



**HAL**  
open science

# Representation learning of multimodal and longitudinal data for renal transplantation monitoring

Léo Milecki

► **To cite this version:**

Léo Milecki. Representation learning of multimodal and longitudinal data for renal transplantation monitoring. Artificial Intelligence [cs.AI]. Université Paris-Saclay, 2024. English. NNT : 2024UP-AST009 . tel-04702855

**HAL Id: tel-04702855**

**<https://theses.hal.science/tel-04702855v1>**

Submitted on 19 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Representation Learning of Multimodal and Longitudinal data for Renal Transplantation monitoring

*Apprentissage de représentations de données multimodales et  
longitudinales pour le suivi de la transplantation rénale*

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°573 interfaces: matériaux, systèmes, usages (INTERFACES)  
Spécialité de doctorat: Ingénierie des systèmes complexes  
Graduate School: Sciences de l'ingénierie et des systèmes  
Référent: CentraleSupélec

Thèse préparée dans l'unité de recherche **Mathématiques et Informatique pour la Complexité et les Systèmes**,  
(CentraleSupélec, Université Paris-Saclay), 91190, Gif-sur-Yvette, France, sous la direction de **Marc-Oliver  
TIMSIT**, professeur des universités et praticien hospitalier, et la co-direction de **Maria VAKALOPOULOU**,  
maîtresse de conférence.

Thèse soutenue à Paris-Saclay, le 17 Janvier 2024, par

**Léo MILECKI**

## Composition du jury

Membres du jury avec voix délibérative

<b>Carole LARTIZIEN</b> Directrice de recherche CNRS CREATIS, Université Lyon 1, INSA Lyon, INSERM, Lyon, France.	Présidente
<b>Dimitris SAMARAS</b> Professeur Department of Computer Science, Stony Brook University, Stony Brook, USA.	Rapporteur & Examineur
<b>Enzo FERRANTE</b> Chercheur CONICET, Universidad Nacional del Litoral, Santa Fe, Argentina.	Rapporteur & Examineur
<b>Mert R. SABUNCU</b> Professeur Cornell Tech, Department of Radiology, Weill Cornell Medicine, New York, USA.	Examineur
<b>Pietro GORI</b> Maître de conférence LTCl, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France.	Examineur
<b>Alexandre LOUPY</b> Professeur des universités et praticien hospitalier Université Paris Cité, INSERM, PARCC, Paris, France.	Examineur

**Titre:** Apprentissage de représentations de données multimodales et longitudinales pour le suivi de la transplantation rénale

**Mots clés:** apprentissage profond, apprentissage auto-supervisé, apprentissage multimodal, données longitudinales, imagerie médicale, transplantation rénale.

**Résumé:**

L'intelligence artificielle (IA) a révolutionné la recherche en vision par ordinateur il y a une décennie. Depuis lors, les algorithmes d'IA ont démontré une pléthore d'études remarquables et d'approches à la pointe de l'état de l'art pour de nombreuses applications, notamment l'analyse d'imagerie médicale. L'objectif de cette thèse est d'explorer les récentes avancées en apprentissage automatique, en mettant l'accent sur les approches d'apprentissage profond, afin de concevoir des algorithmes pour le diagnostic basé sur l'imagerie médicale. En particulier, des méthodologies innovantes exploitant des techniques d'auto-supervision et des approches multimodales sont développées pour générer des prédicteurs robustes capables d'associer les caractéristiques de l'imagerie à la progression du patient. Ces méthodologies sont employées pour une application clinique spécifique, le suivi de patients ayant subi une transplantation rénale.

La transplantation rénale apparaît comme la solution la plus efficace pour l'insuffisance rénale en phase terminale. Cependant, plusieurs mécanismes, dont l'ischémie, l'hyperméthylation, le rejet aigu, la fibrose ou les lésions parenchymateuses des donneurs, peuvent entraîner des lésions rénales et affecter la survie du greffon. À ce jour, la prise en charge clinique des bénéficiaires de la transplantation repose principalement sur des biopsies percutanées de l'organe répétées, entraînant de l'inconfort, des coûts accrus et un risque significatif de complications hémorragiques. Étant donné que la vascularisation des tissus ou l'infiltration des cellules inflammatoires peuvent induire des signaux spécifiques en imagerie par résonance magnétique (IRM) ou en échographie, nous émettons l'hypothèse que la transplantation rénale constitue une source pertinente de données pour les études d'imagerie.

Tout d'abord, nous concevons des outils de détection et de segmentation des reins transplantés basés sur une approche d'apprentissage profond non supervisée, exploitant la différence de rehaussement des greffons dans différentes séquences IRM. Cette étape cruciale permet le calcul de caractéristiques radiomiques prédéfinies dans le greffon et permet d'appliquer une technique de base pour extraire des biomarqueurs robustes à partir des données d'imagerie médicale. Ensuite, nous proposons une approche pour apprendre directement à extraire des caractéristiques pertinentes à partir des données d'IRM de greffons rénaux en utilisant des méthodologies d'apprentissage contrastif. Nous explorons la pertinence de la représentation apprise pour la prédiction de la fonction rénale à partir des examens d'imagerie longitudinaux en utilisant une architecture de transformer adaptée à la gestion des examens manquants. Nous démontrons également le potentiel de nos représentations d'imagerie apprises pour prédire la survie du greffon à partir des examens précoces d'IRM post-transplantation. Enfin, nous explorons des approches multimodales pour intégrer davantage les informations cliniques et biologiques dans l'apprentissage de nos représentations d'image en exploitant l'expressivité des données textuelles avec les récentes avancées en traitement automatique du langage naturel et en utilisant une technique d'apprentissage contrastif multivues.

Nos expériences mettent en évidence la pertinence de la prise en compte des données d'imagerie longitudinales pour extraire des représentations et des informations pertinentes en vue du suivi des transplantations et, par conséquent, dans l'étude des mécanismes de complications et de dysfonctionnements chroniques après la transplantation, ouvrant la voie à des recherches futures dans ce domaine.

**Title:** Representation Learning of Multimodal and Longitudinal data for Renal Transplantation monitoring

**Keywords:** deep learning, self-supervised learning, multimodal learning, longitudinal data, medical imaging, renal transplantation.

**Abstract:**

A decade ago, the field of computer vision was transformed by the advent of Artificial Intelligence (AI). AI algorithms have since made significant strides in various applications, with particular success in medical imaging analysis. This thesis explores recent advancements in machine learning, with a specific focus on leveraging deep learning techniques to develop algorithms for diagnosing medical conditions based on imaging data. The methodologies introduced in this research utilize self-supervision techniques and multimodal approaches, aiming to create robust predictive models that connect imaging features with patient progress. This work is tailored to address a specific clinical need: the monitoring of renal transplant recipients.

Renal transplantation has emerged as a highly effective treatment for end-stage renal disease. However, various mechanisms, including ischemia, hypermethylation, acute rejection, fibrosis, or donor-related parenchymal lesions, can contribute to kidney injury and affect transplant outcomes. Presently, the clinical management of transplant recipients relies heavily on recurrent percutaneous biopsies, which come with drawbacks such as patient discomfort, increased costs, and the risk of hemorrhagic complications. Given that specific Magnetic Resonance Imaging or ultrasound signals can be induced by tissue vascularization and inflammatory cell infiltration, the hypothesis is that kidney transplant data represents a valuable resource for imaging studies.

Our research begins by developing tools for the detec-

tion and segmentation of renal transplants using an unsupervised deep learning method. These tools capitalize on differences in transplant enhancement across multiple MRI sequences, enabling the computation of imaging radiomics features and serving as a baseline for extracting robust imaging biomarkers. Subsequently, an approach that employs contrastive learning methods to extract pertinent features from renal transplant MRI data directly is presented. The utility of these learned representations is explored in forecasting renal transplant function, particularly through longitudinal imaging follow-up examinations. Additionally, attention-based architectures are explored to handle missing data effectively. This study also underscores the potential of these imaging representations in predicting transplant survival using early MRI examinations. Finally, a multimodal approach is investigated to integrate clinical and biological information into the image representations. This is achieved by incorporating textual data and leveraging advancements in Large Language Models alongside multiview contrastive learning.

The experimental findings emphasize the relevance of considering longitudinal imaging data for meaningful representations and insights into transplant monitoring. This approach proves valuable in understanding complex complications and chronic dysfunction mechanisms following transplantation. It paves the way for future research in this domain, both from the methodological and clinical point of view.





# Acknowledgements

My foremost and heartfelt gratitude is extended to Marc-Olivier, Maria, and Vicky for their exceptional roles as thesis directors, supervisors, and mentors. Their guidance has been a source of immense pleasure throughout my PhD journey, and I am grateful for the rewarding experience of collaborating on a variety of intriguing projects under their supervision. I am profoundly grateful to Marc-Olivier for his insatiable curiosity, open-minded approach, and unwavering commitment to the development of such multidisciplinary projects. His responsiveness and dedication as my thesis director have been instrumental in our journey. My utmost gratitude goes to Maria, whose inspirational guidance and boundless enthusiasm have left an indelible mark. I deeply admire her relentless hard work and unwavering dedication to her students. From being a collaborator to a truly encouraging mentor, I would also like to extend my thanks to Vicky for her dedicated supervision, which has played a pivotal role in my academic and research pursuits.

I extend my appreciation to the esteemed members of my thesis jury, including Pr. Dimitris Samaras, Pr. Enzo Ferrante, Carole Lartizien, Pr. Alexandre Loupy, Pr. Mert R. Sabuncu and Pietro Gori, for their gracious acceptance of the invitation to be part of my defense committee and for their diligent evaluation of my research. I would like to express special gratitude to the rapporteurs, Pr. Dimitris Samaras and Pr. Enzo Ferrante, for their meticulous assessment of my manuscript and for the invaluable insights, comments, and feedback they provided. I would like to thank Pr. Arnaud Mejean for accepting the invitation to my defense.

Throughout the course of this thesis, I had the privilege of collaborating with remarkable collaborators, particularly esteemed medical professionals. I extend my gratitude to Pr. Jean-Michel Correas and Pr. Dany Anglicheau for their leadership in the field of clinical research, alongside Pr. Marc-Olivier Timsit, and for the enlightening discussions that shaped our research endeavors. I am profoundly thankful to Sylvain Bordard for his unwavering support, welcoming me almost anytime in his office at Necker Hospital, and assisting me at various stages of my research, from data collection to writing articles together. Establishing a close connection with the clinical setting proved pivotal in our projects. I also wish to express my appreciation to the additional medical collaborators who contributed to data curation, namely Anne-Marie Tissier and Florence Poinard, as well as the gracious staff at Necker and Georges Pompidou Hospitals for their warm hospitality.

I am grateful to the Provost Ultrasound Lab, with special mention to Pr. Jean Provost, for warmly hosting me in Montreal once again. Those three months were truly a breath of fresh air, allowing me to delve deeper into the enchanting city of Montreal, even under its snowy blanket. This research visit was not only productive but also allowed me to reunite with familiar faces, as well as forge new friendships. My initial research experience in Montreal inspired me to embark on this PhD journey, and this return visit only reaffirmed the value of those experiences.

I extend a heartfelt appreciation to the MICS lab for providing a vibrant working environment that brings together exceptional students, engineers, professors, and researchers dedicated to both outstanding research and teaching activities. The lab not only served as a stimulating workspace but also as a place where we engaged in games of table football (the true champion still remains a mystery, but it is not me), shared moments over drinks, and enjoyed some fun on the "Silicon Plateau."

Last but certainly not least, I want to express my deepest gratitude to my family and cherished friends. To my parents, I offer my unending thanks for their unwavering and unconditional support and for always making it possible for me to choose the path I desired. My mom, your tenacious mindset, and the countless acts of kindness and generosity have brightened my days. To my dad, you have been the embodiment of calm strength, and despite not becoming a medical doctor, hopefully, soon, I will be the next Dr. Milecki. Sadly, I only declare this gratitude and love at 27 years old in my PhD thesis and in English! Thanks to my big brothers for watching the little bro. To all my close friends (and even those not so close), you have been instrumental in helping me strike a balance between the rigors of my PhD journey and the joys of life — be it through fun, laughter, parties, music, sports, or endless discussions. I won't enumerate names here (in case I inadvertently overlook someone), nor will I elaborate further (I realize that most of you may not even open this thesis, or actually perhaps only to read the acknowledgments section?). Nevertheless, please know that your friendships are indispensable to my life, and your presence has enriched my journey beyond measure.





# Résumé en français

L'intelligence artificielle (IA) a révolutionné la recherche en vision par ordinateur il y a une décennie. Depuis lors, les algorithmes d'IA ont démontré une pléthore d'études remarquables et d'approches à la pointe de l'état de l'art pour de nombreuses applications, notamment l'analyse d'imagerie médicale. L'objectif de cette thèse est d'explorer les récentes avancées en apprentissage automatique, en mettant l'accent sur les approches d'apprentissage profond, afin de concevoir des algorithmes pour le diagnostic basé sur l'imagerie médicale. En particulier, des méthodologies innovantes exploitant des techniques d'auto-supervision et des approches multimodales sont développées pour générer des prédicteurs robustes capables d'associer les caractéristiques de l'imagerie à la progression du patient. Ces méthodologies sont employées pour une application clinique spécifique, le suivi de patients ayant subi une transplantation rénale.

La transplantation rénale apparaît comme la solution la plus efficace pour l'insuffisance rénale en phase terminale. Cependant, plusieurs mécanismes, dont l'ischémie, l'hyperméthylation, le rejet aigu, la fibrose ou les lésions parenchymateuses des donneurs, peuvent entraîner des lésions rénales et affecter la survie du greffon. À ce jour, la prise en charge clinique des bénéficiaires de la transplantation repose principalement sur des biopsies percutanées de l'organe répétées, entraînant de l'inconfort, des coûts accrus et un risque significatif de complications hémorragiques. Étant donné que la vascularisation des tissus ou l'infiltration des cellules inflammatoires peuvent induire des signaux spécifiques en imagerie par résonance magnétique (IRM) ou en échographie, nous émettons l'hypothèse que la transplantation rénale constitue une source pertinente de données pour les études d'imagerie. L'organisation de cette thèse est la suivante.

Dans un premier chapitre, nous commençons par présenter les motivations méthodologiques de cette thèse, en passant rapidement en revue certains prérequis en apprentissage automatique, ainsi qu'un bref historique sur l'apprentissage profond, puis en nous concentrant sur les problématiques récentes et les orientations de recherche qui marquent la littérature actuelle sur l'apprentissage profond. Ensuite, nous introduisons le contexte médical de cette thèse : le suivi de patients ayant subi une transplantation rénale, notamment par l'imagerie médicale. En troisième lieu, nous présentons la cohorte de données utilisée dans cette thèse afin de saisir les objectifs et les défis de notre travail. Enfin, nous introduisons l'organisation des prochains chapitres en présentant les objectifs et les contributions de cette thèse.

Le second chapitre établit les tâches fondamentales et les méthodologies de base qui sous-tendent cette thèse. Tout d'abord, nous présentons un aperçu des études et des modèles de pointe sur les données de type variables cliniques et biologiques dans le contexte de la transplantation rénale. De plus, nous proposons des expérimentations préliminaires illustratives utilisant notre ensemble de données. Ensuite, nous nous plongeons dans notre première contribution, qui comprend la détection automatique et la segmentation des greffons rénaux dans les données d'IRM. Ces tâches sont essentielles à la progression de notre recherche. Enfin, nous introduisons le concept de radiomique, qui implique l'ingénierie des caractéristiques radiologiques et sert de base pour la découverte et l'analyse de biomarqueurs à partir de méthodes d'imagerie computationnelle.

Dans un troisième chapitre, nous étudions l'utilisation d'approches d'apprentissage contrastif pour apprendre des représentations significatives des données d'IRM de contraste des transplants rénaux. Notre approche vise à incorporer des informations cliniques dans la représentation apprise des images. Ensuite, nous proposons une nouvelle architecture basée sur les transformers adaptée pour traiter les données manquantes dans la tâche complexe de prédiction de la créatinine sérique 2 ans après la transplantation en utilisant des examens d'imagerie de suivi. Une architecture d'encodeur transformer permet d'utiliser les données séquentielles de suivi pour prédire la fonction de la transplantation rénale et modéliser les relations longitudinales, en proposant une méthode innovante pour gérer les examens manquants. Dans une étude plus orientée cliniquement, nous explorons davantage la capacité de nos représentations apprises pour la prédiction de la survie du greffon.

Dans un quatrième chapitre, nous présentons MEDIMP (MEDical Images with clinical Prompts), une approche innovante visant à apprendre des représentations pertinentes à partir de données d'IRM de contraste des greffons en vue de la prédiction de la fonction du greffon dans le contexte de la surveillance des transplantations rénales. Cette approche offre une manière élégante d'incorporer des informations cliniques ou biologiques dans le processus d'apprentissage de l'extraction de caractéristiques des données d'imagerie médicale. Nous avons conçu des "prompts" clinicobiologiques pour exploiter les données textuelles et des encodeurs NLP puissants en utilisant une approche d'alignement multimodal des représentations avec de l'apprentissage contrastif. Nous démontrons la pertinence de l'approche pour la prédiction de la fonction des transplantations rénales.

Dans un dernier chapitre de discussion, nous commençons par mettre en évidence les principales contributions de cette thèse. Ensuite, nous nous plongeons dans une discussion sur des perspectives pertinentes qui ont émergé de notre travail et sur des directions prometteuses pour nos futurs projets de recherche, en mettant l'accent sur l'apprentissage de représentations pour l'analyse d'images médicales, le potentiel des modèles fondamentaux en imagerie médicale, les nouveaux outils d'intelligence artificielle pour la surveillance des transplantations rénales, et deux sujets pertinents en IA qui n'ont pas été abordés dans notre travail. Enfin, nous proposons une conclusion sur l'ensemble de la thèse.

Pour conclure ce résumé, nos expériences mettent en évidence la pertinence de la prise en compte des données d'imagerie longitudinales pour extraire des représentations et des informations pertinentes en vue du suivi des transplantations et, par conséquent, dans l'étude des mécanismes de complications et de dysfonctionnements chroniques après la transplantation, ouvrant la voie à des recherches futures dans ce domaine.







# Contents

<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Notations and conventions</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Methodological motivations . . . . .	5
1.2 Medical context for this thesis . . . . .	23
1.3 Dataset presentation . . . . .	25
1.4 Objectives & Contributions . . . . .	34
<b>2 Classical methods &amp; baselines</b>	<b>39</b>
2.1 Classical data-driven methods for medical tabular data . . . . .	43
2.2 Image-based segmentation of kidney transplant . . . . .	47
2.3 Design of radiomics signature for kidney transplant . . . . .	56
<b>3 Learning meaningful representations of renal transplant DCE MRI data</b>	<b>61</b>
3.1 Introduction . . . . .	65
3.2 Related work . . . . .	66
3.3 Methodology . . . . .	67
3.4 Data . . . . .	71
3.5 Kidney function forecasting experiments . . . . .	71
3.6 Survival analysis experiments . . . . .	76
3.7 Discussion & Conclusion . . . . .	80
<b>4 Leveraging LLMs advances for multimodal representations</b>	<b>85</b>
4.1 Introduction . . . . .	89
4.2 Related work . . . . .	89
4.3 Methodology . . . . .	91
4.4 Data . . . . .	95
4.5 Experiments & Results . . . . .	96
4.6 Discussion & Conclusion . . . . .	101
<b>5 Discussion</b>	<b>105</b>
5.1 Main contributions . . . . .	109
5.2 Perspectives & Future directions . . . . .	110
5.3 Conclusion . . . . .	114
<b>Bibliography</b>	<b>I</b>



# List of Figures

1.1	Chapter 1 cover image	3
1.2	Overview a classical machine learning pipeline.	5
1.3	Brief history of Deep Learning timeline.	9
1.4	Model overview of ResNet18 CNN and the Vision Transformer (ViT).	11
1.5	Overview of self-supervised learning (SSL) techniques categories.	13
1.6	Architecture of the transformer.	16
1.7	Summary of CLIP approach.	17
1.8	Cross-attention mechanism example.	19
1.9	Overview of two multimodal approaches in Medical Imaging Analysis.	21
1.10	Renal transplantation statistics in France.	23
1.11	Clinical study flowchart.	26
1.12	Biological data examples.	28
1.13	Histopathological images results of transplant biopsies examples.	29
1.14	MRI data slice examples of available modalities.	30
1.15	US data examples of renal transplants.	32
2.1	Chapter 2 cover image	41
2.2	Correlation matrix visualization of general clinical variables.	44
2.3	Radar charts on clinical data.	45
2.4	Overview of the proposed method for the automatic segmentation of kidney graft using multimodal MR Imaging.	48
2.5	Region of interest example on the DCE MRI sequence.	50
2.6	Semantic segmentation maps obtained through the iterative process of our unsupervised algorithm.	52
2.7	Evaluation of the different components of the proposed method using different evaluation metrics on the validation set.	53
2.8	One testing subject depicting one slice in axial view and the segmentation results.	54
2.9	Overview of a radiomics approach for renal transplant ultrasound imaging.	57
3.1	Chapter 3 cover image	63
3.2	Overview of the proposed method.	67
3.3	Contrastive learning objective explanation.	68
3.4	Curriculum learning on the weakly-supervised task example.	69
3.5	(A) Training and (B) Validation curves for our contrastive learning scheme based on GFR variable with and without curriculum learning.	71
3.6	Feature visualization of the PCA decomposition of the last layer of the contrastive training using the DCE MRI follow-ups.	72
3.7	Pipeline for generating our CL-based features and the predefined radiomics features.	76
3.8	Heatmaps visualization pipeline.	77
3.9	Boxplots of the evaluation of different CL-based features from different examination timestamps and MRI modalities available in our main cohort.	78
3.10	Kaplan-Meier curves for the two groups obtained by the unsupervised clustering of the selected CL-based features by the survival Cox model.	79
3.11	Heatmaps on the region of interest input to our CL-based model for different patients' examples.	79
4.1	Chapter 4 cover image	87
4.2	Overview of our method MEDIMP – Medical Images with clinical Prompts.	91
4.3	t-SNE visualizations of the features of the last layer of MEDIMP image encoder using the DCE MRI exams.	97
5.1	Chapter 5 cover image	107



# List of Tables

1.1	Patient characteristics and clinical attributes. . . . .	27
1.2	Missing data in the DCE MRI follow-up examinations. . . . .	33
2.1	Transplantation failure prediction with classical ML algorithms from clinical variables. . . . .	46
2.2	Quantitative evaluation of the proposed method and other methods. . . . .	54
2.3	Evaluation of the predefined radiomics features to recover variables linked to graft survival. . . . .	58
3.1	Quantitative evaluation of different methods. . . . .	74
3.2	Quantitative evaluation of different strategies for missing data. . . . .	75
3.3	Quantitative evaluation for an increasing number of missing data for our proposed method and other imputation strategies. . . . .	75
3.4	Concordance index (C-index in %) scores on evaluating our CL-based features to predict graft survival at five years. . . . .	78
4.1	Comparison of MEDIMP vs SOTA. . . . .	98
4.2	Cross-validation results. . . . .	99
4.3	Quantitative evaluation of the proposed method against other text generation methods. . . . .	99
4.4	Ablation results on the way of freezing the text encoder. . . . .	101



# Notations and conventions

## Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CV	Computer Vision
DL	Deep Learning
GPU	Graphical Processing Unit
LLM	Large Language Model
LSTM	Long Short Term Memory
MedIA	Medical Imaging Analysis
ML	Machine Learning
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
SSL	Self-Supervised Learning

## Medical

CT	Computed Tomography
GFR	Glomerular Filtration Rate
MRI	Magnetic Resonance Imaging
US	Ultrasound
sCreat	Serum creatinine

## Metrics

AUC	Area Under the Receiver Operating Characteristic Curve (AUC ROC)
B.Acc	Balanced Accuracy
DSC	Dice Score Coefficient
F1	F1 score
HD95	Hausdorff Distance at percentile 95%
PKVD	Percentage of Kidney Volume Difference
Prec	Precision
Rec	Recall
std	Standard deviation





# Chapter 1

## Introduction

### Contents

---

1.1	Methodological motivations . . . . .	5
1.1.1	Machine learning prerequisites . . . . .	5
1.1.2	The deep learning paradigm . . . . .	8
1.1.3	Self-supervision in computer vision . . . . .	13
1.1.4	The rise of foundation models . . . . .	15
1.1.5	Integrating multimodal learning . . . . .	18
1.1.6	Deep learning for medical imaging . . . . .	19
1.2	Medical context for this thesis . . . . .	23
1.2.1	Renal transplantation . . . . .	23
1.2.2	The role of medical imaging . . . . .	24
1.3	Dataset presentation . . . . .	25
1.3.1	Patients cohort and clinical data . . . . .	25
1.3.2	Medical imaging data . . . . .	29
1.3.3	Translating dataset's limitations to relevant research challenges for data-driven methods . . . . .	32
1.4	Objectives & Contributions . . . . .	34
1.4.1	Objectives of the thesis . . . . .	34
1.4.2	Main contributions . . . . .	35
1.4.3	Scientific production and professional activities . . . . .	36

---





Prompt<sup>1</sup>: "A scene of an introduction to deep learning course in a university's amphitheater with a blackboard and a professor in an old photographic style."

## Summary

In this chapter, we first introduce the methodological motivations to this thesis, going through a rapid overview of machine learning prerequisites and the deep learning trend, then focusing on the recent problematic and research directions that mark today's deep learning literature. Second, we introduce the medical context of this thesis, in which methodological contributions were designed for a specific application: medical imaging monitoring in renal transplantation. Third, we present the dataset of interest of this thesis in order to grasp the objectives and challenges of our work. Finally, we introduce the organization of the next chapters by presenting the objectives and contributions of this thesis.

<sup>1</sup>A PhD thesis in deep learning in 2023 barely discussing generative networks can appear regrettable. So to give a glimpse of their recent advances and capabilities, a cover for each chapter is generated using [Stable Diffusion XL](#), one of the most successful applications of image generation models (along with [Midjourney](#)).



## 1.1 Methodological motivations

### 1.1.1 Machine learning prerequisites

In this section, we introduce the core concepts that are inherent to building Machine Learning (ML) models and that compose most of the methodologies presented in this thesis. ML is a field of Computer Science that consists of developing mathematical and statistical algorithms to solve diverse tasks by optimizing models' parameters using observations, i.e. data. Tasks usually fall into three groups: classification, where models predict the class (category, label) of data samples; regression, where models predict a continuous value variable; and clustering, where models group samples without labels' annotations into clusters according to a certain similarity concept. ML models are *trained* to solve these tasks by *learning*, through an algorithmic or optimization process, from observations available in a *dataset*. When datasets include ground truth annotation to perform the task (e.g. for classification or regression), we use supervised learning. On the contrary, when no labels are available, we use unsupervised learning (e.g. for clustering).

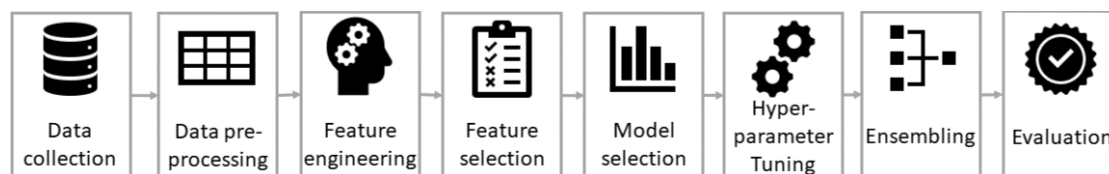


Figure 1.2: **Overview a classical machine learning pipeline.** The main steps of a machine learning model training and evaluation are presented. Different choices for performing each of these steps could be selected and are detailed in this section.

ML pipelines consist of several sequential stages, depending on the specific problem and its complexity. The main components usually are data collection, data preprocessing, feature engineering, feature selection, model selection, hyperparameter tuning, and model evaluation and validation (Figure 1.2). In the next paragraphs, we present the most relevant stages closely aligned with our research.

**Feature engineering.** This first step constitutes a crucial process in transforming raw data into features (variables) that accurately depict the underlying problem [Zheng, 2018]. In essence, it involves the application of domain knowledge to derive analytical representations from raw data, preparing them for utilization in ML pipelines. Feature engineering plays a pivotal role in determining the optimal representation of sample data for effective problem-solving. Effective feature engineering addresses dimensionality and data sparsity issues and contributes to model generalization and robustness. This holds particular significance as the success of artificial intelligence or ML projects often hinges on the quality of data representation. Feature engineering relies on a set of well-defined procedures and methods, the selection of which depends on the specific characteristics of the data. In determining the appropriate procedures, proficiency comes through experience and practical application in a given context. Key steps in the feature engineering process usually include:

- Data Exploration: Understand the characteristics and distribution of the raw data.

- **Encoding Categorical Variables:** Convert categorical variables into numerical format for model compatibility;
- **Scaling and Normalization:** Ensure that numerical features are on a similar scale to prevent certain features from dominating the learning process.
- **Creating Interaction Terms:** Capture relationships between variables by creating new features based on their interactions;
- **Feature Transformation:** Apply mathematical transformations to make data more suitable for modeling, e.g. using logarithmic transformation for variables with skewed distributions;
- **Domain-Specific feature engineering:** Leverage domain knowledge to create more meaningful features for the specific problem. For example, for time series analysis, features based on the spectral analysis of signals are often essential. In Chapter 2, we notably introduce the basis of feature engineering for medical imaging data.

**Feature selection.** One crucial stage is the feature selection procedure [Kotu, 2015b]. Feature selection is the process of choosing a subset of relevant features to improve model performance and reduce complexity. Moreover, the selected features also provide crucial information on the task, as well as an enhanced interpretability of the model. These points become extremely relevant for biomedical applications. Common feature selection methods include:

- **Filter Methods:** These methods select features based on statistical measures like correlation [Pearson, 1895], chi-squared tests [Pearson, 1900], or mutual information [Kraskov, 2004];
- **Wrapper Methods [Guyon, 2002]:** Wrapper methods use a specific ML model to evaluate feature subsets and select the best-performing features. Common techniques include forward selection and backward elimination;
- **Embedded Methods:** Some ML algorithms have built-in feature selection mechanisms. For example, Lasso Regression [Tibshirani, 1996] performs feature selection as it penalizes the absolute values of feature coefficients;
- **Dimensionality Reduction:** Techniques like Principal Component Analysis (PCA) [Hotelling, 1933] and Singular Value Decomposition (SVD) reduce feature dimensionality while preserving as much information as possible.

**Model selection.** We denote "classical machine learning models", common statistical algorithms used for learning tasks, where data are available and we want to build a model to make predictions on new, unseen data. A non-exhaustive list of those models, focusing on the ones that are usually chosen for medical applications, include:

- Linear Regression [Stanton, 2001]: Used for regression tasks, it models the relationship between input features and a continuous target variable by fitting a linear equation;
- Logistic Regression [Cramer, 2002]: For classification tasks, it estimates the probability that a given input belongs to a particular class;
- Decision Trees [Breiman, 1984]: These hierarchical structures make decisions by splitting data based on feature values, leading to a tree-like structure;
- Random Forest [Breiman, 2001]: An ensemble method that combines multiple decision trees to improve predictive accuracy and reduce overfitting;
- Support Vector Machines (SVM) [Cortes, 1995]: Used for both classification and regression, SVMs aim to find a hyperplane that best separates data points;
- K-Nearest Neighbors (KNN) [Cover, 1967]: A simple algorithm that classifies data points based on the majority class among their k-nearest neighbors in feature space;
- Naive Bayes [Hand, 2001]: A probabilistic classifier based on Bayes' theorem with the strong (naive) assumption of conditional independence between every pair of features given the value of the class variable.

**Ensemble models.** Ensemble predictions and ensemble models [Kotu, 2015a] are powerful techniques in ML that involve combining the predictions of multiple base models. Benefits are many-fold by improving: (i) overall predictive accuracy, ensemble methods often outperform individual base models by reducing overfitting and bias; (ii) robustness, ensembles are less sensitive to noise and outliers in the data; (iii) generalization, they tend to have better generalization to unseen data; (iv) interpretability, ensemble models can provide insights into model uncertainty and feature importance. Ensemble methods start with a set of base models, also known as weak learners or base classifiers/regressors. These base models can be of different types or variations of the same algorithm trained on different subsets or with different parameters. Then the predictions of base models are combined in some way to produce a single final prediction. This aggregation can be done through various techniques, such as voting, averaging, or weighted averaging. Examples of existing models integrating this ensemble approach are Random Forest (an ensemble of decision trees) and Adaboost, where the output of the learning algorithms (weak learners) is combined into a weighted sum that represents the final output of the boosted classifier.



**Evaluation.** Model validation is an essential part of ML, whether to select the best model or to assess the performance of a model [Varoquaux, 2023]. Models performances should solely be evaluated on a separated *test set*, usually split from the whole data set available beforehand. Nevertheless, the remaining *train set*, used to learn the model's parameters, can also be split in two to obtain a *validation set*. Split sizes usually correspond to 70-10-20% for the train-validation-test sets. The splitting process is often crucial for small-sized datasets and deeply influences the evaluation. In this context, the stratification of class labels is mandatory, ensuring a balanced distribution of class labels in each split. More recently, discussions surrounding potential biases and fairness in datasets underscore the importance of also considering demographic variables stratification in population datasets. The validation set is often fundamental for monitoring the behavior and the quality of the model's training, optimizing the approach's hyperparameters, or performing feature selection. To go one step further, cross-validation is a technique used to assess the performance of ML models and mitigate issues like overfitting. It consists of dividing the training set into multiple subsets and iteratively training and validating the model on different partitions. Common cross-validation techniques include:

- K-Fold Cross-Validation: The dataset is divided into K subsets (folds). The model is trained on K-1 folds and evaluated on the remaining fold. This process is repeated K times, with each fold serving as the validation set once;
- Stratified K-Fold Cross-Validation: Similar to K-fold, but it ensures that each fold maintains the same class distribution as the original dataset, which is crucial for imbalanced datasets;
- Leave-One-Out Cross-Validation (LOOCV): In LOOCV, each data point is treated as a validation set once while the rest of the data is used for training.

While these notions are non-exhaustive to fully understand ML, we introduced some key concepts to building ML pipelines, and more particularly the ones at the basis of the presented and proposed methodologies in this thesis.

### 1.1.2 The deep learning paradigm

Deep Learning (DL) emerges as a branch of ML algorithms, using models denoted as (deep) neural networks and primarily based on the succession of linear layers and activation functions to provide non-linearity and thus a great expression capability through specific optimization processes. First, we briefly introduce DL history through the main contributions (Figure 1.3) that led to the DL breakthrough near 2012, when those approaches started demonstrating state-of-the-art results on several computer vision tasks. Then, we present more precisely recent DL innovations, focusing on directions that ground and motivate this thesis's work.

### Brief history of deep learning

First neural network models were inspired by the neural network architecture of the human brain [McCulloch, 1943]. Later, Rosenblatt [Rosenblatt, 1958] introduced the basis of Deep Learning architecture: the Perceptron with three layers: input layer, hidden layer, and output layer, now referred to as Multi-Layer Perceptron (MLP) for models with several hidden layers. The first MLP trained by stochastic gradient descent was proposed by Amari [Amari, 1967] to classify non-linearly separable pattern classes. Based on convolution operations instead of linear ones and tailored for vision tasks, the first Convolutional Neural Network (CNN) model trained by backpropagation was introduced by LeCun et al. [LeCun, 1989] and designed to recognize handwritten ZIP code digits.

The DL breakthrough was made possible by the technical innovations and mass-market development of Graphical Processing Units (GPUs). Those highly parallel computing devices were first developed for computer graphics and image processing and then repurposed for calculations involving embarrassingly parallel problems such as deep neural networks' optimization. Thus, the first CNNs to achieve state-of-the-art results in pattern recognition tasks and win different competitions were proposed in 2012, with models such as DanNet [Ciresan, 2012] or AlexNet [Krizhevsky, 2012]. Following those works, DL innovations multiplied in the architecture of models, and strategies to train them, following an exponential increase in both their size and the quantity of data used to train them. Numerous successful applications were designed for various technical fields including Computer Vision (CV), speech recognition, Natural Language Processing (NLP), machine translation, bioinformatics, drug design, Medical Image Analysis (MedIA), environmental sciences, etc. Next, our methodological introduction will focus on CV and MedIA's recent approaches.

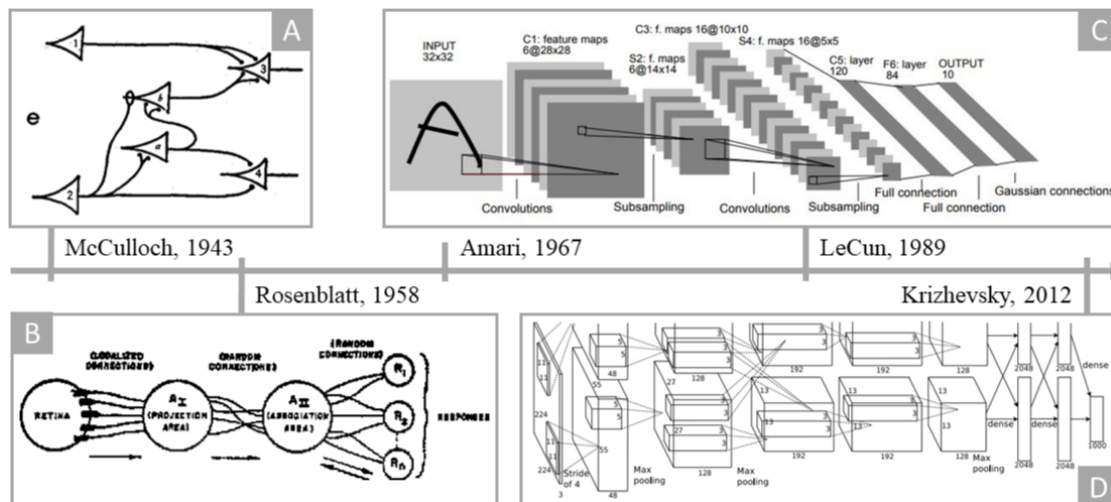


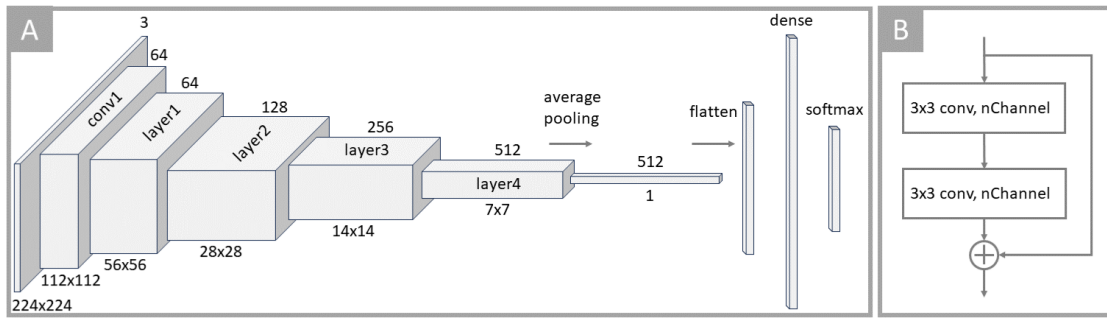
Figure 1.3: **Brief history of Deep Learning timeline.** A: Neural network example from [McCulloch, 1943], B: Organization of the perception from [Rosenblatt, 1958], C: LeNet-5 CNN model overview from [LeCun, 1989], D: AlexNet model overview from [Krizhevsky, 2012].

## Family of approaches & models

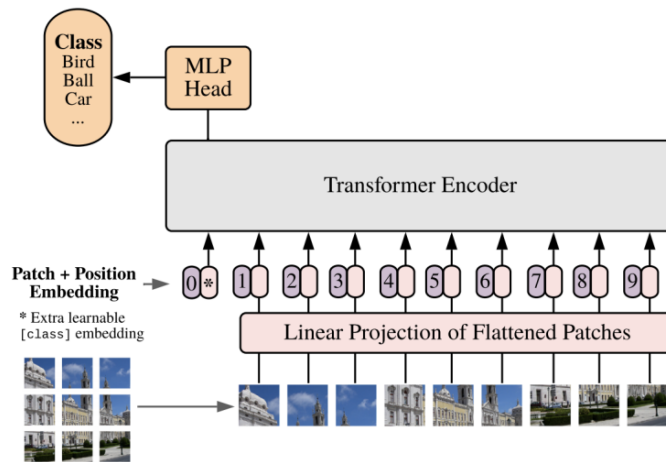
Starting from 2012, the use of deep learning models to solve numerous tasks multiplied. This section introduces the main concepts related to training such models. We present the families of models, the learning paradigm spectrum from supervised to unsupervised learning, models' architectures, and the diversity of downstream tasks. A focus is made on the recent concepts that motivated and ground this thesis, while some other notions, yet fundamental in today's DL literature (e.g. generative models, see last perspectives Section 5.2.4), are only hastily mentioned.

