POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Computer Engineering - Ingegneria Informatica

Authors: Stefano Baroni, Alessia Menozzi

Advisor: Prof. Anna Corti

Co-advisor: Prof. Valentina Corino

Academic year: 2023-2024

1. Introduction

Stroke is one of the leading causes of death and disability in all the world, with about 12 million cases and 6.5 deaths each year [7]. For survivors, it often leads to significant disabilities and requires extensive care. Atherosclerosis, a chronic and progressive inflammatory condition characterized by the accumulation of lipids and fibrous tissue within the arterial wall [6], plays a critical role in the development and cerebrovascular diseases. More specifically, atherosclerotic carotid plaques rupture is estimated to be the cause of the 10–15% of all strokes and transient ischemic attack (TIA) cases. Vessel stenosis has always been the primary parameter for patient risk stratification and indication for treatment [1]. Advances in imaging technologies have enabled a more detailed characterization of the plaque and, particularly, computed tomography angiography (CTA) has emerged as one of the leading methods for carotid artery evaluation [6].
The study of carotid plaques from medical images and their classification into symptomatic and asymptomatic categories has attracted considerable attention in recent scientific literature. One way to achieve this is through radiomics, a rapidly developing field of research consisting in the extraction of quantitative features from medical images, able to offer unique information able to deepen the understanding of disease processes and provide clinical decision support. Scicolone et al. in 2024 [8] summarized radiomics studies based on the image type including several ones on CT scans, as used in this study. Other image-based studies adopt deep learning, a specialized subset of machine learning focused on neural networks with multiple layers of interconnected nodes, referred to as "deep" networks due to the large number of hidden layers. So far, most of the deep learning studies on carotid plaques focused on end-to-end models, as the works by Gui et al. [2] and He et al. [4] based on MRI and US respectively.

The objective of the study is to develop novel image-based approaches for the identification of symptomatic and asymptomatic patients. To this aim, first a radiomic-based model is considered. Then, a method based on the extraction of features from pre-trained deep neural networks is implemented, by considering either the full scans or plaque-centered images as input.

Moreover, an integration of the radiomic and deep learning-based features is also tested. The study was first conducted in 2D, analyzing only the slice with the largest plaque area. Subsequently, the analysis is expanded by considering several slices per patient (termed as 2.5D).

2. Materials and Methods

2.1. Patient Dataset

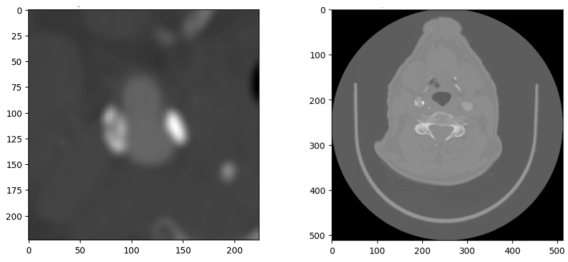
The dataset includes CTA images of 129 patients who underwent elective carotid endarterectomy at the Vascular Surgery Operative Unit of Fondazione IRCSS Ca 'Granda, Ospedale Maggiore Policlinico, Milan. Of the 129 patients, 53 were symptomatic, presenting with either TIA or ischemic strokes. The CTA images employed where acquired using a GE Light Speed VCT 64-slice 3T scanner (GE Healthcare, Little Chalfont, UK). *Table 1* outlines the main clinical characteristics of these patients.

2.2. Radiomic features extraction

Pyradiomics sofware [12] was adopted to extract radiomic features, taking as input the Region of Interest (ROI) and plaque slices. It resulted in a total of 474, of which 102 from the original image and 372 from the wavelet filtered one. Overall, 90 first-order statistical, 9 shape-based and 375 textural features were extracted.

2.3. Deep learning feature extraction

Four different pre-trained networks were utilized for feature extraction: VGG19 [9], trained on Imagenet, ResNet50 [3], InceptionV3 [11], InceptionResNetV2 [10], all trained on RadImagenet, a specialized medical imaging dataset containing millions of images across a variety of radiological categories [5]. The deep learning approach was applied on two types of slices, namely the original full slices and cropped slices, where only the region of the plaque was maintained (*Figure 1*).



