UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

PH.D. THESIS
IN
INFORMATION AND COMMUNICATION TECHNOLOGY FOR HEALTH

ARTIFICIAL INTELLIGENCE AND MEDICAL IMAGING: HANDLING AND MINING MULTIPLE SOURCES

MICHELA GRAVINA

TUTOR: PROF. CARLO SANSONE

COORDINATOR: PROF. DANIELE RICCIO

XXXV CICLO

SCUOLA POLITECNICA E DELLE SCIENZE DI BASE
DIPARTIMENTO DI INGEGNERIA ELETTRICA E TECNOLOGIE DELL'INFORMAZIONE

A me stessa. Alla mia forte determinazione e sfrenata passione per la ricerca. Che io non possa mai perdere la forza e il coraggio di lottare per i miei sogni, indipendentemente dagli sforzi necessari e dagli ostacoli da affrontare.

ARTIFICIAL INTELLIGENCE AND MEDICAL IMAGING: HANDLING AND MINING MULTIPLE SOURCES

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information and Communication Technology for Health
by

MICHELA GRAVINA

October 2022

Approved as to style and content by

Prof. Carlo Sansone, Advisor

Università degli Studi di Napoli Federico II
Ph.D. Program in Information and Communication Technology for Health
XXXV cycle - Chairman: Prof. Daniele Riccio

https://icth.dieti.unina.it

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information and Communication Technology for Health is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 13, 2022

Michela Gravina

Abstract

Medical image computing refers to the process of extracting relevant information from medical images for the characterization of the area under analysis. The large amount of information to consider, and the high complexity of medical images, which make the manual inspection a very tedious and hard task, have prompted research into proposing solutions for the automatic analysis of radiological acquisitions. More recently, Artificial Intelligence (AI), and in particular Machine Learning (ML) and Deep Learning (DL), had a radical spread in medical image computing with surprising results. Moreover, the use of deep neural networks has also enabled the development of DL-based solutions in medical applications characterized by the need of leveraging information coming from *multimodal data sources*, raising the Multimodal Deep Learning (MDL). However, in healthcare, it is very difficult to obtain high-quality, balanced datasets with labels due to privacy and sharing policy issues. Several applications have leveraged DL approaches in data augmentation techniques, proposing models that are able to create new realistic and synthetic samples. As a consequence, it is possible to identify a new source of data, that is denoted as *synthetic data source*. The aim of this thesis is to investigate the DL approaches in medical image computing, considering the presence of multiple data sources. In the case of *multimodal data sources*, a systematic analysis of multimodal data fusion techniques is performed introducing an innovative transfer module that allows the different modalities to influence each other, while in the analysis of *synthetic data source*, a DL-based data augmentation method is proposed that exploits the biological characteristics of the images implementing a physiologically-aware synthetic image generation process.

Keywords: Artificial Intelligence, Medical Image Computing, Deep Learning, Multimodal Learning, Image Synthesis

Sintesi in lingua italiana

L'elaborazione delle immagini mediche si riferisce al processo di estrazione di informazioni rilevanti per la caratterizzazione dell'area in analisi. La grande quantità di dati da considerare e l'elevata complessità delle immagini mediche, che rendono l'ispezione manuale un compito molto complesso, hanno spinto la ricerca a proporre soluzioni per l'analisi automatica delle acquisizioni radiologiche. Recentemente, l'Intelligenza Artificiale (AI), e in particolare il Machine Learning (ML) e il Deep Learning (DL), hanno avuto una diffusione radicale nell'elaborazione delle immagini mediche con risultati sorprendenti. Inoltre, l'uso di reti neurali profonde ha permesso lo sviluppo di soluzioni basate sul DL in applicazioni che sfruttano informazioni provenienti da *fonti di dati multimodali*, dando vita al Multimodal Deep Learning (MDL). Tuttavia, nel settore sanitario è molto difficile ottenere insiemi di dati bilanciati, di alta qualità e etichettati, a causa di problemi di privacy e di politica di condivisione. Diverse applicazioni hanno sfruttato gli approcci DL nelle tecniche di data augmentation, proponendo modelli in grado di creare nuovi campioni realistici e sintetici. Di conseguenza, è possibile identificare una nuova fonte di dati, definita *fonte di dati sintetici*. L'obiettivo di questa tesi è indagare gli approcci di DL nell'elaborazione di immagini mediche, considerando la presenza di più sorgenti di dati. Nel caso di *fonti di dati multimodali* viene effettuata un'analisi sistematica delle tecniche di fusione, introducendo un modulo di trasferimento innovativo che consente alle diverse modalità di influenzarsi reciprocamente, mentre nell'analisi della *fonte di dati sintetici*, viene proposto un metodo di data augmentation basato sulla DL che sfrutta le caratteristiche biologiche delle immagini implementando un processo di generazione fisiologicamente corretto.

Parole chiave: Intelligenza Artificiale, Elaborazione di Immagini Mediche, Deep Learning, Multimodal Learning, Sintesi di Immagini

Contents

Abstract i
Sintesi in lingua italiana ii
Acknowledgements vi
List of Acronyms ix
List of Figures xvi
List of Tables xx

1 Introduction 1

I Artificial Intelligence in Medical Imaging 7

2 Medical Image Computing 11
2.1 Challenges in Medical Image Computing 12
2.2 Computer Aided Detection/Diagnosis System 14
2.3 The Need for Artificial Intelligence 17

3 The Artificial Intelligence Era 21
3.1 The Spread of Deep Neural Networks 22
3.2 The Need for Data Augmentation 27
3.3 Multimodal Deep Learning 29

II Multimodal Data Sources 33

4 Techniques for Multimodal Data Fusion 37
4.1 Intermediate Fusion with Transfer Module 38
4.2 Early Fusion 40
4.3 Late Fusion 42
4.4 CNN architecture 44

5 Multiple Data from an Imaging Exam: Axillary Lymph Node Status Assessment 47
5.1 Radiomic-based approach for axillary lymph nodes evaluation 48
5.2 Population 50
  5.2.1 Axillary Lymph Node Status 52
  5.2.2 Segmentation 53
5.3 Methodology 53
  5.3.1 Pre-processing 53
  5.3.2 IF approach for ALN status assessment 56
5.4 Experimental Set-up 59
5.5 Results 64
5.6 Discussion 67

6 Data from Multiple Imaging Exams: Dementia Status Assessment 73
6.1 Dementia evaluation in MRI and PET acquisitions 75
6.2 Population 78
6.3 Methodology 80
  6.3.1 Pre-Processing 81
  6.3.2 Dementia Severity Assessment 82
6.4 Experimental Set-up 91
6.5 Results 94
6.6 Discussion 98

III Synthetic Data Sources 105

7 Physiologically-aware Synthetic Data Generation 109
7.1 Physiologically Based Pharmacokinetic models 110
7.2 Physiologically-aware Data Generator 113
7.2.1 Latent representation definition 116
7.2.2 Image synthesis 118

8 Synthetic images: Breast Lesion Classification in DCE-MRI 121
8.1 Breast Lesion Classification in DCE-MRI 122
8.2 Population 125
8.3 Methodology 125
8.3.1 PBPK Fitting 126
8.3.2 3TP Slices Extraction 128
8.3.3 3TP Slice Disentangling 129
8.3.4 DAE for data augmentation 131
8.3.5 DAE-CNN Training 133
8.3.6 Lesion Classification 136
8.4 Experimental Setup 137
8.5 Results 140
8.6 Discussions 145

IV Conclusions 151

Bibliography 157

Author's Publications 175

Acknowledgements

The research presented in this thesis was supervised by Professor Carlo Sansone, who deserves special thanks for having been a source of inspiration and an excellent mentor.

Part of data used in this study has been kindly provided by Dr. Domiziana Santucci, and by Professor Paolo Soda from University of Rome Campus Bio-Medico, Unit of Computer Systems and Bioinformatics, Department of Engineering.

The author's work has been partially carried out in collaboration with Professor Consuelo Gonzalo-Martín and Professor Angel M. García Pedrero, from Center for Biomedical Technology (CTB) of Universidad Politécnica de Madrid (Spain).

This work also used data provided by Dr. Antonella Petrillo, Head of Division of Radiology of IRCCS Istituto Nazionale dei Tumori Fondazione G. Pascale" (Naples, Italy)

List of Acronyms

The following acronyms are used throughout the thesis.

2DS	Two-Dimensional Slice
ACC	Accuracy
AI	Artificial Intelligence
AUC	Area under ROC curve
CAD	Computer Aided Detection/Diagnosis
CADe	Computer Aided Detection
CADx	Computer Aided Diagnosis
CC	Classification Core
CL	Clinic
CNN	Convolutional Neural Network
ConvC	Convolutional Core
CT	Computed Tomograph
CV	Cross Validation

DAE	Deforming Autoencoder
DCE-MRI	Dynamic Contrast Enhanced-Magnetic Resonance Imaging
DL	Deep Learning
DNN	Deep Neural Network
DWI	Diffusion Weighted Imaging
DYN	Dynamics
EF	Early Fusion
GEO	Morphological
GP-GPU	General Purpose-Graphic Processing Unit
IF	Intermediate Fusion
LF	Late Fusion
MDL	Multimodal Deep Learning
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PBPK	Physiologically Based Pharmacokinetic
PET	Positron Emission Tomography
SENS	Sensitivity
SFB	Single Fixed-size Box
SIB	Single Isotropic-size Box
SLIB	Single Lesion Isotropic-size Box

SLVB	Single Lesion Variable-size Box
SPE	Specificity
SVB	Single Variable-size Box
T1	T1-weighted
T2	T2-weighted
TM	Transfer Module
TXT	Textural
U	Unimodal

List of Figures

2.1 Example of brain image registration 17

2.2 Example of brain tumor segmentation in different images. Each color corresponds to a specific class 17

3.1 Example of connection between two layers, having 2 and 3 neurons respectively ($n^l = 3$ and $n^{l-1} = 2$) 24

3.2 A simple autoencoder consists of an *encoder* and a *decoder*. In particular, the former is used to provide a compact or "compressed" version of the input, known as latent-space representation (z), while the latter aims to reconstruct the input starting from its encoded version. 28

4.1 Basic schema for Intermediate Furion approach: j is a generic image modality, with j from 1 to M, representing the number of modalities; $ConvC_j$ is the Conv-C of the modality j; and L_j^i indicates the $i-th$ layer in the $ConvC_j$ with i from 1 to N_j that is the number of layers in $ConvC_j$; $F_j^{N_j}$ is the resulting image-specific features map for each ConvC . . . 41

4.2 Architecture of the proposed TM module for M different image modalities. The input consists of the features maps coming from the modality-specific paths, that are further processed by considering two steps acting in a pipeline, namely the *Shared vector computation* and the *Multimodal calibration*. It is worth noting that the index i of the layer in which the module is inserted is omitted to avoid overly complex notation in the image. 42

4.3 TM inserted between layers belonging to M convolutional cores. 43

4.4 Intermediate fusion approach with M image modalities. Each Conv-C is further detailed in Figure 4.3 and represented as a colored box to avoid making the figure too complicated. . 43

4.5 Intermediate fusion approach with M image modalities and clinical features. Each Conv-C is further detailed in Figure 4.3 and represented as a colored box to avoid making the figure too complicated. 44

4.6 Early fusion approach: x_j is the input image belonging to the modality j 44

4.7 Late Fusion approach: x_j is the input image belonging to the modality j, while $Classifier_j$ is the model trained with x_j samples 45

4.8 ResNet architecture: the Convolutional Core (Conv-C) consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four layers containing the blocks implementing the residual network, while the Global Average Pooling and the fully connected layer represent the Classification Core (CC) 46

5.1 Architecture of the proposed TM module for three different image modalities. The input consists of the features maps coming from the modality-specific paths, that are further processed by considering two steps acting in a pipeline, namely the *Shared vector computation* and the *Multimodal calibration*. 57

5.2 BasicNet architecture: the network consists of five *reduction layers*, representing the Convolutional Core (Conv-C), followed by two fully connected layers spaced by ReLU function, constituting the Classification Core (CC) 60

5.3 ResNet architecture: the Convolutional Core (Conv-C) consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four layers containing the blocks implementing the residual network, while the Global Average Pooling and the fully connected layer represent the Classification Core (CC) 60

5.4 Architecture of the IF approach including TM between layers belonging to different convolutional cores and clinic features 61

6.1 Disease progression for patient OAS31045 81

6.2 Extraction of brain region for each patient: the smallest 3D cubical box (in green) is computed from the rectangular box (in orange) surrounding the brain. 82

6.3 Architecture of the 3D CNN proposed for the EF 85

6.4 Architecture of the 3D CNN proposed for the LF 86

6.5 Architecture of the 3D CNN proposed for the IF: it is possible to distinguish three different sub-networks, the MRI-NET, the PET-NET, and the PAIRED-NET. 86

6.6 Architecture of the 3D CNN proposed for the IF with Transfer Module (TM) 87

6.7 In the first iteration, the PAIRED-NET is considered updating the parts highlighted in yellow. 88

6.8 In the second iteration, the MRI-NET (highlighted in yellow) is updated since the input is a set of MRI volumes . . 88

6.9 In the third iteration, the PET-NET is considered updating the parts highlighted in yellow. 89

6.10 BasicNet architecture: the network consists of six *reduction layers*, representing the Convolutional Core (Conv-C), followed by a Global Average Pooling and a fully connected layer, constituting the Classification Core (CC) 90

6.11 ResNet architecture: the Convolutional Core (Conv-C) consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four layers containing the blocks implementing the residual network, while the Global Average Pooling and the fully connected layer represent the Classification Core (CC) 90

7.1 Most commonly measured PBPK metrics. 111

7.2 Architecture of the Intrinsic Deforming Autoencoder, consisting of an encoder and three different decoders. 116

7.3 The considered DAE encoder, highlighting the DBE and TBE elements. 117

7.4 The considered DAE decoder. The decoders D_a, D_s have the same architecture, while D_w differs only for activation functions (a^*), that is ReLU in D_a, D_s or hyperbolic function (tanh) in D_w. The core elements are $DBD(n, k)$ and $TBD(m, n)$. 117

8.1 An illustrative breast DCE-MRI study (on the left) and some descriptive TICs (on the right): *Type I*, indicative of healthy tissues or of benign lesions; *Type II*, typical of borderline/probably malignant lesions; *Type III*, common in malignant lesions. 123

8.2 Proposed DAE-CNN architecture. The DAE, consisting of an encoder (E) and of three different decoders (D_a, D_s, D_w), disentangles a 3TP slice (I_{3TP}) in its *shading (S), albedo (A)* and *deformation field (W)* components. The *Albedo* (A) component with the application of the deformation field is used as input for the 3TP-CNN to determine the lesion class C (benign/malignant). 127

8.3 3TP slice extraction procedure. The 4D volume $V_{x,y,z,t}$ is sectioned along the highest resolution axis (coronal in the example) generating N different t-channels slices, with t equal to the number of temporal acquisitions and N equal to the size of the cutting axis (z in the example). From these t-channels slices, the corresponding 3TP slices (I_{3TP}) are 3-channel images obtained by considering only the channels associated with the three volumes acquired at the time instances closest to those suggested in [29]. 129

8.4 Intrinsic Deforming-Autoencoder (DAE) architecture used for 3TP slice (input) disentangling. The encoder takes a 3TP slice (I_{3TP}) as input and delivers a low-dimensional latent representation, split into three sub-components. Each of these parts is fed to a different decoder. In the middle, the disentangled three components and the resulting texture; on the right, the reconstructed 3TP slice, used as a reference to train the DAE. 131

8.5 DAE-CNN training is implemented by splitting each iteration into four main steps: the updated parts are surrounded by a shadow that is orange if the cross-entropy loss is used or blue when the DAE loss is exploited. 136
8.6 Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the TK model with W arterial input function (TK_W) 142
8.7 Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the TK model with P arterial input function (TK_P) 143
8.8 Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the ETK model with W arterial input function (ETK_W) 144
8.9 Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the ETK model with P arterial input function (ETK_P) 145
8.10 Example of synthetic benign lesion generated by the DAE. The image on the left represents the real I_{3TP} slice from which the albedo and shading components are extracted. The images on the right are the synthetic slices obtained from DAE and DAE-PBPK after steps 3 and 4. 147
8.11 Example of synthetic malignant lesion generated by the DAE. The image on the left represents the real I_{3TP} slice from which the albedo and shading components are extracted. The images on the right are the synthetic slices obtained from DAE and DAE-PBPK after steps 3 and 4. . . 148

List of Tables

2.1 Different classes of features 16

3.1 Examples of activation functions 25

5.1 Details about the collected *clinical features* 52
5.2 Details about the implemented experiments in the Multimodal Approach 64
5.3 Performance achieved in a 10-folds Cross Validation setting considering the solution proposed in [126], describing different bounding box options. 67
5.4 Performance obtained in 10 fold-CV of the experiments conducted to assess how the position of the transfer module influences performance in the case of BasicNet. 67
5.5 Performance obtained in 10 fold-CV of the experiments conducted considering the IF approaches. 68
5.6 Performance obtained in 10 fold-CV of the experiments conducted with unimodal approach and different multimodal fusion modalities. 69
5.7 Comparison of the features selected by the domain expert (D.S), and those obtained with the backward features elimination and SVM. 70

6.1 The progression of patients in the three classes, cognitive normal (C), very mild dementia (M), mild/severe dementia (S) 80

6.2 Information about generated datasets in terms of number of volumes for each class, cognitive normal (C), very mild dementia (M), mild/severe dementia (S) 80

6.3 Details in terms of datasets and tasks for each of the considered approaches 94

6.4 Performance of the EXP.1 (Task B1). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures. 99

6.5 Performance of the EXP.2 (Task B2). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures. 100

6.6 Performance of the EXP.3 (Task A). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures. 101

6.7 Performance of the EXP.4 (Task B). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures. 102

8.1 The mean, the median, the standard deviation (std), the minimum and maximum value of the E_{PBPK} loss computed considering the set of images and PBPK model detailed in columns *Image* and *PBPK Model* respectively. For readability, beside each value the notation $\cdot 10^{-3}$ is omitted. . . . 141

8.2 Performance comparison in 10 fold-CV between the methodology proposed in this thesis and some variants obtained by changing the training iterations. 146

8.3 Performance comparison in 10 fold-CV between the proposed approach and some state-of-the-art competitors, sorted by descending AUC. 146

8.4 Quantitative assessment of the I_{rec} generated by the DAE before and after the DAE-CNN training step. The column *Gen.* reports the involved generator, namely the DAE and DAE-PBPK, namely the DAE and DAE-PBPK, the *Net* gives information about the network used for classification, while *Step* details the number of steps performed during each training iteration. If the network is not specified the DAE or DAE-PBPK obtained after the 3TP Slice Disentangling step is used for the evaluation. For readability, AlexNet, ResNet, and VGG19 are denoted with Al, R, and V. For readability, beside each value the notation $\cdot 10^{-3}$ is omitted. 149

Chapter 1

Introduction

Nowadays, the term "Artificial Intelligence" (AI) is inherent in a very diverse range of application contexts such as automotive and avionics, smart cities, robotics, security, and telemedicine. The relationship between intelligent computer systems and people is becoming more and more noticeable in daily life, and even companies have started to integrate AI in logistics and industrial operations.

AI aims to simulate human intelligence in machines, making computers able to perform the typical human mind referring to the capacity of abstraction and problem-solving. This characteristic highlights that the *intelligent machines* can generate general rules or concepts (abstraction) and exploit them for overcoming obstacles or solving specific tasks, implementing a process similar to what is innate in the human mind. In particular, the term "Machine Learning" (ML) is referred to the ability of AI systems to simulate the human learning process, that is from experience or examples. Indeed, ML models are able to get knowledge from a set of examples and become aware by information or observation. This property completely transforms the way the solution to different tasks is determined. It is no longer necessary to define a precise and unambiguous sequence of steps (algorithm) as the model is able to autonomously learn the concepts and abilities required to solve the specific problem. Among ML models, Artificial Neural Networks (ANNs), have definitely achieved the largest success consisting of a layered structure of computing elements (artificial neurons) that is inspired by the human brain, simulating the complex interconnected structure of biological neurons. More recently, the vastly increased amounts of data, the advances in General-Purpose GPU (GP-GPU) computing, and the development of frameworks enabled the definition of complex and flexible ANNs, determining the rise of *Deep Neural Networks* (DNNs), and, in general, *Deep Learning*, that is a subset of ML characterized by "very deep and complex networks". The layered structure of neurons is able to create a high-level representation of the input while extracting the set of concepts needed to solve a specific problem. The key aspect of DNNs is their ability to autonomously learn the best set of features for the task under analysis, exceeding human capabilities in some applications. This characteristic, known as *features learning*, has played a key role in the recent spread of AI, allowing DL and DNNs also in domains lacking an effective and defined set of features. DL approaches gained surprising performance, overcoming the classical ML models in several fields. Indeed, ML-based solutions require a step of *features definition and selection*, which is a tedious task performed by the domain expert, while DNNs directly determine the best data representation able to capture the main aspects of the specific problem to solve.

Healthcare is one of the domains that is still experiencing a huge impact of AI. It is identified as one of the most promising applications of AI [158] that provides "intelligent" systems to support both patients and physicians. The first AI applications were introduced in the 1970s, implementing rule-based approaches for the diagnosis of diseases [27], treatment selection [132], physicians assistance [96]. However, the definition of rule-based systems was very complex since they require an explicit and well-defined set of rules and human interactions for the updates. Moreover, the performance was limited by the difficulty of being able to encode and model the complex mechanisms affecting two or more entities. The ability to learn complex interactions and autonomously extract the concepts required to solve a specific task without a set of predefined steps made ML increasingly applied in healthcare with very promising results. ML techniques enabled the spread of AI in the medical field and contributed to the discovery of previously unrecognized patterns in the data without the need to specify decision rules [158].

Among all healthcare sectors, medical image analysis or medical image computing is the research field experiencing the greatest impact of

AI-based solutions that particularly exploit ML and DL models. Medical imaging refers to the set of tools that deal with the visualization of areas of the body normally concealed by the sight. In recent years, technological advancements in image acquisition have enabled novel and innovative imaging modalities, such as multi-slice (volumetric) and multi-energy Computed Tomography (CT), multi-parametric and multi-frame (dynamic) Magnetic Resonance Imaging (MRI), dynamic Positron Emission Tomography (PET), or multimodal (hybrid) PET/CT and PET/MRI imaging technologies. Medical image computing consists of the extraction of relevant information from images for the categorization of anatomical structures or diseases. The manual analysis of medical images results in a very tedious and complex task due to the intrinsic characteristics of the data and the huge quantity of information to be considered. As a consequence, physicians make often use of tools, namely Computer Aided Detection/Diagnosis (CAD) systems, supporting them in image processing. More recently, different AI solutions, exploiting ML and in particular, DL approaches, have been proposed for the implementation of a CAD system covering the main steps of *image registration*, aiming to remove noise or artifacts intrinsic to diagnostic tools, *image detection/segmentation*, for the detection of a specific region of interest, and *image classification*, for categorization of the previously defined area.

The complexity of medical image processing strongly depends on the intrinsic nature of the data to be investigated, usually represented by a 3D or 4D volume in the case of temporal dimension (i.e dynamic MRI). Furthermore, in several applications, there is the need of exploiting data coming from multiple sources or *modalities* that provide complementary information, increasing the level of complexity. The idea is that heterogeneous images may highlight different characteristics of the area under analysis that are useful for its characterization. Data from various modalities need to be fused to provide a rich representation of the phenomena to be explored. The *multiple data sources* may include sequences acquired during the same imaging exam or form independent diagnostic tools. In the first case, an example is the MRI that involves the acquisition of heterogeneous sequences, such as the Dynamic-Contrast Enhanced (DCE), the T2-weighted (T2), and the Diffusion-Weighted Imaging (DWI) ones. In the second case, different imaging tools, for instance, T1-weighted (T1)

MRI and dynamic PET showing structural properties and metabolic functions of the tissue under analysis respectively, are exploited in the same task. Although the two scenarios share the need to combine several sources, they differ in the way the images are acquired. Indeed, the independence between the diagnostic tools refers also to the moment of the acquisition, which may not coincide between different images causing the lack of a modality in some patients. The presence of multiple data sources in medical image analysis enables the spread of Multimodal Learning in healthcare, which allows the fusion of complementary information from heterogeneous diagnostic tools. When investigated in conjunction with deep networks, multimodal learning is known as Multimodal Deep Learning [116], leveraging the ability of the DNNs to provide an effective high-level representation of the input. Techniques for multimodal data fusion can be categorized into early, intermediate, and late fusion [116] and differ according to when the integration is performed.

The strength of DNNs consists of their autonomous *features learning*. However, this characteristic comes with a huge number of parameters to learn, resulting in a need for a long training time and a suitable number of annotated samples. Unfortunately, in the medical field, it is very difficult to obtain high-quality, balanced datasets with labels. Indeed, in contrast to natural image processing, the solutions exploiting medical imaging require the consensus of both domain experts and patients, resulting in privacy issues [2]. Data augmentation is a technique introduced to increase the variability of the data used for training, providing an attempt to handle the problem of the limited size of data. It consists of the creation of additional representative images which simulate changes in the acquisitions and anatomical variations of patients. More recently, several applications [19] have leveraged DL approaches for the implementation of data augmentation techniques, proposing models that are able to create new realistic and synthetic samples. Such models, referring as generative networks, learn the distribution of the *real* data, namely the acquired images, and generate samples with the same characteristics. The result is a set of synthetic images that are used to improve the generalization ability of the DNN involved in the specific AI application. As a consequence, during the development of the solution, it is possible to identify a new source of data, that is denoted as *synthetic data source*. However, despite promising, DL naive use may not be effective since *medical images are more than pictures* [38]. In the case of DL-based data augmentation approaches, the synthetic images should preserve the physiological and biological characteristics of the tissue under analysis.

The aim of this thesis is to investigate the AI approaches in medical image computing, considering the presence of multiple data sources, namely multimodal and synthetic ones, and proposing innovative methodologies in comparison with the current literature. The contribution of this thesis can be detailed for each data source and summarized as follows:

- *Multimodal Data Sources*: an innovative module, namely the *Transfer Module*, is proposed to implement the cross-modality calibration of the representations extracted by the DNNs, allowing the different modalities to influence each other; a multi-input multi-output network is implemented to handle the lacking of different modalities in some patients; the analysis of multimodal data fusion techniques is performed in two different scenarios that include sequences acquired during the same imaging exam and from independent diagnostic tools, respectively.
- *Synthetic Data Sources*: an innovative DL-based data augmentation is proposed that exploits the biological characteristics of the images in the generation of new samples, implementing a physiologically-aware synthetic image generation process; a nested training strategy is implemented to handle both real and synthetic images; a DL-based data pre-processing step is exploited to highlight the components of the image that best fit the specific task to solve.

This thesis is organized into four main parts: Part I introduces the principles of medical image computing, and the main concepts of AI, focusing on deep learning and multimodal deep learning in the medical domain; Part II addresses the *multimodal data sources* and analyzes the multimodal data fusion techniques in two different applications; Part III focuses on the *synthetic data source* proposing an innovative DL-based data augmentation technique that considers the physiological characteristics of the images involved in the analysis; finally, Part IV summarizes the content of the work and provides the conclusions.

Part I

Artificial Intelligence in Medical Imaging

In the last years, Artificial Intelligence (AI) has become part of daily life with applications showing surprising results in several domains. In particular, the complexity and the rise of data in healthcare made AI increasingly applied in this field, resulting in a profound transformation in many aspects of patient care as well as administrative processes. Pharmaceutical pharmacies involve AI approaches for the development of drugs, which is a competitive and expensive medical business, reducing the time required for drug discovery and validation. The AI ability to extract information from a set of data is exploited to provide automatic detection, segmentation, and categorization of different diseases and to make a step forward in precision medicine, that proposes treatments and strategies considering the variability among different patients. Healthcare organizations experiment with applications based on natural language processing (i.e chatbots) to support both patients and physicians in administrative processes.

However, biomedical image processing is one research area experiencing the greatest increase of AI-based solutions, providing physicians with systems supporting them in the challenging task of medical image analysis for the detection and characterization of different diseases. Indeed, the *medical image analysis* or *medical image computing* turns out to be a tedious and time-consuming process with intra- and inter-observation variability, that compromises the effectiveness of the results.

In the first part of this thesis, Chapter 2 describes the principles of medical image computing, while Chapter 3 introduces the main concepts of Artificial Intelligence, focusing on Deep Learning and Multimodal Deep Leaning in the medical domain.

Chapter 2

Medical Image Computing

Medical imaging, biomedical imaging, and "diagnostic by imaging" all relate to an area of medicine that deals with seeing bodily parts and structures that are ordinarily out of sight as well as providing physiological data on specific organs and tissues. All imaging techniques exploit one or more physical laws or properties, with the aim of inferring tissues' characteristics from the measured signal: for example, radiography makes use of x-ray radiations, which are absorbed at different rates by different tissue types such as bone, muscle and fat, allowing their recognition.
Novel imaging modalities, such as multi-slice (volumetric) and multi-energy Computed Tomography (CT), multi-parametric and multi-frame (dynamic) Magnetic Resonance Imaging (MRI), dynamic Positron Emission Tomography (PET), that include temporal dimension, or multi-modal (hybrid) PET/CT and PET/MRI imaging technologies, are being introduced in medical practices as a result of ongoing technological advancements in image acquisition. *Medical image analysis* or *medical image computing* refers to the process of extracting relevant information or knowledge from medical images with the aim of developing potential non-invasive biomarkers for the detection and characterization of the disease. Although being closely related to medical imaging, it focuses on the computational analysis of the images, not their acquisition with the aim of improving the interpretability of the depicted contents [121]. Within the wide variety of imaging applications, the problem related to the analysis of medical images can be grouped into three basic tasks, namely, the *visualization*, *registration*, and

segmentation.
The *visualization* plays a key role in medical image computing since it is used to understand and communicate all the information extracted from the data that need to be presented in the most optimal way, supporting the diagnosis and therapy planning. The presence of volumes acquired at different time points (i.e Dynamic MRI or PET) or from different modalities (i.e. CT, MRI and PET), containing complementary information, results in the need of introducing a *registration* operation that determines the spatial relationship between different acquisitions, establishing the correspondence between them. The image registration aims to compensate for the unknown differences in patient positioning in the scanner or for the deformations of tissue and organs between different time points. In medical imaging, the *segmentation* consists of identifying the regions of interest by partitioning the image into different groups, corresponding to organs, tissues, pathologies, or relevant structures. It allows the definition of geometric properties of the objects, such as shape and volume.
However, the processing of the massive volumes of imaging data produced by various modalities has grown to be an enormous challenge for diagnosis, therapeutic planning, follow-up, and biomedical research [117].

2.1 Challenges in Medical Image Computing

The manual analysis of medical images by human experts results in a very tedious and time-consuming task. Moreover, different factors contribute to the complexity of medical imaging processing that depend on the intrinsic characteristics of the data, the specific region of interest, and the difficulty of the validation step.

The *complexity of the data* is a consequence of the fact that medical images are typically 3D or 4D volumes if the temporal dimension is considered together with the spatial characteristics. Although the multi-dimensional nature provides additional information, it introduces another level of complexity [117]. Indeed, instead of processing images slice-by-slice cutting the volumes along a specific dimension, the 3D analysis is more effective, since it exploits the volumetric characteristic of the tissue under analysis. As aforementioned, the imaging techniques consider different physical principles, and the quantification of the images is complicated by the intrinsic limitations of the image acquisition process, in terms of resolution, lack of contrast, noise, or presence of artifacts. For instance, in the modalities that require several time points (i.e Dynamic MRI), even the involuntary patient's movement may compromise image quality. Furthermore, in several applications, there is the need of exploiting data coming from multiple images that provide complementary information. As a consequence, the fusion of different modalities represents another level of complexity.

The objects of interest in medical images typically represent anatomical structures, such as tissues or organs, either normal or pathological (i.e., lesions). They usually present complex shapes that cannot easily be modeled or described by a mathematical model. Moreover, involuntary movements such as breathing-related motions, cause a large intra-patient variability of the anatomical structures that affects the image acquisition process, while inter-subject variability is a consequence of normal biological variation and pathological changes. In general, medical images belonging to the same patient at different time points or to various subjects may show significant variability both in shape and in intensity [117], although related to the same anatomical structure. The *complexity of objects of interest* is a direct result of the difficulty of modeling their variability. Indeed, computational strategies for medical image analysis need to take this variability into account and be sufficiently robust to perform well under a variety of conditions.

Medical imaging allows the extraction of quantitative and qualitative information from structures of the body not visible from the outside. Hence, the evaluation of the analysis is often impossible in most applications due to the lack of a *ground truth*. In clinical practice, the ground truth, such as the manual detection or delineation in the case of image segmentation, is provided by a domain expert, that is the physician involved in the analysis. However, the intra- and inter-observation variability may compromise the effectiveness of the process. As a consequence, the *complexity of the validation* depends on the difficulty to have a well-defined ground truth, not affected by human errors.

The definition of strategies for medical image computing should take the factors of complexity into account with the aim of proposing methods that are able to capture the variability of the anatomical structures under analysis and operate in presence of noise or artifacts while limiting the possibility of human errors. In clinical practice, the use of "double reading" is strongly recommended. It consists of the repetition of the assessments that are executed several times by the same radiologist or by different physicians. This recommendation helps to understand the complexity and sensitivity of the analysis of medical imaging.

2.2 Computer Aided Detection/Diagnosis System

Nowadays, physicians make often use of tools that assist them in the analysis of medical images. The Computer Aided Detection/Diagnosis (CAD) System is frequently used in the analysis of challenging medical examinations both for the huge amount of information to be considered and for an inherent uncertainty of the data due to the scanning process. Additionally, it enables the removal of issues brought by intra- and inter-observation variability, which can be reflected by several assessments of the same location made by the same radiologist at various times or by various clinicians [16]. Indeed, CAD systems evaluate data using precise mathematical patterns that follow predetermined and well-defined procedures.
Essentially, a CAD system is made up of a number of independently implemented steps. The processes that are performed are in line with the system's goals, which might range from straightforward assistance for the doctor to a more sophisticated automatism for the decision (i.e diagnosis of the disease). The bulk of the CAD presented in the literature adheres to the following general schema:

- **Image Pre-processing**: This stage contains a series of low-level preliminary image elaborations with the aim of enhancing quality by lowering noise introduced during the acquisition step or removing any artifacts brought by patient movements. This aspect is essential, particularly if the CAD system is designed for diagnostic instruments that need acquisitions at different times (Dynamic MRI). Image registration should thus be included in this phase to lessen the impact of the motion artefacts. In fact, it enables the transformation of two different images' reference systems for comparison. In medical imaging, registration converts or aligns two acquisitions made by various equipment or acquired at various points in time. There is a large variety of algorithms for achieving image registration [12, 87]. An example of image registration is shown in Figure 2.1.

- **Image Segmentation**: It entails dividing the image into homogenous regions of interest based on predetermined features. The performance of the subsequent phases is impacted by the accuracy and quality of the segmentation findings, making this stage the most crucial. The suggested segmentation methods range from straightforward techniques based on thresholding operations to sophisticated algorithms involving pattern recognition [21]. However, the segmentation process may be complicated by the images' low resolution and the presence of noise or distortions, particularly in cases where the region of interest is tiny relative to the anatomical structure under analysis (i.e. tumor segmentation). An example of image segmentation is shown in Figure 2.2.

- **Feature Extraction and Selection**: It involves the extraction of characteristics ("features") from previously selected areas of interest. As a result, the item to be analyzed is represented by a vector of attributes that are assumed to be pertinent to the particular issue at hand. In the literature, many feature classes have been proposed; they are included in Table 2.1.

- **Image Classification**: In this step, the systems help the physician in the diagnosis of the disease, by collecting the information extracted in the previous stages. In particular, a class or label is associated with each region of interest with a probability, that represents the affinity of each object in a given class. In other words, the previously extracted features are further processed to determine the "type" of the anatomical structure under analysis. In most cases, the classification step involves approaches based on pattern recognition, in which a model, that is a classifier, needs to be trained for the specific task to solve.

It is worth noting that there is no need to implement all the stages in order to classify a system as a CAD. Indeed, the order and the number of steps vary according to the specific medical application. In particular,

Features Class	Symbol	Characterization
Dynamics	DYN	These characteristics use measurements taken straight from the time-intensity curve to assess the temporal dynamics of the signal. This class of characteristics is used on on images that were acquired at different times.
Clinico-Pathologic	CL	These characteristics, which reflect the clinical status and pathological condition of the patient, are taken from their medical records.
Morphological	GEO	These characteristics quantify the structure and geometry of the region of interest.
Textural	TXT	These features measure the perceived texture of an image quantifying the variations in the signal intensity obtained during the image acquisition.
PharmacoKinetic	PBPK	According to PharmacoKinetics (PBPK) Modelling, these characteristics represent certain physiological properties of tissues that were computed using a mathematical model

Table 2.1. Different classes of features

CAD systems are categorized into two main groups, namely Computer Aided Detection (CADe), which focuses on the localization of the disease, and Computer Aided diagnosis (CADx), for the classification of the disease [9].

In recent years, CAD systems using *radiomics* have been developed [9]. Radiomics involves the extraction of quantitative, reproducible information, called *feature* from medical images able to describe complex patterns that are difficult for humans to grasp [93]. The extracted features aim to capture properties of the anatomical structures reflecting their shapes, intrinsic behavioral characteristics and temporal changes. Additionally, they are exploited to predict clinical outcomes such as patient survival or treatment response. The most important aspect of radiomic features derives from the fact that they can be utilized to find previously unidentified indicators of disease evolution and progression, especially when the features are extracted from a sufficiently big dataset. In particular, "radiomic signature" corresponds to the set of characteristics that have predictive or prognostic power. The main idea is that the radiomic features correlate with the molecular characteristics, genotype, and phenotype of the region of interest under study allowing the personalization of the treatment, that is precision medicine.

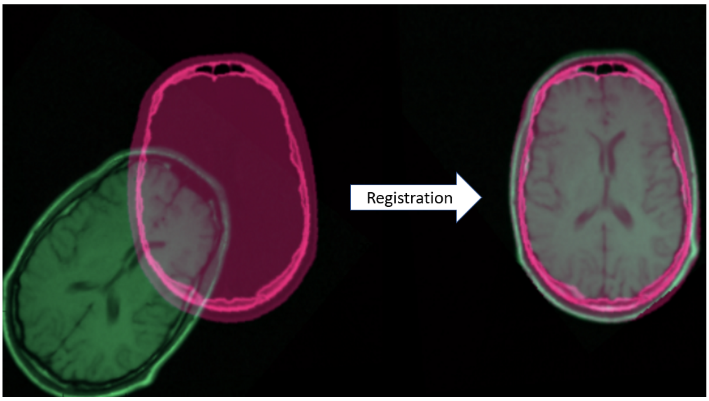


Figure 2.1. Example of brain image registration

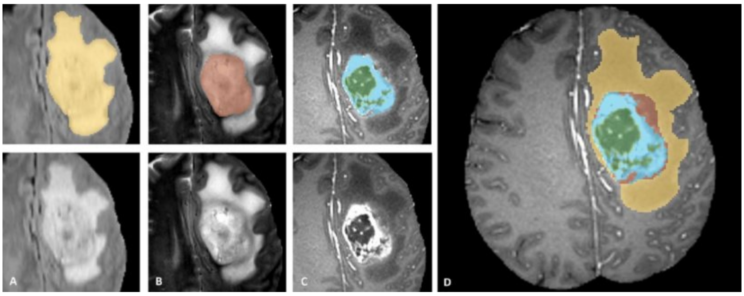


Figure 2.2. Example of brain tumor segmentation in different images. Each color corresponds to a specific class

2.3 The Need for Artificial Intelligence

As aforementioned, medical image computing is a tedious task resulting in the need of implementing systems (CAD) to support physicians in the analysis. Moreover, the methodologies and strategies introduced for medical image processing need to deal with the different factors of complexity to be able to operate under several conditions. The imaging procedures result in complex high-dimensional volumes of data that make the reading and interpreting very challenging for human eyes [106].
The large amount of information to consider, and the high variability and complexity of medical images have prompted research into proposing so-

lutions to automate the analysis of radiological acquisitions exploiting the recent and advanced computational methods developed in computer science [106]. Indeed, the digitization of diagnostic methods has enabled the proposal of new and always more sophisticated software to process them. In recent years, several methodologies have been implemented in the literature focusing on CAD systems that exploit Artificial Intelligence (AI) techniques, which refers to the simulation of human intelligence in machines and it includes a set of strategies and algorithms that are able to discover hidden patterns in data while "learning" how to perform a specific task. Among all AI techniques, Machine Learning (ML) and Deep Learning (DL) are the most widely used approaches. The former is a subset of AI including algorithms that *learn from examples*, extracting from them the general concepts, that is the knowledge, while the latter is a subset of ML that involves the use of Deep Neural Networks.

Many AI applications show very promising performance and cover all the steps implemented in a CAD system, namely the image registration, segmentation, and classification [81], offering an efficient support service that removes the intra- and inter-observation variability [58]. In medical image computing, AI solutions provide a way of finding non-invasive and quantitative assessments of diseases, which is fundamental for early treatment. Moreover, they might highlight pattern changes or intrinsic characteristics that are hidden from the human eye, offering the opportunity to better understand disease processes [103].

Radiomics is one of the most advanced applications for AI. It first extracts a large amount of quantitative features from medical images and, after a careful features selection step, it uses ML models to provide tools to predict different outcomes all with predictive horizons, such as progression-free survival, the raise of metastasis in the case of tumor analysis, response to therapy, etc. A number of recent studies have shown that DL enhances this manually designed workflow by automatically deriving the radiomic signature from images. Indeed, the success of precision medicine, which aims to propose treatments and strategies considering the variability among different patients, strongly depends on the evaluation of the biomarkers [77]. As a consequence, DL methods can explore or create quantitative biomarkers exploiting medical images and integrating different sources of information (i.e clinical characteristics).

Although AI application in medical image computing has shown very promising results, the physiological characteristics of the regions of interest under analysis should always be considered during the proposal of AI-based solutions, since “medical images are more than pictures” [38]. This characteristic implies that diagnostic-related information is not only associated with image texture resulting from the signal intensity. The past medical knowledge needs to be integrated into the implemented system with the aim of providing strategies showing biological and physiological consistency [40, 41].

While many AI solutions have been proposed for medical imaging, it is often challenging to obtain a high-quality set of images that can be used for the evaluation of the methodologies. Indeed, it is not easy to operate with acquisitions coming from different institutions o medical centers due to privacy issues and data-sharing limitations. As a consequence, most models are tailored for a specific institution [75] with a lack of generalization ability in a different context. Moreover, the limited variability in the set of data and the small number of samples, especially in the case of rare diseases, may result in a system with poor performance.

Several AI-based models, especially those exploiting DL approaches, are considered "black-box" since it is not easy to understand how the algorithms compute their decision. As a consequence, research in literature is focusing on the definition of techniques of explainable Artificial Intelligence [75], that aim to highlight the locations of the image contributing to the final outcome. Such techniques help the user to trust the AI system, explaining the reason behind a specific prediction.

Despite these challenges, the field of AI in medical imaging is constantly growing with the recent proposals of methodologies showing surprising results. In particular, ML and DL approaches provide physicians with systems supporting them in the analysis for the detection and characterization of several diseases. It is worth noting that AI-based systems will not replace human clinicians, but rather will help them in their tedious tasks [26]. Regardless of the reported performance, no automated system will be able to replace the human being, the only one capable of empathy and humanity.

