \\ 2 & 10 \end{bmatrix}$
Full	Res	Max	12	0.892	0.866	0.786	$\begin{bmatrix} 22 & 5 \\ 1 & 11 \end{bmatrix}$
Full	Inc	Mean	3	0.772	0.690	0.581	$\begin{bmatrix} 17 & 10 \\ 3 & 9 \end{bmatrix}$
Full	IncRes	Max	21	0.802	0.806	0.714	$\begin{bmatrix} 21 & 6 \\ 2 & 10 \end{bmatrix}$

Table B.4: Test metrics on best model for the best 2D network on cropped and full slices for 2.5D

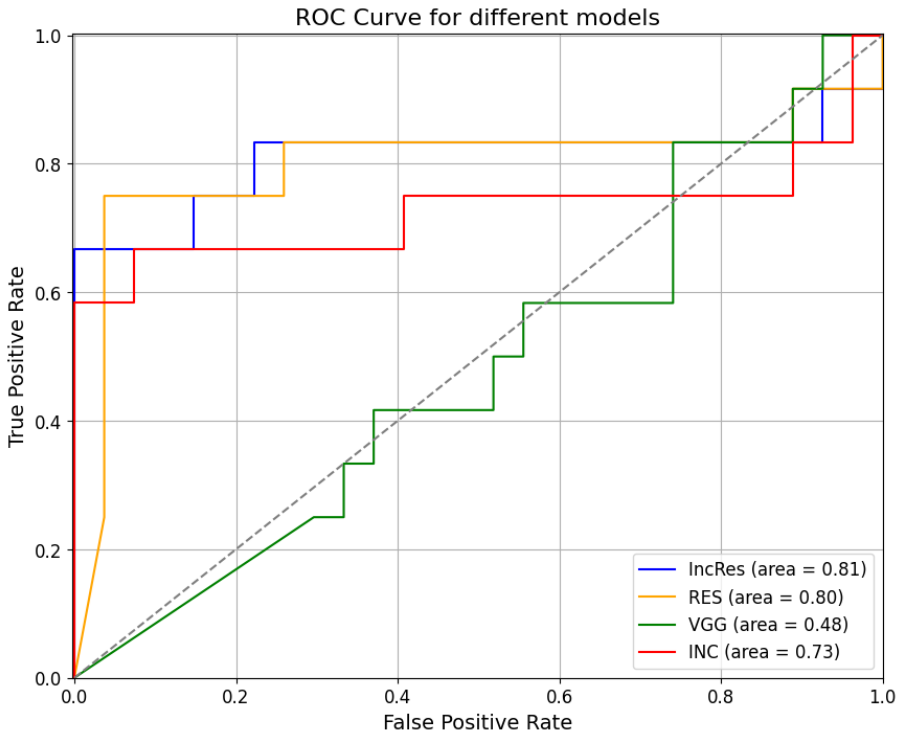


Figure B.11: ROC curve of all networks (Cropped image) 2.5

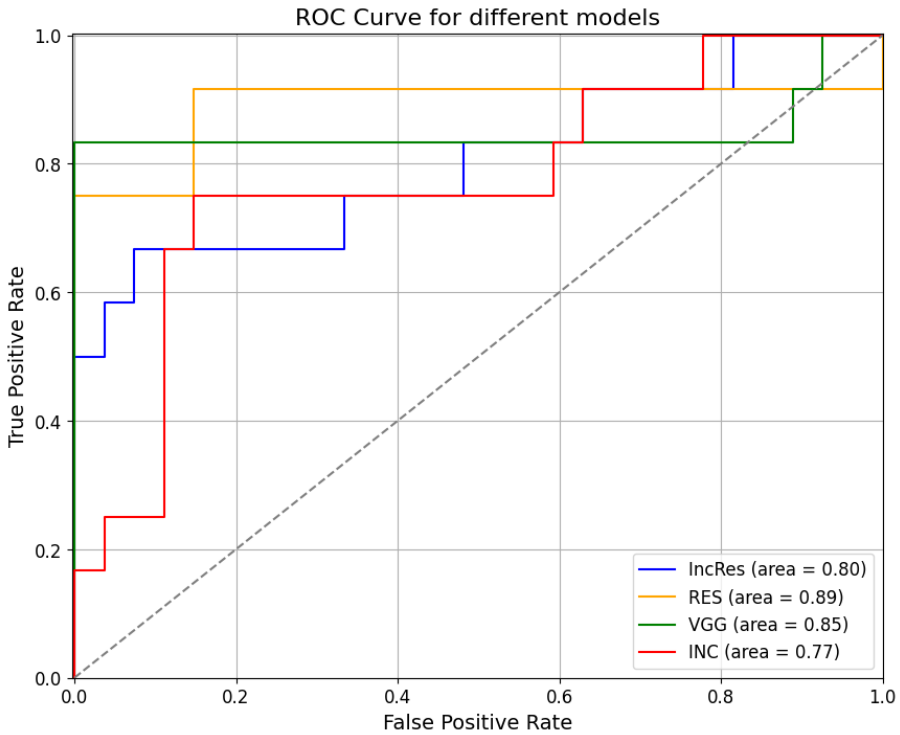


Figure B.12: ROC curve of all networks (Full image) 2.5

B.2.2. Combination deep and radiomic features

Img Type	Network	Classifier	Features number	ROC	Bal Acc	F1	Confusion Matrix
Cropped	VGG	Mean	10	0.725	0.500	0.471	$\begin{bmatrix} 0 & 27 \\ 0 & 12 \end{bmatrix}$
Cropped	Res	Max	5	0.809	0.856	0.818	$\begin{bmatrix} 26 & 1 \\ 3 & 9 \end{bmatrix}$
Cropped	Inc	Mean	5	0.660	0.653	0.500	$\begin{bmatrix} 24 & 3 \\ 7 & 5 \end{bmatrix}$
Cropped	IncRes	Max	11	0.741	0.616	0.514	$\begin{bmatrix} 13 & 14 \\ 3 & 9 \end{bmatrix}$
Full	VGG	Max	21	0.889	0.898	0.870	$\begin{bmatrix} 26 & 1 \\ 2 & 10 \end{bmatrix}$
Full	Res	Max	15	0.917	0.838	0.783	$\begin{bmatrix} 25 & 2 \\ 3 & 9 \end{bmatrix}$
Full	Inc	Mean	7	0.738	0.718	0.609	$\begin{bmatrix} 23 & 4 \\ 5 & 7 \end{bmatrix}$
Full	IncRes	Max	23	0.898	0.824	0.741	$\begin{bmatrix} 22 & 5 \\ 2 & 10 \end{bmatrix}$

Table B.5: Test metrics on best combination model for the best 2D network on cropped and full slices for 2.5