**Discriminative vs. generative models.** Deep learning models can be separated into two main families: discriminative models vs. generative models. The first were designed to separate data points into different classes by learning a decision boundary. More formally, such models learn, by optimizing their parameters, the conditional probability distribution of labels knowing the input (observations, variables, features). Usually, discriminative models are composed of an encoder and a classifier trained end-to-end to perform the desired task. The encoder, also called feature extractor, processes the input to extract a meaningful manifold from the input data, relevant to the task so that the classifier can easily map this representation of the data to labels. This concept of extracting "meaningful features" is crucial as this thesis focuses on producing relevant representations of data, seeking specific characteristics, such as expressiveness, robustness for generalization, or being agnostic to the downstream task. The second family of models was designed to generate new data points by learning probability distributions of the data. Although generative models played an important role in the development of deep learning approaches, notably with Variational Autoencoders (VAEs) [Kingma, 2014a], Generative Adversarial Networks (GANs) [Goodfellow, 2014], and more recently diffusion models [Sohl-Dickstein, 2015; Song, 2019; Ho, 2020], this thesis mainly explored methodologies based on discriminative models (see Chapter 5 for a discussion on perspectives from generative modeling in medical image analysis).

**Learning paradigms.** Several learning paradigms were explored to train those models from data. Between the dominant approaches of supervised and unsupervised learning, different research directions have been explored to form a spectrum of methodologies, notably inspired by real data issues. On the one hand, semi-supervised learning investigates methods that use data in which only a part contains labels or annotations. This set of approaches is explored in scenarios where vast unlabeled data are available, but providing annotations is time-consuming and expensive. On the other hand, weakly-supervised learning explores approaches using noisy, restricted, or inaccurate sources to label data, which are often simpler to supply than hand-labeled data. More recently, self-supervised learning (SSL) has been investigated in CV to produce descriptive and meaningful representations of the training data by defining a pretext or proxy task based on unlabeled inputs. A focus on SSL origins and techniques is presented in Section 1.1.3.



(a) ResNet18. A: Overall ResNet architecture. B: Residual Blocks (2 in each layer for ResNet18).



(b) Vision Transformer (ViT). A sequence of flattened 2D patches obtained from image inputs is fed to a transformer encoder. Image from Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *ICLR*. Curran Associates, Inc., 2020.

Figure 1.4: Model overview of ResNet18 CNN and the Vision Transformer (ViT).

**Models’ architectures.** Models’ architectures also drove and evolved with DL innovations. The main categories are the following:

- **MLPs.** The basic DL classifier is a succession of linear (or dense) layers and activation functions, constituting a block also known as MLP. The activation function provides non-linearity and expressiveness to the models. Usual activation functions are sigmoid, hyperbolic tangent, rectified linear unit (ReLU) [Nair, 2010], leaky ReLU [Maas, 2013], parametric ReLU (PReLU) [He, 2015b]. The expressive power of MLPs, or ability to approximate functions, was studied under certain conditions of depth (number of linear layers) [Barron, 1994; Cybenko, 1989; Funahashi, 1989; Hornik, 1989] or width (dimension’s size of the linear layers) [Lu, 2017]. The universal approximation theorem states that depth-2 networks with suitable activation functions can approximate any continuous function on a compact domain to any desired accuracy. Nevertheless, these models are fundamental as they are often used as building blocks. For example, *linear probing* consists in adding a linear layer, as a classifier, to any feature extractor model in order to evaluate its performance on one or several downstream tasks,

freezing or not (i.e. fine-tuning) the feature extractor model parameters. Linear layers, as well as MLPs, also demonstrated usefulness as *projection heads*, to further embed the feature space and apply some objective function (see Section 1.1.3).

- **CNNs.** For CV, CNNs rapidly became state-of-the-art architectures thanks to their ability and efficiency to extract coherent features from images. The core module of CNNs is the convolutional layer, which applies a parameterized convolution filter to all the input elements. For images, the basic structure combines a series of 2D convolutional, activation function, and max pooling layers to play the role of the feature extractor, followed by fully connected layers, acting as MLP, to perform the classification task, with models such as GoogLeNet [Szegedy, 2014] or the VGG (Visual Geometry Group) family [Simonyan, 2015]. However, these models encounter optimization challenges, specifically experiencing issues like vanishing gradients when delving into deeper architectures. ResNet family [He, 2015a] addresses this issue with residual connections, a particular type of skip connections that combine the signal with a sum and usually skip a few layers (described in Figure 1.4a). On top of that, ResNeXt [Xie, 2016] adds a multi-branch architecture by repeating a building block that aggregates a set of transformations, and DenseNets [Huang, 2016] connects each layer to every other layer in a feed-forward fashion, allowing even deeper architectures. EfficientNets [Tan, 2019] family provides CNN models with better scaling, offering reduced parameters' number while maintaining performances.
- **Sequential models.** Another family of architecture is sequential models, primitively designed to process time series data and improve performances on time series analysis and forecasting tasks, especially when dealing with complex and nonlinear time series patterns. Also known as recurrent neural networks (RNNs) [Rumelhart, 1987], common architectures are Long Short-Term Memory (LSTM) [Hochreiter, 1997] networks and Gated Recurrent Unit (GRU) [Cho, 2014]. They are based on cell and gate mechanisms to keep temporal information and provide a memory component. Nevertheless, such models suffer from backpropagation optimization issues (vanishing or exploding gradients) [Pascanu, 2013] and are able only to model short-term dependencies [Karpathy, 2016]. Natural Language Processing research led to a breakthrough architecture that successfully incorporated the attention mechanism to model sequential dependencies: the Transformer [Vaswani, 2017].
- **Transformers.** The Transformer is part of the sequential models group, but it disrupted DL architectures and deserved a separate presentation. The transformer is based on a noteworthy utilization of the attention mechanism, namely the scaled dot product attention, in a succession of highly expressive layers capable of modeling complex sequential dependencies (detailed in Section 1.1.4, Figure 1.6). Indeed, a successful application of the Transformer architecture to CV was proposed by Dosovitskiy et al. [Dosovitskiy, 2020]. The Vision Transformer (ViT) is given images as an input by making a sequence of flattened 2D patches to a transformer encoder as depicted in Figure 1.4b. In recent years, ViT architectures have shown better scalability and the ability to learn richer features from big amounts of data [Zhai, 2022]. ViT advances investigated knowledge distillation techniques on tokens with DeiT [Touvron, 2020], using hierarchical feature maps in the ViT with Swin Transformers [Liu, 2021], or decoupling the input source of data with latent tokens and cross-attention mechanism with the Perceiver [Jaegle, 2021].

**Downstream tasks.** Deep Learning applications are very wide and even by restricting ourselves to CV tasks, a non-exhaustive list includes: image classification, object recognition, semantic segmentation, pose estimation, image denoising, image super-resolution, and image synthesis. Focusing on the tasks tackled in our studies, image classification is the simplest strategy for extracting semantics from an image and consists of predicting a class from a predefined number of classes, given an input image. Object recognition or detection adds to classification the task of providing the position of the object of interest, usually formalized as the four coordinates of a rectangular bounding box. The finest-grain prediction task for image understanding is semantic segmentation, i.e. predicting for each pixel the class of the object to which it belongs. In some specific contexts of having one class of interest, it can be formulated in binary pixel-level classification with foreground and background classes. For all these tasks, performance can usually be improved with data augmentation techniques [Yaeger, 1996; Simard, 2003; Shorten, 2019]. Common data augmentation techniques include modifying the training samples with hand-designed random transformations that do not change the semantic content of the image, such as cropping, scaling, mirroring, or color changes.

### 1.1.3 Self-supervision in computer vision

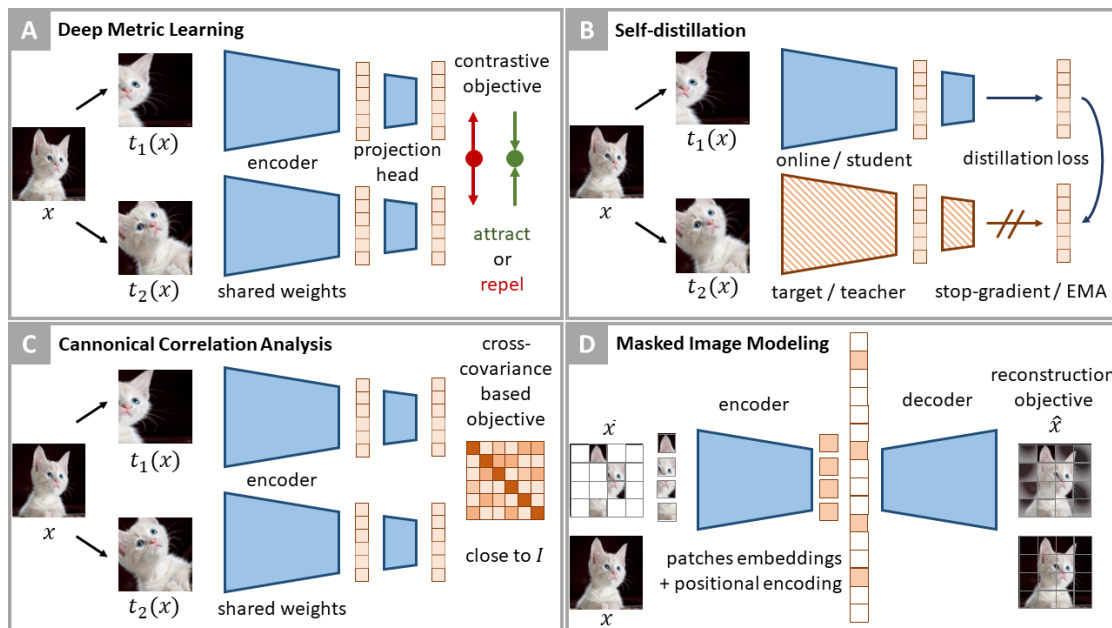


Figure 1.5: **Overview of self-supervised learning (SSL) techniques categories.** A: Deep Metric Learning category: encourage the representations' similarity between transformed versions of an input; B: Self-distillation category: two different views of the same sample are fed to two encoders, and a predictor maps those two views together by distilling the knowledge of one encoder to the other; C: Canonical correlation analysis: infer relationships in the data by analyzing cross-covariance in the batch's samples; D: Masked Image Modeling: mask out portions of an image and teaching a model to inpaint them.

## Learning representations of images

Limited by the availability of labeled data in the supervised learning approaches, self-supervised learning (SSL) methods were primitively designed to learn robust representations from vast unlabeled data [Chen, 2020a; Misra, 2019]. The idea of SSL in CV originates from the development of several methods. Some examples of such tasks include information restoration using colorization [Zhang, 2016], temporal relationships in video [Wang, 2015b], or spatial context in images [Doersch, 2015]. SSL demonstrated particular usefulness in applications such as medicine where labels are costly or the specific task can not be known a priori. Moreover, studies investigated the potential superior robustness of SSL representations to adversarial examples, label corruption, or input perturbations [Hendrycks, 2019; Goyal, 2022]. In practice, learned representations are obtained by the training of a feature extractor encoder and evaluated on different downstream tasks using a simple linear or MLP model, with or without fine-tuning the pretrained model (frozen vs. fine-tuned).

## The different approaches in self-supervised learning

In recent years, SSL has been a field demonstrating growing interest, notably thanks to the availability of large datasets and high-memory GPUs, in addition to the formerly mentioned benefits. To present the main contributions to SSL techniques, approaches can be grouped into four categories [Balestriero, 2023]. An overview of those methods is presented in Figure 1.5.

**Deep Metric Learning.** The first methods were designed to encourage the representations' similarity between transformed versions of an input (Figure 1.5.A). The principle was applied to a learning objective with the idea of a contrastive loss and was referred to as Deep Metric Learning (DML). DML trains a model to predict whether two inputs are from the same class (or not) by making their representation close (or far from each other). Notable DML approaches are SimCLR [Chen, 2020a] where positive pairs are built from data augmentations of the same input and negative samples are other samples in the batch. MeanSHIFT [Koochpayegani, 2021] simply "shift" the embedding of each image to be close to the "mean" of the neighbors of its augmentation. Supervised Contrastive Learning [Khosla, 2020] introduced the use of labels to bag the positive and negative groups on top of SimCLR using an external annotation on the data.

**Self-distillation.** The second category gathers Self-distillation approaches (Figure 1.5.B), where two different views of the same sample are fed to two encoders, and a predictor maps those two views together. A collapse mode [Jing, 2022] can happen by predicting a constant for all inputs, in which all representation vectors cluster at a single point. To prevent collapse, a common approach consists of updating one encoder's weights with a running average of the other encoder's weights. BYOL (Bootstrap Your Own Latent) [Grill, 2020] first introduced self-distillation using image transformations to produce different views of the same image. The online or student network predicts the output and is trained using gradient descent while the target or teacher network produces the target and is updated with an exponential moving average (EMA) update of the student's weights. SimSIAM [Chen, 2020b] replaced the BYOL moving average encoder with a stop-gradient, demonstrating that EMA was not necessary to avoid collapse although leading to a small boost in performance. To avoid sensitivity to the batch size, DINO [Caron, 2021] added a centering of the

output of the student network using a running mean and proposed to discretize the representations by means of a softmax in order to act as an online clustering mechanism from the last layer. Authors also demonstrated that using more than two views, as well as, global and local views, benefits the overall training.

**Canonical Correlation Analysis.** Third, canonical correlation analysis approaches (Figure 1.5.C) aim to infer relationships in the data by analyzing cross-covariance in the batch's samples. Instead of comparing embeddings directly as in DML, SWAV (Swapping Assignments between multiple Views) [Caron, 2020] simultaneously clusters the data while enforcing consistency between cluster assignments produced for different views. BarloTwins [Zbontar, 2021] proposed an objective function that naturally avoids collapse and measures the cross-correlation matrix between the outputs of the two encoders fed with different views of a sample, and making it as close to the identity matrix as possible. VICReg [Bardes, 2021] approach introduced three objectives based on co-variance matrices of representations from two views. Regularizing the variance along each dimension of the representation prevents collapse, the invariance ensures two views are encoded similarly, and the co-variance encourages different dimensions of the representation to capture different features.

**Masked Image Modeling.** Finally, the Masked Image Modeling (MIM) approaches (Figure 1.5.D) are based on the idea of masking out portions of an image and teaching a model to inpaint them, translating the success of Masked Language Modeling (MLM) in Natural Language Processing (NLP) [Devlin, 2018]. The first successful application appeared with BeiT [Bao, 2021] casting the masked token classification task to a regression task to retrieve the masked parts of images. Simplified versions were proposed by directly reconstructing masked image patches rather than discrete image tokens extracted from an encoder with the MAE (masked autoencoders) [He, 2021] and SimMIM [Xie, 2021b] methods.

### 1.1.4 The rise of foundation models

#### Recent advances in natural language processing

Although our work focuses on imaging data, recent deep learning breakthroughs in Natural Language Processing (NLP) influenced CV learning techniques as we have seen in Section 1.1.3 and this trend appears to be gaining momentum. Starting from the breakthrough of the transformer [Vaswani, 2017], first introduced for NLP tasks, the architecture, including a specific attention mechanism (explained in Figure 1.6), rapidly became the basis of all state-of-the-art models for most of the downstream tasks [Tunstall, 2022]. NLP tasks are very diverse, working on both the syntactic and semantic analysis of textual data, a non-exhaustive list being translation, question answering, summarization, and sentiment analysis. Masked Language Model approaches, introduced with the BERT [Devlin, 2018] model, and autoregressive modeling such as in XLNet [Yang, 2019], turned into the key manner to train transformers.



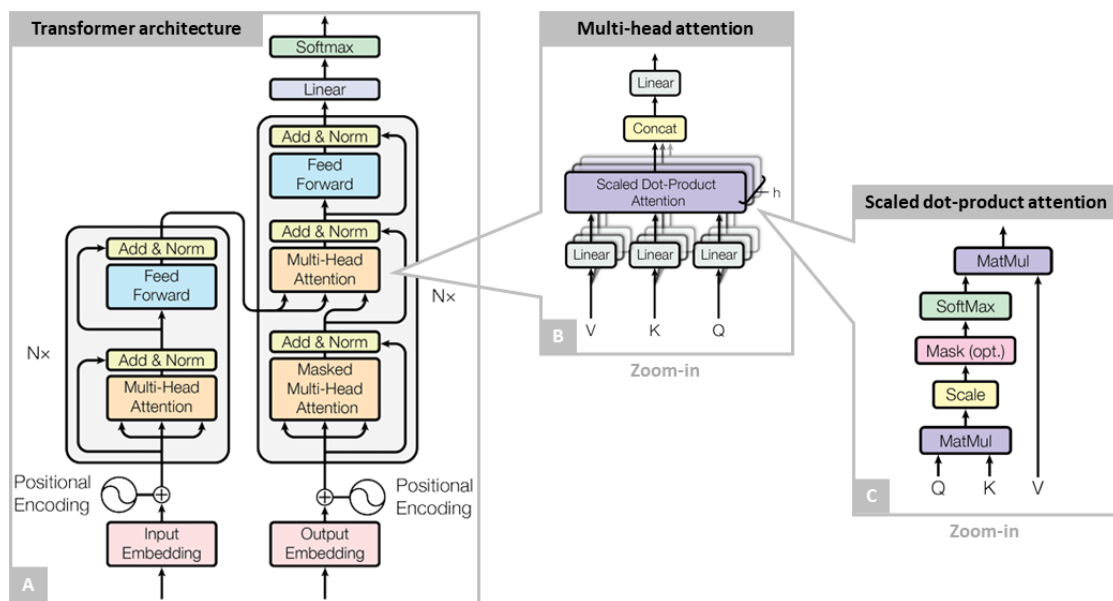


Figure 1.6: **Architecture and core modules of the transformer.** A: The transformer architecture is composed of  $N$  layers of self-attention blocks in the encoder (left) that take as input embeddings added to positional encoding and similarly  $N$  layers in the decoder (right), which in addition performs multi-head attention over the output of the encoder. B: Instead of performing a single attention function, the queries, keys, and values are linearly projected  $h$  (number of heads) times to perform the attention function in parallel. C: This attention mechanism computes the dot products of the query with all keys, divides each by scaling factor, and applies a softmax function to obtain the weights on the values. Images A, B, and C from Ashish Vaswani et al. "Attention Is All You Need". In: *NeurIPS*. Curran Associates, Inc., 2017.

## Large language models

While increasing the amount of large-scale data scrapped from the web and computing resources to train billion-parameter models, performances kept enhancing. Scaling models' size introduced the phenomenon of emerging properties [Wei, 2022a], where larger models (approximately 10+ billion parameters) demonstrated abilities that cannot be predicted simply by extrapolating the performance of smaller models. Notable models are GPT-3 [Brown, 2020], Claude (Anthropic), ChatGPT [OpenAI, 2023], Lambda [Thoppilan, 2022], LLaMa [Touvron, 2023a]. The optimal model size was investigated by [Hoffmann, 2022], denoted as Chinchilla Scaling Law. Recent research aims at having similarly powerful models with reduced size, such as Falcon [Penedo, 2023]. Apart from scalability and model training engineering innovations, the development of those models introduced several important concepts and new methodologies. Model alignment consists of creating agents or models that behave in accordance with what humans want [Kenton, 2021]. The notion of prompting was also introduced as the different ways to input data into those models at training or inference to improve performance. One key idea is the chain-of-thought methodology, where showing steps of reasoning to perform a task to the model improves its ability to solve the task. The "best model" race has just begun with recent models such as PaLM2 [Anil, 2023], Llama 2 [Touvron, 2023b] and impressive numbers such as Falcon-180B, a 180 billion parameters causal decoder-only model and trained on 3,500 billions tokens

of RefinedWeb enhanced with curated corpora. One can note that those models tend to be open-sourced, notably on the [Hugging Face](#) platform.

### Foundation models in computer vision

NLP advances influenced recent CV research in two main ways. First incorporating textual data knowledge and NLP models' capabilities to guide and help the training of CV models, denoted as Vision-Language Modeling (VLM). Radford et al. [Radford, 2021] used multimodal contrastive learning on 400 million image-text pairs from the web to obtain the CLIP model (overview of the approach described in Figure 1.7), demonstrating competitive performances over 30 different computer datasets and numerous tasks such as OCR, action recognition in videos, and many types of fine-grained object classification. ALIGN [Jia, 2021] model scaled up CLIP methodology to leverage a noisy dataset of over one billion image and altered text pairs. Alayrac et al. [Alayrac, 2022] introduced Flamingo, a family of VLM with the ability to ingest images or videos and adapt to novel tasks using only a handful of annotated examples. This research often relies on initiatives to curate large and high-quality datasets such as the LAION project [Schuhmann, 2021], which recently released a 5 billion CLIP-filtered image-text pairs dataset.

Moreover, by trying to reach similar emerging capabilities for imaging tasks, the so-called foundation models aim at learning generic representations rather than specialized embeddings to solve one specific task. An attempt to produce all-purpose visual features was explored with the DINOv2 [Oquab, 2023] model using a diversity of self-supervised techniques and large-scale data curated from diverse sources. To foster research into foundation models for CV, the Segment Anything Model (SAM) [Kirillov, 2023] was trained using an efficient model in a data collection loop, and by building the largest segmentation dataset to date, with over 1 billion masks on 11M licensed and privacy respecting images.

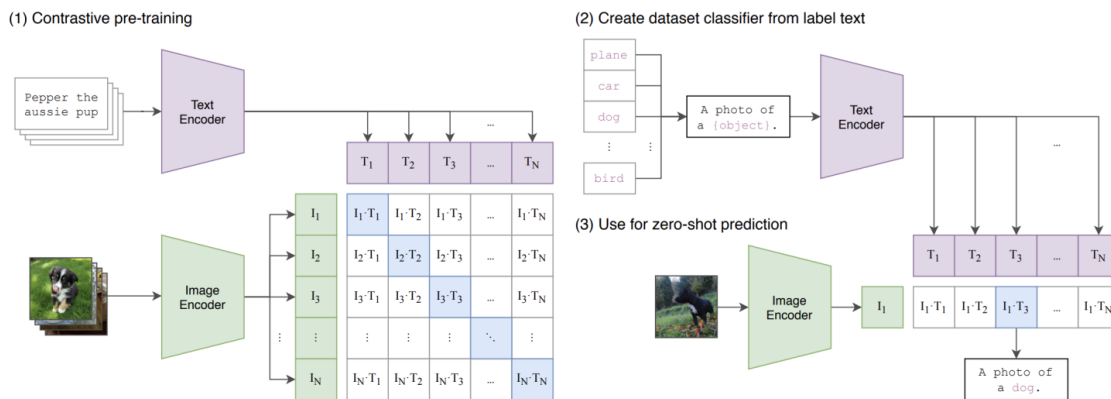


Figure 1.7: **Summary of CLIP approach.** Image and text pairs are given as inputs to their respective image and text encoder, and a multimodal contrastive objective enforces the similarity of the pairs' representations by batch. Image from Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *ICML*. JMLR, 2021.

## 1.1.5 Integrating multimodal learning

### Challenges with multimodal models

Multimodal deep learning combines the strengths of different modalities to create a more complete representation of the data, leading to better performance on various downstream tasks. In general, multimodal architectures consist of three parts: unimodal encoders to represent individual modalities, a fusion network that combines the features extracted from each input modality, and a classifier that accepts the fused data and makes predictions.

Multimodal deep learning aims to solve several core challenges that are active areas of research:

- Representation: feature extraction from heterogeneous data in a way that exploits their synergies is crucial. Moreover, fully exploiting the complementarity of different modalities and not paying attention to redundant information is essential;
- Fusion: the task of joining information from two or more modalities to perform a prediction task. Several techniques have been investigated, namely late fusion by combining the outputs of separate unimodal models, and early fusion by concatenating or merging the raw data. And more recently, cross-modal attention by leveraging attention mechanisms to weigh modalities dynamically (explained in Figure 1.8);
- Alignment: the task of identifying direct relationships between different modalities. Current research in multimodal learning aims to create modality-invariant representations. So when different modalities refer to a similar semantic concept, their representations must be similar/close together in the latent space;
- Translation: the act of mapping one modality to another. The main idea is how one modality (e.g., textual modality) can be translated to another (e.g., visual modalities) while retaining the semantic meaning;
- Co-learning: transferring knowledge learned from one or more modalities to tasks involving other modalities. It becomes especially crucial in scenarios with low-resource target tasks or noisy modalities.

Multimodal models have gained attention with the rise of robust text encoders, as we have seen in the previous section. This approach has enabled significant progress for specific tasks, including visual question answering [Chen, 2022b], text-to-image generation [Ramesh, 2021; Rombach, 2022], and natural language for visual reasoning [Wang, 2023a]. Nevertheless, textual data is not the only type that can bring information to learn better representations of images. Multimodal models are impactful for applications such as autonomous vehicles where models must combine multiple sensor data (image, LiDAR, GPS) for safe navigation. Naturally, the healthcare domain is another significant application of multimodality, where patient data is built from various sources (clinical records, biological, imaging), and integrating tools that are able to gather all this information is crucial, e.g. for disease diagnosis.

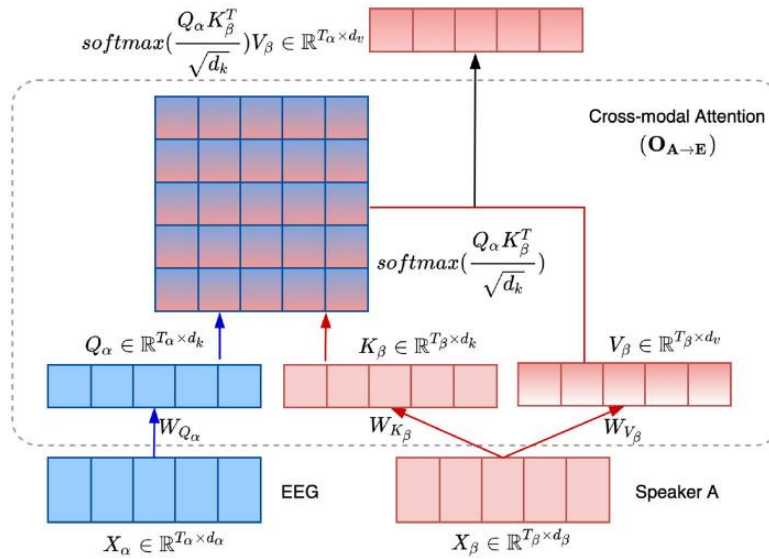


Figure 1.8: **Cross-attention mechanism example.** Multimodal cross-attention is performed by using the scaled dot product attention mechanism from embedding a first modality A (here audio) to the key K and values V, and a second modality B (here EEG) to the queries Q. Image from Siqi Cai et al. “Auditory Attention Detection via Cross-Modal Attention”. In: *Frontiers in Neuroscience* 15 (July 2021).

## 1.1.6 Deep learning for medical imaging

### Challenges and perspectives

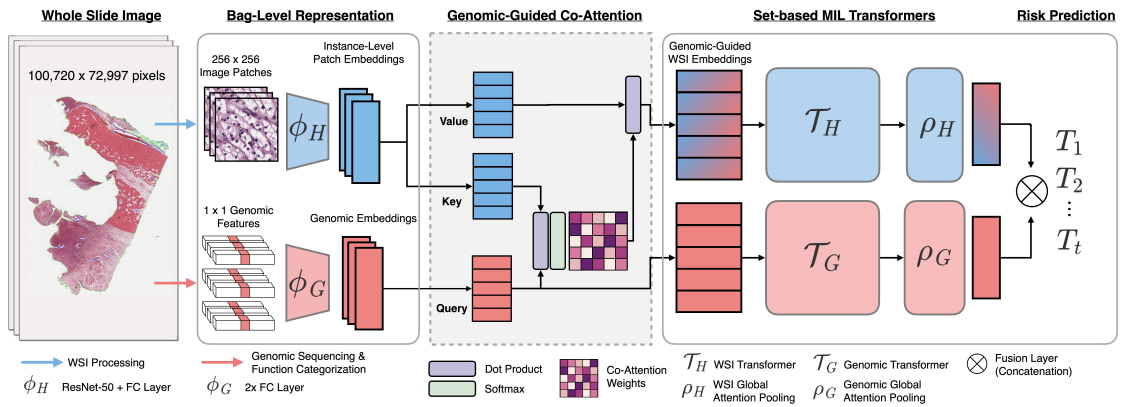
DL in MedIA can appear as a straightforward translation and application of recent CV approaches, but several impactful works show the relevance of this field, such as the U-net [Ronneberger, 2015] architecture, the basis of most of the state-of-the-art studies for segmentation tasks in CV or more recently multimodal contrastive learning to learn robust representations from text-image pairs [Zhang, 2020], prior to the prominent CLIP model [Radford, 2021].

Moreover, the particularities of MedIA data make this domain a flourishing field for innovative methodologies. First, although the mathematical formulation of tasks is similar (classification, detection, segmentation), the objectives (e.g., diagnosis, prognosis, treatment response analysis) go beyond image semantic understanding, with complex biological and medical underlying phenomena. This is, for example, the promise of “radiomics”, computational imaging features discussed in Section 2.3. Second, medical data suffers from several limitations as it is often rare, and expensive to acquire, with a limited amount of samples and labels available, as annotations require medical experts’ knowledge and are time-consuming. Those limitations make it a perfect candidate to develop innovative methodologies in un/weakly/semi-supervised learning. MedIA data are commonly more voluminous, with bigger-sized images (histopathology data), volumetric images such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), and sometimes also longitudinal (follow-up exams). Third, translation to the medical domain makes highly sensitive applications. Thus

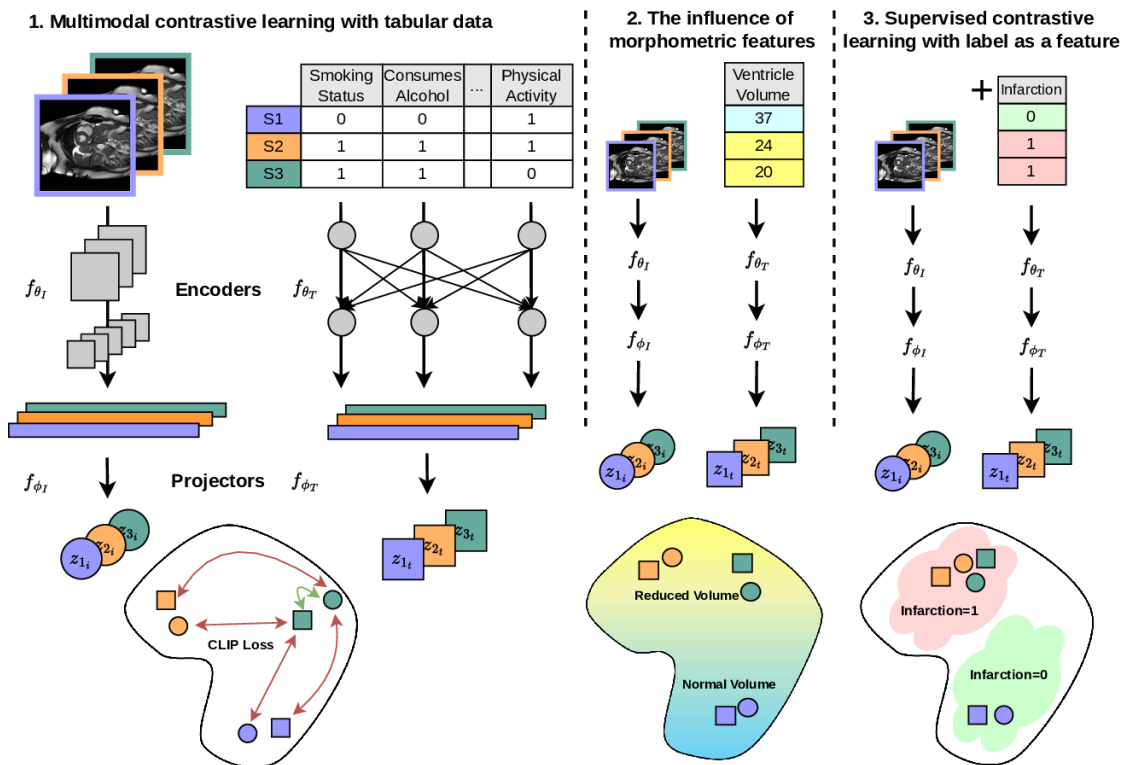
considerations about model calibration [Liu, 2022; Larrazabal, 2023; Murugesan, 2023], and uncertainty estimation of predictions [Belharbi, 2022; Van Waerebeke, 2022] are crucial research directions. Fourth, solving medical-related tasks is not only about achieving state-of-the-art results. Indeed model decision understanding is essential to apprehend our models' ability to solve the desired task, e.g. using and exploring advanced explainability approaches with visual explanations of model predictions [Springenberg, 2015; Selvaraju, 2017; Chattopadhyay, 2018]. Finally, recent studies have shown that AI models can systematically and unfairly be biased against certain populations (e.g. sex, skin color) in multiple scenarios [Zou, 2018] and that MedIA systems make no exception [Ricci Lara, 2022]. Fairness regarding access to healthcare is essential and investigating strategies to mitigate these biases is crucial.

### Self-supervised learning in medical imaging

Having the benefit of learning robust representations from unlabeled data, SSL techniques instantly fit medical applications [Krishnan, 2022]. In MedIA, Azizi et al. [Azizi, 2021] demonstrated that self-supervised learning on ImageNet, followed by additional self-supervised learning on unlabeled domain-specific medical images significantly improves the accuracy of medical image classifiers. Authors also introduced a novel Multi-Instance Contrastive Learning (MICLe) method that uses multiple images of the underlying pathology per patient case, when available, to construct more informative positive pairs for self-supervised learning. Boyd et al. [Boyd, 2021] proposed a generative framework that learns representations of histopathology images' tiles being trained to expand their visual field. This self-supervision proxy task enabled to produce meaningful representations used for different clinical endpoints. Xie et al. [Xie, 2021a] proposed a method to benefit from a wealth of 2D image datasets like chest X-rays as compensation for the lack of 3D data, aiming to build a universal medical self-supervised representation learning framework called UniMiSS. The embedded patches were converted into a sequence regardless of their original dimensions, and the model was trained using self-distillation.



(a) Multimodal Co-Attention Transformer (MCAT) presents a multimodal scheme to combine gigapixel size images with genetic information. Image from Richard J. Chen et al. "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images". In: *ICCV*. IEEE, Inc., 2021, pp. 3995–4005.



(b) Multimodal Contrastive Learning with Tabular and Imaging Data. Image from Paul Hager et al. "Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data". In: *CVPR*. IEEE, Inc., 2023.

Figure 1.9: Overview of two multimodal approaches in Medical Imaging Analysis.

## Multimodality in healthcare

Multimodal deep learning models have significant potential in healthcare. Producing various sources of data is inherent to clinical medicine, such as medical images, electronic health records (EHRs), text-based clinical notes, sensor data, and patient demographics. Straightforward applications are abundant: disease diagnosis and prediction, medical imaging analysis, automatically obtained clinical notes and EHRs, personalized medicine, and drug discovery. Those models also constitute an attempt to mimic the multimodal nature of clinical expert decision-making. Concerning the MedIA field, one can distinguish models with multiple sources of imaging data and models considering another source of data. Indeed, medical imaging itself consists of several modalities such as X-rays, CT, MRI, and Ultrasound (US). The information that captures each imaging system is different and analyzing several at the same time is challenging. In brain MRI data, several MRI sequences as input to models have been exhaustively studied to perform brain tumor segmentation, notably with the BraTS challenge [Menze, 2015; Bakas, 2018]. For multi-imaging modalities models, one main challenge is registration [Baheti, 2021; Weitz, 2023], i.e. ensuring that data are geometrically aligned. Zhang et al. [Zhang, 2020] introduced one of the first text-image models using multimodal contrastive learn-

ing from Chest X-rays images and radiology reports data. Chen et al. [Chen, 2021] proposed Multimodal Co-Attention Transformer (MCAT, overview in Figure 1.9a) to merge whole slide images and genomic data using cross-attention. Hager et al. [Hager, 2023] proposed a multimodal contrastive learning approach to input two-channel medical imaging and clinical (structural) data (overview in Figure 1.9b).

### Foundation models in healthcare

In the medical domain, Med-PaLM [Singhal, 2023a] was the first AI system to surpass the pass mark on USA Medical License Exam (USMLE) style questions and Med-PaLM 2 [Singhal, 2023b] demonstrated a 19% increase on USMLE style questions and investigated the capability of LLMs to encode clinical knowledge. For specific applications, RETFound [Zhou, 2023] is a foundation model that learns generalizable representations from 1.6 million unlabelled retinal images and provides a basis for label-efficient model adaptation for disease detection tasks. OpenPath [Huang, 2023] is a large dataset of 208414 pathology images paired with natural language descriptions from medical Twitter, which was used to pretrain a powerful visual-language foundation model for pathology image analysis. Similarly, CONCH (CONtrastive learning from Captions for Histopathology) [Lu, 2023] is a visual-language foundation model developed using diverse sources of histopathology images and biomedical text (1.17 million image-caption pairs) and evaluated on a suite of 13 diverse benchmarks, achieving state-of-the-art performance on histology image classification, segmentation, captioning, text-to-image, and image-to-text retrieval. For more general purposes, Moor et al. [Moor, 2023] developed through self-supervision on large and diverse datasets, generalist medical AI models (GMAI) to interpret different combinations of medical modalities, including data from imaging, electronic health records, laboratory results, genomics, graphs, or medical text. Models produced expressive outputs such as free-text explanations, spoken recommendations, or image annotations that demonstrate advanced medical reasoning abilities. Similarly, Med-PaLM Multimodal (Med-PaLM M) [Tu, 2023] is a large multimodal generative model that flexibly encodes and interprets biomedical data, including clinical language, imaging, and genomics with the same set of model weights.

Transitioning from these remarkable advancements in foundation healthcare models, exploring their practical implications in clinical settings is imperative. From diagnostic assistance to personalized treatment recommendations, the integration of advanced AI technologies holds the promise of revolutionizing healthcare practices. In the following section, we delve into the medical context of this thesis: monitoring renal transplantation.

## 1.2 Medical context for this thesis

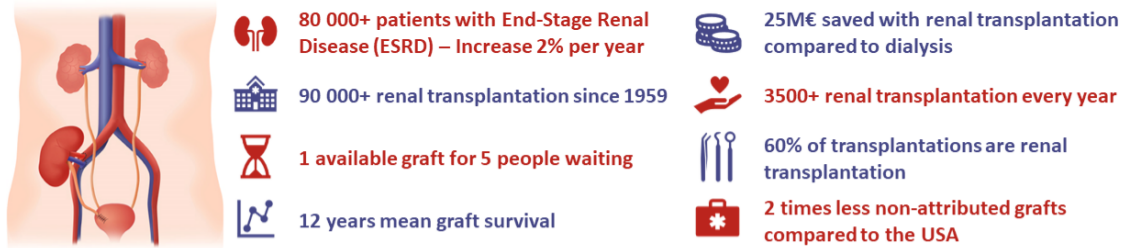


Figure 1.10: **Renal transplantation statistics in France.** Left scheme from CHUV. Statistics from Inserm, Haute Autorité de Santé (HAS), and Agence de la Biomédecine.

### 1.2.1 Renal transplantation

Renal transplantation, also known as kidney transplantation, is a surgical procedure where a healthy kidney is transplanted into a person with kidney failure or end-stage renal disease (ESRD). ESRD is characterized by an irreversible decline in kidney function and is due to various causes, including diabetes, hypertension, chronic glomerulonephritis, and polycystic kidney disease. The renal replacement therapy procedure offers a new lease on life to individuals whose kidneys are no longer functioning effectively [Suthanthiran, 1994]. The two main donor sources are a living donor – a healthy person, usually a family member or close friend, who donates one of their kidneys, and a deceased donor – a kidney is obtained from a deceased individual (cadaveric donor) who has previously consented to organ donation. Both the donor and recipient undergo extensive medical and psychological evaluations to ensure compatibility. Compatibility is determined based on blood type, tissue matching, and other factors to reduce the risk of rejection. The renal transplantation surgery typically lasts 3-4 hours. The donor's kidney is removed, and the recipient's damaged kidney(s) may or may not be removed, depending on the situation. The transplanted kidney is placed in the lower abdomen and connected to the recipient's blood vessels and bladder. Post-transplant medications are principally immunosuppressive drugs, that are prescribed to prevent the recipient's immune system from rejecting the new kidney. Figure 1.10 reports some general renal transplantation statistics in France.