Cropped image Full image

Figure 1: Examples of the two types of slices

These two types of slices were subsequently fed into the neural networks. The features maps were then extracted from the last layers of each network and then subjected to Global Max Pooling, which reduces the spatial dimensions by retaining only the maximum value from each fea-

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 1: Clinical characteristics of the 129 patients, grouped into asymptomatic and symptomatic categories. Missing data are in square brackets.

ture channel, resulting in the final set of features to be classified.
As a result, 512, 2048, 2048 and 1536 features were extracted respectively from VGG19, ResNet50, InceptionV3 and Inception-ResNetV2.

3. Machine learning model development

Once the features were extracted, different steps were performed to obtain the final classification, as shown in *Figure 2*. The training process started by splitting the dataset consisting on the extracted features into training and test sets (90 vs 39 patients), where 30% of the patients were reserved for testing according to the date of surgery, a practice commonly adopted in clinical studies.
The first step to reduce high features dimensionality was Pearson correlation, set to a threshold of 0.9, to eliminate highly redundant features. Next, an ANOVA F-test p-value with a threshold of 0.05 was applied to further refine the features based on their statistical relevance to the symptomatic and asymptomatic classes.
In the second step, a 5-fold stratified cross-validation (CV) was applied along with SMOTE to address class imbalance, aiming to identify the optimal combination of selector and classifier. Specifically, four selectors were considered, namely the LASSO, Minimum Redundancy Maximum Relevance (mRMR), Logistic Regression (LR) and Random Forest (RF), and five classifiers were evaluated, namely RF, LR, Support Vector Machine (SVM), MultiLayer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost). Additionally, an ensemble model, combining the predictions of RF, SVM, and LR through soft voting was also considered. Features that appeared in at least three out of five folds were selected and retained. The best combination was identified based on the evaluation of the balanced accuracy, with ROC AUC used as secondary measure. Then, the selected combination of selector-classifier was applied and trained on the entire dataset, and evaluated on the test set.
For the combined approach, features from the best radiomic model and the best model for each neural network were integrated. This set of combined features was then refined using an importance-based filter, which retains features with at least 50% of the importance of the most significant feature, as measured by the Gini index from a RF. Additionally, the best classification algorithm among the six used was determined using a single stratified validation set, including 27 out of the 90 patients of the training set.
Finally, an ensemble approach was tested, combining the predictions of the four best models for each network, with and without the inclusion of the radiomics model. The ensemble used a hard voting mechanism, where ties are resolved by assigning 1, thereby prioritizing symptomatic classification.
This study was conducted using both a 2D and a 2.5D approach. The 2D approach involved selecting a single slice per patient, specifically the one with the largest plaque area as determined by the ROI. In contrast, the 2.5D approach leveraged the same trained model found for the 2D to perform classification across a set