Chapter 3

The Artificial Intelligence Era

As already said, Artificial Intelligence (AI) refers to the simulation of human intelligence in machines, including methods that enable computers to carry out operations that are similar to the common human mental processes. However, AI applications are not limited to a specific technology, but rather a collection of them. The term "Pattern recognition" refers to the recognition and finding of patterns in a set of data. It has several fields of application, from data analysis with statistical approaches, to signal processing, image analysis, and Machine Learning (ML). Indeed ML is a subset of Pattern Recognition and encompasses the group of algorithms that can "learn from examples," which refers to their capacity to pick up new skills via practice. The term "to learn" is defined as getting knowledge by study, experience, or being taught, and it highlights that ML algorithms are able to extract concepts (or knowledge), namely the "things to be learned", from a set of data. An example is usually denoted as an "instance" and represents the "example of the concept to be learned". In healthcare, the most common application of traditional ML is precision medicine – predicting what treatment protocols are likely to succeed on a patient considering his/her attributes and the treatment context. The algorithms require a set of examples, that is the training set, for which the outcome variable is known, with the aim of learning through a process known as *the training step*, the concepts needed to perform a specific task. Artificial Neural Networks (ANNs), which are inspired by the human brain's network of biological neurons, have unquestionably enjoyed the greatest degree of success among all ML models. An artificial neuron consists of a mathematical operation that represents the basic building blocks of an ANN, which is a parallel structure whose neurons are arranged in layers and interact with each other to carry out the desired task after an appropriate training phase. The value passed to a neuron is calculated by taking all the values of the neurons in the previous layer, multiplying them by the appropriate weights, and summing the results. This sum plus an extra offset, known as the bias, is passed as the input to a function, known as the transfer function and the output from this function is the value passed to the successive neuron. The weights, biases, and transfer functions determine how inputs are transformed into outputs. The term Deep Learning (DL) is a subset of ML characterized by Deep Neural Networks (DNNs), consisting of deep architectures organized in several stacked layers. DL approaches have become more popular in recent years for several pattern recognition tasks, beating earlier state-of-the-art ML models in a variety of fields. This characteristic leads the research to explore DL approaches in the development of CAD systems, obtaining excellent results [106] and allowing more complex tasks to be performed than classical hand-crafted radiomics [54].

3.1 The Spread of Deep Neural Networks

DL approaches have spread to many research areas, thanks to four main elements:

- *Vastly increased amounts of data*: A significant amount of data from many sources has been gathered thanks to new technologies and growing Internet usage. The evolution of "Big Data" in late 2000, a term used to describe a collection of data that is too massive and complicated to be handled by conventional software, demonstrates the limitations of traditional machine learning techniques and the need for more intricate models.

- *Deeper and larger network architectures*: Deeper, more intricate, and more adaptable models are possible as evaluation data volume rises. In particular, researchers began to suggest novel approaches using various methodologies to enhance the generalization capacity of the deployed networks.

- *Accelerated training using GPU techniques*: The recently suggested networks need high computational power to optimize the huge number of parameters, typical of deep architectures, and to elaborate the huge amount of data. As a consequence, the advances in General-Purpose GPU (GP-GPU) computing strongly supported the adoption of deep learning models, enabling a large decrease in the necessary training periods.
- *Proper evaluation of machine learning methods*: In the past, it was common practice for several groups to evaluate the performance using the same data sets. However, results were sometimes challenging to compare since different research groups utilized diverse experimental approaches, which frequently resulted in poor baselines. The introduction of machine learning challenges with large common test sets makes outcomes more directly comparable and encourages teams to focus their time and effort on developing their unique approach.

Deep neural networks (DNNs) compose computations performed by many layers, consisting of neurons organized in a hierarchical architecture. Typically, each neuron in a layer is connected to all the neurons in the previous one, creating a *fully connected deep neural network*. Let L^l be the l-th layer, with l from 1 to N, representing the number of layers, and n^l be the number of the neurons in L^l. Denoting with x the input vector for the DNN and the output the the l-th layer x^l is computed as follows:

$$\begin{cases} x^l = f^l(W^l x^{(l-1)} + b^l) & l > 1 \\ x^1 = f^1(W^1 x^{+} b^1) & l = 1 \end{cases} \tag{3.1}$$

where $x^l \in R^{n^l \times 1}$, $x^{l-1} \in R^{n^{l-1} \times 1}$, W^l is a $n^l \times n^{l-1}$ matrix in which the w^l_{ij} element represents the weights between the i-th neuron in the layer l with the j-th neuron of the layer $l-1$, $b^l \in R^{n^l \times 1}$ represents the bias of the layer l and f^l is its activation function. In particular, the activation function helps the network learn the complex and nonlinear pattern in the data, influencing what is transferred to the next level.
The described DNN is called *feedforward* because information flows from

the input x, through the intermediate layers, and finally to the output y. There are no feedback connections in which outputs of the model are fed back into itself. Figure 3.1 shows an example of connection between two layers, having 2 and 3 neurons respectively ($n^l = 3$ and $n^{l-1} = 2$). The Equation 3.1 is detailed as follows:

$$\begin{pmatrix} x_1^l \\ x_2^l \\ x_3^l \end{pmatrix} = f^l(\begin{pmatrix} w_{11}^l & w_{21}^l \\ w_{12}^l & w_{22}^l \\ w_{13}^l & w_{23}^l \end{pmatrix} \cdot \begin{pmatrix} x_1^{l-1} \\ x_2^{l-1} \end{pmatrix} + \begin{pmatrix} b_1^l \\ b_2^l \\ b_3^l \end{pmatrix}) \tag{3.2}$$

The *rectify* function (ReLU), defined as $f(x) = max(0, x)$ is an example of activation function that yelded superior results in many different settings. Since it is 0 for negative argument values, some units in the model will yield activations that are 0, giving a “sparseness” property that is useful in many contexts. Table 3.1 gives a list of different activation functions.

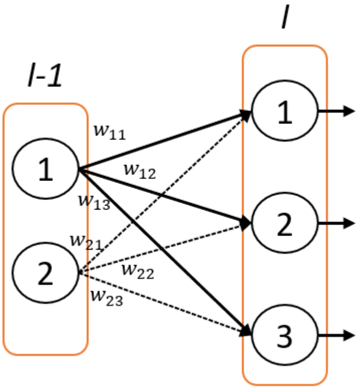


Figure 3.1. Example of connection between two layers, having 2 and 3 neurons respectively ($n^l = 3$ and $n^{l-1} = 2$)

Convolutional Neural Networks (CNNs) are a special kind of feedforward networks that has proven extremely successful when the input is a multi-dimensional array or data with a grid-like topology [11]. Examples include image data, which can be thought of as a 2D grid of pixels. The name "Convolutional Neural Network" indicates that the network implements a mathematical operation called *convolution* that is used in place of the general matrix multiplication described in Equations 3.1 and 3.2. This

Name	**Function**
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Hyperbolic	$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
Softplus	$f(x) = log(1+e^{x})$
ReLU	$f(x) = max(0, x)$
Leaky ReLU	$\text{f(x)} = \begin{cases} x & x \geq 0 \\ ax & x > 0 \end{cases}$

Table 3.1. Examples of activation functions

operation is used to apply filters to the input data obtaining a set of organized features. The filtering operation can be implemented by multiplying a relatively small spatial zone of the image by a set of learnable weights and feeding the result to an activation function like those discussed above. The operation is repeated around the image using the same weights (parameters sharing).

When viewed as a layer in a neural network, called convolutional layer, that takes an image as input, this filtering operation can be viewed as a set of constraining spatially organized neurons that compute features within a limited region of the input known as the neuron's *receptive field.* In a convolutional, layer the filter used in the convolutional operation is denoted as kernel, while the output is the features map, containing the filter's response to each spatial location.

Instead of using a set of predefined filters, CNNs jointly learn sets of convolutional filters, demonstrating the remarkable capacity to "autonomously learn" during the training stage the optimum input representation for the particular task under investigation. This characteristic, known as feature learning, has been essential to the recent rise of AI since it allowed the technology to be applied in fields like medicine where expert-designed features were unsuccessful. Indeed, it removes the need for the hard job of features engineering by eliminating the features extraction stage that is usual in standard ML systems.

Although the deep hierarchical design of a CNN, which can learn features at many levels of abstraction [34, 141], is its key strength, it results in a huge number of parameters to learn, requiring a lengthy training stage and a sufficient amount of annotated examples. Unfortunately, assembling a large dataset is a challenging, expensive, and time-consuming operation, particularly in fields where it is difficult to obtain a lot of samples. Biomedical imaging is an example where gathering large datasets is not only technically challenging (privacy-related difficulties, various methods, etc.), but also complicated due to the extreme class imbalance (e.g. between positive and negative oncology patients). An increasingly common approach to face this issue is to use *Transfer Learning*, which refers to the process of applying what has been learned in one task (e.g., utilising pre-trained CNN parameters) with an adequate quantity of training data to another (sometimes very different) task. Transfer learning is being used more and more in current research to develop high-performance solutions in a variety of sectors [105]. Specifically, two separate methods may be used to look for transfer learning:

- Fine-tuning: It involves applying the previously trained CNN to a new task. To do this, a pre-trained network is used as a starting point, followed by a training step using data from the new task.
- CNN as a feature extractor: The assumption is that the pre-trained CNN has acquired how to extract a set of characteristics that should be useful also for the new task. Thus, by feeding the pre-trained network with the new data and using the output of one of the convolutional layers to create a new set of features, it is possible to take advantage of this expertise. Any type of machine learning model may then be trained using these features to perform the new task.

In the medical field, DL approaches are widely exploited for the implementation of CAD systems [77, 140]. Moreover, the use of transfer learning helped the implementation of models for various diseases and different image modalities. Several CAD systems have been proposed in the literature for breast cancer [153], lung cancer [73, 50], and Alzheimer's disease [138], showing promising results. The spread of CNNs improved the performance for the tasks of image classification, image segmentation, and generation [64]. The image classification is exploited to differentiate several diagnoses, the presence or absence of a disease, or the type of malignancy. Applications for image segmentation include approaches for the segmentation of lungs, tumors, biological cells, or bone tissue. However, in the medical field, it is difficult to obtain high-quality, balanced datasets with labels.

To overcome the issue, several proposals in the literature exploit CNN for synthetic image generation.

Autoencoders are another type of feedforward network that learns an efficient coding of its input. The objective is simply to reconstruct the input, but through the intermediary of a compressed or reduced dimensional representation. Figure 3.2 shows a simple autoencoder consisting of several layers. It is worth noting that an autoencoder can be implemented considering fully connected o convolutional layers. The architecture consists of an *encoder* and a *decoder*. In particular, the former is used to provide a compact or "compressed" version of the input, known as latent-space representation (z), while the latter aims to reconstruct the input starting from its encoded version.

The characteristic of the autoencoder to make the output equal to the input may sound useless. However, different applications are more interested in the hidden representation than the decoder's output. Indeed, the aim is to compute a compressed version that retains all the useful properties that are the salient features of the data distribution. The quality of the representation is determined according to the ability of the decoder to reconstruct the input. Moreover, recently, theoretical connections between autoencoders and latent variable models have brought autoencoders to the forefront of generative modeling, exploiting a variation of the latent representation to create new instances belonging to the same distribution of the input data.

3.2 The Need for Data Augmentation

Deep neural approaches require a large amount of data in the training stage with the aim of preventing overfitting that limits the generalization ability of the networks [19]. However, the application of DL in the medical field is challenged by the limited size of the data, since it is very difficult to obtain high-quality, balanced datasets with labels. Indeed, in contrast to natural image processing, the solutions exploiting medical imaging require the consensus of both domain experts and patients, resulting in privacy issues [2].

Data augmentation is a technique introduced to increase the variability of the data used for training, providing an attempt to handle the problem of

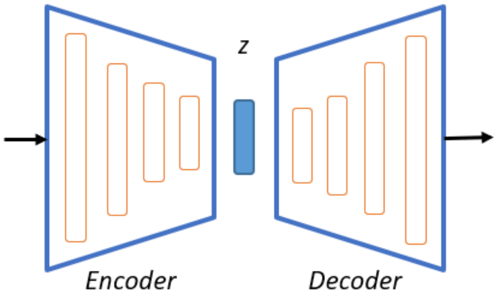


Figure 3.2. A simple autoencoder consists of an *encoder* and a *decoder*. In particular, the former is used to provide a compact or "compressed" version of the input, known as latent-space representation (z), while the latter aims to reconstruct the input starting from its encoded version.

the limited size of data. It consists of the creation of additional representative instances which simulate changes in the acquisitions and anatomical variations of patients by applying some transformation to the training images. This procedure prevents the model from specializing too much in the training set, improving its generalization ability and avoiding overfitting. In the literature, different basic data augmentation techniques are implemented in medical image analysis:

- Geometric transformations: they represent the most common techniques and include image scaling, translation, rotation, and shearing.
- Cropping: this technique randomly extracts patches from an existing image. These patches are then added to the original training set, increasing the size.
- Occlusion: it implements a concept similar to cropping since it selects patches from an image. However, they are removed to generate augmented images.
- Intensity operation: it manipulates the values of the pixels within the image by modifying the brightness or contrast.
- Noise injection: it simulates the noisy images by introducing some noise that is sampled from a random distribution.

- Filtering: this transformation is implemented by using a convolutional operation. By changing the filter an image can be sharpened, blurred, or smoothed to produce an augmented image.

- Combination: it creates a new image by combining two samples of the data set [160]. It implements a weighted combination of random image pairs from the training data.

Despite the presence of different techniques, basic data augmentation may not be sufficient in generating the variability needed to improve the generalization of the model. As a consequence, the *deformable augmentation techniques* simulate a range of possible variations during a clinical scan, such as patient movement, artifacts, or tissue deformation. In particular, the Randomised displacement field uses a generated deformation field to create a variation in the geometric shape, while the Spline interpolation consists of a mathematical operation that allows the smooth deformation of images. In applications where there is the need to compare different image modalities (i.e CT/MRI), a common use is the deformable image registration techniques that exploit the process of image registration modifying the content of an image in order to match it with a reference one[70]. The surprising results of DL in image processing have led several studies to exploit DL approaches for the implementation of data augmentation techniques. The purpose is to propose models able to create new realistic samples, namely *synthetic images*, learning the distribution of the input data. Most of the works in the literature focusing on medical imaging exploit Generative Adversarial Networks (GAN) that learn how to map a random distribution to the distribution of real samples [19].

3.3 Multimodal Deep Learning

The term *modality* refers to a single, specific source of data. In medical imaging, MRI, PET, or CT are examples of image modalities. The use of multiple and different sources provides complementary information that improves the detection and analysis of diseases, resulting in the need by physicians for techniques that aim to combine and simultaneously analyze different image modalities.

Multimodal Learning allows the fusion of complementary information coming from heterogeneous sources with the aim of providing a richer data representation than the unimodal approach. When investigated in conjunction with deep networks, multimodal learning is also known as Multimodal Deep Learning (MDL) [116]. In recent years, several data fusion techniques have been investigated in the research community [8, 62] resulting in three main levels: early fusion or data-level, late fusion or decision-level, intermediate or joint fusion.
The early fusion combines data from different sources, including techniques that aim to remove the correlation between them or propose a lower-dimensional common space. It exploits principal component analysis, independent analysis, and correlation analysis before using a classical ML algorithm. The late fusion began popular when researchers started to focus on ensemble learning, leveraging the ability of different models while the intermediate fusion directly follows from the layered structure of DNNs, providing a flexible approach with promising results in different contexts.

Early fusion consists of the integration of different and heterogeneous sources of data in a single structure that is then used as input for a DL model. In the case of medical image analysis using CNNs, the simplest strategy involves concatenating the acquisitions in a single volume. However, the inherent characteristics of each image modality, such as different resolutions or sampling times, may make the creation of a single structure very complex. As a consequence, it is possible to extract a high-level representation from each input modality before performing the fusion. In this case, handcrafted features and those extracted exploiting a DNN are merged in a single layer. Indeed, the simplest form to implement early fusion is to concatenate several features vectors as proposed in [112]. Since the features are first extracted from each imaging modality and then merged in a single structure, the early fusion may not fully take into account the complementary nature of the images, creating vectors with redundancy. This characteristic requires the implementation of techniques for features reduction like the principal component analysis.

Late fusion integrates the decision coming from different models, each trained on a specific image modality. In other words, this technique combines the decision of independent "experts", exploiting the fact that errors from multiple models tend to be uncorrelated. There are different combining strategies such as majority voting, averaged-fusion, Bayes'rule, or those exploiting the use of a meta-model. Indeed, late fusion started to be popular with the growth of solutions involving ensemble learning. It is worth noting that, among all multimodal fusion techniques, late fusion represents the simplest technique to implement.

Intermediate fusion exploits the characteristic of DNNs to be able to transform raw inputs into higher-level representations. As aforementioned, a DNN consists of several layers and the idea is to fuse in a single representation, known as shared representation, units coming from multiple modality-specific paths [116] (modality-specific DNN). In particular, in an MDL model, the representations for each modality are fused into a hidden layer, that is the *shared representation layer*. Since in DL approaches it is possible to implement end-to-end training, the resulting architecture autonomously learns the shared representation that well fits the specific task to solve. In the literature, the shared representation layer can be implemented according to the two main methodologies, proposed in [101] and [159] respectively. The first simple and effective approach implements a concatenation of the extracted vectors from a given layer of each unimodal model [101]. The resulting representation is processed by the following hidden layers. On the other hand, the approach described in [101] is a multiplicative method that computes the Kronecker product between all the modality-specific features vectors, extracted from a given hidden layer, considering also the unimodal representation. In the intermediate approach, it is not easy to understand when the modality-specific representation should be merged into a shared representation. In the literature, while different solutions proposed a single fusion layer, several approaches [60, 99] implement a gradual fusion strategy. The choice of which modality to fuse at which depth of representation can be very challenging, especially in cases where more than two sources are present.

DL-based multimodal learning offers several advantages when compared to classical ML techniques, where the data fusion usually consists of a handcrafted rigid architecture. ML approaches typically implement early or late fusion, explicitly performing features selection and dimensionality reduction. Since features are manually designed from domain experts requiring prior knowledge about the problem to be solved, the fusion techniques may be sensitive to data pre-processing. On the other hand, in MDL, the DNNs support the development of solution based on early, late and intermediate fusion and autonomously learn both the best features representation, making the features extraction step useless.

The medical field is a complex and heterogeneous scenario of multimodal sources, that require expert physicians for interpretation. The surprising results shown by DL approaches prompted many works to explore MDL in medical image analysis for tissue and organ segmentation [67], image registration [134] and CAD implementation [85].

Part II

Multimodal Data Sources

In the last years, Deep Learning (DL) approaches have been employed in medical imaging achieving promising results to support medical decisions [86]. A key role has been played by Convolutional Neural Networks (CNNs), which are networks made up of numerous convolutional layers that learn on their own the features that best match the particular problem to be solved. Moreover, the use of deep neural networks has also enabled the development of DL-based solutions in domains characterized by the need of leveraging information coming from multiple data sources, raising the Multimodal Deep Learning (MDL). The healthcare domain is characterized by the need to exploit information from heterogeneous and complementary sources for the prediction of different diagnostic outcomes. Data from various modalities related to the same phenomena require to be integrated or *fused* with the aim of providing a richer representation than the unimodal approach involving a specific source of information. Techniques for multimodal data fusion can be categorized into early, intermediate, and late fusion [116] and differ according to *when* the integration is performed.

Chapter 4 describes the proposed methodology involving the systematic analysis of early, late, and joint fusion techniques in medical image processing. In particular, the presented approaches are applied to two different cases of study, detailed in Chapters 5 and 6 respectively. The former exploits the presence of different sequences acquired during the same MRI exam, namely the Dynamic-Contrast Enhanced (DCE), the T2-weighted (T2), and the Diffusion-Weighted Imaging (DWI), while the latter considers two independent image modalities that are the T1-weighted (T1-w) MRI and the Pittsburgh Compound B (C-PiB) PET. As a consequence, Chapter 5 presents a scenario in which for each patient all three image modalities are available, offering an example with a *complete dataset* where complementary images obtained from the same exam are considered. In Chapter 6, instead, the independence between the T1-w MRI and C-PiB PET, which are acquired on different days, may cause the lack of a modality in some patients, resulting in a dataset with *incomplete acquisition*. As a result, data from multiple imaging exams are exploited.

Chapter 4

Techniques for Multimodal Data Fusion

Multimodal Learning allows the fusion of complementary information coming from heterogeneous sources (i.e different diagnostic tools) with the aim of providing a richer data representation than the unimodal approach. When investigated in conjunction with deep networks, multimodal learning is also known as Multimodal Deep Learning (MDL) [116]. Techniques for multimodal data fusion can be categorized into early, intermediate, and late fusion. Early fusion (EF) combines different modalities from a *data fusion perspective* [116], integrating different sources into a single structure. On the other hand, the late fusion (LF) acts in *decision-level*, in which the prediction is obtained by considering the output of different models, each trained on a single image modality. In this case, this fusion architecture aims to overcome the errors in the predictions, exploiting the uncorrelated characteristics of the involved classifiers. An alternative approach is joint or intermediate fusion (IF), which creates a *shared representation* by merging features coming from multiple modality-specific paths [116], resulting in a multi-input network that simultaneously processes different image modalities.

The techniques implemented for multimodal data fusion are described in the following sections.

4.1 Intermediate Fusion with Transfer Module

Intermediate fusion (IF), exploits the deep neural networks to transform raw inputs into higher-level and shared representations, which are constructed, for instance, by merging into a single layer, units coming from multiple modality-specific paths [116]. Among all deep neural networks, CNNs are widely used in biomedical image processing [86, 140] with surprising results.
A typical CNN consists of stacked *relatively complex layers* [11], with each of them having a convolutional stage, a non-linearity function (i.e ReLU), and a pooling operation. The set of *complex layers* constitutes the Convolutional Core (Conv-C), responsible for the features extraction step, while the classification is performed by a Classification Core (CC), typically including fully connected layers. In IF approach, the features maps coming from Conv-Cs related to different image modalities are merged into a single structure before feeding the CC, as shown in Figure 4.1.

Let j be a generic image modality, with j from 1 to M, representing the number of modalities, $ConvC_j$ be the Conv-C of the modality j, and L_j^i indicate the $i-th$ layer in the $ConvC_j$ with i from 1 to N_j that is the number of layers in $ConvC_j$. Denoting with x_j the input image belonging to the modality j, the resulting image-specific features map $F_j^{N_j}$ for each ConvC is formalized as follows:

$$F_j^{N_j} = ConvC_j(x_j) = L_j^{N_j}(...L_j^i(...L_j^2(L_j^1(x_j))...)...), \tag{4.1}$$

The IF approach exploits the characteristics of the neural networks to transform the inputs into higher-level representations with the aim of creating a *shared representation.* In particular, the fusion operation can be implemented by concatenating the extracted features maps $F_j^{N_j}$ [101], resulting in

$$F_{img} = [F_1^{N_1} \frown F_2^{N_2} ... \frown F_j^{N_j} ... \frown F_M^{N_M}] \tag{4.2}$$

where F_{img} is the *shared features map* and $\frown$ corresponds to the concatenation operation. The obtained representation is the input for the following layers, providing a unique prediction. However, Equation 4.1 highlights that each image-specific features map is computed by only considering a single modality. As a consequence, the complementary information coming from different sources is exploited after the concatenation operation reported in Equation 4.2, without affecting the features extraction process. During the training phase, the loss is back-propagated to all the convolutional cores, offering the only practical way to make the CNNs provide a shared representation that is well suited to the task to be solved. In this thesis, the aim is to propose an IF approach, in which the different specific-modality paths influence each other while extracting the feature maps of the various layers. Inspired by the works presented in [59, 48], we introduce a Transfer Module (TM) to implement a *cross-modality calibration* of the features. In particular, TM is inserted between layers belonging to different convolutional cores to take into account the complementary characteristics of the images in the determination of the resulting features maps. For each modality j, we denote the output of the L_j^i as $F_j^i \in \mathbb{R}^{X_j^i \times Y_j^i \times Z_j^i \times C_j^i}$, where $X_j^i \times Y_j^i \times Z_j^i$ is the spatial dimension, while ${C_j}^i$ is the number of channels. The TM inserted in the layer i is a multi-input multi-output module since it considers as input M features maps F_j^i and provides M outputs $\tilde{F}_j^i$, corresponding to the calibrated version of the features maps, that are computed in two different steps, namely the *Shared vector computation* and the *Multimodal calibration*, as summarised in Figure 4.2. The first one aims to determine the shared representation z_s^i considering the vectors computed from the features maps. As suggested in [48], the channel descriptor vector $s_j^i \in \mathbb{R}^{1 \times C_j^i}$ for each features map is obtained by using a Global Average Pooling operation $G(\cdot)$. The concatenation of the three s_j^i determines the vector $z^i \in \mathbb{R}^{1 \times C^i}$, where $C^i = \sum_{j=1}^{M} C_j^i$. Then, z^i is further processed by considering it as input for the fully connected layer f_c, followed by the ReLU function $R(\cdot)$, that introduces a nonlinear map between the elements of the input vector. The result is the shared representation $z_s^i \in \mathbb{R}^{1 \times C^t}$, where $C^t = C^i/4$ as suggested in [59]. The *Shared vector computation* can be summarised as follows:

1. s_j^i, with j from 1 to M
2. $z^i = [s_1^i \frown s_2^i ... \frown s_j^i \frown s_M^i]$
3. $z_s^i = R(f_c(z^i))$

The *Multimodal calibration* uses the shared representation to calibrate the features maps F_j^i, thus exploiting information coming from different data sources. z_s^i is considered as input for M fully connected layers f_j, each for a specific modality j. As suggested in [59, 48], the Sigmoid activation function $\sigma(\cdot)$ is used to force the output in the range [0,1], resulting in M calibration vectors $c_j \in \mathbb{R}^{1\times C_j^i}$. The features maps $\tilde{F}_j^i$ are obtained by implementing a channel-wise product between the input F_j^i and the corresponding c_j, creating a gating mechanism, where the contribution of the selected filters is reduced. This allows each F_j^i to be influenced by the others during the features extraction step. The *Multimodal calibration* step can be summarised as follows:

1. $c_j = \sigma(f_j(z_s^i))$ with j from 1 to M
2. $\tilde{F}_j^i = c_j \odot F_j^i$ with j from 1 to M, where $\odot$ represents the channel-wise product

Figure 4.3 shows the TM inserted between layers belonging to the three convolutional cores. It is worth noting that the module can be inserted at any level, without particular constraints on the architectures of the networks involved.

The *shared features map* F_{img}, obtained by concatenating the outputs of each Conv-C [101] according to Equation 4.2, is further processed by the following layers with the aim of providing a unique prediction. In particular, the *shared layer* L_s receives F_{img} as input, generating the features vector F_s that is used to feed the CC, as shown in Figure 4.4.

In healthcare, images belonging to different patients acquired with diverse diagnostic tools are often associated with clinical features (CL) in the form of tabular data, resulting in the need of including them in the IF approach. As a consequence, a Multilayer Perceptron MLP is used to process CL obtaining the features vector F_{cl}. Then, $\dot{F}_{img}$ that is the output of L_s is concatenated with F_{cl} resulting in the features vector F_s, as described in Figure 4.5, representing the input for the CC.

4.2 Early Fusion

In the early fusion (EF) approach the M image modalities are organized in a single structure before being considered as an input for the single classifier. In the case of medical image analysis using CNNs, the simplest

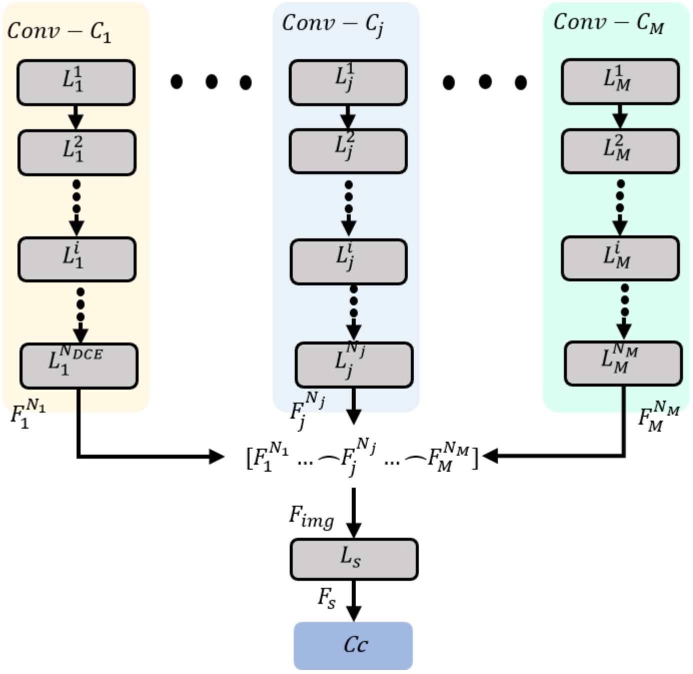


Figure 4.1. Basic schema for Intermediate Furion approach: j is a generic image modality, with j from 1 to M, representing the number of modalities; $ConvC_j$ is the Conv-C of the modality j; and L_j^i indicates the $i-th$ layer in the $ConvC_j$ with i from 1 to N_j that is the number of layers in $ConvC_j$; $F_j^{N_j}$ is the resulting image-specific features map for each ConvC

strategy involves concatenating the acquisitions in a multi-channels volume, as described in Figure 4.6. However, the inherent characteristics of each image modality, such as different resolutions or sampling times, may make the creation of a single structure very complex, especially when one or more sources involve tabular data (i.e clinic features). As a consequence, higher-level representations, which might be either manually created features or learnt representations (i.e CNN used a features extractor) can be extracted from each modality first and then combined before being fed into ML model to help with some of the problems associated with fusing raw data [116].

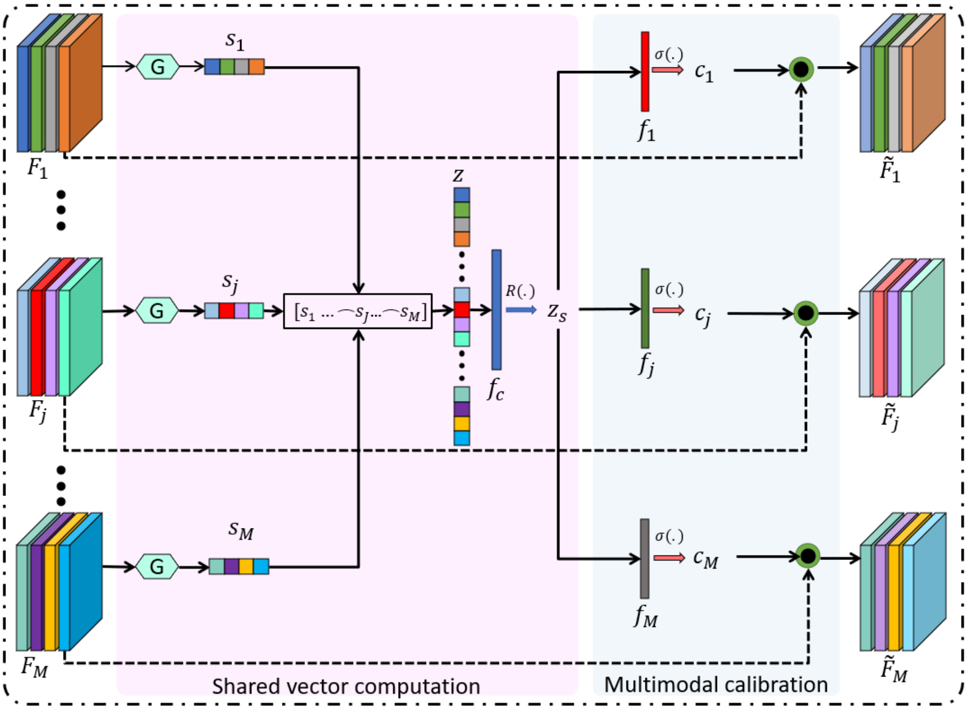


Figure 4.2. Architecture of the proposed TM module for M different image modalities. The input consists of the features maps coming from the modality-specific paths, that are further processed by considering two steps acting in a pipeline, namely the *Shared vector computation* and the *Multimodal calibration*. It is worth noting that the index i of the layer in which the module is inserted is omitted to avoid overly complex notation in the image.

4.3 Late Fusion

The late fusion (LF) completely relies on the unimodal approach since it aggregates the prediction coming from the classifiers implemented for each modality separately, as shown in Figure 4.7. To this goal, the predictions coming from different models are combined using a voting strategy. Among all the combining rules, the Weighted Majority Voting (WMV) is used, in which a weight is assigned to each prediction according to the model output probability. However, in LF each classifier acts independently, not taking

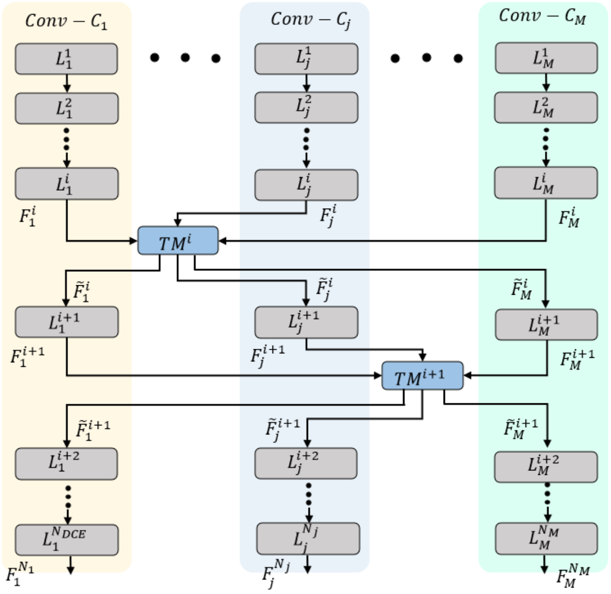


Figure 4.3. TM inserted between layers belonging to M convolutional cores.

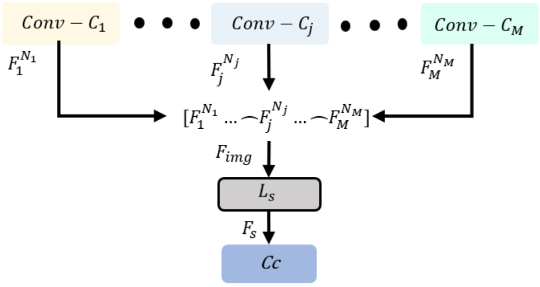


Figure 4.4. Intermediate fusion approach with M image modalities. Each Conv-C is further detailed in Figure 4.3 and represented as a colored box to avoid making the figure too complicated.

advantage of the complementary characteristics of the different modalities that do not influence each other during the prediction.

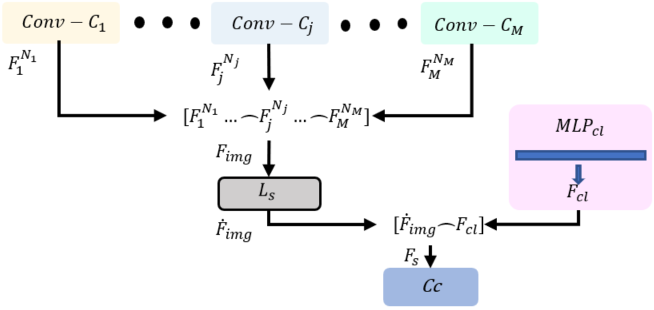


Figure 4.5. Intermediate fusion approach with M image modalities and clinical features. Each Conv-C is further detailed in Figure 4.3 and represented as a colored box to avoid making the figure too complicated.

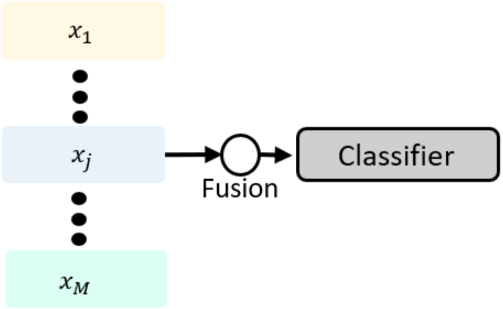


Figure 4.6. Early fusion approach: x_j is the input image belonging to the modality j

4.4 CNN architecture

In this thesis, two different deep models are presented for the convolutional part (Conv-C) and for the classification part (CC) of each architecture, regardless of the specific data modality considered as input and regardless of the type of fusion adopted, i.e., early, late and joint. Such two networks are named in the following as BasicNet and ResNet.
BasicNet consists of a set of *reduction layers* followed by a CC including fully connected layers. A reduction layer is a block with a 3D-convolutional operation, followed by batch normalization and ReLU function. Each convolution reduces the input feature map and doubles the number of

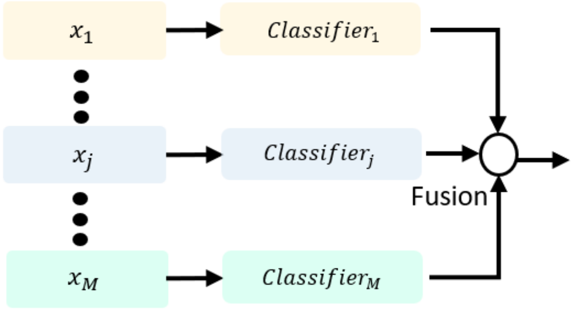


Figure 4.7. Late Fusion approach: x_j is the input image belonging to the modality j, while $Classifier_j$ is the model trained with x_j samples

channels, and the chain of reduction layers constitutes the Conv-C of the network, which is responsible for the features extraction step. In the implemented methodology BasicNet is trained from scratch.

Inspired by the state-of-art network proposed in [46], ResNet is a 3D CNN, adapted to handle 3D volume data. Furthermore, the solution implemented in [18] offers a set of 3D ResNet architectures pre-trained with medical images for segmentation tasks. The architecture consists of a backbone, which is an encoder representing the ResNet basic structure, responsible for features extraction, and a set of decoders for the generation of the segmentation masks. Hence, the ResNet is adapted for the classification task as proposed in [18]: the backbone is retained while the set of decoders is replaced with a CC consisting of a global average pooling and a fully connected layer. Figure 4.8 shows the resulting ResNet architecture highlighting the backbone, that is the Conv-C, consisting of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four blocks containing the layers implementing the residual network as proposed in [46]. Differently from [18], the stride of the convolution kernels in blocks 3 and 4 is restored to 2 allowing downsampling the features maps to better tailor the network to a classification problem.

Note that on this one side, transfer learning is used, considering the pre-trained backbone as a starting point and implementing the fine-tuning for adapting the ResNet for the specific task to solve. On the other hand, the classification core is trained from scratch.

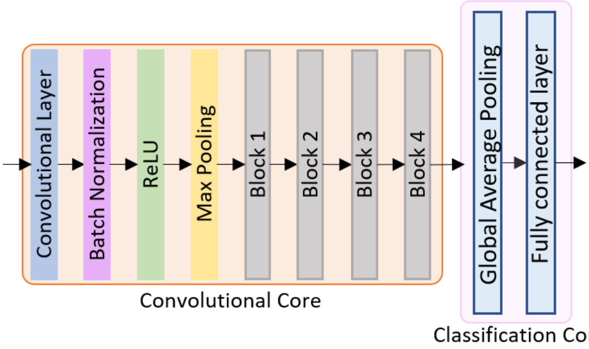


Figure 4.8. ResNet architecture: the Convolutional Core (Conv-C) consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four layers containing the blocks implementing the residual network, while the Global Average Pooling and the fully connected layer represent the Classification Core (CC)

Chapter 5

Multiple Data from an Imaging Exam: Axillary Lymph Node Status Assessment

Among women, breast cancer (BC) is the most frequent form of tumor [119], and the axillary lymph nodes status (ALNS) is considered a crucial indicator, representing one of the most influencing and independent prognostic factors [90]. All the procedures involved in the evaluation of the axillary cable are invasive and may have long-term effects. No imaging modality, including Ultrasound, Computer Tomography, and Magnetic Resonance, has been shown to be particularly accurate [130, 150, 108] and the obtained morphological characteristics do not seem to be adequate enough to distinguish between normal and malignant lymph nodes. Magnetic Resonance Imaging (MRI) is always performed for BC stage definition [95, 22, 28] and plays a key role in primary tumor examination since it provides both qualitative and quantitative information. The most important MRI sequence is the dynamic-contrast enhanced (DCE) that, thanks to the high contrast resolution, provides information about the tumor morphology, size, and perfusional behavior allowing the distinction between benign and malignant lesions, the prediction of biological aggressiveness and the prognostic evaluation [95, 149]. The T2-weighted (T2) imaging is a standard component of breast MRI exams. It is most prominently utilized for the identification of cysts but also contributes to the characterization of lesions in the evaluation of their aggressiveness [88, 97]. In particular, T2 sequence allows a better depiction of lesion morphology and perifocal or prepectoral edema within the breast, which improves the distinction between a malignant and benign tumor. Moreover, Santucci et al. demonstrated in [124] that the evaluation of the edema increases the accuracy in the prediction of prognostic factors. The Diffusion-Weighted Imaging (DWI) is another sequence acquired during the MRI exam. It reflects the mobility of water molecules diffusing in tissues, revealing tissue organization at the microscopic level and providing complementary information for lesion assessment in comparison with the DCE scan.

Assessment of axillary lymph nodes (ALNS) indicates inherent primary tumour properties, whose examination enables the discovery of minimally invasive solutions for the sentinel node biopsy currently being utilized. This thesis exploits DL approaches, for ALN metastasis prediction focusing on different MRI acquisitions. The DNNs, i.e CNNs, remove the need for the features extraction stage that is characteristic of ML approaches by autonomously learning the collection of features that best matches the particular problem to solve.

In particular, in this work, the presence of multiple and complementary sequences acquired during the MRI exam, namely DCE, T2 and DWI, enables the evaluation of multimodal deep learning approaches to exploit information coming from heterogeneous sources.

5.1 Radiomic-based approach for axillary lymph nodes evaluation

The majority of literature solutions for DCE sequences use radiomics to extract handcrafted characteristics from breast lesions while classifying the data with ML techniques. The absence of a clearly defined, practical collection of features in the area of BC has, however, prompted researchers to investigate broad and varied features computed from the main tumor and then pick the best discriminatory ones. While many studies utilize radiomics of the MRI primary tumor to predict the histological nature [78], there are still relatively few studies that use this data to predict the condition of the axillary cavity [14]. The work known to date are highly heterogeneous in terms of the features extraction/selection step and the trained classifiers. The majority of the proposals focus on first-order, morphological, shape, texture, and first-order characteristics when extracting handcrafted features. Additionally, the approach suggested in [84] makes use of pharmacokinetic variables, whereas a prior study [125] investigated the 3D extension of Local Binary Patterns (LBPs) to enhance texture description. On the one side, shallow learners are taken into account by several methods in the literature to forecast ALN metastasis using radiomics, including Support Vector Machines [83, 25, 44], Logistic Regression [84, 82, 33], Linear Discriminant Analysis [17, 7], and Random Forest [125, 123, 23]. The majority of research only examines the primary tumor's imaging features and excludes the tissue around it. However, research has demonstrated that the peritumoral area has important clues about the possible aggressiveness of the tumor and lymphatic dissemination [65, 124, 139]. As a result, while assessing the ALN condition, the tissue around the breast lesion should be taken into consideration. In a previous study [125], authors used patient clinical data with information on the primary breast tumor's histology and MRI radiomics characteristics (First-Order, 3D Gray Level Co-Occurrence Matrix, Three Orthogonal Planes-Local Binary Patterns) to predict LN metastasis. They took into account both breast lesions and peritumoral areas for the features extraction step since the tissues around tumors can be used to determine how aggressive they are. Before utilizing a Random Forest (RF) classifier to make the prediction, the wrapper features selection approach was required due to the large dimensionality of the problem to be solved (257 features).