Patients require close monitoring and follow-up care to ensure the success of the transplant. Regular blood tests and check-ups are necessary to detect and manage any potential complications. Various complications can arise in renal transplants, including urologic disorders, vascular complications, and immunologic adverse effects. Medical imaging plays a crucial role, and various imaging techniques have been explored for monitoring renal transplants [Sharfuddin, 2014]. Clinicians can access several imaging modalities, such as ultrasonography (US), nuclear medicine studies, CT, and MRI. However, these examinations primarily provide clinicians with relevant information concerning the current status of the graft and the gold-standard method to specify and investigate graft status is needle biopsy, an invasive surgical operation. Moreover, a risk of chronic



dysfunction of the transplanted kidney remains, which can result in the loss of the graft or, ultimately, the patient's death [Hariharan, 2021].

## 1.2.2 The role of medical imaging

### Modern radiology in the era of artificial intelligence

In Section 1.1.6, we have introduced the methodological challenges and perspectives of Deep Learning in Medical Imaging from our computer scientist point of view. Now, let us try to discuss the implications of Artificial Intelligence (AI) from a radiologist, or clinician view. AI algorithms have shown tremendous potential in various medical applications, especially imaging, due to their ability to extract relevant information from medical data. The key aspects of modern radiology in the AI era are the following:

- Automated image interpretation: Computer-aided detection (CAD) and computer-aided diagnosis (CADx) systems help radiologists detect abnormalities and provide diagnostic insights;
- Faster diagnosis and treatment: AI-powered software can process images in a fraction of the time it takes for a human radiologist. This acceleration in diagnosis enables quicker initiation of treatment and improved patient outcomes, especially in critical conditions;
- Image enhancement and reconstruction: algorithms can enhance the quality of medical images by reducing noise, improving resolution, and enhancing contrast. This can be especially beneficial in cases where image quality is suboptimal due to patient movement or limited imaging capabilities;
- Predictive analytics: AI can analyze vast amounts of patient data and images to predict disease progression, treatment response, and potential complications;
- Workflow optimization: AI-driven tools can help radiologists manage their workflow more efficiently by prioritizing critical cases and automating routine tasks;
- Personalized medicine: AI enables the customization of treatment plans based on a patient's unique characteristics and response to therapy, also denoted as precision medicine and essential in fields such as oncology;
- Quality assurance: AI can assist in quality control by identifying discrepancies or anomalies in medical images and flagging them for review;
- Data integration: AI facilitates the integration of radiology data with electronic health records (EHRs) and other clinical data sources, providing a comprehensive view of the patient's health.

More specifically, we can observe how AI is pouring into impactful clinical studies. Esteva et al. [Esteva, 2017] were among the first to propose an artificial intelligence system (based on CNNs) capable of performing a medical imaging analysis task with a level of competence comparable to medical experts, on the application of the classifying skin cancer. Sun et al. [Sun, 2018] demonstrated that automatically extracted features using a radiomics approach in CT images could be robust biomarkers for assessing tumor-infiltrating CD8 cells from RNA-seq genomic data. Sammut et al. [Sammut, 2021] collected clinical, digital pathology, genomic and transcriptomic profiles of pre-treatment biopsies of breast tumors to build a ML predictor of breast cancer therapy response.

### Medical imaging in renal transplantation monitoring

Kidney transplant dysfunction is primarily indicated by a rising serum creatinine level, decreasing urine output, and pain and tenderness over the graft, but often it is clinically asymptomatic and presents only with an isolated increase in serum creatinine. In opposition to solid biopsy, most imaging techniques are non-invasive and thus appear as highly desired tools, used and needed by the transplant community. Imaging modalities enable radiology experts to primarily evaluate functional information such as transplant perfusion with gadolinium-based DCE MRI [Loubeyre, 1996] or the resistance index (RI) value [Ghonge, 2021] to diagnose transplant renal artery stenosis.

Nevertheless, computational imaging algorithms, and more particularly AI-based ones, offer the potential to provide models that go one step further, providing tools to retrieve structural or molecular information or to discover new biomarkers. Becker et al. [Becker, 2020], in their review on AI in nephropathology, anticipate the gradual integration of such technologies into routine diagnostics and research. Furthermore, Yang et al. [Yang, 2016] recently proposed a deep-learning-based approach to detect and quantify pathological lesions in post-transplant biopsies, showcasing its superior ability to predict post-transplant graft loss. This approach holds potential applications in prevention, risk stratification, and monitoring tools. Lastly, Yoo et al. [Yoo, 2023] presented an automatic decision-support system that assigns kidney allograft diagnoses, extending the international Banff classification system [Roufosse, 2018] and enabling to improve risk stratification of long-term allograft outcomes.

## 1.3 Dataset presentation

In this section, we describe our main private dataset. For this thesis, a unique dataset has been collected and used, as no other publicly available dataset exists for this task. At the same time, we will present the main aspects of this specific application so that the reader can grasp the particularities and challenges of this application that will be at stake in the next chapters.

### 1.3.1 Patients cohort and clinical data

#### Patients' cohort

The data cohort used in this thesis is based on the clinical trial *Imag-NCT*, which underwent review and approval by the Institutional Review Board and waived the requirement for obtaining patients' consent. The private dataset corresponds to the reference ID-RCB 2012-A01070-43 and is registered on ClinicalTrials.gov under the identifier NCT02201537. 136 patients subject to renal transplantation between 2013 and 2015 at Necker Hospital and Georges-Pompidou European Hospital (HEGP) were eligible. Inclusion, non-inclusion, and exclusion criteria are presented in Figure 1.11. The follow-up lasted up to July 2019, or until the withdrawal from the study (e.g. will of the patient), the transplantation failure (return to dialysis), or the patient's death.

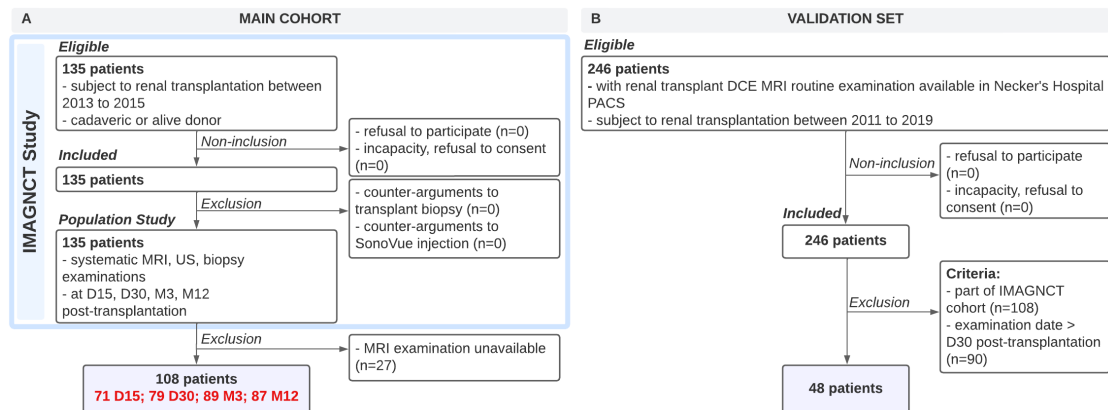


Figure 1.11: **Clinical study flowchart.** A: Our main data cohort of 135 patients subject to renal transplantation from 2013 to 2015 was obtained from **Imag-NCT** study. Systematic follow-up MRI, US, and biopsy examinations were performed on day 15 (D15), day 30 (D30), month 3 (M3), and month 12 (M12). B: A validation set was obtained from routine first-month MRI examinations from 48 kidney transplant recipients from 2011 to 2019 to further validate results.

The data collected corresponds to clinical variables, such as demographics, transplant allocation, surgery complication, and post-transplantation complication variables. Moreover, resulting from blood or urinary tests, biological variables (serum creatinine, urinary creatinine, and proteinurias) were collected regularly from a few days before the transplantation to the last follow-up. And the core of our interest, several medical imaging examinations were performed at specific timestamps, as well as a transplant biopsy, having the Banff classification variables.

To further validate some of our results during the thesis, a validation set of 138 was extracted from the Necker Hospital data system (PACS) from routine MRI examinations (Figure 1.11 right) up to one year and two months post-transplantation to match our main cohort examinations' timestamps range. The patients part of the Imag-NCT cohort were excluded. The validation cohort underwent review and approval by the Institutional Review Board (CRM-2308-367). Only a few patients had several examinations (for the longitudinal analysis) and to evaluate our tasks, available serum creatinine levels, complication events, and status of the graft 5 years post-transplantation were also collected. To form a sub-group of 48 early exams, the examinations with a date bigger than thirty days post-transplantation were also excluded in a second phase.

### Clinical attributes

Clinical variables collected in the study were divided into two groups. Clinicobiological variables correspond to demographics, inclusion, non-inclusion criteria, transplant allocation (e.g. donor's age, transplant incompatibility), transplant surgery (e.g. cold ischemic time, surgery complications), and unclamping biopsy variables [Wang, 2015a]. Progress variables correspond to complication events that occurred post-transplantation (e.g. acute rejection, stenosis). Table 1.1 reports information and statistics of some of these variables in our main cohort and in the supplementary validation set, curated in a second phase. The transplant incompatibility variable summarizes (integral values from 0 to 6) the factors that may hinder the successful transplantation of a transplanted kidney into the recipient's body. These factors can lead to an immune response against the transplanted organ, potentially resulting in rejection. The cold ischemic time corresponds to the period of time when the organ is preserved in a hypothermic state prior to transplantation into the recipient; an acceptable cold ischemic time for renal transplantation is less than 20 hours (1200 minutes).

Table 1.1: Patient characteristics and clinical attributes.

Variable	Main cohort (mean $\pm$ std or counts)	Validation set (mean $\pm$ std or counts)
Sex	Male: 82; Female: 46	Male: 30; Female: 18
Age at transplantation (Years (y))	51 $\pm$ 14	54 $\pm$ 13
Height (Meters (cm))	169 $\pm$ 9	170 $\pm$ 11
Weight (Kilogram (kg))	72 $\pm$ 16	72 $\pm$ 14
BMI (kg.m-2)	25 $\pm$ 5	25 $\pm$ 4
Previous transplants	0: 99; 1: 22; 2: 4; 3: 2	0: 40; 1: 7; 2: 1
Transplant type	Alive donor: 33; Cadaveric: 95	Alive donor: 11; Cadaveric: 37
Transplant side	Right: 101; Left: 27	Right: 28; Left: 20
Initial nephropathy	Glomerulopathy: 23	Glomerulopathy: 12
	Diabetes: 10	Diabetes: 2
	Cystic / hereditary / congenital disease: 38	Cystic / hereditary / congenital disease: 11
	Secondary glomerulopathy: 8	Hypertension: 6
	Hypertension: 9	Interstitial nephropathy: 5
Transplant incompatibility (0-6)	Interstitial nephropathy: 3	Unknown cause: 12
	Other: 11	
	Unknown cause: 26	
Donor's age (Years (y))	0: 9; 1: 2; 2: 9; 3: 26; 4: 38; 5: 26; 6: 18	0: 3; 1: 3; 2: 9; 3: 8; 4: 10; 5: 9; 6: 6
Cold ischemic time (Minutes (min))	54 $\pm$ 15	55 $\pm$ 14
Surgery complications	1068 $\pm$ 637	927 $\pm$ 667
	No: 115; Yes: 13	No: 20; Yes: 28

*BMI: Body Mass Index; std: standard deviation. Statistics with missing data*

### Biological data

In clinical practice, creatinine [Kasiske, 1989; First, 2003] and proteinuria [Cherukuri, 2010; Ponticelli, 2012] levels are used as primary indicator of renal function, measuring wastes kidneys are supposed to filter, and easily accessible as they result from blood (serum creatinine) or urinary (urinary creatinine, proteinuria) tests. Figure 1.12 presents serum creatinine, urinary creatinine, and proteinuria levels for different patients. Using the creatinine level with age, gender, and other factors, clinicians compute an estimation of the glomerular filtration rate (eGFR). There are different equations for estimating GFR, with the Modification of Diet in Renal Disease (MDRD) [Levey, 1999] and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) [Levey, 2009] equations being commonly used. The glomerular filtration rate (GFR) is a measure of the rate at which the glomeruli in the kidneys filter waste and excess substances from the blood to form urine. The glomeruli are tiny structures within the kidneys responsible for this vital function. Increasing and unstable variations are the two main criteria for monitoring these biological levels. Nevertheless, creatinine or eGFR is considered a late biomarker, as well as a renal biopsy is often needed to investigate the reason for its fluctuation. Peaks and high fluctuations correspond to complication events occurring in the transplant, such as acute rejection events, and have to be carefully monitored in order to rapidly adapt the patient's treatment or complete further investigations (usually renal biopsy).

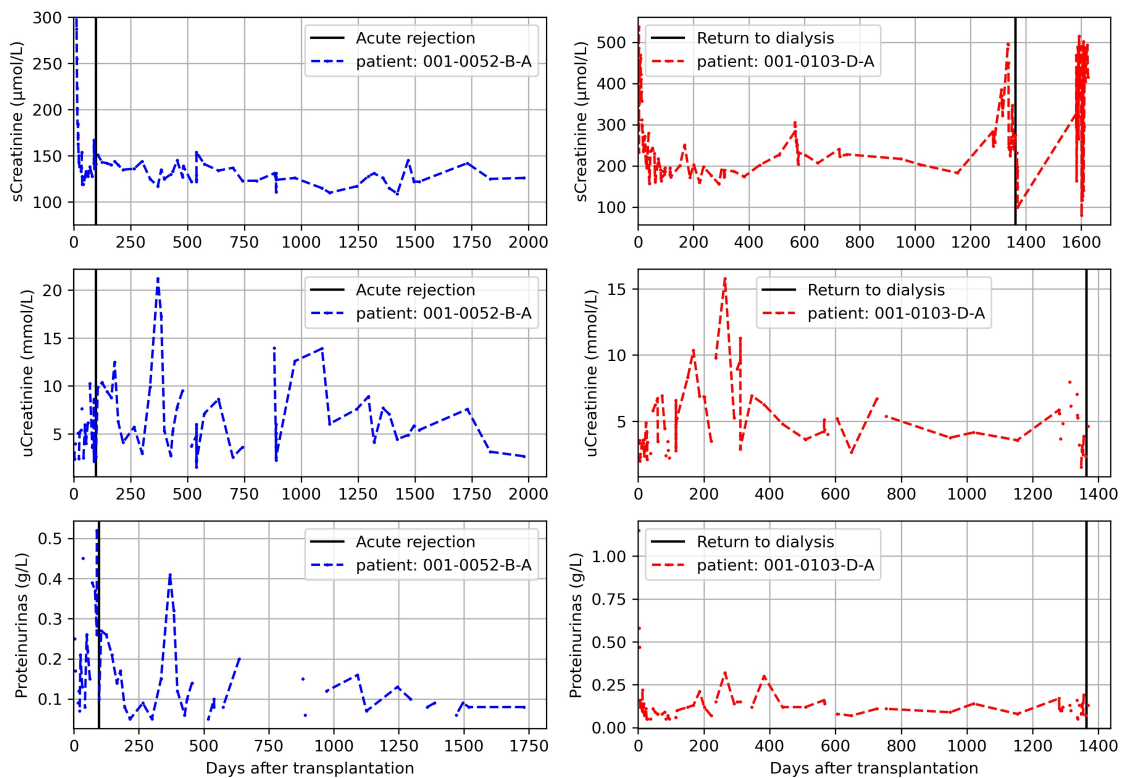


Figure 1.12: **Biological data examples.** Plot examples of the serum creatinine (sCreatinine in  $\mu\text{mol.L}^{-1}$ , line 1), urinary creatinine (uCreatinine in  $\text{mmol.L}^{-1}$ , line 2), and proteinurinas (in  $\text{g.L}^{-1}$ , line 3) for two patients with (left) and without (right) transplantation failure event.

### 1.3.2 Medical imaging data

#### Follow-up examinations

The goal of this thesis is to better integrate medical imaging data into the analysis of the follow-up of the patients, aiming to explore its capability to bring information toward renal transplantation monitoring and events. The medical imaging examinations are already integrated into the clinical flow to detect and monitor specific events. In this clinical study, examinations were systematic and obtained at specific follow-up timestamps post-transplantation, namely fifteen days (D15), thirteen days (D30), three months (M3), and twelve months (M12) post-transplantation. At each follow-up, patients underwent MRI, US examinations, and renal biopsy. In clinical practice, renal biopsy is the gold standard method to make a diagnosis on complication events of transplants [Serón, 2007]. The extracted tissues are analyzed by microscope to obtain histopathological images (examples in Figure 1.13). Nephropathology experts analyze these images using the Banff classification scores [Roufosse, 2018] to further make a diagnosis on the state of the transplant. In our study, only these classification variables were collected as results of the systematic biopsies. However, this surgical procedure is invasive and thus is not favorable as it is time-consuming, expensive, and frequently causes injurious events such as bleeding or inflections.

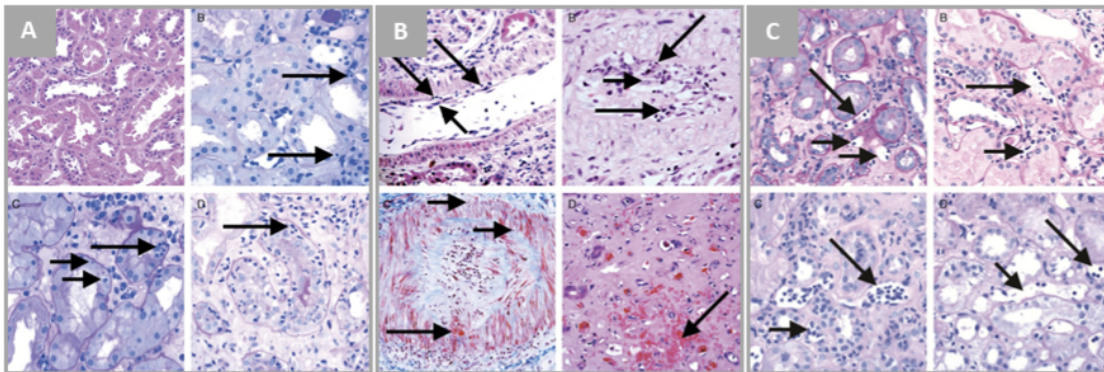


Figure 1.13: **Histopathological images results of transplant biopsies examples.** Solid renal biopsies are performed to extract transplant tissue samples. Then nephropathology experts analyse those microscopic images and provide the Banff lesion scores (0-3 value). Sample examples are provided for A: lesion score t (tubulitis), B: lesion score v (intimal arteritis), and C: lesion score ptc (peritubular capillaritis). Images from Candice Roufosse et al. "A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology". In: *Transplantation* 102.11 (Nov. 2018), pp. 1795–1814.

### MRI modalities

For the monitoring of kidney transplants, several MRI sequences are investigated. In this study, three sequences were acquired, namely Dynamic-Contrast Enhanced (DCE), Diffusion-weighted (DW), and Blood Oxygenation Level Dependent (BOLD). The two first are used in routine examinations while the third is rare in the context of renal transplantation. Examples of images for each sequence are shown in Figure 1.14).

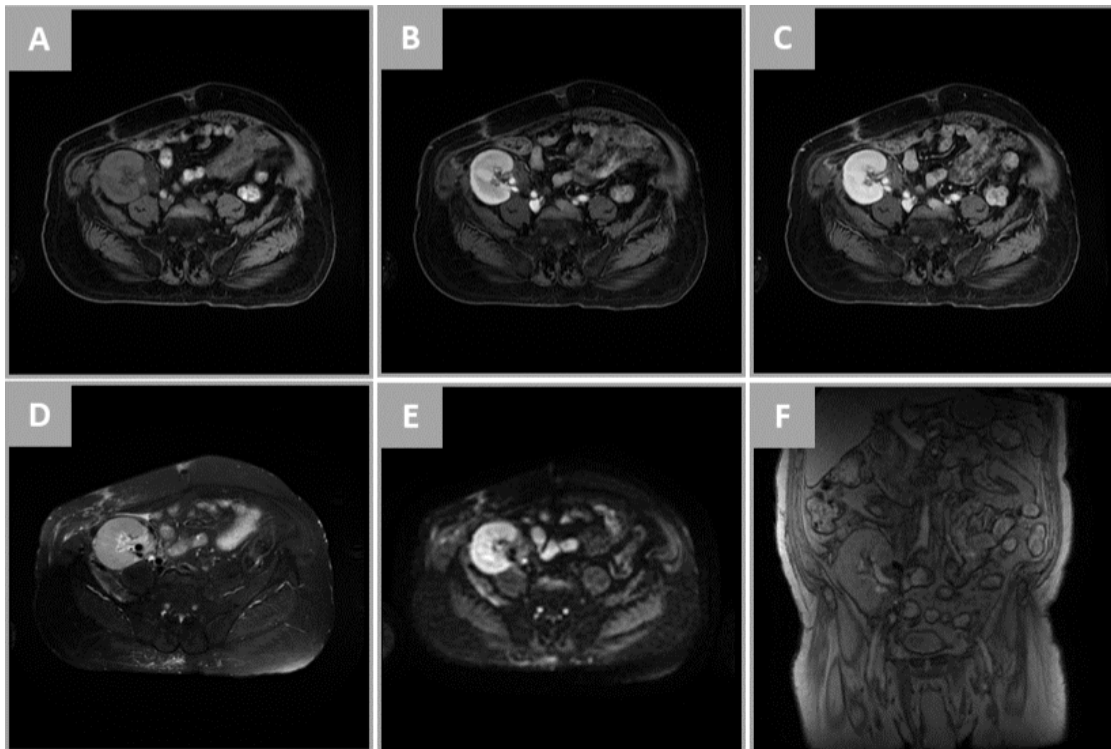


Figure 1.14: **MRI data slice examples of available modalities.** A: Axial DCE MRI without injection (T1 acquisition); B: Axial DCE MRI at arterial time; C: Axial DCE MRI at tubular time; D: Axial T2 acquisition; E: Axial Diffusion MRI; F: Coronal BOLD MRI.

**DCE MRI** is a dynamic sequence and includes time-series acquisition while the contrast agent flows into the organ. Clinicians usually adopt a gadolinium chelate that is not nephrotoxic. However, in cases of  $GFR < 30 \text{ ml} \cdot \text{min}^{-1} \text{ per } 1.73 \text{ m}^2$ , gadolinium-based contrast agents cannot be used because of their risk of causing nephrogenic systemic fibrosis. Two main times are noted by radiology technologists: the arterial time where the arterial part of the kidney is enhanced, and the tubular time where the whole organ is enhanced. DCE MRI allows for the assessment of blood flow within the transplanted kidney. This is crucial in monitoring the immediate post-transplant period to ensure that the newly transplanted kidney is receiving adequate blood supply. Poor blood flow can be a sign of complications such as vascular issues or acute rejection. DCE MRI can also help in the early detection of complications such as acute rejection, renal artery stenosis, and vascular thrombosis.



**Diffusion-weighted MRI** is obtained by varying the b-value (acquisition parameters) to modify the molecular diffusion in tissues (generally water). In renal tissues, this movement, known as diffusion, is sensitive to the microstructural integrity of the renal parenchyma. By assessing diffusion characteristics, DW MRI can help evaluate renal function and detect changes in the renal tissue, such as injury or disease. Moreover, DW MRI can be particularly sensitive to early ischemic injury and cellular damage, which are common in the early post-transplant period. It can aid in the early detection of complications such as acute rejection or ischemia-reperfusion injury, allowing for prompt intervention.

**BOLD MRI** is based on the different magnetic properties of hemoglobin, in its oxygenated and deoxygenated form. Thus, BOLD MRI can assess the oxygenation levels in the renal tissue. This is particularly important because adequate oxygen supply is crucial for the normal functioning of the transplanted kidney. BOLD MRI could detect changes in renal oxygenation at an early stage, potentially allowing for the early identification of complications. For example, a decrease in oxygenation levels can be an early sign of ischemia-reperfusion injury or acute rejection, allowing for timely intervention to protect the transplanted kidney.

The dataset used for this thesis has been mainly acquired by an MR450w GEM machine (General Electric Healthcare, Chicago, IL) was used to acquire a  $1.5T$  MRI scan. The standardized protocol included an axial gradient-echo T1-weighted sequence, reconstructed as in-phase, out-of-phase, water, and fat-suppression images before and after an intravenous bolus of  $0.1\text{mmol.kg}^{-1}$  ( $0.5\text{mmol.mL}^{-1}$ ) of gadoteric acid (Dotarem®, Guerbet, Villepinte, France) at a rate of  $2\text{mL.s}^{-1}$ , followed by a  $25\text{mL}$  saline flush; and axial and coronal, fat-suppressed, T2-weighted, diffusion-weighted (DW) sequences with b-values of 0, 400, and  $1200\text{s.mm}^2$ ; and Blood oxygenation level-dependent (BOLD) sequences.

### Ultrasound modality

Ultrasound imaging (or sonography, denoted as US) is commonly used in the setting of renal transplantation. This technology allows its use at numerous times post-transplantation: in the first few hours or days post-transplantation, in routine surveillance, and notably when there is a need for evaluation of graft dysfunction. Ultrasound is convenient, rapid, and relatively cheap. US is performed in real-time, provides multi-planar views, is non-invasive, and does not involve ionizing radiation, nor does it need iodinated contrast, thus avoiding the further risk of contrast nephropathy.

US evaluation of the renal allograft requires the use of B-mode, Color Doppler and spectral Doppler technologies [Baxter, 2001]. B-mode ultrasound provides a morphologic evaluation, while color and spectral Doppler assesses blood flow. Newer applications such as Contrast-Enhanced Ultrasound (CEUS) [Schwenger, 2006; Chung, 2014; Kazmierski, 2020] and Shear Wave Elastography (SWE) [Grenier, 2012; Grenier, 2013] are being investigated in the renal transplant setting. On the one hand, CEUS provides a useful technique for improved quantitative analysis of kidney allograft perfusion, early prediction of chronic allograft nephropathy, and loss of graft function. On the other hand, SWE helps in the evaluation of renal fibrosis, which can occur in



response to various factors, including chronic kidney disease and transplant rejection. This technique provides a quantitative measurement of tissue stiffness, aiding in the assessment of graft health. One can also note that ultrasound imaging is frequently used to guide renal biopsies, especially when performing percutaneous (through the skin) kidney biopsies. Ultrasound guidance ensures that the biopsy needle is accurately placed within the kidney, increasing the safety and precision of the procedure. Images examples are presented in Figure 1.15.

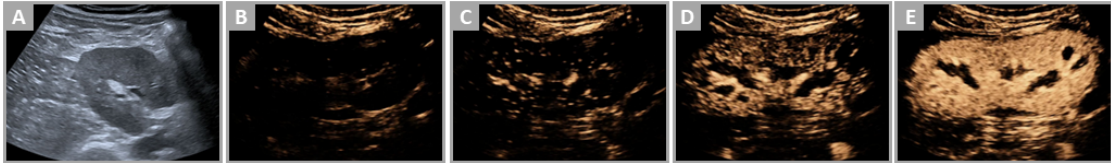


Figure 1.15: **US data examples of renal transplants.** A: US B-mode image; B, C, D, E: Successive frames of CEUS video.

### 1.3.3 Translating dataset's limitations to relevant research challenges for data-driven methods

This section provides a succinct overview of the limitations inherent in medical applications and datasets, as elaborated in Section 1.1.6. These limitations, particularly pertinent to the specific issue addressed in this thesis (outlined in Section 1.3), have significantly influenced the trajectory of our research directions. Consequently, these challenges have driven the formulation of innovative methodologies tailored to effectively address and overcome those challenges.

#### A limited amount of imbalanced data

Our primary dataset comprises a modest number of 136 patients, a quantity that falls significantly short of the ideal data volume required for ML algorithms, a challenge exacerbated when considering specific sub-groups, such as the 108 patients with at least one MRI available. This circumstance presents a substantial challenge when adopting the conventional supervised learning framework, where the dataset consists of a mere 108 pairs of inputs (e.g., MRI volumes) and labels (e.g., transplantation outcomes). This challenge is particularly pronounced in the context of deep learning techniques, which typically demand substantial volumes of data for optimal performance. Furthermore, the dataset, being derived from real-world clinical records, exhibits significant class imbalance, a characteristic that is especially prevalent in the context of systematic examinations, resulting in a preponderance of cases with favorable outcomes. Nevertheless, the incorporation of prior information holds substantial value, as detailed in subsequent sections, offering insights into how we can enrich the feature learning process. It is worth noting that when considering the Dynamic Contrast-Enhanced (DCE) MRI examinations solely, these data include time series acquisitions featuring between five and nine MRI volumes per examination, thereby supplementing our dataset with additional information for our learning algorithms.

### Longitudinal analysis

One noteworthy and critical aspect of our dataset is the inclusion of follow-up examinations, a form of longitudinal data. This component offers valuable opportunities to enhance our data modeling capabilities by incorporating the temporal dimension, allowing us to monitor the progression of data and underlying biomedical events. Longitudinal data plays a pivotal role in medical applications due to its relevance in capturing the evolving nature of pathologies. It serves as a key element in the decision-making process for medical professionals and is an essential input for predictive analyses. The analysis of longitudinal imaging data has been extensively explored in fields like neuroimaging [Couronné, 2021; Ren, 2022; Ouyang, 2022a; Ouyang, 2022b], aiding in the understanding of diverse pathology trajectories inter- and intra-patients. Moreover, recent developments in self-supervised approaches have shown promise in utilizing longitudinal data for further investigation [Zhao, 2021; Ouyang, 2021].

### Missing data

While not explicitly addressed in the dataset description, our dataset is notably affected by a substantial amount of missing data, particularly concerning the imaging data, as detailed in Table 1.2 for the Dynamic Contrast-Enhanced (DCE) MRI examinations. This is especially prevalent in the medical imaging segment, where patients may have missed follow-up appointments for various reasons. Effectively managing missing data is of utmost importance, particularly in real-world data scenarios, such as those encountered in medical applications. Data imputation methods offer solutions for filling in these gaps using statistical or generative techniques. For unidimensional data like clinical variables (e.g., demographic or comorbidity variables), conventional statistical approaches may suffice. However, when dealing with high-dimensional data, such as 3D MRI (e.g.,  $512 \times 512 \times 88$  vectors), employing simple interpolation techniques or generative models to create synthetic data from learned representations becomes notably critical.

Table 1.2: **Missing data in the DCE MRI follow-up examinations.** We summarize the counts of available DCE MRI examinations for the different follow-up examinations (D15, D30, M3, and M12), and the patients having  $n+$  examinations ( $n = 2, 3, 4$ ) available (Exam(s)). Our main cohort includes 136 patients and the missing value is calculated in percentage relative to the number of patients having at least one DCE MRI examination available (108).

Exam(s)	D15	D30	M3	M12	2+ Exams	3+ Exams	4 Exams
<b>Total Number</b>	71	79	89	87	70	63	44
<b>Missing (%)</b>	34	27	18	19	35	42	59

*D15: day fifteen, D30: day thirteen, M3: month three, M12: month twelve.*

## Multimodal learning

Our dataset encompasses a variety of data sources, primarily comprising three distinct modalities: structural or tabular data (clinical variables), temporal series (biological data), and imaging data (MRI and ultrasound). Many research efforts have explored the potential of addressing different tasks (such as detection and prediction) separately for each modality to gain a deeper understanding of the unique insights they offer. However, recent research trends aim to introduce models or methodologies capable of effectively handling diverse modalities as inputs. This presents a host of challenges, as each data source contributes varying amounts of information. The integration of multiple data modalities is of paramount importance in the context of renal transplantation monitoring. By combining clinical, biological, and imaging data, we can gain a more holistic understanding of a patient's health and treatment outcomes. Challenges arise in the harmonization of multimodal data due to differences in data dimensions, scales, and information content. For example, merging a  $512 \times 512 \times 88$  vector representing an MRI volume with a clinical feature vector containing several variables requires careful consideration and advanced techniques. In the medical imaging domain, several sub-modalities can be distinguished, such as MRI vs. ultrasound, Dynamic Contrast-Enhanced (DCE) MRI vs. Diffusion-Weighted (DW) MRI, each providing unique insights into the transplanted kidney's health. Recent research in multimodal learning is introducing innovative techniques, including attention mechanisms, data fusion strategies, and transfer learning approaches. These methods are enabling researchers to harness the combined power of different data modalities, resulting in more comprehensive and accurate insights and demonstrating high promise for the complex dynamics of renal transplantation.

## 1.4 Objectives & Contributions

### 1.4.1 Objectives of the thesis

This thesis focuses on developing and designing novel deep learning methodologies for renal transplantation monitoring. On the methodological component, we investigate approaches based on un- and self-supervised learning to tackle and make the most of the scarcity of our data to learn meaningful and robust representations. On the renal transplantation application component, the objective is to demonstrate the potential of the imaging data to hold comprehensive and predictive information regarding the complex biological and medical phenomenon at stake when monitoring renal transplants. In particular,

- **Chapter 2** aims to provide the key baselines when dealing with clinical and medical imaging data focusing on classical machine learning algorithms. The baselines developed focus on establishing a radiomics-based approach, using a variety of machine learning algorithms.
- **Chapter 3** addresses the issue of learning meaningful representations of medical imaging data when having a low number of observations (i.e. patients). This chapter also provides a tailored method to incorporate clinical knowledge to guide the learning of the images' representations and to handle missing follow-up examinations.

- **Chapter 4** intends to go one step further to integrate structural clinical variables in multimodal representations. To this end, we propose an elegant way to input this part of the data with medical prompts by leveraging the recent advances in Natural Language Processing with Large Language Models.
- **Chapter 5** initially summarizes the primary contributions of this thesis, followed by a discussion of compelling research avenues uncovered by our work and potential directions for future research projects. The chapter concludes with a comprehensive synthesis of the thesis.

### 1.4.2 Main contributions

Each chapter of this thesis provides distinct original contributions to demonstrate the relevance of data-driven approaches in medical imaging analysis for renal transplantation monitoring and to set the path for future research in this area. The main contributions are the following.

In **Chapter 2**, we establish the foundational tasks and baseline methodologies that ground this thesis. First, we provide an overview of state-of-the-art studies and models pretraining to clinical and biological data types in the context of renal transplantation. Additionally, we offer illustrative preliminary experiments utilizing our dataset. Subsequently, we delve into our contributions, which include the automatic detection and segmentation of renal transplants within MRI data, published in [Milecki, 2021a]. These tasks are integral to the progression of our research. Lastly, we introduce the concept of radiomics, which involves the engineering of radiological features and serves as the foundation for computational imaging biomarkers.

In **Chapter 3**, we study the use of contrastive learning approaches to learn meaningful representations of renal transplants DCE MRI data. Our approach aims at incorporating clinical information into the learned representation. Then, we propose a novel transformer-based architecture tailored to deal with missing data for the challenging task of serum creatinine prediction 2 years post-transplantation using follow-up imaging modalities. A transformer encoder architecture enables to input the sequential features data per follow-up in order to forecast the renal transplant function and model longitudinal relationships, including a custom method to handle missing exams. In a more clinically oriented study, we further investigate the capacity of the learned representations for graft survival prediction. This work has been published in [Milecki, 2022c; Milecki, 2022a] and a clinical article that is under submission.

In **Chapter 4**, we introduce MEDIMP (**MEDical IMages with clinical Prompts**), an approach to learn powerful manifolds of renal transplant DCE MRI data toward transplant function forecasting in the context of renal transplantation monitoring. This approach provides an elegant way to incorporate clinical or biological information into the learning process of feature extraction of medical imaging data. We designed clinicobiological prompts to leverage textual data and powerful NLP encoders using multimodal alignment of representations with contrastive learning. We demonstrate the relevance of the approach for renal transplant function forecasting. This work has been published in [Milecki, 2023a].

### 1.4.3 Scientific production and professional activities

#### Conference papers

- **Leo Milecki**, Sylvain Bodard, Jean Michel Correas, Marc Olivier Timsit, and Maria Vakalopoulou. “3D unsupervised kidney graft segmentation based on deep learning and multi-sequence mri”. In: *ISBI*. vol. 2021-April. IEEE Computer Society, Apr. 2021, pp. 1781–1785.
- **Leo Milecki**, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean Michel Correas, Marc Olivier Timsit, and Maria Vakalopoulou. “Contrastive Masked Transformers for Forecasting Renal Transplant Function”. In: *MICCAI*. vol. 13437 LNCS. Springer, 2022, pp. 244–254.
- **Leo Milecki**, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. “MEDIMP: 3D Medical Images with clinical Prompts from limited tabular data for renal transplantation”. In: *MIDL*. PMLR, Apr. 2023.

#### In submission

- **Leo Milecki**, Sylvain Bodard, Vicky Kalogeiton, Florence Poinard, Anne-Marie Tissier, Idris Boudhabhay, Jean-Michel Correas, Dany Anglicheau, Maria Vakalopoulou, and Marc-Olivier Timsit “Renal transplant survival prediction from unsupervised deep learning-based radiomics on early DCE MRI”, under submission to Kidney International journal as Technical Notes format.

#### Abstracts

- **Leo Milecki**, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. “Contrastive Learning for Kidney Transplant Analysis using MRI data and Deep Convolutional Networks”. In: *MIDL Short Paper Track*. 2022.

#### Professional activities

- **Research visit** of three months (Sep-Dec 2022) at **Provost Ultrasound Lab**, Polytechnique Montréal, Montréal, Canada. Prospecting the potential of our US data with an expert team in ultrasound imaging, and more particularly contrast-enhanced.
- **Teacher assistant** for “Introduction to Deep Learning” Course (Fall 2020 & 2021), *Master MVA (Mathématiques, Vision, Apprentissage)*, ENS Paris-Saclay and CentraleSupélec, “Foundation of Deep Learning” Course (Fall 2021) *Master DSBA (Data Sciences & Business Analytics)*, CentraleSupélec, and “Foundation of Deep Learning” Course (Fall 2023) *Master IA (Intelligence Artificielle)*, CentraleSupélec.
- **Project Supervision** for AI Hub undergraduate students of CentraleSupélec with **Hopia** (Spring 2021) on AI for hospital beds management and with **Aura** (Spring 2022) on benchmarking ML algorithms for epilepsy crisis detection.
- **Reviewer** for conferences: 2024 WACV, 2023 MICCAI (Honorable Mention), 2023 ICCV, 2022 MICCAI (Honorable Mention - Top 9%), and journals: Medical Image Analysis, Computer Vision and Image Understanding, Computers in Biology and Medicine.
- **Student volunteer** for ICCV 2023 in Paris.

**Training courses**

- Participation in AI4Health Winter Workshop 2021 (including a poster presentation).
- Lectures and Training on Fundamentals of Accelerated Computing with CUDA Python by NVIDIA Deep Learning Institute Team.
- Scientific English by the Academic Writing Center of CentraleSupélec.
- Scientific integrity in research professions, Fun MOOC by Bordeaux University.



