New CNN stacking model for classification of medical imaging modalities and anatomical organs on medical images

Mamar Khaled[a], Djamel Gaceb[a], Fayçal Touazi[a], Chakib Ammar Aouchiche[a], Youcef Bellouche[a], Ayoub Titoun[b]

[a] *LIMOSE laboratory, Computer science department, University M'hamed Bougara, Independence Avenue, 35000 Boumerdes, Algeria*
[b] *F2i Institute, School of Computer Science, Digital & Commerce, Vincennes, France*

Abstract

Decision making in medical diagnosis is tedious and very rigorous task, hence the requirement to use more advanced and intelligent medical imaging diagnostic support systems. The automation of the recognition of medical imaging modalities and human anatomical organs gives these systems the possibility of processing, in an automatic and adapted manner, different types of images in consideration of different medical imaging modalities. It also offers better support to clinicians and patients allowing them to access to more effective image analysis and diagnostic tools. In this context, three deep learning approaches were developed and tested on six different CNN models (VGG16, VGG19, ResNet-50, Xcpetion, Inception and NASNet). Two deep transfer learning modes and an ensemble deep learning algorithm based on stacking were used. The experiments carried out on two datasets of medium and high challenges show very interesting results with F-score reaching 99% for the classification of image modalities and 98% for the classification of anatomical organs.

Keywords 1

Anatomy organs, medical imaging modalities, deep transfer learning, ensemble deep learning, medical image processing, computer-aided diagnosis.

1. Introduction

The medical imaging field has undergone spectacular evolution in recent decades, offering unprecedented opportunities for the early and accurate diagnosis of various pathologies. However, manual or conventional interpretation of medical images and segmentation of anatomical organs or lesions remain complex and time-consuming tasks for radiologists and clinicians. In this context, the use of machine learning techniques, and in particular deep learning, has shown promise in improving the efficiency and accuracy of these processes. Today, the scientific community uses deep learning algorithms to improve diagnosis and help doctors in their work [1]. These algorithms offer more relevant automatic characterization and are capable of learning by developing broad knowledge on a large volume image datasets. The possibility of transferring learning in an incremental and scalable manner offers a great advantage to these algorithms in recognition, prediction or classification tasks with better precision. These properties are particularly interesting in the medical field, which is very demanding in terms of precision on datasets, which are often limited. By leveraging models already pre-trained on image classification tasks, we can capitalize on learned visual features to aid in the automatic identification of medical imaging modalities and anatomical organs.

A large number of deep learning methods use deep convolutional neural networks (CNN). They are successfully applied in medical image analysis, giving promising results. The application area

IDDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, November 17 - 19, 2023, Bratislava, Slovakia
EMAIL: m.khaled@univ-boumerdes.dz (A. 1); d.gaceb@univ-boumerdez.dz (A. 2), f.touazi@univ-boumerdez.dz (A. 3)
ORCID: 0000-0002-6178-0608 (A. 2); 0000-0001-5949-5421 (A. 3);

covers the entire spectrum of medical image analysis, including detection, segmentation, classification and computer-aided diagnosis [2].

The automatic classification of medical imaging modalities and anatomical organs will enable the development of medical diagnosis support systems and the appropriate automatic processing of a large corpus of images. It also facilitates the research work of doctors and healthcare professionals, by automatically knowing the image modality, clinicians can correctly interpret the image and make precise medical decisions. For example, it allows clinicians to quickly find the medical images they need to make comparisons, diagnoses, plan treatments and track the progression of diseases.

This work presents several original contributions by addressing the problem of recognition of medical imaging modalities and anatomical organs. A problem that is rarely addressed in the literature where public datasets are very rare. The main contributions of this work are as follows:

- construction of a new dataset with different challenges, created from several sources,
- comparisons of six existing CNN models with simple and complex architecture according to different transfer learning strategies (VGG 16, VGG 19, ResNet-50, Xcpetion, Inception and NASNet networks). We also studied the ability to transfer knowledge and features, learned initially on the basis of ImageNet dataset containing more than 14 million non-medical images to medical images constituting a set of images of relatively small size. Therefore, we chose to explore different levels of fine tuning allowing partial transfer of the learned features from ImageNet dataset to the images from our targeted dataset. This type of transfer can be very useful when deep neural networks are pre-trained on datasets very different from the subject area and when we have a very small and insufficient image dataset.
- Application of ensemble deep learning using the stacking of different CNN sub-models, for the combination of knowledge of different models and the pooling of their complementarity.

The rest of this paper is organized as follows: Section 2 presents related works on classifications of medical image modalities and anatomical organs using deep learning. In the third section, we present the different proposed approaches. The experimental results are presented and discussed in the fourth section.

2. Related works

2.1. Overview of existing deep learning-based approaches for classification of medical imaging modalities

The classification of medical imaging modalities is a very important preliminary step to bring more autonomy to intelligent medical diagnostic support systems and to help clinicians access the required medical imaging in the system. Among the existing works that focus on deep learning, we find the work of Yu et al. [3, 4], which focuses on the combination of two CNN architectures (VGG16 and ResNet-50), already pre-trained on ImageNet dataset, using deep transfer learning and a voting system. The experiments, which were carried out on two medical datasets (ImageCLEF2015 and ImageCLEF2016), showed that the proposed combination approach offers the best accuracy (90.22% on the ImageCLEF2015 dataset and 88.40 on the ImageCLEF2016 dataset) in comparison with the VGG16 architectures (87.27% on ImageCLEF2015 and 85.13 on ImageCLEF2016) and ResNet-50 (89.34% on ImageCLEF2015 and 87.47 on ImageCLEF2016).