On the other hand, not many studies investigate CNNs to forecast ALN state. Some of them make use of breast ultrasonography [79, 164, 162], while the MRI-based studies take into account images of axillary lymph nodes without primary tumor [120, 42]. The research proposed by Nguyen et al. [100] is the first effort to establish if ALN metastasis can be predicted using the preoperative DCE-MRI of the primary tumor and CNNs. In that study, DCE-MRI images are processed using a 3D CNN in a subtractive approach employing the third, fourth, and fifth post-contrast volumes. Each DCE-MRI data is cropped using a 3D cuboidal bounding box of size $50 \times 50 \times 50$ that includes the tumor area. DCE-MRI images and clinical data are used in the methodology suggested in [100], which is the first work exploiting the multimodal approach.

5.2 Population

All breast MRI exams were performed for preoperative evaluation at the Central Radiology Department of Policlinico Umberto I between January 2017 and January 2020 and retrospectively reviewed. A written informed consensus was obtained before the execution of a contrast-MRI for all examinations. All patients met the following inclusion criteria: three Tesla magnetic field MRI examinations, post-contrast sequences, mass-like tumors, histopathological confirmation of invasive breast cancer, a complete histological analysis, and definitive lymph-node status of the ipsilateral axilla. In cases of BC bilateral lesions, the two lesions were evaluated separately since the two breasts can be considered a single part. Patients with an incomplete MRI examination or damaged images and patients without a complete histopathological analysis were excluded. The patients were excluded if they had breast implants or expanders, were in follow-up neo- or adjuvant chemotherapy, or the MRI images were not of excellent diagnostic quality.

A total of 153 patients (average age 55 years; range 30–85) met the inclusion criteria. In two patients who had bilateral breast cancer, the two breasts were considered as a single, independent part. Therefore, a total of 155 malignant breast cancer lesions were included in total. The LNS status was assessed as positive if at least one lymph node involved by metastasis was present in the definitive histopathological axillary cable sample (LN+); the LNS was considered negative if all axillary lymph nodes were safe (LN−).

All MRI examinations were performed using a 3T magnet (Discovery 750; GE Healthcare, Milwaukee, WI, USA). Three orthogonal localizer sequences were employed, then images were acquired following this protocol:

- T2-weighted axial single-shot fast spin echo sequence with a modified Dixon technique (IDEAL) for intravoxel fat-water separation (TR/TE 3500–5200/120–135 ms, matrix 352 × 224, FoV 370 × 370, NEX 1, slice thickness 3.5 mm).

- Diffusion weighted axial single-shot echo-planar with fat suppression sequence (TR/TE 2700/58 ms, matrix 100 × 120, FOV 360 × 360, NEX 6, slice thickness 5 mm) with diffusion-sensitizing gradient applied along the three orthogonal axes and with a b-value of 0, 500, and 1000 s/mm2.

- T1-weighted axial 3D dynamic gradient echo sequence with fat suppression (VIBRANT) (TR/TE 6.6/4.3 ms, flip angle 10°, matrix 512 × 256, NEX 1, slice thickness 2.4 mm), before and n times after intravenous contrast medium injection, with n from 5 to 12. An amount of 0.2 mmol/kg of Gadobenate-dimeglumine (Multihance®; Bracco Imaging, Milan, Italy) was used as the contrast agent, injected through a 20G intravenous cannula at a rate of 2 mL/s plus 15 mL of saline solution at the same speed. For each acquisition, the relative subtracted images were automatically generated and used for tumor analysis.

The images were analyzed by two radiologists with 10 and 3 years of experience, respectively. The tumors were described as unifocal when only one lesion was present; multifocal when more than one tumor lesion was present in the same breast quadrant/region; and multicentric when multiple tumor lesions were present in different breast quadrants/regions. For each lesion, the target dimensions, margins (regular, irregular, lobulated, or spiculated), and intensity signal timing curve (I, II, or III, based on wash-in and wash-out) were reported.

The patient's clinical data were collected, and, according to these data, the population was split into subgroups: age, familiarity (considered positive if at least one familiar member was affected by breast cancer at any age), hormone therapy (considered positive if the patient performed at least 3 continuous months of hormone therapy including any kind contraceptive, replacement, or therapeutic therapy), and menopausal status.

The samples were obtained by a core biopsy or surgery and analyzed by an anatomic pathologist with more than 15 years of experience. The tumor histotype classification followed the WHO classification [136]. The tumor histological grade was assigned in accordance with the NGS, and a score from one to three was given for these tumor characteristics: tubular formation, nuclear pleomorphism, and the number of mitoses. Furthermore,

Class	Details
Clinical	age, familiarity, hormone therapy, menopausal status
Histological	ER, PgR, HER2, ki-67, grading, tumor class, histotype
Image-Derived	visibility on T2, visibility on DWI, apparent diffusion coefficient (ADC), signal timing curve, dimensions, margins, breast quadrant, multifocality

Table 5.1. Details about the collected *clinical features*

the estrogen receptor (ER), progesterone receptor (PgR), human epidermal growth factor receptor (HER2), and the proliferation index Ki67 were assessed for immunohistochemical analysis. A cut-off of 10% was used to consider the ER and PgR expression as positive; while HER2 was considered positive when $>+2$, and ki67 was considered positive when $>14\%$. Moreover, other histological data were collected: histotype (including ductal carcinoma (IDC) and invasive lobular carcinoma (ILC)), grading (divided into G1, G2, or G3), and tumor class, which includes the hormone receptor status and the proliferation index percentage (Luminal A: ER+, HER2− and low ki67; Luminal B: ER+, HER2 −/+ and high ki67; HER2 overexpressed; Triple Negative (TN): ER−, PgR−, HER2−).
Table 5.1 details the features acquired for each patient. It is possible to distinguish three classes of features, namely clinical, histological and image-derived features. For the sake of simplicity, the three classes are indicated with *clinical features*.

5.2.1 Axillary Lymph Node Status

The axillary lymph node status was considered as the final output. The LNS was assessed after an invasive breast cancer diagnosis using definitive surgical characterization (sentinel node dissection, sampling dissection, or total lymphadenectomy, based on surgeon decision but curative in all cases). The LNS was simply classified as positive, if there was at least a sentinel LN involved, or negative if there was no positive lymph node. On this basis, the dataset accounts for 27 positive and 128 negative patients, which are referred to as LN+ and LN− in the following.

5.2.2 Segmentation

The images were anonymized and uploaded on a dedicated open-source software (3D Slicer, version 4.8, November 2012). An identification number (ID) was assigned to each patient. Bilateral tumors were considered with two different IDs. For each case, the subtracted post-contrast T1w-MRI was selected. The second phase (60–120 s) was selected for ROI segmentation, due to its higher contrast resolution. Then, a label map was created. The lesions were manually drawn through manual and assisted thresholding segmentation techniques on the axial projection . When present, necrosis was avoided by segmentation. For multifocal or multicentric tumors, all lesions, even the smallest, were segmented.

5.3 Methodology

In this thesis, a solution based on MDL approach is proposed for the ALN status assessment, which considers complementary image acquisitions, anamnestic information, and histological characteristics of the primary tumour. The implemented methodology consists of two main steps: the *Pre-processing*, used to prepare data belonging to DCE, T2, and DWI sequences of different patients, the *IF approach for ALN status assessment*, introducing the implemented methodology based on intermediate fusion technique to exploit the characteristics of the neural networks to transform the inputs into higher-level representations with the aim of creating a *shared representation*.

5.3.1 Pre-processing

In MRI examination, the DCE requires the intravenous administration of a contrast agent (CA), characterized by specific wash-in and wash-out times. Indeed, the DCE is composed of MRI scans that were acquired at different times and were obtained before (pre-contrast) and after (post-contrast) CA injection. For each patient p, the result is a 4D data with three (x,y,z) spatial and one temporal (t) dimensions, represented by a volume of size $X^p \times Y^p \times Z^p \times N^p$ where $X^p \times Y^p \times Z^p$ is the size of each acquired MRI image and N^p the number of acquisitions. Denoting with t_i the i_{th} acquisition, where t_0 and t_{N^p-1} are the pre-contrast and the last acquisition in the sequence of MRI images respectively, the subtractive series is obtained by considering $t_i - t_0$ with i raging from 1 to $N^p - 1$. Although MRI exams are acquired with the same instrument, patients may present a different number of acquired volumes, resulting in the need of selecting a subset of them. Four specific subtractive volumes are selected: the first (t_1), the second (t_2), the last (t_{N^p-1}) volume, and the median index (t_m) volume between the third and the second-to-last volume. In this way, information about the wash-in and wash-out of CA flowing is preserved. The result of the pre-processing step for each patient p is a 4D volume of size $X^p \times Y^p \times Z^p \times 4$, obtained by considering the four subtractive acquisitions (t_1,t_2,t_m,t_{N^p-1}) from her DCE series.

In the study proposed in [126], different tumor bounding options are analyzed considering the characteristics of the DCE dataset for the assessment of ALN status. In particular, the work described in [126] evaluates how the amount of the included non-tumor tissue impacts the ALN assessment, taking into account that patients differ in size and number of lesions. A total of six different tumour bounding options are proposed, namely *Single Fixed-size Box*, *Single Variable-size Box*, *Single Isotropic-size Box*, *Single Lesion Variable-size Box*, *Single Lesion Isotropic-size Box* and *Two-Dimensional Slice.*
In *Single Fixed-size Box (SFB)* a fixed-size 3D bounding box centered in the lesion center and completely encompassing the tumour region is used to crop each subject. It is applied to each of the four subtractive selected acquisitions (t_1,t_2,t_m,t_{N^p-1}) and is patient-independent.
To limit the amount of non-tumour tissue to include, in *Single Variable-size Box (SVB)*, the smallest 3D rectangular bounding box circumscribed to the tumour region is considered for each patient. In contrast to SFB, the cubical box in SVB option is patient-dependent, and the amount of non-tumour tissue depends on the shape of each patient's tumour region. However, in the case of multifocal and multicentric tumours, the parenchyma between lesions is included in the extracted 4D volume.
Single Isotropic-size Box (SIB) considers the fact that different patients may show acquisitions that vary in terms of resolution, namely the information regarding the measurement of each voxel in millimetres (mm). As a consequence, all the volumes are re-sampled to obtain isotropic voxels with dimension $1 \times 1 \times 1 mm^3$, before selecting the smallest box surround-

ing the tumour area.
Single Lesion Variable-size Box (SLVB) analyses how much the parenchyma between tumours impacts the ALN assessment, applying the SVB option to each lesion of the patient p. Hence, a box for each lesion is extracted and therefore, the tissue between lesions is excluded in the prediction of ALN status.
The *Single Lesion Isotropic-size Box (SLIB)* acts as SLVB with the difference that for each patient the 3D volumes re-sampled to obtain isotropic voxels.
Finally, the *Two-Dimensional Slice (2DS)* option proposes to apply the SVB procedure and then to cut the sequence of the 3D cropped volumes along the projection with the highest spatial resolution, resulting in a series of two-dimensional slices with four temporal instants.
According to [126], the SIB is considered the best bounding box option since it focuses on the smallest area surrounding the tumor while reducing the impact of the different resolutions among the three axes.

Differently from DCE, T_2 and DWI scans consist of the acquisition of a 3D volume without the temporal information, which can be considered as a 3D image with 1 channel. Since the breast lesion is segmented considering the second post-contrast volume, the T_2 and DWI acquisitions are aligned to the segmentation mask generated from the DCE volume exploiting the *Slice Location* attribute that corresponds to the relative position of each slice. In particular, slices belonging to different acquisitions with the same *Slice Location* value are considered aligned. As a consequence, it is possible to apply the segmentation mask to all the sequences by taking into account only the aligned slices. Moreover, before applying the information about the lesion localization, the DWI scan is co-registered to the DCE volume considering the second post-contrast acquisition as a reference and using mutual information as a similarity metric [92, 115].
The three image modalities differ in terms of resolution, that is the information associated with each voxel in millimetres (mm), resulting in the need of re-sampling all the volumes to obtain isotropic voxels with dimension $1 \times 1 \times 1 mm^3$. Furthermore, according to the results reported in [126], the SIB bounding option is selected and applied to DCE, T_2, and DWI volumes.

5.3.2 IF approach for ALN status assessment

In clinical trials, the DCE, T_2 and DWI acquisitions are used to evaluate the aggressiveness of the primary malignant tumor. In this thesis, the different and heterogeneous information provided by the three image modalities is exploited for the prediction of ALN metastasis. In particular, a solution based on Intermediate fusion (IF) is proposed, considering complementary image acquisitions, anamnestic data, and histological characteristics of the primary tumor. Moreover, CNNs are exploited, by implementing a Multi-Input Single-Output network that simultaneously processes DCE, T_2 and DWI acquisitions and clinical features (CL) in the form of tabular data.
As reported in Section 4.1, each image modality is associated with a specific Convolutional Core (Conv-C), resulting in architecture with three Conv-Cs, responsible for the features extraction step.

Let $j \in \{DCE; T_2; DWI\}$ represent single image modality, $ConvC_j$ be the Conv-C of the modality j, N_j denote the number of layers in each Conv-C, x_j be the input image belonging to the modality j, the Equation 4.1 presented in Section 4.1 that computes the resulting image-specific features map is formalized as follows:

- $F_{DCE}^{N_{DCE}} = ConvC_{DCE}(x_{DCE})$;
- $F_{T_2}^{N_{T_2}} = ConvC_{T_2}(x_{T_2})$;
- $F_{DWI}^{N_{DWI}} = ConvC_{DWI}(x_{DWI})$;

The fusion operation implemented according 4.2 results in $F_{img} = [F_{DCE}^{N_{DCE}} \frown F_{T_2}^{N_{T_2}} \frown F_{DWI}^{N_{DWI}}]$, that is the *shared features map*. To make the different specific-modality paths, represented by the different convolutional cores, influence each other while extracting the feature maps of the various layers, the Transfer Module (TM) introduced in Section 4.1, is included implementing the *cross-modality calibration* of the features. Figure 5.1 shows the TM customized for the specific task to solve presented in this thesis, detailing the three input features maps and the two processing steps, namely the *Shared vector computation* and the *Multimodal calibration.*

Before processing the *shared features map* F_{img} with the aim of providing a unique prediction, it is necessary to merge the information coming

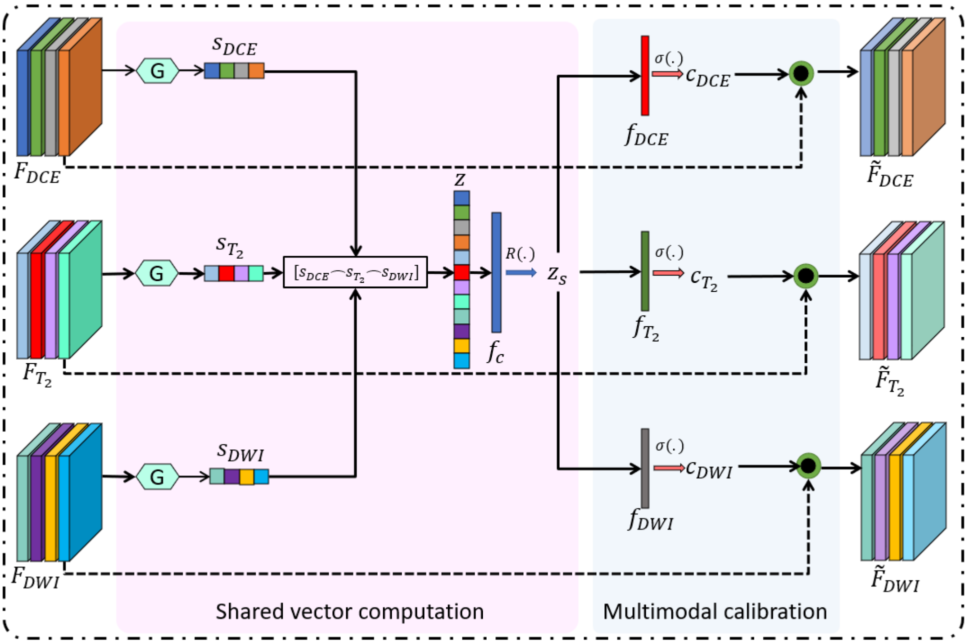


Figure 5.1. Architecture of the proposed TM module for three different image modalities. The input consists of the features maps coming from the modality-specific paths, that are further processed by considering two steps acting in a pipeline, namely the *Shared vector computation* and the *Multimodal calibration*.

from the images with the clinical features (CL) acquired as described in Sections 5.2. In particular, all the collected data are described in Table 5.1 consisting of 19 features. However, a subset of them is selected with the domain expert (D.S), obtaining the subset CL-S1. The basic idea is to remove all the features which can be derived from the images, retaining those that provide additional information to the acquisitions. Inspired by [100], the CL-S1 includes age, familiarity, hormone therapy, menopausal status, dimensions, ER, PgR, HER2, ki-67, and grading, resulting in a set of 10 features.

A Multilayer Perceptron (MLP) is used to process CL-S1 obtaining the features vector F_{cl}. Then, $\dot{F}_{img}$ that is the output of L_s is concatenated with F_{cl} resulting in the features vector F_s, as described in Figure 4.5, representing the input for the CC.

CNN Architectures

In this thesis, two different deep models are presented for the convolutional part (Conv-C) and for the classification part (CC) of each architecture, regardless of the specific data modality considered. Such two networks are named in the following as BasicNet and ResNet.
BasicNet consists of five reduction layers followed by a CC including fully connected layers. The input volume represents the smallest cubical box surrounding the tumour region. A reduction layer is a block with a 3D-convolutional operation, followed by batch normalization and ReLU function. The chain of five reduction layers constitutes the Conv-C of the network, which is responsible for the features extraction step. Each reduction layer consists of a 3D convolutional layer with $4 \times 4 \times 4$ kernels (the number of kernels depends on the output channels) and a stride set to 2 in order to extract features from the input volume and at the same time to have a gradual dimensionality reduction. The padding is set to 1 in each layer, excluding the last one where it is set to 0. Moreover, each layer doubles the number of channels, while the convolutional operation in the first block presents 8 output channels. The classification core consists of two fully connected layers, resulting in the architecture shown in Figure 5.2. In the implemented methodology BasicNet is trained from scratch.
Inspired by the state-of-art network proposed in [46], ResNet is a 3D CNN, adapted to handle 3D volume data. Furthermore, the solution implemented in [18] offers a set of 3D ResNet architectures pre-trained with medical images for segmentation tasks. The architecture consists of a backbone, which is an encoder representing the ResNet basic structure, responsible for features extraction, and a set of decoders for the generation of the segmentation masks. Hence, the ResNet is adapted for the classification task as described in Section 4.4. Figure 5.3 shows the resulting ResNet architecture highlighting the backbone, that is the Conv-C, consisting of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four blocks containing the layers implementing the residual network as proposed in [46]. The CC consists of a global average pooling and a fully connected layer.
Note that on this one side, transfer learning is used, considering the pre-

trained backbone as a starting point and implementing the fine-tuning for adapting the ResNet for the specific task to solve. On the other hand, the classification core is trained from scratch.
In both networks, the number of input channels of the first convolutional layer depends on the imaging modality. In particular, the first layer in the $ConvC_{DCE}$ presents 4 channels, while $ConvC_{T_2}$ and $ConvC_{DWI}$ consider 1 channel volumes as input.
The MLP used to process CL-S1 presents a fully connected layer with 10 input and 4 output features, followed by the ReLU function. In the multimodal scenario, described in Figure 5.4, the first concatenation operation integrates $F_{DCE}^{N_{DCE}}$, $F_{T_2}^{N_{T_2}}$, $F_{DWI}^{N_{DWI}}$, obtaining the F_{img} features map, with C channels, where $C = \sum_{j=1}^{M} C_j$ and C_j is the number of channels of the features map related to the modality j at the end of its convolutional core. Then, the *shared image features map* $\dot{F}_{img}$ is generated by the block L_s that includes a convolutional operation with a number of input and output channels set to C, a $1 \times 1 \times 1$ kernel, values of stride and padding set to 1 and 0 respectively, followed by batch normalization and ReLU function. Moreover, when the ResNet architecture is exploited for the ConvC, L_s also includes a Global Average Pooling layer to obtain a features vector. The second concatenation operation merges the $\dot{F}_{img}$ with features vector F_{cl} coming from the MLP that processes the clinical information. The resulting representation F_s is considered as input for the CC, which consists of two fully connected layers spaced by ReLU function. In particular, the first layer in the classification core considers an input features vector of $C + 4$ elements, with $C/3$ output features, while the second one is responsible of the prediction, presenting two output neurons.

5.4 Experimental Set-up

As reported in Section 5.3.1, the selected tumor bounding option (SIB) considers a box whose size varies according to each patient's region of interest. Moreover, the presence of multiple image modalities with different resolutions causes the creation of volumes with different dimensions. As a consequence, a resize stage is used to give the volumes a common size of $64 \times 64 \times 64$, before feeding them to the involved CNNs. In experiments involving ResNet, the ResNet10, the ResNet18, the ResNet34, and the

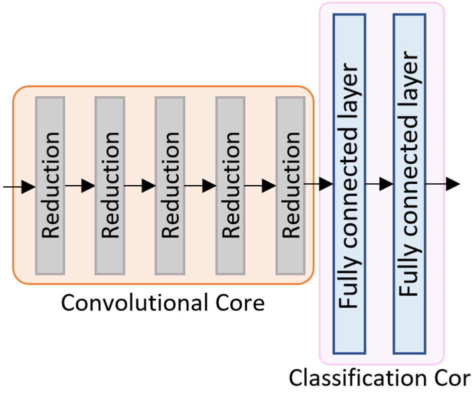


Figure 5.2. BasicNet architecture: the network consists of five *reduction layers*, representing the Convolutional Core (Conv-C), followed by two fully connected layers spaced by ReLU function, constituting the Classification Core (CC)

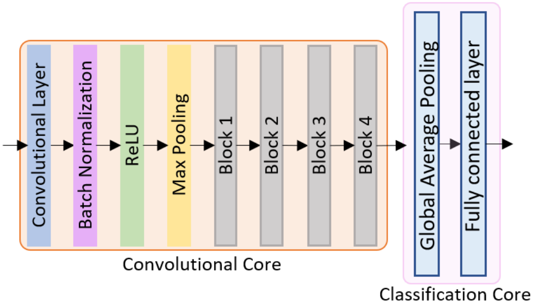


Figure 5.3. ResNet architecture: the Convolutional Core (Conv-C) consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four layers containing the blocks implementing the residual network, while the Global Average Pooling and the fully connected layer represent the Classification Core (CC)

ResNet50 architectures are used.
In this thesis, a complete set of experiments is provided since the study focuses on IF technique in which the TM is inserted and evaluates the unimodal (U) approach and the other multimodal fusion strategies, namely the Early Fusion (EF) and the Late Fusion (LF).

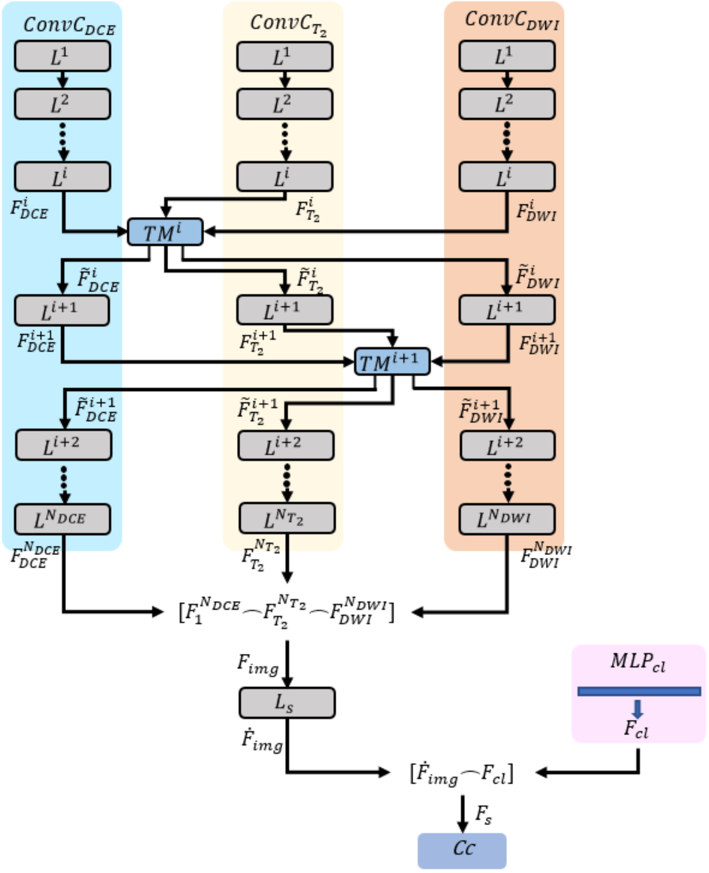


Figure 5.4. Architecture of the IF approach including TM between layers belonging to different convolutional cores and clinic features

In the Unimodal Approach (U), the heterogeneous image modalities and the clinical features are exploited to build different models that do not co-operate for the determination of a single prediction. Indeed, the result is a set of 4 classifiers, each of them trained on a specific source of data. In the case of MRI sequences, the BasicNet and the ResNet are trained considering the DCE, the T2, and the DWI volumes separately, obtaining the U(DCE), U(T2), and U(DWI) configurations. As reported in Section 5.2, clinical features are reported in the form of tabular data.

This thesis focuses both on CL and CL-S1 sets, providing a model for each of them and obtaining the U(CL) and U(CL-S1) configuration.
The CL set consists of 19 features described in Table 5.1, including both numeric and categorical information that is used to train a ML model. In the pre-processing step, the categorical features are transformed into numerical data by exploiting the Integer Encoding if the values of the features have a natural ordered relationship with each other (i.e grading) and the One-Hot Encoding, if there is not a relationship between the categories (i.e margin). The Support Vector Machines Classifier (SVM) [24] is the selected ML model, widely used in the literature [83, 25, 44] for the ALN status evaluation. The features selection step is performed by implementing a backward elimination, while the Adaptive synthetic sampling (Adasyn) [45] algorithm is exploited to handle the high imbalance between the LN+ and LN- classes. The best SVM hyper-parameters settings are determined through a bayesian optimization stage varying the SVM kernel, the polynomial order in [1;5], and the cost value in [0;1.5]. The search reported a cost value equal to 0.347 and a kernel set to linear. During the steps for features selection and hyper-parameter optimization, the objective function aims to maximize the Area under the ROC curve.
The CL-S1 set consists of 10 features, that, similarly to the process described in Section 5.3.2, are considered as input for a MLP, consisting of two fully connected layers, spaced by ReLU function. The first hidden layer presents 4 output features, while the output consists of two neurons.

In the Early Fusion (EF) the different image modalities are organized in a single structure before being considered as an input for the single classifier. The simplest strategy involves concatenating the acquisitions in a multi-channels volume. However, the described fusing approach can not be applied in this work for two main reasons: i) it is not possible to integrate clinical features, ii) the characteristic of the DCE to be an image in which the temporal information is concatenated on the channels makes the idea of concatenating all the images along the channels unfeasible. As a consequence, a higher-level representation is extracted from each imaging modality by using the networks trained in the unimodal approach as features extractors. Then, the representations are concatenated with the CL, before being considered as input for a ML model. In particular, the Principal Component Analysis (PCA) is used to implement the reduction of the resulting features vector, and the SVM classifier is used for the classification as described in the U approach.

The late fusion (LF) completely relies on the unimodal approach since it aggregates the prediction coming from the classifiers implemented for each data source. To this goal, the predictions coming from the three networks and the SVM are combined using the Weighted Majority Voting (WMV), in which a weight is assigned to each prediction according to the model output probability. However, in LF each classifier acts independently, not taking advantage of the complementary characteristics of the image modalities and clinical data that do not influence each other during the prediction.

In the IF approach, the influence of the TM is studied by considering solutions in which that module is excluded. More in detail, the use of the transfer module is denoted with IF-MT configuration. When the TM is included, it is inserted after each block containing the layers implementing the residual network in the case of ResNet. As a consequence, the number of TM is 4 and it does not depend on the specific architecture involved. On the other hand, in the case of BasicNet, a study is conducted to assess how the position of the transfer module influences performance. The location of the module is indicated in round brackets near the IF-MT configuration. In the IF technique, fine-tuning is exploited by initialising the convolutional cores with the weights determined in the U approach.

Moreover, the EF, LF and IF approaches are also evaluated excluding the CL-S1 set, considering only the features coming from MRI scans. When the tabular data is not exploited, the resulting configurations are shown with "IMG". Table 5.2 gives the details about the implemented experiments.

In the experiments involving DL approaches, during the training data augmentation is used by applying random rotations and flippings, while the dataset is balanced by replicating some randomly chosen volumes belonging to the minority class. Moreover, the greyscale intensity in each extracted volume is normalized in $[0; 1]$ to ensure that, in the classification step, the CNNs operate with volumes having the same scale across different patients. During the experiments, the maximum number of epochs is set to 500, the batch size is set to 32 for BasicNet and all the ResNet architectures. The learning rate for the cross-entropy loss was set to 10^{-6}.

Experiment	**Details**
U	Unimodal approach. In brackets the data modality is reported
EF	Early Fusion considering images and clinical data
EF-IMG	Early Fusion without clinical data
LF	Late Fusion considering images and clinical data
LF-IMG	Late Fusion without clinical data
IF	Intermediate fusion without TM
IF-TM	Intermediate fusion with TM
IF-IMG	Intermediate fusion without TM and clinical data
IF-IMG-TM	Intermediate fusion with TM but without clinical data

Table 5.2. Details about the implemented experiments in the Multimodal Approach

Adam optimizer is used with a weight decay set to 10^{-4}. To find the appropriate hyper-parameters, a grid search is implemented by varying the batch size in [8, 64], the learning rate in $[10^{-7}, 10^{-3}]$ and the weight decay in $[0, 10^{-4}]$.

Performance is evaluated in terms of Accuracy (ACC), Sensitivity (SENS), Specificity (SPE) and Area under ROC curve (AUC). All the experiments were run in a 10-folds Cross Validation (CV) setting, to better assess the generalization ability of each approach. In particular, a patient-based cross validation is performed, to reliably compare the performance of different models by avoiding the use of volumes or 3D slices from the same patient during the training and evaluation phase.

All the DL experiments were carried out using Pytorch (version 1.10), while the pre-processing step, solutions with SVM, and the different bounding box options were implemented in MATLAB 2020b.

A Linux workstation equipped with Intel(R) Core(TM) i7-10700KF CPU, 64 GB of DDR4 RAM and a Nvidia RTX 3090 GPU is used.

5.5 Results

Table 5.3 focuses on the preliminary work proposed in [126], showing how the different bounding box options impact the performance of axillary lymph node metastasis prediction on the DCE sequence and explaining the choice of the SIB procedure in the pre-processing. The networks proposed in [126], namely the SFB-NET, VB-NET, 2DS-NET, present the same structure of the BasicNet described in this thesis since they consist of a series of reduction layers in the convolutional cores, followed by two fully connected layers in the classification cores. The 2DS creates a set of images, one for each slice having lesions, as opposed to the SFB, SVB, and SIB bounding settings, which extract one volume for each subject entirely covering the tumour region. Moreover, SLVB and SLIB produce a set of volumes based on the patient's number of lesions. Each volume or image receives a prediction from the involved CNN. However, the objective is to assign each patient with a label that indicates the risk of axillary lymph node metastases. As a result, when the SLVB, SLIB, and 2DS bounding options are investigated, the estimated classes of all volumes that belong to the same patient must be pooled. Majority Voting (MV), and Weighted Majority Voting (WMV), which acts as MV but assigns a weight to each volume according to the network output probability, are two of the combining strategies (CS) that are explored.

The aim of the work proposed in this thesis is to propose a solution based on IF that includes the TM to implement a *cross-modality calibration* of the features map. In each table, the performance is reported in terms of ACC, SPE, SENS and AUC, specifying the fusion modality (Mod.) and the network (Net.). For each metric, the average rate computed adopting 10-fold CV and its standard deviation are shown. In each section, the best performance for each metric is reported in bold. As reported in Section 5.4, in the case of BasicNet, a study is conducted to assess how the position of the transfer module influences performance. Table 5.4 reports the performance obtained by varying the position of the transfer module in the BasicNet. In particular, the first row illustrated the results gained when TM is inserted after the second, the third, and the fourth level of the CNN, the second experiment involves the proposed model after the third, and the fourth layer while the last row reports the performance when TM is inserted only after the fourth layer. It is possible to note that the configuration IF - MT (3,4) achieves the highest value in each metric. In particular, it presents a value of ACC equal to 87.01±0.08% while a value of SENS equal to 81.48±0.17%.

Table 5.5 reports the performance of all the experiments involving the IF modality, namely IF, IF-TM, IF-IMG, IF-IMG-TM, for each network.

In particular, the first second shows the performance on the BasicNet considering in the IF-TM configuration the transfer module after the third and the fourth layer of the CNN. It is possible to note that the solution with IF-TM (3,4) outperforms the other in all metrics. The second part focuses on ResNet10 where the IF-TM approach has the best performance in terms of ACC, SPE, and SENS achieving values of 90.91±0.08%, 92.91±0.07% and 81.48±0.17% respectively. The results obtained with ResNet18 are reported in the third section, confirming the IF-TM configuration as the best one. Indeed, it reports 89.61±0.06% and 81.48±0.17% in terms of ACC and SENS respectively. The third part focuses on ResNet34, where the IF-TM approach achieves a value of ACC equal to 87.66±0.10% and a value of SENS equal to 77.78±0.16%. The last section reports the performance of the ResNet50. In this case, the IF-TM solution presents values of 84.42±0.05% and and 77.78±0.16% in terms of ACC and SENS respectively.

Table 5.6 reports the performance of all the experiments detailed in Table 5.2, selecting for the IF configuration the one that in Table 5.5 shows the highest performance in each network, that is the solution that includes the transfer module.Table 5.6 presents a section for each network, reporting the performance with the unimodal approach, early, late and intermediate fusion modalities. It is possible to note that the IF approach is always the preferred solution. Moreover, the LF approach outperforms the EF solution.

In the first two rows, the results of the solutions involving only the clinical feature are detailed. In particular, the configuration U(CL) refers to the SVM classifier, while the U(CL-S1) uses the MLP. The implemented methodology is also compared with the solution proposed in the literature by Nguyen et al.[100] that represents the first attempt to apply multimodal learning for ALN metastasis prediction. The authors propose a solution that exploits a CNN, DCE sequence of the primary tumor and four clinical features, namely age, ER, ki-67 and HER2, representing the set CL-4. In that work a 3D CNN is implemented to process DCE-MRI images using a subtractive approach that works with the third, the fourth and the fifth post-contrast volumes. A 3D cuboidal bounding box of size $50 \times 50 \times 50$, encompassing the tumor region, is used to crop each DCE-MRI data. The CL-4 set is inserted in the first fully connected layer of the classification core of the implemented network. It is possible to note that the implemented methodology outperforms the solution proposed by Nguyen et al.[100] in all metrics.

Finally, Table 5.7 reports a comparison between the CL-S1 set and features selected from the solution involving the SVM and the CL in the unimodal approach.

Model	**Option**	**CS**	**ACC**	**SPE**	**SENS**	**AUC**
SFB-NET	SFB	-	70.62±0.17	75.92±0.20	43.33±0.31	69.52±0.14
VB-NET	SVB	-	76.79±0.06	78.17±0.06	71.67±0.27	75.05±0.15
	SIB	-	78.06±0.11	78.91±0.12	**74.07±0.15**	**78.53±0.09**
	SLVB	MV	52.71±0.21	53.24±0.30	48.33±0.31	53.11±0.16
		WMV	53.33±0.22	54.79±0.31	45.00±0.34	57.75±0.20
	SLIB	MV	62.58±0.17	71.79±0.20	18.33±0.20	47.34±0.19
		WMV	63.21±0.16	72.56±0.19	18.33±0.20	49.49±0.24
2DS-NET	2DS	MV	**78.63±0.08**	**85.75±0.08**	46.67±0.31	77.86±0.14
		WMV	**78.63±0.08**	**86.52±0.08**	43.33±0.26	77.31±0.13

Table 5.3. Performance achieved in a 10-folds Cross Validation setting considering the solution proposed in [126], describing different bounding box options.

Mod.	**ACC**	**SPE**	**SENS**	**AUC**
IF - TM (2,3,4)	81.17±0.09	81.10±0.10	**81.48±0.17**	82.15±0.12
IF - TM (3,4)	**87.01±0.08**	**88.19±0.08**	**81.48±0.17**	**89.82±0.11**
IF - TM (4)	86.36±0.10	87.40±0.11	**81.48±0.17**	81.51±0.14

Table 5.4. Performance obtained in 10 fold-CV of the experiments conducted to assess how the position of the transfer module influences performance in the case of BasicNet.

5.6 Discussion

Table 5.3 highlights that the approach involving the SIB option can be considered the best one, reporting the highest value in terms of sensitivity and acceptable performance on the remaining metrics. The other experiments, characterized by accuracy or AUC higher than those of the selected solution, show a lower sensitivity, i.e., percentage of axillary lymph node metastasis correctly predicted. The obtained result is completely in accordance with the specific problem to solve. In the SFB option, the fixed-sized

Net.	Mod.	ACC	SPE	SENS	AUC
BasicNet	IF - MT (3,4)	**87.01±0.08**	**88.19±0.08**	**81.48±0.17**	**89.82±0.11**
	IF	84.42±0.11	85.04±0.12	**81.48±0.17**	81.28±0.15
	IF -IMG- TM (3,4)	82.47±0.12	82.68±0.14	**81.48±0.17**	86.56±0.15
	IF -IMG	81.82±0.08	82.68±0.12	77.78±0.22	80.40±0.10
Resnet10	IF - MT	**90.91±0.08**	**92.91±0.07**	**81.48±0.17**	87.17±0.14
	IF	90.26±0.08	92.13±0.08	**81.48±0.17**	**87.43±0.12**
	IF -IMG- TM	89.61±0.08	91.34±0.08	**81.48±0.17**	85.42±0.14
	IF -IMG	84.42±0.10	85.83±0.09	77.78±0.16	85.97±0.16
Resnet18	IF - MT	**89.61±0.06**	**91.34±0.06**	**81.48±0.17**	**85.04±0.14**
	IF	86.36±0.10	88.19±0.10	77.78±0.16	84.54±0.14
	IF -IMG- TM	85.71±0.10	87.40±0.10	77.78±0.16	80.78±0.15
	IF -IMG	83.12±0.08	84.25±0.08	77.78±0.16	81.69±0.14
Resnet34	IF - MT	**87.66±0.10**	**89.76±0.09**	**77.78±0.16**	83.79±0.15
	IF	87.01±0.07	88.98±0.05	**77.78±0.16**	**84.49±0.12**
	IF -IMG- TM	87.01±0.06	88.98±0.06	**77.78±0.16**	80.84±0.15
	IF -IMG	86.36±0.08	88.98±0.07	74.07±0.15	80.55±0.13
Resnet50	IF - MT	**84.42±0.11**	**85.83±0.11**	**77.78±0.16**	**83.96±0.16**
	IF	81.82±0.11	82.68±0.12	77.78±0.16	81.31±0.13
	IF -IMG- TM	**84.42±0.05**	**85.83±0.07**	**77.78±0.22**	80.43±0.17
	IF -IMG	79.87±0.11	81.10±0.12	74.07±0.15	79.24±0.11

Table 5.5. Performance obtained in 10 fold-CV of the experiments conducted considering the IF approaches.

bounding box may result in the inclusion of an excessive amount of healthy tissue with respect to the lesioned one, especially in patients with a small tumour region. On the other hand, the SVB and SIB options consider the smallest 3D cubical bounding box circumscribed to the patient's tumour region, limiting the amount of healthy tissue to include. However, the SVB extracts volumes whose voxels have different dimensions (in terms of mm) along the three spatial axes, resulting in the need of introducing the SIB procedure that considers volumes with isotropic pixels. The considerations made between SVB and SIB can be also used for SLVB and SLIB where the presence of multiple lesions is not exploited for the prediction of axillary lymph node metastasis.

Results show that for each network the solution involving the IF approach with the transfer module has the best performance. The analysis conducted in Table 5.6 shows that in most experiments, the LF performed better than the EF, thus supporting the preference of the former over the latter. Indeed, the LF approach exploits different models, each trained for

Model	Mod.	ACC	SPE	SENS	AUC
SVM	U(CL)	76.77±0.12	78.13±0.13	70.37±0.19	71.53±0.21
MPL	U(CL-S1)	75.97±0.11	78.74±0.10	62.96±0.22	75.61±0.17
BasicNet	U (DCE)	78.06±0.11	78.91±0.12	74.07±0.15	78.53±0.09
	U (T2)	74.84±0.11	78.13±0.11	59.26±0.28	62.36±0.21
	U (DWI)	79.87±0.13	85.04±0.12	55.56±0.25	66.00±0.23
	IF - TM (3,4)	**87.01±0.08**	**88.19±0.08**	**81.48±0.17**	**89.82±0.11**
	EF	62.34±0.15	62.99±0.14	59.26±0.22	61.13±0.19
	EF -IMG	55.19±0.14	49.61±0.15	81.48±0.23	65.54±0.18
	LF	84.52±0.11	86.72±0.11	74.07±0.14	78.62±0.15
	LF -IMG	84.52±0.11	86.72±0.10	74.07±0.15	75.23±0.17
ResNet10	U (DCE)	85.16±0.07	87.50±0.07	74.07±0.17	81.34±0.12
	U (T2)	85.16±0.08	89.84±0.09	62.96±0.08	76.71±0.14
	U (DWI)	83.77±0.08	88.98±0.09	59.26±0.11	71.60±0.15
	IF - TM	**90.91±0.08**	92.91±0.07	**81.48±0.17**	87.17±0.14
	EF	85.06±0.08	**98.43±0.08**	22.22±0.25	60.32±0.23
	EF -IMG	72.08±0.08	76.38±0.08	51.85±0.25	64.11±0.22
	LF	89.03±0.07	92.19±0.07	74.07±0.05	92.85±0.10
	LF -IMG	85.81±0.07	88.28±0.07	74.07±0.06	**92.97±0.10**
ResNet18	U (DCE)	84.52±0.08	86.72±0.08	74.07±0.15	80.93±0.10
	U (T2)	78.71±0.13	82.03±0.16	62.96±0.08	72.60±0.16
	U (DWI)	81.17±0.11	85.83±0.12	59.26±0.11	71.30±0.13
	IF - TM	**89.61±0.06**	**91.34±0.06**	**81.48±0.17**	85.04±0.14
	EF	66.88±0.10	67.72±0.12	62.96±0.14	65.34±0.19
	EF -IMG	77.27±0.11	88.19±0.11	25.93±0.15	57.06±0.20
	LF	83.23±0.08	85.94±0.08	70.37±0.10	86.98±0.10
	LF -IMG	83.87±0.08	87.50±0.08	66.67±0.10	**87.88±0.11**
ResNet34	U (DCE)	78.71±0.10	80.47±0.12	70.37±0.15	74.13±0.12
	U (T2)	76.77±0.10	80.47±0.13	59.26±0.19	67.85±0.16
	U (DWI)	74.03±0.08	77.95±0.10	55.56±0.20	64.36±0.13
	IF - TM	**87.66±0.10**	**89.76±0.09**	**77.78±0.16**	**83.79±0.15**
	EF	74.03±0.07	81.89±0.10	37.04±0.25	59.46±0.21
	EF -IMG	70.13±0.08	74.02±0.08	51.85±0.27	62.93±0.23
	LF	81.29±0.10	82.03±0.11	**77.78±0.14**	82.96±0.19
	LF -IMG	80.00±0.11	81.25±0.11	74.07±0.15	80.64±0.17
ResNet50	U (DCE)	78.06±0.08	79.69±0.08	70.37±0.20	82.26±0.11
	U (T2)	67.10±0.10	68.75±0.13	59.26±0.11	64.06±0.09
	U (DWI)	72.08±0.12	75.59±0.13	55.56±0.25	64.01±0.17
	IF - TM	**84.42±0.11**	**85.83±0.11**	**77.78±0.16**	83.96±0.16
	EF	67.53±0.15	68.50±0.15	62.96±0.21	65.73±0.17
	EF -IMG	39.61±0.15	31.50±0.15	**77.78±0.15**	54.64±0.21
	LF	80.00±0.11	81.25±0.11	74.07±0.11	**85.19±0.15**
	LF -IMG	78.06±0.10	80.47±0.10	66.67±0.12	82.90±0.11
Nguyen et al.[100]	DCE+CL-4	68.18±0.15	70.87±0.21	55.56±0.33	67.07±0.20

Table 5.6. Performance obtained in 10 fold-CV of the experiments conducted with unimodal approach and different multimodal fusion modalities.