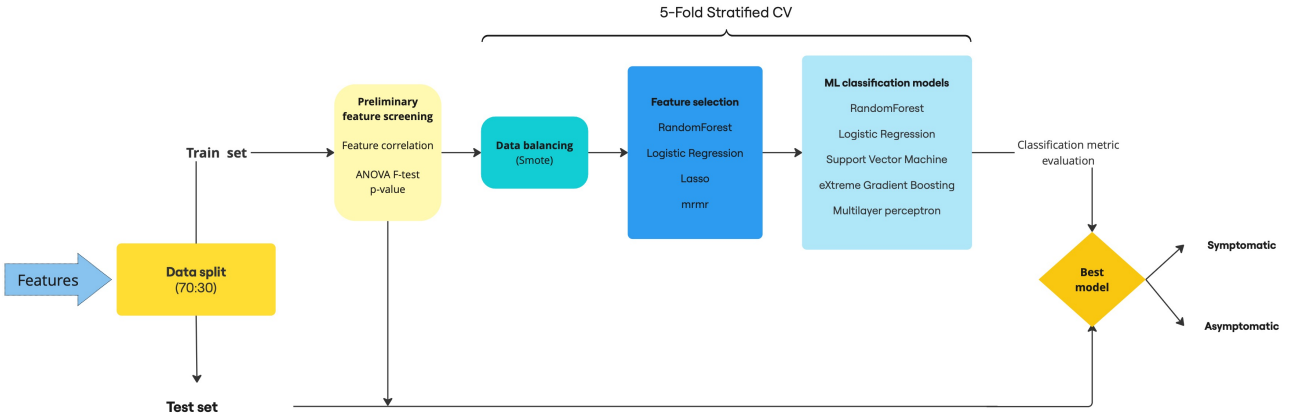


Figure 2: Scheme representing the steps of the training process

of slices for each patient, with a plaque area of at least 30% of the largest plaque area for that patient. The same validation set was used to define the modality to aggregate predictions coming from the different slices. These methods are: (1) the *Mean* approach, which calculates the average predicted probabilities across all slices for each patient; (2) *Majority Voting* (MV), where each slice casts a "vote" and the final prediction is determined by the majority of these votes; (3) the *Max* approach, which bases the prediction on the slice with the highest probability score.

4. Results

Table 2 shows the number of features selected by the Pearson Correlation and the ANOVA F-test p-value.

Network	Img type	Init# feat	Corr	p-value
Radiomics		474	111	39
VGG	Cropped	512	512	6
ResNet	Cropped	2048	915	142
Inception	Cropped	2048	1225	191
InceptionResNet	Cropped	1536	852	162
VGG	Full	512	483	67
ResNet	Full	2048	1862	268
Inception	Full	2048	1600	380
InceptionResNet	Full	1536	1151	158

Table 2: Dimensionality reductions results

4.1. 2D Results

The chosen combination for radiomics is SVM - mRMR showing a balanced accuracy across the CV folds of 0.732 ± 0.120, ROC AUC of 0.726 ± 0.135 and F1 score of 0.715 ± 0.118. These 4 features are kept after feature selection: wavelet-HL_glcm_ldn, wavelet-HL_glrlm_GrayLevelNonUniformity, wavelet-HL_gldm_DependenceVariance, wavelet-LH_gldm_DependenceVariance. Test metrics are shown in *Table 3*.

Bal. acc	AUC	F1	C.M.
0.532	0.620	0.300	$\begin{bmatrix} 22 & 5 \\ 9 & 3 \end{bmatrix}$

Table 3: Test metrics of the radiomic

Table 4 present the validation metrics for the best classifier-selector combination for each deep learning network for cropped images. InceptionResnetV2 was selected as best network, with LR-RF as selector-classifier combination, leading to a final features set of 19 features, as shown in *Table 5* along with the test metrics.

Net.	Bal. acc	ROC AUC	F1
VGG	0.657 ±0.065	0.703 ±0.121	0.603 ±0.065
Res	0.644 ±0.005	0.625 ±0.004	0.593 ±0.055
Inc	0.679 ±0.080	0.755 ±0.120	0.605 ±0.078
IncRes	0.724 ±0.150	0.756 ±0.130	0.703 ±0.166

Table 4: Validation metrics of the deep features approach on cropped slices

Net(#feat)	Bal. acc	AUC	F1	C.M.
IncRes (19)	0.787	0.807	0.690	$\begin{bmatrix} 20 & 7 \\ 2 & 10 \end{bmatrix}$

Table 5: Test metrics of the deep features approach on cropped slices

Table 6 present the validation metrics in the full image case, where VGG19 was selected as best network, with RF-RF as selector-classifier combination, with a final features set of 20. Metrics on the test set are detailed in *Table 7*.

Net.	Bal. acc	ROC AUC	F1
VGG	0.759 ±0.176	0.746 ±0.153	0.718 ±0.209
Res	0.671 ±0.080	0.754 ±0.110	0.624 ±0.087
Inc	0.726 ±0.008	0.770 ±0.008	0.716 ±0.008
IncRes	0.662 ±0.150	0.694 ±0.160	0.614 ±0.175

Table 6: Validation metrics of the deep features approach on full slices

Net(#feat)	Bal. acc	AUC	F1	C.M.
VGG (20)	0.824	0.877	0.741	$\begin{bmatrix} 22 & 5 \\ 2 & 10 \end{bmatrix}$

Table 7: Test metrics of the deep features approach on full slices

The selected features of the two models were then combined with the selected features in the radiomic model. *Table 8* reports the test results obtained for the two type of slices respectively.