Kim et al. [5] developed a new method called Class-selective Relevance Mapping (CRM), to locate and visualize RoI (regions of interest) in a medical image in order to improve the predictions of CNN models for medical imaging classification. In addition to this model, a pre-trained VGG16 was used to classify seven different types of image modalities. An accuracy of 0.98% is obtained on the Access Biomedical Image Search engine dataset from the United States National Library of Medicine (NLM) and on the ImageCLEF2013 dataset.

In another study, Remedios et al. [6], explored a CNN architecture known as Φ-Net to classify MRI images into different categories according to the acquisition modality (T1, T2, FLAIR and subclasses T1 pre, T1 post, FLAIR pre and FLAIR post). This model was created by combining several CNN architectures with the concept of residual learning [7]. The experiments carried out on a

dataset of 3418 MRI images showed that the Φ-Net model had an average accuracy of 97.57% for classification (T1, T2, FLAIR), compared to 95.47% obtained by the ResNet architecture.

Chiang et al. [8] used a CNN model for the classification of 4 classes of medical imaging (CT of the abdomen, CT of the brain, MRI of the brain and MRI of the spine). The experiments carried out on the dataset from the Taiwanese Shuang-Ho hospital (700 images per class) showed an accuracy of 99.5% and an F-score of 99%, in the same direction we find the work of Laribi and al. [9], where the authors developed a new progressive deep transfer learning approach to diagnose Alzheimer Disease, applied on the same dataset (Brain MRI dataset), and they achieved best results.

Recent work of Atrey and al. [10], who developed a hybrid deep learning bimodal CAD algorithm for the classification of breast lesions using mammogram and ultrasound imaging modalities combined, A combined CNN and LSTM model was implemented using different images obtained from both mammogram and ultrasound modalities to improve the early diagnosis of Breast Cancer. The proposed bimodal approach achieved a 99.35% of accuracy for the classification.

According to the literature, we notice the intensive need to explore new ways and approaches based on the transfer learning and on the ensemble deep learning in order to achieve better results in medical diagnostic support systems and especially in the context of classification of medical imaging modalities as well as for the classification of anatomical organs which represents the object of our study.

2.2. Overview of existing deep learning-based approaches for anatomical organs classification

Automated classification of anatomical organs is an important step and a prerequisite for many medical diagnostic support systems. Spatial complexity and variability of anatomy throughout the human body make classification difficult. In the literature we can find the review of Jiang and al. [11], where they reviewed in-depth and analyzed some deep learning-based methods utilized in multiple-lesion recognition, they were interested to the multiple-lesion recognition in diverse body areas and recognition of whole-body multiple diseases. Holger and al. [12] trained a CNN model to identify anatomical organs (neck, lungs, liver, pelvis and legs) on axial tomography images. An accuracy of 0.998% was achieved on images from Hospital PACS Dataset.

Takiyama et al. [7] worked on the classification of endoscopic (esophagogastroduodenal) medical images to recognize the locations of anatomical organs. Images were categorized into four anatomical locations (larynx, esophagus, duodenum, and stomach) and three additional sublocations of the stomach (upper, middle, and lower), allowing for accurate anatomical classification of the images. The experiments were carried out on a dataset of 27,335 endoscopic gastroesophageal (EGD) images from a Japanese hospital. An accuracy of 97% was achieved using the GoogleNet model. In the study done by Kolbinger and al. [13], We see the combination of two well-known methods (DeepLabv3 and SegFormer) on a new dataset of 13195 laparoscopic images, in the aim to develop segmentation models for the anatomical structures, they concluded that ML methods can improve the assistance in anatomy recognition. Khan et al. [14] proposed a new CNN architecture (compared to three existing CNN architectures: LeNet, AlexNet and GoogLeNet) for the classification of images of different parts of the human body (head, neck, thorax, abdomen, pelvis, upper and lower limbs) coming from different medical imaging modalities, including CT, MRI, PET, ultrasound and X-rays. The proposed architecture gave a Test Accuracy rate of 81%, the best rate in comparison with three existing CNN architectures (LeNet 59%, AlexNet 74% and GoogLeNet 45%) on a dataset of 37,198 images of various anatomical organs. This work shows the interest of existing work in more powerful CNN architectures.

Deep learning-based approaches have yielded promising results in the classification of anatomical organs. However, they often require large amounts of data, which can be difficult to obtain in the medical field.

3. Proposed approaches

As part of this work, we present different approaches based on deep transfer learning that we have developed for the classification of medical imaging modalities and anatomical organs. Such classifications present several challenges that should not be overlooked during development. The main difficulties are often posed by the intra-class variability of medical images, the diversity of imaging modalities used, the complexity of anatomical structures and the unavailability of datasets of sufficient size in the medical field. The use of deep transfer learning is a better choice to design a more robust approach to these constraints. This involves the use of pre-trained CNN models on generic image datasets of sufficient volume, to benefit from representations that focus on generic and low-level image features, learned on massive and diverse data. These models are thus re-trained (on our small dataset) and refined by seeking the best level of fine tuning, making it possible to complete the initial low-level representation, valid for all kinds of images, with a second high-level representation, specific to our problem of classification of image modalities and anatomical organs.

In this context, we chose to develop six very popular CNN models (VGG16, VGG19, ResNet-50, Inception, Xception and NASNet) with performance already demonstrated in the medical field. Based on our previous study on CNN models combination [15] and it's benefits, we also developed an ensemble deep learning which involves combining several CNNs to take advantage of their complementarities. With the stacking mechanism, we propose to use a softmax meta-model which learns the best weighting and combination of these sub-models.