Set of features	Details
CL-S1	age, familiarity, hormone therapy, menopausal status dimensions, ER, PgR, HER2, ki-67, and grading
SVM features	age, familiarity, hormone therapy, menopausal status dimensions, ER, PgR, HER2, ki-67, grading apparent diffusion coefficient (ADC), signal timing curve margins, tumor class, histotype, multifocality

Table 5.7. Comparison of the features selected by the domain expert (D.S), and those obtained with the backward features elimination and SVM.

a specific data modality. As a consequence, the four classifiers learn to extract features that reflect the distinctive characteristics of each modality, delaying the combination of the results in a post-processing step. On the other hand, in the EF solution, the networks trained in the U approach are used as features extractors, creating a high-dimensionality shared representation, that may make it difficult to capture the complementary characteristics of each data modality. In solutions involving the IF approach, the shared representation is created by concatenating features coming from the convolutional cores at an intermediate level, thus preserving the distinctiveness of the different image modalities, which is then exploited in the classification core. In particular, the presence of the transfer module makes the layers of the convolutional cores influence each other during the features extraction step, implementing the *cross modality calibration*. As a result, the IF approach is able to overcome the disadvantages identified in the EF and LF solutions, introducing an improvement in the performance obtained.

Table 5.4 shows that the best performance is obtained when the TM module is inserted after the third and the fourth layer of the BasicNet, that is after the mid-point of the CNN. This implies that the features extracted in the first layers strongly depend on the specific image modality. When the TM is inserted only in the fourth, layer the obtained performance is quite similar to the results of the approach obtained when that module is excluded (Tables 5.4 and 5.5).

Table 5.5 reports that the CL-S1 features affect the results since for each network the solutions exploiting only MRI scans present lower performance. Finally, in Table 5.6 it is possible to note that the features CL-S1 are selected by the backward features elimination implemented when the SVM model is used.

Chapter 6

Data from Multiple Imaging Exams: Dementia Status Assessment

Dementia disease consists in the loss of cognitive functioning (thinking, remembering, and reasoning), which interferes with daily life and activities of a person. It describes a syndrome of generalized mental deterioration that also causes personality change and makes the subject vulnerable in many ways (i.e., injuries from falls and accidents) [76]. Alzheimer's disease (AD) is the most common cause of dementia, affecting millions of elderly people around the world [69]. The progression of the disease spans in different stages, from very mild cognitive impairment, where subjects experience memory loss or cognitive difficulties, to mild and severe stages, where damage occurs in the areas of the brain that control languages and conscious thought [69].

AD is a neurodegenerative disorder, and early detection is a key element to improve the quality of life of affected patients and their families [4]. The Clinical Dementia Rating (CDR) [52] is used in clinical trials to classify the severity of AD. It is derived from an interview with the patient and rates impairment in each of the six cognitive categories (memory, orientation, judgment, community affairs, hobbies, and personal care). The CDR value consists of a five-point scale in which 0 means a cognitive normal condition, 0.5 indicates a questionable impairment, whilst the values 1, 2, and 3 correspond to a mild, moderate, and severe impairment respectively. In clinical trials, the Magnetic Resonance Imaging (MRI) is the standard diagnostic tool [129, 151] due to the fact that the acquired images have a strong relationship with the topology of the brain showing the alteration of the morphology. In particular, the T1-weighted (T1-w) MRI provides information about the brain structure, making it possible to evaluate its volumetric characteristics and atrophy. Positron Emission Tomography (PET) is another imaging technique that consists of the injection of a tracer capable of revealing the metabolic functions of the tissue under analysis. AD is characterized by the presence of β-amyloid ($A\beta$) plaques that may develop many years before the onset of dementia [122]. As a consequence, amyloid imaging is widely used in the early diagnosis of AD. The Pittsburgh Compound B (C-PiB) [68] is the first PET tracer specific for $A\beta$ plaques and it is used to identify plaques in brain tissue.

There is the need of combining information from heterogeneous and complementary sources, such as MRI and PET, to evaluate the structural and metabolic characteristics of the brain. This makes the Multimodal Deep Learning (MDL) approaches well suited in the case of dementia assessment. Although the majority of papers exploit a single modality, that is the MRI, few work have been proposed to provide a model able to process multiple data sources considering the joint or intermediate learning.

However, when working with a multimodal dataset in the medical field, it is not easy to have images of all the involved modalities, belonging to the same patient. For each subject, *paired acquisition* consists of images coming from all the different sources and collected at the same time or in a specific range. In a real scenario, patients may have *incomplete acquisitions*, in which some modalities are missed. In the literature, some implemented methodologies discard the incomplete instances, considering only the paired acquisitions and limiting the amount of data to be considered [91, 51, 131]. On the other hand, few work propose to fill the missing modalities with black images [91] or interpolation operations [1].

In this thesis, a systematic analysis of early, late, and joint approaches in fusion for dementia severity assessment is conducted on the publicly available OASIS-3 dataset [74], which is the latest release in the Open Access Series of Imaging Studies (OASIS) and it includes two different image modalities, the T1-w MRI and the C-PiB PET. 3D CNNs are used to exploit the volumetric features of the involved images, including in the training step strategies to handle a high imbalance and incomplete dataset. In particular, this work provides an analysis of the effects of the incomplete dataset in each multimodal fusion technique, and in the case of intermediate fusion, a Multiple Input - Multi Output 3D CNN is proposed whose training iteration changes according to the characteristics of the input volumes.

6.1 Dementia evaluation in MRI and PET acquisitions

Most of the work in the literature propose solutions that take advantage of DL approaches and MRI for the assessment of dementia. The methodologies implemented differ for the classification task, for the dataset used and for the characteristics of the neural networks (3D CNN or 2D CNN) [63, 144]. With reference to the classification task, some work address task distinguishing a normal brain condition from a damaged one [113, 3, 148, 61, 114], whilst others deal with a solution for a multiclass classification, also introducing the condition of mild cognitive impairment [55, 127, 102, 10, 47, 98, 157]. With reference to the dataset, most of the papers use the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [56], whereas few work use the OASIS-3 collection [74] to exploit the fine-tuning of state-of-the-art 2D-CNNs (AlexNet [71], ResNet [46], Inception [141], VGG [135]) [113, 3, 55, 61, 102, 98]. With reference to the neural network characteristics, since MRI is a 3D acquisition, the involvement of state-of-the-art networks introduces the problem of selecting the right number and portion of slices to use. Some articles use the median slice or select some central slices [55, 61]. However, each slice is classified independently from the others, and therefore, volumetric information is not exploited.

To overcome this limitation, the 3D CNNs are then introduced permitting to extract the features from the whole 3D brain volume [114, 10, 47]. However, they require more training time since they are characterized by a huge number of parameters and it is not easy to find pre-trained 3D networks, especially considering biomedical images. The implemented 3D networks in the literature consist of a sequence of 3D convolutional layers, separated by pooling operation, batch normalization, and activation functions.

In the literature, few solutions considering PET acquisitions have been proposed, presenting the same heterogeneity as seen in work focusing on MRI. Some papers focus on the ADNI dataset [56], exploiting fluorodeoxyglucose (FDG) PET. In particular, the solutions presented in [32, 161] use state-of-the-art CNNs (AlexNet [71] and InceptionV3 [141]), while the methodology proposed in [20] relies on a 3D CNN. Only one paper [15] uses the OASIS-3 dataset [74], proposing a 3D CNN trained from scratch, focussing on amyloid imaging.

In the recent years, implemented solutions exploited joint learning for dementia assessment, without considering the late and early fusion. The authors of [91] proposed a methodology that combines T1-w MRI, Diffusion Tensor Imaging (DTI-MRI) and 2D ResNet18 [46], representing the first article that focuses on a multimodal approach with the OASIS-3 dataset [74].

The majority of papers that use both MRI and PET images focus on ADNI dataset [56]. In [1] authors used a 2D CNN for the classification of features extracted from MRI, PET and genetic data, while in [51] authors proposed a multi-input 3D CNN considering as input the hippocampal area selected from MRI and PET. In [152] sparse autoencoder and 3D CNN were introduced for classification, while in [131] stacked Deep Polynomial Networks (DPNs) were used to extract features from the two image modalities. The solutions proposed in [1, 51, 152, 131] report more than 90% of accuracy in separating a normal brain condition from a damaged one.

When working with a multimodal environment, it is not easy to have images of all the involved modalities, that, in turn, will result in the need to manage *incomplete acquisitions*. Authors in [91, 51, 131] used only a paired dataset, reducing the number of images available. In [1] the authors used linear interpolation to fill in the missing modalities. However, this process makes the different acquisitions dependent on each other, especially in the case of longitudinal studies. Authors in [145] handled the incomplete dataset proposing a fully connected multi-input network, consisting of three fully connected layers, with three outputs (first modality only, second modality only, paired datasets). The features are extracted from MRI and PET before being fed into the network that is updated in different parts according to the available input data, obtaining an accuracy

equal to 65%.
As a further remark, it is important to note that the OASIS-3 [74] and ADNI [56] datasets consist of longitudinal studies, so that they include images of the same patient collected at different time. This implies that a time-dependent relationship between images belonging to the same subject exists and has to be taken into account, especially when designing the validation step of the implemented solutions. However, although some work in the literature got very high precision (more than 90%), most of them conducted an image level rather than a patient-level split of the dataset [113, 91, 3, 55, 61, 10, 47, 152, 98, 157, 20, 1, 131], not guaranteeing complete independence between the train, validation, and test sets, as images belonging to the same patient may not be included in the same set. Moreover, some work applied data augmentation techniques before splitting the dataset [3, 10, 47, 98]. These characteristics may introduce bias in the performance evaluation step.
This work focuses only on the OASIS-3 [74] dataset, using the CDR value to provide an assessment of the dementia status rather than a specific diagnosis. Indeed, as reported in [74], the diagnostic information is separated from the clinical assessment that includes the CDR value. It is worth noting that in the ADNI [56] dataset the information about the CDR value is not available for all the subjects, since the study focuses on the progression of mild cognitive impairment and early AD, presenting a specific diagnosis for each image. This makes the dataset suited for tasks that aim to determine a diagnosis or a specific Alzheimer's stage rather than an assessment of dementia status.
In comparison with the existing literature, the contribution of this thesis can be summarised as follows:

- To perform a systematic analysis of MDL approaches for dementia status assessment considering both MRI and PET images. To this end, different fusion techniques such as early, late and intermediate fusion are exploited.

- To implement a solution completely based on 3D CNN, extracting features from the whole 3D brain volume.

- To propose a training strategy able to handle a high imbalance and incomplete dataset. In particular, in the case of the intermediate approach, 3D CNN architecture is implemented whose training iteration changes according to the characteristics of its input.

6.2 Population

This work considers the OASIS-3 dataset [74], consisting in a compilation of MRI, PET imaging, and clinical data for 1098 patients. The study collects 605 cognitively normal patients and 493 individuals at various stages of cognitive decline. Each MRI, PET, and clinical data acquisition session report information about when they are acquired expressed as the number of days since the subject's entry date into the study. The status of dementia was assessed using the Clinical Dementia Rating Scale (CDR). For each assessment, information on the CDR value is available, together with a diagnosis of dementia (i.e., 'cognitively normal', 'AD dementia', 'vascular dementia'). This work focuses only on the most common cause of dementia, Alzheimer's disease, selecting among subjects with CDR> 0, individuals with AD diagnosis, reducing the number of patients to 1025. In this thesis, T1-w MRI and C-PiB PET acquisitions are exploited, since they are the most used diagnostic tools for dementia diagnosis and assessment [129, 151, 68].

The CDR value is used to distinguish between three different classes, as proposed in [113, 127, 55, 13, 31]: the CDR value set at 0 ($CDR = 0$) represents a normal cognitive function, determining the cognitive normal class (C); the CDR value set at 0.5 ($CDR = 0.5$) indicates very mild dementia, determining the mild class (M); the CDR value greater than 0.5 ($CDR > 0.5$), represents mild / severe dementia, suggesting the severe class (S). Hence, the task to solve consists in the classification of each MRI, PET, and MRI-PET pair in three different classes (C, M, and S), which is addressed considering two different approaches:

- **Three-class classification (Task A)**: it directly classifies samples into one of the three different classes (C, M, S)
- **Hierarchical classification (Task B)**: it consists of two steps:
  - *Cognitive Normal Classification (Task B1)*: which distinguishes between healthy (C) and diseased (NC) instances.

- *Disease Classification (Task B2)*: which discriminates between M and S that is applied if the instance is classified as NC.

Scan sessions and clinical data acquisition do not always occur at the same assessment. As a consequence, there is the need to find a criteria to match up MRI, PET images, and clinical data, with the aim of attributing the CDR value to each MRI and PET acquisition. A common criterion is to consider all clinical information within a year before and after the date of the MRI (or PET) session valid. In particular, this thesis considers the clinical data closest to each MRI (or PET) session to be a match if the difference in days is less than one year. The MRI and PET scans are matched to clinical assessments separately, and the result consists of three different datasets, namely MRI, PET and PAIRED datasets. The MRI and PET datasets include all the MRI and PET volumes, respectively, with the related CDR value, while the PAIRED dataset contains only MRI-PET pairs, represented by the MRI and PET volumes associated with the same clinical information date. Straightforwardly, both the MRI and PET datasets include the acquisitions of the PAIRED dataset. It is worth noting that the set of data involved in this paper represents an incomplete dataset in which, for each assessment, all image modalities are not always available.

During the data analysis step, 46 patients were deleted because the preprocessing phase of the raw acquisitions failed, as reported by the authors in [74]. Moreover, Figure 6.1 shows the case where MRI or PET images are not available for each clinical assessment. In particular, the patient OAS31045 should move from a normal condition to a mild/severe dementia according to the information coming from clinical data. However, only volumes representing the normal condition are provided. This characteristic implies that the available volumes should be considered when studying the distribution of patients in the different classes. This study discards all the patients that according to the clinical information change their condition from CDR $= 0$ (C class) to CDR > 0 (M and S classes), but only images for C are available. As a consequence 100 subjects are deleted. Then, with the aim of maintaining the same age distribution in the three classes, patients younger than 50 years are deleted.

This study considers a total of 628 patients, including 1263 MRI and 711 PET acquisitions, representing the MRI and PET datasets respectively, while the number of MRI-PET pairs is 628. Table 6.1 shows the progression of the patients in the three classes. The small number of patients in class S does not allow a further division of the CDR value. Table 6.2 reports the information on the generated datasets considering the number of volumes for each class. It is worth noting that the MRI and PET datasets include the volume in the PAIRED one.

	End with			
Start with	**C**	**M**	**S**	**TOTAL**
C	340	67	3	410
M		145	10	155
S			63	63
TOTAL	340	212	76	628

Table 6.1. The progression of patients in the three classes, cognitive normal (C), very mild dementia (M), mild/severe dementia (S)

Dataset	**Classes**			**TOTAL**
	C	**M**	**S**	
MRI	926	258	79	1263
PET	627	75	16	711
Paired	556	58	14	628

Table 6.2. Information about generated datasets in terms of number of volumes for each class, cognitive normal (C), very mild dementia (M), mild/severe dementia (S)

6.3 Methodology

This thesis aims to develop a decision system supporting dementia severity assessment, and exploiting information coming from two different image modalities, namely the T1-w MRI and the C-PiB PET. To this end, different MDL approaches belonging to all the three paradigms (early, late and joint) are investigated. As described in section 6.2, the distinction in three classes is addressed considering two different approaches, namely task A and B. Hierarchical classification is introduced to explicitly exploit the

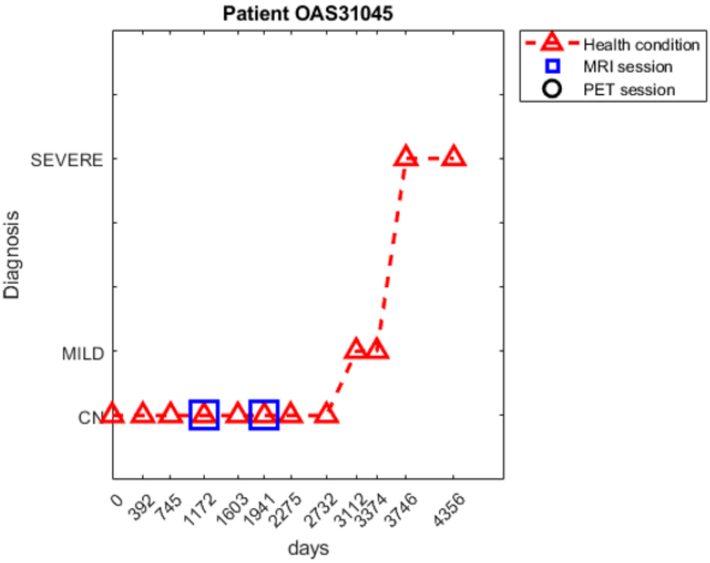


Figure 6.1. Disease progression for patient OAS31045

relationship between classes related to different CDR levels (C is $CDR = 0$, M is $CDR = 0.5$ and S is $CDR > 0.5$). Moreover, this work proposes to analyse the differences with the three-class classification approach, where the three classes are considered as independent, delegating the model the task of learning their relationships.
The proposed methodology consists of two steps: *Pre-processing*, used to prepare MRI and PET, and the *Dementia Severity Assessment*, which introduces the implemented solutions considering multimodal approaches.

6.3.1 Pre-Processing

T1w-MRI volumes are processed according to the procedure reported in [74], including motion correction, removal of non-brain tissue (skull-stripping) and intensity normalization. The result is a set of volumes of size $256 \times 256 \times 256$ with isotropic dimension.
A PET scan consists in the acquisition of 3D volumes after the injection of the tracer, resulting in 4D data. In OASIS dataset subjects underwent a 60 minute dynamic PET scan in 3D mode (24 x 5 sec frames; 9 x 20 sec

frames; 10 x 1 min frames; 9 x 5 min frames) generating a maximum of 52 3D volumes acquired over time. Since the analysis in interested in C-PiB retention, only volumes acquired after tracer absorption are considered. Indeed, the acquisitions obtained 40 minutes after tracer injection [94] are averaged to create a *static* PET scan, which is then rigidly registered to the MRI volume acquired during the dynamic PET scan session using mutual information as a similarity metric [92, 115]. The result is a PET volume of size 256 × 256 × 256 with isotropic dimension. To reduce the amount of non-brain tissue to include, the smallest 3D cubical box including the patient's brain is extracted from each MRI and PET volume. The cubical box is patient-dependent and its dimension strongly depends on the y axis, as shown in Figure 6.2. The resulting volumes are then normalised in [0,1] to ensure that, in the next stage, the involved CNNs operate on images of the same scale across different acquisitions.

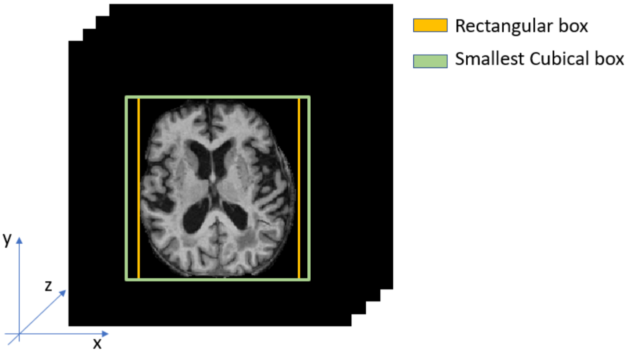


Figure 6.2. Extraction of brain region for each patient: the smallest 3D cubical box (in green) is computed from the rectangular box (in orange) surrounding the brain.

6.3.2 Dementia Severity Assessment

In clinical trials, T1-w MRI and PET are the standard diagnostic tools used to assess the severity of dementia. They provide different information since the former focuses on the volumetric and structural characteristics of the brain, while the latter reveals its metabolic functions. MDL allows the fusion of complementary information coming from heterogeneous sources

with the aim of providing a richer data representation than the unimodal approach.
Among all deep neural networks, CNNs are widely used in biomedical image processing [86, 140] with surprising results. A typical CNN consists of stacked *relatively complex layers* [11], with each of them having a convolutional stage, a non-linearity function (i.e ReLU), and a pooling operation. The set of *complex layers* constitutes the Convolutional Core (Conv-C), responsible for the features extraction step, while the classification is performed by a Classification Core (CC), typically including fully connected layers.
CNNs are exploited to process MRI and PET scans for the dementia severity assessment. Each acquisition represents a 3D volume, with three spatial dimensions (x, y, z) resulting in the need of using networks with 3D convolutional layers. Information coming from complementary data sources is exploited considering MDL techniques.

In the early fusion (EF) approach, described in Figure 6.3, each pair of MRI-PET is organized in a single structure before being considered as an input for a CNN. To this end, each MRI-PET pair is converted in a 2-channel volume by concatenating the different paired acquisitions. Straightforwardly, the first layer of the 3D-CNN is changed, by modifying the number of the input channels. In EF approach, two image modalities simultaneously contribute to prediction, using different information at the same time. Furthermore, the network is trained using the PAIRED dataset, which has a reduced amount of available data.

The late fusion (LF) completely relies on the unimodal approach since it aggregates the prediction coming from the classifiers implemented for the MRI and the PET separately, as shown in Figure 6.4. To this goal, the predictions coming from the two networks are combined using the Weighted Majority Voting (WMV), in which a weight is assigned to each prediction according to the network output probability. However, in LF each CNN acts independently, not taking advantage of the complementary characteristics of the two image modalities that do not influence each other during the prediction.

In the case of intermediate or joint fusion (IF), a Multi Input-Multi Output 3D-CNN architecture is proposed, that simultaneously processes MRI and PET scans. Furthermore, a training strategy able to handle the case where one of the inputs of the two modalities is missing is introduced. This permits the training dataset not to be reduced, thus exploiting all the available volumes. Figure 6.5 shows the architecture of the proposed network, which consists of two inputs, one for the MRI and one for the PET volumes, respectively; then there are two convolutional cores, namely the MRI Conv-C and the PET Conv-C, which extract abstract representations that are the inputs of two classification cores, namely the MRI-CC and the PET-CC. Next, the joint fusion concatenates the outputs of the two convolutional cores, which fed a PAIRED classification core, named as PAIRED-CC (Figure 6.5). The TM introduced in Section 4.1 is used to implement the *cross-modality calibration* between the layers of the two Conv-C, resulting in the architecture reported in Figure 6.6. It is possible to note that the resulting CNN differs from the one presented in Figure 4.4 as the incomplete dataset requires the use of three classification cores. Therefore, the result is a 3D-CNN with three outputs, with only one of them being considered according to the nature of the input: the PAIRED-CC is considered only for *paired acquisitions*, while the MRI-CC and the PET-CC act in the case of incomplete inputs. In particular, it is worth noting that it is possible to distinguish three different sub-networks, namely the MRI-NET, the PET-NET, and the PAIRED-NET that, during the training, are separately updated as follows:

1. When both modalities are available, the volumes of each MRI-PET pair fed MRI Conv-C and PET Conv-C, respectively, whose outputs are concatenated and considered as input for the PAIRED-CC for the classification. In particular, the transfer modules introduced between the layers of the MRI Conv-C and PET Conv-C contribute to the prediction. The described chain represents the PAIRED-NET, whose classification loss is used to update the two convolutional core, including the transfer modules, and the PAIRED-CC element (Figure 6.7).

2. When the PET data is missing, the MRI volumes fed into the MRI-NET, consisting of MRI Conv-C and MRI-CC. The output of the MRI-CC is used to compute the classification loss, updating the network weights (Figure 6.8). In this case, the transfer modules do not contribute to the prediction.

3. Similarly when the MRI data is missing, the PET volumes fed PET-NET consisting of the PET Conv-C and PET-CC chain. The output of the PET-CC is used to compute the classification loss, updating the network weights (Figure 6.9). In this case, the transfer modules do not contribute to the prediction.

It is worth noting that the joint fusion approach described so far overcomes the aforementioned disadvantages of both the early and the late fusions. Indeed, on the one side, it allows to train the network exploiting all the data available in the dataset. On the other side, it is also beneficial in the test step, as it is possible to classify instances where one modality is missing. Furthermore, if the PAIRED-CC is removed, it is possible to consider the MRI-NET and the PET-NET as independent networks, obtaining the unimodal approach (U). In this respect in the following, the networks trained with the MRI and PET volumes respectively are denoted as U MRI-NET and U PET-NET.

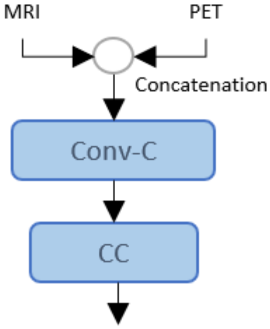


Figure 6.3. Architecture of the 3D CNN proposed for the EF

CNN architecture

In this thesis, two different deep models are implemented for the convolutional part (Conv-C) and for the classification part (CC) of each architecture, regardless of whether it receives MRI or PET images as input, and regardless of the type of fusion adopted, i.e., early, late and joint. Such two networks are named in the following as BasicNet and ResNet, introduced

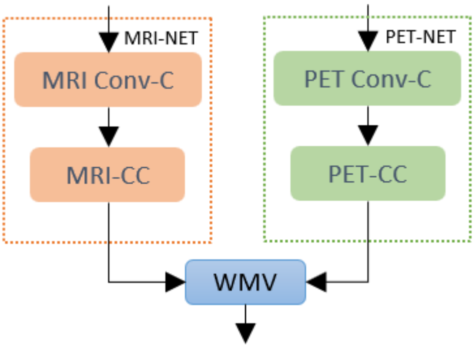


Figure 6.4. Architecture of the 3D CNN proposed for the LF

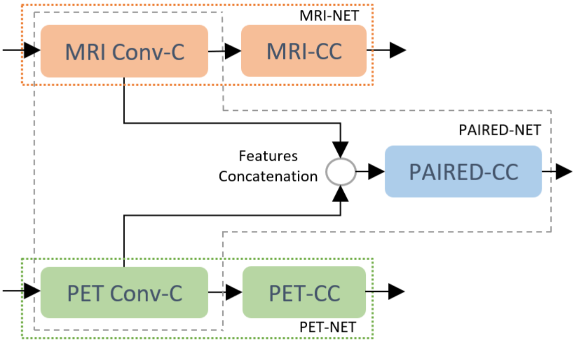


Figure 6.5. Architecture of the 3D CNN proposed for the IF: it is possible to distinguish three different sub-networks, the MRI-NET, the PET-NET, and the PAIRED-NET.

in Section 4.4.

BasicNet consists of six *reduction layers* followed by a global average pooling and a fully connected layer. A reduction layer is a block with a 3D-convolutional operation consisting of $3 \times 3 \times 3$ kernel, stride, and padding set to 2 and 1 respectively, followed by batch normalization and ReLU function. Each convolution reduces the input feature map and doubles the number of channels, excluding the first convolutional layer with 16 output channels. The stride set to 2 makes pooling operations (max or average pooling) not necessary [137]. The chain of the six reduction layers constitutes the Conv-C of the network, which is responsible for the

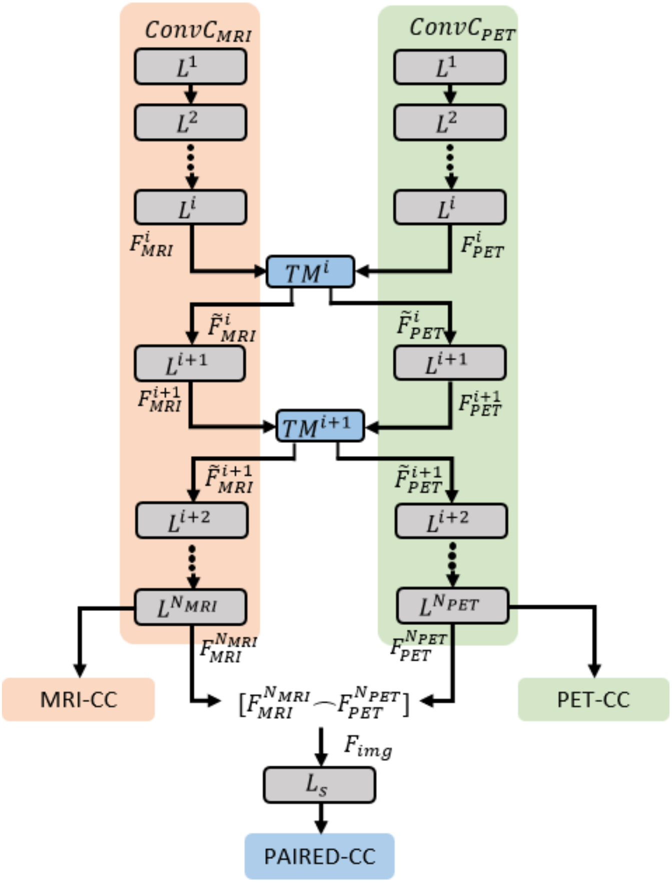


Figure 6.6. Architecture of the 3D CNN proposed for the IF with Transfer Module (TM)

features extraction step, while the global average pooling and the fully connected layer represent the CC as shown in Figure 6.10. In the implemented methodology BasicNet is trained from scratch.
Inspired by the state-of-art network proposed in [46], ResNet is a 3D CNN, adapted to handle 3D volume data. Furthermore, the solution implemented in [18] offers a set of 3D ResNet architectures pre-trained with medical images for segmentation tasks. The architecture consists of a backbone, that is an encoder representing the ResNet basic structure, re-

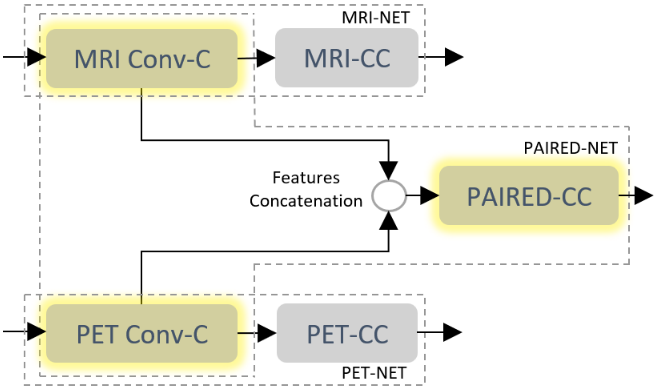


Figure 6.7. In the first iteration, the PAIRED-NET is considered updating the parts highlighted in yellow.

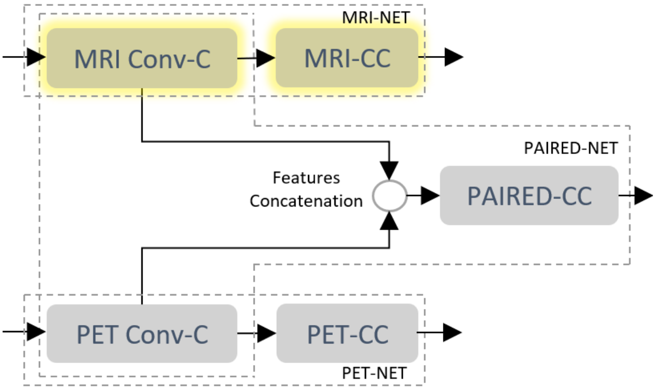


Figure 6.8. In the second iteration, the MRI-NET (highlighted in yellow) is updated since the input is a set of MRI volumes

sponsible for features extraction, and a set of decoders for the generation of the segmentation masks. Hence, the ResNet is adapted for the classification task as described in Section 4.4. The result is an architecture similar to the one proposed in [46], with some modifications: 3D convolutional layers are used to process MRI and PET acquisitions, and the input is a 1-channel 3D image. Note that on this one side, transfer learning is used, considering the pre-trained backbone as a starting point and implementing

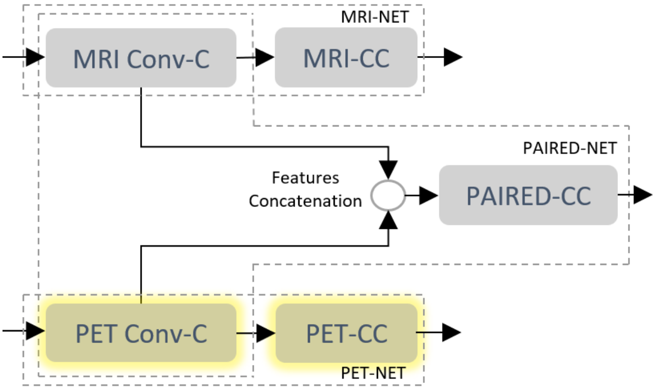


Figure 6.9. In the third iteration, the PET-NET is considered updating the parts highlighted in yellow.

the fine-tuning for adapting the ResNet for the specific task to solve. On the other hand, the classification core is trained from scratch.
All the techniques described in Section 6.3.2 are tested with the two 3D-CNNs. In particular, the U and LF approaches use the architectures as presented above, while in the EF the first convolutional layer is modified to handle a 2-channel volume. The BasicNet and the ResNet are also exploited in the IF implementing the MRI-NET (MRI Conv-C and MRI-CC) and the PET-NET (PET Conv-C and PET-CC). The PAIRED-NET is constituted by the MRI Conv-C, the PET Conv-C, and the PAIRED-CC, consisting of a global average pooling and two fully connected layers separated by the ReLU function. When the TM is used, it is inserted after each block containing the layers implementing the residual network in the case of ResNet. On the other hand, inspired by the analysis conducted in Section 5.5, it is inserted in the last three layers in the case of BasicNet.

In the medical field, it is very difficult to have large and balanced datasets as the number of patients usually involved in the study is small and the number of acquisitions belonging to healthy subjects is greater than the number of unhealthy ones. In this work, to improve the network's robustness the variability in the set of data used for training is introduced by applying data augmentation techniques which include translation, rotation, and scaling [18]. Straightforwardly, the implemented techniques

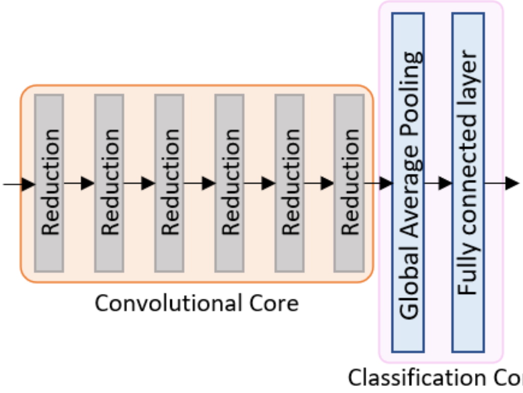


Figure 6.10. BasicNet architecture: the network consists of six *reduction layers*, representing the Convolutional Core (Conv-C), followed by a Global Average Pooling and a fully connected layer, constituting the Classification Core (CC)

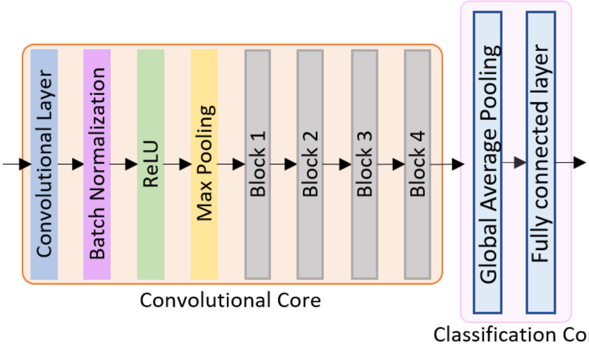


Figure 6.11. ResNet architecture: the Convolutional Core (Conv-C) consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four layers containing the blocks implementing the residual network, while the Global Average Pooling and the fully connected layer represent the Classification Core (CC)

avoids completely overturning the position of the brain areas when setting up the augmentation operations. Since the extracted bounding box is strongly influenced by the y axis, volumes are translated within $[-10, 10]$ pixels in x and z dimensions, and within $[-5, 5]$ in the remaining one. The

rotation angle is selected within $[-10, 10]$ for the y axis and within $[-5, 5]$ for the remaining ones, to reproduce natural head positions. The scaling factor, applied in each dimension, is chosen within $[0.9, 1.1]$, to simulate different brain sizes by introducing moderate modifications.
During the training, a strategy to handle data imbalance is also implemented, ensuring the creation of balanced batches in the various iterations. Denoting with b the batch size, and with N_c the number of classes, each batch contains b/N_c instances of each class. As a consequence, in each iteration, the network receives a balanced batch. However, since the number of batches to create in an epoch strongly depends on the size of the majority class, different batches may share samples belonging to the minority classes, resulting in an approach similar to that achieved by the simple replication of instances.

6.4 Experimental Set-up

The 3D CNN presented for IF consists of three different sub-networks, the MRI-NET, the PET-NET and the PAIRED-NET, as described in Section 6.3.2 and Figure 6.5. As a consequence, it is possible to evaluate the performance of the implemented network on the three datasets, considering the respective classification cores. The presence of the TM in the CNN implemented for the IF approach is denoted with IF-TM.
When the input consists in a MRI-PET pair, the three sub-networks can also be used in parallel, resulting in three outputs. To provide a unique prediction, a WMV combining strategy is used. The described configuration is referred with IF-WMV or IF-TM-WMV. Furthermore, the IF approach can be evaluated on a dataset obtained by merging the MRI, PET, and PAIRED datasets (COMPLETE dataset). Indeed, if the input is an MRI (or PET) volume, the output of the MRI-NET (or PET-NET) is considered, while if an MRI-PET pair is considered, the results of the three sub-nets are aggregated using WMV. As a consequence, a performance evaluation can be obtained considering all available volumes.

The EF and LF approaches can be used only with the PAIRED dataset, since the first requires an input of two channels, while the latter combines the results of two networks. However, it is possible to evaluate the performance by considering the COMPLETE dataset exploiting the network trained in the U approach for incomplete inputs. More in detail, in the LF the output combination is performed only for the MRI-PET pairs, while in the EF the outputs coming from the three CNNs are aggregated using WMV in the case of paired input.

This thesis provides a complete set of experiments since it exploits the MDL fusion methods in each task, namely the three classes (Task A) and hierarchical (Task B) classification, also detailing the results on the Tasks B1 (C/NC classification) and B2 (M/S classification). Moreover, the EF, LF and IF techniques are tested with the two CNNs, BasicNet and ResNet.

To assess the effectiveness of the proposed methodology, The result of the IF, EF and LF are compared with the U approach, considering also a proposal from the literature focusing on the OASIS-3 dataset [74].
The U approach is used as a baseline, evaluating its performance on the MRI and PET datasets. In particular, the BasicNet and the ResNet are trained considering the MRI and PET volumes separately, obtaining the U MRI-NET and the U PET-NET.
The solution described in [91] is selected as it represents the first article that focuses on MDL considering the OASIS-3 dataset [74]. Since the authors use T1-w and DTI MRI, the proposed methodology is implemented according to the information provided in [91], exploiting T1-w MRI and C-PiB PET for the classification of the three classes (Task A). Indeed, Massalimova et al.[91] explored the fine-tuning of the state-of-the-art RestNet18 [46] in the U approach, proposing in the IF a multi-input 2D CNN that leverages the network trained in the solution with a single modality. The input consists of a 3-channel image obtained concatenating the median slices of the sagittal, axial, and coronal views from each 3D scan. In the IF approach, the authors replace the missing modalities with black images. The methodology is compared with the solution proposed in [91] implementing both the U and IF approaches.

Table 6.3 gives details in terms of datasets and tasks for each of the considered approaches.

The experiments are organized in four different groups, as summarised as follows:

- Experiment 1 (EXP.1): that focuses on Task B1 (C/NC classification), evaluating the effectiveness of the MDL and U approaches on the four datasets (MRI, PET, PAIRED, COMPLETE).

- Experiment 2 (EXP.2): that considers Task B2 (M/S classification), exploring both the MDL and U solutions

- Experiment 3 (EXP.3): that focuses on Task A (C/M/S classification), evaluating the effectiveness of the MDL and U approaches on the four datasets (MRI, PET, PAIRED, COMPLETE).

- Experiment 4 (EXP.4): similarly, it focuses on Task B (hierarchical classification)

As described in Section 6.3.1, the smallest 3D cubical box including the patient's brain is extracted from each MRI and PET volume. Since the resulting volume in patient-dependent, a resize stage is used before feeding data into the involved 3D CNNs. More in detail, the BasicNet and ResNet receive as input volumes of size 128 × 128 × 128. In the IF technique, fine-tuning is exploited by initialising the MRI-NET and the PET-NET with the weights determined in the U approach, while the PAIRED-CC is trained from scratch. In experiments involving ResNet, the ResNet34 architecture is used. The involved networks are trained by minimizing the cross entropy loss. The maximum number of epochs is 200, the batch size is 32 for BasicNet and 16 for ResNet. The learning rate for the cross-entropy loss was set to 10^{-5} and 10^{-6} for the experiments involving BasicNet and ResNet, respectively. Adam optimizer is used with a weight decay set to 10^{-4}.

Performance is evaluated in terms of Accuracy (ACC), Precision, and Recall for each class, and Area under the ROC Curve (AUC). For the three-class classification, the average AUC weighted by each class support, that is the number of true instances for each label, is reported. All experiments were run in a 5-fold cross-validation (CV) setting, to better assess the generalization ability of each approach. More in detail, patient-based cross-validation is performed, avoiding the use of volumes from the same patient during the training and evaluation phase.

All the experiments were carried out using Pytorch (version 1.10), while the pre-processing step was implemented in MATLAB 2020b. A Linux workstation equipped with two NVIDIA RTX 3090 GPUs, AMD Ryzen Threadripper 3960X 24-Core Processor, 64 GB of DDR4 RAM is used.