# Chapter 2

## Classical methods & baselines

### Contents

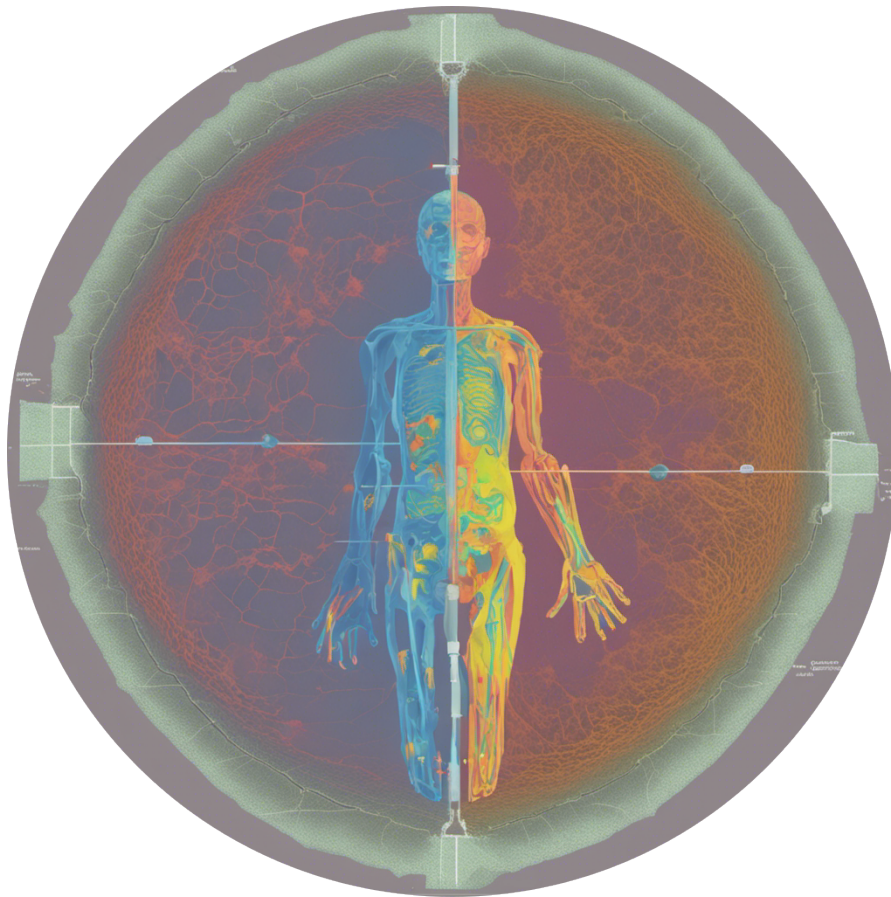
---

2.1	Classical data-driven methods for medical tabular data . . . . .	43
2.1.1	Clinically relevant tools . . . . .	43
2.1.2	Experiments on clinical data . . . . .	43
2.2	Image-based segmentation of kidney transplant . . . . .	47
2.2.1	Introduction & Related work . . . . .	47
2.2.2	Methodology . . . . .	49
2.2.3	Data . . . . .	51
2.2.4	Experiments & Results . . . . .	52
2.2.5	Discussion & Conclusion . . . . .	55
2.3	Design of radiomics signature for kidney transplant . . . . .	56
2.3.1	Identifying biomarkers from imaging data . . . . .	56
2.3.2	Radiomics for renal transplantation . . . . .	57

---







Prompt<sup>1</sup>: "A computer vision segmentation map of a medical imaging system in a surrealist style."

## Summary

This chapter establishes the foundational tasks and baseline methodologies that ground this thesis. First, we provide an overview of state-of-the-art studies and models pretraining to clinical and biological data types in the context of renal transplantation. Additionally, we offer illustrative preliminary experiments utilizing our dataset. Subsequently, we delve into our contributions, which include the automatic detection and segmentation of renal transplants within MRI data, published in [Milecki, 2021a]. These tasks are integral to the progression of our research. Lastly, we introduce the concept of radiomics, which involves the engineering of radiological features and serves as the foundation for computational imaging biomarkers.

---

<sup>1</sup>Generated using [Stable Diffusion XL](#).



## 2.1 Classical data-driven methods for medical tabular data

### 2.1.1 Clinically relevant tools

While considered "classical" in the machine learning community, the models discussed in Section 1.1.1 have seen relatively recent widespread adoption and remain relatively uncommon in clinical studies aimed at comprehending renal transplantation phenomena and mechanisms. These studies often lean on two primary types of models: linear/logistic regression and survival analysis models, such as the Cox model [Cox, 1972] and Kaplan-Meier curves [Kaplan, 1958]. Survival analysis encompasses a suite of statistical tools designed to analyze the anticipated time until a specific event occurs, be it complications like acute rejection, the onset of chronic dysfunction, or the ultimate occurrence of transplantation failure in renal transplantation cases. A recent study [Truchot, 2023] advises caution against the blind adoption of machine learning algorithms, highlighting lower predictive and calibration results than classical statistical approaches in modeling transplant failure using donor, recipient, and transplant-related parameters.

Several studies [Moore, 2011; Shabir, 2014; Gonzales, 2016] investigated the ability of different variables to predict graft survival. Individual variables, such as estimated glomerular filtration rate (eGFR) [Kaplan, 2003; He, 2009], proteinuria [Naesens, 2016], or histology [Yilmaz, 2003] failed to produce clinically relevant performances. Moreover, efforts at developing prognostic systems in renal transplantation have been hampered by small sample sizes, the absence of proper validation, limited phenotypic details, and the failure to include key prognostic factors that affect allograft outcomes (e.g., donor-derived factors, or disease recurrence). The recent iBOX [Loupy, 2019] model demonstrated the most complete and comprehensive study, validated on large international cohorts. Specifically, the study drew upon a derivation cohort comprising 4000 kidney recipients from four French centers. Additionally, validation cohorts incorporated data from 2129 kidney recipients across three European centers and 1428 recipients from three North American centers. This collective dataset was instrumental in the identification of eight key prognostic factors spanning functional, histological, and immunological domains. These factors were then combined into a risk prediction score, which exhibited robust confirmation in the validation cohorts across both European (C-index 0.81) and U.S. (0.80) settings.

### 2.1.2 Experiments on clinical data

A first dive into this thesis was to analyze the predictive power of the available clinical variables to detect or forecast events, such as transplant failure, at the scale of having a low number of patients and more particularly positive events<sup>2</sup>. Nevertheless, those preliminary experiments enabled us to better apprehend the medical context, the challenges related to our data, and to drive our research directions and objectives.

---

<sup>2</sup>In practice, only these data were available during the beginning of the project and the imaging data curation arrived in a second phase (mainly due to COVID19 restrictions).

## Data analysis and visualization

The first task we tackled was to analyze the data regarding the "transplantation failure" event. Transplantation failure was defined as a return to dialysis. From a machine learning point of view, it can be addressed as a binary classification task. Among the 126 patients for whom a substantial portion of clinical variables were available, 14 individuals experienced transplantation failure. We performed a first analysis to highlight the relationships between these clinically relevant attributes and the occurrence of transplantation failure. In particular, the correlation of the attributes is presented in Figure 2.2 using Pearson correlation analysis. Notably, our data analysis revealed that the post-transplantation complication count exhibits the strongest correlation with the incidence of transplantation failure. Furthermore, our examination uncovered noteworthy statistical associations within the dataset. For instance, we observed a correlation between the year of birth and the number of previous transplantations, suggesting that older individuals are more likely to have undergone prior transplant procedures. Additionally, we identified a relationship between the type of transplant and the duration of cold ischemic time. In particular, the cold ischemic time, signifying the delay before the transplant's reperfusion during the transplantation surgery, tends to be longer for organs from deceased donors compared to those from living donors. This correlation visualization provided valuable insights into the interplay between clinical variables and transplantation outcomes, offering a foundation for further analysis and understanding of renal transplantation dynamics.

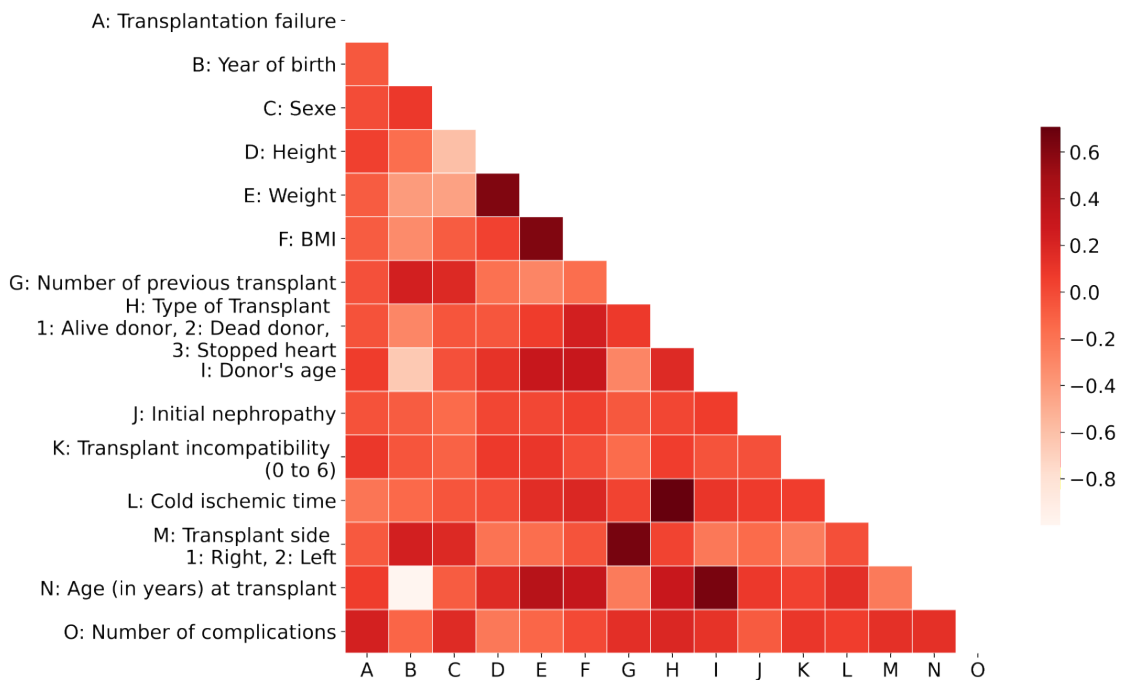


Figure 2.2: **Correlation matrix visualization of general clinical variables.** Pearson correlation between each variable was computed and outputs a value in  $[-1, 1]$ . The darker the red color is, the more correlated the two variables are.

In Figure 2.3a, we observed radar plots of the mean and standard deviation (std) values of those variables

when separating the patients into two groups according to the transplantation failure annotation. Figure 2.3b corresponds to the same radar plot for the complications during the surgery variables. Complications are clearly more important for the group of patients who had transplantation failure. Whereas for other clinical variables, the difference between the two groups did not appear indicative.

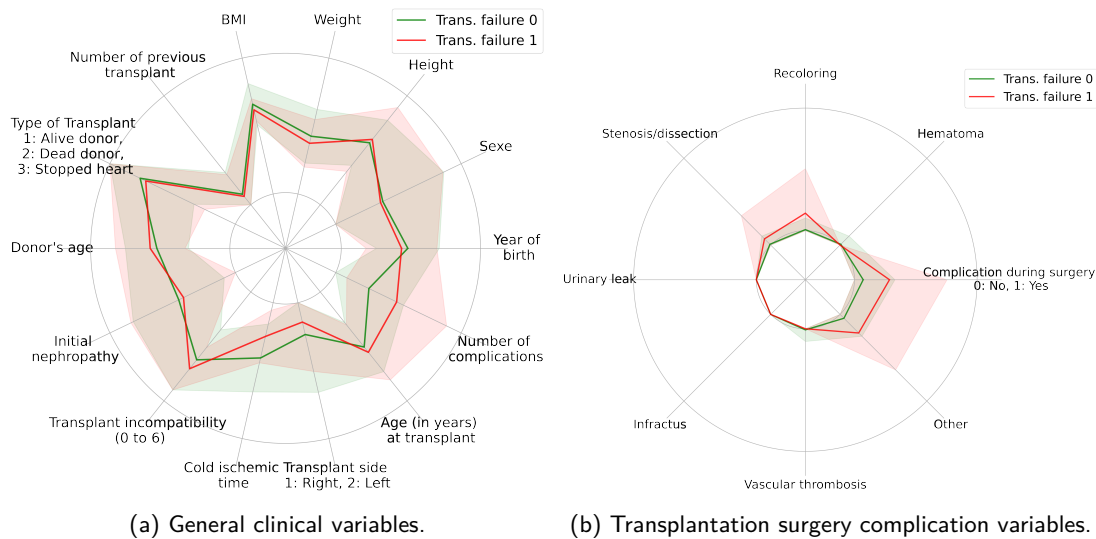


Figure 2.3: **Radar charts on clinical data.** Visualization of some general clinical variables statistics (mean  $\pm$  std) separated into two groups of patients, having transplantation failure (Trans. failure) or not.

### Data-driven methods for clinical data

After the comprehensive analysis of our available clinicobiological attributes, our objective was to assess the performance of classical machine learning algorithms in predicting transplantation failure. To achieve this, we designed a structured pipeline consisting of several key steps, each carefully tailored to maximize predictive accuracy:

- **Data Split:** We partitioned the dataset into training and testing subsets, with an equitable distribution of 50% and stratified regarding the transplantation failure label.
- **Model Selection:** We considered a range of machine learning models, including Random Forest, Decision Tree, Lasso, Linear Support Vector Machine (SVM), Radial Basis Function (RBF) SVM, Adaboost, and Gaussian Naive Bayes.
- **Hyperparameter Specification:** For each selected model, we defined a set of hyperparameters and identified those that warranted fine-tuning.
- **Wide Parameter Grid:** A broad parameter grid was formulated, encompassing the hyperparameters selected for optimization.
- **Randomized Grid Search:** Employing a randomized grid search methodology, we executed a 4-fold stratified cross-validation process on the training subset, utilizing 10% of the total grid size.

- Analysis of Results: To identify the optimal hyperparameters, we examined the outcomes of the randomized grid search. This involved constructing box plots to visualize the validation scores against parameter values on the grid.
- Narrow Parameter Grid: Focusing on the best hyperparameters determined in the previous step, we established a more constrained parameter grid.
- Full Grid Search: A comprehensive grid search was performed through 4-fold stratified cross-validation, using the narrowed parameter grid.
- Evaluation: Ultimately, we evaluated the model's performance on both the training and testing sets.

Table 2.1: **Transplantation failure prediction with classical ML algorithms from clinical variables.** The "Select." column indicates how the best model hyperparameters were chosen, either by selecting the model with the best cross-validation F1 score ("Highest") or by employing an ensembling method that considered models performing above the median cross-validation F1 score ("Median"). **Bold**, Underlined indicates the top 1, 2 performing combinations, respectively.

Model	Select.	Train				Test			
		B.Acc	Rec	Prec	F1	B.Acc	Rec	Prec	F1
Random Forest	Highest	87.5	75.0	<b>100</b>	85.7	68.8	41.7	71.4	<u>52.6</u>
	Median	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<u>69.3</u>	<u>44.4</u>	57.1	50.0
Decision Tree	Highest	66.8	35.7	<u>83.3</u>	50.0	63.4	31.2	71.4	43.5
	Median	55.1	14.3	<u>83.3</u>	24.4	50.3	11.6	71.4	20.0
Lasso	Highest	61.0	26.1	75.0	38.7	61.9	26.7	80.0	40.0
	Median	54.1	15.6	62.5	25.0	52.3	13.3	40.0	20.0
Linear SVM	Highest	71.7	45.5	<u>83.3</u>	58.8	53.2	16.7	28.6	21.1
	Median	71.7	45.5	<u>83.3</u>	58.8	53.2	16.7	28.6	21.1
RBF SVM	Highest	<u>92.9</u>	<u>85.7</u>	<b>100</b>	<u>92.3</u>	<b>88.5</b>	0	0	0
	Median	10.0	10.0	<b>100</b>	18.2	11.5	11.5	<b>100</b>	20.6
Adaboost	Highest	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	67.7	42.9	42.9	42.9
	Median	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	72.2	<b>50.0</b>	57.1	<b>53.3</b>
G. Naives Bayes	Highest	68.8	37.5	<b>100</b>	54.6	60.1	23.1	<u>85.7</u>	36.4
	Median	68.6	42.9	50.0	46.2	59.7	28.6	28.6	28.6

Table 2.1 presents a summary of the training and testing scores obtained in our analysis. In the "Select." column, we indicate how the best model hyperparameters were chosen, either by selecting the model with the best cross-validation F1 score ("Highest") or by employing an ensembling method that considered models performing above the median cross-validation F1 score ("Median"). It is notable that many of these machine learning models exhibit a tendency to overfit, evident from the substantial discrepancy between the near-perfect training scores and the test scores. Notably, the RBF SVM model (line 10) exemplifies this behavior most prominently. This behavior is largely attributed to the limited number of training samples available. Among the models, the Adaboost and Random Forest algorithms appear to yield the most promising results

on the test set, achieving balanced accuracies of 72.2% and 68.8%, respectively, as well as F1 scores of 53.3% and 52.6%, respectively. The selection procedure for hyperparameters favored "Median" for the Adaboost model, while "Highest" proved to be the optimal choice for other models. Despite our persistent efforts to construct a robust machine learning pipeline aimed at mitigating these challenges, the test results fell short of achieving the desired level of satisfaction.

To conclude this first section focusing on classical ML algorithms applied to clinical and biological variables, while this portion of our dataset did not reveal clinically significant patterns, it did provide valuable insights into the renal transplantation context and fostered productive discussions with our medical colleagues. In a subsequent phase, we meticulously curated the imaging data within our dataset, and the following section will present the outcomes of our preliminary efforts in this domain. To provide new information and enhance the clinical attributes, characteristics from the medical images are needed to be extracted from the data. To this end, we need algorithms for the accurate localization of the transplants in the images. This includes our initial contribution to object detection and segmentation of renal transplants in Dynamic Contrast-Enhanced (DCE) MRI data.

## 2.2 Image-based segmentation of kidney transplant

### 2.2.1 Introduction & Related work

Segmentation of organs or other structures such as vessels or cells in biomedical imaging is often an important first step or even a prerequisite to analyzing medical volumes. They are essential to describe and monitor shapes, and volume characteristics [Lee, 2011] or are further used in more complex pipelines, as in radiomics analysis [Rizzo, 2018] and precision medicine. In recent years, automatic segmentation in biomedical images has been subject to extensive research, with deep learning-based techniques holding state-of-the-art performances in various applications and modalities, including Magnetic Resonance Imaging (MRI) [Lundervold, 2019]. However, one of the main limitations of common deep learning supervised approaches is the need for expensive and time-consuming annotations.

Extracting information and analyzing biomedical data is not very trivial due to the very high dimensionality and complexity. Moreover, one more limitation that needs to be addressed when it comes to these types of data is the absence or a limited number of annotations that are usually available. This is one of the main challenges that need to be tackled when it comes to machine learning and deep learning approaches. Nowadays, unsupervised approaches investigate methods that are able to generate models without the need for annotations. In particular, concerning unsupervised kidney segmentation, image-based methods have been proposed to obtain a segmentation through an image processing pipeline using tools such as thresholding [Sandmair, 2016] or morphological operators [Li, 2016; Yu, 2012]. Some other methods propose clustering frameworks based on K-means [Yang, 2016] or on wavelet-based clustering [Li, 2012]. Deformable models were also designed in level-set methods [Abdulahi, 2015] or exploiting active contours [Yu, 2013]. Advanced



deformable boundary methods are also explored in [Shehata, 2018] with however the use of training samples for parts of their formulations. Applications of unsupervised deep learning to segmentation focused mainly on generative models.

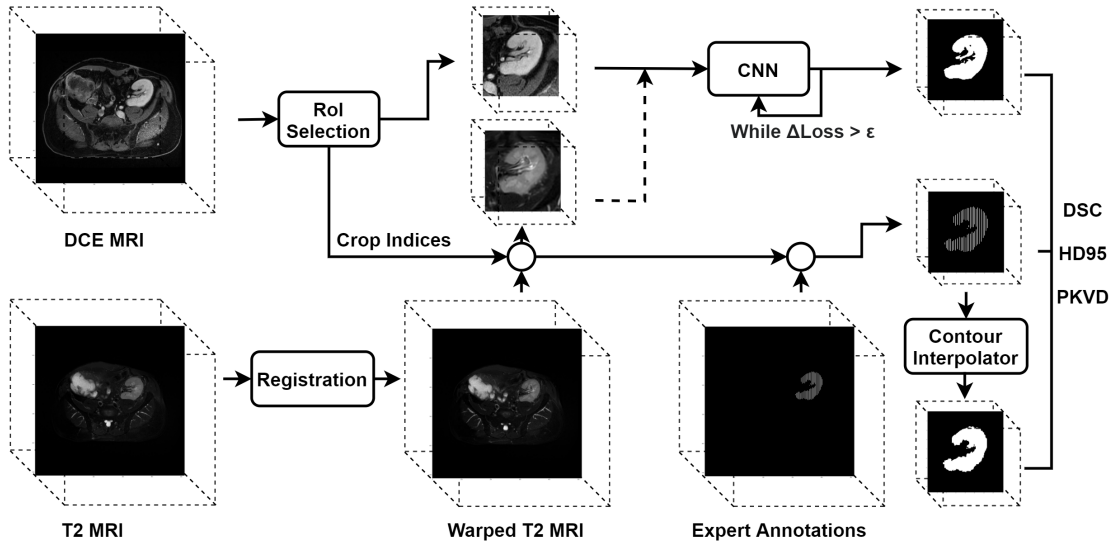


Figure 2.4: **Overview of the proposed method for the automatic segmentation of kidney graft using multimodal MR Imaging.** T2 MRI volumes are aligned to the DCE MRI volumes. The Region of Interest (RoI) is obtained from the DCE MRI using thresholding and morphological operations and applied to the T2 MRI. The predicted segmentation map is obtained from our unsupervised semantic segmentation algorithm, taking as input both modalities and optimized to reach a tolerance criteria  $\epsilon$  from two constraints. Finally, expert annotations enable us to evaluate our results using the dice score (DSC), Hausdorff distance at 95 percentile (HD95), and the percentage of kidney volume difference (PKVD).

This work proposes an unsupervised method for the segmentation of kidney grafts from Dynamic Contrast-Enhanced (DCE) and T2 sequences obtained 12 months after the transplantation. Our method consists of a two-step procedure that is summarised in Figure 2.4. First, we propose an unsupervised method to detect the kidney region using thresholding techniques and morphological operators. The second step corresponds to the pixel-wise segmentation of the kidney graft. Our method is based on [Kim, 2020] extending the formulation to 3D to explore all the available information from the MRI modalities. Moreover, it proposes a multimodal formulation where multiple MRI sequences are integrated towards a robust kidney graft segmentation tool.

## 2.2.2 Methodology

### Selection of the region of interest

The first step of the proposed method is the automatic detection of possible areas of interest. This step is essential to reduce the search space and boost the performance of the unsupervised segmentation technique eliminating the number of false positives. The detection of the region of interest is based on a relatively simple pipeline taking full advantage of the contrast-enhanced MRI. In particular, a thresholding operation is first applied to the entire DCE volume highlighting the highest intensities of the volume corresponding to regions that depict the graft kidney and blood vessels. In the pixels' volume histogram, the threshold is obtained as the abscissa of maximum intensities. After the volume thresholding, morphological operations are applied to remove artifacts and fill holes in the detected regions. The morphological step consists of a binary closing followed by a binary opening. For both operations, spheres of radius 5 and 4 have been chosen as structural elements, respectively. Connected component analysis has been used in the thresholded volume. A bounding box indicating the highest volume component was selected as the region of interest (see Figure 2.5) and further processed. The same area is also indicated in the T2 sequence. Areas of interest ranged for  $N_x \times N_y \times N_z \in [125, 173] \times [118, 199] \times [49, 75]$ .

### Unsupervised segmentation model

The proposed unsupervised segmentation model then processes the bounding boxes of various sizes containing the region of interest. Our method is based on differentiable feature clustering [Kim, 2020]. A simple straight-forward 3D convolutional neural network (CNN) is used to extract features and optimize two loss functions without any need for ground truth annotations. The first loss is designed to refine and extract relevant features, and the second is to favor smoothness on the obtained labels. The first loss will be defined as similarity and the second as continuity loss.

Let us denote  $X \in \mathbb{R}^{2 \times N_x \times N_y \times N_z}$  the regions of interest extracted from the T2 and DCE sequences. Our unsupervised method is based on a 3D feature extractor CNN denoted as  $h(\cdot, \theta)$ , where  $\theta \in \Theta$  defines its trainable parameters. The  $p$ -dimensional feature vector obtained by the CNN is denoted as  $Y = h(X, \theta) \in \mathbb{R}^{p \times N_x \times N_y \times N_z}$ . From this feature vector, we can propose a mapping strategy defined by a function  $g$  to obtain class labels for each of the voxels of the volume  $C = g(Y) \in \mathbb{Z}^{N_x \times N_y \times N_z}$ . Two different loss functions are defined for the optimization of the method,

$$L_{sim}(Y, C) = \sum_{i=1}^{N_x \times N_y \times N_z} \sum_{j=1}^p -\delta(j - c_i) \ln(y_{j,i}) \quad (2.1)$$

where  $Y = \{y_{j,i}\}_{j \in [1,p], i \in [1, N_x \times N_y \times N_z]}$  and  $C = \{c_{j,i}\}_{j \in [1,p], i \in [1, N_x \times N_y \times N_z]}$  and  $\delta$  is the kronecker function. This loss corresponds to the cross entropy between the network output and class labels used as intermediate targets that are obtained by an argmax operation.



Figure 2.5: **Region of interest example on the DCE MRI sequence.** The MRI volume is displayed slice by slice in axial view. The obtained region of interest is formalized by the red bounding box.

$$L_{cont}(Y) = \sum_{k=1}^{N_x-1} \sum_{l=1}^{N_y-1} \sum_{m=1}^{N_z-1} \|y_{k+1,l,m} - y_{k,l,m}\|_1 + \|y_{k,l+1,m} - y_{k,l,m}\|_1 + \|y_{k,l,m+1} - y_{k,l,m}\|_1 \quad (2.2)$$

where  $y_{k,l,m}$  corresponds to the  $p$ -dimensional feature vector of the voxel at  $(k, l, m)$ . This loss aims to enforce spatial smoothness on the obtained segmentation masks.

Finally, the overall optimization is achieved by,

$$L = L_{sim}(Y, C) + \lambda L_{cont}(Y) \quad (2.3)$$

where  $\lambda$  is a weight that balances the contribution of each component to the final optimization.

The network parameters are optimized iteratively until the absolute difference of loss between two successive iterations is lower than a predefined value  $\epsilon$ . Using this criterion, the number of classes obtained can be superior to 2 as the non-kidney elements can rarely be assigned in one class. Using the final predicted clustering, the kidney class can easily be obtained by selecting the first connected component class.

### Implementation details

In this study, we first applied bias field correction to all the MRI images as a preprocessing step. The number of iterations was set to 2 and the number of fitting levels to 8, enabling the enhancement of peaks in volumes' histograms and, more particularly, contrast-enhanced regions. Moreover, even if the two sequences (DCE and T2) correspond to the same follow-up exam, they were acquired a few minutes apart, introducing misalignments. To address this issue and project properly one sequence to the other, we used an unsupervised deformable registration method [Ferrante, 2017]. The DCE modality has been used as the target, and the T2 has been warped to it, using a variety of similarity metrics, including normalized cross-correlation, mean square error, and mutual information.

Our 3D CNN architecture consists of  $N$  convolutional blocks with  $p$  channels. Each block consists of a 3D convolution layer with a kernel size of  $3 \times 3 \times 3$  and stride 1, followed by a ReLU activation and a 3D batch normalization layer. The two different sequences were concatenated channel-wise and were given into the network. The choice to include as stacked channels the two sequences – early fusion – has been made and tested over middle (classification level fusion) or late fusion (decision level fusion) strategies [Fang, 2020].

Concerning the other parameters of our network, we set  $\lambda = 5$  and  $\epsilon = 0.005$ . Our CNN model was trained using classic stochastic gradient descent with a momentum set to 0.9 and a learning rate of 0.01. The overall optimization of a single volume needed approximately 1 minute on a GeForce GTX 1080 TI GPU (11.2 Go).

### 2.2.3 Data

The dataset used in this study corresponds to a total of 32 patients subject to a kidney transplant. All the MRIs were obtained 12 months after the transplantation. Our dataset consists of 2 different MRI sequences, namely T2 and DCE. DCE volumes being  $512 \times 512 \times [64 - 88]$  included spacing ranging from  $0.78 \times 0.78 \times 1.9$  to  $0.94 \times 0.94 \times 2.5$  and T2 volumes being  $512 \times 512 \times [25 - 36]$  from  $0.70 \times 0.70 \times 5.0$  to  $0.98 \times 0.98 \times 6.0$ . For the optimization of the network's hyperparameters, 3 patients were used, while all the rest (29 patients) were used for the evaluation of the proposed method. The 3 patients were selected as a sample of low, medium, and high segmentation difficulty volumes from preliminary studies. Pixel-wise kidney annotation was performed by a medical expert (a board-certified radiologist), by tracing the kidney boundaries using the DCE MRI in sagittal view. The annotation was available in every 2 or 3 slices, resulting in 945 annotated slices. We obtained the 3D annotation by interpolating the obtained manual annotations per subject to assess the method's performance.

### 2.2.4 Experiments & Results

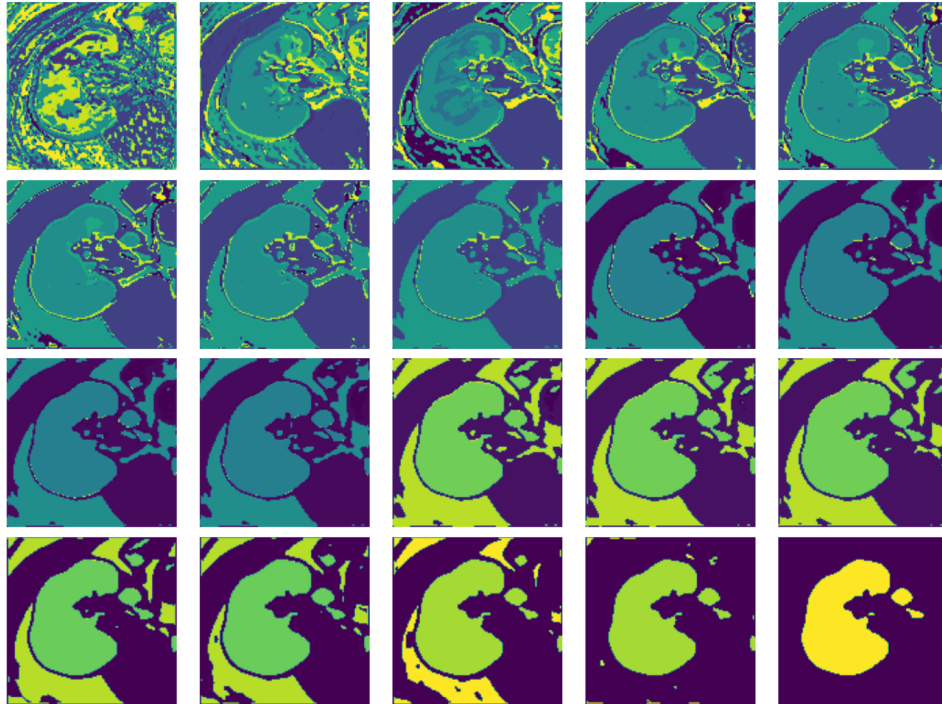


Figure 2.6: **Semantic segmentation maps obtained through the iterative process of our unsupervised algorithm.** Our approach was applied to the selected region of interest. We plot the segmentation maps obtained every five iterations for one slice from one MRI volume and colors correspond to all the detected labels.

We evaluated the graft kidney's segmentation accuracy using five metrics commonly used to assess biomedical segmentation tasks. First, the mean dice coefficient (DSC) metric is used to indicate the similarity between the manual annotation and prediction. Secondly, we used the mean of Hausdorff distance at percentile 95% (HD95) to measure how far two surfaces are from each other. Finally, the mean percentage of kidney volume difference (PKVD), precision (Prec), and recall (Rec) between our predictions and the annotations were used to evaluate different methods' performance. We visualize in Figure 2.6 the semantic segmentation maps obtained on a region of interest example through the iterative process of our unsupervised learning algorithm.

Starting with the evaluation of the proposed method, an ablation study benchmarking the influence of the parameter  $\lambda$  and the 3D CNN architecture has been performed for both 1 sequence (DCE) and 2 sequences (DCE & T2) models (Figure 2.7). Starting with the influence of  $\lambda$ , we used grid search and benchmark values in  $\{0.01, 0.1, 1, 5, 10, 25, 50, 100\}$ . The  $\lambda$  parameter shows the influence of the  $L_{cont}$  in the overall optimization. The superiority of the 2 sequence model seems to be indicated with the best performances to be found in [1,25]. Moreover, for the 1 sequence model, the variations are less monotone, and the selection

of an optimum value is more difficult.

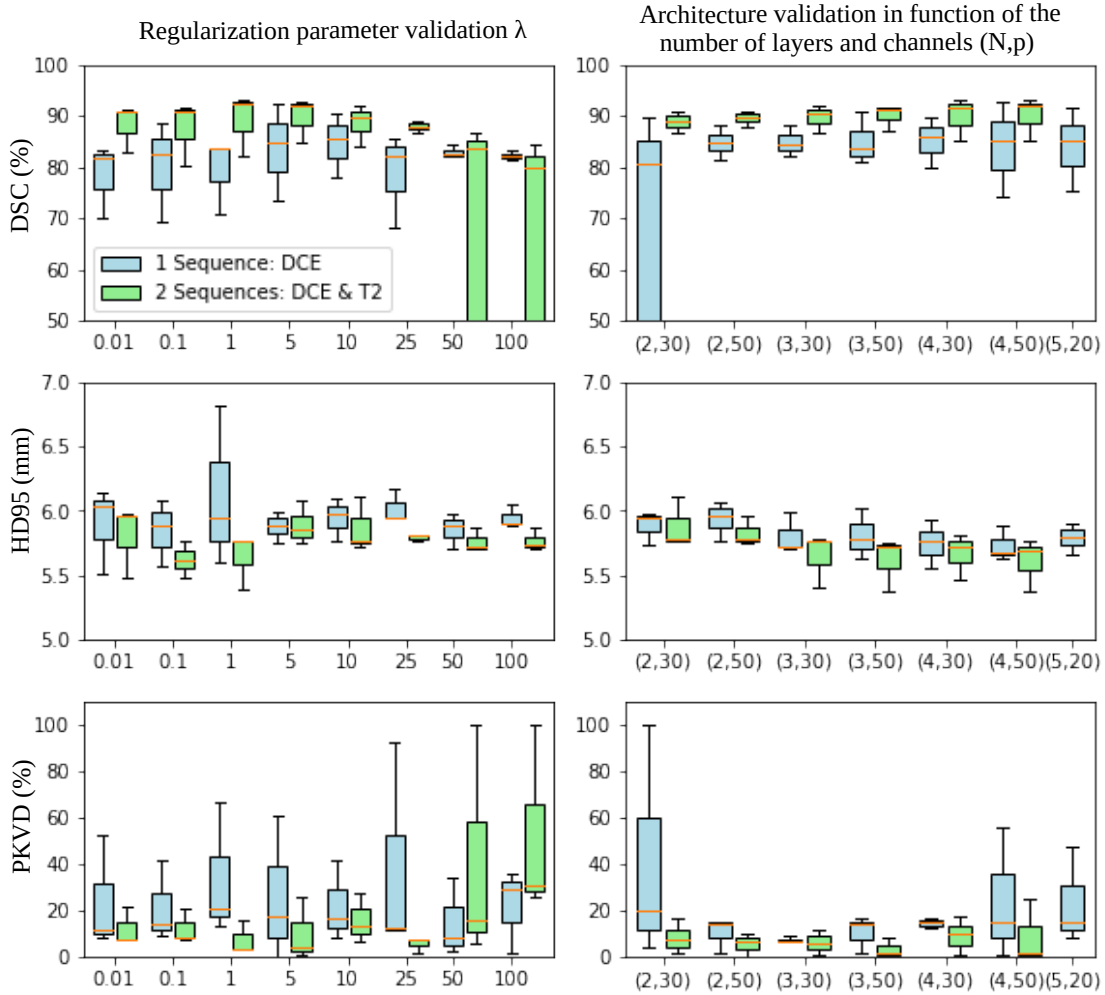


Figure 2.7: **Evaluation of the different components of the proposed method using different evaluation metrics on the validation set.** The first column indicates the influence of  $\lambda$  and the second column indicates the influence of the used architecture for different numbers of channels and layers.

Moreover, we analyzed the impact of the network's architecture in depth (number of convolutional blocks) and width (number of channels in each block) to extract relevant features. We evaluated these parameters ( $N$  convolutional blocks,  $p$  number of channel per block) in the set of  $\{(2,30), (2,50), (3,30), (3,50), (4,30), (4,50), (5,20)\}$ . The specific parameters seem to not really influence each model's performance, with again the 2 sequences one reporting more stable performance for all the experiments. Overall, the (2,50) configuration seems to report more stable performance, with less variance, on the validation set. This indicates that a relatively simple architecture is enough to address our problem.

Table 2.2: **Quantitative evaluation of the proposed method and other methods reporting:** dice coefficient (DSC), Hausdorff distance at percentile 95% (HD95), percentage of kidney volume difference (PKVD), precision and recall. **Bold**, Underlined indicates the top 1, 2 performing combinations, respectively.

Method	DSC (%)		HD95 (mm)		PKVD (%)		Prec (%)		Rec (%)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Otsu Thresholding	71.04	10.97	6.55	0.65	84.71	57.34	57.07	13.03	<b>97.53</b>	<b>0.91</b>
Watershed	84.06	23.30	<b>5.75</b>	<u>0.42</u>	16.35	23.99	<u>89.25</u>	23.91	80.05	23.18
Proposed 2D ( <i>DCE only</i> )	69.09	20.34	7.22	1.67	51.30	65.00	74.57	25.77	69.79	20.48
Proposed ( <i>DCE only</i> )	<u>86.48</u>	<u>4.01</u>	5.92	<u>0.42</u>	<u>11.08</u>	<u>10.13</u>	83.93	<u>8.32</u>	<u>89.97</u>	<u>3.67</u>
Proposed 2D ( <i>DCE &amp; T2</i> )	75.75	10.80	6.39	1.11	32.81	41.56	81.65	14.24	81.65	14.66
Proposed ( <i>DCE &amp; T2</i> )	<b>89.81</b>	<b>3.17</b>	<u>5.82</u>	<b>0.41</b>	<b>5.88</b>	<b>5.74</b>	<b>91.00</b>	<b>5.18</b>	88.20	5.82

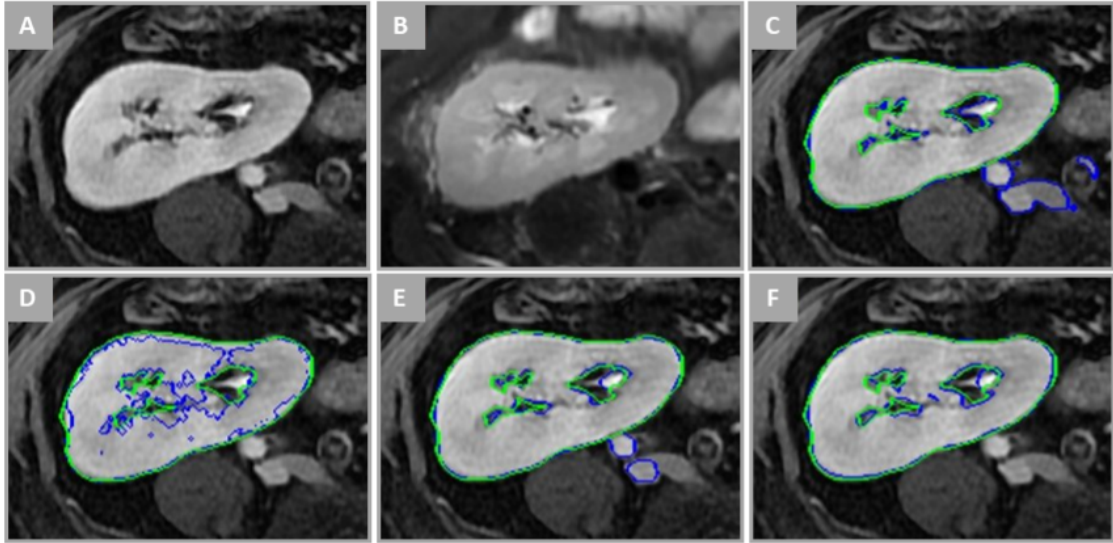


Figure 2.8: **One testing subject depicting one slice in axial view and the segmentation results as overlay for each of the evaluated methods (in blue) and the annotation from the expert (in green):** A: DCE, B: Warped T2, C: Otsu thresholding, D: watershed-based method, E: Proposed (*DCE*), F: Proposed (*DCE & T2*).

To highlight our method’s soundness, we compare it with other commonly used unsupervised methods in the literature and its 2D implementation [Kim, 2020]. In particular, our results were evaluated against the kidney graft segmentations obtained by Otsu thresholding [Bindu, 2012] and watershed 3D [Yu, 2012] methods that are commonly used for unsupervised segmentation in a variety of studies.