3.1. Proposed approaches for the classification of medical imaging modalities

At this level, we have developed three different approaches to determine the best behavior to follow, the first two approaches concern the development of six CNN models using two transfer learning modes (features extractor and fine-tuning modes). Transfer learning makes it possible to solve the problem of the reduced size of a dataset. It consists of reusing a pre-trained model on another large dataset (even outside the medical field), preserving part of it for relevant extraction of generic characteristics and fine tuning the remaining part on our small target dataset to extract specific characteristics. This approach allows for faster learning and a more reliable model from a very small dataset.

The third approach consists of seeking the best combination of different CNN models with the stacking technique in order to take advantage of the complementarity between them. This combination has the advantage of being able to aggregate very different classifiers and significantly improve the quality of the final prediction. The use of ensemble deep learning methods is necessary when we want to take a step forward in obtaining better prediction results of medical imaging modalities.

- **Approach 1:** is based on transfer learning in features extractor mode. The convolutional part (features extractor) of the pre-trained CNN is completely frozen in order to preserve all the knowledge already acquired on the initial (very large) dataset. With this mode, only the classifier part (Softmax) will be adapted to the new image modalities classification task. The use of pre-trained CNN models, aims to extract high-level characteristics from medical images. Then, these features were used to train modality-specific classifiers. This approach allows us to benefit from knowledge learned from large generic databases. Transfer learning with this mode is faster than that based on fine tuning, however, it requires the presence of certain similarities between the original dataset images and the target dataset images. Six CNNs are compared using this approach (see Figure 1).

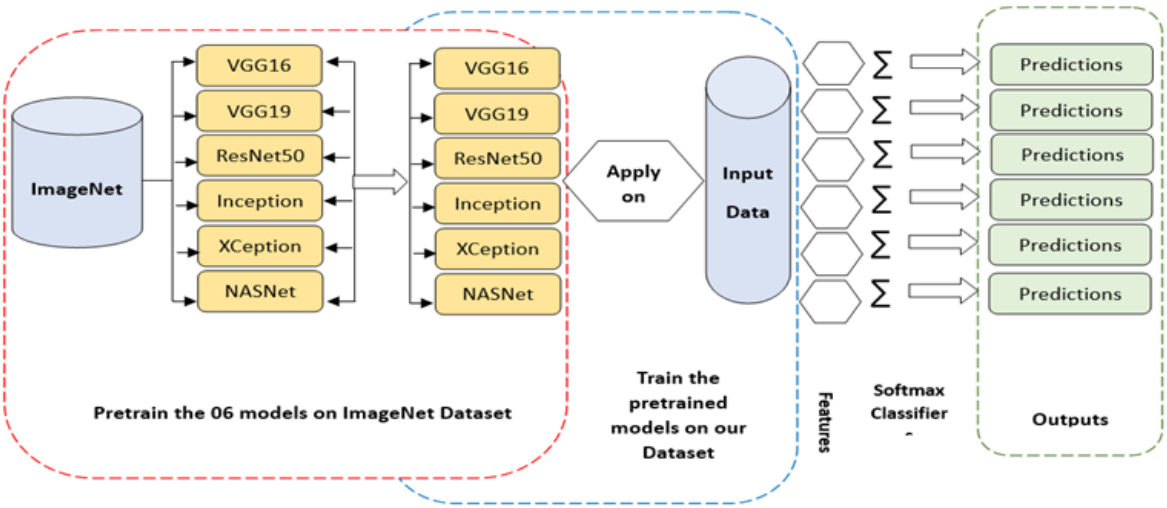

Figure 1: Approach 1 architecture

Approach 2: is based on transfer learning in fine-tuning mode. In this mode, a pre-trained model is used as a starting point, but unlike the features extractor mode, a set of the last layers of the convolutional part of the model are fine-tuned during training on the new dataset, specific to medical imaging. The layers which are not refined are frozen (this concerns the layers closest to the input) to preserve certain generic knowledge of the pre-trained CNN model which has already learned on a large dataset to extract low level features (which concerns all kinds of images). For the convolution part, the number of frozen blocks must be fixed empirically in order to have a better score. Since the number of classes is different in the target dataset compared to that of the original dataset, the structure of the classifier part (Full connected layers) must be adapted to recognize the new classes of different image modalities. This method is composed of three steps, the first one concern the fine-tuning of the parameters of the non-fixed layers, secondly extraction of image features, finally, the last step consists of the generation of predications for the classification using the Softmax classifier.