Approach	Dataset	Task
EF (BasicNet, ResNet)	PAIRED, COMPLETE	B1, B2, A
LF (BasicNet, ResNet)	PAIRED, COMPLETE	B1, B2, A
IF (BasicNet, ResNet)	PAIRED, COMPLETE, MRI, PET	B1, B2, A
IF-MT (BasicNet, ResNet)	PAIRED, COMPLETE, MRI, PET	B1, B2, A
U (BasicNet, ResNet)	MRI, PET	B1, B2, A
ResNet18[91]	PAIRED, COMPLETE, MRI, PET	A

Table 6.3. Details in terms of datasets and tasks for each of the considered approaches

6.5 Results

The results are organized for each class of experiments (EXP.1, EXP2, EXP3, EXP4) that are evaluated on each dataset (MRI, PET, PAIRED and COMPLETE) considering both networks to provide a performance comparison among different approaches.

Tables 6.4, 6.5, 6.6, 6.7 report the performance of EXP.1, EXP.2, EXP.3, EXP.4 respectively, specifying the dataset (Data), the network (Net), and the fusion modality (Mod.). Moreover, the BasicNet and ResNet architectures are denoted as Ba and Re. Each table consists of four main sections. The first one shows the results of the U, IF and IF-MT approaches on the MRI dataset, while the second section focuses on the PET volumes. In the IF and IF-MT cases, only the outputs of the MRI-CC and the PET-CC separately are considered, while in the U approach the performance of the U MRI-NET and U PET-NET, trained considering only the MRI and PET dataset respectively is evaluated. The third part considers the PAIRED dataset, evaluating all the MDL fusion techniques (EF, LF and IF). In particular, the section includes also the IF-WMV as mentioned in Section 6.4 and the performance of the U solution computed only on the MRI and PET volumes of the PAIRED dataset, resulting in the U-MRI and U-PET configurations. Finally, the last section shows the performance of the EF, LF, and IF modalities on the COMPLETE dataset. In the EF and LF cases, the U MRI-NET and the U PET-NET trained in the U approach are exploited, while in IF the three outputs of the implemented CNN are considered. Moreover, in the case of Task A in Table 6.6, the results of the solution proposed in [91] is included in each section.

For each metric, the average rate computed adopting 5-fold CV and its standard deviation are reported, omitting 0. after the ± symbol and showing only 2 significant digits. In each section, the best performance for each metric is reported in bold.

When considering the results of the EXP.1 in Table 6.4, it is possible to note that for each dataset the IF approach has the best performance. On the MRI dataset, it achieves 78.78±.05% of ACC and 80.67±.03% of AUC on the BasicNet, and 84.80±.01% and 84.30±.02% of ACC and AUC respectively on the ResNet, outperforming the U approach also in terms of precision and recall for each class. When the TM is used, the IF-TM approach achieves 76.96±.05% of ACC and 81.80±.03% of AUC on the BasicNet, and 87.33±.01% and 93.32±.02% of ACC and AUC respectively on the ResNet. In the second section, the IF has values of accuracy equal to 80.45±.01% and 89.59±.04% on the BasicNet and ResNet respectively and values of AUC equal to 79.39±.03% and 84.50±.04% on the two CNNs, showing higher performance than the U approach in all the metrics. The introduction of TM increases the performance on BasicNet obtaining 82.42±.04% and 84.45±.09% in terms of ACC and AUC respectively while gaining 86.50±.04% in terms of AUC on ResNet. In the third section, the PAIRED dataset is analyzed, where the IF modality achieves at least 90% of accuracy on both networks. Indeed, it obtains 90.13±.03% and 92.20±.01% in that metric for the BasicNet and ResNet respectively, whilst it achieves 83.76±.06% and 90.13±.01% of AUC for the two networks, thus outperforming the U-MRI, the U-PET approaches, and the EF and LF fusion modalities. The TM increases the performance in terms of ACC in both networks. In the last section, it is possible to note that although the EF and LF show good results, the IF achieves 81.80±.03% and 82.77±.03% of accuracy and 86.70±.02% and 87.24±.01% of AUC on the BasicNet and ResNet respectively. Moreover, for both networks, it also has the highest performance in terms of precision and recall for each class. In the IF-TM configuration, ResNet achieves the best performance on each metric.

Table 6.5 focuses on the EXP.2, showing the results on the Task B2. On the MRI dataset, the IF modality outperforms the U approach, achieving values of ACC equal to 73.29±.03% and 82.20±.04% and values of AUC equal to 70.52±.05% and 75.44±.05% on the BasicNet and ResNet respectively. The IF-TM approach obtains 74.78±.04% and 74.29±.04% in terms of ACC and AUC respectively on BasicNet, while values equal to 85.4±.05 of ACC and 81.99±0.08% of ACC on ResNet. It also reports the best performance in terms of precision and recall for each class. In the second section, the PET dataset is considered, on which the IF achieves 84.62±.05% and 80.42±.07% for the accuracy and AUC respectively on the BasicNet, whilst it has 89.01±.07% and 80.75±.13% in the same metrics for the ResNet. Moreover, it also reports an improvement in precision and recall, when compared to the U approach. The TM improves the performance on both networks, obtaining 87.91±.05% and 88.08±.08% on ACC and AUC respectively on BasicNet, while 91.11±.06% and 82.42±.12% in those metrics on ResNet. On the PAIRED dataset, the IF-WMV-TM achieves the best performance on the BasicNet obtaining 95.00±.05% and 97.29±.09% on accuracy and AUC respectively. The IF-WMV-TM and the IF-TM report the same results in terms of accuracy on the ResNet, whilst the former outperforms the latter in terms of AUC. Indeed, the IF-WMV-TM achieves 94.44±.03% and 92.73±.07% of ACC and AUC respectively. In the last section, among all the MDL fusion modalities, the IF-TM introduces an improvement in all the metrics for both networks, achieving 78.09±.03% and 87.36±.03% in accuracy and the 76.22±.05% and 83.28±.05% of AUC for BasicNet and ResNet respectively.

When considering the results of the EXP.3 in Table 6.6, it is possible to note that the IF and IF-TM modalities introduce an improvement on each dataset when compared to the other approaches. On the MRI dataset, the IF achieves 72.60±.03% of ACC and 79.39±.04% of AUC on the BasicNet, and 75.67±.01% and 78.78±.02% of ACC and AUC respectively on the ResNet. Moreover, the IF-TM configuration achieves values of ACC equal to 74.27±.03% and 76.47±.01% on BasicNet and ResNet respectively, while values of ACC equal to 81.40±.05% and 80.09±.02% on BasicNet and ResNet respectively. In the second section, the results on the PET dataset show that the IF approach achieves values of ACC equal to 79.32±.04% and 82.84±.03% and values of AUC equal to 77.73±.05% and 81.80±.05% on the BasicNet and ResNet respectively. The IF-TM approach outperforms the IF configuration on ResNet, obtaining 85.55±.03% and 83.66±.05% on ACC and AUC respectively. The third section focuses on the PAIRED dataset, where the IF obtains 88.06±.04% in ACC for the two networks, while it achieves 86.15±.05% and 89.45±.04% of AUC for BasicNet and ResNet respectively, thus outperforming the U-MRI, the U-PET approaches, and the EF and LF fusion modalities. However, the IF-TM configuration achieves 89.81±.05% of ACC and 74.50±.13% of AUC on the BasicNet, and 93.15±.04% and 89.49±.03% of ACC and AUC respectively on the ResNet. In the last section, it is possible to note that the IF achieves 76.89±.03% and 77.93±.02% of accuracy and 80.53±.02% and 82.81±.01% of AUC on the BasicNet and ResNet respectively. Moreover, the IF-TM configuration achieves values of ACC equal to 77.71±.03% and 80.61±.03% on BasicNet and ResNet respectively, while values of ACC equal to 80.54±.02% and 84.60±.03% on BasicNet and ResNet respectively.

Table 6.7 focuses on the EXP.4, showing the results on the Task B. On the MRI dataset, the IF modality outperforms the U approach, achieving values of ACC equal to 72.53±.03% and 80.29±.01% and values of AUC equal to 80.51±.03% and 82.00±.02% on the BasicNet and ResNet respectively. The IF-TM configuration outperforms the IF one in the case of ResNet, obtaining a value of ACC equal to 83.53±.01% and a value of AUC equal to 80.51±.03%. In the second section, the PET dataset is considered, on which the IF achieves 78.62±.03% and 83.49±.04% for the accuracy and AUC respectively on the BasicNet, whilst it has 88.33±.03% and 84.13±.05% in the same metrics for the ResNet. Moreover, it also reports an improvement in recall for each class, when compared to the U approach. When the TM is included, it is possible to note an improvement in terms of accuracy and AUC on both networks. Indeed, the IF-TM approach achieves values of ACC equal to 81.01±.04% and 88.47±.03% and values of AUC equal to 86.80±.05% and 84.72±.04% on the BasicNet and ResNet respectively. The third section focuses on the PAIRED dataset, where both IF, IF-WMV, IF-TM, and IF-WMV-TM modalities achieve good performance. In particular, IF and IF-WMV show a value of ACC equal to 88.85±.04% and values of AUC equal to 84.67±.04% and 89.23±.04% respectively on the BasicNet. Moreover, the IF and IF-WMV approaches obtain 91.08±.01% and 89.01±.01% of ACC respectively on the ResNet, achieving values of AUC equal to 88.73±.04% and 94.38±.01%. The IF-TM and IF-WMV-TM show values of ACC equal to 93.47±.05% and 90.76±.03% and values of AUC equal to 73.97±.05% and 88.72±.04%

respectively on the BasicNet. Moreover, the IF-TM and IF-WMV-TM approaches obtain 98.09±.01% and 96.66±.01% of ACC respectively on the ResNet, achieving values of AUC equal to 96.36±.04% and 97.21±.01%. On the COMPLETE dataset, it is possible to note that the IF modality outperforms both EF and LF in all metrics. Indeed, it obtains 76.97±.02% and 82.32±.02% of ACC and 82.81±.02% and 84.97±.01% of AUC on the BasicNet and ResNet respectively. Furthermore, the IF-TM configuration improves the performance in terms of ACC on BasicNet and in terms of ACC and AUC on ResNet. Indeed, it achieves values of ACC equal to 78.23±.02% and 88.71±.02% and values of AUC equal to 82.51±.02% and 91.00±.01% on the BasicNet and ResNet respectively.

6.6 Discussion

Results show that for each experiment and for both investigated networks (BasicNet and ResNet), the solutions that involve the IF and the IF-TM configurations have the best performance, even when the three subnetworks are evaluated separately. Moreover, ResNet outperforms BasicNet in each experiment, confirming, as expected, the effectiveness of finetuning in the medical field. In contrast to BasicNet, ResNet can exploit the knowledge learnt during the previous training phase.

The analysis conducted with the PAIRED and COMPLETE datasets shows that in most experiments, the LF performed better than the EF, thus supporting the preference of the former over the latter. Indeed, the LF approach exploits two different networks, each trained for a specific image modality, leveraging a set of data larger than the one used in the EF. As a consequence, the two CNNs learn to extract features that reflect the distinctive characteristics of each modality, delaying the combination of the results in a post-processing step. On the other hand, in the EF solution, the MRI-PET pairs simultaneously contribute to the prediction, as they are organized in a series of 2-channel 3D volumes. However, the limited amount of data and the structure of the input do not allow the network to learn a shared features representation that at the same time is able to capture the distinctive characteristics of MRI and PET. In solutions involving the IF approach, the shared representation is created by concatenating features coming from the convolutional cores at an intermediate level, thus

Data	Net	Mod.	ACC	Precision		Recall		AUC
				C	NC	C	NC	
MRI	Ba	U	76.96±.02	89.54±.01	55.00±.01	77.65±.02	75.07±.01	79.80±.01
		IF	78.78±.05	89.35±.03	58.08±.06	80.67±.08	73.59±.09	80.67±.03
		IF-TM	76.96±.05	89.54±.03	55.00±.06	77.65±.08	75.07±.09	81.80±.03
	Re	U	80.36±.01	91.96±.01	59.78±.02	80.24±.02	80.71±.03	81.71±.02
		IF	84.80±.01	93.28±.01	67.47±.05	85.42±.02	83.09±.03	84.30±.02
		IF-TM	**87.33±.01**	**96.94±.03**	**69.80±.04**	**85.42±.02**	**92.58±.08**	**93.32±.06**
PET	Ba	U	78.90±.03	94.51±.01	33.88±.06	80.48±.04	68.13±.06	77.86±.05
		IF	80.45±.01	96.34±.01	37.50±.04	80.65±.01	79.12±.04	79.39±.03
		IF-TM	82.42±.04	96.96±.01	40.47±.08	82.42±.04	82.42±.08	84.45±.09
	Re	U	88.33±.04	95.90±.02	53.17±.07	90.48±.03	73.63±.13	79.00±.07
		IF	**89.59±.04**	**96.91±.01**	**56.59±.18**	**90.97±.05**	**80.22±.07**	84.50±.04
		IF-TM	**89.59±.04**	**96.91±.01**	**56.59±.18**	**90.97±.05**	**80.22±.07**	**86.50±.04**
PAIRED	Ba	U-MRI	80.89±.02	96.58±.01	35.00±.01	81.29±.02	77.78±.01	82.77±.01
		U-PET	78.50±.03	95.66±.01	31.14±.06	79.32±.04	72.22±.06	79.94±.05
		EF	82.96±.03	93.42±.03	34.23±.08	86.87±.05	52.78±.20	78.85±.08
		LF	82.48±.02	96.65±.02	37.33±.07	83.09±.05	77.78±.20	87.37±.70
		IF	90.13±.03	96.25±.01	55.32±.05	92.45±.03	72.22±.10	83.76±.06
		IF-WMV	88.22±.02	96.53±.01	49.09±.04	89.93±.03	75.00±.11	87.52±.05
		IF-TM	93.95±.04	96.42±.01	74.29±.22	96.76±.04	72.22±.08	79.14±.09
		IF-WMV-TM	88.85±.05	96.55±.01	50.94±.17	90.65±.04	75.00±.11	87.72±.08
	Re	U-MRI	82.17±.01	97.03±.01	37.18±.02	82.37±.02	80.56±.03	83.66±.02
		U-PET	88.22±.04	96.53±.02	49.09±.07	89.93±.03	75.00±.13	80.67±.07
		EF	79.30±.07	93.65±.03	29.29±.11	82.19±.08	56.94±.17	77.96±.07
		LF	82.01±.05	96.83±.02	36.77±.11	82.37±.08	79.17±.18	88.14±.07
		IF	92.20±.01	98.29±.01	61.17±.07	92.81±.02	87.50±.05	90.13±.01
		IF-WMV	90.61±.04	98.44±.01	55.65±.10	90.83±.05	88.89±.10	94.54±.03
		IF-TM	**98.57±.01**	98.41±.01	**100.00±.00**	**100.00±.00**	87.50±.05	96.41±.05
		IF-WMV-TM	96.97±.01	**98.91±.01**	83.54±.07	97.66±.01	**91.67±.10**	**97.20±.03**
COMPLETE	Ba	EF	80.53±.01	88.64±.02	61.63±.03	84.34±.02	69.94±.03	81.81±.01
		LF	79.72±.01	89.27±.02	59.58±.42	82.32±.06	72.47±.03	81.25±.02
		IF	81.80±.03	90.62±.02	62.94±.04	83.94±.05	75.84±.08	82.77±.03
		IF-TM	82.32±.03	90.69±.03	63.98±.04	84.65±.04	75.84±.08	82.73±.04
	Re	EF	82.24±.02	91.58±.15	63.21±.05	83.54±.02	78.65±.04	83.63±.02
		LF	80.83±.02	91.78±.14	60.43±.04	81.21±.01	79.78±.03	83.73±.01
		IF	86.70±.02	93.74±.01	71.12±.07	87.78±.03	83.71±.03	87.24±.01
		IF-TM	**91.83±.01**	**97.11±.02**	**79.85±.03**	**91.62±.02**	**92.42±.06**	**94.51±.03**

Table 6.4. Performance of the EXP.1 (Task B1). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures.

preserving the distinctiveness of the different image modalities, which is then exploited in the classification core in the case of MRI-PET pair. However, in the implemented training strategy, the loss is propagated back to both the MRI and PET convolutional cores when a *paired acquisition* is

Data	**Net**	**Mod.**	**ACC**	**Precision**		**Recall**		**AUC**
				M	**S**	**M**	**S**	
MRI	Ba	U	70.62±.09	84.12±.05	40.38±.16	75.97±.12	53.16±.14	70.05±.06
		IF	73.29±.03	85.90±.39	44.66±.06	77.91±.06	58.23±.14	70.52±.05
		IF-TM	74.78±.04	87.77±.04	47.22±.07	77.91±.06	64.56±.15	74.29±.04
	Re	U	68.84±.08	86.96±.02	40.00±.10	69.77±.10	65.82±.02	68.80±.06
		IF	82.20±.04	91.25±.02	59.79±.07	84.88±.05	73.42±.04	75.44±.05
		IF-TM	**85.46±.05**	**92.31±.01**	**66.67±.15**	**88.37±.07**	**75.95±.04**	**81.99±.08**
PET	Ba	U	74.73±.07	88.24±.08	34.78±.10	80.00±.11	50.00±.29	68.92±.07
		IF	84.62±.05	90.67±.03	56.25±.23	90.67±.06	56.25±.14	80.42±.07
		IF-TM	87.91±.05	94.44±.05	63.16±.19	90.67±.06	**75.00±.26**	**88.08±.08**
	Re	U	74.73±.03	91.94±.01	37.93±.09	76.00±.03	68.75±.11	77.83±.06
		IF	89.01±.07	93.33±.04	68.75±.23	**93.33±.06**	68.75±.16	80.75±.13
		IF-TM	**91.11±.06**	**94.59±.03**	**70.59±.22**	**93.33±.06**	**75.00±.15**	82.42±.12
PAIRED	Ba	U-MRI	80.56±.09	92.31±.05	50.00±.16	82.76±.12	71.43±.14	82.88±.06
		U-PET	75.00±.07	87.04±.08	38.89±.10	81.03±.11	50.00±.29	69.58±.07
		EF	79.17±.10	89.09±.04	47.06±.28	84.48±.11	57.14±.09	77.34±.06
		LF	84.72±.09	89.83±.06	61.54±.28	91.38±.11	57.14±.10	84.61±.07
		IF	86.11±.07	91.38±.01	64.29±.28	91.38±.094	64.29±.12	85.71±.12
		IF-WMV	88.89±.05	91.67±.01	75.00±.22	94.83±.07	64.29±.11	88.42±.06
		IF-TM	94.44±.03	93.55±.04	**100.00±.00**	**100.00±.00**	71.43±.21	92.24±.22
		IF-WMV-TM	**95.00±.05**	91.67±.04	91.67±.15	98.28±.03	71.43±.21	**97.29±.09**
	Re	U-MRI	79.17±.08	92.16±.02	47.62±.10	81.03±.10	71.43±.02	77.83±.06
		U-PET	70.83±.03	89.36±.01	36.00±.09	72.41±.03	64.29±.11	74.51±.06
		EF	70.83±.14	87.76±.08	34.78±.22	74.14±.19	57.14±.36	71.55±.09
		LF	81.94±.08	92.45±.03	52.63±.09	84.48±.02	71.43±.11	84.48±.08
		IF	90.28±.06	93.22±.04	76.92±.17	94.83±.05	71.43±.18	81.90±.20
		IF-WMV	90.28±.06	93.22±.04	76.92±.17	94.83±.05	71.43±.18	90.89±.08
		IF-TM	94.44±.03	93.55±.04	**100.00±.00**	**100.00±.00**	71.43±.18	81.90±.21
		IF-WMV-TM	94.44±.06	**95.00±.05**	91.67±.15	98.28±.04	**78.57±.22**	92.73±.07
COMPLETE	Ba	EF	70.22±.07	83.4±.04	37.86±.12	76.73±.10	48.15±.13	68.57±.06
		LF	70.22±.07	83.4±.04	37.86±.11	76.73±.09	48.15±.13	68.53±.06
		IF	76.12±.03	86.82±.04	47.96±.10	81.45±.07	58.02±.12	72.26±.07
		IF	76.12±.03	86.82±.04	47.96±.10	81.45±.07	58.02±.12	72.26±.07
		IF-TM	78.09±.03	88.63±.04	51.49±.09	82.18±.07	64.20±.14	76.22±.05
	Re	EF	70.22±.05	87.89±.02	40.60±.07	71.27±.07	66.67±.02	69.13±.07
		LF	70.51±.06	87.95±.02	40.91±.07	71.64±.07	66.67±.02	69.58±.07
		IF	83.71±.42	89.84±.03	55.00±.07	83.64±.04	67.90±.07	77.50±.06
		IF-TM	**87.36±.03**	**92.91±.02**	**70.45±.12**	**90.55±.05**	**76.54±.05**	**83.28±.05**

Table 6.5. Performance of the EXP.2 (Task B2). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures.

considered as input, leading the two features extracting processes to create a shared representation that is suitable for the task to solve. Moreover, the presence of the transfer module affects the features extraction process,

Data	Net	Mod.	ACC	Precision			Recall			AUC
				C	M	S	C	M	S	
MRI	Ba	U	70.86±.02	86.96±.04	37.42±.07	34.31±.09	80.67±.04	43.80±.02	44.30±.16	76.53±.04
		IF	72.60±.03	87.96±.02	41.37±.04	34.71±.16	82.07±.06	44.57±.09	53.16±.13	79.39±.04
		IF-TM	74.27±.03	89.36±.02	45.58±.07	36.00±.17	**82.51±.05**	50.00±.08	56.96±.16	**81.40±.05**
	Re	U	70.23±.03	91.81±.03	41.34±.05	32.82±.01	73.87±.03	62.02±.03	54.43±.07	78.53±.03
		IF	75.67±.01	91.75±.02	49.55±.05	38.66±.04	80.54±.01	**63.57±.04**	58.23±.05	78.78±.02
		IF-TM	**76.47±.01**	**92.09±.03**	**50.62±.03**	**43.41±.04**	80.54±.01	**63.57±.04**	**70.89±.15**	80.09±.02
	Re[91]	U	61.36±.04	84.47±.06	27.12±.06	16.33±.07	72.14±.10	32.17±.19	30.38±.02	71.11±.03
		IF	48.38±.05	86.84±.04	26.06±.02	15.17±.06	46.33±.08	59.69±.17	34.18±.22	68.66±.06
PET	Ba	U	73.14±.04	94.46±.02	21.89±.05	16.22±.39	76.94±.04	49.33±.08	37.50±.07	75.99±.06
		IF	79.32±.04	93.85±.02	26.95±.07	41.18±.29	83.71±.05	50.67±.08	43.75±.14	77.73±.05
		IF-TM	78.62±.03	93.77±.01	26.21±.15	**45.00±.30**	82.58±.04	50.67±.07	56.25±.18	77.73±.06
	Re	U	79.32±.03	**96.76±.01**	33.10±.09	20.45±.09	81.94±.02	62.67±.11	56.25±.03	80.70±.03
		IF	82.84±.03	96.35±.01	38.28±.12	34.29±.32	85.16±.03	**65.33±.07**	75.00±.23	81.80±.05
		IF-TM	**85.55±.03**	96.43±.02	**45.37±.07**	37.84±.26	**87.97±.05**	**65.33±.08**	**87.50±.14**	**83.66±.05**
	Re[91]	U	65.96±.09	95.53±.02	18.54±.07	8.73±.07	69.68±.08	37.33±.14	56.25±.37	74.76±.05
		IF	76.93±.07	89.13±.03	3.70±.04	6.49±.43	87.26±.09	1.33±.02	31.25±.38	70.34±.05
PAIRED	Ba	U-MRI	78.18±.02	95.03±.04	21.01±.07	26.92±.10	82.55±.04	43.10±.15	50.00±.11	80.33±.04
		U-PET	72.29±.04	95.65±.02	19.75±.05	14.71±.39	75.18±.04	53.45±.08	35.71±.07	78.32±.06
		EF	78.82±.05	93.45±.03	18.60±.09	21.05±.18	84.71±.07	27.59±.14	57.14±.37	76.97±.07
		LF	76.11±.04	96.88±.02	24.05±.08	27.27±.45	78.06±.05	65.52±.19	42.86±.35	84.06±.08
		IF	88.06±.04	94.85±.14	40.28±.28	66.67±.06	92.81±.12	50.00±.09	57.14±.05	86.15±.05
		IF-WMV	87.74±.05	94.99±.01	40.26±.15	66.67±.33	92.09±.06	53.45±.16	57.14±.09	87.34±.04
		IF-TM	89.81±.05	94.94±.01	46.97±.22	88.89±.15	94.42±.05	53.45±.11	57.14±.09	74.50±.13
		IF-WMV-TM	86.78±.04	95.09±.01	38.37±.23	**69.23±.31**	90.47±.05	56.90±.08	64.29±.09	85.29±.07
	Re	U-MRI	74.68±.03	97.22±.03	26.58±.05	20.51±.12	75.36±.03	**72.41±.03**	57.14±.08	82.31±.03
		U-PET	79.62±.03	97.23±.01	30.25±.09	20.00±.09	82.01±.02	62.07±.11	57.14±.27	80.39±.03
		EF	63.54±.26	95.05±.01	15.38±.07	8.16±.08	65.65±.30	51.72±.21	28.57±.21	74.81±.05
		LF	78.98±.07	96.81±.03	28.18±.36	20.83±.13	81.83±.06	53.45±.16	**71.43±.29**	88.84±.08
		IF	88.06±.01	97.10±.01	47.19±.10	38.10±.19	90.47±.02	**72.41±.08**	57.14±.09	89.45±.04
		IF-WMV	86.78±.02	96.69±.01	43.33±.08	40.00±.33	89.21±.03	67.24±.12	**71.43±.25**	91.70±.02
		IF-TM	**93.15±.04**	**97.27±.02**	**66.67±.15**	53.33±.34	**96.22±.05**	**72.41±.12**	57.14±.25	89.43±.03
		IF-WMV-TM	90.76±.04	96.66±.01	54.93±.23	55.56±.31	93.71±.05	67.24±.08	**71.43±.09**	**93.73±.02**
	Re[91]	IF	77.87±.06	93.79±.02	16.67±.14	15.79±.12	84.17±.09	20.69±.31	64.29±.25	77.10±.10
		IF-WMV	78.03±.06	93.60±.02	17.11±.12	17.31±.13	84.17±.09	22.41±.29	64.29±.25	76.98±.08
COMPLETE	Ba	EF	73.70±.03	87.54±.03	41.78±.06	37.89±.13	83.74±.03	46.18±.14	44.44±.16	77.02±.03
		LF	70.43±.03	87.56±.03	37.39±.06	34.65±.12	78.89±.03	48.00±.12	43.21±.12	76.36±.03
		IF	76.89±.03	87.93±.02	47.73±.09	43.75±.14	**87.58±.05**	45.82±.08	51.85±.11	80.53±.02
		IF-TM	77.71±.03	88.90±.02	51.45±.14	44.34±.14	86.57±.05	51.64±.06	58.02±.17	80.61±.03
	Re	EF	72.88±.01	91.73±.02	43.55±.02	33.59±.10	78.38±.02	58.91±.03	53.09±.07	81.19±.01
		LF	72.66±.01	91.90±.02	44.20±.02	31.94±.06	77.98±.05	58.18±.02	56.79±.06	81.43±.01
		IF	77.93±.02	91.80±.02	51.35±.06	45.05±.12	83.64±.28	**62.18±.03**	61.73±.04	82.81±.01
		IF-TM	**80.54±.02**	**92.30±.02**	**55.88±.05**	**51.33±.10**	86.40±.02	**62.18±.03**	**71.60±.08**	**84.60±.03**
	Re[91]	IF	62.70±.04	87.13±.02	33.76±.03	16.47±.06	69.09±.07	48.00±.16	34.57±.21	76.98±.04

Table 6.6. Performance of the EXP.3 (Task A). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures.

implementing the *cross-modality calibration* of the features maps. The result is a network in which the different specific modality paths influence each other, introducing an improvement in the performance obtained. When comparing Tables 6.6, and 6.7, it is possible to note that the hierarchical classification (Task B) has better performance than the three-class

Data	**Net**	**Mod.**	**ACC**	**Precision**			**Recall**			**AUC**
				C	**M**	**S**	**C**	**M**	**S**	
MRI	Ba	U	70.70±.01	89.54±.01	46.91±.03	24.32±.18	77.65±.02	50.00±.07	56.96±.16	77.92±.01
		IF	72.53±.03	89.35±.02	42.62±.03	31.97±.15	80.67±.05	50.39±.09	49.37±.13	80.51±.03
		IF-TM	70.86±.02	89.54±.02	44.18±.03	27.98±.13	77.65±.04	50.00±.08	59.49±.12	80.75±.01
	Re	U	73.00±.03	91.96±.01	46.26±.05	28.16±.05	80.24±.02	50.39±.09	62.03±.06	78.93±.03
		IF	80.29±.01	93.28±.02	55.08±.03	**50.00±.04**	**85.42±.01**	65.12±.03	69.62±.05	82.00±.02
		IF-TM	**83.53±.01**	**96.94±.02**	**68.23±.03**	40.54±.03	**85.42±.05**	**79.07±.03**	**75.95±.04**	**86.93±.02**
PET	Ba	U	76.09±.04	94.51±.02	26.87±.04	12.24±.39	80.48±.04	48.00±.08	37.50±.07	76.87±.04
		IF	78.62±.03	96.34±.02	28.81±.06	53.33±.28	80.65±.05	68.00±.08	50.00±.13	83.49±.04
		IF-TM	81.01±.04	96.96±.02	32.93±.05	55.00±.22	82.42±.04	72.00±.08	68.75±.14	**86.80±.05**
	Re	U	85.51±.03	95.90±.01	43.37±.09	25.58±.08	90.48±.02	48.00±.10	68.75±.06	78.58±.03
		IF	88.33±.03	96.91±.01	47.32±.12	**64.71±.29**	**90.97±.02**	**70.67±.06**	68.75±.23	84.13±.05
		IF-TM	**88.47±.03**	96.91±.01	**68.83±.09**	23.08±.25	**90.97±.01**	**70.67±.03**	**75.00±.22**	84.72±.04
PAIRED	Ba	U-MRI	78.98±.02	96.58±.03	28.33±.06	25.00±.10	81.29±.04	58.62±.13	**71.43±.10**	81.40±.04
		U-PET	75.96±.04	95.66±.02	24.39±.04	13.64±.42	79.32±.04	51.72±.08	42.86±.08	78.87±.06
		EF	81.21±.04	93.42±.03	21.43±.08	46.15±.43	86.87±.05	36.21±.18	42.86±.25	79.42±.08
		LF	77.87±.03	96.72±.02	26.57±.07	29.63±.47	79.68±.05	65.52±.19	57.14±.38	87.21±.08
		IF	88.85±.04	96.25±.03	57.38±.09	27.27±.05	92.45±.11	60.34±.09	64.29±.04	84.67±.04
		IF-WMV	88.85±.04	96.60±.01	55.88±.13	29.03±.29	91.91±.05	65.52±.11	64.29±.10	89.23±.04
		IF-TM	93.47±.05	96.42±.03	67.24±.09	83.33±.04	96.76±.12	67.24±.09	71.43±.04	73.97±.05
		IF-WMV-TM	90.76±.03	96.83±.01	53.25±.11	66.67±.25	93.35±.03	70.69±.11	71.43±.10	88.75±.04
	Re	U-MRI	80.10±.02	97.03±.03	33.02±.04	20.00±.12	82.37±.03	60.34±.03	71.43±.05	83.84±.03
		U-PET	85.35±.04	96.53±.02	38.03±.08	23.08±.08	89.93±.02	46.55±.12	64.29±.23	80.28±.03
		EF	76.59±.07	93.65±.02	24.62±.38	10.67±.15	82.19±.07	27.59±.19	57.14±.36	78.31±.08
		LF	80.25±.07	96.65±.02	29.46±.34	25.64±.15	82.91±.04	56.90±.15	71.43±.29	88.58±.07
		IF	91.08±.01	98.29±.01	53.33±.09	61.54±.16	92.81±.02	82.76±.10	57.14±.09	88.73±.04
		IF-WMV	89.01±.01	98.43±.01	45.71±.09	69.23±.24	90.29±.02	82.76±.16	64.29±.20	94.38±.01
		IF-TM	**98.09±.01**	98.41±.01	**94.44±.09**	**100.00±.00**	**100.00±.00**	87.93±.10	64.29±.09	96.36±.04
		IF-WMV-TM	96.66±.01	**98.91±.01**	81.25±.09	78.57±.24	97.84±.02	**89.66±.13**	**78.57±.20**	**97.21±.01**
COMPLETE	Ba	EF	74.37±.03	88.90±.03	44.08±.04	32.38±.12	84.14±.03	48.73±.11	41.98±.15	79.52±.03
		LF	71.99±.02	89.14±.03	40.71±.04	30.70±.11	80.40±.03	50.18±.11	43.21±.11	78.73±.02
		IF	76.97±.02	90.73±.01	54.81±.07	31.08±.11	85.05±.04	53.82±.07	56.79±.10	82.81±.02
		IF-TM	78.23±.02	90.94±.01	54.71±.07	37.12±.11	86.16±.04	54.91±.07	60.49±.10	82.51±.02
	Re	EF	75.71±.01	91.66±.02	50.19±.02	30.00±.08	84.34±.02	48.36±.03	62.96±.07	80.88±.02
		LF	73.92±.02	91.70±.03	45.82±.03	30.54±.06	81.52±.03	49.82±.02	62.96±.05	80.96±.01
		IF	82.32±.02	93.72±.01	58.07±.04	**55.00±.11**	87.47±.20	68.00±.02	67.90±.03	84.97±.01
		IF-TM	**88.71±.02**	**97.11±.01**	**77.78±.04**	50.41±.11	**91.72±.20**	**81.45±.02**	**76.54±.03**	**91.00±.01**

Table 6.7. Performance of the EXP.4 (Task B). First and second sections report the performance obtained on the MRI and PET datasets respectively, considering the U and IF approaches. Third section shows the performance of the MDL fusion techniques on the PAIRED dataset in comparison with the U-MRI and U-PET configurations. The last section reports the results obtained with the COMPLETE dataset. Ba and Re denote the BasicNet and ResNet architectures.

classification (Task A). This happens because in Task B the existing relationship between classes is explicitly exploited, while Task A delegates the model the role of understanding the connection between the labels. However, the small size and high imbalance of the dataset make this task very complex.

The work proposed in [91] is also considered since it represents the first article considering MDL approaches for the OASIS-3 dataset [74]. Table 6.6 shows that the implemented methodology outperforms the solution proposed in [91] in all metrics and for each dataset. Indeed, the U approach proposed in [91] obtains 61.36±.04% and 71.11±.03% in terms of ACC and AUC respectively on MRI dataset, and 65.96±.09% and 74.76±.05% in those metrics on the PET dataset. In the third section, the IF approach proposed in [91] achieves a value of accuracy equal to 77.87±.06% and a value of AUC equal to 77.10±.10%, while on the COMPLETE dataset it has 62.70±.04% and 76.98±.04% in terms of ACC and AUC respectively. The implemented approach differs from that presented in [91] for the characteristic of the CNN involved and for the strategy implemented to handle the incomplete dataset, consisting of replacing missing modalities with black images. These characteristics negatively affect performance. Furthermore, the 2D CNN implemented in [91] only considers the median slice as input without exploiting the volumetric information of the involved acquisitions, while the use of black images may affect the process of features learning, especially if the number of incomplete inputs is greater than the number of paired acquisitions. As a consequence, the solution described in this thesis takes these aspects into account by proposing a training strategy that handles both paired and incomplete data.

When comparing the approach with other proposals in the literature, it is possible to note that the IF methodology has performance comparable to the solution proposed in [127] for EXP.1, EXP.3, and EXP.4 on the MRI dataset. The authors in [127] proposed a 3D CNN trained with MRI volumes and reported only the accuracy and the balanced accuracy[1] to assess the performance. In particular, in the case of Tasks B1 and A the authors achieved values of accuracy equal to 84% and 77% respectively; moreover, in the same tasks, the balanced accuracy is 84% and 76%, respectively. In the first section of Table 6.4, note that the implemented IF approach achieves 84.80±.01% in accuracy and 84.26±.03% in balanced accuracy for Task B1, while for the three classes classification, the solution in [127] is compared with both EXP.3 and EXP.4. The first section of Table 6.6 shows that the IF methodology obtains 75.67±.01% and 67.45±.02% in accuracy and balanced accuracy respectively, whilst the first section of Table 6.7 reports 80.29±.01% and 73.39±.03% in the same metrics. On the other hand, the IF-TM approach outperforms the solution proposed in [127] for

---

[1] In our experiments, the balanced accuracy is computed by averaging the recall for each class.

EXP.1 (Task B1) and for EXP.4. (Task B). The proposal is not compared with the other state-of-art solutions focusing on the OASIS-3 dataset [74] [113, 3, 55, 102, 61, 15] since they conducted an image-level rather that a patient-level split of the dataset, introducing a bias in the performance evaluation step.

Part III

Synthetic Data Sources

The ability of Convolutional Neural Network (CNNs) to autonomously learn during the training stage the best input representation for the specific task under analysis, has enables their applications in domains lacking well-defined, effective sets of features. This is the case of medical imaging that refers to the generic process through which it is possible to observe an area of a body not visible from the outside. Data augmentation is an essential part of training a discriminative CNN [53] and it is introduced with the aim of achieving more generalizable and accurate models based on relatively small labeled datasets, typical of the medical domain. A variety of augmentation strategies including flips, intensity operations, random rotation, and crops have been proposed to apply transformations to images improving the model's performance. Moreover, recent applications leverage DL approaches for the implementation of augmentation techniques. In particular, generative networks automatically learn the representation of images and create realistic samples to introduce variability in the set of data used for training. This process generates a set of *synthetic* images that aims to improve the generalization ability of the implemented model. As a consequence, the training process considers two distributions or sources of data, namely the *real* and the *synthetic* one.

However, despite promising, DL naive use may not be effective since *medical images are more than pictures* [38]. A notable example is Dynamic Contrast Enhanced-Magnetic Resonance Imaging (DCE-MRI), in which the kinetic of the injected Contrast Agent (CA) is crucial for the analysis of the disease. In the case of DL-based augmentation approaches, the generative model is designed to create samples that reflect the characteristics of the real data, making the difference in the two distributions as small as possible. In the medical domain, this implies that the physiological properties of the tissue under analysis need to be preserved, i.e the kinetic behavior of the CA flowing in the case of DCE-MRI. The result is a set of generated images that can be integrated with the available dataset.

Chapter 7 illustrates the physiologically based pharmacokinetic (PBPK) models used to describe the contrast agent kinetic behavior and the basic notions for the generation of synthetic images through a process that aims to preserve the biological characteristics. Furthermore, the presented concepts are further exploited in Chapter 8, dealing with a case of study that considers the breast lesion classification in DCE-MRI. In particular, a nested deep architecture is introduced to disentangle the CA effect from all the other image components while learning how to classify breast lesions with a CNN. Moreover, the physiologically-aware synthetic image generation process is exploited to propose an innovative data augmentation approach that introduces shape variability improving the classification task.

Chapter 7

Physiologically-aware Synthetic Data Generation

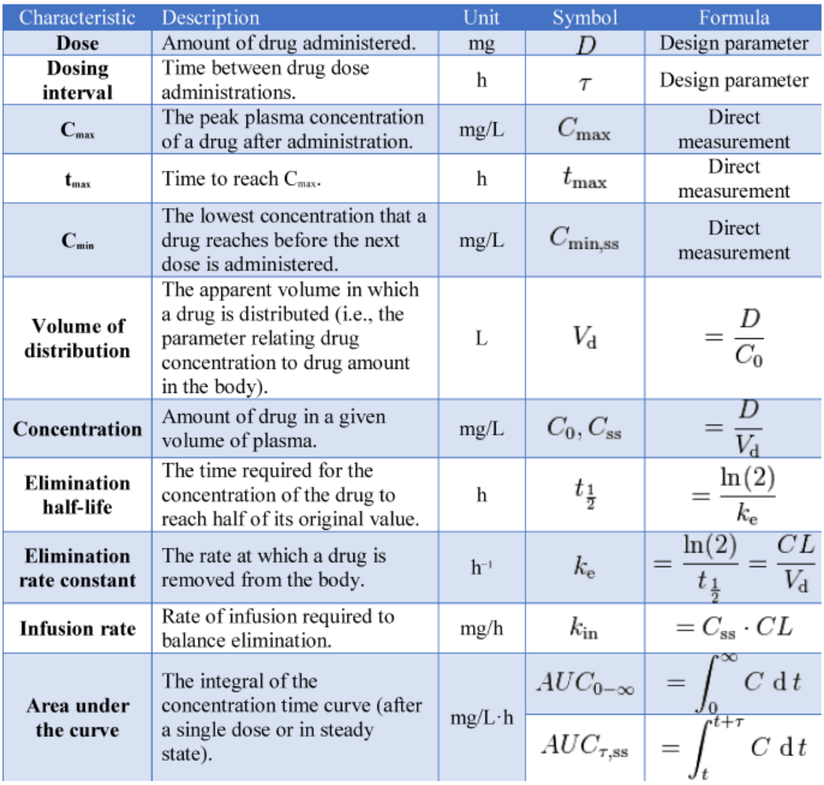
One of the research areas where deep learning-based solutions are being used most frequently is biomedical image processing, where the number of papers published each year nearly doubles. Although there are several tasks for which image processing plays an important role, cancer analysis is definitely one of the most crucial. Indeed, cancer is one of the leading causes of death in the western world [35], with the total number of cases undergoing a sizeable increase in the last years. Therefore, early diagnosis is still crucial for a favorable prognosis. Since the World Health Organization (WHO) suggests the use of imaging modalities in a variety of cancer screening guidelines [154, 104], several studies have been proposing to leverage Convolutional Neural Networks (CNNs) to analyze biomedical imaging techniques. However, despite some works show very promising results [156, 165, 167], the use of Deep Learning (DL) without considering all the physiological characteristics of medical imaging may not be the most effective solution since. as already said "*medical images are more than pictures*" [38]. This is especially true for all medical imaging techniques where the diagnostic-related information is not only associated with the image texture but, for example, with the variations due to the flowing of a Contrast Agent (CA). In particular, Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) has shown great effectiveness in the diagnosis of several diseases and consists of the intravenous injection of a paramagnetic CA whose flowing within tissue highlights both morphological and physiological characteristics. The Physiologically Based Pharmacokinetic (PBPK) modeling is a mathematical modeling technique for describing the diffusion of the CA from the blood pool into the extracellular space, enabling quantitative analysis of contrast agent distributions in the body and its relation to the characteristics of tumors. Indeed, in addition to an assessment of morphological features, the kinetic behavior of the CA has diagnostic potential [72].

7.1 Physiologically Based Pharmacokinetic models

Physiologically Based Pharmacokinetic (PBPK) modeling is a technique developed in order to model (estimate and predict) the absorption, distribution, metabolism and excretion (ADME) of substances (including contrast agents, hormones, nutrients, toxins, etc.) in animal species. The majority of these concepts are represented in a drug concentration graph, where the involved mathematical formulations are reported in Figure 7.1. PBPK modeling is performed by compartmental or non-compartmental methods:

- **Compartmental methods**: assume that any organism is formed by a number of related compartments (organs or tissues) and interconnections (blood or lymph flows). The drug concentration in each compartment is uniform and may be estimated by solving a system of differential equations, where its parameters represent blood flows, pulmonary ventilation rate, organ volumes, etc. Pharmacokinetic models are used in the compartment of the organ or tissue under study to determine the concentration-time graph at any time. The disadvantage consists of the difficulty in developing and validating a proper model.