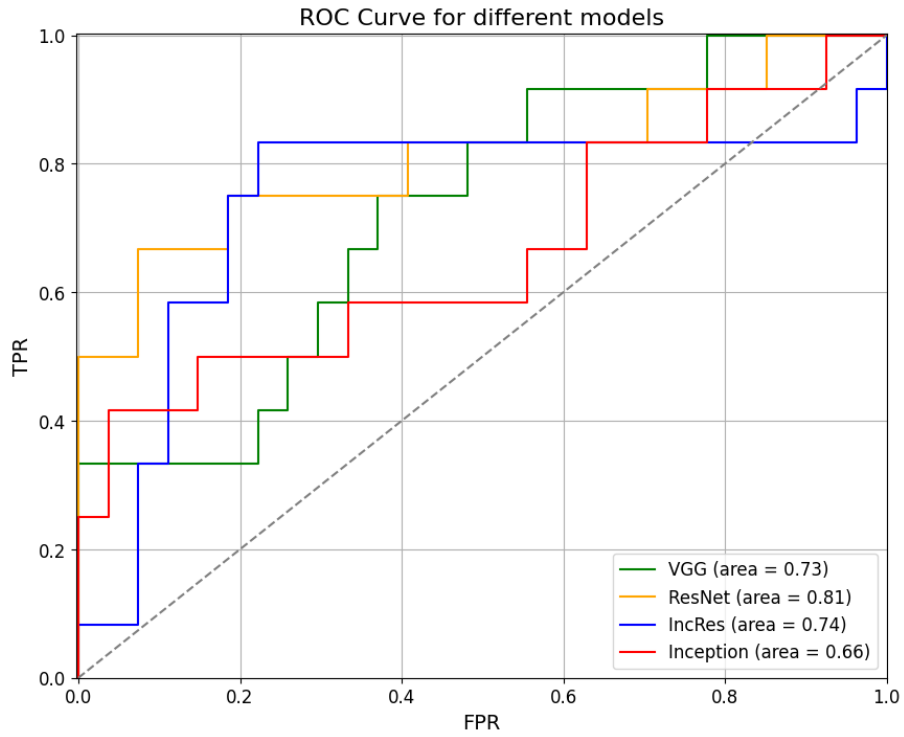


Figure B.13: ROC curve of all networks combined (Cropped image) 2.5

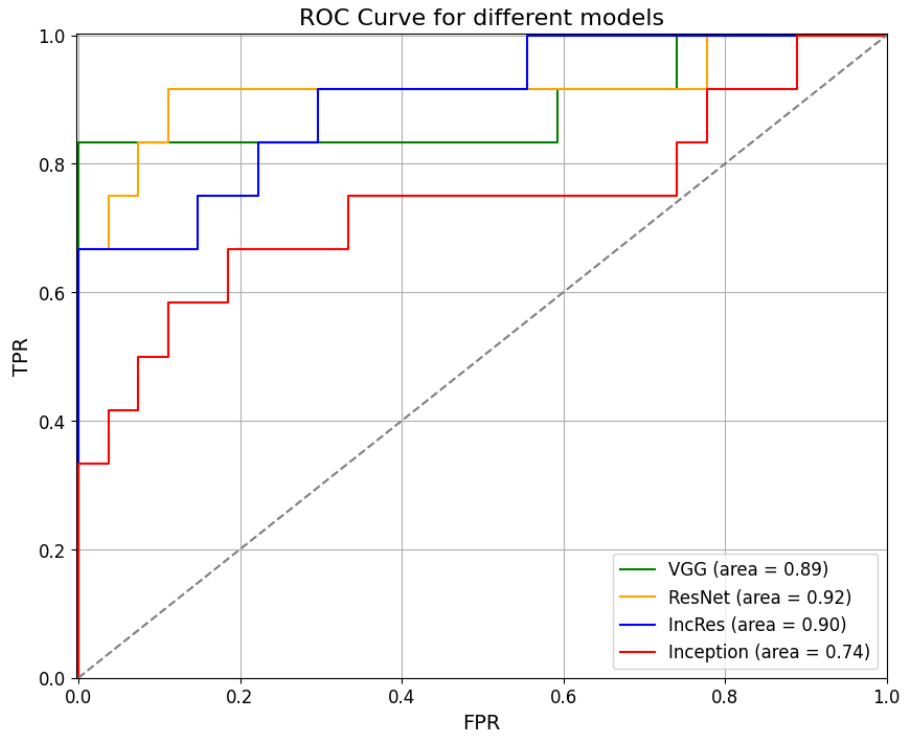


Figure B.14: ROC curve of all networks combined (Full image) 2.5

C | Appendix C

C.1. Slices selection for 2.5D approach

Patient	Original Slices	30% Slices	Patient	Original Slices	30% Slices
1	39	28	29	33	30
2	36	22	30	47	32
3	49	39	31	33	19
4	22	18	32	48	36
5	33	28	33	39	29
6	29	22	34	49	39
7	42	30	35	45	32
8	57	46	36	37	31
9	51	48	37	24	19
10	33	24	38	47	37
11	34	29	39	39	28
12	60	48	40	38	27
13	40	31	41	20	16
14	26	21	42	52	42
15	29	20	43	33	22
16	63	36	44	43	35
17	31	28	45	31	26
18	34	27	46	30	19
19	26	18	47	39	27
20	26	17	48	36	28
21	41	34	49	38	31
22	50	37	50	31	20
23	18	13	51	29	22
24	33	28	52	57	53
25	30	21	53	74	57
26	49	42	54	40	34
27	50	38	55	31	28
28	41	35	56	113	97

Patient	Original Slices	30% Slices	Patient	Original Slices	30% Slices
57	40	33	94	45	39
58	36	28	95	44	34
59	34	24	96	66	60
60	48	35	97	25	21
61	31	19	98	32	22
62	26	21	99	36	30
63	24	14	100	42	34
64	34	25	101	33	28
65	42	32	102	38	30
66	38	29	103	22	17
67	42	33	104	34	25
68	54	40	105	43	32
69	46	36	106	51	37
70	45	35	107	74	61
71	38	31	108	35	30
72	30	25	109	31	25
73	44	27	110	43	28
74	27	21	111	52	46
75	41	32	112	50	42
76	49	43	113	42	36
77	26	18	114	39	32
78	46	43	115	54	46
79	27	23	116	65	59
80	37	32	117	52	46
81	44	26	118	40	29
82	41	33	119	43	36
83	33	25	120	40	34
84	46	32	121	29	21
85	38	31	122	31	20
86	55	33	123	44	39
87	28	18	124	51	36
88	40	19	125	37	31
89	42	33	126	55	48