The obtained results are summarised in Table 2.2 presenting the mean and standard deviation (std) of each score on the whole test set. Even if additional postprocessing methods are presented in the literature [Will, 2014; Yu, 2012] in our study, we did not perform any additional postprocessing in any of the exploit methods. The lowest performance in terms of all the metrics is reported by the Otsu thresholding, which reaches a mean precision of  $57.07 \pm 3.03\%$ , highlighting the high number of false positives. Watershed seems to perform better, reporting a mean DSC of 84.06% with, however, a high standard deviation. The proposed method

performs better in all the cases. Using only the DCE MRI as input, we obtain better (in terms of mean values) and more stable (in terms of std) performance than the previous two methods. The performance of the proposed method is further increased with the integration of the T2 sequence. The 2-sequence model reports 3% higher DSC than the model that uses only one sequence. The method's superiority is also indicated by the higher numbers of mean precision that increases from  $83.93 \pm 8.32\%$  for the DCE to  $91.00 \pm 5.18\%$  for the DCE & T2 models. Even if the two models report similar mean HD95, the DCE & T2 model's impact is further indicated by the significantly lower mean PKVD metric. Finally, our experiments indicated that the 2D [Kim, 2020] implementation is not performing well, highlighting the need for 3D formulations for this problem.

In Figure 2.8, we present some qualitative analysis of the obtained segmentation masks. One can observe that the accurate detection of the kidney between structures – essentially between the kidney and its irrigating vessels – is very challenging for unsupervised methods. However, the proposed formulation is able to separate the kidney graft from the different structures and provide an accurate segmentation.

### 2.2.5 Discussion & Conclusion

In this work, we proposed a 3D deep learning-based unsupervised kidney segmentation framework using multi-modal MRI. Our framework consists of two losses, one that optimizes the cross-entropy between the predicted labels and the labels of clusters, while the second one is enforcing spatial continuity, producing smooth segmentation maps. Our 3D formulation works better than other commonly used unsupervised methods in the literature, including its 2D implementation [Kim, 2020].

Our experiments indicated that integrating more than one MRI sequence could significantly help remove detected structures close to the kidney graft, such as vessels irrigating the contrast agent. Our proposed method limitations include the number of hyperparameters that need to be defined, which are the same in all the deep learning-based frameworks, and a lack of prior kidney information. Our quantitative results indicate that our tool could help clinical practice provide fast and robust monitoring of kidney grafts. In the future, we would like to investigate the impact of additional loss functions that could integrate more shape-specific information. Moreover, we plan to investigate our segmentation framework's adaptation to the analysis of follow-up exams for patients subject to transplantation toward potential rejection.



## 2.3 Design of radiomics signature for kidney transplant

### 2.3.1 Identifying biomarkers from imaging data

"Omics" domains correspond to biomedical fields that generate big data, such as genomics, proteomics, or metabolomics. Medical Imaging has been used increasingly and so its quantitative analysis by means of advanced, and sometimes nonintuitive mathematical tools, constituting the radiomics approaches. The concept of radiomics was primitively broadly applied in the field of oncology. The assumptions of radiomics presume that biomedical images contain information on disease-specific processes that are imperceptible by the human eye [Mannil, 2018]. Thus, such a quantitative approach could unveil information that is not accessible through traditional visual inspection, even by radiology experts.

More precisely, radiomics is based on a key concept of machine learning approaches: feature engineering, designing features from prior knowledge of the data, and reducing its dimensionality. The goal is to analyze with statistical or more advanced machine learning tools the influence (correlation, power of prediction) of these features on some endpoints, such as diagnosis, drug treatment, or prognosis. The features correspond to the extraction of the spatial distribution of signal intensities, pixel interrelationships, shapes, and textures. Using predefined formulas, radiomics address also the issue of overcoming the subjective nature of image interpretation, and the variability in radiology experts' knowledge and training.

Van Griethuysen et al. [Van Griethuysen, 2017] introduced `pyradiomics`, an open-source Python package for the extraction of radiomics features from medical imaging and hence aimed at establishing a reference standard for radiomic analysis by providing a tested and maintained open-source platform for easy and reproducible radiomic features extraction. Authors divided features in classes: First Order Statistics, Shape-based (3D), Shape-based (2D), Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighbouring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM). This work constitutes our main baseline for extracting predefined radiomic features from our medical imaging data. It is important to note that the radiomics term was also extended to more advanced AI methodologies where the feature extraction process is part of the learning of the model, e.g. by using deep neural network encoders, which we will explore in the next chapters.

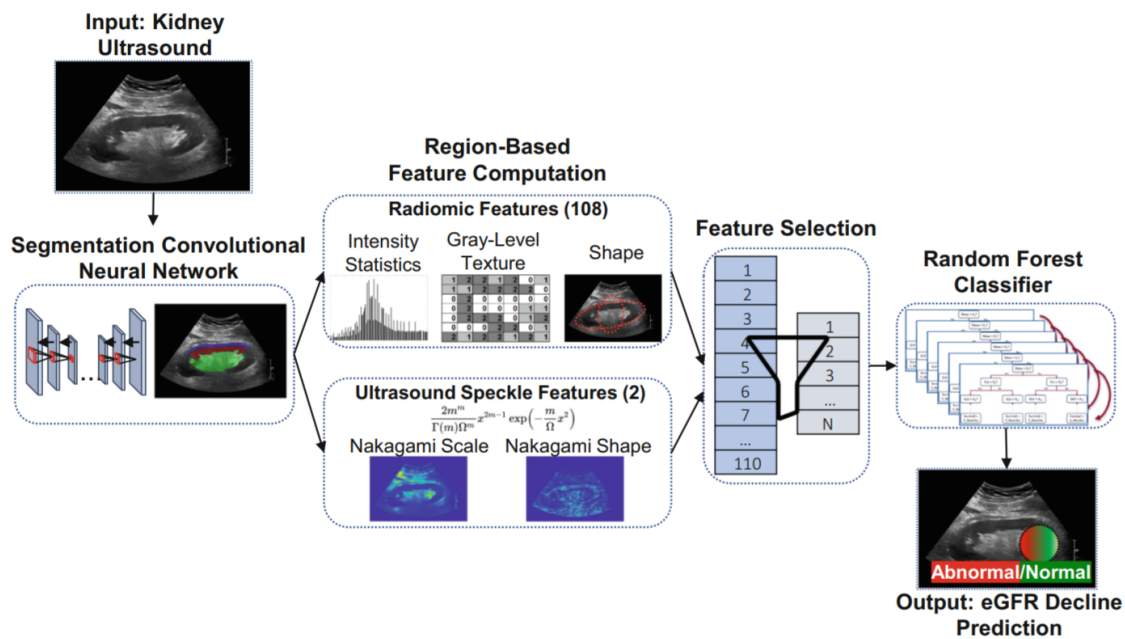


Figure 2.9: **Overview of a radiomics approach for renal transplant ultrasound imaging.** Image from Ricky Hu et al. “Prediction of Kidney Transplant Function with Machine Learning from Computational Ultrasound Features”. In: *MICCAI Workshop ASMUS*. vol. 13565 LNCS. Springer, 2022, pp. 34–43.

### 2.3.2 Radiomics for renal transplantation

Renal transplantation imaging datasets are rare and often contain limited samples, but the promises of such approaches rapidly seduced transplant experts. Indeed phenomena at stake when having transplant complications are wide and such an approach could bring a more comprehensive understanding of transplant status or even investigate its ability to predict complex events such as chronic dysfunction and ultimately the transplantation failure [Lim, 2023]. Several studies have investigated the analysis of MRI data to discover new biomarkers to detect acute rejection events early post-transplantation [Shehata, 2016; Shehata, 2020]. More recently, ultrasound imaging radiomic features demonstrated promising results to evaluate kidney function post-transplantation [Hu, 2022; Zhu, 2022]. Authors developed a prediction model utilizing machine learning algorithms and computational image features to predict a decline in estimated glomerular filtration rate (eGFR) using 819 transplant patients. A multi-stage pipeline was built to first automatically segment the cortex, medulla, and central echo complex from ultrasound. Imaging features (104 total) related to shape, intensity statistics, texture, and ultrasound speckle were computed. A random forest classifier was trained to predict 5-year eGFR decline from the feature set (Figure 2.9).

### Relation of imaging features with clinical attributes

While conducting our first experiments on DL feature extractors to encode renal transplant MRI data, we investigated the use of the predefined radiomics features by Van Griethuysen et al. [Van Griethuysen, 2017] as baseline and comparison features. As we have seen in Section 2.1.2, directly aiming at predicting transplantation failure in a fully supervised setup is challenging. As alternative tasks to better apprehend the capacity of expressiveness of computational imaging features, we decided, along with our medical collaborators, to attempt to predict clinical and biological variables known to be linked to graft survival, as described in Chapter 1. Specifically, the GFR, the donor’s age, the transplant incompatibility, the number of previous transplants, and the cold ischemic time variables were considered. The task was built by fixing a threshold on these variables to perform binary classification. To compare to simple deep learning-based generated features, we used the common transfer learning technique, consisting of using a pretrained model (here with a ResNet18 model pretrained on ImageNet and duplicating its weights to 3D) to exploit its representations, features denoted as Transfer Learning-based (TL-based).

Table 2.3: **Evaluation of the predefined radiomics features to recover variables linked to graft survival.** We investigate the ability of predefined radiomics features vs. TL-based features (Transfer Learning-based features from ResNet18 pretrained on ImageNet) to predict clinical variables linked to graft survival. The variables are the M12 Glomerular Filtration Rate (GFR,  $< 45 \text{ mL} \cdot \text{min}^{-1}$ ), the Donor’s Age ( $< 65 \text{ years}$ ), the transplant incompatibility (Trans. Inc.,  $< 3$ ), number of previous transplant (Prev. Trans.,  $> 1$ ), and the cold ischemic time (C. Ischem. T.,  $< 20 \text{ hours}$ ). In the Target var. column, we show the number of positive examples ( $n_{pos}$ ) in the test set of 20 patients. **Bold** indicates the top 1 performing features for each target variable.

Features	Target var.	B.Acc	Prec	Recc	F1	AUC
Radiomics	M12 GFR (12)	68.6 ± 7.7	73.5 ± 4.8	<b>84.9</b> ±8.6	<b>78.7</b> ±6.5	<b>76.6</b> ±1.1
TL-based	M12 GFR (12)	<b>72.9</b> ±5.9	<b>80.0</b> ±7.1	75.0 ± 0.0	77.3 ± 3.2	73.3 ± 0.5
Radiomics	Donor’s Age (6)	53.9 ± 1.8	<b>37.8</b> ±3.1	<b>33.3</b> ±0.0	<b>35.4</b> ±1.4	<b>66.2</b> ±2.4
TL-based	Donor’s Age (6)	<b>55.6</b> ±12.5	30.3 ± 23.4	27.8 ± 28.3	27.9 ± 25.4	59.1 ± 5.5
Radiomics	Trans. Inc. (13)	50.0 ± 0.0	65.0 ± 0.0	<b>100</b> ±0.0	<b>78.8</b> ±0.0	48.8 ± 0.0
TL-based	Trans. Inc. (13)	<b>51.8</b> ±3.4	<b>66.1</b> ±1.9	84.6 ± 0.0	74.2 ± 1.2	<b>64.5</b> ±12.1
Radiomics	Prev. Trans. (3)	55.2 ± 1.5	21.7 ± 2.4	33.3 ± 0.0	26.2 ± 1.7	39.6 ± 3.0
TL-based	Prev. Trans. (3)	<b>62.8</b> ±0.0	<b>22.2</b> ±0.0	<b>66.7</b> ±0.0	<b>33.3</b> ±0.0	<b>58.8</b> ±1.6
Radiomics	C. Ischem. T. (7)	42.1 ± 5.9	33.8 ± 6.4	<b>47.6</b> ±13.5	39.4 ± 9.1	26.7 ± 1.8
TL-based	C. Ischem. T. (7)	<b>63.6</b> ±3.0	<b>80.6</b> ±14.2	33.3 ± 6.7	<b>46.3</b> ±6.1	<b>64.5</b> ±3.7

These preliminary experiments were conducted utilizing a dataset comprising the 87 M12 DCE MRI examinations. The predefined radiomics features were precomputed through the application of the `pyradiomics` Python package on the transplants from the DCE MRI volumes, specifically at the tubular time point. The segmentation maps, derived from the algorithm described in the preceding section, enabled the computation of these predefined radiomics features. The proposed unimodal approach was adopted when the T2 sequence was unavailable for a subset of examinations. A test set comprising 20 patients was partitioned, with the remaining set allocated for training and validation purposes. Subsequently, a Random Forest classifier was

trained, and hyperparameter tuning was accomplished through 3-fold cross-validation. The hyperparameters yielding the best mean validation F1 score were selected for evaluating the model on the test set. Moreover, the identical pipeline was applied to the features acquired through transfer learning, involving inference from a pre-trained ResNet18 on ImageNet, focusing on the region of interest around the transplants within the DCE MRI at the tubular time. The weights of the ResNet18 model were duplicated to a 3D configuration in the depth dimension.

Table 2.3 provides a comprehensive summary of the outcomes derived from the evaluation on the test set comprising 20 patients. Given the relatively limited size of our dataset, necessitating the division of a comparatively small test set to facilitate effective ML model training, it is essential to acknowledge the potential for strong variability in the scores, as evidenced by high standard deviation values. Despite these considerations, a notable observation emerges in the context of M12 Glomerular Filtration Rate (GFR), a crucial biological indicator of kidney function, wherein both sets of features demonstrate comparable performance levels. In contrast, the prediction of transplant incompatibility (Trans. Inc.) exhibits a discernible imbalance between the two main metrics, F1 score and ROC AUC, and between the two sets of features. On the other hand, for other variables (Donor's age, number of previous transplants, and cold ischemic time) these approaches exhibit relatively poor performance, characterized by low F1 scores and ROC AUC values hovering around 60%. The interpretability of these results is correspondingly diminished in comparison to the M12 GFR predictions.

To conclude, this chapter has introduced baseline ML-based methodologies, underscored by the incorporation of key concepts such as radiomics features, which collectively constitute an appropriate set of comparison methods. We also presented a study on the localization of a region of interest around kidney transplants and the generation of a segmentation map of the organ. This segmentation is crucial for computing predefined radiomics features, and the localization holds significance in reducing the dimensionality of our data while retaining pertinent information regarding the transplants by selecting a region of interest. These preliminary experiments serve as a foundational step toward our desired contributions. Specifically, this thesis hypothesizes that learning representations directly from raw MRI data, employing modern deep learning-based feature extractor approaches, can capture more comprehensive, expressive, and robust information for renal transplant monitoring. Subsequent chapters will expound upon the incorporation of clinical attributes into our imaging representations, leveraging this pertinent information to guide and enhance our learning pipelines.



# Chapter 3

## Learning meaningful representations of renal transplant DCE MRI data

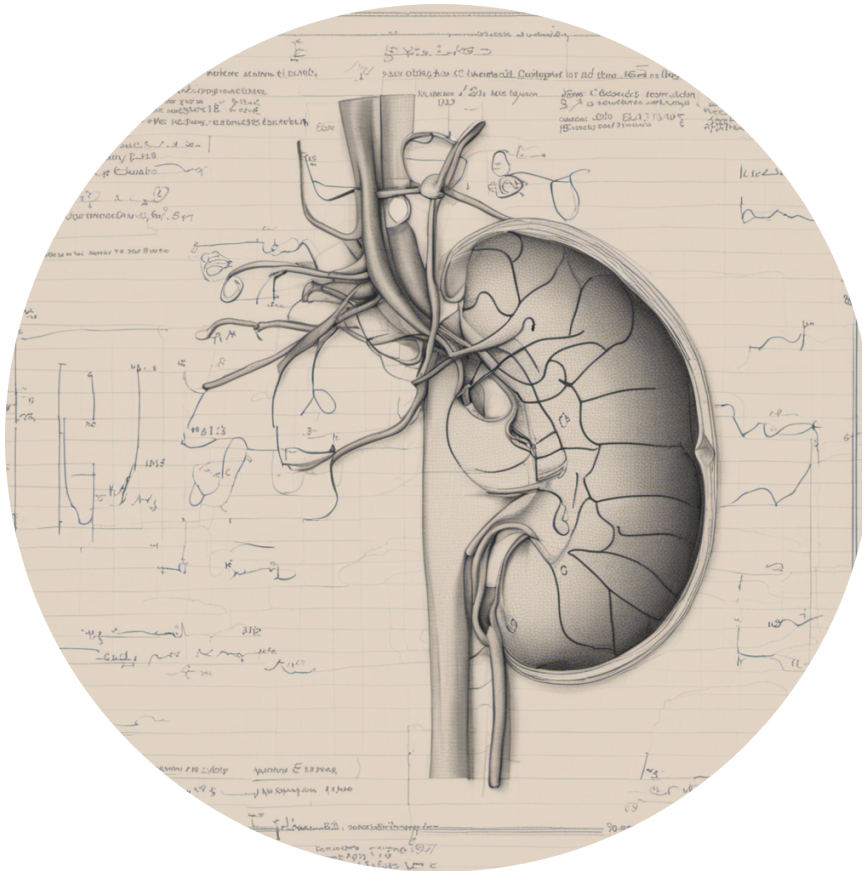
### Contents

---

3.1	Introduction . . . . .	65
3.2	Related work . . . . .	66
3.2.1	Medical imaging in renal transplantation . . . . .	66
3.2.2	Dealing with missing data . . . . .	66
3.3	Methodology . . . . .	67
3.3.1	Contrastive learning for renal transplantation . . . . .	67
3.3.2	Sequential model architecture handling missing data . . . . .	69
3.3.3	Implementation details . . . . .	70
3.4	Data . . . . .	71
3.5	Kidney function forecasting experiments . . . . .	71
3.5.1	Training strategy with curriculum learning . . . . .	72
3.5.2	Feature visualization of the contrastive pretraining . . . . .	73
3.5.3	Renal function forecasting . . . . .	73
3.5.4	Ablation study for missing data strategies . . . . .	74
3.5.5	Increasing the number of missing data . . . . .	75
3.6	Survival analysis experiments . . . . .	76
3.6.1	Survival analysis from imaging features . . . . .	76
3.6.2	CL-based features results . . . . .	77
3.7	Discussion & Conclusion . . . . .	80
3.7.1	Novel methodologies for renal function forecasting . . . . .	80
3.7.2	Deep learning-based imaging features into the clinic . . . . .	81

---





Prompt<sup>1</sup>: "A scientific drawing of a mathematical model for medical data, including a kidney-shaped object in the foreground."

## Summary

In this chapter, we study the use of contrastive learning approaches to learn meaningful representations of renal transplants DCE MRI data. Our approach aims at incorporating clinical information into the learned representation. Then, we propose a novel transformer-based architecture tailored to deal with missing data for the challenging task of serum creatinine prediction 2 years post-transplantation using follow-up imaging modalities. A transformer encoder architecture enables to input the sequential features data per follow-up in order to forecast the renal transplant function and model longitudinal relationships, including a custom method to handle missing exams. In a more clinically oriented study, we further investigate the capacity of the learned representations for graft survival prediction. This work has been published in [Milecki, 2022c; Milecki, 2022a] and a clinical article that is under submission.

<sup>1</sup>Generated using [Stable Diffusion XL](#).





## 3.1 Introduction

As we have in the introduction Chapter 1, renal transplantation appears as the most effective solution for end-stage renal disease and highly improves patients' quality of life, mainly by avoiding periodic dialysis [Suthanthiran, 1994]. However, a substantial risk of transplant chronic dysfunction or rejection persists and may lead to graft loss or ultimately the patient death [Hariharan, 2021]. The genesis of such events takes place in heterogeneous causes, complex phenomena, and results from a gradual decrease in kidney function. In clinical practice, the primary indicator of kidney function is based on blood tests and urine sampling (serum creatinine, creatinine clearance). However, when results are irregular, the gold standard method is needle biopsy, an invasive surgical operation. Thus, the need for a non-invasive alternative that could provide valuable information on transplant function post-transplantation through time is crucial.

Medical imaging plays a significant role in renal transplantation. Beyond the respective limitations of the several imaging modalities, such as the necessity of radiations or the intrinsic trade-off on resolution, to our knowledge, there are no studies focusing on monitoring the evolution of kidney grafts using imaging data. On the other hand, the recent transformer models [Vaswani, 2017] offer new directions in processing sequential data. Moreover, recent advances in self-supervised learning [Taleb, 2020] enable the training of powerful deep learning representations with a limited amount of data. Renal transplantation datasets usually belong to this case, making the use of such methods the way to move forward. Our study is among the first that explore such methods for renal transplantation, solving challenging clinical questions.

In this work, we propose a method to forecast renal transplant function through serum creatinine prediction from follow-up exams of Dynamic Contrast-Enhanced (DCE) MRI data post-transplantation. The main contributions of this work are twofold. First, we propose the use of contrastive schemes, generating informative manifolds of DCE MRI exams of patients undergoing renal transplantation. Different self-supervised and weakly-supervised clinical pertinent tasks are explored to generate relevant features using the cosine similarity. Secondly, we introduce a transformer-based architecture for forecasting serum creatinine score, while proposing a tailored method to deal with missing data. In particular, our method uses a key mask tensor that highlights the missing data and does not take them into account for the training of the sequential architecture. Such a design is very robust with respect to the position and number of missing data, while it provides better performance than other popular data imputation strategies. To the best of our knowledge, our study is among the first that propose a novel, robust, and clinically relevant framework for forecasting serum creatinine directly from imaging data.

This chapter is organized as follows: we first introduce related work in Section 3.2, with a focus on medical imaging in transplantation and approaches to deal with missing data. In Section 3.3, we present our methodology on self-supervised contrastive learning for renal transplantation. Then we describe the data used in this work in Section 3.4 and the implemented experiments and obtained results to demonstrate the soundness of our proposed approach in Section 3.5. Subsequently, Section 3.6 introduces a clinical study built

from the previously presented models to investigate graft survival analysis from the learned representations of the imaging data. Section 3.7 provides a discussion and conclusion to this work.

## 3.2 Related work

### 3.2.1 Medical imaging in renal transplantation

Several medical imaging approaches investigated the diagnosis of renal transplant dysfunction. In [Sharfudin, 2014], diverse imaging modalities have been investigated to assess renal transplant functions in several studies. Moreover, we used in section 2.2 multiple Magnetic Resonance Imaging (MRI) modalities for the unsupervised kidney graft segmentation. Recent studies focused on detecting specific events such as renal fibrosis [Orlacchio, 2014] or acute rejection [Khalifa, 2013]. In [Shehata, 2020], multimodal MRI and clinical data are explored to assess renal allograft status at the time of the different exams. Most of those approaches seek to, indirectly through related events or directly through complex automated systems, non-invasively retrieve structural, functional, and molecular information to diagnose chronic kidney disease [Alnazer, 2021].

### 3.2.2 Dealing with missing data

When it comes to real clinical settings, a limited amount and missing data is one of the most important issues during data curation. Handling of missing data has been thoroughly studied by data imputation methods, which mainly propose approaches to fill the missing data as a pre-processing step to some downstream task [Mackinnon, 2010]. Beyond simple statistical approaches such as sampling the mean or median of available data, methods can be categorized into two groups: discriminative and generative approaches. The former is mainly developed for structural data (discrete or continuous) with methods such as structured prediction [Keshavan, 2009]. On the other hand, generative approaches include expectation-maximization algorithms [García-Laencina, 2010] or deep learning models such as Generative Adversarial Imputation Nets (GAIN) [Yoon, 2018]. Those latest approaches showed very good performance for medical image tasks, as proposed in [Dalca, 2019; Xia, 2021]. However, the training of such models usually is subjective to a large amount of data that are not available all the time [Kazemina, 2020], especially in a clinical setting.

Considering the use of the transformer models, the attention mechanism showed promising results in missing data imputation for structural [Wu, 2020] and trajectory data [Becker, 2021; Giuliari, 2020]. In particular, the attention mask was used to investigate the robustness of a vanilla encoder-decoder transformer and a Bidirectional Transformer (BERT) model [Devlin, 2018] while missing 1 to 6 point's coordinates out of 32 for forecasting the people trajectories. Among all these methods, our method is the first to handle in an efficient and robust way missing data with high dimensionality, tested on sequences with long-time dependencies.

### 3.3 Methodology

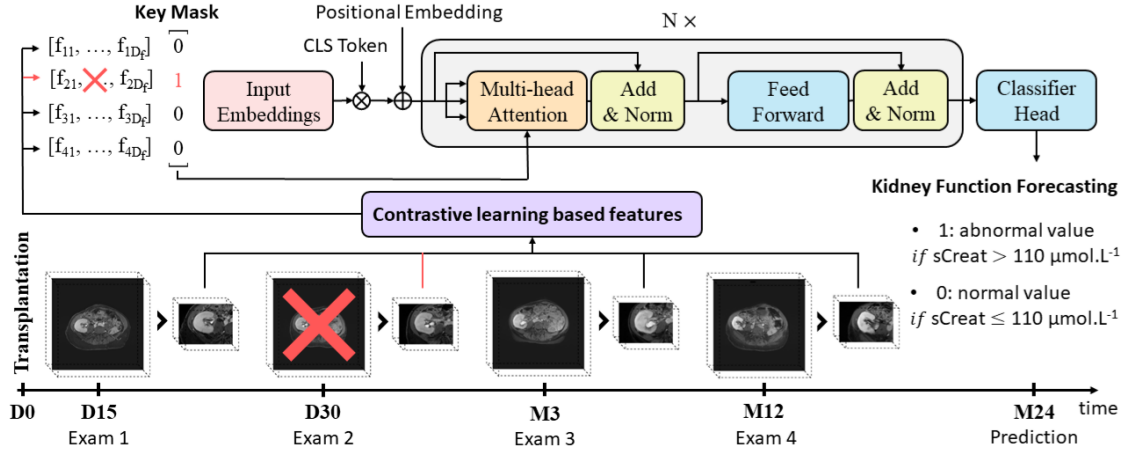


Figure 3.2: **Overview of the proposed method.** Different contrastive schemes are used to represent the different MRIs. These features are used to train a sequential model coupled with a key mask tensor to mark the missing data.

In this study, we focus on the prediction of serum creatinine from imaging data and in particular DCE MRI, exploring both anatomical and functional information. An overview of our method is presented in Figure 3.2.

#### 3.3.1 Contrastive learning for renal transplantation

In this work, we propose two contrastive learning schemes to explore meaningful data representations: (a) a self-supervised scheme, where we learn meaningful features by solving the proxy task of determining if two MRI volumes belong to the same patient, and (b) a weakly-supervised scheme, where we discriminate samples based on the differences in the value of various clinical variables.

Let us denote  $(v_1, v_2) \in (\mathbb{R}^{N_x \times N_y \times N_z})^2$  a pair of MRI regions of interest. Each stream  $i = 1, 2$  consists of a ResNet model to extract a latent representation from the MRI volumes, which takes  $v_i$  as input and outputs features  $z_i \in \mathbb{R}^{D_f}$ , with  $D_f = 512$  for ResNet18. Then, a feature embedding head associates these features with the underlying task. This is modeled by a linear layer or a Multi-Layer Perceptron (MLP) mapping the features to  $(z'_1, z'_2) \in \mathbb{R}^{D_{fe}}$ , with  $D_{fe} = 256$ .

### Self-supervised pretraining

Our first strategy relies on a self-supervised task at the patient level, i.e., we train a model to distinguish if a pair of volumes comes from the same patient or not.  $P_j = \{v \in \mathbb{R}^{N_x \times N_y \times N_z} | v \text{ from patient } j\}$  for  $j \in \llbracket 1, N_p \rrbracket$ , where  $N_p$  denotes the number of patients, the set of available volumes from MRI series for each exam and patient. Then, our proxy task is to discriminate pairs by knowing if they belong or not to the same patient, i.e.,  $y = 1$  if  $\exists j (v_1, v_2) \in (P_j)^2$ ; else  $y = 0$ .

### Weakly-supervised various clinical pretraining

Our second strategy discriminates samples based on the difference of certain clinical variable's value, i.e.,  $y = 1$  if  $\|\text{Var}(v_1) - \text{Var}(v_2)\| < \theta$ ; else  $y = 0$ , where  $\text{Var}(\cdot)$  is a clinicobiological variable and  $\theta$  a clinically relevant threshold. The clinicobiological variables are suggested by nephrology experts to encode clinical priors and information, as they are significantly linked to graft survival [Loupy, 2019]. In this study, we investigate three variables: (1) the transplant incompatibility, (2) the age of the transplant's donor, and (3) the Glomerular Filtration Rate (GFR) value.

### Training loss

From the embedded features  $(z'_1, z'_2)$ , the optimization is done by the following cosine embedding loss:

$$\text{CosEmbLoss}(z'_1, z'_2, y) = \begin{cases} 1 - \cos(z'_1, z'_2), & \text{if } y = 1, \\ \max(0, \cos(z'_1, z'_2) - m), & \text{if } y = 0, \end{cases} \quad (3.1)$$

where  $\cos$  refers to the cosine similarity and  $m$  a margin scalar between  $-1$  and  $1$ . This loss enforces the model to build relevant features that adequately express the kidney transplant imaging and define the way to create strategies to label  $y$  each pair (see Figure 3.3).

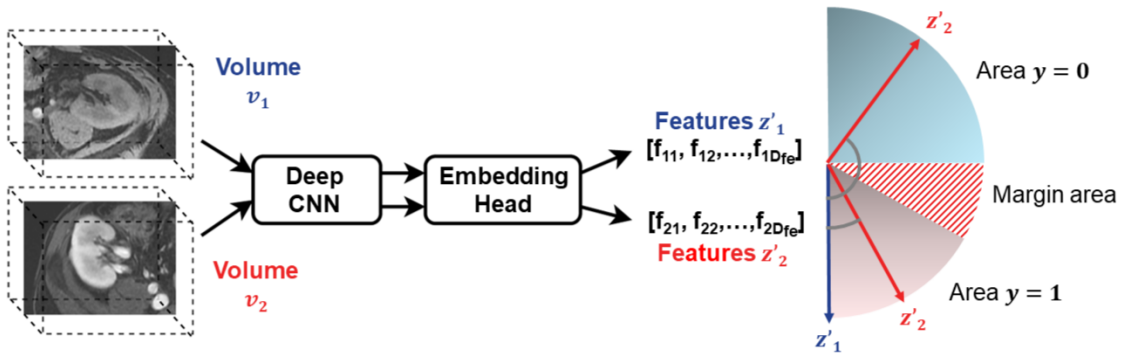


Figure 3.3: **Contrastive learning objective explanation.** Our pretraining phase is based on a two stream approach. Pairs of MRI volumes, labeled  $y = 0, 1$ , are fed to a Deep CNN and an embedding head to produce features. According to the label  $y$ , our CosEmbLoss enforces the paired features to be near or far using the cosine similarity distance, as depicted in 2D with the angles between the normalized feature vectors.

### Training scheme and curriculum learning

Since the dimensionality of our data is very high and the tasks we investigate are very challenging, we apply curriculum learning to facilitate the training process. In particular, for the self-supervised task at the patient level, pairs from the same exam of each patient are enabled in the beginning until half of the training, while they are discarded in the second half.

For the weakly-supervised task based on clinicobiological variables, the perplexity of the task is determined by the thresholds  $\theta$ . More specifically, the training labels are adjusted every  $e_i$  epochs following the rule:  $y = 1$  if  $|\text{Var}(v_1) - \text{Var}(v_2)| < \theta_{i,1}$ ;  $y = 0$  if  $|\text{Var}(v_1) - \text{Var}(v_2)| > \theta_{i,2}$ ; else discard the pair  $(v_1, v_2)$ , where  $\theta_{i,1}, \theta_{i,2}$  are set in the image of  $\text{Var}(\cdot)$  satisfying  $\forall i$  (1)  $\theta_{i,1} \leq \theta_{i,2}$ ; (2)  $\theta_{i+1,1} \leq \theta_{i,1}$ ; and (3)  $\theta_{i,2} \leq \theta_{i+1,2}$ . Our loss enforces the feature pairs to be near or far in the feature embedding space, depending on the label  $y$ . The condition (1) enables to form a grey area between the two cases, while the conditions (2) and (3) strengthen the constraint through epochs on the difference of value  $\text{Var}(\cdot)$  between the two pairs to be correctly arranged (examples in Figure 3.4).

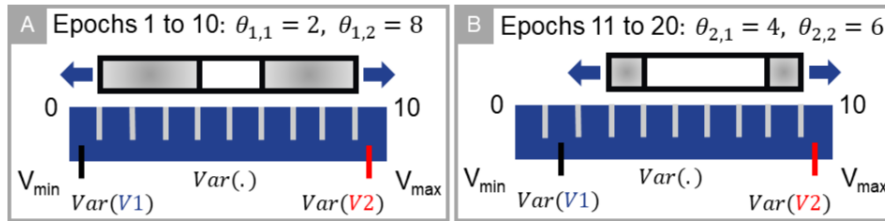


Figure 3.4: **Curriculum learning on the weakly-supervised task example.** In this example, the procedure to set the pairs' labels from the variable  $\text{Var}(\cdot)$  is less restrictive in the first phase of training (A: first ten epochs) than in the second phase (B: next ten epochs) where the grey area reduces and the weakly-supervised task becomes more challenging.

### 3.3.2 Sequential model architecture handling missing data

Our forecasting model takes as input  $T = 4$  features  $z \in \mathbb{R}^{D_f}$  corresponding to the different follow-ups and relies on a transformer encoder architecture [Vaswani, 2017]. First, these features are mapped to embeddings of size  $D_{model}$  using a linear layer, while a special classification token (CLS) is aggregated in the first position to generate an embedded sequence. Then, the core of the transformer encoder architecture stacks  $N$  layers on top of learned positional embeddings added to the embedded sequences. Each layer is first composed of a multi-head self-attention sub-layer, which consists of  $h$  heads running in parallel. Each head is based on the scaled dot-product attention. Then, a position-wise fully connected feed-forward sub-layer applies an MLP of hidden dimension  $D_{model}$  to each position separately and identically. Finally, to perform the classification task, the CLS token output is fed to a linear layer.

### Strategy for missing data

Our proposed strategy to deal with missing data is applied to the scaled dot product operation, the core of each multi-head self-attention sub-layer. For simplicity, we consider here a sub-layer with one head,  $h = 1$ . The operation takes as input the query  $Q$ , key  $K$  and value  $V$ , which are linear projections of the embedded sequences, with  $d_q$ ,  $d_k$  and  $d_v$  dimensions, respectively and performs  $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^t}{\sqrt{d_k}})V$ . In this work, we build a key mask tensor  $m_k \in \mathbb{R}^T$  based on the availability of exams for each patient so that zero attention is given to missing data both during the training and inference times, i.e.  $\forall t \in \llbracket 1, T \rrbracket$   $m_k[t] = -\infty$  if exam  $t$  is available else 0. Thus, our mask cancels the attention on missing exams by  $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^t}{\sqrt{d_k}} + M_k)V$  where  $M_k = \llbracket m_k m_k \dots m_k \rrbracket \in \mathbb{R}^{T \times d_k}$ . For  $h > 1$ , keys, values, and queries are linearly projected  $h$  times with different, learned linear projections, concatenated, and once again projected after the scaled-dot product.

### 3.3.3 Implementation details

Starting with the contrastive learning, we used data augmentation with horizontal flipping and random affine transformation with a 0.5 probability, as well as random Gaussian blur ( $\sigma \in [0, 0.5]$ ) and random Gaussian noise ( $\sigma \in [0, 0.05]$ ), using TorchIO python library [Pérez-García, 2021]. Having approximately a set of pairs of  $\binom{V}{2} = \frac{V(V-1)}{2}$ , where  $V$  is the number of available volumes, we proposed to fix the training set size to  $V_t = 5000$ . We decided to fix the number of positive samples, as well as its balance to 25%, and to randomly sample every epoch the remaining from the negative samples.

Concerning the optimization of our models, a 10% dropout has been used for the linear layers of both the contrastive and sequential models. For the contrastive model, the Stochastic Gradient Descent optimizer with a momentum equal to 0.9 was used with a starting learning rate of  $10^{-2}$  following a cosine schedule and preceded by a linear warm-up of 5 epochs. The batch size was set to 20 and the model trained for 60 epochs on 4 NVIDIA Tesla V100 GPU using Pytorch [Paszke, 2019]. For the transformer, a binary cross-entropy loss (BCE) was used when binarizing the serum creatinine value using a threshold of  $110 \mu\text{mol.L}^{-1}$ , specified by nephrology experts, as a clinically relevant value to assess normal/abnormal renal transplant function at a specific time point. Adam optimizer [Kingma, 2014b] was used with a starting learning rate of  $10^{-4}$  following the same learning rate scheduler. The batch size was set to 32 and the model was trained for 30 epochs on 1 NVIDIA Tesla V100 GPU. The architecture's hyperparameters were set by grid search and 10-fold cross-validation, providing  $N = 2$ ,  $h = 2$ ,  $D_{model} = 768$ .

### 3.4 Data

Our study was approved by the Institutional Review Board, which waived the need for patients' consent. The data cohort corresponds to study reference ID-RCB: 2012-A01070-43 and ClinicalTrials.gov identifier: NCT02201537. All the data used in this study were anonymized. Overall, our imaging data are based on the DCE MRI series collected from 89 subjects at 4 follow-up exams which took place approximately 15 days (D15), 30 days (D30), 3 months (M3), and 12 months (M12) after the transplant surgery, resulting in respectively 68, 75, 87, and 83 available scans at each follow-up.

The MRI volumes sized  $512 \times 512 \times [64 - 88]$  voxels included spacing ranging in  $[0.78 - 0.94] \times [0.78 - 0.94] \times [1.9 - 2.5]$  mm. All volumes were cropped around the transplant using the automatic selection of the region of interest presented in section 2.2 in order to reduce dimensionality while no information about the transplant is discarded. Intensity normalization was executed to each volume independently by applying standard normalization, clipping values to  $[-5, 5]$ , and rescaling linearly to  $[0, 1]$ .

As a primary indicator of the kidney function assessment, all patients were subject to blood tests regularly a few days before the transplantation to several years after, to measure the serum creatinine level in  $\mu\text{mol.L}^{-1}$ . The serum creatinine target prediction value is calculated as the mean over an interval of two months before and after the prediction date, 2-year post-transplantation (M24).

### 3.5 Kidney function forecasting experiments

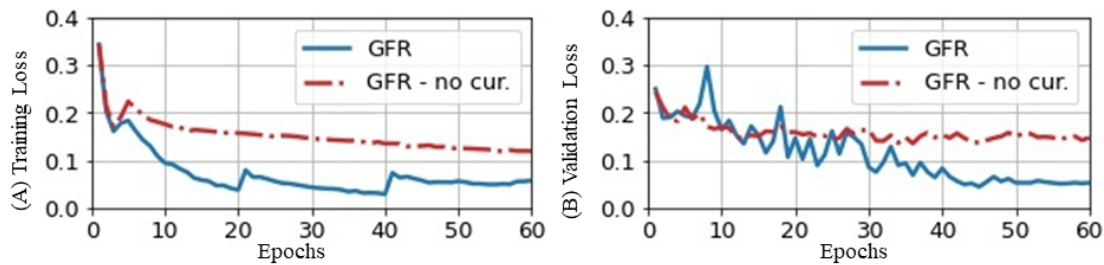


Figure 3.5: **(A) Training and (B) Validation curves for our contrastive learning scheme based on GFR variable with and without curriculum learning.** No-cur. stands for no curriculum learning at training. Overall, the curriculum learning contributes to lower training and validation losses and helps the training process.



### 3.5.1 Training strategy with curriculum learning

In Figure 3.5, we plot the training and validation loss curves of our contrastive learning pretraining scheme using the GFR variable. We demonstrate the relevance of using curriculum learning (in blue) in our contrastive objective as we obtain significantly lower training and validation losses at the end of the training. When we update the threshold on the GFR difference to set positive and negative pairs, we observe peaks at epochs 20 and 40 on the training loss when the contrastive objective becomes more complex. The validation loss was calculated on the same set (the most complex one) during the whole training. We clearly see the validation loss stabilizing after epoch 40 when the objective of the validation set matches the one of the training set.