- **Non-compartmental methods**: compute the exposure to a drug by estimating the area under the curve of a concentration-time graph. Since they do not assume any specific compartmental model, they are often more versatile with accurate results. However, the estimation of total drug exposure, which is frequently assessed using the area under the curve (AUC) approaches (numerical integration), is crucial for PBPK non-compartmental analysis. Additionally, models that are non-compartmental in nature can only predict the concentration for a single time instant.



Characteristic	Description	Unit	Symbol	Formula
Dose	Amount of drug administered.	mg	D	Design parameter
Dosing interval	Time between drug dose administrations.	h	τ	Design parameter
C_{max}	The peak plasma concentration of a drug after administration.	mg/L	$C_{\max}$	Direct measurement
t_{max}	Time to reach C_{max}.	h	$t_{\max}$	Direct measurement
C_{min}	The lowest concentration that a drug reaches before the next dose is administered.	mg/L	$C_{\text{min,ss}}$	Direct measurement
Volume of distribution	The apparent volume in which a drug is distributed (i.e., the parameter relating drug concentration to drug amount in the body).	L	V_d	$= \frac{D}{C_0}$
Concentration	Amount of drug in a given volume of plasma.	mg/L	C_0, C_{ss}	$= \frac{D}{V_\text{d}}$
Elimination half-life	The time required for the concentration of the drug to reach half of its original value.	h	$t_{\frac{1}{2}}$	$= \frac{\ln(2)}{k_\text{e}}$
Elimination rate constant	The rate at which a drug is removed from the body.	h^{-1}	k_e	$= \frac{\ln(2)}{t_{\frac{1}{2}}} = \frac{CL}{V_\text{d}}$
Infusion rate	Rate of infusion required to balance elimination.	mg/h	k_{in}	$= C_{\text{ss}} \cdot CL$
Area under the curve	The integral of the concentration time curve (after a single dose or in steady state).	mg/L·h	$AUC_{0-\infty}$	$= \int_0^{\infty} C \, \mathrm{d}t$
			$AUC_{\tau,\text{ss}}$	$= \int_t^{t+\tau} C \, \mathrm{d}t$

Figure 7.1. Most commonly measured PBPK metrics.

For the ability to predict the contrast agent concentration at any given time, this thesis makes use of compartmental PBPK modeling for tumor tissue. The choice of modeling only tumour tissues relays on the higher reliability of tumour tissues models with respect to healty ones. Indeed, the tumor area is characterized by a large vascularization. It's important to note that these models don't always accurately depict the exact situation within an organism. For example, the drug will be distributed more slowly

in some body tissues than in others with a higher blood supply. Additionally, according to the characteristics of the drugs, some tissues might be penetrated more easily than others. For these reasons, the PBPK compartmental model that is employed must be carefully adapted to the particular tissue. This indicates that even if the suggested technique is generally applicable, the evaluation is strongly related to the specific organ.
All of the compartmental models used in literature make some basic assumptions

- The contrast agent (tracer) concentration $C_t(t)$ is uniform within a compartment.
- The contrast agent flux between two compartments is proportional to the concentration difference between them
- The parameters used to describe compartments are considered fixed.
- The contrast agent can flow from nearby capillaries to the extracellular space (EES) and vice-versa; however diffusion through the EES from more distant capillaries is possible and (if present) would render the simple modeling invalid.
- The variation of the MR T1 constant is proportional to the contrast agent concentration $C_t(t)$.
- The $C_t(t)$ is the real contrast agent concentration, while the $C_t^{measured}(t)$ is the value derived from DCE-MRI signal intensity value.

Many approaches differ from each other in the basic physiological assumption about blood plasma and extra-vascular extracellular space (EES) and the consequent mathematical models to represent them.
The Tofts-Kermode (TK) model [146] uses a simple compartmental approach to represent blood plasma and the extra-vascular extracellular space (EES), in which the intra-vascular compartment can be neglected.
According to this model, the contrast liquid concentration $C_t(t)$ might be calculated as the convolution between the signal $C_p(t)$, (the specific Arterial Input Function, also known as AIF) and the exponential impulsive response, following the law:

$$C_t(t) = K^{trans} \int_0^t e^{-\frac{K^{trans}}{v_e}(t-s)} \cdot C_p(s) ds \tag{7.1}$$

in which K^{trans} is a combined measure of blood flow and capillary permeability and v_e is the EES volume per tissue volume unity.
In later work, the Tofts-Kermode model was extended in several ways. The most known one is the Extended Tofts-Kermode model [147], an improvement of the basic TK model in which the plasma volume fraction vp has been added assuming that the intra-vascular compartment influence can not be neglected:

$$C_t(t) = K^{trans} \int_0^t e^{-\frac{K^{trans}}{v_e}(t-s)} \cdot C_p(s) ds + v_p C_p(s) \tag{7.2}$$

Generally, the two AIF used in the DCE-MRI context are:

- **Weinmann AIF**[155]: a bi-exponential AIF

  $$C_p(t) = D \cdot \sum_{i=1}^{2} a_i e^{-m_i t} = D \cdot (a_1 e^{-m_1 t} + a_2 e^{-m_2 t})$$

  where D is the contrast agent injected quantity, a_i the exponential impulses extent and m_i the decay time-constant.

- **Parker AIF** [107]: a model composed by the sum of two gaussians and one exponential multiplied by one sigmoid

  $$C_p(t) = \sum_{i=1}^{2} \frac{A_i}{\sigma_i \sqrt{2\pi}} e^{\frac{-(t-T_i)^2}{2\sigma_i^2}} + \alpha \frac{e^{-\beta t}}{1 + e^{-s(t-\tau)}}$$

  where A_i is the scaling constant, σ_i and T_i the extent and the centers of two gaussians, α and β the extent and the decay time-constant of the exponential, s and τ the extent and the center of the sigmoid.

7.2 Physiologically-aware Data Generator

Generative models are able to create new samples considering a latent representation as input, implementing a process known as *image synthesis* that aims to artificially generate *synthetic instances*.

In image processing, the generative model, or generator, is typically a deep network (i.e CNN) that learns to map a fixed latent distribution p_z to the distribution of real data p_x. Denoting with f_g the mapping function represented by the model, the process of synthetic instance generation can be formalized as follows:

$$I_s = f_g(Z) \tag{7.3}$$

where Z is a sample from the distribution p_z ($Z \sim p_z(Z)$) corresponding to a latent representation and I_s represents the generated instance belonging to the distribution p_g, that is the generative model's distribution. The goal of the generative modeling algorithm is to learn p_g which approximates p_x as closely as possible [39] with the aim of creating new samples preserving the intrinsic characteristics of the real ones, especially the physiological properties in the case of the medical domain.

The distribution of real data includes all the available samples (i.e training images), while the generation process depends on the fixed latent distribution p_z. Several approaches (i.e Generative Adversarial Networks) consider p_z as a random distribution, However, in this thesis, an Intrinsic Deforming Autoencoder (Intrinsic-DAE, or DAE), is used as a generative model for its ability to provide a latent distribution p_z while disentangling factors of variations that are exploited in the process of image manipulation.

The Intrinsic DAE is a generative network proposed in [133] for images that disentangles shape from appearance. It makes use of the basic notion that creating an image involves combining two processes: a synthesis of appearance on a coordinate system with no distortion (referred to as a "template"), followed by a second deformation that includes shape diversity. Moreover, the appearance represents the "texture" in the image that can be further decomposed in the components of *albedo* and *shading*, corresponding to the overall brightness of an object and to the depth information respectively. Denoting with $T(p)$ the value of the synthesized appearance at coordinate $p = (x, y)$, with $A(p)$ and $S(p)$, the albedo and shading components and with $W(p)$ the estimated deformation field, an image I_{rec} is reconstructed as follows:

$$I_{rec} = T(W(p)) \tag{7.4}$$

where

$$T(p) = A(p) \circ S(p) \tag{7.5}$$

with $\circ$ indicating the Hadamard product.

The DAE is an unsupervised encoder-decoder architecture designed to disentangle an image in its main components of albedo (A), shading (S), and deformation field (W). In this thesis, the network is trained according to the ability to reconstruct the input image starting from its components. Moreover, the physiological characteristics of the medical images are preserved during the disentangling process by introducing the evaluation of the biological consistency of the generated images (i.e consistency with the PBPK model in the case of CA flowing).

The network architecture consists of an Encoder (E) and three different decoders (D_a, D_s, D_w), for the synthesis of A, S and W. The architecture is shown in Figure 7.2. In particular, E is represented in Figure 7.3 and consists of a batch normalization layer (BN), followed by ReLU function, convolutional operation (CONV) with 32 output channels, and a chain of dense encoder blocks and encoder transition blocks. $DBE(n, k)$ is a dense encoder block consisting of k stacked $3 \times 3 \times n$ convolutions with stride and padding values set to 1, each preceded by batch normalization and ReLU function. In this case, n represents the number of input and output channels. On the other hand, $TBE(m, n, p)$ is an encoder transition block composed of batch normalization, leaky ReLU function (LeakyReLU), $1 \times 1 \times n$ convolutions with m input and n output channels, and max-pooling operation.

The decoders D_a, D_s have the same architecture, while D_w differs only for activation functions. In particular, each decoder, shown in Figure 7.4, consists of a batch normalization layer (BN), followed by an activation function that is ReLU in D_a, D_s or hyperbolic function (tanh) in D_w. Then the network presents a transpose convolutional operation (ConvTranspose) with 256 output channels, and a chain of dense decoder block and decoder transition block. $DBD(n, k)$ is a dense decoder block consisting of k stacked $3 \times 3 \times n$ transposed convolution, with stride and padding values set to 1, each preceded by batch normalization and ReLU function. $TBD(m, n)$ is a decoder transition block composed of batch normalization, ReLU function and a $4 \times 4 \times n$ transpose convolution with m input channels, with stride and padding values set to 2 and 1 respectively. The final layers consist of a batch normalization layer (BN) followed by an activation function, that is ReLU in D_a, D_s or hyperbolic function (tanh) in D_w, and a transpose convolution operation.

The resulting Intrinsic DAE is used as a generator of synthetic images, exploiting a methodology that consists of two main steps, namely the *Latent representation definition*, in which the p_z distribution is determined, and *Image synthesis* that aims to produce new samples.

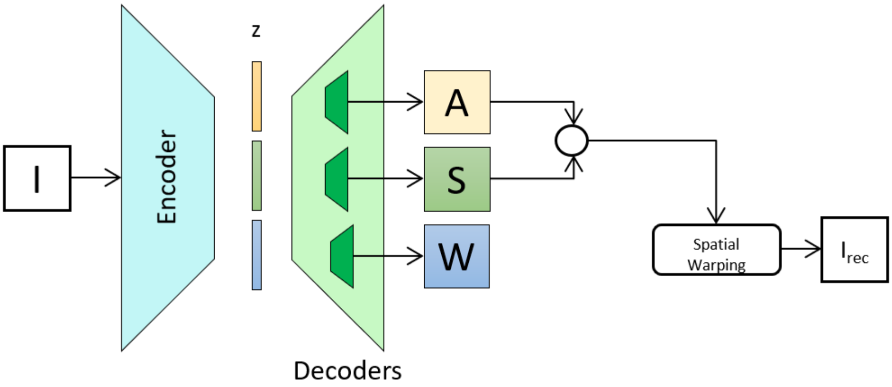


Figure 7.2. Architecture of the Intrinsic Deforming Autoencoder, consisting of an encoder and three different decoders.

7.2.1 Latent representation definition

This step focuses on the real data distribution p_x with the aim of defining the latent distribution p_z during the disentangling process. In particular, the real input image $I \sim p_x$ feeds the encoder (E) that delivers the low-dimensional latent representation Z that is split into three parts $Z = [Z_s, Z_a, Z_w]$ providing the clear separation of components. Each of these parts is fed to a different decoder $(D_s, D_a,\ D_w)$ that generates A, S and W. The texture element (T) is then computed as the Hadamard product ($\circ$) between the albedo and the shading while the output of the network is the reconstructed image I_{rec}, belonging to the distribution of reconstructed images $p_{I_{rec}}$.

It is possible to note that while E is responsible for the low-dimensional representation Z, the three decoders act in the process of image recon-

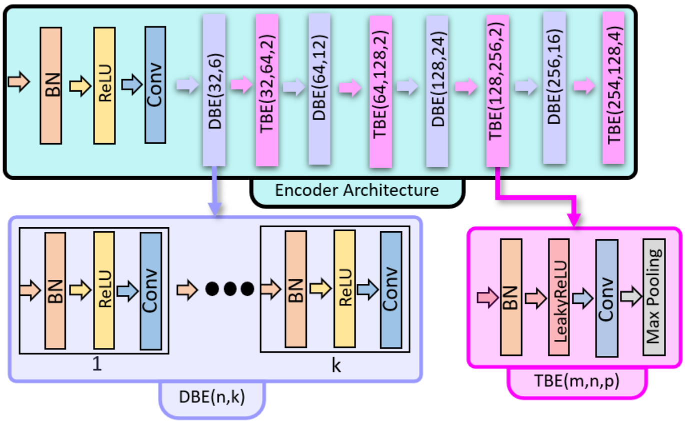


Figure 7.3. The considered DAE encoder, highlighting the DBE and TBE elements.

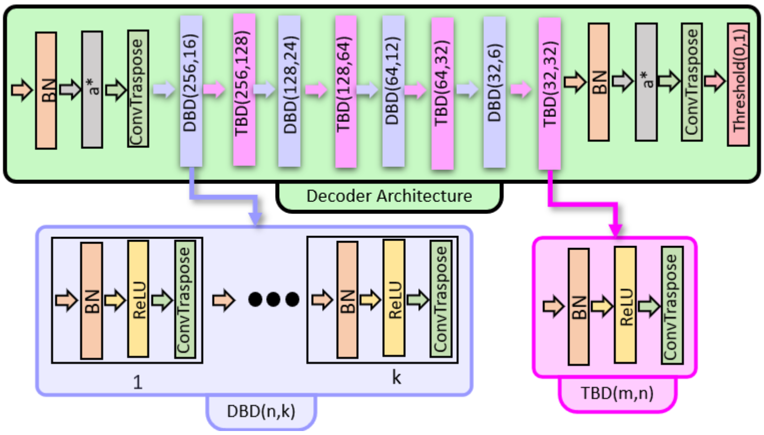


Figure 7.4. The considered DAE decoder. The decoders D_a, D_s have the same architecture, while D_w differs only for activation functions (a^*), that is ReLU in D_a, D_s or hyperbolic function (tanh) in D_w. The core elements are $DBD(n,k)$ and $TBD(m,n)$.

struction, as shown in Figure 7.2. In particular, the encoder transforms the input images in a latent representation Z learning the map between p_x and p_z. In other words, it generates the latent distribution that is responsible for the image generation process. Since each vector Z is split

into three parts ($Z = [Z_s, Z_a, Z_w]$), different distributions are generated, namely p_a, p_s and p_w, representing the latent distribution for the albedo, shading and deformation field components, with $Z_s \sim p_s(Z_s)$, $Z_a \sim p_a(Z_a)$ and $Z_w \sim p_w(Z_w)$. Each $I \sim p_x$ is associated with a specific $Z \sim p_z(Z)$ vector introducing a consistency constraint that the distributions p_s , p_a and p_w need to respect. In other words, the three distributions present a dependency relationship. In particular, each $Z_s \sim p_s(Z_s)$ corresponds to a specific $Z_a \sim p_a(Z_a)$ that is associated with a unique $Z_w \sim p_w(Z_w)$.
The reconstruction process that generates $I_{rec} \sim p_{I_{rec}}$ is formalized as a function f_g that exploits the components Z_s, Z_a, Z_w, the three decoders D_s, D_a, D_w and includes the Equations 7.4 and 7.5. In particular, the I_{rec} is computes according to Equation 7.3 that is detailed as follows:

$$I_{rec} = f_g(Z_a, Z_s, Z_w) \tag{7.6}$$

In the *Latent representation definition* step, the DAE aims to generate the input image from the latent representation acting as an autoencoder, while performing the decomposition. In particular, the loss function proposed in [133] [1] for training includes the reconstruction loss, as the standard l_2 loss, the warping and adversarial losses that aim to make the network able produce realistic images by preventing the introduction of inconsistent deformations. To make the disentangling process aware of the physiological characteristics of the images, the loss function is modified by introducing a component E_{PBPK} that performs the evaluation of the biological consistency of the generated image. In the case of DCE-MRI images, the E_{PBPK} considers the kinetic behavior of the CA flowing exploiting the pharmacokinetic models.

7.2.2 Image synthesis

The Encoder (E) trained in the previous step provides the distribution of the latent representation p_z, which in turn generates p_a, p_s and p_w, associated by a strict consistency constraint. The *Image synthesis* process exploits the three latent distributions, the decoders D_s, D_a, D_w and the generation function f_g to produce synthetic images $I_s \sim p_g$, that represent

---

[1] Please refer to [133] for a mode detailed description

new samples. Moreover, the E_{PBPK} component in the loss function prevents biological inconsistency in the generated images.

According to Equation 7.6, an image is generated starting from the latent representations of albedo ($Z_a \sim p_a(Z_a)$), shading ($Z_s \sim p_s(Z_s)$) and deformation field ($Z_w \sim p_w(Z_w)$). When the constraints between p_a, p_s and p_w are considered as described in Section 7.2.1, the generated image is the I_{rec} that does not correspond to a new sample since the network is trained with the aim of making $I_{rec} \sim p_{I_{rec}}$ equal to the real image $I \sim p_x$.

The ability of the DAE to create instances is exploited to generate synthetic images starting from the latent distributions p_a, p_s, and p_w. In particular, the generation function f_g considers a latent representation Z as input, split into three main components. As a consequence, the process of generating new instances results in the challenge of introducing variability into latent representations. According to the consistency constraint, the selection of Z_a automatically defines the Z_s and Z_w components to select. However, the variability in the Z representation can be introduced by removing the dependency constraint between p_w and the other two distributions p_a, p_s. As a result, when Z_a and Z_s are chosen form p_a and p_s respectively, a random element Z_w from p_w is used generating the synthetic representation $\tilde{Z}$. For each Z_a and Z_s paired components, different $\tilde{Z}$ can be obtained simply by changing the information related to the deformation field. It is worth noting that the dependency between p_a and p_s should be retained as the albedo and shading equally contribute to the texture of the image, providing information about the illumination from different yet correlated perspectives: albedo corresponds to the brightness while the shading to the depth characteristics. The effect is a generator network that creates synthetic images by applying different deformation fields to a "template" object to which a texture is applied.

The generation of the synthetic image I_s, belonging to the distribution of the generated instances p_g, exploits the vector $\tilde{Z}$ and the function f_g since $I_s = f_g(\tilde{Z})$. As aforementioned, the aim of the generative model is to learn how to create p_g which approximates p_x as closely as possible [39] with the aim of producing new samples preserving the intrinsic characteristics of the real ones, especially the physiological properties in the case of the medical domain.

During the *Image synthesis*, the encoder trained in the previous step can used to the definition of the p_a, p_s and p_w from which the Z_a, Z_s, and Z_w vectors are selected, without further training iterations. On the other hand, the decoders are initialized with the weights determined in the previous step and further trained considering a loss function that consists of the E_{PBPK} components, the warping, and the adversarial losses. Indeed, the absence of a reference image makes it impossible to use the reconstruction loss while the presence E_{PBPK} forces the model to generate physiologically consistent images.

Chapter 8

Synthetic images: Breast Lesion Classification in DCE-MRI

Since the 1970s, there has been a marked increase in the number of breast cancer occurrences, which has become one of the most widespread forms of tumor [57]. Among supplemental diagnostic approaches recommended by WHO [166], Dynamic Contrast Enhanced-Magnetic Resonance Imaging (DCE-MRI) is one of the most successful techniques in detecting different types of tumors, becoming increasingly popular in breast cancer [80]. Since the kinetic behavior of the contrast agent (CA) provides potential information about the lesion aggressiveness [72], the work described in [40] proposes the 3TP-CNN approach which is a straightforward yet effective solution to make any CNN able to properly exploit the CA effects in breast DCE-MRI. In particular, it considers the Three Time Points (3TP) method [29] to transform any DCE multi-channel slice into a 3TP slice (i.e. a 3-channels image). This thesis wants to make a step further toward the design of a domain-aware DL approach in the contest of breast DCE-MRI, *introducing a new nested deep architecture designed to learn how to disentangle the CA effects from all the other image components while preserving the physiological characteristics of the tissue under analysis and classifying breast lesions with a CNN*. Moreover, an innovative data augmentation approach is proposed implementing a *physiologically-aware synthetic im-*

age generation process that introduces shape variability with the aim of improving the classification task. In particular, the physiological characteristics of the slices are preserved by including the physiologically based pharmacokinetic (PBPK) models in the training process. The main contributions of the proposed approach can be summarized as follows:

- Exploiting an intrinsic Deforming Autoencoder (DAE) [133] to learn how to disentangle the effects of the CA dynamics in breast DCE-MRI 3TP slices;
- Designing a Generative Adversarial Network (GAN)-like architecture, where the discriminator is a 3-channels CNN and the generator is the previously mentioned DAE;
- Defining a multi-stage training strategy to make the deep architecture learn how to classify breast lesions while preserving all the physiological characteristics of the tissue under analysis;
- Introducing an innovative data augmentation approach that exploits the DAE as a generator to implement a physiologically-aware synthetic image generation process.

8.1 Breast Lesion Classification in DCE-MRI

DCE-MRI requires the intravenous administration of a paramagnetic Contrast Agent (CA) characterized by specific wash-in and wash-out times. The CA dynamics highlight both morphological and physiological characteristics of the tissues vascularization, allowing the identification of damaged tissues with respect to the surrounding healthy ones. Indeed, solid tumors present an abnormal and irregular growth of blood vessels. A DCE-MRI scan consists of several 3D MRI volumes (at least two), taken before (pre-contrast) and after (post-contrast) the intravenous injection of the CA. The result is a 4D volume with three spatial and one temporal dimensions. As a consequence, each voxel is associated with a Time Intensity Curve (TIC) reflecting the variation of the signal intensity due to the absorption and the release of the CA over the different acquired 3D series (Figure 8.1). Although this seems to suggest that the lesion malignancy could be easily estimated by visually analysing the TIC [36], the strong differences between real and illustrative TICs, the huge amount of data to analyze and the large intra/inter subject variability make the accurate and precise quantification of lesion malignancy a very hard, error-prone and time-consuming task. To face these problems, physicians rely on Computer-Aided Detection and Diagnosis (CAD) systems, namely tools designed to assist in the analysis of cancerous lesions by means of computer vision and, more recently, pattern recognition algorithms.

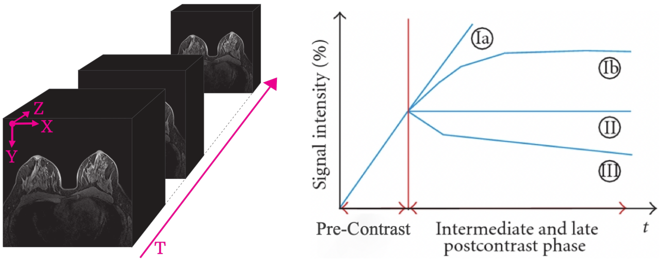


Figure 8.1. An illustrative breast DCE-MRI study (on the left) and some descriptive TICs (on the right): *Type I*, indicative of healthy tissues or of benign lesions; *Type II*, typical of borderline/probably malignant lesions; *Type III*, common in malignant lesions.

The task of lesion diagnosis consists of categorizing lesions, automatically or semi-automatically segmented into Regions of Interest (ROIs) by radiologists, according to their aggressiveness. It can be approached as the binary classification problem of differentiating between benign and malignant lesions at its extreme. Despite the fact that many literature ideas rely on hand-crafted characteristics, several research has lately begun investigating DL-based alternatives. Convolutional Neural Networks (CNNs) in particular have gained widespread use due to their hierarchical structure, which enables them to automatically learn the appropriate set of features for the problem being analysed. In general, the results are encouraging and perform better than the aforementioned traditional techniques, albeit at the cost of the need for more training samples to converge. To face this, a common approach is to use "transfer learning" as a pre-training technique to exploit the features learned on a different task with a huger number of available training samples. The so extracted high-level features are then fed to other classification approaches (not necessarily deep) or used as a starting point to fine-tune the same deep neural network architecture on the medical data [143]. An example of the former can be found in [6], with authors adopting an AlexNet [71] model pre-trained on ImageNet [30] as a feature extractor to feed a Support Vector Machine (SVM) trained for the binary lesion classification task. For the latter case, in [89] and [43] the authors proposed to directly fine-tune the pre-trained deep network, with [89] relying on AlexNet and [43] on ResNet34 [46].

Similarly, Hu et al. [49] adopted the VGG19 [135] pre-trained on ImageNet [30] to extract features from the DCE-MRI, represented as a 4D volume consisting of both volumetric and temporal information. To handle the 4D data, volumetric information was collapsed into two dimensions by taking the maximum intensity projection at the image level or feature level within the CNN architecture. In the work proposed by Rasti et al. [118], a mixture ensemble of CNNs is used to perform the breast lesion classification considering each network as an expert. However, the cited approaches are mainly based on a textural analysis of the lesioned tissues, totally ignoring the fact that *medical images are more than pictures* [38]. Indeed, as aforementioned, the temporal dimension is one of the DCE-MRI key characteristics that makes it very sensitive to the staging of the malignancy of a tumour. Therefore, in [163] the authors exploited the CA dynamic and the three-dimensional dependency between slices extracted from the same lesion, by proposing two different approaches based on VGG19 [135] and on InceptionV3 [142] fine-tuning respectively. On the same line, in a previous work [40] authors proposed 3TP-CNN, a simple but effective approach to make any Image-Net pre-trained CNN able to properly leverage the contrast agent temporal dynamic. In particular, the work introduced the concept of 3TP slices, namely 3-channels images representing a given slice at three different time points, uniquely identified by means of the Three Time Points (3TP) [29] approach. This allows identifying a fixed set of acquisition time instants (in seconds after the CA injection), making the proposed approach more general and applicable also to DCE-MRI protocols considering a different number of acquired series.

8.2 Population

A women breast DCE-MRI dataset provided by "Istituto Nazionale Tumori, Fondazione G. Pascale" of Naples were used. It consists of 34 patients[1] (ages in 16-69) undergone imaging with a 1.5T scanner (Magnetom Symphony, Siemens Medical System, Erlangen, Germany) equipped with breast coils. DCE FLASH 3D T1-weighted coronal (the axis having the highest resolution) images were acquired (Transversal Relaxation Time: 9.8ms, TE: 4.76ms; Flip Angle: 25°; Field of View 370x185 mm^2; Image: 256x128 pixels; Thickness: 2mm; Gap: 0; Acquisition time: 56s; 80 slices spanning entire breast volume). One series (t_0) was acquired before the intravenous injection of the CA and 9 series (t_1-t_9) after the intravenous injection of 0.1 mmol/kg positive paramagnetic contrast agent (gadolinium-diethylene-triamine penta-acetic acid, Gd-DOTA, Dotarem, Guerbet, Roissy CdG Cedex, France). An experienced radiologist manually segmented all the lesions by using original and *subtractive* I_s image series (where $I_s = t_1 - t_0$). Lesions malignancy has been histopathologically proven.

8.3 Methodology

Despite CNNs have shown surprising performance even for biomedical image processing, the experience gained by developing hand-crafted features should be taken into account to design more effective DL-based solutions. On this line, as aforementioned, the solution proposed in [40] introduced 3TP-CNN to properly take into account the CA flowing in DCE-MRI data when using CNNs. This thesis exploits the methodology presented in [41] to *make a step further towards the design of a domain-aware DL approach* for DCE-MRI data processing. To this aim, it proposes to use a non-supervised Deforming Autoencoder (DAE) [133] to disentangle the signal intensity variation due to contrast agent flowing in lesioned tissues from the signal components associated with texture and deformation while retaining the biological characteristics of the tumor area. Moreover, the DAE is evaluated according to its ability to reconstruct the input image starting with the disentangled components, namely *shading (S), albedo*

---

[1] All patients have agreed to the use of the data for research purposes.

(A), and *deformation field (W)*. This characteristic allows the DAE to be used as a generator of synthetic images which are obtained by modifying the different components and exploited in the classification step for data augmentation. As the DAE expects a 3-channels image while DCE-MRI produces t-channels slices (with t equal to the number of acquired series as described in Section 8.1), this work relies on 3TP slices [40, 111] to make the approach invariant to the considered DCE-MRI acquisition timeline. Finally, the actual lesion classification is performed by a CNN trained on the disentangled slices. The resulting *"DAE-CNN"* architecture is thus able to learn how to extract salient information while removing those not associated with the CA flowing or with other physiological characteristics of the tissues.

It is worth noting that DAE-CNN is not a simple stack of a CNN and DAE, but a GAN-like architecture, where the discriminator is a CNN and the generator is a DAE (Figure 8.2). In particular: the input are 3TP slices (I_{3TP}); the DAE disentangles each 3TP slice in its *shading (S)*, *albedo (A)* and *deformation field (W)* components; finally a CNN is used to classify the disentangled *albedo* (A) component after the application of the deformation fiels W. Next sections detail each of the cited aspects, including 3TP slices and networks (pre)training strategy. Finally, the used procedure for combining slice-level predictions is introduced and motivated.

8.3.1 PBPK Fitting

As already mentioned in Section 7.1, $C_t^{Measured}(t)$ represents the concentration of contrast liquid, derived from the value of the signal acquired through the DCE-MRI. Its computation depends on the following elements:

- the signal intensity $SI(t)$, representing the brightness of a voxel at a time instant t;
- the signal intensity relative enhancement $RE(t)$, which represents the percentage variation of the brightness of the voxel at instant t relative to the observed pre-injection values, therefore: $RE(t) = SI(t) - SI$

Taking up what was reported by Schabel [128], $C_t^{Measured}(t)$ can be computed as follows:

$$C_t^{Measured}(t) = \frac{RE(t)}{T_1 \cdot r_1}\frac{mMol}{l} \tag{8.1}$$

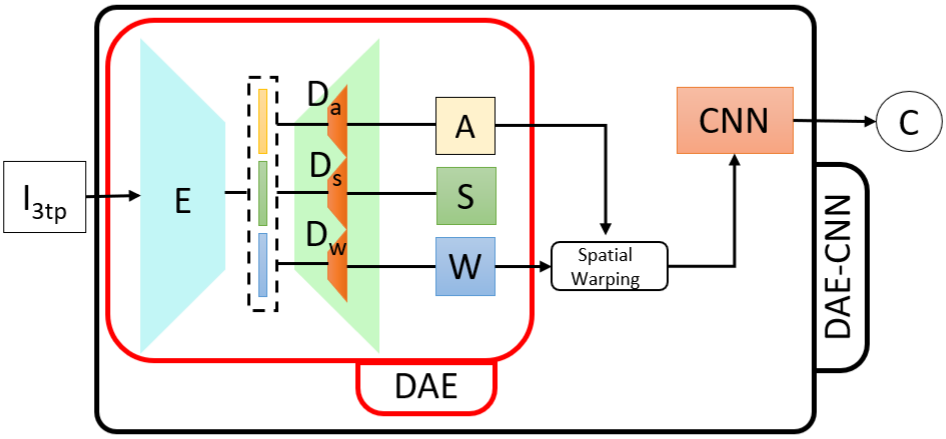


Figure 8.2. Proposed DAE-CNN architecture. The DAE, consisting of an encoder (E) and of three different decoders (D_a, D_s, D_w), disentangles a 3TP slice (I_{3TP}) in its *shading (S)*, *albedo (A)* and *deformation field (W)* components. The *Albedo* (A) component with the application of the deformation field is used as input for the 3TP-CNN to determine the lesion class C (benign/malignant).

where T_1 is the tissue longitudinal relaxation time before contrast agent injection (expressed in ms) and r_1 is the contrast agent longitudinal relaxation time ($\frac{mMol}{ms}$).
As introduced in Section 8.1, a DCE-MRI is a 4D volume having 3 spatial (x, y, z) and 1 temporal (t) dimensions representing the acquisition of 3D volumes over time. In particular it can be interpreted as a volume of size $X \times Y \times Z \times N$, where $X \times Y \times Z$ is the spatial dimension, while N is the number of temporal acquisitions (10 in this work). The presence of the segmentation mask for each patient makes it possible to select only slices containing lesions along the z axes, resulting in a series of 3D slices of size $X \times Y \times N$. In particular, for each slice of the lesion, the median pixels for the N temporal instants is calculated among the pixel belonging to the tumour area, resulting in a vector on N element that is used to compute the $C_t^{Measured}(t)$ for the ten time instants.
The aim of the fitting process is to find the parameters of the pharmacokinetic models introduced in Equations 7.1 and 7.2, assuming the $C_t^{Measured}(t)$ as a reference. The Tofts-Kermode[146] needs two parame-

ters, namely the K^{trans} and v_e while in the case of the Extended Tofts-Kermode[147] model K^{trans}, v_e and v_p should be found. The method of least squares is a standard approach in regression analysis, commonly used in data fitting. Defined a residual as the difference between an observed value and the fitted value provided by a model, the best fit in the least-squares sense minimizes the sum of squared residual. In this thesis, it is applied to fit the measured contrast agent concentration $C_t^{Measured}(t)$ to the desired tumour tissue model.
At the end of this process, for each slice the $C_t^{fitted}(t)$ that represents a theoretical trend of the signal $C_t^{Measured}(t)$ and the set of parameters are computed.

8.3.2 3TP Slices Extraction

During the DCE-MRI, the number of volumes acquired (N) may strongly vary across different medical centres, making harder the design of a more general approach. To address this problem, a previous [40] work introduced the concept of 3TP slices to standardize the considered number of series. The idea is to fix the number of temporal acquisitions by taking into account the 3TP method [29] according to which the lesion classification can be faced by only taking DCE images at three-time points identified by the time (in seconds) passed after the contrast agent (CA) injection: pre-contrast (t_0); 2 minutes after the CA injection (t_1, corresponding to the pick of CA levels in tissues); 6 minutes after the CA injection (t_2, corresponding to the end of the CA washout). Then, 3TP slices are obtained by extracting slices over the projection with the higher spatial resolution (i.e. $[x, y]$, $[x, z]$ or $[y, z]$) from the three volumes acquired at the time instances closest to t_0, t_1 and t_2 (Figure 8.3). In other words, each 3D slice of size $X \times Y \times N$ is transformed in an image of size $X \times Y \times 3$. This process generates a set of multi-channel images (i.e. the 3TP slices), each representing the same portion of tissue seen at different temporal instants. It is worth noting that this work only considers slices containing a lesion. This is possible since lesion classification is a stage executed after the lesion detection/segmentation [110] that can be made both manually or by an automatic procedure [111]. Finally, the 3TP slices are further processed by extracting only the portion of the image within a squared box centred in the lesion centre and having size 1.5x the lesion diameter

[40]. The resulting images are then normalized in $[0, 1]$ to ensure that, in the next stage, the DAE operates on images having the same scale across different lesions.

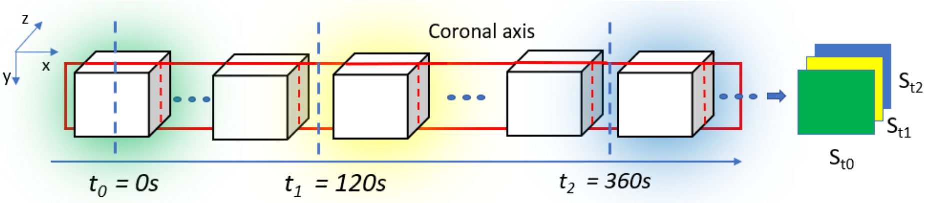


Figure 8.3. 3TP slice extraction procedure. The 4D volume $V_{x,y,z,t}$ is sectioned along the highest resolution axis (coronal in the example) generating N different t-channels slices, with t equal to the number of temporal acquisitions and N equal to the size of the cutting axis (z in the example). From these t-channels slices, the corresponding 3TP slices (I_{3TP}) are 3-channel images obtained by considering only the channels associated with the three volumes acquired at the time instances closest to those suggested in [29].

8.3.3 3TP Slice Disentangling

3TP slices were designed to preserve all the information needed to describe the CA course within tissues by relying on a fixed number of time series. Despite this, they still carry other information that may “daze” the CNN during the classification (possibly resulting in a higher number of labelled training samples to converge). In this step, *the aim is to enhance the useful information while weakening the impact of the others*, exploiting the fact the generation of an image consists of two processes: the synthesis of a texture (how the image appears) on a template (usually a deformation-free coordinate system), and the deformation to introduce shape variability. Texture can be further disentangled in *albedo* and *shading*, where albedo is the overall brightness of an object (i.e. the reflecting power of a material) while shading represents depth information. In [133], the authors introduced the "intrinsic Deforming-Autoencoder (DAE)", an unsupervised encoder-decoder architecture designed to disentangle an image in its main components. The network consists of a single encoder (E) and of three independent decoders (D_a, D_s, D_w) for the synthesis of albedo, shading and deformation functions respectively (as in Figure 8.2).

It is worth to note that the DAE loss (E_{tot}) proposed in [133] consists of the sum of three parts[2]: the reconstruction loss (E_{rec}), the warping loss (E_{warp}) and the adversarial loss (E_{adv}). The first is the standard l_2 loss, used to penalize reconstruction errors. The second is used to both penalize quickly-changing deformations encoded by local warping field and to remove any bias introduced by the fitting process. The third is used to generate visually realistic images. In this thesis, the DAE training is modified by adding to the loss function the component related to the pharmacokinetic model (E_{PBPK}), exploiting the fact that "*medical images are more than pictures*". The idea consists in making the disentangling process dependent on the PBPK compartmental model, avoiding the loss of information related to the contrast agent flowing. This thesis *leverages the DAE to disentangle 3TP slices, with the aim of compacting into a single component all the information associated with the CA flowing*. To do so, the encoder (E) takes a 3TP slice (I_{3TP}) as input and delivers a low-dimensional latent representation Z, split in three sub-components $Z = [Z_s, Z_a, Z_w]$. Each of these parts is fed to a different decoder (D_s, D_a, D_w), providing the decomposition of the input 3TP slice in its *shading (S)*, *albedo (A)* and *deformation field (W)* components (Figure 8.4). The texture element (T) is then computed as the Hadamard product between the albedo and the shading. The output of the network is a reconstructed image I_{rec} and the DAE is trained to be able to reconstruct the input considering the three features components $Z = [Z_s, Z_a, Z_w]$. To avoid the elimination of PBPK features, the E_{PBPK} is introduced, acting on each I_{rec} as follows:

$$E_{PBPK}(I_{3TP}, I_{rec}, p_{I_{3TP}}) = \frac{1}{3}\sum_{i=0}^{2}(C^{fitted}_{t_{I_{3TP}}}(t_i) - C^{Measured}_{t_{I_{rec}}}(t_i))^2 \qquad (8.2)$$

where $C^{fitted}_{t_{I_{3TP}}}$ is the $C^{fitted}_t(t)$ computed on the input I_{3TP} by applying to Equations 7.1 or 7.2 the parameters $p_{I_{3TP}}$ found during the PBPK Fitting step described in Section 8.3.1, while $C^{Measured}_{t_{I_{rec}}}$ is the $C^{Measured}_t(t)$ on the I_{rec} image obtained by using Equation 8.1 and t_i with i form 0 to 2 are the time points selected during the 3TP Slice Extraction step (Section 8.3.2). The aim of the E_{PBPK} is to make the DAE aware of the biological char-

---

[2] Please refer to [133] for a more detailed description.

acteristics of the images.
The shading component is discarded, using only the albedo with the application of the deformation field as input for the classification step to help the model to focus only on the CA course. This choice is due to the strong correlation between the brightness and the signal intensity variations associated with the CA flowing. Moreover, the W is retained since the information about the shape can be useful for the distinction between a malignant or benign lesion.

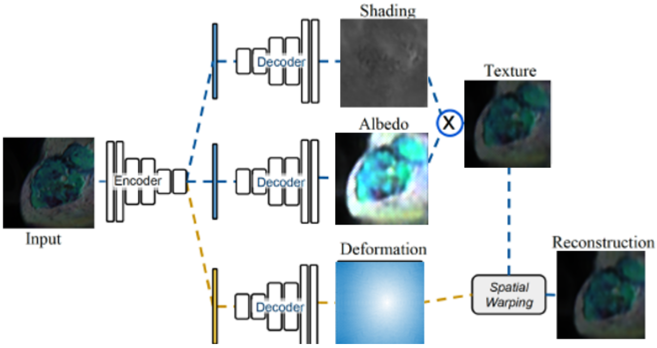


Figure 8.4. Intrinsic Deforming-Autoencoder (DAE) architecture used for 3TP slice (input) disentangling. The encoder takes a 3TP slice (I_{3TP}) as input and delivers a low-dimensional latent representation, split into three sub-components. Each of these parts is fed to a different decoder. In the middle, the disentangled three components and the resulting texture; on the right, the reconstructed 3TP slice, used as a reference to train the DAE.

8.3.4 DAE for data augmentation

As aforementioned in Section 8.3.3, the generation of an object represented by an image consists of the synthesis of a texture on a template and the application of a deformation field. In other words, an image can be obtained by exploiting the A, S, and W components determined by the decoders D_a, D_s, D_w considering the vector Z, split into three subcomponents $Z = [Z_s, Z_a, Z_w]$. Denoting with $Z_{a_{p_i}}$, $Z_{s_{p_i}}$, $Z_{w_{p_i}}$, the components of the i-th slice of the generic patient p, the $I_{rec_{p_i}}$, that is her reconstructed image can be formalized as follows:

$$I_{rec_{p_i}} = f_g(Z_{a_{p_i}}, Z_{s_{p_i}}, Z_{w_{p_i}}) \tag{8.3}$$

where f_g exploits the DAE decoders D_a, D_s, D_w.
In this thesis, an *innovative image generation process* for data augmentation is implemented by considering the main elements of an object together with the biological characteristics. In particular, data variability is introduced by changing the $Z_{w_{p_i}}$ component in Equation 8.3 with the aim of creating a lesion with a different shape.
The DAE trained in the previous step gives a distribution of Z representation, whose Z_w sub-component is responsible for the W element. The *image generation process* exploits the Z_w vector to apply the deformation that introduces the shape variability. Each *synthetic image* takes into account the components coming from slices of two different patients p and q, with $p \neq q$ belonging to the same class. It is worth noting that benign and malignant lesions present different biological characteristics resulting in the need of preserving class separation. In particular, denoting with $I_{3TP_{p_i}}$ the i-th 3TP slice of the generic patient p, from which the $Z_{a_{p_i}}$, $Z_{s_{p_i}}$ components are extracted discarding the $Z_{w_{p_i}}$ element, with $I_{3TP_{q_j}}$ the j-th 3TP slice of the patient q ($p \neq q$), from which the $Z_{w_{q_j}}$ vector is considered, the *synthetic image* of the class c, $I^c_{s_{p_i q_j}}$ is formalized as follows:

$$I^c_{s_{p_i q_j}} = f_g(Z_{a_{p_i}}, Z_{s_{p_i}}, Z_{w_{q_j}} | c) \tag{8.4}$$

where $c \in \{benign; malignant\}$. It is worth noting that from each $I_{3TP_{p_i}}$ different *synthetic images* can be generated by variyng the patient q and the slice j.
As described in Section 8.3.2, each patient consists of a set of slices whose number varies in respect of the extension of the lesion along the z-axis. Denoting with Be and Ma the total number of patients belonging to the benign and malignant classes respectively, with n_p the number of slices of the patient p, and with p and q two generic patients, the cardinality of the

set of synthetic images T_s is computed as follows:

$$T_s = \sum_{p=1}^{Be}(n_p \cdot \sum_{q=1;q\neq p}^{Be} n_q) + \sum_{p=1}^{Ma}(n_p \cdot \sum_{q=1;q\neq p}^{Ma} n_q) \tag{8.5}$$

The E_{PBPK} loss introduced in Section 8.3.3 is used to control the synthetic image generation process, making the process aware of the biological characteristics of the images to be generated. In particular, denoting with $I^c_{p_i}$ the i-th 3TP slice of the patient p belonging to the class c from which the components $Z_{a_{p_i}}, Z_{s_{p_i}}$ are extracted, with $I^c_{s_{p_i q_j}}$ the *synthetic image* of the class c, obtained by applying to $Z_{a_{p_i}}, Z_{s_{p_i}}$ the $Z_{w_{q_j}}$ extracted from the j-th 3TP slice of the patient q as described in Equation 8.4, the E_{PBPK} loss in Equation 8.2 is adjusted as follows:

$$E_{PBPK}(I^c_{p_i}, I^c_{s_{p_i q_j}}, p_{I^c_{p_i}}) = \frac{1}{3}\sum_{i=0}^{2}(C^{fitted}_{t_{I^c_{p_i}}}(t_i) - C^{Measured}_{t_{I^c_{s_{p_i q_j}}}}(t_i))^2 \tag{8.6}$$

where $C^{fitted}_{t_{I^c_{p_i}}}$ is the $C^{fitted}_t(t)$ computed on the input $I^c_{p_i}$ by applying to Equations 7.1 or 7.2 the parameters $p_{I^c_{p_i}}$ found during the PBPK Fitting step described in Section 8.3.1, while $C^{Measured}_{t_{I^c_{s_{p_i q_j}}}}$ is the $C^{Measured}_t(t)$ on the $I^c_{s_{p_i q_j}}$ image obtained by using Equation 8.1. As reported in Section 8.3.1, the $C^{Measured}_t(t)$ is calculated by considering the median pixels of the tumour area in each slice. As a consequence, the synthetic lesion mask for each slice $I^c_{s_{p_i q_j}}$ is computed by applying to the real mask of $I^c_{p_i}$ the deformation field resulting from $Z_{w_{q_j}}$.