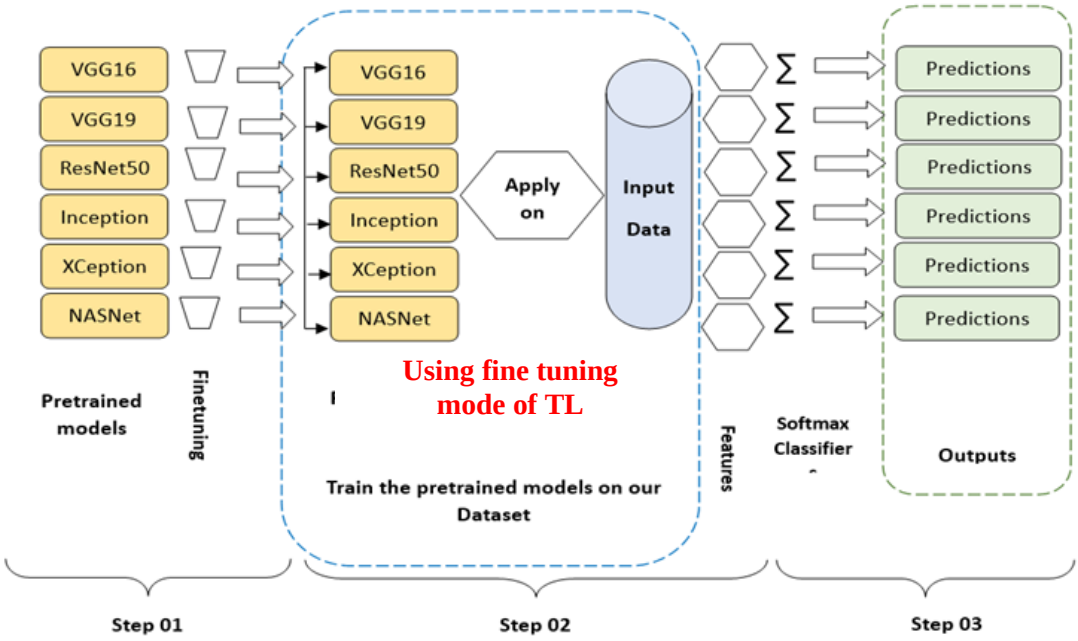

Figure 2: Approach 2 architecture

- **Approach 3:** is based on the combining mode of different CNN models with the stacking technique. In this mode, several pre-trained models are used as a starting point, we train the models on a database which only contains MRI images of different types (ex. Flair, T1w, T1wCE and T2w), we will thus obtain trained models, then we combined them with stacking.

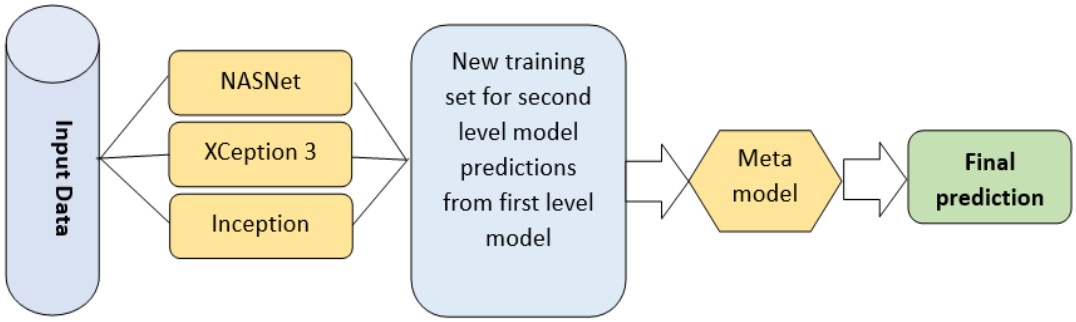

Figure 3: Synoptic diagram of approach 3, based on stacking of CNNs

Stacking process consists of training several CNN models independently on the same dataset. Each sub-model may have a different architecture with different settings. Once the sub-models are trained individually, a meta-model is added to the output of these models. It receives predictions from different CNNs as input and learns to combine these predictions to produce a final prediction. This meta-model can be based on any machine learning model. In our method, a softmax classifier is used as a stacking meta-model. It receives as input the different prediction probabilities coming from the output of different stacked CNN sub-models. The objective of this step is to train a new model (softmax) to learn how to best combine the contributions from each CNN sub-model.

Stacking process allows you to take advantage of the diversity of individual models by combining their strengths and mitigating their weaknesses. This approach can result in better predictive performance than any single contributing model.

It is also important to note that it is possible to distribute this approach in a real-time framework and make the CNN sub-models work in multitasking, multiprocessors, multicores or parallelism. This allows better management of the complexity generated by the stacking of sub-models.

3.2. Anatomical organs classification

For the classification of anatomical organs, we will use the transfer learning in feature extractor mode. We used it by the same idea as that of the approach 1 to the classification of medical imaging modalities.

4. Experiments and results

In this section, we will present the experiments that we carried out as part of our study, as well as the results obtained. First, we will present the evaluation metrics used. Then we will present in detail the datasets used with samples of each of them. After that, a description of the data augmentation technique used. Then, the results of our experiments for each approach illustrated in tables followed by comparisons and comments.

4.1. Evaluation metrics

In the literature, there are several evaluation metrics, in our case, and to evaluate the different proposed approaches, we used the following metrics: Accuracy, Precision, Recall and F1-score.

- **Confusion matrix:** Confusion matrix or error matrix is one of the key concepts when we talk about classification problems. This matrix is a two-dimensional array (“actual” and “predicted”)

and sets of "classes" in both dimensions. Our actual classifications are columns and the predicted ones are rows as shown in the table below:

Table 1
Confusion Matrix

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	TP	FP
	Negative (0)	FN	TN

Almost all performance measures are based on the confusion matrix and the numbers it contains.

- True positive (TP): Real class = True and the prediction=True.
- True negative (TN): Real class = False and the prediction = False.
- False positive (FP): Real class =False and prediction =True.
- False negative (FN): Real class = True and the prediction=False.

- **Accuracy:** Number of correct predictions divided by the total number of samples. Is a good measure when the classes of target variables in the data are almost balanced.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ samples} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

- **Precision:**

$$P = \frac{TP}{TP + FP} \quad (2)$$

- **Recall:**

$$R = \frac{TP}{TP + FN} \quad (3)$$

- **F1-score:**

$$F1 - Score = \frac{2 \times PR}{P + R} \quad (4)$$

The difference between Precision and Recall in the classification problem is that Recall gives us information about the performance of a classifier against false negatives (how many did we miss), while precision gives us information about its performance versus false positives (how many did we catch).