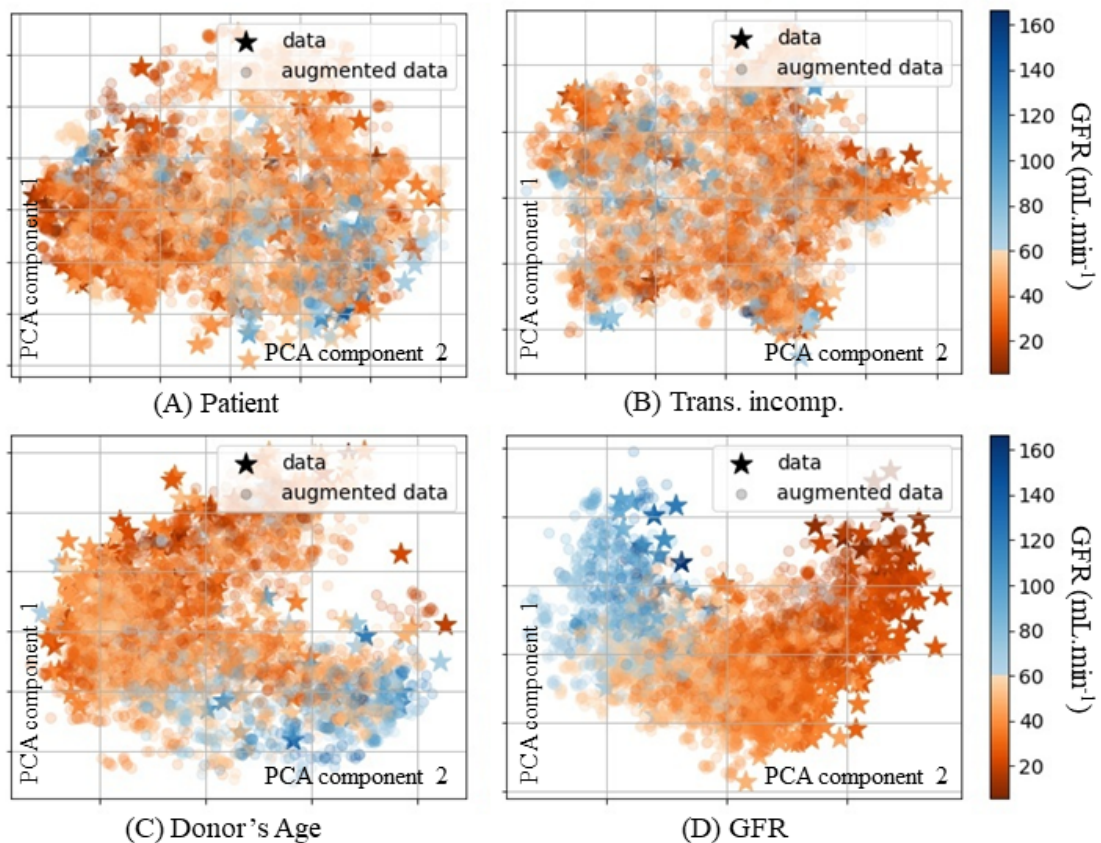


Figure 3.6: **Feature visualization of the PCA decomposition of the last layer of the contrastive training using the DCE MRI follow-ups.** Our contrastive tasks, including (A) identifying exams of the same patient, (B) transplant incompatibility (Tran. incompat.), (C) donor's age, and (D) GFR variables. Colormap is set by the GFR value for each follow-up,  $< 60 \text{ mL}\cdot\text{min}^{-1}$  corresponds to abnormal kidney function. With stars, we indicate the real data, and with circles the augmented ones. Overall, the Donor's Age and the GFR trainings seem to provide better representations with respect to the GFR.

### 3.5.2 Feature visualization of the contrastive pretraining

We visualize the representations of DCE MRI from the trained image encoder using PCA decomposition. Figure 3.6 shows the different reduced feature spaces obtained by the models trained using the different clinicobiological variables. Colormap is set by the GFR value obtained at the corresponding MRI follow-up exam and is used as an indicator of renal function ( $< 60 \text{ mL}\cdot\text{min}^{-1}$  corresponds to abnormal kidney function). We evaluate the clustering of the imaging features toward their renal function evaluation ability, adding augmented images to check the tendency better. We observe that the (C) Donor's Age and (D) GFR pretrainings seem to provide the best representations toward kidney function evaluation. While the (B) Trans. incomp. and the unsupervised (A) Patient pretrainings only demonstrate some group patterns with less visible clusters.

### 3.5.3 Renal function forecasting

To evaluate the performance of our proposed method and compare it with other strategies for the forecasting of serum creatinine, four evaluation metrics are used: recall, precision, F1 score, and the area under the receiver operating characteristic curve (ROC AUC). A testing set of 20 patients is separated from the train set and used to validate the performance of our models. We perform a 10-fold cross-validation (CV) on the train set (69 patients) and report the mean (standard deviation) scores in % for each fold. During CV, the model reaching the minimum loss is saved, and an ensemble approach is used to make the final prediction on the test set from models, that reach more than 50% ROC AUC out of the 10 folds.

We compare our sequential model to an LSTM model, which is a commonly used architecture for sequential data, and which architecture was set using the same approach as our main model, resulting in 2 LSTM cells and a hidden size of 768. Additional sets of feature representations were used to compare the significance of our approach. First simple statistics from the serum creatinine captured from the available blood test results between each follow-up (number of points, mean, median, standard deviation, minimum, maximum) are calculated and used as input to the models. Second, a set of predefined radiomics features [Van Griethuysen, 2017] is obtained from the segmentation of the kidney transplant following the unsupervised method presented in section 2.2. Finally, we investigate generating MRI features from SimCLR [Chen, 2020a] contrastive scheme, while we report the performance of different transfer-learning approaches, pretrained on ImageNet [Deng, 2010] by duplicating the weights to 3D, Kinetics [Smaira, 2020], and medical image datasets MedicalNet [Chen, 2019].

Quantitative results for all the methods are reported in Table 3.1. Our proposed approach outperforms the rest of the methods for the test. Both LSTMs and transformers architectures report good performances, with only a few models reporting performance lower than 60% on every metric. Interestingly, our method outperforms the sCreat model which models directly the serum creatinine level. Moreover, our GFR contrastive-based features report the best performance among all the other features for both LSTMs and transformer formulations. Limitations appear as our model seems to misclassify cases where the patient's serum creatinine is

stable and close to the used threshold, during the first two years post-transplantation.

Table 3.1: **Quantitative evaluation of different methods.** Our contrastive tasks, including: self-supervised on identifying exams of the same patient (patient), weakly-supervised on transplant incompatibility (Tran. inc.), donor’s age and GFR variables vs comparison methods: serum creatinine (sCreat), predefined radiomics features, pretrainings on imagenet, kinetics, medicalnet, and another self-supervised contrastive loss (SimCLR). We report precision (Prec), recall (Rec), F1 score, and ROC AUC (AUC) for both LSTM and Transformer methods. Overall, the GFR with Transformers reports the highest metrics for the test set. **Bold** indicates the best-performing methods.

Method	Features	Validation				Test			
		Prec	Rec	F1	AUC	Prec	Rec	F1	AUC
LSTM	sCreat	80.5(12.3)	62.9(21.0)	71.1(13.8)	80.4(22.4)	83.3	76.9	80.0	83.5
	Radiomics	86.2(14.9)	73.5(15.5)	77.3(8.2)	80.7(16.0)	90.9	76.9	83.3	84.6
	Imagenet	85.5(15.0)	68.0(17.7)	74.0(12.8)	91.0(10.8)	90.9	76.9	83.3	81.3
	Kinetics	<b>90.7(9.4)</b>	74.0(21.5)	78.5(11.3)	<b>91.4(8.5)</b>	92.3	92.3	92.3	85.7
	MedicalNet	86.5(13.9)	78.5(21.2)	79.8(13.2)	82.7(18.8)	57.1	61.5	59.3	41.8
	SimCLR	79.8(15.9)	86.5(24.1)	80.9(17.2)	91.8(13.7)	72.2	<b>100.0</b>	83.9	64.8
	Patient	83.5(15.7)	68.0(21.0)	73.7(18.0)	84.2(16.3)	90.9	76.9	83.3	81.3
	Tran. inc.	81.0(29.1)	56.0(25.5)	63.8(24.9)	82.9(14.1)	75.0	46.2	57.1	68.1
	Donor’s age	79.4(15.5)	87.5(13.7)	81.8(10.3)	81.6(13.8)	84.6	84.6	84.6	90.1
GFR	82.8(9.6)	<b>95.5(9.1)</b>	<b>88.3(7.7)</b>	88.3(13.1)	86.7	<b>100.0</b>	92.9	<b>98.9</b>	
Transformer	sCreat	79.0(28.7)	60.2(31.1)	65.4(29.3)	71.6(24.2)	81.3	<b>100.0</b>	89.7	86.8
	Radiomics	81.3(15.7)	66.0(28.6)	69.1(20.2)	65.3(30.5)	90.9	76.9	83.3	91.2
	Imagenet	58.4(22.4)	76.5(34.8)	65.8(27.5)	45.5(21.6)	65.0	<b>100.0</b>	78.8	58.2
	Kinetics	53.2(35.8)	66.0(44.8)	58.3(38.9)	64.0(19.7)	65.0	<b>100.0</b>	78.8	83.5
	MedicalNet	65.5(27.9)	58.0(33.2)	58.3(28.3)	64.8(19.6)	75.0	46.2	57.1	50.6
	SimCLR	58.9(30.9)	75.5(38.7)	65.6(33.2)	64.8(23.5)	68.4	<b>100.0</b>	81.3	72.5
	Patient	84.3(15.8)	73.0(21.6)	74.8(14.8)	58.2(32.3)	90.9	76.9	83.3	83.5
	Tran. inc.	73.2(27.9)	62.0(23.0)	66.2(23.7)	68.3(16.1)	88.9	61.5	72.7	72.5
	Donor’s age	79.1(17.3)	73.0(26.6)	71.9(18.4)	49.7(30.8)	91.7	84.6	88.0	89.0
GFR	86.3(20.9)	71.5(22.7)	77.4(20.6)	79.7(20.7)	<b>92.9</b>	<b>100.0</b>	<b>96.3</b>	<b>98.9</b>	

### 3.5.4 Ablation study for missing data strategies

The proposed key mask padding approach for handling missing data is specific to the attention mechanism, hence the transformer model. Thus, we investigate 3 other missing data strategies applicable to both the transformer and LSTM model: (1) filling with zeros strategy (None), (2) filling with the nearest available exam (N.A.), and (3) taking the mean for intermediate exams and fill for first and last (M.+N.A.). Results presented in Table 3.2 are obtained with the best-performing imaging features (proposed using the GFR value).

Our proposed approach to handling missing data reports the best precision, recall, and F1 score and the second-best ROC AUC on the test set. Overall, the different strategies report better performance on transformer-based architectures than the LSTM ones indicating the interest in using such models for this task. Moreover, the M.+N.A. strategy reports a lower precision rate for both LSTM and our sequential model, affirming the difficulty of interpolating imaging features. Both None and N.A. strategies appear to report competitive results, lower however than our proposed.

Table 3.2: **Quantitative evaluation of different strategies for missing data.** With none we denote the filling with zero strategy, N.A. the filling with the nearest neighbor exam, and with M.+N.A. the filling with the mean and nearest neighbor exam. **Bold** indicates the top-performing combination.

Method	Strategy	Prec	Validation			AUC	Test			
			Rec	F1			Prec.	Rec	F1	AUC
LSTM	None	80.5(11.5)	81.0(14.3)	80.0(9.6)	73.6(16.9)	86.7	<b>100.0</b>	92.9	98.9	
	N.A.	82.8(9.6)	<b>95.5(9.1)</b>	<b>88.3(7.7)</b>	<b>88.3(13.1)</b>	86.7	<b>100.0</b>	92.9	98.9	
	M.+N.A.	81.1(10.8)	93.5(10.0)	86.1(6.8)	84.2(11.0)	81.3	<b>100.0</b>	89.6	96.7	
Transformer	None	86.2(12.9)	78.5(23.2)	79.7(18.2)	71.5(25.3)	92.3	92.3	92.3	98.9	
	N.A.	88.8(20.6)	75.5(21.5)	81.3(20.8)	80.5(22.2)	92.3	92.3	92.3	96.7	
	M.+N.A.	<b>90.5(12.3)</b>	73.5(17.3)	80.0(12.7)	80.0(18.3)	76.5	<b>100.0</b>	86.7	<b>100.0</b>	
	Proposed	86.3(20.9)	71.5(22.7)	77.4(20.6)	79.7(20.7)	<b>92.9</b>	<b>100.0</b>	<b>96.3</b>	98.9	

### 3.5.5 Increasing the number of missing data

To further analyze the relevance of our approach to the missing data strategy, we artificially increased the proportion of missing exams in two setups: first both in the train and test set, and second only in the train set. 23% corresponds to the inherent missing data proportion in our training dataset. We compare to the strategies of filling with zeros (None) and filling with the nearest available exam strategy (N.A.). In Table 3.3, we observe a lower decrease in both F1 and ROC AUC metrics in both setups and that our proposed approach of handling missing exams with an attention mechanism demonstrates better robustness.

Table 3.3: **Quantitative evaluation for an increasing number of missing data for our proposed method and other imputation strategies.** 23% corresponds to the inherent missing data in our training dataset. With none, we denote the filling with zero strategy, N.A. the filling with the nearest neighbor. We report F1 score and ROC AUC. The first five columns correspond to the percentage of missing values on both training and test, with the last five for only the training set while the test set has 25% missing data. Overall, our method reports better robustness than the other two commonly used strategies for both experiments.

Strategy	Missing data Metric	23%	30%	40%	50%	60%	23%	30%	40%	50%	60%
		in train & test set						in train set			
None	F1	92.3	92.3	92.3	85.7	81.8	92.3	88.0	88.0	87.0	87.0
	ROC AUC	98.9	95.6	97.8	82.4	90.1	98.9	97.8	96.7	98.9	96.7
N.A.	F1	92.3	91.7	87.0	88.9	85.7	92.3	96.0	91.7	86.7	86.7
	ROC AUC	96.7	98.9	96.7	91.2	91.2	96.7	95.6	98.9	97.8	96.7
Proposed	F1	96.3	91.7	88.0	92.3	89.7	96.3	91.7	96.0	92.3	92.3
	ROC AUC	98.9	98.9	97.8	95.6	93.4	98.9	100.0	96.7	97.8	96.7

### 3.6 Survival analysis experiments

In a second phase, we investigated further use of the models developed previously in a more clinically relevant oriented context. This section introduces a clinical study, built on top of the contrastive learning models presented in this chapter, that aims at investigating the ability of the produced deep learning-based features of renal transplants to predict graft survival. In this study, we hypothesized that imaging of the entire allograft, particularly using MRI, can capture comprehensive information about potential insults. Consequently, we proposed an innovative approach suggesting that early dynamic contrast-enhanced (DCE) MRI examinations of renal transplants can offer pertinent information for predicting graft survival.

#### 3.6.1 Survival analysis from imaging features

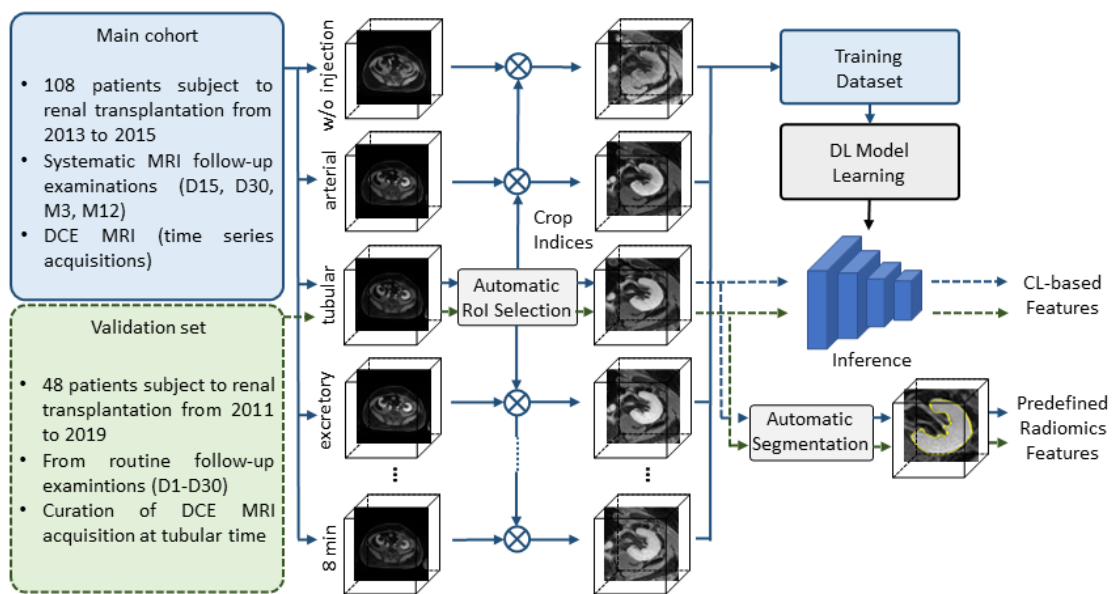


Figure 3.7: **Pipeline for generating our CL-based features and the predefined radiomics features.** In our main cohort, all the DCE MRI time series are available and the one at tubular time is used to select the region of interest around the kidney and applied to the other volumes of the same examination. All these volumes build our training set for our contrastive learning models. The tubular time volume is again used to generate the features. In the validation set, only the tubular time is available and the same processing is applied. Comparisons with the predefined radiomics features are obtained from the segmentation masks on the transplant as described in 2.

We denote CL-based (Contrastive Learning-based) features the generated imaging features from our CNN models pretrained using contrastive learning (see Figure 3.7). For this study, we curated a separated validation set, as explained in Section 1.3 and Figure 1.11. After generating our CL-based features features at inference from our model, the survival analysis of transplant failure event was performed using classical penalized Cox models [Cox, 1972] and Kaplan-Meier curves [Kaplan, 1958]. The survival analysis was performed in a classical statistical learning pipeline: preprocessing (standard normalization of the features), hyperparameter tuning

(regularization parameters) and feature selection using 3-fold cross-validation on the main cohort, and testing in the validation set. To obtain two groups of patients from our CL-based features, we used the k-nearest neighbors [Cover, 1967] unsupervised clustering algorithm. Then we analyzed the statistical significance of those two groups regarding transplantation failure event using Kaplan-Meier curves and the logrank test.

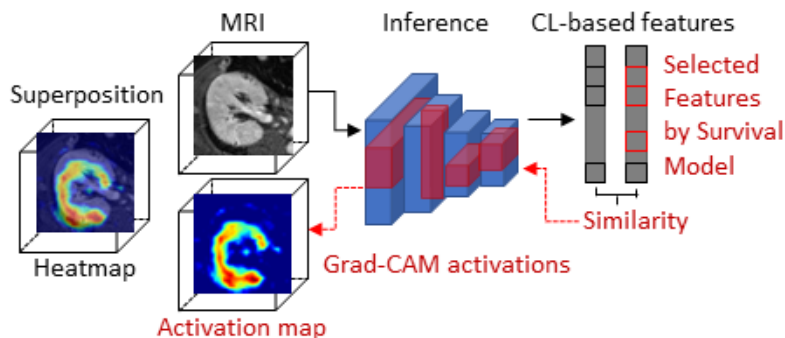


Figure 3.8: **Heatmaps visualization pipeline.** The similarity between the complete set of features and the ones selected by the survival model was used to compute the Grad-CAMS [Selvaraju, 2017] activations of our model. The activation map was then superposed to the corresponding input MRI volume.

To further demonstrate the clinical relevance of our CL-based model, we visualized heatmaps of the activation of our model from the image input using explainability in AI approaches [Selvaraju, 2017]. For each MRI volume, we used the cosine similarity between the initial feature vector and the selected feature vector (zero-ing the non-selected ones) to monitor the gradient activation maps in the model at different layers. Figure 3.8 presents an overview of the approach to obtain those heatmaps.

### 3.6.2 CL-based features results

#### Predictive capability of our CL-based features

Preliminary extensive experiments on our main cohort (Figure 3.9) led us to further design this study toward the DCE MRI modality at early-stage post-transplantation. The predictive capability of our CL-based features is presented in Table 3.4. The concordance index (C-index) scores were obtained through 3-fold cross-validation on the training set and testing on the validation set. Baseline comparisons were made with the predefined radiomics features [Van Griethuysen, 2017] extracted using segmentation masks on the grafts, both with and without pre-transplant clinical attributes.

Our CL-based features demonstrated superior performance compared to clinical attributes (listed in Section 1.3.1) and the predefined radiomics features, achieving a C-index of 72.7% in the validation set. Additionally, we investigated the combination of our CL-based features with clinical attributes and found that their concatenation (line 5) marginally improved the results, yielding a C-index of 73.5%. This comparison suggests that our features hold complementary predictive information to those clinical attributes.

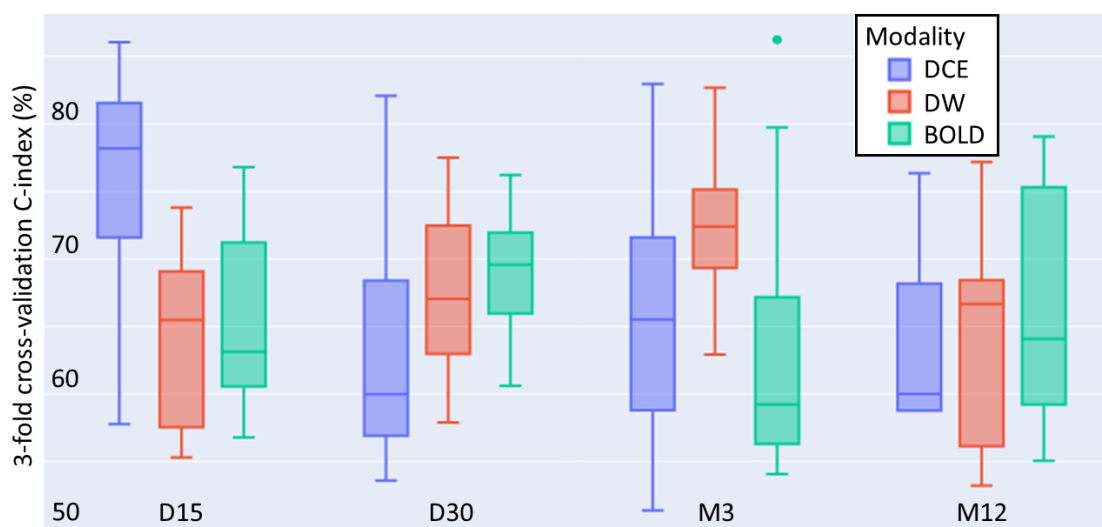


Figure 3.9: **Boxplots of the evaluation of different CL-based features from different examination timestamps and MRI modalities available in our main cohort.** Results on the main cohort for the different timestamps (D15, D30, M3, and M12) and MRI modalities (DCE, DW, and BOLD) represented as boxplots. We report the mean 3-fold cross-validation C-index in percentage (%) on the main cohort.

Table 3.4: **Concordance index (C-index in %) scores on evaluating our CL-based features to predict graft survival at five years.** We report the mean  $\pm$  std on 3-fold cross-validation on the train set and a test score on the validation set.

Features	3-fold CV train set (mean $\pm$ std in %)	Validation set (in %)
Clinical attributes	69.4 $\pm$ 17.8	66.5
Predefined radiomics features	65.4 $\pm$ 1.6	55.5
Predefined radiomics features + Clin. att.	76.9 $\pm$ 11.0	61.7
Our CL-based features	84.6 $\pm$ 13.1	72.7
Our CL-based features + Clin. att.	86.1 $\pm$ 8.6	73.5

*CV=Cross-Validation; Clin. att.=Clinical attributes.*

### Unsupervised stratification of transplants' outcome

Unsupervised stratification of transplant outcomes was performed using our CL-based features and unsupervised clustering methods. The patients were divided into two groups based on this stratification. Figure 3.10 illustrates the obtained stratification regarding graft survival outcomes using Kaplan-Meier curves.

We observed significant stratification of patients concerning transplant survival events in both the training set used for training the CL-based model and the validation set. The p-values associated with these stratifications were 0.057 and 0.029, respectively. These results demonstrate the potential capability of our model to discriminate between different pathological trajectories effectively.



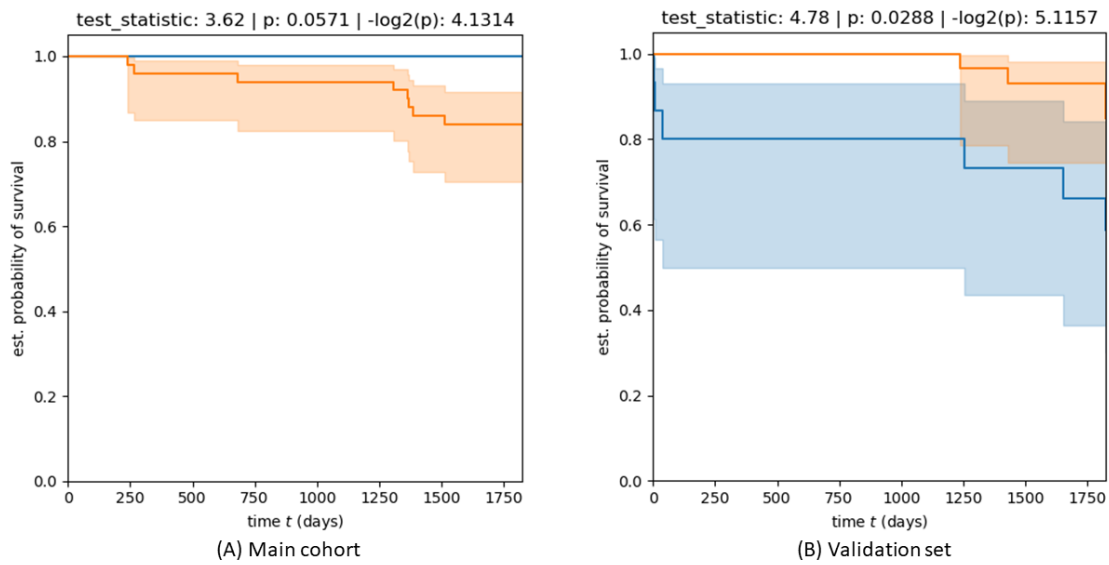


Figure 3.10: Kaplan-Meier curves for the two groups obtained by the unsupervised clustering of the selected CL-based features by the survival Cox model. (A) is obtained on the main cohort (N=71) and (B) on the validation set (N=48). Colors correspond to the two groups obtained in a unsupervised manner.

### Visual explanation from the CL-based model

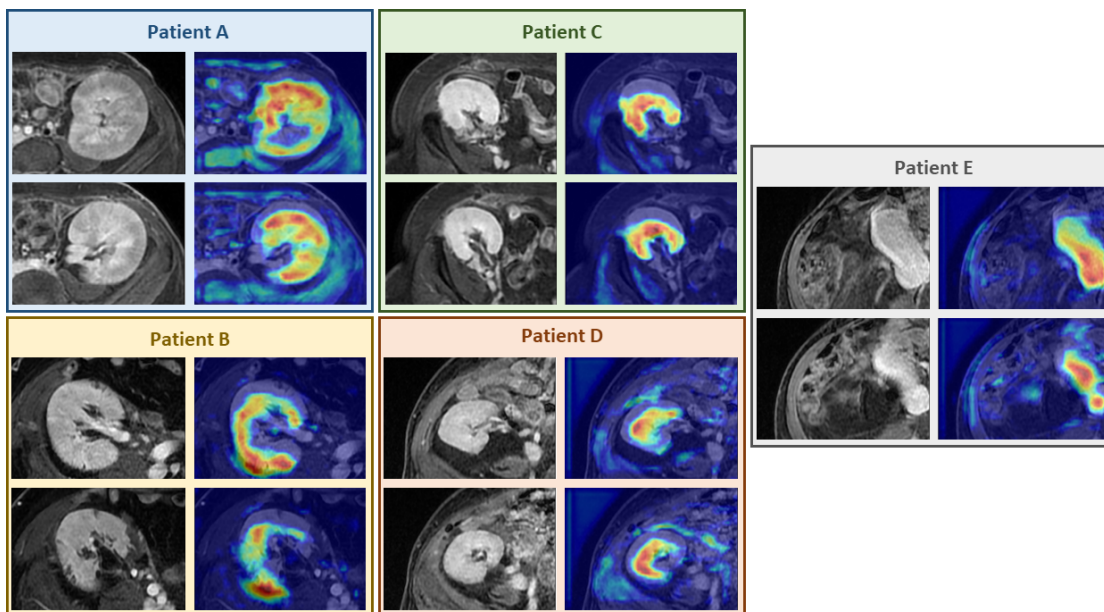


Figure 3.11: Heatmaps on the region of interest input to our CL-based model for different patients' examples. For each patient A, B, C, D, and E, we show two transversal slices of the MRI region of interest (left) and its superposition with our model activations map to obtain the CL-based model heatmap (right).



Explanability in AI approaches enabled us to visualize the activations of our CL-based model from the selected most predictive features through heatmaps on the corresponding MRI input (Figure 3.8). Figure 3.11 presents five examples of kidney recipient MRI exams, including 3 having transplantation failure (patients A, D, and E) and two not (B, C). These cases illustrate the ability of the automated imaging analysis to capture perfusion abnormalities and differentiate cortical (B), medullary phenomenon (C, D), or both (A). Patient A presented with severe transplant pyelonephritis and humoral rejection requiring intravenous immunoglobulin during one year post-transplant. Patient B had polar artery thrombosis with cortical infarct. Patient C underwent humoral rejection with a transplant biopsy revealing V3 vascular lesions without interstitial fibrosis and tubular atrophy (IF/TA) lesions. Patient D received a kidney from an ABO-incompatible living donor with early severe thrombotic microangiopathy. Lastly, patient E illustrates a limit of our model with an intense heatmap outside the transplant (saccular male aneurysm dilation).

## 3.7 Discussion & Conclusion

### 3.7.1 Novel methodologies for renal function forecasting

This chapter proposes a novel transformer-based architecture tailored to deal with missing data for the challenging task of serum creatinine prediction 2 years post-transplantation using imaging modalities. First, we show the significant use of contrastive learning schemes for this task. Our trained representations outperform common transfer learning and contrastive approaches. Then, a transformer encoder architecture enables to input the sequential features data per follow-up in order to forecast the renal transplant function, including a custom method to handle missing data. Our strategy performs better than other commonly used data imputation techniques.

The proposed approach attempts to demonstrate the relevance of using recent methodologies, such as contrastive learning schemes and transformer-based models, to learn meaningful representations of renal transplant data to forecast kidney function and better apprehend transplant complications, and ultimately transplantation failure. However several limitations must be addressed. First, although the longitudinal and dynamic dimensions of our MRI data enabled us to provide enough samples to successfully train such deep neural networks, our dataset contains a relatively small number of distinct patients. The separated test set remains relatively small and our results provide a proof of concept of the relevance of such an approach for renal transplantation applications. Second, the longitudinal information on the follow-up exams is used only in the second stage of the model. Several studies have investigated pretraining approaches of imaging data to incorporate longitudinal information [Couronné, 2021; Ren, 2022; Ouyang, 2022a] and assimilating this temporal information sooner in our model could be desirable. Finally, our proposed contrastive learning scheme could incorporate the information of only one clinicobiological variable at a time. Thus we obtained several models and evaluated their ability to provide meaningful representations for kidney function forecasting downstream tasks separately. Providing further investigations on which variable provides the most relevant features for the imaging data is essential, but a model that would be able to incorporate all those relevant

variables at the same time would be crucial.

Nevertheless, we designed an approach that successfully learn meaningful representations of renal transplants DCE MRI and a tailored downstream architecture to model longitudinal dependencies and handle missing data. Our promising results encourage the use of medical imaging over time to assist clinical practice for fast and robust monitoring of kidney transplants.

### 3.7.2 Deep learning-based imaging features into the clinic

After focusing on novel methodologies for renal function forecasting, we investigated a clinically relevant approach based on CL-based (Contrastive Learning-based) features to extract pertinent information from renal transplant DCE MRI examinations toward graft survival. Our model was trained using unsupervised contrastive learning to obtain meaningful renal transplant MRI data representations. We demonstrated that these learned CL-based features hold valuable information for 5-year graft survival analysis. We showed the potential prediction capacity of our CL-based features from medical imaging, obtaining a 72.7% C-index and their ability to stratify patients regarding this crucial event ( $p=0.0288$ ). Additionally, our model enabled us to retrieve explainability directly in the images, highlighting the promising clinical relevance of the approach and its potential to stratify patients based on graft survival.

Computer-aided diagnosis tools have been investigated to assess renal transplants' current status [Shehata, 2020]. At the same time, our study seeks to go one step further and suggests that early MRI examinations contain comprehensive information that can be used to predict overall graft survival. These results hold a promising direction for future research to predict complex events such as chronic dysfunction better. At the same time, recent DL approaches provide powerful tools for extracting this information from such complex and high-dimensional data, and such cutting-edge learning algorithms offer rising successful applications in Medical Imaging [Krishnan, 2022].

However, it is important to note that this study serves as a primary proof of concept, and several limitations must be addressed. Firstly, the validation was performed on MRIs from standard diagnostic procedures, which may introduce bias compared to systematic examinations in the main cohort. Although this inclusion potential bias, consistent results on regular exams demonstrate the robustness of the learned features. Furthermore, it highlights the value of healthy transplants and the diversity of pathological paths in the learning process. Secondly, to validate our approach's robustness and generalization capability, it is necessary to include larger patient cohorts and data from different sites and hospitals. This study reveals our most relevant findings from a longitudinal, multimodal retrospective cohort of medical imaging data from 108 patients who received a renal transplant. Nevertheless, systematic medical imaging examination data for multiple modalities (or multiple MRI sequences) and longitudinally are scarce.

After further demonstrating the predictive quality of DL-based features, the next crucial step is to determine how to incorporate this relevant information into clinical practice. The development of multimodal

models capable of handling multiple sources of information would benefit the potential results. Indeed, models that could integrate all the available information (clinical, biopsy, imaging) or combine existing models, such as the ones for nephropathological data [Yi, 2022; Yoo, 2023] or the iBOX [Loupy, 2019] model from clinical data, would highly benefit the clinic. Moreover, while we were able to provide clinically relevant interpretations of the activations heatmaps from our DL-based model, it is essential to address the issue of interpretability. Future clinical tools should not solely rely on results; better explainability is crucial. Collaborative efforts between medical doctors and engineers are necessary to shed light on the capability of such models, which can sometimes be perceived as black boxes. A recent study [Truchot, 2023] warns against the blind adoption of machine learning algorithms, highlighting lower predictive and calibration results than classical statistical approaches in modeling transplant failure using donor, recipient, and transplant-related parameters. Finally, medical imaging techniques to monitor the kidney are evolving [Alnazer, 2021], and other advancing modalities, such as Ultrasound Localization Microscopy [Bodard, 2023], could provide valuable data to apply DL methodologies to overall graft survival.

To conclude, we designed in this study a novel DL-based methodology to extract pertinent imaging features from early MRI examinations of renal transplants. Our promising results demonstrated the relevance of such approaches to predict graft survival in future research.





# Chapter 4

## Leveraging LLMs advances for multi-modal representations of renal grafts

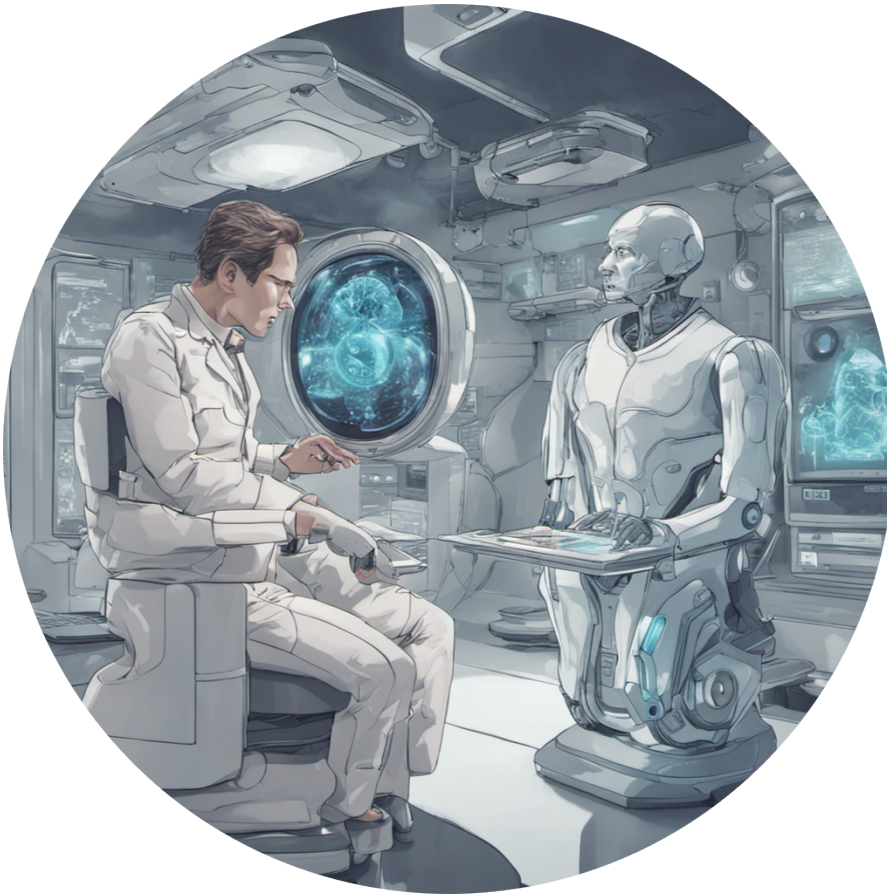
### Contents

---

4.1	Introduction . . . . .	89
4.2	Related work . . . . .	89
4.2.1	Learning representation of medical imaging . . . . .	89
4.2.2	Multimodal approaches . . . . .	90
4.2.3	Vision-language models . . . . .	90
4.3	Methodology . . . . .	91
4.3.1	Contrastive learning from joint text-image pairs . . . . .	92
4.3.2	Medical prompts from structural clinicobiological data . . . . .	92
4.3.3	Implementation details . . . . .	95
4.4	Data . . . . .	95
4.5	Experiments & Results . . . . .	96
4.5.1	Learned representations visualization . . . . .	96
4.5.2	Renal function forecasting . . . . .	96
4.5.3	Medical prompt generation . . . . .	99
4.5.4	Freezing the text encoder . . . . .	100
4.6	Discussion & Conclusion . . . . .	101

---





Prompt<sup>1</sup>: "A scene of a discussion between an Artificial Intelligence Large Language Model and a medical imaging system in a futuristic style."

## Summary

In this chapter, we introduce MEDIMP (**MED**ical **IM**ages with clinical **P**rompts), an approach to learn powerful manifolds of renal transplant DCE MRI data toward transplant function forecasting in the context of renal transplantation monitoring. This approach provides an elegant way to incorporate clinical or biological information into the learning process of feature extraction of medical imaging data. We designed clinicobiological prompts to leverage textual data and powerful NLP encoders using multimodal alignment of representations with contrastive learning. We demonstrate the relevance of the approach for renal transplant function forecasting. This work has been published in [Milecki, 2023a].

---

<sup>1</sup>Generated using [Stable Diffusion XL](#).





## 4.1 Introduction

End-stage renal disease is characterized by an irremediable reduction in kidney function, and renal replacement therapy is required to save the patient's life. Being more cost-effective than long-term dialysis and highly improving quality of life, renal transplantation emerges as the most effective solution [Suthanthiran, 1994]. However, a substantial risk of transplant chronic dysfunction persists and may lead to graft loss or patient death [Hariharan, 2021]. In clinical practice, the graft health status is primarily indicated by calculating the glomerular filtration rate (GFR) from the creatinine level resulting from blood tests. Medical imaging plays a substantial role in further examinations, and diverse imaging modalities have been investigated to monitor renal transplants [Sharfuddin, 2014].

In this work, we introduce MEDIMP (**MED**ical **IM**ages with clinical **P**rompts). This method learns relevant DCE MRI representations of renal transplants using contrastive learning from pairs of 3D images and clinicobiological prompts. The learned manifold enabled us to outperform state-of-the-art methods in the challenging task of kidney function prediction 2, 3, and 4 years post-transplantation from 4 DCE MRI follow-up exams. Our contributions are:

- (i) We propose a semi-automatic medical prompt generation from tabular data; to the best of our knowledge, this is the first work to propose such an approach for augmenting medical textual data;
- (ii) We extend existing approaches that combine text and imaging data by integrating 3D medical inputs and fine-tuning strategies; our approach allows using pretrained NLP models on a limited amount of textual data.