8.3.5 DAE-CNN Training

As aforementioned, DAE-CNN is a GAN-like architecture, where the discriminator and the generator are respectively a CNN and a DAE (Figure 8.2). In particular, the DAE and the CNN are trained simultaneously in order to *make the CNN learn how to classify the disentangled albedo (A) slices while influencing the DAE into producing disentangled slices more suited for the lesion classification task.* Thus, the work proposed in [41] is extended by introducing the DAE-CNN training that is implemented by splitting each iteration into four main steps:

1. Albedo 3TP slices (from here on 3TP-A) are computed by feeding the DAE with 3TP slices. The resulting 3TP-A images are then classified by using the CNN, whose classification loss is used to update both DAE and CNN weights with a learning rate α_1;
2. The same 3TP images used as input in the previous step are used again to generate the 3TP-A slices, but this time considering the DAE weights updated in the previous step. The DAE loss (see Section 8.3.3), including the E_{PBPK}, is then re-estimated and used to re-updated only the DAE weights with a learning rate $\alpha_2 \ll \alpha_1$.
3. For each class c, a set of synthetic 3TP slices is generated by changing the Z_w sub-component as described in Section 8.3.4. As reported in step 1, the resulting synthetic-3TP-A images are then classified by using the CNN, whose classification loss is used to update the weights of the three decoders D_a, D_s, D_w and the CNN with a learning rate α_1;
4. The same synthetic 3TP slices are used again to generate the synthetic-3TP-A slices, but this time considering the weights of the decoders updated in the previous step. The DAE loss (see section 8.3.3), including only the E_{PBPK}, the E_{warp}, and E_{adv}, is then re-estimated and used to re-updated only the weights of the three decoders D_a, D_s, D_w with the learning rate α_2.

The CNN is trained by minimising the cross-entropy loss

$$L(y, \hat{y}) = \sum_i y_i log(\hat{y}_i)$$

where $\hat{y}_i$ is model prediction, while y is the target class. It is worth noting that only the decoders are updated in steps 3 and 4 since they are the networks responsible for image generation. Moreover, the absence of a reference image makes it impossible to use the E_{rec} component in the loss function when step 4 is performed. The training iteration is summarised in Figure 8.5. In each step the updated parts are surrounded by a shadow that is orange if the cross-entropy loss is used or blue when the DAE loss

is exploited.
This four-steps training helps the DAE learn how to generate more valuable (from the CNN point of view) 3TP-A slices while preventing it from overfitting "under" the CNN loss. In particular, the steps 1 and 2 focus on *real* images, that are directly obtained from Section 8.3.2, while steps 3 and 4 exploits the process described in Section 8.3.4 for the generation of *synthetic* slices.
According to Equation 8.5, the number of synthetic images that the DAE is able to generate is much larger than the size of the dataset used in this work. As a consequence, in each iteration, the DAE-CNN is trained considering the same amount of *real* and *synthetic* slices obtained by randomly selecting from each class the 3TP images and Z_w sub-component to be used in steps 3 and 4. Moreover, in the implemented training process, two different iterations consider non-overlapping sub-sets of both *real* and *synthetic* images. A training epoch ends when all *real* images have been processed. However, since the number of *benign synthetic* images differs from the total number of *malignant synthetic* slices, it is necessary to introduce two additional concepts of *epoch*. In particular, the *malignant epoch* ends when all the *malignant synthetic* slices have been processed, while after a *benign epoch* the DAE-CNN has been fed by all *benign synthetic* images. The different cardinality of the set of images implies that the *epoch*, the *malignant epoch*, and *benign epoch* may not coincide.
The final aspect to consider is the characteristic of the training dataset. Indeed, medical imaging datasets are usually small and very unbalanced. The former is mostly the results of the reduced number of patients usually involved in DCE-MRI programs, while the latter is because of the different number of slices per lesion type (benign/malignant). As a consequence, in the course training, random rotations and flipping operations are applied both on real images (steps 1 and 2) and during the generation of synthetic images (steps 3 and 4), while only the real dataset is balanced by replicating some randomly chosen slices belonging to the minority class.
Although this might sound like a limitation, it has been proved [143] that for medical images leveraging transfer learning "outperforms or, in the worst case, performs as well as training a CNN from scratch". Therefore, the CNN is pre-trained on a huge dataset (i.e. ImageNet [30]) before starting the proposed training algorithm. *It is worth noting i) that this thesis*

does not fix any CNN since the proposed approach applies to any 3-channels input CNN, and ii) there is nothing in the approach preventing a training from scratch (if the dataset size allows it). Unfortunately, this pre-training strategy is not very effective on the DAE, since the reconstruction loss is very domain sensitive. Therefore, the DAE directly is pre-trained on the considered medical data, as reported in Section 8.3.3. This and the fact that the DAE behaves like the generator help enforce the absence of bias or performance over-estimation despite this full-dataset pre-training.

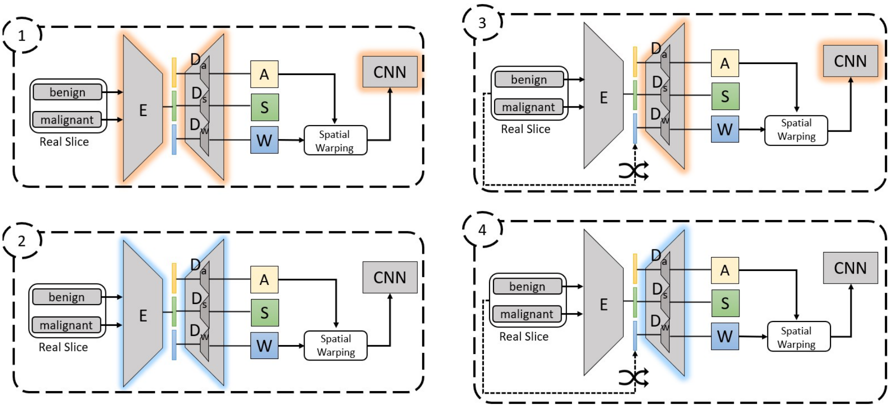


Figure 8.5. DAE-CNN training is implemented by splitting each iteration into four main steps: the updated parts are surrounded by a shadow that is orange if the cross-entropy loss is used or blue when the DAE loss is exploited.

8.3.6 Lesion Classification

The DAE-CNN architecture works on a slice base, producing the probability for it to contain a malignant or a benign lesion. However, since the final aim is to classify each lesion as a whole, as a final step the classes of all the slices from a given lesion are combined into a single class. This thesis introduces a combining strategy that takes into account both the probability produced by the model and the slice size in terms of the number of pixels belonging to the lesioned tissue. More in details, denoting with dim_i^l the size of the i_{th} slice of the lesion l, with n^l the total number of the slices of the lesion l, with m_i^l and b_i^l the predicted probability of containing a malignant or a benign tissue respectively, the probabilities for the lesion l to be a benign (b^l) or malignant (m^l) lesion are computed as

$$m^l = \frac{\sum_{i=1}^{N^l} m_i^l \cdot dim_i^l}{\sum_{i=1}^{n^l} m_i^l \cdot dim_i^l + \sum_{i=1}^{n^l} b_i^l \cdot dim_i^l} \tag{8.7}$$

$$b^l = \frac{\sum_{i=1}^{n^l} b_i^l \cdot dim_i^l}{\sum_{i=1}^{n^l} m_i^l \cdot dim_i^l + \sum_{i=1}^{n^l} b_i^l \cdot dim_i^l} \tag{8.8}$$

It is worth noting that dim_i^l in equations 8.7 and 8.8 is in the range $]0;1]$ since it is normalized with respect to the size of the largest slice of the lesion l. This slice-level prediction combining strategy allows using the prediction for all the slices belonging to a lesion while taking into account that the model may be not very confident on slices containing only a very little portion of the lesion (e.g. border slices).

8.4 Experimental Setup

In the 3TP Slice Disentangling step (Section 8.3.3) the DAE is trained by adding to the loss introduced in [133] the E_{PBPK} loss resulting in DAE-PBPK configuration, while the version without the pharmacokinetic component is denoted with DAE. In particular, the two compartmental methods, namely the Tofts-Kermode[146] (TK) and the Extended Tofts-Kermode[147] (ETK) models are tested, considering both the Weinmann [155] (W) and Parker[107] (P) arterial input function (AIF) introduced in Section 7.1, resulting in four different configurations: TK_W, TK_P, ETK_W, and ETK_P. In particular, in Weinmann [155] AIF, D is the contrast agent injected quantity, a_i the exponential impulses extent and m_i the decay time-constant. Standard used values in the present study will be [155]: $D = 0.1\frac{mmole}{kg}$, $a_1 = 3.99\frac{kg}{l}$, $a_2 = 4.78\frac{kg}{l}$, $m_1 = 0.144\frac{1}{min}$, $m_2 = 0.0111\frac{1}{min}$. In Parker AIF, A_i is the scaling constant, σ_i and T_i the extent and the centers of two gaussians, α and β the extent and the decay time-constant of the exponential, s and τ the extent and the center of the sigmoid. The parameters are obtained through a population-averaged fitting operation [107]: $A_1 = 0.809 mmole \cdot min$, $A_2 = 0.330 mmole \cdot min$,

$T_1 = 0.17046min$, $T_2 = 0.365min$, $\sigma_1 = 0.0563min$, $\sigma_2 = 0.132min$, $\alpha = 1.050mmole$.

As described in section 8.3.5, the proposed architecture consists of two networks jointly trained by using a four stages training schema that exploits the DAE both for image disentangling and synthetic slice generation. Although any three-channels CNN can be used in the designed architecture, this work reports experiments made by using AlexNet [71], ResNet34 [46] and VGG19 [135]. All the considered CNNs were pre-trained on ImageNet [30] and then fine-tuned on the considered dataset.

To assess the effectiveness of the proposed training strategy, some variants are explored by removing one step at a time from the methodology proposed in Section 8.3.5.

The maximum number of epochs was set to 300, the patience to 15. The batch size was set to 32 for AlexNet and to 16 for ResNet34 and VGG19. The learning rate for the cross-entropy loss was set to 10^{-6}, while the one for the DAE loss to $2 \cdot 10^{-8}$. Adam [66] optimizer is used with a weight decay set to 10^{-4}. Since the DAE generates 64×64 images while the tested CNNs expect images of 224×224 pixels, a resizing stage is used before feeding the 3TP albedo slices to the CNN. The DAE has been pretrained using the Adam optimizer, without any weight decay strategy, by setting the maximum number of epochs to 2000, the patience to 50, the batch size to 40 and the learning rate to $2 \cdot 10^{-4}$. Performance is evaluated by measuring Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), F1-Score (F1) and Area under ROC curve (AUC). All the experiments were run in a 10-fold cross-validation fashion to better assess the approach generalization ability. More specifically, it is of crucial importance to execute patient-based cross-validation to reliably compare the performance of different models, avoiding the use of slices from the same patient both during the training and the evaluation phase. For each repetition, 8 folds were used as the training set, 1 as the validation set and 1 as the test set.

To assess the effectiveness of the proposed approach, this thesis compares it against some literature proposals, considering two classical (non-deep) and three DL-based solutions:

- Fusco et al.[37], proposing to leverage Dynamic and Morphological features together (to consider both CA variations and lesion shape) using a Multi-Classifier System;

- Piantadosi et al.[109], relying on Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) descriptor as textural features together with a Random Forest classifier;

- Antropova et al. [5], using AlexNet (pre-trained on ImageNet) as a feature extractor for a Support Vector Machine (SVM) classifier. To match the 3-channels input layer, the authors replicate slices extracted from the second post-contrast series. Since the authors did not provide any information about the chosen SVM hyperparameters settings, an optimization stage is performed to find the best set for parameters. To do so, a grid search is implemented by varying the value of cost value in [0,1] with a step of 0.1, and the degree of the polynomial kernel in [1,3] with a step of 1. The search reported a cost value equal to 1 and degree equal to 1 as the best set of values;

- Haarburger et al.[43], fine-tuning a ResNet34 [46] CNN. To match the 3-channels input layer, the authors perform a grid-search among all the possible time-series combinations, resulting in the selection of $[T_1, T_7, T_8]$;

- Zhou et al. [163], proposing two lesion-level approaches based on VGG19[135] and InceptionV3 [142]. The authors added a max-pool and two fully connected layers at the top of the two networks. To match the 3-channels input layers, they consider the pre-contrast image (I^{pre}), the peak-enhancement image (I^{peak}), the initial uptake image (I^{early}) and the delayed response image (I^{delay}): first channel is I^{peak}, the second is $I^{peak} - I^{pre}$ and the third channel is $I^{early} - I^{delay}$. In this work, the VGG-19 version is considered as the authors proved to be the best.

- Gravina et al. [40]. proposing the 3TP method to select the three-time points to generate the 3TP slices. The authors exploited the fine-tuning of AlexNet.

All the experiments were carried out using Pytorch (version 1.10), while 3TP slices extraction and non-deep competitors have been implemented in MATLAB 2020b. A Linux workstation equipped with Intel(R) Core(TM) i7-10700KF CPU, 64 GB of DDR4 RAM and a Nvidia RTX 3090 GPU is used.

8.5 Results

This section reports the results of the implemented experiments with the aim of evaluating the proposed approach for the task of breast lesion classification exploiting synthetic image generation for data augmentation. Table 8.1 reports the results of the PBPK Fitting and 3TP Slice Disentangling steps showing the mean, the median, the standard deviation (std), the minimum and maximum values of the E_{PBPK} loss computed considering the set of images and PBPK model detailed in columns *Image* and *PBPK Model* respectively. For readability, beside each value the notation $\cdot 10^{-3}$ is omitted. The table consists of three main sections and for each of them the smallest values are reported in bold. In the first part, when the set of images consists of I_{3TP} slices, the E_{PBPK} represents the fitting error, that is the difference between the $C_t^{measured}$ and C_t^{fitted} obtained during the PBPK Fitting step. In other words, the E_{PBPK} is computed as $E_{PBPK}(I_{3TP}, I_{3TP}, p_{I_{3TP}})$. The second section reports the values of the E_{PBPK} computed exploiting the DAE trained without the pharmacokinetic component in the loss function. The aim is to evaluate the error on the images reconstructed by the network (I_{rec}), considering the I_{3TP} slices as reference for the C_t^{fitted}. In particular, the E_{PBPK} value is computes as $E_{PBPK}(I_{3TP}, I_{rec}, p_{I_{3TP}})$ (Equation 8.2). The last section shows the value of the pharmacokinetic component exploiting the DAE trained with the loss function proposed in this thesis. It is possible to note that the values are reduced when the E_{PBPK} component is included in the training phase, improving the fitting process. Moreover, in each section, the experiments with the TK_w PBPK model show the lowest mean and median values. As a consequence, TK_w model is selected to perform further analysis. Figures 8.6, 8.7, 8.8, 8.9 illustrate the box plots representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with DAE and DAE-PBPK.

Table 8.2 shows the results of the experiments implemented for breast lesion classification obtained by changing the training iterations to assess the effectiveness of the strategy introduced in Section 8.3.5 and the E_{PBPK}

Image	PBPK Model	mean	median	std	min	max
I_{3TP}	ETK_P	5.012	2.260	8.128	0.018	67.762
	ETK_W	4.970	2.608	7.451	0.031	59.140
	TK_P	5.322	2.118	10.282	0.014	96.284
	TK_W	**4.073**	**1.438**	**7.291**	**0.002**	**59.139**
$I_{rec}DAE$	ETK_P	5.334	2.269	8.564	0.017	65.267
	ETK_W	5.230	2.586	7.796	0.062	64.353
	TK_P	5.668	2.259	10.678	**0.001**	103.182
	TK_W	**4.409**	**1.618**	**7.618**	0.008	**64.352**
$I_{rec}DAE - PBPK$	ETK_P	1.025	0.482	1.909	0.015	19.740
	ETK_W	1.005	0.394	1.470	0.001	10.086
	TK_P	0.144	0.073	**0.201**	**0.000**	**1.391**
	TK_W	**0.107**	**0.028**	0.745	**0.000**	14.910

Table 8.1. The mean, the median, the standard deviation (std), the minimum and maximum value of the E_{PBPK} loss computed considering the set of images and PBPK model detailed in columns *Image* and *PBPK Model* respectively. For readability, beside each value the notation $\cdot 10^{-3}$ is omitted.

loss described in Section 8.3.3. In particular, the columns *Net* and E_{PBPK} detail the involved CNN and the inclusion of the PBPK component in the loss function respectively, while the steps of each training iteration are defined in the column *Steps*. In particular, it is possible to note four different configurations. The *1-2-3-4*, *1-2-3*, and *1-2* ones represent the method proposed in Section 8.3.5, implementing the steps specified by the numbers. In the *DAE+CNN* configuration the two networks are not jointly trained: the DAE obtained from 3TP Slice Disentangling step is used while the CNN, pre-trained on ImageNet [30], is then fine-tuned on the considered dataset. Table 8.2 consists of three sections; for each of them, the best values are reported in bold. The first one shows the results of the experiments obtained by using the AlexNet [71] gaining values equal to 93.94% and 97.69% in terms of ACC and AUC respectively when all the four steps and the E_{PBPK} component is used. The second part exploits the ResNet34[46] that achieves 87.88% and 92.69% in ACC and AUC respectively when the proposed training strategy is implemented. The experiments reported in the third section exploit the VGG19[135] gaining values equal to 87.88% and 91.15% in terms of ACC and AUC when the proposed approach is involved. It is possible to note that for each CNN the solution including steps 1-2-3-4 and the E_{PBPK} component has the best performance. More-

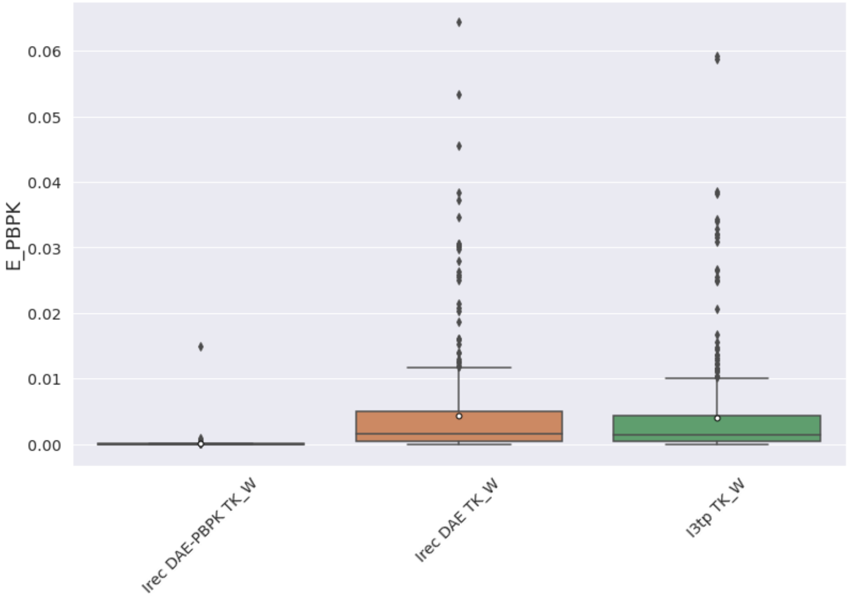


Figure 8.6. Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the TK model with W arterial input function (TK_W)

over, the configuration with the steps 1-2 represents the work proposed in [41].
Table 8.2 compare the implemented methodology with some literature proposals and clearly shows that the proposed approach outperforms all the competitors by a large margin, with the variant based on AlexNet resulting to be the most effective.

Figures 8.10 and 8.11 provide a qualitative assessment of the images generated by the DAE and DAE-PBPK for data augmentation. The image on the left represents the real I_{3TP} slice from which the albedo and shading components are extracted. The images on the right are the synthetic slices obtained from DAE and DAE-PBPK after steps 3 and 4. In both cases, it is possible to note that when the E_{PBPK} component is not considered, the synthetic images do not differ significantly from the one used as a reference. On the other hand, the presence of the biological component enables the network to create new lesions with a different shape but pharmacokinetic

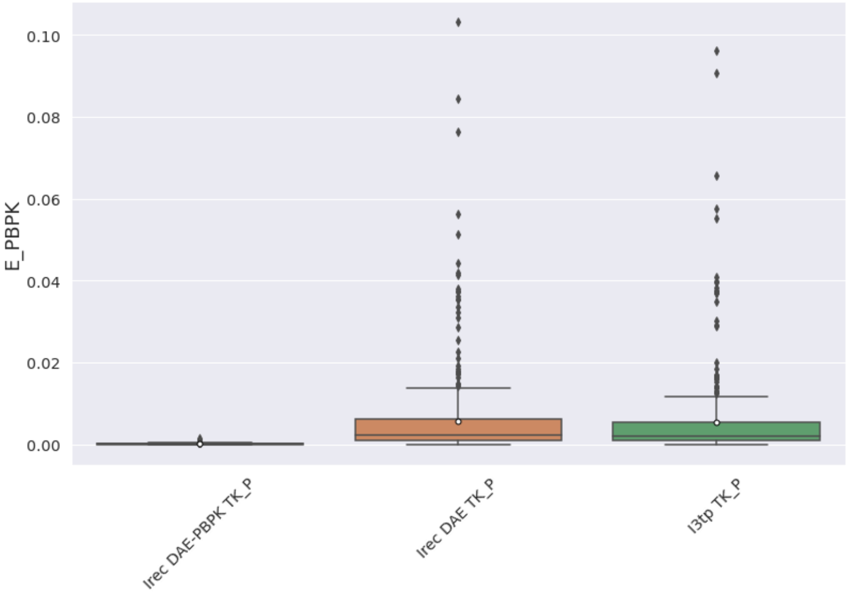


Figure 8.7. Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the TK model with P arterial input function (TK_P)

characteristics similar to those of the baseline image. The aim of Table 8.4 is to provide a quantitative assessment of the I_{rec} generated by the DAE before and after the DAE-CNN training step. In particular the column *Gen.* reports the involved generator, namely the DAE and DAE-PBPK, the *Net* gives information about the network used for classification, while *Step* details the number of steps performed during each training iteration. If the network is not specified the generator obtained after the 3TP Slice Disentangling step is used for the evaluation. For readability, AlexNet, ResNet and VGG19 are denoted with Al, R and V and beside each value the notation $\cdot 10^{-3}$ is omitted. Table 8.4 consists of two parts. The first one focuses on the real images, while the second part on the synthetic slices. In the case of real images, the pharmacokinetic component is computed as $E_{PBPK}(I_{3TP}, I_{rec}, p_{I_{3TP}})$ as proposed in Equation 8.2, exploiting the fact that each I_{rec} corresponds to I_{3TP}. In other words, the aim is to evaluate the biological characteristics of the images reconstructed by the DAE with

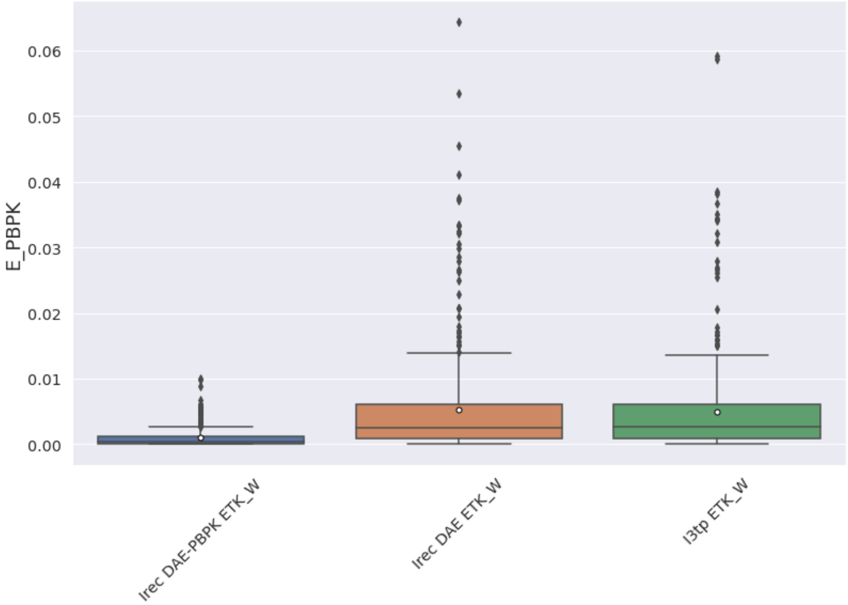


Figure 8.8. Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the ETK model with W arterial input function (ETK_W)

reference to real slices. In the case of synthetic images, generated according to the methodology proposed in Section 8.3.4, the E_{PBPK} component is computed according to Equation 8.6. Moreover, Table 8.4 is organized into two main sections; for each of them the lowest and highest values are reported in bold and italics respectively. In the first one, the DAE trained considering the loss function as proposed in [133] is exploited. In the part focusing on the real images, although the configuration involving AlexNet with Steps 1-2-3 shows an increase in values in comparison with the DAE obtained after the 3TP Slice Disentangling step (first row), it is possible to note that the E_{PBPK} loss does not suffer any significant alteration. On the other hand, in the case of synthetic images, the implemented training strategy leads to a reduction in values regardless of the number of steps performed. The second section considers the DAE trained by exploiting the proposed loss function (DAE-PBPK). When the real slices are con-

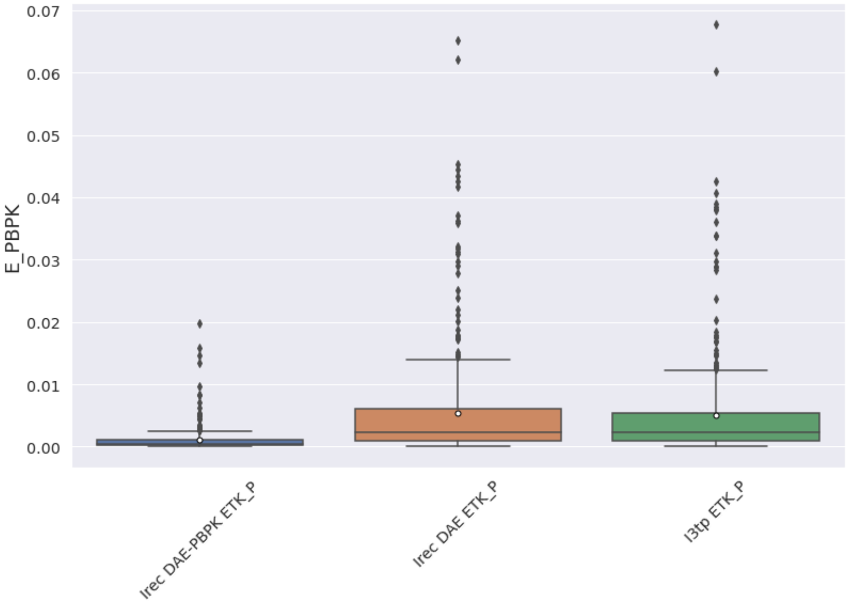


Figure 8.9. Box plot representing the distribution of the E_{PBPK} computed on I_{3TP} images, and the I_{rec} slices obtained with the DAE and the DAE-PBPK that considers the ETK model with P arterial input function (ETK_P)

sidered, the DAE-PBPK obtained after the 3TP Slice Disentangling step (first row) presents the lowest mean and median values. The training strategy proposed in Section 8.3.5 introduces a reduction in the values reported in the *max* column. Moreover, the performance obtained by varying the CNN and the number of steps are comparable. In the case of synthetic images, the values achieved are comparable with those computed considering the configuration without CNN (first row). In conclusion, when the two sections of Table 8.4 are compared, it is possible to note that the presence of the PBPK component always leads to better performance.

8.6 Discussions

The aim of this thesis was to propose and assess a domain-aware architecture for the automatic breast lesion malignancy classification in DCE-

Net	E_{PBPK}	Steps	ACC	SPE	SENS	F1	AUC
AlexNet[71]	Yes	1-2-3-4	**93.94%**	**92.31%**	**95.00%**	**95.00%**	**97.69%**
		1-2-3	90.91%	84.62%	**95.00%**	92.68%	94.23%
		1-2	87.88%	84.62%	90.00%	90.00%	91.15%
		DAE+CNN	78.79%	76.92%	80.00%	82.05%	83.46%
	No	1-2-3-4	81.82%	69.23%	90.00%	85.71%	81.15%
		1-2-3	81.82%	69.23%	90.00%	85.71%	78.85%
		1-2	87.88%	84.62%	90.00%	90.00%	85.00%
		DAE+CNN	75.76%	69.23%	80.00%	80.00%	85.38%
ResNet34[46]	Yes	1-2-3-4	**87.88%**	**84.62%**	90.00%	**90.00%**	**92.69%**
		1-2-3	81.82%	76.92%	85.00%	85.00%	88.46%
		1-2	78.79%	69.23%	85.00%	82.93%	81.15%
		DAE+CNN	72.73%	53.85%	85.00%	79.07%	73.46%
	No	1-2-3-4	69.70%	46.15%	85.00%	77.27%	72.31%
		1-2-3	72.73%	46.15%	90.00%	80.00%	80.77%
		1-2	81.82%	61.54%	**95.00%**	86.36%	77.69%
		DAE+CNN	72.73%	53.85%	85.00%	79.07%	70.77%
VGG19[135]	Yes	1-2-3-4	**87.88%**	**84.62%**	**90.00%**	**90.00%**	**91.15%**
		1-2-3	84.85%	84.62%	85.00%	87.18%	85.38%
		1-2	81.82%	84.62%	80.00%	84.21%	85.38%
		DAE+CNN	69.70%	61.54%	75.00%	75.00%	82.69%
	No	1-2-3-4	75.76%	53.85%	**90.00%**	81.82%	69.62%
		1-2-3	72.73%	46.15%	**90.00%**	80.00%	68.46%
		1-2	84.85%	76.92%	**90.00%**	87.80%	85.77%
		DAE+CNN	69.70%	61.54%	75.00%	75.00%	63.22%

Table 8.2. Performance comparison in 10 fold-CV between the methodology proposed in this thesis and some variants obtained by changing the training iterations.

Approach	CNN	ACC	SPE	SEN	F1	AUC
DAE-CNN (Steps 1-2-3-4 with PBPK)	AlexNet	**93.94%**	**92.31%**	**95.00%**	**95.00%**	**97.69%**
DAE-CNN (Steps 1-2-3-4 with PBPK)	ResNet34	87.88%	84.62%	90.00%	90.00%	92.69%
DAE-CNN (Steps 1-2-3-4 with PBPK)	VGG19	87.88%	84.62%	90.00%	90.00%	91.15%
3TP-CNN [40]	AlexNet	75.76%	61.54%	85.00%	80.95%	83.08%
Haarburger et al. [43]	ResNet34	67.65%	30.77%	90.48%	77.55%	83.15%
Piantadosi et al. [109]	-	76.47%	53.85%	90.48%	82.61%	72.16%
Fusco et al. [37]	-	70.91%	53.33%	77.50%	79.49%	65.42%
Zhou et al. [163]	VGG19	61.76%	76.92%	52.38%	62.86%	64.65%
Antropova et al. [5]	AlexNet	64.10%	38.89%	85.71%	72.00%	62.30%

Table 8.3. Performance comparison in 10 fold-CV between the proposed approach and some state-of-the-art competitors, sorted by descending AUC.

MRI. The idea was to develop a solution joining the radiomic knowledge (i.e. past experience in hand-made feature engineering) and deep learning techniques. In particular, this work introduced *DAE-CNN*, a new GAN

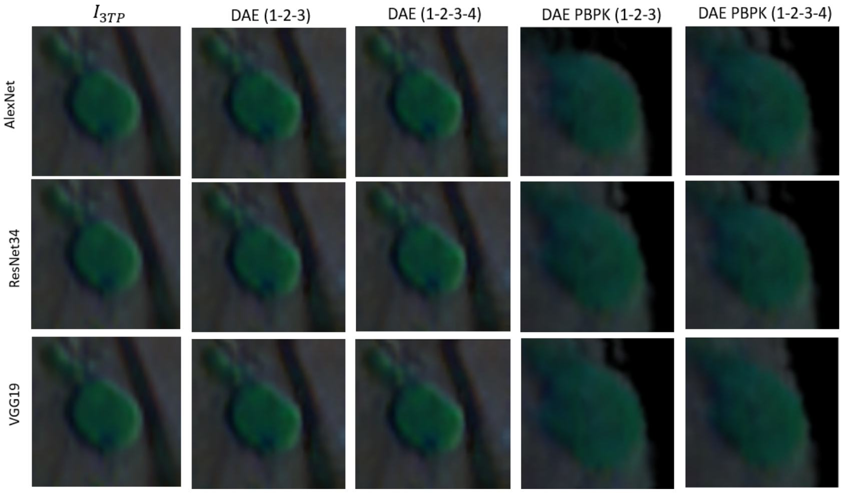


Figure 8.10. Example of synthetic benign lesion generated by the DAE. The image on the left represents the real I_{3TP} slice from which the albedo and shading components are extracted. The images on the right are the synthetic slices obtained from DAE and DAE-PBPK after steps 3 and 4.

like architecture designed to disentangle the contrast agent effects from all the other image components while learning how to perform the lesion classification task. The proposed model is based on an intrinsic Deforming Autoencoders (DAE) and on a CNN, simultaneously trained to adapt both networks to the specific task to solve. Moreover, a new approach of data augmentation step is implemented, exploiting the ability of the DAE to generate new slices.

Table 8.1 compares the fitting error, computed by varying the PBPK model on the I_{3TP} slices (first section), with the E_{PBPK} loss of the I_{rec} images representing the output of the DAE and DAE-PBPK respectively. It is possible to note that when the DAE trained with the loss proposed in [133] is used, the values describing the fitting error are higher than those in the first section since the reconstruction process does not use biological information. On the other hand, the inclusion of the E_{PBPK} component in the third section introduces an improvement of the fitting procedure, resulting in a DAE able to remove from the original I_{3TP} the elements, i.e noise or artifacts, limiting the PBPK model in the description of the

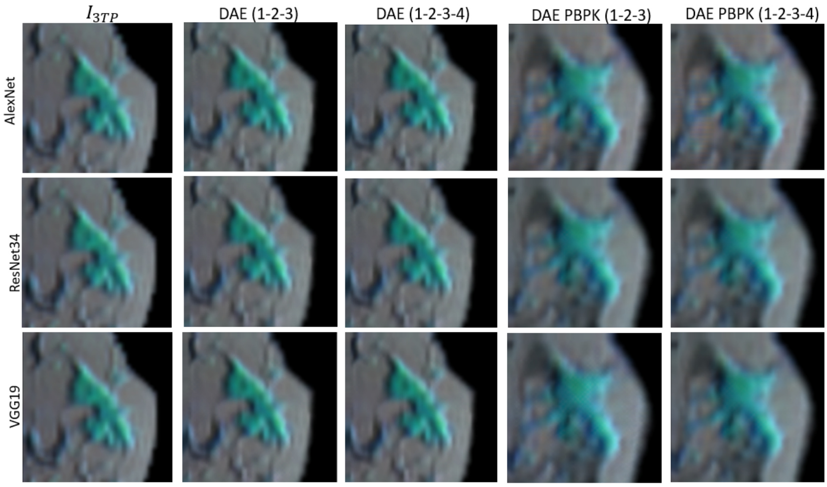


Figure 8.11. Example of synthetic malignant lesion generated by the DAE. The image on the left represents the real I_{3TP} slice from which the albedo and shading components are extracted. The images on the right are the synthetic slices obtained from DAE and DAE-PBPK after steps 3 and 4.

absorption, distribution, metabolism, and excretion (ADME) of contrast agent. In other words, the DAE-PBPK represents a physiologically-aware motion correction method. Indeed, MRI scan protocol provides different series acquisitions that have a duration of several minutes, and therefore it is mostly affected by motion artifacts. Cardiac pulsation, anatomical structures in constant motion (lung, blood vessel walls, eyes), patient breath or involuntary movements are the main accidental disturbances that can occur during scanning. With specific reference to DCE-MRI, the aim of a motion correction technique is to re-align each voxel in the post-contrast images to the corresponding voxel in the pre-contrast (reference) image, trying to minimize an objective function that, in the case of this thesis, is the fitting error.

Table 8.2 shows that for each network the use of the E_{PBPK} loss and the implementation of the training procedure consisting of four steps leads to the best performance. In particular, when the biological characteristics are exploited, steps 3 and 4 represent a physiologically-aware generation process that produces slices able to improve the classification phase.

Gen	**Net**	**Step**	**Real Images**					**Synthetic Images**				
			mean	**median**	**std**	**min**	**max**	**mean**	**median**	**std**	**min**	**max**
DAE	-	-	4.409	1.618	7.618	0.008	64.352	*16.232*	*7.255*	*24.959*	*0.004*	*272.735*
	Al[71]	1-2	3.875	1.596	7.603	0.009	109.812	13.702	5.656	22.081	**0.000**	237.089
		1-2-3	*4.527*	*1.994*	*11.903*	*0.013*	*185.438*	**11.801**	**4.421**	21.604	**0.000**	238.308
		1-2-3-4	4.178	1.903	9.602	0.004	138.913	12.229	4.535	21.249	**0.000**	229.772
	R[46]	1-2	3.875	1.596	7.603	0.009	109.812	13.703	5.656	22.082	**0.000**	237.089
		1-2-3	3.690	1.506	**6.503**	**0.002**	**53.191**	14.603	5.980	22.685	0.001	248.946
		1-2-3-4	3.510	1.374	6.907	0.005	73.183	13.889	5.735	**21.072**	0.001	228.451
	V[135]	1-2	**3.496**	**1.302**	8.006	**0.002**	120.083	13.335	5.389	21.703	0.001	238.574
		1-2-3	3.703	1.367	8.975	0.013	124.587	14.216	6.246	21.945	0.001	226.905
		1-2-3-4	3.583	1.353	7.295	0.007	87.221	13.883	5.795	21.246	0.002	**223.035**
DAE-PBPK	-	-	**0.107**	**0.028**	0.745	**0.000**	*14.910*	9.273	**1.435**	*21.177*	***0.000***	259.176
	Al[71]	1-2	0.400	0.209	0.554	**0.000**	**4.660**	10.300	1.773	21.382	***0.000***	268.180
		1-2-3	0.676	0.352	0.891	**0.000**	8.804	*10.830*	2.173	21.628	***0.000***	268.658
		1-2-3-4	*0.751*	0.369	*0.972*	0.001	7.042	10.792	*2.180*	21.287	***0.000***	*274.395*
	R[46]	1-2	0.433	0.326	0.424	0.001	4.742	9.252	1.542	20.389	***0.000***	252.402
		1-2-3	0.457	0.355	0.425	*0.003*	5.020	9.241	1.511	20.420	***0.000***	253.774
		1-2-3-4	0.440	0.332	0.426	**0.000**	4.908	**8.886**	1.464	**19.906**	***0.000***	**250.037**
	V[135]	1-2	0.437	0.351	**0.397**	0.001	4.750	8.893	1.467	20.157	***0.000***	252.003
		1-2-3	0.648	*0.575*	0.558	0.002	5.459	9.171	1.452	20.865	***0.000***	261.022
		1-2-3-4	0.443	0.360	0.410	0.001	4.739	9.267	1.484	20.734	***0.000***	257.592

Table 8.4. Quantitative assessment of the I_{rec} generated by the DAE before and after the DAE-CNN training step. The column *Gen.* reports the involved generator, namely the DAE and DAE-PBPK, namely the DAE and DAE-PBPK, the *Net* gives information about the network used for classification, while *Step* details the number of steps performed during each training iteration. If the network is not specified the DAE or DAE-PBPK obtained after the 3TP Slice Disentangling step is used for the evaluation. For readability, AlexNet, ResNet, and VGG19 are denoted with Al, R, and V. For readability, beside each value the notation $\cdot 10^{-3}$ is omitted.

Indeed, Figures 8.10 and 8.11 highlight that the use of E_{PBPK} loss makes the network able to introduce shape variability while preserving the physiological characteristics. Moreover, when the pharmacokinetic component is not exploited, the generated images in steps 3 and 4 appear similar to the reference ones, thus explaining the reduction in the performance reported in Table 8.2. Indeed, the implemented data augmentation process provides a small shape variability without considering the biological characteristics that contribute to the distinction between a benign and malignant lesion.

Table 8.4 provides a quantitative assessment of the I_{rec} generated by the DAE before and after the DAE-CNN training step. When the DAE is used (first section), the results on both real and synthetic images do not suffer significant variations, showing that the training steps do not completely destroy biological information. In the case of DAE-PBPK, the results on the synthetic slices report that the E_{PBPK} component helps the network limit the value of the fitting error. However, when real images are considered, it is possible to note an increase in values compared to the experiment without CNN. This characteristic is explained by the fact that when the DAE and CNN are jointly trained during the methodology described in Section 8.3.5, they cooperate to produce an image disentangling that aims to enhance the information useful for the classification while preserving the pharmacokinetic characteristics with the consequence that the E_{PBPK} is not reduced, even if it maintains a low range. Indeed, it is possible to note that the values obtained with the DAE-PBPK represent the smallest ones in the Table 8.4. In other words, when the combined training strategy is used, the purpose is to find the best set of features for the classification task while exploiting pre-processing phase that considers the biological information. On the other hand, the DAE-PBPK trained as described in Section 8.3.3 on real images aims to provide an image decomposition that preserves the pharmacokinetic information and makes it possible to reconstruct the input image from the extracted components without considering the further classification task.