4.2. Datasets used

Our experiments were carried out on two different datasets with different sizes and challenges.

- **First dataset: called MC4_Dataset (Size and average challenges):**

This image dataset was created by us from nine public image datasets (size of 35150 images). Then, we divided this dataset into 4 classes depending on the imaging modality used (MRI 7023 images, ultrasound 2116 images, CT-Scan 12988 images, X-ray: 12023 images). Each class is divided into subclasses according to the different anatomical organs (see Table 2). This dataset is created from different datasets and image sources presenting different degradations, resolutions, complexities, etc. This increases the challenges of our dataset which will be confronted with the developed models.

Table 2
List of public datasets combined for each medical imaging modality

IRM 7023 images	Ultrasound 2116 images	CT-Scan 12988 images	X-ray 12023 images
- Public Dataset "Brain Tumor MRI Dataset" [16] 7023 Images	Combination of two datasets: 1) 780 medical images of women's breasts.[17] 2)1336 ultrasound images of the fetal head [18]	Combination of two datasets: 1) 988 images of the torso [19]. 2) 12,000 medical images of the kidneys [20].	Combination of three datasets: 1) Chest X-ray Dataset: 5856 images [21]. 2) Chest Xray Masks and Labels: 1600 images of human torsos [21]. 3) Figure-detection: 5567 hand images [21]

This table contains an unbalanced dataset, ex. the Ultrasound subset (2116 images) is a minority class compared to the CT-scan class (12988 images). The significant difference between the sizes of the classes can destabilize the model, limit its generalization and/or cause overfitting. To reduce these problems, we balanced the dataset by oversampling each minority class to give it the same weight as the majority class, using data augmentation: image rotation, zooming, shifting, scaling and shearing (see section 4.3).

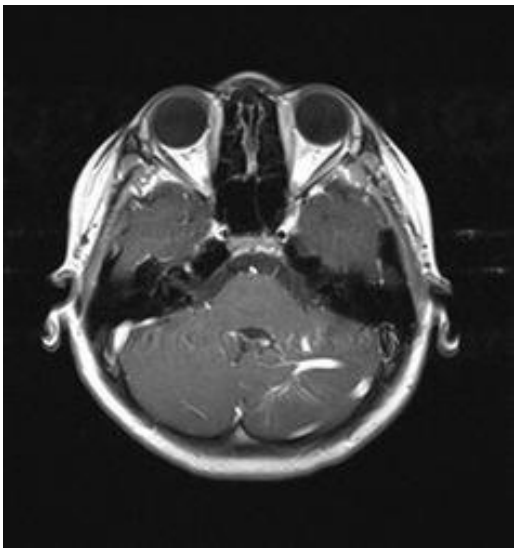

Figure 4: Example of MRI images [16]

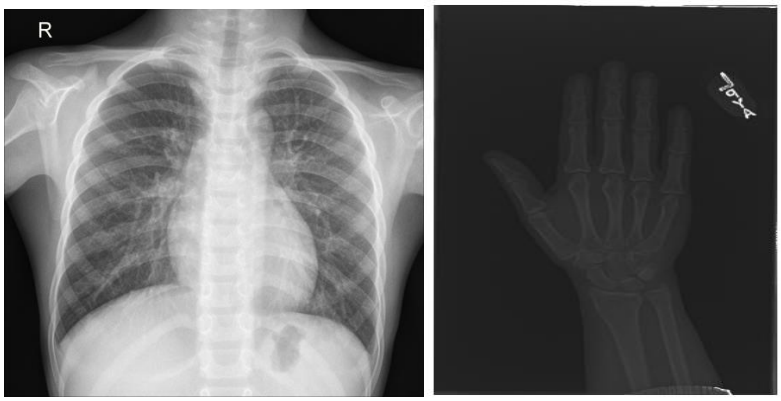

Figure 5: Example of X-ray images, a) medical image of the torso, b) medical image of the hand [21]

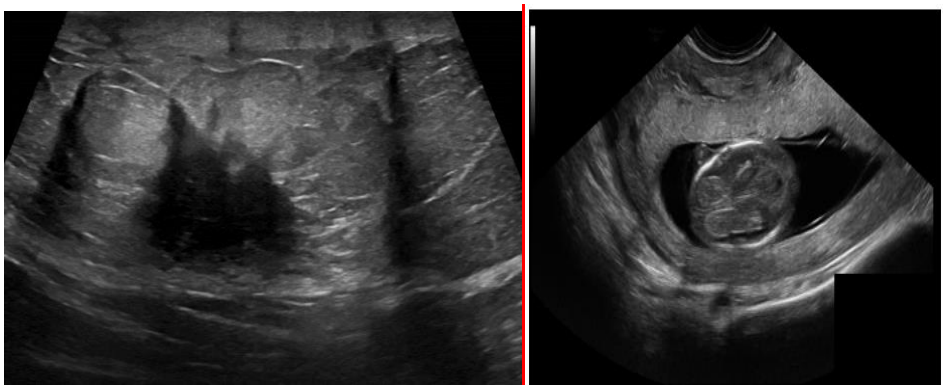

Figure 6: Example of ultrasound images a) medical image of a woman's chest, b) medical image of the fetal head by ultrasound [18]