This chapter is split into five sections. First, we present related work on learning representations of medical imaging, multimodal approaches, and vision language modeling in Section 4.2. Then, in Section 4.3, we present our methodology based on two main components: contrastive learning from joint image-text pairs and medical prompts from structural clinicobiological data. We briefly describe the data used in this work in Section 4.4. In Section 4.5, we detail the set of experiments and achieve results to demonstrate the relevance of the proposed approach. Section 4.6 concludes and brings discussion on the limitations and perspectives of this work.

## 4.2 Related work

### 4.2.1 Learning representation of medical imaging

Learning powerful representations of medical imaging is of utmost importance, given the usual small size and limited annotations available. In such a setting, learning is performed in two stages. In the first stage, different self-supervised or weakly-supervised learning methods [Taleb, 2020; Krishnan, 2022] are used on the available imaging datasets, applying different types of learning, such as contrastive or adversarial learning [Sowrirajan, 2020; Azizi, 2021; Boyd, 2021]. Such representations are then frozen or fine-tuned for different downstream tasks, for which the amount of information is insufficient for fully supervised learning. Such pretrainings

could provide better representations and outperform ImageNet pretrained networks when applied to medical imaging. However, they may produce suboptimal representations for the downstream tasks that merely capture spurious correlations [Arjovsky, 2019]. More particularly, for renal transplantation, we proposed in section 3.3 weakly-supervised tasks from clinical information to learn rich representations of Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE MRI) using a single continuous attribute, confirming that combining imaging and clinical data leads to powerful biomarkers for prognosis.

## 4.2.2 Multimodal approaches

Proposing the best multimodal framework to integrate all the available sources of information successfully is of utmost importance in the medical field, where the amount of annotated data is very limited, with possibly also missing attributes. Going beyond simple concatenation of multimodal variables, Sahasrabudhe et al. [Sahasrabudhe, 2021] propose using a mixture-of-experts model to learn the contribution of each provided modality for the diagnosis of lymphocytosis. The method outperformed other single-modal and naive multimodal methods. Moreover, Chen et al. [Chen, 2021] proposed the use of a multimodal co-attention transformer for the combination of whole slide imaging and genomic information.

## 4.2.3 Vision-language models

Recent advances in Natural Language Processing (NLP) make textual data a potent candidate for designing weakly-supervised tasks to train computer vision models. Multiview contrastive learning [Bachman, 2019] has been investigated to take advantage of jointly training an image and text encoder [Zhang, 2020; Radford, 2021; Jia, 2021; Müller, 2021]. For natural images, Radford et al. [Radford, 2021] produced robust representations using 400 million (image, text) pairs, reporting competitive results on several downstream tasks on unseen datasets compared to fully supervised baselines. Zhang et al. [Zhang, 2020] used chest X-rays and pathology descriptions from radiology experts' diagnoses. Müller et al. [Müller, 2021] extended this joint image-text representation learning for localized tasks like semantic segmentation or object detection. All these studies consider 2D images and the medical ones used the MIMIC-CXR database, the largest dataset containing paired medical images and radiology reports. However, such data curation is arduous and highly time-consuming for medical experts. Moreover, such reports mainly contain information about the corresponding imaging exam and do not focus on other comorbidities.

Therefore, we propose to go one more step forward by generating representations using paired imaging and clinicobiological attributes in a relevant clinical setup with limited data. We explore recent NLP advances in Large Language Models (LLMs). In particular, ChatGPT [OpenAI, 2023], a 175B parameters model, offers a powerful tool to produce textual data. Specifically, textual data allow advantages as opposed to tabular data for medical applications:

1. Contextual information: Textual data contains rich contextual information, helping to understand the underlying patterns in the data better;

2. Better representation: text can provide a more expressive representation of the information contained in the data, leading to improved performances as our model better captures the complexities of the clinicobiological data;
3. Transferability and Interpretability: text is often more transferable across domains than tabular data. Moreover, text is more interpretable by humans, which is valuable for validating and understanding the decisions made by the proposed method.

## 4.3 Methodology

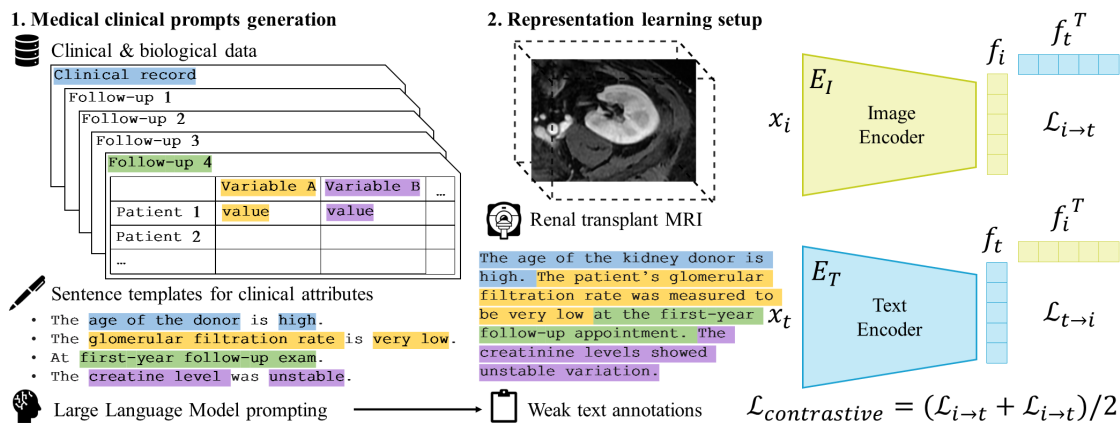


Figure 4.2: Overview of our method MEDIMP – Medical Images with clinical Prompts.

1. Medical prompts are generated from clinicobiological data using predefined templates of sentences, given as inputs to Large Language Models to produce augmented text data. 2. The medical prompts are used to learn multimodal representations of renal transplants DCE MRI using contrastive learning from image-text pairs.

Our multimodal representations are based on contrastive learning, coupling imaging, and text embeddings. Our text relies on attributes that are easily accessible, widely used in clinical practice, and supplementary to imaging data. Our goal is to use the learned manifold of renal transplant DCE MRI for the prognosis of the transplant or patient status after 2, 3, and 4 years post-transplantation. An overview of the method is presented in Figure 4.2.

### 4.3.1 Contrastive learning from joint text-image pairs

The first component of MEDIMP is pretraining an image encoder  $E_I$ , and a text encoder  $E_T$ , following a contrastive learning scheme using image-text pairs. Let us denote  $(x_i, x_t) \in \mathbb{R}^{B \times N_x \times N_y \times N_z} \times \mathbb{R}^{B \times T}$  a batch of  $B$  corresponding pairs of an 3D MRI volume  $x_{ib}$  and a tokenized text  $x_{tb}$  for  $b \in \llbracket 1, B \rrbracket$ . Both encoders transform  $x_i$  and  $x_t$ , in  $f_i = E_I(x_i)$  and  $f_t = E_T(x_t)$  respectively, to a batch of  $D$  dimensional embeddings. Both encoders are jointly trained to maximize the cosine similarity between the  $B$  pairs of image and text embeddings by optimizing the two following losses:

$$\mathcal{L}_{i \rightarrow t} = \sum_{b=1}^B -\log \frac{e^{\cos(f_{ib}, f_{tb})/\tau}}{\sum_{k=1}^B e^{\cos(f_{ib}, f_{tk})/\tau}}, \quad (4.1)$$

where  $\cos(\cdot, \cdot)$  is the cosine similarity function and  $\tau \in \mathbb{R}^+$  a learned temperature parameter. Such loss was first proposed as the InfoNCE loss [Oord, 2018] to maximize a lower bound on mutual information and is widely used in recent uni-modal contrastive learning frameworks [Chen, 2020a].  $\mathcal{L}_{i \rightarrow t}$  enforces the image embeddings to align to the text embeddings and is, therefore, asymmetric. Similarly, we define  $\mathcal{L}_{t \rightarrow i}$ :

$$\mathcal{L}_{t \rightarrow i} = \sum_{b=1}^B -\log \frac{e^{\cos(f_{tb}, f_{ib})/\tau}}{\sum_{k=1}^B e^{\cos(f_{tb}, f_{ik})/\tau}}. \quad (4.2)$$

The total loss is obtained by averaging Eq.4.1 and Eq.4.2, denoted as  $\mathcal{L}_{\text{contrastive}}$ :

$$\mathcal{L}_{\text{contrastive}} = \frac{\mathcal{L}_{i \rightarrow t} + \mathcal{L}_{t \rightarrow i}}{2} \quad (4.3)$$

$\mathcal{L}_{\text{contrastive}}$  learns a multimodal feature space by jointly optimizing  $E_I$  and  $E_T$  to maximize the cosine similarity of the embeddings  $f_i$  and  $f_t$  between the  $B$  true pairs per batch and minimizing the cosine similarity of the  $B^2 - B$  false pairs.

### 4.3.2 Medical prompts from structural clinicobiological data

To exploit image-text pairing with contrastive learning, as well as the expression and encoding capability of recent NLP model advances, such as LLMs [Brown, 2020; OpenAI, 2023; Radford, 2021]. We propose a framework to generate textual data from structural clinicobiological data that describe variables used in clinical practice and linked to graft survival. The process is displayed on the left side of Figure 4.2. First, medical experts guided us to set thresholds to categorize continuous variables into text labels such as “low”, “high”, “stable”, and “unstable” and to produce one *template sentence* per variable of interest, e.g., “the GFR of the patient is very low at the first-year follow-up exam”. However, templates offer only one way of expressing the information of the variables. Indeed, the richness of language vocabulary can provide a variety of descriptions for the same information, such as “During the first-year follow-up visit, the transplant patient’s GFR is found to be very low.”, or “The transplant patient’s GFR is assessed as very low at the date follow-up examination.”, thus generating descriptive text to train the proposed contrastive scheme. This richness was leveraged by recent advances in LLMs at training; hence, they offer robust NLP tools. Specifically, we use the dialogue

LLM ChatGPT [OpenAI, 2023] to produce  $N = 10$  textual data augmentations for each template sentence. All generated prompts are reported. The main conversation with the dialogue LLM ChatGPT [OpenAI, 2023] to generate our medical prompts was the following:

- Input 1: "I want to find synonyms or text augmentations of full sentences in the context renal transplantation. Are you familiar with these concepts?"

Output 1: "Yes, I am familiar with the concept of finding synonyms or text augmentations of full sentences, also known as text augmentation or data augmentation. This is a technique used to artificially increase the size and diversity of a dataset by applying various natural language processing (NLP) techniques such as synonym replacement, random insertion, deletion, and substitution of words. In the context of renal transplantation, it could be used to expand the number of examples of medical reports or other texts related to renal transplantation by generating similar but slightly different versions of the original text."

- Input 2: "The full sentence template is "The age of the donor is {age}. The glomerular filtration rate (GFR) of the patient is {gfr} at {date} follow-up exam. And the creatinine levels variation were {adj}" where {age}, {gfr}, {date}, {adj} can be changed by me but not by you, do you understand?"

Output 2: "Yes, I understand. You have provided a sentence template with placeholders for age, glomerular filtration rate (GFR), date, and an adjective that can be changed by you. And you are looking for text augmentation to generate similar but slightly different versions of the original text with the help of natural language processing(NLP) techniques such as synonym replacement, random insertion, deletion, and substitution of words."

- Input 3: "Provide full sentence synonyms of the sentences I provided, you are allowed to give more context information and not especially respect the exact number of sentences."

Output 3: "Here are some examples of synonyms or text augmentations of the sentence template you provided:

1. "The age of the kidney donor is {age}. The patient's glomerular filtration rate was measured to be {gfr} at the {date} follow-up appointment. The creatinine levels showed {adj} variation."
2. "The individual who provided the kidney for transplant was {age} years old. The recipient's GFR was determined to be {gfr} during the {date} check-up. Creatinine levels fluctuated {adj}."
3. "At the time of transplant, the donor was {age} years of age. The patient's GFR was evaluated as {gfr} at the {date} follow-up examination. Creatinine levels demonstrated {adj} changes."
4. "The age of the individual who donated the kidney was {age}. The GFR of the transplant recipient was {gfr} at the {date} examination. There were {adj} variations in creatinine levels."
5. "The kidney donor was {age} years old. The GFR of the transplant patient was measured as {gfr} at the {date} follow-up visit. The creatinine levels showed {adj} fluctuations."
6. "The donor's age was {age} at the time of transplant. The patient's glomerular filtration rate was found to be {gfr} during the {date} appointment. Creatinine levels were {adj}."

Please note that the sentences are not exact synonyms, but variations that convey similar information, and that it's possible to generate a lot more variations by changing the words, phrases or structures of the sentences."

- Input 4: "Give ten more synonym examples of the provided sentence and you can generate more variations as you proposed."

Output 4: "Sure, here are ten more examples of text augmentations of the sentence template you provided:

1. "The kidney donor was {age} years of age. The patient's GFR was found to be {gfr} at the {date} check-up. Creatinine levels displayed {adj} fluctuations."
2. "At the time of transplant, the age of the kidney donor was {age}. The patient's glomerular filtration rate was {gfr} at the {date} follow-up. Creatinine levels varied {adj}."
3. "The individual who provided the kidney for transplant was {age} years old. The GFR of the recipient was determined to be {gfr} during the {date} examination. There were {adj} changes in creatinine levels."
4. "The age of the kidney donor was {age}. The GFR of the transplant recipient was {gfr} at the {date} follow-up visit. Creatinine levels fluctuated {adj}."
5. "The individual who donated the kidney was {age} years of age. The patient's glomerular filtration rate was measured as {gfr} at the {date} check-up. The creatinine levels demonstrated {adj} variations."
6. "At the time of transplant, the donor's age was {age}. The GFR of the patient was evaluated as {gfr} during the {date} follow-up. The creatinine levels showed {adj} fluctuations."
7. "The kidney donor was {age} years old at the time of transplant. The patient's glomerular filtration rate was {gfr} at the {date} examination. Creatinine levels varied {adj}."
8. "The individual who provided the kidney for transplant was {age} years of age. The recipient's GFR was determined to be {gfr} at the {date} follow-up appointment. The creatinine levels displayed {adj} changes."
9. "At the time of transplant, the age of the kidney donor was {age}. The GFR of the transplant patient was {gfr} during the {date} check-up. Creatinine levels fluctuated {adj}."
10. "The individual who donated the kidney was {age} years old. The glomerular filtration rate of the transplant recipient was {gfr} at the {date} examination. The creatinine levels showed {adj} variation."

Please note that the above sentences are not exact synonyms, but variations that convey similar information, and that there are many more variations that can be generated by changing the words, phrases or structures of the sentences."

### 4.3.3 Implementation details

The image encoder followed a 3D ResNet50 architecture initialized with CLIP [Radford, 2021] weights, a model pretrained on 400 million (image, text) pairs collected from the internet. We extended the attention-based pooling layer of CLIP to 3D and duplicated the weights to 3D in depth to match our data dimension. For the text encoder, we used the BERT [Devlin, 2018] architecture initialized with the Bio+Clinical BERT [Alsentzer, 2019] model pretrained on the MIMIC clinical notes [Johnson, 2016]. The first 11 layers of Bio+Clinical BERT were frozen, fine-tuning the last layer of the transformer with our contrastive task. Section 4.5.4 summarises the ablation of fine-tuning more layers for our task. The temperature parameter  $\tau$  was initialized to 0.07, incorporated into the model as a learnable parameter, and clipped to prevent scaling the logits by more than 100, following the recommendations of CLIP training. In our experiments, we use the Adam [Kingma, 2014b] optimizer with decoupled weight decay regularization [Loshchilov, 2017] of 0.02 with a starting learning rate of  $5e^{-5}$  following a cosine schedule and preceded by a linear warm-up of 40 epochs. The batch size was set to 88 and the model trained for 200 epochs with mixed-precision on 4 NVIDIA Tesla V100 GPU using Pytorch [Paszke, 2019].

For the image data, we used data augmentation with the sequential application with each a 0.5 probability of horizontal flipping, random affine transformation, random Gaussian blur ( $\sigma \in [0, 0.5]$ ), random Gaussian noise ( $\sigma \in [0, 0.05]$ ), random contrast perturbation ( $\log \gamma \in [-0.3, 0.3]$ ) using TorchIO python library [Pérez-García, 2021].

## 4.4 Data

Our study was approved by the Institutional Review Board, which waived the need for patients' consent. The data cohort corresponds to study reference ID-RCB: 2012-A01070-43 and ClinicalTrials.gov identifier: NCT02201537. The data used in this study are anonymized. Our imaging data are based on DCE MRI series collected from 105 subjects (split as 72/5 training/validation, and 28 test). Each subject underwent up to 4 follow-up exams, taking place approximately 15 days, 30 days, 3 months, and 1 year post-transplantation.

The DCE MRI volumes sized  $512 \times 512 \times [64 - 88]$  voxels included spacing ranging in  $[0.78 - 0.94] \times [0.78 - 0.94] \times [1.9 - 2.5]$  mm. All volumes were cropped around the transplant using the automatic and unsupervised method for selecting the region of interest and reducing dimensionality presented in section 2.2. Intensity normalization was executed to each volume independently by applying standard normalization, clipping values to  $[-5, 5]$ , and rescaling linearly to  $[0, 1]$ .

To provide the clinicobiological data used to generate text annotations and the endpoints, all 77 patients in the train set were regularly subjected to blood tests before the transplantation and several years after to measure the serum creatinine (Creat) level in  $\mu\text{mol.L}^{-1}$ . The donor's age variable and the GFR value at each follow-up exam were also collected. For the 28 test subjects, these clinicobiological attributes were not accessible during this study. As a result of blood tests, Creat is a primary indicator of kidney function used in



clinical practice. The binary classification downstream task is obtained when binarizing the Creat value using a threshold of  $110 \mu\text{mol.L}^{-1}$  at different prediction dates. The Creat target prediction value is calculated as the mean over three months before and after the prediction dates.

## 4.5 Experiments & Results

### 4.5.1 Learned representations visualization

First, we visualize the representations of DCE MRI from the trained image encoder using t-SNE decomposition [Maaten, 2008]. Figure 4.3 shows the different reduced feature spaces obtained by the model trained on all available clinicobiological variables ( $n_{cl} = 4$ ), projected specifically and adding colormaps for each attribute. We evaluate the clustering of the imaging features toward the clinicobiological information covered by our weak text annotations, adding augmented images to check the tendency better. While the continuous variables were categorized and transformed to medical prompts, MEDIMP image encoder demonstrates relevant representations towards (B) the GFR and (C) the Creat. The t-SNE decomposition does not reveal favorable representations regarding (D) Donor's Age. On the contrary, the obtained feature space serves well the (A) Exam date information, where we retrieve better clustered very early exams (D15, red) and late exams (M12, blue) due to their respective distance to the transplantation surgery.

### 4.5.2 Renal function forecasting

#### Downstream task and metrics

We evaluate MEDIMP on the downstream task of serum creatinine (Creat) prediction from the imaging features of 4 follow-up exams using a light transformer architecture tailored for missing follow-up exams, presented in section 3.5. Following the authors' evaluation, 10-fold cross-validation was performed on the training set, and results for the main models are summarised in Section 4.5.2. To make the task more challenging, we evaluate the performance of the representations at 2, 3, and 4 years post-transplantation and also report the mean over the three predictions for the 28 test subjects. The two evaluation metrics used were the F1 score and the area under the receiver operating characteristic curve (AUC).

#### Ablation

First, we ablate all information used in our method MEDIMP by adding different combinations of the clinicobiological measurements in the medical prompts, i.e., the glomerular filtration rate (GFR) at the exam date, the timestamp of the patient's exam (Exam), the creatinine levels variation from the previous exam (Creat), and the age of the donor (D.A.). We report the results in the bottom part of Table 4.1 (MEDIMP). We observe that the best mean scores over the three predictions are obtained using all the available medical prompts (last row). Moreover, the AUC decreases over the prediction date, showing the increasing prognosis difficulty with time. Using only the GFR prompts (row 8), AUC scores are just above random, indicating the need for more descriptive text.

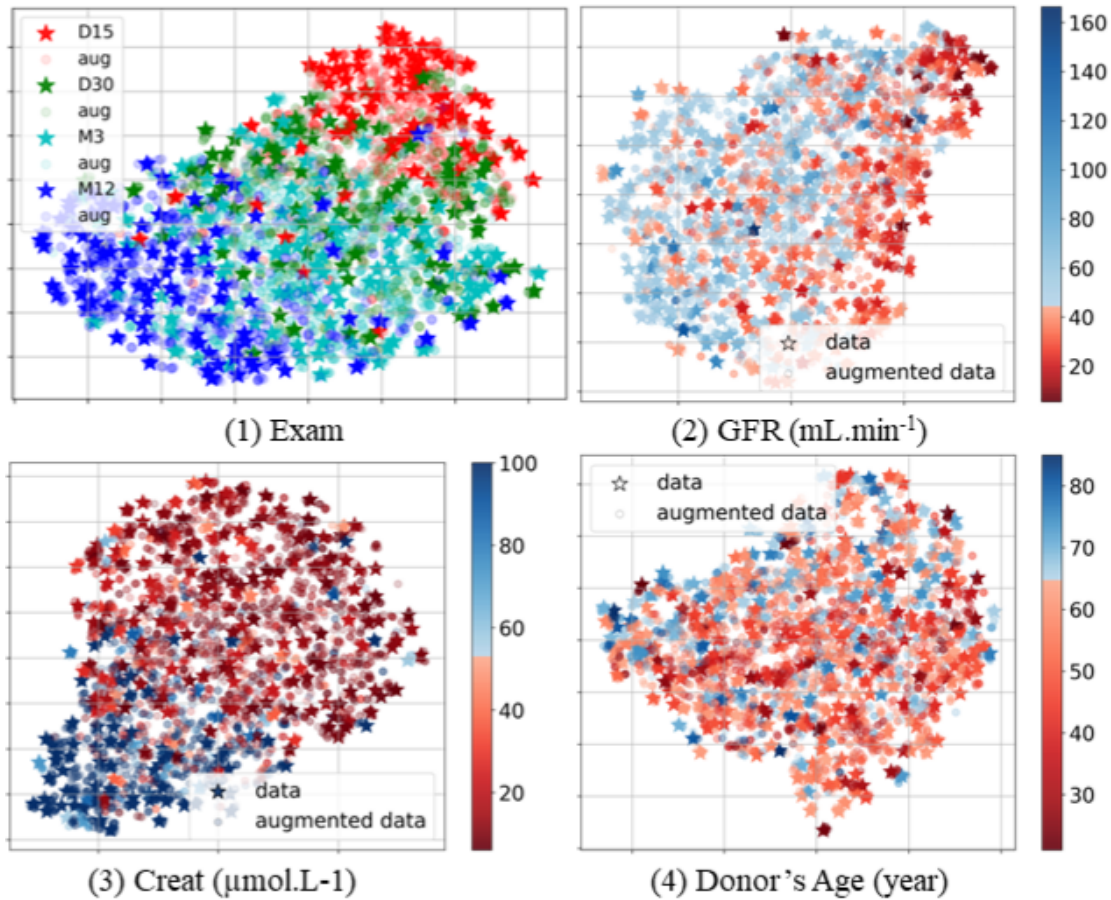


Figure 4.3: t-SNE visualizations of the features of the last layer of MEDIMP image encoder using the DCE MRI exams. Colormaps are set by the 4 variables of interest value used for the medical prompts: (1) Exam (exam date), (2) GFR ( $\text{mL}\cdot\text{min}^{-1}$ ), (3) Creat ( $\mu\text{mol}\cdot\text{L}^{-1}$ ), and (4) donor's age (year). Stars symbol display the real data while circles the augmented (aug) data. D15, D30, M3, and M12 are the four exam timestamps.

### Comparison to the state of the art

Table 4.1 reports the results when comparing the proposed MEDIMP with our previous work for this task presented in section 3.5, denoted as CosEmbLoss (row 2 & 3). Note that the main model of CosEmbLoss uses only GFR information. Hence, it is directly comparable to MEDIMP when using only GFR (8th row). These experiments reveal that smaller and more compact models, such as the CosEmbLoss perform better than big models when the text information is not very rich. However, when more variables are integrated, our proposed methods outperform the CosEmbLoss. For a fair comparison, we also compare MEDIMP to several baselines with the same level of information. In particular, we evaluate against four baselines, denoted as the CosEmbLoss++, where we gradually add the same level of information as in MEDIMP. In practice, we optimize the same two-stream approach by averaging several cosine embedding losses based on the number

Table 4.1: **Comparison of MEDIMP vs SOTA.** We evaluate the performance at 2,3,4 years post-transplantation and report the mean. Ablations in weak annotations from either the comparison CosEmbLoss pretraining or our proposed generated textual data are denoted as GFR (GFR at exam), Exam (which follow-up), Creat (creatinine variations from the previous exam), and D.A. (the donor’s age). We report F1 score (F1), and ROC AUC (AUC). **Bold**, Underlined indicates the top 1, 2 performing combinations, respectively.

Method	Weak annotations				2 years		3 years		4 years		Mean	
	GFR	Exam	Creat	D.A.	AUC	F1	AUC	F1	AUC	F1	AUC	F1
CLIP weights					62.6	73.7	52.5	78.1	51.3	54.6	55.5	68.8
CosEmbLoss	✓				76.2	86.4	77.8	70.6	67.0	77.3	73.6	78.1
CosEmbLoss				✓	75.5	81.1	75.6	68.8	66.1	78.1	72.4	76.0
CosEmbLoss++	✓			✓	<u>84.4</u>	<u>88.9</u>	82.5	<b>86.4</b>	73.9	<u>85.7</u>	<u>80.3</u>	87.0
CosEmbLoss++	✓		✓		81.6	87.8	71.3	85.1	71.3	<u>90.2</u>	74.7	<u>87.7</u>
CosEmbLoss++	✓		✓	✓	78.2	87.0	75.0	83.3	<u>74.8</u>	<u>87.0</u>	76.0	85.8
CosEmbLoss++	✓	✓	✓	✓	75.5	85.7	62.0	69.8	<u>63.5</u>	80.9	67.0	78.8
MEDIMP	✓				56.5	83.3	51.9	79.1	49.6	<b>90.2</b>	52.6	84.2
MEDIMP	✓	✓			81.0	<b>89.4</b>	81.9	80.0	<u>74.8</u>	84.4	79.2	84.6
MEDIMP	✓	✓		✓	76.9	73.2	<b>86.3</b>	<u>85.7</u>	<u>74.8</u>	<b>90.2</b>	79.3	83.0
MEDIMP	✓	✓	✓		72.8	86.4	71.9	81.0	71.3	71.8	72.0	79.7
MEDIMP	✓	✓	✓	✓	<b>85.0</b>	<b>89.4</b>	<u>84.4</u>	83.7	<b>75.7</b>	<b>90.2</b>	<b>81.7</b>	<b>87.8</b>

of variables of interest incorporated. We report these results in rows 4-7 of Table 4.1. We observe that CosEmbLoss++ achieves its best performance when using a combination of 2 variables. The best mean AUC is 80.3% with GFR and D.A., and the best F1 is 87.7% with GFR and Creat, which are lower than the best MEDIMP performances. Note, the Exam information was only added to one CosEmbLoss++ combination (7th row) as this variable is less adapted to CosEmbLoss approach. Nevertheless, combining the 3 variables of interest with CosEmbLoss++ lowers the performance to 76.0% AUC and 85.8% F1 in Mean. Overall, MEDIMP with all medical prompts results in the best predictions at 2 and 4 years post-transplantation.

### Cross-validation results

We evaluated our representations on the downstream task of kidney function prediction 2 years post-transplantation presented in section 3.5. We performed the task on two more prediction dates, namely 3 and 4 years post-transplantation, to better highlight the significance of our proposed approach. Nevertheless, following the evaluation of section 3.5.3, we also performed 10-fold cross-validation on the training set. We report below in Table 4.2 those cross-validation results (ROC AUC, F1 as mean  $\pm$  standard deviation) for our validation set, for the best combination (denoted with \*) of CosEmbLoss presented in section 3.3.1, CosEmbLoss++, and our proposed MEDIMP, for the three different tasks. While demonstrating similar cross-validation results over the mean of the three prediction tasks, MEDIMP enables lower variation in the validation sets.

Table 4.2: **Cross-validation results.** We report F1 score (F1), and ROC AUC (AUC) as mean  $\pm$  standard deviation for the best combinations of CosEmbLoss, CosEmbLoss++ and our proposed MEDIMP, denoted with \*.

Method Validation set	2 years		3 years	
	AUC	F1	AUC	F1
CosEmbLoss*	93.3 $\pm$ 12.0	86.4 $\pm$ 12.2	81.7 $\pm$ 15.9	74.4 $\pm$ 23.3
CosEmbLoss++*	91.7 $\pm$ 14.6	88.8 $\pm$ 10.4	84.1 $\pm$ 14.3	71.4 $\pm$ 20.5
MEDIMP*	89.3 $\pm$ 11.4	80.1 $\pm$ 11.8	87.5 $\pm$ 4.1	81.7 $\pm$ 6.5

Method Validation set	4 years		Mean	
	AUC	F1	AUC	F1
CosEmbLoss*	84.5 $\pm$ 16.5	64.6 $\pm$ 26.9	86.5 $\pm$ 14.8	75.1 $\pm$ 20.8
CosEmbLoss++*	83.3 $\pm$ 13.9	69.1 $\pm$ 27.3	86.4 $\pm$ 14.3	76.4 $\pm$ 19.4
MEDIMP*	81.4 $\pm$ 13.9	71.9 $\pm$ 29.6	86.1 $\pm$ 9.1	77.9 $\pm$ 16.0

### 4.5.3 Medical prompt generation

To demonstrate the relevance of the proposed approach for medical prompt generation, we compare our main model with two other approaches that produce text information. The first one is noted as “Manual” and comprises all the templates indicated by the medical experts, corresponding to only one sentence per variable of interest. Note that this is the base of our proposed medical prompting without using the prompt expansion method described in Section 4.3.2. The second one uses an existing NLP model, T5 [Raffel, 2020], to produce sentences from structural data. For a fair comparison, we train the T5 model on the WebNLG 2020 data [Gardent, 2017] and infer it on our data to generate text, denoted as “T5 WebNLG”. The results are summarized in Table 4.3, highlighting the superiority of our method. The “T5 WebNLG” approach offers a competitive F1 for all the predictions, although the AUC is the lowest except for the 2 years prediction. We show in Section 4.5.3 examples of generated texts from these three approaches. “Manual” approach lacks diversity in the text data, and therefore the training process of our proposed approach without text augmentations is more challenging.

Table 4.3: **Quantitative evaluation of the proposed method against other text generation methods.** All medical prompts were used. We report F1 score (F1), and ROC AUC (AUC). **Bold**, Underlined indicates the top 1, 2 performing combinations, respectively.

Method	2 years		3 years		4 years		Mean	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
MEDIMP	<b>85.0</b>	<b>89.4</b>	<b>84.4</b>	<b>83.7</b>	<u>75.7</u>	<b>90.2</b>	<b>81.7</b>	<b>87.8</b>
Manual	74.2	76.2	<u>80.6</u>	62.1	<b>80.0</b>	76.9	<u>78.3</u>	71.7
T5 WebNLG	<u>74.8</u>	<u>85.7</u>	78.8	<u>83.3</u>	74.8	<u>85.7</u>	76.1	<u>84.9</u>

### Textual data generation

We compare the proposed approach for medical prompt generation with two other approaches that produce text annotations. The first one is noted as “Manual” and comprises all the templates indicated by the medical experts, corresponding to only one sentence per variable of interest. Note that this is the base of our proposed medical prompting without using the LLM augmentation method. The second one uses an existing NLP model, T5 [Raffel, 2020], to produce sentences from structural data. We train the T5 model on the WebNLG 2020 data [Gardent, 2017] and infer it on our data to generate text, denoted as “T5 WebNLG”. We observe that the “Manual” approach lacks diversity in the textual data, as no text augmentations are performed for this straightforward process. “T5 WebNLG” offers more variability in words used, but the structure of the sentences remains similar and straightforward. Moreover, some incorrect generations occur, e.g., “The age of the donor” is replaced by “The age of the patient”. Such errors introduce anomalies in the data, a highly sensitive issue in such a medical context.

- “Manual” – from one sentence template:  
“The age of the donor is low. The glomerular filtration rate (GFR) of the patient is high at one month follow-up exam. And the creatinine levels variation were stable.”;
- “T5 WebNLG” – pretraining a model to generate textual data from structural data:
  - correct generation example: “The age of the donor is high. The glomerular filtration rate is medium. The creatinine levels of a patient are unstable.”;
  - incorrect generation example: “The age of the patient was low. The glomerular filtration rate of GFR is an extrem low rate. The creatinine levels of a patient are unstable.”
- MEDIMP: see Section 4.3.2.

#### 4.5.4 Freezing the text encoder

For the text encoder, we used the BERT [Devlin, 2018] architecture initialized with the Bio+Clinical BERT [Alsentzer, 2019] model pretrained on the MIMIC clinical notes. BERT is based on the transformer [Vaswani, 2017] architecture and comprises 12 transformer blocks. Our main models were obtained by freezing the first 11 layers of the Bio+Clinical BERT model, fine-tuning only the last layer of the transformer with our contrastive task.

Benefiting from a dataset of 400 million (image, text) pairs collected from the internet, Radford et al. [Radford, 2021] trained both their image and text encoder from scratch. While we used the same initialization as Zhang et al. [Zhang, 2020], they froze their text encoder’s first half (6 layers). In recent NLP work, Lu et al. [Lu, 2021] suggested only finetuning normalization layers (LN) in the transformer blocks, without finetuning the self-attention and feedforward layers of the residual blocks. Table 4.4 reports the ablation results using this latest strategy, denoted as not LN, and gradually freezing the 6, 9, and 11 first layers of the text encoder. The ablation was evaluated using two sets of weak annotations from our proposed method, first, the GFR and date, denoted as MEDIMP A, and second, the GFR, Exam, and Donor’s Age, denoted as MEDIMP B.

We observe that freezing the first 11 layers gives us the best performances, which is the strategy we used for our approach.

Table 4.4: **Ablation results on the way of freezing the text encoder.** We report F1 score (F1), and ROC AUC (AUC). **Bold**, Underlined indicates the top 1, 2 performing combinations, respectively.

Method	Freezing $E_T$	2 years		3 years		4 years		Mean	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1
MEDIMP A	First 11	<u>81.1</u>	<b>89.4</b>	81.9	80.0	74.8	84.4	79.2	<b>84.6</b>
MEDIMP A	First 9	74.8	75.7	81.9	81.0	76.5	68.6	77.7	75.1
MEDIMP A	First 6	74.2	74.4	70.0	82.1	<u>83.5</u>	68.6	75.9	75.0
MEDIMP A	not LN	73.5	70.6	77.5	80.0	73.0	71.8	74.7	74.1
MEDIMP B	First 11	76.9	73.2	<b>86.3</b>	<b>85.7</b>	74.8	<b>90.2</b>	<u>79.3</u>	<u>83.0</u>
MEDIMP B	First 9	<b>83.7</b>	64.5	78.1	<u>82.9</u>	75.7	64.7	79.2	70.7
MEDIMP B	First 6	75.5	70.6	<u>84.4</u>	80.0	<b>84.4</b>	82.9	<b>81.4</b>	77.8
MEDIMP B	not LN	66.7	<u>81.0</u>	79.4	80.9	60.9	<u>85.7</u>	69.0	82.5

## 4.6 Discussion & Conclusion

Our experiments have shown improvements in the representation learning capabilities of deep image encoders for renal transplantation MRI compared to the previous state-of-the-art approach for the specific application of renal transplant function forecasting. MEDIMP aimed at enhancing representation learning approaches using external data, leveraging the power of deep NLP models, such as LLMs, and introducing a novel process to incorporate relevant clinicobiological medical information. We deem that such an approach crossing modalities in medical research would highly improve the capacity to understand complex biological and medical phenomena.

However, some limitations of our framework remain to be analyzed. (1) First, although limited data is part of the challenges of this study, supplementary test data would indubitably support validating our method. To our knowledge, no public medical imaging dataset offers simultaneously longitudinal imaging, biological, as well as clinical data for each patient for prognosis at different times. Nevertheless, our proposed framework could be easily translated to similar datasets having imaging data and any type of tabular data. (2) Second, this work constitutes a first attempt to generate medical prompts as text information from a few clinicobiological variables of interest, which are crucial to apprehend complex medical concepts. Thus, we would seek to extend this work to exploit more variables to guide the training of the image encoders. Using our proposed framework this can be easily implemented since only a few templates will need to be defined to automatically generate text augmentations. (3) Finally, the study of the development and comprehension of LLMs is recent. We expect such NLP models to increase the emphasis on using textual data. However, the main drawback of these models is their reliability, in the sense that inaccurate generation might be challenging to detect. In this work, we use ChatGPT to enable relevant and robust text augmentations. No information from the sensitive medical data was leaked into the prompts, which is an essential process in the medical domain. One could

think of several ways to further use such models as powerful tools, for instance, through prompt engineering tuned for medical cases.

We have introduced MEDIMP, an approach to learn powerful manifolds of renal transplant DCE MRI data toward transplant function forecasting in the context of renal transplantation monitoring. This approach provides an elegant way to incorporate clinical or biological information into the learning process of feature extraction of medical imaging data. The presented representation learning strategy enabled us to outperform the state of the art in the challenging task of creatinine prediction. These promising results advocate using textual data from emerging LLMs to assist in training robust medical imaging models.







# Chapter 5

## Discussion

### Contents

---

5.1	Main contributions . . . . .	109
5.2	Perspectives & Future directions . . . . .	110
5.2.1	Representation learning for medical image analysis . . . . .	110
5.2.2	Developing foundation models for medical imaging . . . . .	111
5.2.3	Next AI tools for renal transplant monitoring . . . . .	112
5.2.4	Going beyond this thesis . . . . .	113
5.3	Conclusion . . . . .	114

---





Prompt<sup>1</sup>: "A wild discussion between scientists (including women) at an international conference with coffee mugs in cartoon style."

## Summary

In this chapter, we start by outlining the main contributions of this thesis. We then delve into a discussion of compelling perspectives that have emerged from our work and exciting future directions in forthcoming research projects, with a focus on representation learning for medical image analysis, the potential of foundation models in medical imaging, AI frameworks for renal transplantation monitoring, and two relevant AI topics not covered in our work. Lastly, we provide a concluding perspective on the entire thesis.

---

<sup>1</sup>Generated using [Stable Diffusion XL](#).



## 5.1 Main contributions

In this thesis, we investigated different methodologies to advocate the use of AI algorithms in the monitoring of renal transplantation, to better detect or predict transplant complications, risk of chronic dysfunction, and ultimately graft failure. We built our research on the recent development of deep learning-based representation learning and un-/weakly-/self-supervised frameworks. We worked with multiple medical data sources, focusing on the medical imaging modalities.

Chapter 2 introduced the baseline methodologies to develop machine learning models for renal transplantation-related tasks from clinical, biological, and imaging data. We investigated these classical approaches and observed their limitations with limited data. To compute predefined radiomics from medical imaging data, we designed an unsupervised and deep-learning based algorithm to automatically detect and segment renal transplants from MRI data. We investigated the use of multi-sequence MRI, notably with the DCE and T2 sequences, to alleviate the difference in enhancement of the organ in the scans in order to perform the task. Our quantitative results indicate that our tool could help clinical practice to provide fast and robust monitoring of kidney grafts.

In Chapter 3, we provided medical-specific pretrainings to learn meaningful representations of our renal transplant DCE MRI data. These pretrainings leverage contrastive learning objectives based on clinical variables linked to graft survival to guide the training of our deep CNNs encoders. We demonstrated the relevance of our generated features by designing a renal function forecasting downstream task from the imaging follow-up examinations. Our approach introduced the use of transformer architecture to model time dependencies and allowed us to handle missing examinations to perform the prediction. In a second phase, we studied the relevance of our imaging encoders, providing AI-based radiomics, for different MRI sequences and different timestamp follow-ups, and their capability to predict overall graft survival at five years, notably from early DCE MRI examinations. Our promising results encourage the use of medical imaging over time to assist clinical practice for fast and robust monitoring of kidney transplants.

In Chapter 4, we investigated multimodal contrastive learning to incorporate clinical and biological structural data into the learning of renal transplant MRI representations by leveraging the expressiveness of textual data, notably with the recent success of Natural Language Processing innovations. We designed medical clinical prompts from the emerging LLMs to constitute image-text pairs. Moreover, we provided tailored multimodal methodologies to handle limited tabular data and 3D medical imaging data, notably with transfer learning and fine-tuning techniques. Our encouraging findings advocate using textual data from emerging LLMs to assist in training robust medical imaging models.