Net(#feat)	Bal. acc	AUC	F1	C.M.
IncRes (11) *Cropped*	0.583	0.698	0.444	$\begin{bmatrix} 18 & 9 \\ 6 & 6 \end{bmatrix}$
VGG (21) ***Full***	0.787	0.907	0.690	$\begin{bmatrix} 20 & 7 \\ 2 & 10 \end{bmatrix}$

Table 8: Test metrics of the combined radiomics-deep features approach

Finally the ensemble models obtained by combining the 4 best models, one for each network

type, with and without the radiomic model, are shown in *Table 9*.

Type	Bal. acc	F1	C.M.
4 net *Cropped*	0.856	0.818	$\begin{bmatrix} 26 & 1 \\ 3 & 9 \end{bmatrix}$
4 net + rad *Cropped*	0.708	0.588	$\begin{bmatrix} 19 & 8 \\ 2 & 10 \end{bmatrix}$
4 net *Full*	0.727	0.621	$\begin{bmatrix} 27 & 0 \\ 7 & 5 \end{bmatrix}$
4 net + rad *Full*	0.796	0.727	$\begin{bmatrix} 25 & 2 \\ 4 & 8 \end{bmatrix}$

Table 9: Test metrics ensembles

4.2. 2.5D Results

2.5D results were obtained using the classifier trained in the 2D approach. A single validation fold has been used to choose the best aggregation method (*Table 10*).

Net(mode)	Bal. acc	AUC	F1	C.M.
Rad (Mean)	0.542	0.654	0.154	$\begin{bmatrix} 27 & 0 \\ 11 & 1 \end{bmatrix}$
IncRes (Max) *Cropped*	0.745	0.809	0.643	$\begin{bmatrix} 20 & 7 \\ 3 & 9 \end{bmatrix}$
VGG (Max) *Full*	0.880	0.849	0.833	$\begin{bmatrix} 25 & 2 \\ 2 & 10 \end{bmatrix}$

Table 10: Test metrics 2.5D

Finally, *Table 11* shows the results on the test set for the radiomic-deep features combinations, and *Table 12* shows the ensembles' metrics.

Net(#feat)	Bal. acc	AUC	F1	C.M.
IncRes (11) *Cropped*	0.616	0.741	0.514	$\begin{bmatrix} 13 & 14 \\ 3 & 9 \end{bmatrix}$
VGG (21) *Full*	0.898	0.889	0.870	$\begin{bmatrix} 26 & 1 \\ 2 & 10 \end{bmatrix}$

Table 11: Test metrics of the combined radiomics-deep features approach 2.5D

Type	Bal. acc	F1	C.M.
4 net *Cropped*	0.727	0.621	$\begin{bmatrix} 19 & 8 \\ 3 & 9 \end{bmatrix}$
4 net + rad *Cropped*	0.824	0.741	$\begin{bmatrix} 25 & 2 \\ 3 & 9 \end{bmatrix}$
4 net *Full*	0.838	0.783	$\begin{bmatrix} 22 & 5 \\ 2 & 10 \end{bmatrix}$
4 net + rad *Full*	0.856	0.818	$\begin{bmatrix} 26 & 1 \\ 3 & 9 \end{bmatrix}$