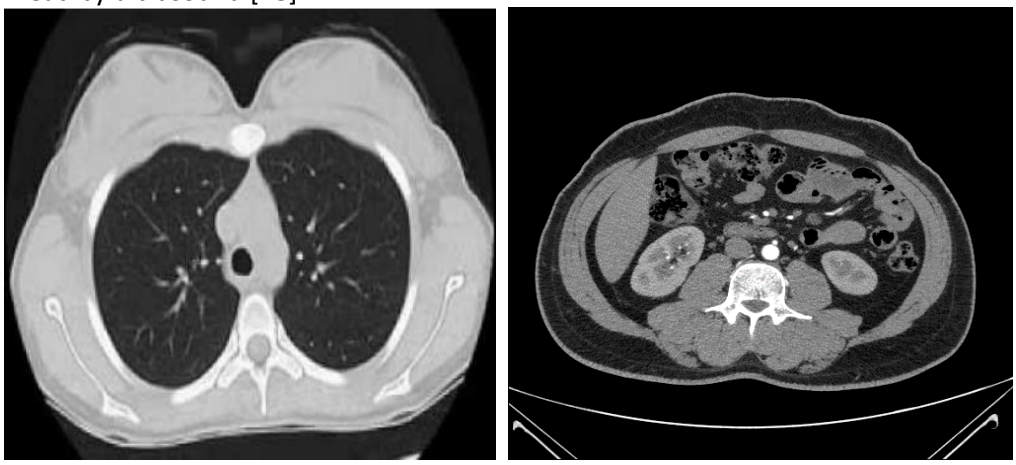

Figure 7: Example of CT-Scan image a) Medical image of the torso, b) medical image of the kidney acquired by CT scan [19].

- **Second RSNA-MICCAI dataset (high size and challenges)**

The second used dataset, presents more challenges and more diversity compared to the first dataset. This is the public dataset "RSNA-MICCAI Brain Tumor Radio-genomic Classification" (size = 290923 images) [22]. It is used to test CNN combination in approach 3. The latter is divided into four classes of modalities (FLAIR, T1w, T1wCE and T2w). The following table summarizes the image distribution in this dataset according to the different classes. This dataset is balanced like the first dataset using data augmentation.

Table 3
RSNA-MICCAI Brain Tumor Radio-genomic Classification dataset table [22].

Dataset	Repartition	Class	Number of images	Size of resized images
RSNA MICCAI PNG	Train	FLAIR	50682	224x224x3
		T1w	55440	
		T1wCE	68012	
		T2w	67722	
		Total		241856
	Valid	FLAIR	3000	224x224x3
		T1w	3000	
		T1wCE	3000	
		T2w	3000	
		Total	12000	

	Test	FLAIR	7926	224x224x3
		T1w	8791	
		T1wCE	10482	
		T2w	9863	
		Total	37067	
Total			290923	

4.3. Data Augmentation

To improve deep learning on small datasets, there is a technique called data augmentation. It consists of creating new training data samples by applying simple transformations to existing data, which increases the size of the available dataset in the aim to improve model training. It has several advantages for deep learning, it allows generalization of models, by exposing the model to a variety of transformations, it becomes more resilient to variations such. The transformations used in our study for the data augmentation are rotation, zooming, shifting, scaling, cropping, etc. By using data augmentation, it is possible to improve the performance of deep learning models in an extraordinary way. Data augmentation makes it possible to avoid overfitting on complex CNN architectures (with a large number of parameters) applied to a small dataset. It also makes it possible to offer better stability and model generalization on an unbalanced dataset like the MC4_Dataset.

4.4. Experiments and results: Classification of medical imaging modalities

In order to study the relevance of deep learning for the classification of medical imaging modalities, we tested the three proposed approaches using different CNN models with transfer learning on our datasets.

4.4.1. Results of approach 1: Features extractor mode

This approach was applied on each of six CNN models (VGG16, VGG19, ResNet-50, Inception 3, Xception and NASNet) already pre-trained on the ImageNet dataset. The dataset used is "MC4_Dataset" with data augmentation technique to overcome the problem of the small size in this dataset. The features extraction mode is applied with the adaptation of the Softmax classifier part to classification in 4 modalities (MRI, ultrasound, CT-scan and X-ray).

The following table summarizes the results obtained by the six CNNs:

Table 4

Comparative table of the results of different models in features extraction mode for the classification of modalities on the first MC4_Dataset dataset.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.96	0.96	0.96	0.96
VGG19	0.94	0.95	0.94	0.95
ResNet-50	0.81	0.82	0.80	0.81
Inception	0.97	0.97	0.97	0.97
Xception	0.976	0.976	0.976	0.976
NASNet	**0.99**	**0.99**	**0.99**	**0.99**

According to this table, we see the effectiveness of the NASNet model compared to all the other models with a value of 0.99 for each of the evaluation metrics, then comes the Xception model with values less than the NASNet but very convincing values.

4.4.2. Results of approach 2: Fine Tuning mode

The goal of this experiment is to determine whether the performance of the first five CNNs (VGG 16, VGG 19, ResNet50, Inception 3, XCeption) will be enhanced by the fine-tuning mode and make them also competitive compared to the NASNet CNN.