As a final consideration, it is worth noting that the proposed methodology can be generalized to different organs or tissues (i.e liver or brain), considering medical imaging tools whose protocols provide at least three acquisitions. However, the mathematical physiological model should be adapted to the specific organ under examination.

Part IV

Conclusions

More recently, novel imaging modalities are being introduced in medical practices as a result of ongoing technological advancements in image acquisition. *Medical image analysis* or *medical image computing* refers to the process of extracting relevant information or knowledge from medical images with the aim of developing potential non-invasive biomarkers for the detection and characterization of the tissues and the anatomical structures under analysis. The manual investigation of medical images by human experts results in a very tedious and time-consuming task. Moreover, different factors contribute to the complexity of medical imaging processing that depends on the intrinsic characteristics of the data and the specific region of interest. The definition of strategies for medical image computing should take the factors of complexity into account with the aim of proposing methods that are able to capture the variability of the anatomical structures under analysis and operate in presence of noise or artifacts while limiting the possibility of human errors. The large amount of information to consider, and the high complexity of medical images have prompted research into proposing solutions for the implementation of systems supporting physicians in the automatic analysis of radiological acquisitions. Artificial Intelligence (AI), referring to the simulation of human intelligence in machines, has been widely used in healthcare with surprising results. In particular, among all AI techniques, Machine Learning (ML) and Deep Learning (DL) are the most widely used approaches. The former includes the set of algorithms that learn from examples, extracting from them the general concepts, that is the knowledge, while the latter is a subset of ML that involves the use of Deep Neural Networks (DNNs). Medical image computing represents the field experiencing the greatest impact of AI solutions. A key role is played by Convolutional Neural Networks (CNNs), a class of DNNs that consists of several convolutional layers that autonomously learn the set of features that well fits the specific task to solve.

This thesis focused on the investigation of AI approaches, and in particular CNNs, in medical image computing, considering the presence of multiple data sources, namely multimodal and synthetic ones. The *multimodal data sources* (Part II) consider the need for several medical applications of exploiting information coming from multiple *modalities*, to create a rich data representation of the phenomena to be explored. The idea is that heterogeneous images may highlight different characteristics of the area under analysis that are useful for its characterization. This thesis leverages Multimodal Learning for the fusion of information from heterogeneous diagnostic tools, proposing in Chapter 4, a systematic analysis of early, late, and intermediate fusion techniques in medical image processing. Moreover, an innovative Transfer Module (TM) was introduced with the aim of implementing in the intermediate approach the cross-modality calibration of the extracted features. In particular, the presented approaches were applied to two different cases of study, detailed in Chapters 5 and 6 respectively. The former exploited the presence of different sequences acquired during the same Magnetic Resonance Imaging (MRI) exam, namely the Dynamic-Contrast Enhanced (DCE), the T2-weighted (T2), and the Diffusion-Weighted Imaging (DWI), while the latter considered two independent image modalities that are the T1-weighted (T1-w) MRI and the Positron Emission Tomography with Pittsburgh Compound B (C-PiB PET). As a consequence, Chapter 5 presented a scenario in which for each patient all three image modalities are available, offering an example with a *complete dataset* where complementary images obtained from the same exam are considered. In Chapter 6, instead, the independence between the T1-w MRI and C-PiB PET, which were acquired on different days, caused the lack of a modality in some patients, resulting in a dataset with *incomplete acquisition*.

On the other hand, the *synthetic sources of data* (Part III) arise when the AI-based solutions implement techniques for the generation of synthetic images that are integrated with the available set of data (real images) with the aim of improving the generalization ability of the implemented model. This thesis focused on DCE-MRI illustrating in Chapter 7 the physiologically based pharmacokinetic (PBPK) models used to describe the contrast agent kinetic behavior and the basic notions for the generation of synthetic images through a process that aims to preserve the biological characteristics. Furthermore, the presented concepts were further exploited in Chapter 8, dealing with a case of study that considered the breast lesion classification in DCE-MRI. In particular, a nested deep architecture was introduced to disentangle the contrast agent effect from all the other image components with an intrinsic Deforming-Autoencoder (DAE) while learning how to classify breast lesions with a CNN. Moreover, the physiologically-aware synthetic image generation process was exploited to propose an innovative data augmentation approach that introduces shape variability improving the classification task. During the training phase, both the image synthesis and the classification steps were influenced by a specific component in the loss function E_{PBPK} that introduced a biological integrity constraint on the processed images.

In all the applications described in this thesis, the use of multiple sources of data results in the improvement of the generalization ability of the implemented models.

The best solutions proposed in the case of *multimodal data sources* exploit the intermediate data fusion with the presence of the transfer module. Indeed, the shared representation is created by concatenating features coming from the convolutional cores at an intermediate level, thus preserving the distinctiveness of the different image modalities while determining the interaction between them. In the implemented training strategy, the loss is propagated back to each modality-specific convolutional core leading the features extracting processes to create a shared representation that is suitable for the task to solve.

In the case of *synthetic data sources*, the solution exploiting the proposed DL-based data augmentation process, that leverages the latent representation provided by the DAE introducing shape variability, and the multistage training strategy, that considers both real and synthetic images, shows surprising results. Moreover, the use of the E_{PBPK} component makes the generative models aware of the biological and physiological characteristics of the data, improving the classification performance and, at the same time, the quality of the generated images.

The approaches presented in this thesis suggest several developments in future research within the field of medical imaging analysis. In particular, the solutions proposed for multimodal data fusion only focused on classification tasks. As a consequence, future work will explore the applicability of Multimodal Deep Learning in different applications (i.e tumor segmentation). Moreover, the results obtained with the DAE introducing a new loss function based on physiologically based pharmacokinetic (PBPK) models will prompt new research into exploring its ability in image decomposition for the removal of artifacts and noise in the acquisition process (motion correction).

Despite promising, AI-based approaches in healthcare and, in particular, in medical image processing, need to face different challenges. ML techniques require a huge amount of high-quality training data, especially in the case of DL and DNNs. Unfortunately, although data augmentation improves the generalization ability of the models, it may not be enough, particularly if the number of network parameters is very large. Moreover, the same data modality, if acquired with different devices or in different medical centers, may show heterogeneous characteristics which may cause a system implemented for a center to fail to generalize in another context. Indeed, images from a particular hospital can contain noise, artifacts, or various types of bias that make it very difficult to integrate data from several organizations. In many applications, the available quantity of data is limited by the fact that the definition of the ground truth results in a very tedious and hard task for physicians. Furthermore, the intra- and inter-observation variability contribute to increasing the complexity of medical image computing in the validation step. The interpretability of the results is another issue that AI-based solutions require dealing with. Although ML models, especially those based on DNNs, are able to achieve very promising performance, it is not easy to understand the reason behind a specific outcome. In other words, the networks are identified as "black box systems". To overcome this issue, recent approaches based on explainable AI have been implemented focusing on image classification [158]. However, the task model interpretability still remains a challenge for networks trained with data coming from different sources (multimodal learning), especially if one of them is not an imaging modality. Finally, ethical implications should be considered around the use of AI in healthcare since the involvement of intelligent machines raises problems in terms of accountability, permission, and privacy. It is not easy to establish accountability for the implemented systems in case of mistakes. As a consequence, there is the need to provide clear guidance defining the entity that holds the liability. Moreover, ML models require to process huge amounts of data during the training step, which may contain sensitive information about the patients, raising the issue of determining a well-defined privacy policy for data collection and sharing.

Bibliography

[1] Mohammed Abdelaziz, Tianfu Wang, and Ahmed Elazab. Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks. *Journal of Biomedical Informatics*, 121:103863, 2021.

[2] Behnaz Abdollahi, Naofumi Tomita, and Saeed Hassanpour. Data augmentation in training deep learning models for medical image analysis. In *Deep learners and deep learner descriptors for medical applications*, pages 167–180. Springer, 2020.

[3] Sitara Afzal, Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, Khalid M Awan, Irfan Mehmood, and Oh-Young Song. A data augmentation-based framework to handle class imbalance problem for alzheimer's stage detection. *IEEE Access*, 7:115528–115539, 2019.

[4] Ane Alberdi, Asier Aztiria, and Adrian Basarab. On the early diagnosis of alzheimer's disease from multimodal signals: A survey. *Artificial intelligence in medicine*, 71:1–29, 2016.

[5] N Antropova, B Huynh, and M Giger. SU-D-207B-06: Predicting Breast Cancer Malignancy On DCE-MRI Data Using Pre-Trained Convolutional Neural Networks. *Medical physics*, 43(6):3349–3350, jun 2016.

[6] Natalia O Antropova, Benjamin Q. Huynh, and Maryellen L. Giger. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical physics*, 44 10:5162–5171, 2017.

[7] Dooman Arefan, Ruimei Chai, Min Sun, Margarita L Zuley, and Shandong Wu. Machine learning prediction of axillary lymph node metastasis in breast cancer: 2d versus 3d radiomic features. *Medical physics*, 47(12):6334–6342, 2020.

[8] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[9] Michele Avanzo, Joseph Stancanello, and Issam El Naqa. Beyond imaging: the promise of radiomics. *Physica Medica*, 38:122–139, 2017.

[10] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, Alzheimer's Disease Neuroimaging Initiative, et al. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. *NeuroImage: Clinical*, 21:101645, 2019.

[11] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[12] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.

[13] Sebastian Buvari and Kalle Pettersson. A comparison on image, numerical and hybrid based deep learning for computer-aided ad diagnostics, 2020.

[14] Alessandro Calabrese, Domiziana Santucci, Roberta Landi, Bruno Beomonte Zobel, Eliodoro Faiella, and Carlo de Felice. Radiomics mri for lymph node status prediction in breast cancer patients: the state of art. *Journal of Cancer Research and Clinical Oncology*, 147(6):1587–1597, 2021.

[15] Giovanna Castellano, Andrea Esposito, Marco Mirizio, Graziano Montanaro, and Gennaro Vessio. Detection of dementia through 3d convolutional neural networks based on amyloid pet. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE, 2021.

[16] Ronald A Castellino. Computer aided detection (cad): an overview. *Cancer Imaging*, 5(1):17, 2005.

[17] Ruimei Chai, He Ma, Mingjie Xu, Dooman Arefan, Xiaoyu Cui, Yi Liu, Lina Zhang, Shandong Wu, and Ke Xu. Differentiating axillary lymph node metastasis in invasive breast cancer patients: a comparison of radiomic signatures from multiparametric breast mr sequences. *Journal of Magnetic Resonance Imaging*, 50(4):1125–1132, 2019.

[18] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[19] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[20] Hongyoon Choi, Yu Kyeong Kim, Eun Jin Yoon, Jee-Young Lee, and Dong Soo Lee. Cognitive signature of brain fdg pet based on deep learning: domain transfer from alzheimer's disease to parkinson's disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(2):403–412, 2020.

[21] Chiranji Lal Chowdhary and D Prasanna Acharjya. Segmentation and feature extraction in medical imaging: a systematic review. *Procedia Computer Science*, 167:26–36, 2020.

[22] Valentina Cipolla, Domiziana Santucci, Daniele Guerrieri, Francesco Maria Drudi, Maria Letizia Meggiorini, and Carlo de Felice. Correlation between 3 t apparent diffusion coefficient values and grading of invasive breast carcinoma. *European journal of radiology*, 83(12):2144–2150, 2014.

[23] Ermanno Cordelli, Rosa Sicilia, Domiziana Santucci, Carlo de Felice, Carlo Cosimo Quattrocchi, Bruno Beomonte Zobel, Giulio Iannello, and Paolo Soda. Radiomics-based non-invasive lymph node metastases prediction in breast cancer. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 486–491. IEEE, 2020.

[24] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[25] Xiaoyu Cui, Nian Wang, Yue Zhao, Shuo Chen, Songbai Li, Mingjie Xu, and Ruimei Chai. Preoperative prediction of axillary lymph node metastasis in breast cancer using radiomics features of dce-mri. *Scientific reports*, 9(1):1–8, 2019.

[26] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[27] FT De Dombal, DJ Leaper, John R Staniland, AP McCann, and Jane C Horrocks. Computer-aided diagnosis of acute abdominal pain. *Br Med J*, 2(5804):9–13, 1972.

[28] C De Felice, V Cipolla, D Guerrieri, D Santucci, A Musella, LM Porfiri, and ML Meggiorini. Apparent diffusion coefficient on 3.0 tesla magnetic resonance imaging and prognostic factors in breast cancer. *Eur J Gynaecol Oncol*, 35(4):408–414, 2014.

[29] Hadassa Degani, Vadim Gusis, Daphna Weinstein, Scott Fields, and Shalom Strano. Mapping pathophysiological features of breast tumors by mri at high spatial resolution. *Nature Medicine*, 3(7):780–782, 1997.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR, 2009*, pages 248–255. IEEE, 2009.

[31] K Devika and V Ramana Murthy Oruganti. A machine learning approach for diagnosing neurological disorders using longitudinal resting-state fmri. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 494–499. IEEE, 2021.

[32] Yiming Ding, Jae Ho Sohn, Michael G Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W Jenkins, Dmytro Lituiev, Timothy P Copeland, Mariam S Aboian, Carina Mari Aparici, et al. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290(2):456, 2019.

[33] Yuhao Dong, Qianjin Feng, Wei Yang, Zixiao Lu, Chunyan Deng, Lu Zhang, Zhouyang Lian, Jing Liu, Xiaoning Luo, Shufang Pei, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of t2-weighted fat-suppression and diffusion-weighted mri. *European radiology*, 28(2):582–591, 2018.

[34] David Eigen, Jason Rolfe, Rob Fergus, and Yann LeCun. Understanding deep architectures using a recursive convolutional network. *arXiv preprint arXiv:1312.1847*, 2013.

[35] Guy B Faguet. A brief history of cancer: age-old milestones underlying our current knowledge database. *International journal of cancer*, 136(9):2022–2036, 2015.

[36] R Fusco, A Petrillo, M Petrillo, and M Sansone. Use of tracer kinetic models for selection of semi-quantitative features for dce-mri data classification. *Applied Magnetic Resonance*, 44(11):1311–1324, 2013.

[37] Roberta Fusco, Mario Sansone, Antonella Petrillo, and Carlo Sansone. A multiple classifier system for classification of breast lesions using dynamic and morphological features in dce-mri. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 684–692, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[38] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015.

[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[40] Michela Gravina, Stefano Marrone, Gabriele Piantadosi, Mario Sansone, and Carlo Sansone. 3tp-cnn: Radiomics and deep learning for lesions classification in dce-mri. In *ICIAP*, pages 661–671. Springer, 2019.

[41] Michela Gravina, Stefano Marrone, Mario Sansone, and Carlo Sansone. Dae-cnn: exploiting and disentangling contrast agent effects for breast lesions classification in dce-mri. *Pattern Recognition Letters*, 145:67–73, 2021.

[42] Richard Ha, Peter Chang, Jenika Karcich, Simukayi Mutasa, Reza Fardanesh, Ralph T Wynn, Michael Z Liu, and Sachin Jambawalikar. Axillary lymph node evaluation utilizing convolutional neural networks using mri dataset. *Journal of digital imaging*, 31(6):851–856, 2018.

[43] Christoph Haarburger, Peter Langenberg, Daniel Truhn, Hannah Schneider, Johannes Thüring, Simone Schrading, Christiane K. Kuhl, and Dorit Merhof. Transfer learning for breast cancer malignancy classification based on dynamic contrast-enhanced mr images. In *Bildverarbeitung für die Medizin 2018*, pages 216–221, Berlin, Heidelberg, 2018. Springer Berlin Heidelberg.

[44] Lu Han, Yongbei Zhu, Zhenyu Liu, Tao Yu, Cuiju He, Wenyan Jiang, Yangyang Kan, Di Dong, Jie Tian, and Yahong Luo. Radiomic nomogram for prediction of axillary lymph node metastasis in breast cancer. *European radiology*, 29(7):3820–3829, 2019.

[45] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[47] Hadeer A Helaly, Mahmoud Badawy, and Amira Y Haikal. Deep learning approach for early detection of alzheimer's disease. *Cognitive Computation*, pages 1–17, 2021.

[48] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[49] Qiyuan Hu, Heather M Whitney, Hui Li, Yu Ji, Peifang Liu, and Maryellen L Giger. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection mri in breast cancer diagnosis using dynamic contrast-enhanced mri. *Radiology: Artificial Intelligence*, 3(3), 2021.

[50] Kai-Lung Hua, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Yu-Jen Chen. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, 8, 2015.

[51] Yechong Huang, Jiahang Xu, Yuncheng Zhou, Tong Tong, Xiahai Zhuang, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. *Frontiers in neuroscience*, page 509, 2019.

[52] Charles P Hughes, Leonard Berg, Warren Danziger, Lawrence A Coben, and Ronald L Martin. A new clinical scale for the staging of dementia. *The British journal of psychiatry*, 140(6):566–572, 1982.

[53] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA annual symposium proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.

[54] A Ibrahim, S Primakov, M Beuque, HC Woodruff, I Halilaj, G Wu, T Refaee, R Granzier, Y Widaatalla, Roland Hustinx, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods*, 188:20–29, 2021.

[55] Emimal Jabason, M Omair Ahmad, and MNS Swamy. Classification of alzheimer's disease from mri data using an ensemble of hybrid deep convolutional neural networks. In *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 481–484. IEEE, 2019.

[56] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[57] Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90, 2011.

[58] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[59] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020.

[60] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[61] Bijen Khagi, Bumshik Lee, Jae-Young Pyun, and Goo-Rak Kwon. Cnn models performance analysis on mri images of oasis dataset for distinction between healthy and alzheimer's patient. In *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4. IEEE, 2019.

[62] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.

[63] M Khojaste-Sarakhsi, Seyedhamidreza Shahabi Haghighi, SMT Fatemi Ghomi, and Elena Marchiori. Deep learning for alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, page 102332, 2022.

[64] Mingyu Kim, Jihye Yun, Yongwon Cho, Keewon Shin, Ryoungwoo Jang, Hyun-jin Bae, and Namkug Kim. Deep learning in medical imaging. *Neurospine*, 16(4):657, 2019.

[65] Sun-Ah Kim, Nariya Cho, Eun Bi Ryu, Mirinae Seo, Min Sun Bae, Jung Min Chang, and Woo Kyung Moon. Background parenchymal signal enhancement ratio at preoperative mr imaging: association with subsequent local recurrence in patients with ductal carcinoma in situ after breast conservation surgery. *Radiology*, 270(3):699–707, 2014.

[66] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[67] Ryan Kiros, Karteek Popuri, Dana Cobzas, and Martin Jagersand. Stacked multiscale feature learning for domain independent medical image segmentation. In *International workshop on machine learning in medical imaging*, pages 25–32. Springer, 2014.

[68] William E Klunk, Henry Engler, Agneta Nordberg, Yanming Wang, Gunnar Blomqvist, Daniel P Holt, Mats Bergström, Irina Savitcheva, Guo-Feng Huang, Sergio Estrada, et al. Imaging brain amyloid in alzheimer's disease with pittsburgh compound-b. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 55(3):306–319, 2004.

[69] WHAT DO WE KNOW. What is alzheimer's disease? 1986.

[70] Egor Krivov, Maxim Pisov, and Mikhail Belyaev. Mri augmentation via elastic registration for brain lesions segmentation. In *International MICCAI Brainlesion Workshop*, pages 369–380. Springer, 2017.

[71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[72] Christiane Katharina Kuhl, Peter Mielcareck, Sven Klaschik, Claudia Leutner, Eva Wardelmann, Jurgen Gieseke, and Hans H Schild. Dynamic breast mr imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211(1):101–110, 1999.

[73] Devinder Kumar, Alexander Wong, and David A Clausi. Lung nodule classification using deep features in ct images. In *2015 12th conference on computer and robot vision*, pages 133–138. IEEE, 2015.

[74] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019.

[75] Curtis P Langlotz, Bibb Allen, Bradley J Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S Cook, Adam E Flanders, Matthew P Lungren, David S Mendelson, Jeffrey D Rudie, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 291(3):781, 2019.

[76] Eric B Larson, Walter A Kukull, and Robert L Katzman. Cognitive impairment: dementia and alzheimer's disease. *Annual review of public health*, 13:431–449, 1992.

[77] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.

[78] Seung-Hak Lee, Hyunjin Park, and Eun Sook Ko. Radiomics in breast imaging from techniques to clinical applications: a review. *Korean Journal of Radiology*, 21(7):779, 2020.

[79] Yan-Wei Lee, Chiun-Sheng Huang, Chung-Chih Shih, and Ruey-Feng Chang. Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks. *Computers in Biology and Medicine*, 130:104206, 2021.

[80] Constance D. Lehman, Constantine Gatsonis, Christiane K. Kuhl, R. Edward Hendrick, Etta D. Pisano, Lucy Hanna, Sue Peacock, Stanley F. Smazal, Daniel D. Maki, Thomas B. Julian, Elizabeth R. DePeri, David A. Bluemke, and Mitchell D. Schnall. Mri evaluation of the contralateral breast in women with recently diagnosed breast cancer. *New England Journal of Medicine*, 356(13):1295–1303, 2007. PMID: 17392300.

[81] Sarah J Lewis, Ziba Gandomkar, and Patrick C Brennan. Artificial intelligence in medical imaging practice: looking to the future. *Journal of Medical radiation sciences*, 66(4):292–295, 2019.

[82] Chunling Liu, Jie Ding, Karl Spuhler, Yi Gao, Mario Serrano Sosa, Meghan Moriarty, Shahid Hussain, Xiang He, Changhong Liang, and Chuan Huang. Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced mri. *Journal of Magnetic Resonance Imaging*, 49(1):131–140, 2019.

[83] Jia Liu, Dong Sun, Linli Chen, Zheng Fang, Weixiang Song, Dajing Guo, Tiangen Ni, Chuan Liu, Lin Feng, Yuwei Xia, et al. Radiomics analysis of dynamic contrast-enhanced magnetic resonance imaging for the prediction of sentinel lymph node metastasis in breast cancer. *Frontiers in oncology*, 9:980, 2019.

[84] Meijie Liu, Ning Mao, Heng Ma, Jianjun Dong, Kun Zhang, Kaili Che, Shaofeng Duan, Xuexi Zhang, Yinghong Shi, and Haizhu Xie. Pharmacokinetic parameters and radiomics model based on dynamic contrast enhanced mri for the preoperative prediction of sentinel lymph node metastasis in breast cancer. *Cancer Imaging*, 20(1):1–8, 2020.

[85] Siqi Liu, Sidong Liu, Weidong Cai, Hangyu Che, Sonia Pujol, Ron Kikinis, Dagan Feng, Michael J Fulham, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. *IEEE transactions on biomedical engineering*, 62(4):1132–1140, 2014.

[86] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[87] JB Antoine Maintz and Max A Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998.

[88] Ritse M Mann, Nariya Cho, and Linda Moy. Breast mri: state of the art. 2019.

[89] Stefano Marrone, Gabriele Piantadosi, Roberta Fusco, Antonella Petrillo, Mario Sansone, and Carlo Sansone. An investigation of deep learning for lesions malignancy classification in breast dce-mri. In *Image Analysis and Processing - ICIAP 2017*, pages 479–489, Cham, 2017. Springer International Publishing.

[90] Gabriele Martelli, Patrizia Boracchi, Michaela De Palo, Silvana Pilotti, Saro Oriana, Roberto Zucali, Maria Grazia Daidone, and Giuseppe De Palo. A randomized trial comparing axillary dissection to no axillary dissection in older patients with t1n0 breast cancer: results after 5 years of follow-up. *Annals of surgery*, 242(1):1, 2005.

[91] Aidana Massalimova and Huseyin Atakan Varol. Input agnostic deep learning for alzheimer's disease classification using multimodal mri images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2875–2878. IEEE, 2021.

[92] David Mattes, David R Haynor, Hubert Vesselle, Thomas K Lewellyn, and William Eubank. Nonrigid multimodality image registration. In *Medical imaging 2001: image processing*, volume 4322, pages 1609–1620. Spie, 2001.

[93] Marius E Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4):488–495, 2020.

[94] Rebecca L McNamee, Seong-Hwan Yee, Julie C Price, William E Klunk, Bedda Rosario, Lisa Weissfeld, Scott Ziolko, Michael Berginc, Brian Lopresti, Steven DeKosky, et al. Consideration of optimal time window for pittsburgh compound b pet summed uptake measurements. *Journal of Nuclear Medicine*, 50(3):348–355, 2009.

[95] FAOCR Michelle C. Walters, D.O. and FACR Lennard Nadalo M.D. Mri breast clinical indications: A comprehensive review, 2015.

[96] Randolph A Miller, Melissa A McNeil, Sue M Challinor, Fred E Masarie Jr, and Jack D Myers. The internist-1/quick medical reference project—status report. *Western Journal of Medicine*, 145(6):816, 1986.

[97] Catherine J Moran, Brian A Hargreaves, Manojkumar Saranathan, Jafi A Lipson, Jennifer Kao, Debra M Ikeda, and Bruce L Daniel. 3d t2-weighted

spin echo imaging in the breast. *Journal of Magnetic Resonance Imaging*, 39(2):332–338, 2014.

[98] Saeeda Naz, Abida Ashraf, and Ahmad Zaib. Transfer learning using freeze features for alzheimer neurological disorder detection using adni dataset. *Multimedia Systems*, 28(1):85–94, 2022.

[99] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[100] Son Nguyen, Dogan Polat, Paniz Karbasi, Daniel Moser, Liqiang Wang, Keith Hulsey, Murat Can Çobanoğlu, Basak Dogan, and Albert Montillo. Preoperative prediction of lymph node metastasis from clinical dce mri of the primary breast tumor using a 4d cnn. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 326–334. Springer, 2020.

[101] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288, 2016.

[102] Modupe Odusami, Rytis Maskeliunas, Robertas Damaševičius, and Sanjay Misra. Comparable study of pre-trained model on alzheimer disease classification. In *International Conference on Computational Science and Its Applications*, pages 63–74. Springer, 2021.

[103] Ohad Oren, Bernard J Gersh, and Deepak L Bhatt. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*, 2(9):e486–e488, 2020.

[104] World Health Organization et al. Guide to cancer early diagnosis. 2017.

[105] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[106] Vishwa S Parekh and Michael A Jacobs. Deep learning and radiomics in precision medicine. *Expert review of precision medicine and drug development*, 4(2):59–72, 2019.

[107] Geoff JM Parker, Caleb Roberts, Andrew Macdonald, Giovanni A Buonaccorsi, Sue Cheung, David L Buckley, Alan Jackson, Yvonne Watson, Karen Davies, and Gordon C Jayson. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function

for dynamic contrast-enhanced mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(5):993–1000, 2006.

[108] Anna Perrone, Pietro Guerrisi, Luciano Izzo, Ilaria D'Angeli, Simona Sassi, Luigi Lo Mele, Marina Marini, Dario Mazza, and Mario Marini. Diffusion-weighted mri in cervical lymph nodes: differentiation between benign and malignant lesions. *European journal of radiology*, 77(2):281–286, 2011.

[109] Gabriele Piantadosi, Roberta Fusco, Antonella Petrillo, Mario Sansone, and Carlo Sansone. Lbp-top for volume lesion classification in breast dce-mri. In *Image Analysis and Processing — ICIAP 2015*, pages 647–657, Cham, 2015. Springer International Publishing.

[110] Gabriele Piantadosi, Stefano Marrone, Roberta Fusco, Mario Sansone, and Carlo Sansone. Comprehensive computer-aided diagnosis for breast t1-weighted dce-mri through quantitative dynamical features and spatio-temporal local binary patterns. *IET Computer Vision*, 12(7):1007–1017, 2018.

[111] Gabriele Piantadosi, Stefano Marrone, Antonio Galli, Mario Sansone, and Carlo Sansone. Dce-mri breast lesions segmentation with a 3tp u-net deep convolutional neural network. In *2019 IEEE 32nd International Symposium on CBMS*, pages 628–633. IEEE, 2019.

[112] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.

[113] Alejandro Puente-Castro, Enrique Fernandez-Blanco, Alejandro Pazos, and Cristian R Munteanu. Automatic assessment of alzheimer's disease diagnosis based on deep learning techniques. *Computers in Biology and Medicine*, 120:103764, 2020.

[114] Shangran Qiu, Prajakta S Joshi, Matthew I Miller, Chonghua Xue, Xiao Zhou, Cody Karjadi, Gary H Chang, Anant S Joshi, Brigid Dwyer, Shuhan Zhu, et al. Development and validation of an interpretable deep learning framework for alzheimer's disease classification. *Brain*, 143(6):1920–1933, 2020.

[115] Smriti Rahunathan, Don Stredney, P Schmalbrock, and Bradley D Clymer. Image registration using rigid registration and maximization of mutual information. In *13th Annu. Med. Meets Virtual Reality Conf*, 2005.

[116] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[117] Erik R Ranschaert, Sergey Morozov, and Paul R Algra. *Artificial intelligence in medical imaging: opportunities, applications and risks.* Springer, 2019.

[118] Reza Rasti, Mohammad Teshnehlab, and Son Lam Phung. Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. *Pattern Recognition*, 72:381–390, 2017.

[119] Ahmedin Jemal Rebecca L Siegel, Kimberly D Miller. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.

[120] Thomas Ren, Renee Cattell, Hongyi Duanmu, Pauline Huang, Haifang Li, Rami Vanguri, Michael Z Liu, Sachin Jambawalikar, Richard Ha, Fusheng Wang, et al. Convolutional neural network detection of axillary lymph node metastasis using standard clinical breast mri. *Clinical breast cancer*, 20(3):e301–e308, 2020.

[121] Felix Ritter, Tobias Boskamp, André Homeyer, Hendrik Laue, Michael Schwier, Florian Link, and H-O Peitgen. Medical image analysis. *IEEE pulse*, 2(6):60–70, 2011.

[122] Christopher C Rowe and Victor L Villemagne. Brain amyloid imaging. *Journal of nuclear medicine technology*, 41(1):11–18, 2013.

[123] Sanaz Samiei, Renée WY Granzier, Abdalla Ibrahim, Sergey Primakov, Marc BI Lobbes, Regina GH Beets-Tan, Thiemo JA van Nijnatten, Sanne ME Engelen, Henry C Woodruff, and Marjolein L Smidt. Dedicated axillary mri-based radiomics analysis for the prediction of axillary lymph node metastasis in breast cancer. *Cancers*, 13(4):757, 2021.

[124] Domiziana Santucci, Eliodoro Faiella, Ermanno Cordelli, Alessandro Calabrese, Roberta Landi, Carlo de Felice, Bruno Beomonte Zobel, Rosario Francesco Grasso, Giulio Iannello, and Paolo Soda. The impact of tumor edema on t2-weighted 3t-mri invasive breast cancer histological characterization: A pilot radiomics study. *Cancers*, 13(18):4635, 2021.

[125] Domiziana Santucci, Eliodoro Faiella, Ermanno Cordelli, Rosa Sicilia, Carlo de Felice, Bruno Beomonte Zobel, Giulio Iannello, and Paolo Soda. 3t mri-radiomic approach to predict for lymph node status in breast cancer patients. *Cancers*, 13(9):2228, 2021.

[126] Domiziana Santucci, Eliodoro Faiella, Michela Gravina, Ermanno Cordelli, Carlo de Felice, Bruno Beomonte Zobel, Giulio Iannello, Carlo Sansone, and Paolo Soda. Cnn-based approaches with different tumor bounding options for lymph node status prediction in breast dce-mri. *Cancers*, 14(19):4574, 2022.

[127] Cristina L Saratxaga, Iratxe Moya, Artzai Picón, Marina Acosta, Aitor Moreno-Fernandez-de Leceta, Estibaliz Garrote, and Arantza Bereciartua-Perez. Mri deep learning-based solution for alzheimer's disease prediction. *Journal of personalized medicine*, 11(9):902, 2021.

[128] Matthias C Schabel, Glen R Morrell, Karen Y Oh, Cheryl A Walczak, R Brad Barlow, and Leigh A Neumayer. Pharmacokinetic mapping for lesion classification in dynamic breast mri. *Journal of Magnetic Resonance Imaging*, 31(6):1371–1378, 2010.

[129] Philip Scheltens, Nick Fox, Frederik Barkhof, and Charles De Carli. Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. *The Lancet Neurology*, 1(1):13–21, 2002.

[130] Mahesh K Shetty and Wendy S Carpenter. Sonographic evaluation of isolated abnormal axillary lymph nodes identified on mammograms. *Journal of ultrasound in medicine*, 23(1):63–71, 2004.

[131] Jun Shi, Xiao Zheng, Yan Li, Qi Zhang, and Shihui Ying. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease. *IEEE journal of biomedical and health informatics*, 22(1):173–183, 2017.

[132] Edward H Shortliffe, Randall Davis, Stanton G Axline, Bruce G Buchanan, C Cordell Green, and Stanley N Cohen. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system. *Computers and biomedical research*, 8(4):303–320, 1975.

[133] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, September 2018.

[134] Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention*, pages 10–18. Springer, 2016.

[135] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[136] Hans-Peter Sinn and Hans Kreipe. A brief overview of the who classification of breast tumors. *Breast care*, 8(2):149–154, 2013.

[137] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[138] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014.

[139] Qiuchang Sun, Xiaona Lin, Yuanshen Zhao, Ling Li, Kai Yan, Dong Liang, Desheng Sun, and Zhi-Cheng Li. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. *Frontiers in oncology*, 10:53, 2020.

[140] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.

[141] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[142] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[143] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[144] Alexandra-Maria Tăuţan, Bogdan Ionescu, and Emiliano Santarnecchi. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. *Artificial Intelligence in Medicine*, 117:102081, 2021.

[145] Kim-Han Thung, Pew-Thian Yap, and Dinggang Shen. Multi-stage diagnosis of alzheimer's disease with incomplete multimodal data via multi-task deep learning. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 160–168. Springer, 2017.

[146] Paul S Tofts. Modeling tracer kinetics in dynamic gd-dtpa mr imaging. *Journal of magnetic resonance imaging*, 7(1):91–101, 1997.

[147] Paul S Tofts, Gunnar Brix, David L Buckley, Jeffrey L Evelhoch, Elizabeth Henderson, Michael V Knopp, Henrik BW Larsson, Ting-Yim Lee, Nina A Mayr, Geoffrey JM Parker, et al. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted mri of a diffusable tracer: standardized quantities and symbols. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 10(3):223–232, 1999.

[148] Selene Tomassini, Nicola Falcionelli, Paolo Sernani, Henning Müller, and Aldo Franco Dragoni. An end-to-end 3d convlstm-based framework for early diagnosis of alzheimer's disease from full-resolution whole-brain smri scans. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 74–78. IEEE, 2021.

[149] Lindsay W Turnbull. Dynamic contrast-enhanced mri in the diagnosis and management of breast cancer. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, 22(1):28–39, 2009.

[150] Shigeto Ueda, Hitoshi Tsuda, Hideki Asakawa, Jiro Omata, Kazuhiko Fukatsu, Nobuo Kondo, Tadaharu Kondo, Yukihiro Hama, Katsumi Tamura, Jiro Ishida, et al. Utility of 18 f-fluoro-deoxyglucose emission tomography/computed tomography fusion imaging (18 f-fdg pet/ct) in combination with ultrasonography for axillary staging in primary breast cancer. *BMC cancer*, 8(1):1–10, 2008.

[151] MW Vernooij, FB Pizzini, Reinhold Schmidt, Marion Smits, TA Yousry, N Bargallo, GB Frisoni, Sven Haller, and Frederik Barkhof. Dementia imaging in clinical practice: a european-wide survey of 193 centres and conclusions by the esnr working group. *Neuroradiology*, 61(6):633–642, 2019.

[152] Tien Duong Vu, Hyung-Jeong Yang, Van Quan Nguyen, A-Ran Oh, and Mi-Sun Kim. Multimodal learning using convolution neural network and sparse autoencoder. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 309–312. IEEE, 2017.

[153] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[154] Jane Wardle, Kathryn Robb, Sally Vernon, and Jo Waller. Screening for prevention and early diagnosis of cancer. *American psychologist*, 70(2):119, 2015.

[155] Hanns-Joachim Weinmann, M Laniado, and W Mützel. Pharmacokinetics of gddtpa/dimeglumine after intravenous injection into healthy volunteers. *Physiological chemistry and physics and medical NMR*, 16(2):167–172, 1984.

[156] Jimmy Wu, Bolei Zhou, Diondra Peck, Scott Hsieh, Vandana Dialani, Lester Mackey, and Genevieve Patterson. Deepminer: Discovering interpretable representations for mammogram classification and explanation. *arXiv preprint arXiv:1805.12323*, 2018.

[157] Kushpal Singh Yadav and Krishna Prasad Miyapuram. A novel approach towards early detection of alzheimer's disease using deep learning on magnetic resonance images. In *International Conference on Brain Informatics*, pages 486–495. Springer, 2021.

[158] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[159] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[160] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[161] Chuanchuan Zheng, Yong Xia, Yuanyuan Chen, Xiaoxia Yin, and Yanchun Zhang. Early diagnosis of alzheimer's disease by ensemble deep learning using fdg-pet. In *International Conference on Intelligent Science and Big Data Engineering*, pages 614–622. Springer, 2018.

[162] Xueyi Zheng, Zhao Yao, Yini Huang, Yanyan Yu, Yun Wang, Yubo Liu, Rushuang Mao, Fei Li, Yang Xiao, Yuanyuan Wang, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nature communications*, 11(1):1–9, 2020.

[163] Leilei Zhou, Zuoheng Zhang, Xindao Yin, Hong-Bing Jiang, Jie Wang, Guan Gui, Yu-Chen Chen, and Jin-Xia Zheng. Transfer learning-based dce-mri method for identifying differentiation between benign and malignant breast tumors. *IEEE Access*, 8:17527–17534, 2020.

[164] Li-Qiang Zhou, Xing-Long Wu, Shu-Yan Huang, Ge-Ge Wu, Hua-Rong Ye, Qi Wei, Ling-Yun Bao, You-Bin Deng, Xing-Rui Li, Xin-Wu Cui, et al. Lymph node metastasis prediction from primary breast cancer us images using deep learning. *Radiology*, 294(1):19–28, 2020.

[165] Yongjin Zhou, Weijian Huang, Pei Dong, Yong Xia, and Shanshan Wang. D-unet: a dimension-fusion u shape network for chronic stroke lesion segmentation. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[166] Roger Zoorob, Russell Anderson, Charles Cefalu, and Mohamad Sidani. Cancer screening guidelines. *American family physician*, 63(6):1101–1112, 2001.

[167] Hasib Zunair and A Ben Hamza. Melanoma detection using adversarial training and deep transfer learning. *Physics in Medicine & Biology*, 2020.

Author’s Publications

1. Galli, A., Gravina, M., Marrone, S., Piantadosi, G., Sansone, M., Sansone, C. (2019, September). Evaluating impacts of motion correction on deep learning approaches for breast DCE-MRI segmentation and classification. In International Conference on Computer Analysis of Images and Patterns (pp. 294-304). Springer, Cham.

2. Gravina, M., Marrone, S., Piantadosi, G., Sansone, M., Sansone, C. (2019, September). 3TP-CNN: radiomics and deep learning for lesions classification in DCE-MRI. In International Conference on Image Analysis and Processing (pp. 661-671). Springer, Cham.

3. Galli, A., Gravina, M., Moscato, V., Picariello, A., Sansone, C., Sperlí, G. (2019, September). A Business Reputation Methodology Using Social Network Analysis. In International Symposium on Pervasive Systems, Algorithms and Networks (pp. 96-106). Springer, Cham.

4. De Santo, A., Galli, A., Gravina, M., Moscato, V., Sperlì, G. (2020). Deep Learning for HDD health assessment: An application based on LSTM. IEEE Transactions on Computers, 71(1), 69-80

5. Gravina, M., Marrone, S., Piantadosi, G., Moscato, V., Sansone, C. (2021, January). Developing a smart PACS: CBIR system using deep learning. In International Conference on Pattern Recognition (pp. 296-309). Springer, Cham

6. Gravina, M., Gragnaniello, D., Verdoliva, L., Poggi, G., Corsini, I., Dani, C., Meneghin, F., Lista, G., Aversa, S., Raimondi, F., Migliaro, F., Sansone, C. (2021, January). Deep learning in the ultrasound evaluation of neonatal respiratory status. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 10493-10499). IEEE.

7. Gravina, M., Marrone, S., Sansone, M., Sansone, C. (2021). DAE-CNN: Exploiting and disentangling contrast agent effects for breast lesions classification in DCE-MRI. Pattern Recognition Letters, 145, 67-73
8. Bollino, R., Bovenzi, G., Cipolletta, F., Docimo, L., Gravina, M., Marrone, S., Parmeggiani, D., Sansone, C. (2022). Synergy-Net: Artificial Intelligence at the Service of Oncological Prevention. In Handbook of Artificial Intelligence in Healthcare (pp. 389-424). Springer, Cham.
9. Gravina, M., Marrone, S., Docimo, L., Santini, M., Fiorelli, A., Parmeggiani, D., Sansone, C. (2022). Leveraging CycleGAN in Lung CT Sinogram-free Kernel Conversion. In International Conference on Image Analysis and Processing (pp. 100-110). Springer, Cham.
10. Sannino, C., Gravina, M., Marrone, S., Fiameni, G., Sansone, C. (2022). LessonAble: Leveraging Deep Fakes in MOOC Content Creation. In International Conference on Image Analysis and Processing (pp. 27-37). Springer, Cham.
11. Pontillo, G., Penna, S., Cocozza, S., Quarantelli, M., Gravina, M., Lanzillo, R., Marrone, S., Costabile, T., Inglese, M., Morra Brescia, V., Riccio, D., Elefante, A., Petracca, M., Sansone, C., Brunetti, A. (2022). Stratification of multiple sclerosis patients using unsupervised machine learning: a single-visit MRI-driven approach. European Radiology, 1-10
12. Gravina, M., Spirito, L., Celentano, G., Capece, M., Creta, M., Califano, G., Collà Ruvolo, C., Morra, S., Imbriaco, M., Di Bello, F., Sciuto, A., Cuocolo, R., Napolitano, L., La Rocca, La Rocca, R., Mirone, V., Sansone, C., Longo, N. (2022). Machine Learning and Clinical-Radiological Characteristics for the Classification of Prostate Cancer in PI-RADS 3 Lesions. Diagnostics, 12(7), 1565.
13. Gravina, M., Cordelli, E., Santucci, D., Soda, P., Sansone, C., (2022, August). Evaluating Tumour Bounding Options for Deep Learning-based Axillary Lymph Node Metastasis Prediction in Breast Cancer. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 4335-4342). IEEE
14. Santucci, D., Faiella, E., Gravina, M., Cordelli, E., de Felice, C., Beomonte Zobel, B., Iannello, G., Sansone, C., Soda, P. (2022). CNN-Based Approaches with Different Tumor Bounding Options for Lymph Node Status Prediction in Breast DCE-MRI. Cancers, 14(19), 4574.
15. Capuozzo, S., Gravina, M., Gatta, G., Marrone, S., Sansone, C. (2022). A Multimodal Knowledge-Based Deep Learning Approach for MGMT Promoter Methylation Identification. Journal of Imaging, 8(12), 321.

16. Calabrese, A., Santucci, D., Gravina, M., Faiella, E., Cordelli, E., Soda, P., Iannello G., Sansone, C., Beomonte Zobel, B., Catalano, C., de Felice, C. (2022). 3T-MRI Artificial Intelligence in Patients with Invasive Breast Cancer to predict distant metastasis status: a pilot study. Cancers, (ISSN 2072-6694)