Table 12: Test metrics ensembles 2.5D

5. Discussion

This study highlights the advantages of deep learning over radiomics for classifying symptomatic and asymptomatic carotid plaques in CTA images. Radiomics achieved reasonable validation performance, with a balanced accuracy of 0.732, but showed significant limitations on the test set, where accuracy dropped to 0.532. This drop highlights radiomics' limitations in effectively generalizing to unseen data. In contrast, deep learning models like VGG and InceptionResNet consistently outperformed radiomics. The best-performing model, VGG in the 2.5D approach, reached a balanced accuracy of 0.880, a ROC AUC of 0.849, and a F1 score of 0.833. While 2.5D improved results for full images, its impact on cropped images was not as evident. By comparison, the 2D approach remains a simpler and faster alternative, achieving a balanced accuracy of 0.824, ROC AUC of 0.877 and F1 Score of 0.741 for VGG, surpassing the 2.5D model in terms of ROC AUC. Its strong performance highlights the diagnostic significance of the largest slice, which often encompasses most of the critical features required for classification. This is further supported by the fact that the chosen aggregation mode for 2.5D classification, *Max*, still relies on the prediction of a single slice. This makes the 2D method useful for scenarios prioritizing efficiency, as it reduces segmentation effort while maintaining competitive results. Radiomics, although not effective alone, demonstrated value as a complementary approach. When combined with deep learning, particularly in the 2.5D VGG model, radiomics achieved a balanced accuracy of 0.898, ROC AUC of 0.889 and F1 score of 0.870. Ensemble methods showed mixed results, with some combinations enhancing performance but placing greater emphasis on asymptomatic predictions, whereas in a clinical context, accurately identifying symptomatic cases is of greater importance. Full images consistently outperformed cropped images, which could be explained by the preservation of contextual information beyond the plaque itself. This also avoids the need for manual cropping, making the process quicker and easier in medical practice.
Our results are in line with findings in the literature, while introducing novel methods. As regards radiomics, most of the state-of-the-art studies focused on 3D analysis. Le et al. [8] used a single and multi-slice approach, but considered only CV and achieved a mean AUC of 0.67, lower than the one obtained by our method. Addition-

ally, to the best of the authors' knowledge, only end-to-end studies have been proposed to date, while machine learning models based on deep learning features remain unexplored. Moreover, our results mostly outperformed previous findings from end-to-end analyses. For example, He et al. [4] achieved a AUC of 0.854 on US images (compared to our best AUC of 0.889). Gui et al. [2] reached greater performance (AUC of 0.930) on HRMRI images, by augmenting the dataset 60 times. However, similar results on CTA images have not been reported so far.

Overall, this study introduces several innovations, including the application of deep learning to CTA images of carotid plaques, the combination of deep and radiomic features and the ensemble modeling, the integration of 2D and 2.5D methods, and a focus on patient-level predictions.

Future research should address the limitations identified in this study, including dataset size and model interpretability. Expanding the dataset would improve robustness and generalizability, while interpretability tools, such as heatmaps, could provide insights into model decisions, increasing clinical trust. Exploring 3D approaches and additional network architectures trained on medical-specific datasets like RadImageNet could further improve results.

6. Conclusions

This study demonstrates the effectiveness of combining deep learning and radiomics for carotid plaque classification. The 2D approach offers simplicity and efficiency, while the 2.5D method improves performance through multi-slice analysis, particularly for full images. The results are competitive with the state-of-the-art, highlighting the clinical relevance of the proposed framework. This work provides a valuable contribution to the classification of symptomatic and asymptomatic carotid plaques, with the potential to improve diagnostic accuracy and supporting decision-making.

References

[1] Sobreira Avelar Fingerhut Stein Guillaumon Geiger, Flumignan. Carotid plaque composition and the importance of non-invasive in imaging stroke prevention. *Front Cardiovasc Med.*, 2022.

[2] Zhang Zhang Ni Gui, Cao and Ming. Radiomics and artificial neural networks modelling for identification of high-risk carotid plaques. *Front Cardiovascular Medicine*, 2023.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Wang Chen Diao Wang Yuan Li Zhang He Shen He, Yang. A deep learning algorithm to identify carotid plaques and assess their stability. *Frontiers Artificial Intelligence*, 2024.

[5] Robson Marinelli Huang Doshi Jacobi Mei, Liu and Wan Greenspan Deyer Fayad Yang Cao Link, Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 2022.

[6] Therasse Robillard Giroux Arsenault Cloutier Soulez Naim, Douziech. Vulnerable atherosclerotic carotid plaque evaluation by ultrasound, computed tomography angiography, and magnetic resonance imaging: an overview. *Canadian Association of Radiologists*, 2014.

[7] World Stroke Organization. Impact of stroke. https://www.world-stroke.org.

[8] Pisu Benson Nardi Lanzino Suri Saba Scicolone, Vacca. Radiomics and artificial intelligence: General notions and applications in the carotid vulnerable plaque. *European Journal of Radiology*, 2024.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[10] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[12] J.J.M. Van Griethuysen et al. Computational radiomics system to decode the radiographic phenotype, 2017.