Still working on the MC4_Dataset dataset with four output classes (MRI, ultrasound, CT-Scan and x-ray), we trained these five CNN models (already pre-trained on ImageNet) in fine-tuning mode on our medical image dataset. Each model is trained in 5 epochs with a batch-size of 100, where we have frozen a certain percentage of layers (the leftmost layers which are closer to the initial image), this process can be described as follows:

- **VGG16:** By default, this model contains 5 convolution blocks, but in our case, we generated two models from the base model, the first was fixed at 80% (4 blocks out of 5) and the second was fixed at 60 % (3 blocks out of 5).
- **VGG19:** In the same way as the VGG16 model, we generated two models from this model where the first was fixed at 80% and the second one at only 60%.
- **ResNet-50:** This model contains 50 layers, while we generated only one single model from it, with a percentage of 80%.
- **Inception 3:** We thus generated two models from this model which contains 48 layers, the first one was fixed at 90% (43 layers) and the other at 80% (38 layers).
- **XCeption:** Based on its number of layers which is 71 layers, we thus generated two models from the base model, the first one we frozen 90% (64 layers) of its layers and the other at 80% (57 layers).

The obtained results are displayed in this table:

Table 5
Comparison of the results of the different models in fine-tuning mode for the classification of image modalities.

Model	Accuracy	Precision	Recall	F1-Score
VGG16 (60%)	0.42	0.37	0.47	0.41
VGG16 (80%)	0.94	0.95	0.94	0.94
VGG19 (60%)	0.46	0.46	0.46	0.46
VGG19 (80%)	0.96	0.96	0.96	0.96
Resnet 50 (80%)	0.84	0.87	0.82	0.84
Inception 3 (80%)	0.99	0.99	0.99	0.99
Inception 3 (90%)	0.99	0.99	0.99	0.99
XCeption (80%)	0.99	0.99	0.99	0.99
NASNet (100%) Features extraction mode	**0.99**	**0.99**	**0.99**	**0.99**

According to these results, we can see the effectiveness of the fine-tuning approach, and especially for the two models Inception 3 and XCeption which gave very high values in terms of accuracy, precision, recall and F1-Score (0.99) and which are also competitive with the NASNet architecture.

4.4.3. Results of approach 3: Stacking mode

Given that the performance of the CNNs (approaches 1 and 2) on the MC4_Dataset dataset (average challenges) was 0.99, we considered that it was unnecessary to use the third approach (combination of CNNs) which was dedicated on all to higher challenge datasets. This is why we tested approach 3 only on the second dataset ("RSNA-MICCAI Brain" with 4 classes Flair, T1w, T1wCE, T2w) which presents greater diversity and difficulties.

Initially we tested each of the six CNNs separately (using approach 2). Subsequently we combined three best architectures (Inception 3, XCeption and NASNet), relying on Stacking mode in order to achieve a more efficient model with very promising and convincing results which are displayed in the following table:

Table 6

Comparison of the results of the different models in fine-tuning mode for the classification of modalities and combination of models.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.58	0.75	0.24	0.37
VGG19	0.50	0.73	0.24	0.36
Resnet 50	0.63	0.70	0.41	0.51
Inception 3	0.84	0.89	0.81	0.85
XCeption	0.88	0.89	0.87	0.88
NASNet	0.86	0.91	0.82	0.87
Stacking Model	0.91	0.95	0.89	0.92

As we see in the table, the combination of models is efficient and gave us better results comparing to other models separately.

4.5. Experiments and results: Classification of anatomical organs

For the classification of anatomical organs, we tested on the six CNN models, which are already pre-trained based on ImageNet dataset. The transfer learning technique was adopted in features extraction mode in order to increase the learning results, and to take advantage of the power of the ImageNet dataset. In this case, the MC4_Dataset dataset is subdivided into 8 classes of anatomical organs: human torso acquired by scanner, human kidney acquired by scanner, brain acquired by MRI, torso acquired by MRI, female chest acquired by ultrasound, fetal head acquired by ultrasound and finally of the torso and hand both acquired by x-ray.

Table 7

Comparative table of the results of the different models in features extraction mode for the classification of anatomical organs

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.94	0.94	0.93	0.95
VGG19	0.64	0.51	0.38	0.78
Resnet 50	0.66	0.54	0.39	0.89
Inception 3	0.98	0.99	0.96	0.97
XCeption	0.98	0.99	0.97	0.98
NASNet	0.98	0.99	0.96	0.98

The obtained results mentioned in this table, showing the effectiveness of the three models (NASNet, Inception 3 and XCeption) well recognized in the literature, giving very convincing values.

4.6. Discussion

In order to select the best architecture among the six architectures seen precisely, we adopted the F1-Score as the best performance comparison criterion, due to the costs of false positives and false negatives which differ in number, which leads to obtaining additional false positives (false alerts) rather than saving false negatives.

In the first approach, we saw that the NASNet model outperformed the other models with an F1-score of 0.99, but it should be noted that the NASNet in features extraction mode did not have a high challenge dataset.

In the second approach, we noticed that the results increased or decreased, depending on the number of layers frozen by the fine-tuning mode. The XCeption architecture beat the other models in fine-tuning by 80% where it improved its F1-score compared to the first approach in features extraction mode, this shows that the choice of fine-tuning levels had a significant impact to overcome the problem of the small size of the image database.

The third approach which serves to combine the three CNN architectures (Inception 3, XCeption and NASNet) where the choice was justified by the best results obtained by these models on the basis of medical imaging of different types of MRI, this approach gave us very good results on a very high challenge dataset. The model that combines the three models improved the F1-score of the best performing model (XCeption) by 5%. This leads us to the conclusion that the deep learning stacking approach is very powerful on high challenge datasets.

For the classification of anatomical organs, we noted the effectiveness of the two models (XCeption and NASNet) with the softmax classifier which had a score of 0.98. Considering the sufficient results, the application of approach three was not necessary.