90	24	16	127	54	47
91	41	32	128	51	44
92	44	36	129	51	35
93	40	29	**TOTAL**	5244	4086

Table C.1: Number of slices of each patients before and after selecting the 30%. At the end the total number of slices.

Acknowledgements

Arrivati finalmente alla fine del nostro percorso universitario ci teniamo a ringraziare le persone che ci hanno permesso di arrivare fino a questo splendido, e sudato, traguardo.

Vogliamo iniziare ringraziando la professoressa Anna Corti e la professoressa Valentina Corino per averci dato l'opportunità di prendere parte a una ricerca così interessante quanto importante che ha le basi per diventare utile in un contesto clinico reale. Nonostante i risultati non sempre ottimi, e gli svariati tentativi con gli autoencoder, hanno saputo consigliarci ed aiutarci in ogni momento della ricerca rendendo più piacevole e stimolante il percorso di tesi. Grazie a questa oportunità abbiamo avuto un assaggio di cosa sia il mondo della ricerca e ci piacerebbe in futuro prenderne ancora parte.

Vorremmo poi ringraziare i nostri colleghi e amici, che sono stati fondamentali durante questi due anni milanesi. Siete riusciti a rendere tutto più leggero, dalle pause studio agli aperitivi in Porta Venezia. Un ringraziamento speciale va anche a tutti gli amici di sempre, grazie per averci sempre sopportato, vi vogliamo bene.

Infine, il grazie più importante alle nostre famiglie, senza di voi.. chi lo pagava l'affitto? A parte gli scherzi, senza di voi e il vostro supporto nulla nella nostra vita sarebbe mai possibile.