5. Conclusion

This work focuses on the classification of medical imaging modalities and the classification of anatomical organs. For this, six CNN architectures were tested and compared according to three different approaches, which we proposed. The objective was to explore deep transfer learning in two modes (features extraction and fine-tuning) and ensemble deep learning using the stacking technique which combines and complements several models (the best) CNNs. The experiments were carried out on two datasets with different challenges: unbalanced MC4_Dataset (created from nine existing datasets, medium size and challenges) and RSNA-MICCAI Brain (very high size and challenges). The experimental results showed that the NASNet architecture is very powerful compared to the other five models on small or medium challenge datasets. Its performance on challenges datasets with larger sizes is significantly increased when using combinations with other models.

Overall, our approach represents a significant advancement in the classification of medical image modalities and anatomical organs via the use of deep transfer learning. These results open new perspectives for the automation and improvement of medical image analysis tools, thus contributing to the improvement of healthcare and medical decision-making.

In future work, we plan to test stacking on higher challenge datasets by combining CNN models with ViT model.

6. References

[1] P.K. Mall, P.K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, M. Ganzha, A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities, Healthcare Analytics 4 (2023) 100216.

[2] A.S. Panayides, A. Amini, N.D. Filipovic, A. Sharma, S.A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, K. Huang, K.S. Nikita, B.P. Veasey, M. Zervakis, J.H. Saltz, C.S. Pattichis, AI in Medical Imaging Informatics: Current Challenges and Future Directions, IEEE journal of biomedical and health informatics 24(7) (2020) 1837-1857.

[3] F. Sica, G. Gobbi, P. Rizzoli, L. Bruzzone, F-Net: Deep Residual Learning for InSAR Parameters Estimation, IEEE Transactions on Geoscience and Remote Sensing 59(5) (2021) 3917-3941.
[4] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, Z. Zhao, Deep Transfer Learning for Modality Classification of Medical Images, Information, 2017.
[5] I. Kim, S. Rajaraman, S. Antani, Visual Interpretation of Convolutional Neural Network Predictions in Classifying Medical Image Modalities, Diagnostics (Basel, Switzerland) 9(2) (2019).
[6] S. Remedios, D.L. Pham, J.A. Butman, S. Roy, Classifying magnetic resonance image modalities with convolutional neural networks, Medical Imaging 2018: Computer-Aided Diagnosis, SPIE, 2018, pp. 558-563.
[7] H. Takiyama, T. Ozawa, S. Ishihara, M. Fujishiro, S. Shichijo, S. Nomura, M. Miura, T. Tada, Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks, Scientific reports 8(1) (2018) 7497.
[8] C.H. Chiang, C.L. Weng, H.W. Chiu, Automatic classification of medical image modality and anatomical location using convolutional neural network, PloS one 16(6) (2021) e0253205.
[9] N. Laribi, D. Gaceb, A. Benmira, S. Bakiri, A. Tadrist, A. Rezoug, A. Titoun, F. Touazi, A Progressive Deep Transfer Learning for the Diagnosis of Alzheimer's Disease on Brain MRI Images, in: M. Salem, J.J. Merelo, P. Siarry, R. Bachir Bouiadjra, M. Debakla, F. Debbat (Eds.) Artificial Intelligence: Theories and Applications, Springer Nature Switzerland, Cham, 2023, pp. 65-78.
[10] K. Atrey, B.K. Singh, N.K. Bodhey, R. Bilas Pachori, Mammography and ultrasound based dual modality classification of breast cancer using a hybrid deep learning approach, Biomedical Signal Processing and Control 86 (2023) 104919.
[11] H. Jiang, Z. Diao, T. Shi, Y. Zhou, F. Wang, W. Hu, X. Zhu, S. Luo, G. Tong, Y.-D. Yao, A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation, Computers in Biology and Medicine 157 (2023) 106726.
[12] H.R. Roth, C.T. Lee, H.-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, R.M.J.I.t.I.S.o.B.I. Summers, Anatomy-specific classification of medical images using deep convolutional nets, (2015) 101-104.
[13] F.R. Kolbinger, F.M. Rinner, A.C. Jenke, M. Carstens, S. Krell, S. Leger, M. Distler, J. Weitz, S. Speidel, S. Bodenstedt, Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise – an experimental study, International Journal of Surgery (9900).
[14] S. Khan, S.P. Yong, A deep learning architecture for classifying medical images of anatomy object, 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 1661-1668.
[15] M. Khaled, D. Gaceb, F. Touazi, A. Otsmane, F. Boutoutaou, Progressive and Combined Deep Transfer Learning for pneumonia diagnosis in chest X-ray images. IDDM'2022: 5th International Conference on Informatics & Data-Driven Medicine, 2022, pp. 160-173.
[16] M. Nickparvar, Brain Tumor MRI Dataset, 2021. https://www.kaggle.com/dsv/2645886, https://doi.org/10.34740/KAGGLE/DSV/2645886
[17] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data in Brief 28 (2020) 104863.
[18] Fetal Head UltraSound Dataset For Image Segment, 2023. https://www.kaggle.com/datasets/ankit8467/fetal-head-ultrasound-dataset-for-image-segment.
[19] Chest CT-Scan images Dataset, 2020. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.
[20] MD NAZMUL ISLAM & Md Humaion Kabir Mehedi. (2021). CT KIDNEY Dataset, https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone.
[21] L. Rubini, Soundarapandian,P., and Eswaran,P.. Chronic_Kidney_Disease, 2015.

[22] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M.Y.L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A.N. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M.A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning, Cell 172(5) (2018) 1122-1131.e9.

[23] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F.C. Kitamura, S.J.a.p.a. Pati, The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, (2021).