## 5.2 Perspectives & Future directions

This thesis presented distinct and original contributions to the field of deep learning and representation learning with un-/weakly-/self-supervised techniques applied to renal transplantation monitoring. Each chapter incorporated a discussion including limitations and perspectives for each of our distinct studies. Nevertheless, beyond the respective limitations of the proposed methodologies, we would like to provide broader perspectives and appealing future directions in the domains of deep learning for medical image analysis that we explored (more or less) during the thesis.

### 5.2.1 Representation learning for medical image analysis

First, we developed innovative deep learning methodologies that were experimented only on our specific renal transplantation application and dataset. Although some aspects of the proposed approaches were tailored for the renal transplantation context, such as using specific clinical and biological variables to incorporate clinical knowledge guidance to the learning of our imaging representations, we deem that our proposed approaches can easily be adapted to other applications, using other imaging modalities and other clinical variables of interest specific to the task. In forthcoming research endeavors, we plan to expand our multimodal methodology to multimodal imaging by incorporating both MRI and US modalities. Our exploration aims to discern the advantages derived from both modalities through multiview contrastive learning, evaluating their respective contributions to the task of forecasting renal transplant function. Additionally, we are implementing late fusion modules based on self-attention in the scenario of missing examinations. These modules facilitate the combination of features from distinct modalities, yielding unified embeddings agnostic to modality variations. Analogous to the architecture introduced in Chapter 3 for handling longitudinal dependencies, our fusion module captures cross-modal dependencies while accommodating the absence of certain modalities during both training and testing phases. In this context, we are experimenting with knowledge distillation techniques to establish a teacher-student framework. The teacher model is exclusively exposed to input pairs without missing modalities, and it imparts knowledge to the student model, which is trained utilizing inputs featuring missing modalities. This methodology holds promise for diverse applications and various modalities, such as detecting alterations of consciousness in fighter jet pilots based on video and biological signals data.

Moreover, recent studies have highlighted a limitation in the classical supervised learning setup, where models are directly optimized from input-label pairs. This limitation affects the expressiveness of their internal representations. The emergence of self-supervised approaches, which design proxy tasks based on prior data knowledge or the desired task, has shown significant promise and success in mitigating this issue. A notable example of this shift is the AlphaFold project [Jumper, 2021], a collaborative effort involving experts from various disciplines, including biology, computer science, and engineering. This project developed a powerful AI model grounded in the principles of physics and biological mechanisms to predict protein structures. It illustrates the potential of incorporating prior knowledge about data or tasks as a valuable guide for representation learning techniques. Such prior knowledge can take various forms, including expertise in specific fields like medicine and biology, physics-based principles, or longitudinal data. Incorporating this

information into AI models not only enhances their performance but also offers insights into the underlying mechanisms and phenomena, potentially exceeding the capabilities of human comprehension. Following this direction, our objective is to enhance the integration of medical follow-up mechanisms to refine the modeling of longitudinal imaging data. This refinement would enable a more accurate representation and inference of trajectories associated with transplants, facilitating comprehensive analyses of inter and intra-patient statistics concerning these trajectories. A specific focus on the examination of our longitudinal imaging data in the context of complications events and adjustments in drug treatments holds significant potential for clinical applications.

### 5.2.2 Developing foundation models for medical imaging

In our previous chapter, we introduced an innovative approach that involves integrating structural data into medical imaging models using Large Language Models (LLMs). This strategy harnesses the power of textual data and leverages state-of-the-art NLP encoders, including large transformer models like BERT, which have been trained on extensive collections of clinical and biological notes. We anticipate a growing role for LLMs in the realm of medical imaging, as they offer robust tools that can be seamlessly integrated into AI frameworks. Our research utilized a limited set of clinical variables to generate prompts and textual data augmentations. Expanding the diversity and richness of prompts derived from LLMs could potentially enhance our training processes and the quality of our learned representations. Furthermore, our evaluation of the pretrained image encoder was primarily focused on the renal function forecasting task. Future investigations could explore additional downstream tasks, such as text or image retrieval, to further underscore the value of multimodal approaches. Regarding the modeling and architecture aspects, we established a strong baseline by training one encoder per modality using multiview contrastive learning. Future work involves potential architectural enhancements, such as integrating cross-attention mechanisms to better capture complex cross-modal relationships or directly incorporating prompt mechanisms into our optimization loop. Recent studies have introduced appealing methodologies for adapting existing foundation models, including the use of adapter modules [Wu, 2023] in the Segment Anything Model [Kirillov, 2023] and the exploration of medical and clinical knowledge within existing LLMs [Singhal, 2023a]. Moreover, the literature on LLMs continues to introduce novel innovations for training our models. Techniques like prompt tuning [Lester, 2021] and the chain of thoughts [Wei, 2022b] approach have shown promise in optimizing LLM performance. Additionally, this research contributes to a deeper understanding of transformer architecture at a larger scale, with emerging technical advancements such as rotary positional embeddings [Su, 2021] and flash attention [Dao, 2022], which hold the potential to enhance our methodologies.

Another potential avenue of exploration involves the development of dedicated foundation models tailored specifically for medical imaging, rather than relying on foundation models from other domains (primarily NLP and CV) and employing transfer learning or fine-tuning strategies. In this case, the primary challenge revolves around data scarcity. Given the diversity, intricacy, and specificity of medical applications, an essential consideration is whether it is realistic to expect a single, all-encompassing model to perform exceptionally across a wide range of medical tasks and organs. Recent studies in the field of medical imaging have started



to address this issue by initially focusing on specialized applications. These endeavors have given rise to foundation models for applications such as histopathology whole slide images [Chen, 2022a] and endoscopic videos [Wang, 2023b]. A future direction for our work would be to design a foundation imaging model for renal transplant status monitoring and prediction. The next section describes the challenges and potential directions to tackle this appealing direction.

### 5.2.3 Next AI tools for renal transplant monitoring

Our studies advocate the potential of AI algorithms as robust tools for the analysis of medical imaging data in the context of apprehending complications in renal transplants, forecasting kidney function, and overall graft survival. However, it is imperative to acknowledge the limitations imposed by the scale of our cohorts, which originate from a single hospital. A critical direction for improvement lies in collaborating across multiple sources of data from various sites and distinct research groups. Such collaboration holds the promise of bolstering the credibility of our findings. Incorporating data from diverse populations, utilizing different imaging devices, and, potentially, accommodating various acquisition protocols would serve the dual purpose of mitigating biases in our research and analysis and illustrating the resilience and adaptability of the models we employ. This collaborative approach would be invaluable in advancing the field of medical imaging analysis and demonstrating its utility in clinical settings.

While our research has been primarily centered on the analysis of imaging data and the extraction of pertinent information, we have also made efforts to integrate clinical and biological information within the proposed frameworks. However, renal transplantation monitoring is inherently multidisciplinary, encompassing nephrology, urology, radiology, and potentially a range of other medical specialties, each tailored to individual patient needs. In the pursuit of precision (or personalized) medicine, the development of models capable of processing multiple modalities to aid in diagnosis and prognosis holds significant promise for clinical practitioners. The integration of multimodality and the fusion of diverse data sources are paramount to effectively emulate the multi-disciplinary nature of medical decision-making. In the context of AI frameworks, the ongoing challenges lie in the intricacies of curating complex multimodal datasets and, equally importantly, in designing approaches that can manage and interpret this wealth of information in a balanced and comprehensible manner. These efforts are instrumental in realizing the potential of multimodality for our clinical application.

Finally, in addition to addressing potential biases and enhancing robustness through data, there is a pressing need to focus on improving the robustness and generalizability of our models and methodologies. In our initial Chapter 1, we introduced several approaches aimed at producing better-calibrated models and elucidated methodologies for model explainability. One notable example is the utilization of deep network model activations to provide predictions, coupled with heatmap visualization (as demonstrated in Chapter 3). The development of such techniques is crucial in the context of medical applications. Furthermore, the work presented in this thesis, encompassing ongoing research and experimentation, primarily revolved around methodological innovations. Beyond those we have employed, a plethora of other tools, particularly statistical

ones, are readily available to facilitate more comprehensive analyses of the results obtained. Leveraging these tools opens up the possibility of addressing clinically relevant questions and ensuring the applicability and integrity of the proposed models from a clinical point of view.

#### 5.2.4 Going beyond this thesis

Beyond the valuable contributions of this thesis, some of its limitations stem from unexplored directions. The focus of this thesis has revolved around recent deep learning-based methodologies in the realm of medical image analysis, encompassing areas such as representation learning, self-supervision, and contrastive learning. It also delved into aspects of multimodality, the interpretability of deep networks, sequential data modeling, and the fusion of vision and language models. While our community is actively exploring numerous other research directions, we would like to highlight two specific prospects that hold both excitement and significance in contemporary literature. These prospects not only offer exciting directions for further exploration but also underscore pertinent issues in the context of medical applications.

The first area of interest pertains to generative modeling, particularly the recent advances in diffusion models. These models have their roots in the concept of learning to reverse a noising or diffusion process [Sohl-Dickstein, 2015; Song, 2019; Ho, 2020], offering a novel paradigm for learning data distributions and generating new data points. Diffusion models have demonstrated state-of-the-art capabilities in content generation, particularly in computer vision tasks, and have even played a role in crafting the visually appealing chapter covers within this thesis. Much like the competition for superior Large Language Models (LLMs), innovation in diffusion models is progressing rapidly, with novel approaches aimed at enhancing the generation process [Dhariwal, 2021; Ho, 2020] and providing editing tools for both existing and generated images [Hertz, 2022; Brooks, 2023]. In the realm of medical imaging, a growing body of recent studies highlights the relevance of such models across a range of tasks, including artifact restoration and image denoising [He, 2023; Shen, 2023], image acquisition and reconstruction [Pan, 2023; Han, 2023], medical image synthesis [Jiang, 2023; Peng, 2023], and certification of segmentation models [Laousy, 2023]. However, given the complexity of the data and the intricate phenomena at stake in medical imaging applications, the application of generative models for clinical purposes must be approached with caution. These generative approaches give rise to critical considerations for clinical practice. Following this direction in our work, diffusion models present a promising avenue for modeling longitudinal data, particularly in addressing challenges related to varying sequence lengths and missing data. This potential has been explored in a recent study by Yoon et al. [Yoon, 2023], which focused on a 3D longitudinal cardiac MRI dataset.

The second critical domain of consideration involves fairness, privacy, ethics, social responsibility, and transparency within AI frameworks. A notable study has shed light on the systemic biases that can influence AI models, resulting in unfair discrimination against certain demographic groups (e.g., based on sex or skin color) across various scenarios [Zou, 2018], including machine translation and image classification. The field of medical image analysis is no exception to this challenge, with issues such as sex imbalance observed in medical imaging datasets [Ricci Lara, 2022]. The origins of biased systems can be traced to three key factors:

the data used for model training, design choices made in model development, and the individuals responsible for creating these systems. Ensuring fairness in healthcare access is paramount, and exploring strategies to mitigate those biases is imperative. The aforementioned studies and a vast body of literature offer valuable solutions to address bias and promote the development of more equitable algorithms. While it may seem demanding to incorporate these solutions, along with other principles such as calibration and explainability, into our research methodologies, a simple initial step is to consistently report the demographic characteristics of the data used in our studies. Privacy is another critical facet of consideration, as discussed by Kaissis et al. [Kaissis, 2020] in the context of machine learning algorithms for medical imaging. Medical data are inherently sensitive, and ensuring the security of systems deployed in clinical settings is essential.

### 5.3 Conclusion

In this thesis, we provided novel deep learning approaches to tackle several challenges in medical image analysis in the context of renal transplantation monitoring. Our methodologies focused on innovative and medical-oriented representation learning techniques to obtain pertinent manifolds of our data and provide relevant information for different downstream tasks for renal transplantation. We hypothesized kidney transplants to be a significant source of data for imaging investigations. Our studies, focusing on innovative methodological contributions, demonstrated the relevance of imaging data for renal transplant function forecasting and graft survival prediction. Having the chance to pursue this multi-disciplinary research with active medical and computer scientists collaborators, our projects led to original publications in top-tier international conferences. Although our private dataset provided numerous stimulating aspects, such as various imaging modalities, availability of clinical attributes, and longitudinal examinations, to pursue challenging research and tackle clinically relevant investigations, the limited number of observations (patients) and no access to external site validation sets hinder the potential impact of our results. Nevertheless, our research supports a strong proof of concept for such approaches and advocates the use of cutting-edge deep learning methodologies for medical image analysis. We aspire to set the path for future research directions from collaborating in a multi-disciplinary environment with clinicians, computer scientists, and engineers toward more expressive, robust, and generalizable features and information extractor models for medical applications and ultimately improved patient care.





# Bibliography

- [Abdulahi, 2015] Wehib. A. Abdulahi and Jules. R. Tapamo. "Fast Chan-Vese without edges and connected component analysis for kidney segmentation in MRI images". In: *AFRICON*. IEEE, Inc., 2015, pp. 1–5 (cit. on p. 47).
- [Alayrac, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, et al. "Flamingo: a Visual Language Model for Few-Shot Learning". In: *NeurIPS*. Curran Associates, Inc., 2022 (cit. on p. 17).
- [Alnazer, 2021] Israa Alnazer, Pascal Bourdon, Thierry Urruty, Omar Falou, Mohamad Khalil, Ahmad Shahin, and Christine Fernandez-Maloigne. "Recent advances in medical image processing for the evaluation of chronic kidney disease". In: *Medical Image Analysis* 69 (Apr. 2021), p. 101960 (cit. on pp. 66, 82).
- [Alsentzer, 2019] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B A Mcdermott. "Publicly Available Clinical BERT Embeddings". In: *ClinicalNLP*. Association for Computational Linguistics (ACL), July 2019, pp. 72–78 (cit. on pp. 95, 100).
- [Amari, 1967] Shunichi Amari. "A Theory of Adaptive Pattern Classifiers". In: *IEEE Transactions on Electronic Computers* EC-16.3 (1967), pp. 299–307 (cit. on p. 9).
- [Anil, 2023] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, et al. "PaLM 2 Technical Report". In: (May 2023) (cit. on p. 16).
- [Arjovsky, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. "Invariant Risk Minimization". In: (July 2019) (cit. on p. 90).
- [Azizi, 2021] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, et al. "Big Self-Supervised Models Advance Medical Image Classification". In: *ICCV*. IEEE, Inc., 2021, pp. 3458–3468 (cit. on pp. 20, 89).
- [Bachman, 2019] Philip Bachman, R Devon Hjelm, and William Buchwalter. "Learning Representations by Maximizing Mutual Information Across Views". In: *NeurIPS*. Curran Associates, Inc., 2019 (cit. on p. 90).
- [Baheti, 2021] Bhakti Baheti, Diana Waldmannstetter, Satrajit Chakrabarty, Hamed Akbari, Michel Bilello, et al. "The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients". In: (Dec. 2021) (cit. on p. 21).
- [Bakas, 2018] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, et al. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge". In: (Nov. 2018) (cit. on p. 21).
- [Balestrieri, 2023] Randall Balestrieri, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, et al. "A Cookbook of Self-Supervised Learning". In: (Apr. 2023) (cit. on p. 14).
- [Bao, 2021] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. "BEiT: BERT Pre-Training of Image Transformers". In: *ICLR*. Curran Associates, Inc., 2021 (cit. on p. 15).
- [Bardes, 2021] Adrien Bardes, Jean Ponce, and Yann LeCun. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning". In: *ICLR*. Curran Associates, Inc., 2021 (cit. on p. 15).
- [Barron, 1994] Andrew R Barron. "Approximation and estimation bounds for artificial neural networks". In: *Machine Learning* 14.1 (Jan. 1994), pp. 115–133 (cit. on p. 11).
- [Baxter, 2001] G. M. Baxter. "Ultrasound of Renal Transplantation". In: *Clinical Radiology* 56.10 (Oct. 2001), pp. 802–818 (cit. on p. 31).
- [Becker, 2020] Jan U. Becker, David Mayerich, Meghana Padmanabhan, Jonathan Barratt, Angela Ernst, Peter Boor, Pietro A. Cicalese, Chandra Mohan, Hien V. Nguyen, and Badrinath Roysam. "Artificial intelligence and machine learning in nephropathology". In: *Kidney International* 98.1 (July 2020), pp. 65–75 (cit. on p. 25).
- [Becker, 2021] Stefan Becker, Ronny Hug, Wolfgang Huebner, Michael Arens, and Brendan Tran Morris. "MissFormer: (In-)Attention-Based Handling of Missing Observations for Trajectory Filtering and Prediction". In: *ISVC*. Vol. 13017 LNCS. Springer, 2021, pp. 521–533 (cit. on p. 66).

- [Belharbi, 2022] Soufiane Belharbi, Jérôme Rony, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. “Deep Interpretable Classification and Weakly-Supervised Segmentation of Histology Images via Max-Min Uncertainty”. In: *IEEE Transactions on Medical Imaging* 41.3 (Mar. 2022), pp. 702–714 (cit. on p. 20).
- [Bindu, 2012] Hima Bindu and Satya K. Prasad. “An Efficient Medical Image Segmentation Using Conventional OTSU Method”. In: *International Journal of Advanced Science and Technology* 38 (Jan. 2012) (cit. on p. 54).
- [Bodard, 2023] Sylvain Bodard, Louise Denis, Vincent Hingot, Arthur Chavignon, Olivier Hélénon, Dany Anglicheau, Olivier Couture, and Jean Michel Correas. “Ultrasound localization microscopy of the human kidney allograft on a clinical ultrasound scanner”. In: *Kidney International* 103.5 (May 2023), pp. 930–935 (cit. on p. 82).
- [Boyd, 2021] Joseph Boyd, Mykola Liashuha, Eric Deutsch, Nikos Paragios, Stergios Christodoulidis, and Maria Vakalopoulou. “Self-Supervised Representation Learning using Visual Field Expansion on Digital Pathology”. In: *ICCVW*. Vol. 2021-October. IEEE, Inc., Oct. 2021, pp. 639–647 (cit. on pp. 20, 89).
- [Breiman, 2001] Leo Breiman. “Random forests”. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32 (cit. on p. 7).
- [Breiman, 1984] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Routledge, 1984, pp. 1–358 (cit. on p. 7).
- [Brooks, 2023] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. “InstructPix2Pix: Learning to Follow Image Editing Instructions”. In: *CVPR*. IEEE, Inc., 2023 (cit. on p. 113).
- [Brown, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. “Language Models are Few-Shot Learners”. In: *NeurIPS*. Curran Associates, Inc., 2020 (cit. on pp. 16, 92).
- [Cai, 2021] Siqi Cai, Peiwen Li, Enze Su, and Longhan Xie. “Auditory Attention Detection via Cross-Modal Attention”. In: *Frontiers in Neuroscience* 15 (July 2021) (cit. on p. 19).
- [Caron, 2020] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: *NeurIPS*. Curran Associates, Inc., 2020 (cit. on p. 15).
- [Caron, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. “Emerging Properties in Self-Supervised Vision Transformers”. In: *ICCV*. IEEE, Inc., Apr. 2021 (cit. on p. 14).
- [Chattopadhyay, 2018] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. “Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks”. In: *WACV*. IEEE, Inc., 2018 (cit. on p. 20).
- [Chen, 2022a] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. “Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning”. In: *CVPR*. IEEE, Inc., 2022 (cit. on p. 112).
- [Chen, 2021] Richard J. Chen, Ming Y. Lu, Wei Hung Weng, Tiffany Y. Chen, Drew F.K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. “Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images”. In: *ICCV*. IEEE, Inc., 2021, pp. 3995–4005 (cit. on pp. 20, 22, 90).
- [Chen, 2019] Sihong Chen, Kai Ma, and Yefeng Zheng. “Med3D: Transfer Learning for 3D Medical Image Analysis”. In: (Apr. 2019) (cit. on p. 73).
- [Chen, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. JMLR, 2020 (cit. on pp. 14, 73, 92).
- [Chen, 2022b] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, et al. “PaLI: A Jointly-Scaled Multilingual Language-Image Model”. In: *ICLR*. Curran Associates, Inc., 2022 (cit. on p. 18).

- [Chen, 2020b] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning". In: *CVPR*. IEEE, Inc., 2020 (cit. on p. 14).
- [Cherukuri, 2010] Aravind Cherukuri, Matthew P. Welberry-Smith, James E. Tattersall, Niaz Ahmad, Chas G. Newstead, Andrew J. P. Lewington, and Richard J. Baker. "The Clinical Significance of Early Proteinuria After Renal Transplantation". In: *Transplantation* 89.2 (Jan. 2010), pp. 200–207 (cit. on p. 28).
- [Cho, 2014] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *EMNLP*. Association for Computational Linguistics, 2014 (cit. on p. 12).
- [Chung, 2014] Yong Eun Chung and Ki Whang Kim. "Contrast-enhanced ultrasonography: advance and current status in abdominal imaging". In: *Ultrasonography* 34.1 (Sept. 2014), pp. 3–18 (cit. on p. 31).
- [Cireşan, 2012] Dan Cireşan, Ueli Meier, and Juergen Schmidhuber. "Multi-column Deep Neural Networks for Image Classification". In: *CVPR*. IEEE, Inc., 2012 (cit. on p. 9).
- [Cortes, 1995] Corinna Cortes, Vladimir Vapnik, and Lorenza Saitta. "Support-vector networks". In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297 (cit. on p. 7).
- [Couronné, 2021] Raphaël Couronné, Paul Vernhet, and Stanley Durrleman. "Longitudinal Self-supervision to Disentangle Inter-patient Variability from Disease Progression". In: *MICCAI*. Vol. 12902. Springer, 2021, pp. 231–241 (cit. on pp. 33, 80).
- [Cover, 1967] T. M. Cover and P. E. Hart. "Nearest Neighbor Pattern Classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27 (cit. on pp. 7, 77).
- [Cox, 1972] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (Jan. 1972), pp. 187–202 (cit. on pp. 43, 76).
- [Cramer, 2002] J.S. Cramer. "The Origins of Logistic Regression". In: *Tinbergen Institute Working Paper* 119.4 (Dec. 2002) (cit. on p. 7).
- [Cybenko, 1989] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314 (cit. on p. 11).
- [Dalca, 2019] Adrian V. Dalca, Katherine L. Bouman, William T. Freeman, Natalia S. Rost, Mert R. Sabuncu, and Polina Golland. "Medical Image Imputation From Image Collections". In: *IEEE Transactions on Medical Imaging* 38.2 (Feb. 2019), pp. 504–514 (cit. on p. 66).
- [Dao, 2022] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness". In: *NeurIPS*. Curran Associates, Inc., 2022 (cit. on p. 111).
- [Deng, 2010] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *CVPR*. IEEE, Inc., 2010, pp. 248–255 (cit. on p. 73).
- [Devlin, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT*. Oct. 2018 (cit. on pp. 15, 66, 95, 100).
- [Dhariwal, 2021] Prafulla Dhariwal and Alex Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *NeurIPS*. Curran Associates, Inc., 2021 (cit. on p. 113).
- [Doersch, 2015] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction". In: *ICCV*. IEEE, Inc., May 2015 (cit. on p. 14).
- [Dosovitskiy, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. Curran Associates, Inc., 2020 (cit. on pp. 11, 12).
- [Esteva, 2017] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (Jan. 2017), pp. 115–118 (cit. on p. 24).



- [Fang, 2020] Lingling Fang, Xin Wang, and Lujie Wang. “Multi-modal medical image segmentation based on vector-valued active contour models”. In: *Information Sciences* 513 (Mar. 2020), pp. 504–518 (cit. on p. 51).
- [Ferrante, 2017] Enzo Ferrante, Puneet K. Dokania, Rafael Marini, and Nikos Paragios. “Deformable Registration Through Learning of Context-Specific Metric Aggregation”. In: *MLMI*. 2017, pp. 256–265 (cit. on p. 51).
- [First, 2003] M. R. First. “Renal function as a predictor of long-term graft survival in renal transplant patients”. In: *Nephrology Dialysis Transplantation* 18.90001 (Jan. 2003), pp. 3i–6 (cit. on p. 28).
- [Funahashi, 1989] Ken Ichi Funahashi. “On the approximate realization of continuous mappings by neural networks”. In: *Neural Networks* 2.3 (Jan. 1989), pp. 183–192 (cit. on p. 11).
- [García-Laencina, 2010] Pedro J. García-Laencina, José Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. “Pattern classification with missing data: A review”. In: *Neural Computing and Applications* 19.2 (Sept. 2010), pp. 263–282 (cit. on p. 66).
- [Gardent, 2017] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. “The WebNLG Challenge: Generating Text from RDF Data”. In: *INLG*. Association for Computational Linguistics, 2017, pp. 124–133 (cit. on pp. 99, 100).
- [Ghonge, 2021] Nitin P Ghonge, Nidhi Goyal, Sandeep Vohra, and Veena Chowdhury. “Renal transplant evaluation: multimodality imaging of post-transplant complications”. In: *The British Journal of Radiology* 94.1124 (Aug. 2021), p. 20201253 (cit. on p. 25).
- [Giuliani, 2020] Francesco Giuliani, Irtiza Hasan, Marco Cristani, and Fabio Galasso. “Transformer Networks for Trajectory Forecasting”. In: *ICPR*. Mar. 2020 (cit. on p. 66).
- [Gonzales, 2016] Manuel Moreno Gonzales, Andrew Bentall, Walter K. Kremers, Mark D. Stegall, and Richard Borrows. “Predicting individual renal allograft outcomes using risk models with 1-year surveillance biopsy and alloantibody data”. In: *Journal of the American Society of Nephrology* 27.10 (2016), pp. 3165–3174 (cit. on p. 43).
- [Goodfellow, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Networks”. In: *NeurIPS*. Curran Associates, Inc., 2014 (cit. on p. 10).
- [Goyal, 2022] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. “Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision”. In: (Feb. 2022) (cit. on p. 14).
- [Grenier, 2013] N. Grenier, J. L. Gennisson, F. Cornelis, Y. Le Bras, and L. Couzi. “Renal ultrasound elastography”. In: *Diagnostic and Interventional Imaging* 94.5 (May 2013), pp. 545–550 (cit. on p. 31).
- [Grenier, 2012] Nicolas Grenier, Séverine Poulain, Sébastien Lepreux, Jean Luc Gennisson, Benjamin Dal-laudière, et al. “Quantitative elastography of renal transplants using supersonic shear imaging: A pilot study”. In: *European Radiology* 22.10 (Oct. 2012), pp. 2138–2146 (cit. on p. 31).
- [Grill, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, et al. “Bootstrap your own latent: A new approach to self-supervised Learning”. In: *NeurIPS*. Curran Associates, Inc., 2020 (cit. on p. 14).
- [Guyon, 2002] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. “Gene selection for cancer classification using support vector machines”. In: *Machine Learning* 46.1-3 (2002), pp. 389–422 (cit. on p. 6).
- [Hager, 2023] Paul Hager, Martin J. Menten, and Daniel Rueckert. “Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data”. In: *CVPR*. IEEE, Inc., 2023 (cit. on pp. 21, 22).
- [Han, 2023] Zeyu Han, Yuhan Wang, Luping Zhou, Peng Wang, Binyu Yan, Jiliu Zhou, Yan Wang, and Dinggang Shen. “Contrastive Diffusion Model with Auxiliary Guidance for Coarse-to-Fine PET Reconstruction”. In: *MICCAI*. Vol. 14229. Springer, 2023, pp. 239–249 (cit. on p. 113).

- [Hand, 2001] David J. Hand and Keming Yu. "Idiot's Bayes: Not So Stupid after All?" In: *International Statistical Review* 69.3 (Dec. 2001), p. 385 (cit. on p. 7).
- [Hariharan, 2021] Sundaram Hariharan, Ajay K. Israni, and Gabriel Danovitch. "Long-Term Survival after Kidney Transplantation". In: *New England Journal of Medicine* 385.8 (Aug. 2021), pp. 729–743 (cit. on pp. 24, 65, 89).
- [He, 2021] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked Autoencoders Are Scalable Vision Learners". In: *CVPR*. IEEE, Inc., 2021 (cit. on p. 15).
- [He, 2015a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *CVPR*. IEEE, Inc., 2015 (cit. on p. 12).
- [He, 2015b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *ICCV*. IEEE, Inc., 2015 (cit. on p. 11).
- [He, 2009] Xiang He, Jason Moore, Shazia Shabir, Mark A. Little, Paul Cockwell, Simon Ball, Xiang Liu, Atholl Johnston, and Richard Borrows. "Comparison of the predictive performance of eGFR formulae for mortality and graft failure in renal transplant recipients". In: *Transplantation* 87.3 (Feb. 2009), pp. 384–392 (cit. on p. 43).
- [He, 2023] Zhenqi He, Junjun He, Jin Ye, and Yiqing Shen. "Artifact Restoration in Histology Images with Diffusion Probabilistic Models". In: *MICCAI*. Springer, 2023, pp. 518–527 (cit. on p. 113).
- [Hendrycks, 2019] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. "Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty". In: *NeurIPS*. Curran Associates, Inc., 2019 (cit. on p. 14).
- [Hertz, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. "Prompt-to-Prompt Image Editing with Cross Attention Control". In: (Aug. 2022) (cit. on p. 113).
- [Ho, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *NeurIPS*. Curran Associates, Inc., 2020 (cit. on pp. 10, 113).
- [Hochreiter, 1997] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780 (cit. on p. 12).
- [Hoffmann, 2022] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. "Training Compute-Optimal Large Language Models". In: *NeurIPS*. Curran Associates, Inc., 2022 (cit. on p. 16).
- [Hornik, 1989] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366 (cit. on p. 11).
- [Hotelling, 1933] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (Sept. 1933), pp. 417–441 (cit. on p. 6).
- [Hu, 2022] Ricky Hu, Rohit Singla, Cailin Ringstrom, Zoe Hu, Victoria Lessoway, Janice Reid, Timothy Murray, Christopher Nguan, and Robert N. Rohling. "Prediction of Kidney Transplant Function with Machine Learning from Computational Ultrasound Features". In: *MICCAI Workshop ASMUS*. Vol. 13565 LNCS. Springer, 2022, pp. 34–43 (cit. on p. 57).
- [Huang, 2016] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *CVPR*. IEEE, Inc., 2016 (cit. on p. 12).
- [Huang, 2023] Zhi Huang, Federico Bianchi, Mert Yuksekogul, Thomas J. Montine, and James Zou. "A visual-language foundation model for pathology image analysis using medical Twitter". In: *Nature Medicine* 29.9 (Aug. 2023), pp. 2307–2316 (cit. on p. 22).
- [Jaegle, 2021] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. "Perceiver: General Perception with Iterative Attention". In: *ICML*. JMLR, 2021 (cit. on p. 12).

- [Jia, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision". In: *ICML*. JMLR, 2021 (cit. on pp. 17, 90).
- [Jiang, 2023] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. "CoLa-Diff: Conditional Latent Diffusion Model for Multi-modal MRI Synthesis". In: *MICCAI*. Springer, 2023, pp. 398–408 (cit. on p. 113).
- [Jing, 2022] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. "Understanding Dimensional Collapse in Contrastive Self-supervised Learning". In: *ICLR*. Curran Associates, Inc., 2022 (cit. on p. 14).
- [Johnson, 2016] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.1 (May 2016), pp. 1–9 (cit. on p. 95).
- [Jumper, 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (July 2021), pp. 583–589 (cit. on p. 110).
- [Kaissis, 2020] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. "Secure, privacy-preserving and federated machine learning in medical imaging". In: *Nature Machine Intelligence* 2.6 (June 2020), pp. 305–311 (cit. on p. 114).
- [Kaplan, 2003] Bruce Kaplan, Jesse Schold, and Herwig Ulf Meier-Kriesche. "Poor Predictive Value of Serum Creatinine for Renal Allograft Loss". In: *American Journal of Transplantation* 3.12 (Dec. 2003), pp. 1560–1565 (cit. on p. 43).
- [Kaplan, 1958] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (June 1958), p. 457 (cit. on pp. 43, 76).
- [Karpathy, 2016] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. "Visualizing and Understanding Recurrent Networks". In: *ICLR Workshop*. Curran Associates, Inc., 2016 (cit. on p. 12).
- [Kasiske, 1989] Bertram L. Kasiske. "Creatinine excretion after renal transplantation". In: *Transplantation* 48.3 (Sept. 1989), pp. 424–427 (cit. on p. 28).
- [Kazemina, 2020] Salome Kazemina, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. "GANs for medical image analysis". In: *Artificial Intelligence in Medicine* 109 (Sept. 2020), p. 101938 (cit. on p. 66).
- [Kazmierski, 2020] Brittany J. Kazmierski, Kedar G. Sharbidre, Michelle L. Robbin, and Edward G. Grant. "Contrast-Enhanced Ultrasound for the Evaluation of Renal Transplants". In: *Journal of Ultrasound in Medicine* 39.12 (Dec. 2020), pp. 2457–2468 (cit. on p. 31).
- [Kenton, 2021] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. "Alignment of Language Agents". In: (Mar. 2021) (cit. on p. 16).
- [Keshavan, 2009] Raghunandan H. Keshavan, Sewoong Oh, and Andrea Montanari. "Matrix completion from a few entries". In: *ISIT*. IEEE, Inc., 2009, pp. 324–328 (cit. on p. 66).
- [Khalifa, 2013] Fahmi Khalifa, Garth M. Beache, Mohamed Abou El-Ghar, Tarek El-Diasty, Georgy Gimel'Farb, Maiying Kong, and Ayman El-Baz. "Dynamic contrast-enhanced MRI-based early detection of acute renal transplant rejection". In: *IEEE Transactions on Medical Imaging* 32.10 (2013), pp. 1910–1927 (cit. on p. 66).
- [Khosla, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. "Supervised Contrastive Learning". In: *NeurIPS*. Curran Associates, Inc., 2020 (cit. on p. 14).
- [Kim, 2020] Wonjick Kim, Asako Kanezaki, and Masayuki Tanaka. "Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8055–8068 (cit. on pp. 48, 49, 54, 55).

- [Kingma, 2014a] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *ICLR*. Curran Associates, Inc., 2014 (cit. on p. 10).
- [Kingma, 2014b] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. Curran Associates, Inc., 2014 (cit. on pp. 70, 95).
- [Kirillov, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, et al. "Segment Anything". In: *ICCV*. IEEE, Inc., 2023 (cit. on pp. 17, 111).
- [Koochpayegani, 2021] Soroush Abbasi Koochpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. "Mean Shift for Self-Supervised Learning". In: *ICCV*. IEEE, Inc., May 2021 (cit. on p. 14).
- [Kotu, 2015a] Vijay Kotu and Bala Deshpande. "Data Mining Process". In: *Predictive Analytics and Data Mining*. Morgan Kaufmann, Jan. 2015, pp. 17–36 (cit. on p. 7).
- [Kotu, 2015b] Vijay Kotu and Bala Deshpande. "Feature Selection". In: *Predictive Analytics and Data Mining*. Morgan Kaufmann, Jan. 2015, pp. 347–370 (cit. on p. 6).
- [Kraskov, 2004] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Physical Review E* 69.6 (June 2004), p. 16 (cit. on p. 6).
- [Krishnan, 2022] Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. "Self-supervised learning in medicine and healthcare". In: *Nature Biomedical Engineering* 6.12 (Aug. 2022), pp. 1346–1352 (cit. on pp. 20, 81, 89).
- [Krizhevsky, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *NeurIPS*. Curran Associates, Inc., 2012 (cit. on p. 9).
- [Laousy, 2023] Othmane Laousy, Alexandre Araujo, Guillaume Chassagnon, Nikos Paragios, Marie-Pierre Revel, and Maria Vakalopoulou. "Certification of Deep Learning Models for Medical Image Segmentation". In: *MICCAI*. Springer, 2023, pp. 611–621 (cit. on p. 113).
- [Larrazabal, 2023] Agostina Larrazabal, Cesar Martinez, Jose Dolz, and Enzo Ferrante. "Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation". In: *MICCAI*. Springer, 2023 (cit. on p. 20).
- [LeCun, 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (Dec. 1989), pp. 541–551 (cit. on p. 9).
- [Lee, 2011] Jong Hoon Lee, Je Hwan Won, and Chang Kwon Oh. "Impact of the ratio of graft kidney volume to recipient body surface area on graft function after live donor kidney transplantation". In: *Clinical Transplantation* 25.6 (Nov. 2011), E647–E655 (cit. on p. 47).
- [Lester, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning". In: *EMNLP*. Association for Computational Linguistics, 2021 (cit. on p. 111).
- [Levey, 1999] Andrew S. Levey. "A More Accurate Method To Estimate Glomerular Filtration Rate from Serum Creatinine: A New Prediction Equation". In: *Annals of Internal Medicine* 130.6 (Mar. 1999), p. 461 (cit. on p. 28).
- [Levey, 2009] Andrew S. Levey, Lesley A. Stevens, Christopher H. Schmid, Yaping (Lucy) Zhang, Alejandro F. Castro, et al. "A New Equation to Estimate Glomerular Filtration Rate". In: *Annals of Internal Medicine* 150.9 (May 2009), p. 604 (cit. on p. 28).
- [Li, 2016] Hong Li, Nan Bao, Xieping Xu, Yaonan Zhang, Shikai Jin, Yueming Jin, and Haoran Sun. "A renal vascular compartment segmentation method based on dynamic contrast-enhanced images". In: *Technology and Health Care* 24.s2 (Jan. 2016), S631–S639 (cit. on p. 47).
- [Li, 2012] Sheng Li, Frank G. Zöllner, Andreas D. Merrem, Yinghong Peng, Jarle Roervik, Arvid Lundervold, and Lothar R. Schad. "Wavelet-based segmentation of renal compartments in DCE-MRI of human kidney: Initial results in patients and healthy volunteers". In: *Computerized Medical Imaging and Graphics* 36.2 (Mar. 2012), pp. 108–118 (cit. on p. 47).
- [Lim, 2023] Ee Jean Lim, Jie Yen, Khi Yung Fong, Ho Yee Tiong, Edwin Jonathan Aslim, et al. "Radiomics in Kidney Transplantation: A Scoping Review of Current Applications, Limitations, and Future Directions". In: *Transplantation* (June 2023) (cit. on p. 57).

- [Liu, 2022] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. “The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration”. In: *CVPR*. IEEE, Inc., 2022 (cit. on p. 20).
- [Liu, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *ICCV*. IEEE, Inc., Mar. 2021 (cit. on p. 12).
- [Loshchilov, 2017] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *ICLR*. Curran Associates, Inc., 2017 (cit. on p. 95).
- [Loubeyre, 1996] Pierre Loubeyre, Remy Cahen, Florence Grozel, Pierre Trolliet, Claire Pouteil-Noble, Michel Labeeuw, and Van Andre Tran Minh. “Transplant renal artery stenosis. Evaluation of diagnosis with magnetic resonance angiography compared with color duplex sonography and arteriography”. In: *Transplantation* 62.4 (Aug. 1996), pp. 446–450 (cit. on p. 25).
- [Loupy, 2019] Alexandre Loupy, Olivier Aubert, Babak J. Orandi, Maarten Naesens, Yassine Bouatou, et al. “Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study”. In: *BMJ* 366 (Sept. 2019) (cit. on pp. 43, 68, 82).
- [Lu, 2021] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. “Pretrained Transformers as Universal Computation Engines”. In: *AAAI*. Mar. 2021 (cit. on p. 100).
- [Lu, 2023] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, et al. “Towards a Visual-Language Foundation Model for Computational Pathology”. In: (July 2023) (cit. on p. 22).
- [Lu, 2017] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. “The Expressive Power of Neural Networks: A View from the Width”. In: *NeurIPS*. Curran Associates, Inc., 2017 (cit. on p. 11).
- [Lundervold, 2019] Alexander Selvikvåg Lundervold and Arvid Lundervold. “An overview of deep learning in medical imaging focusing on MRI”. In: *Zeitschrift für Medizinische Physik* 29.2 (May 2019), pp. 102–127 (cit. on p. 47).
- [Maas, 2013] Andrew L. Maas. “Rectifier Nonlinearities Improve Neural Network Acoustic Models”. In: *ICML*. JMLR, 2013 (cit. on p. 11).
- [Maaten, 2008] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605 (cit. on p. 96).
- [Mackinnon, 2010] A. Mackinnon. “The use and reporting of multiple imputation in medical research – a review”. In: *Journal of Internal Medicine* 268.6 (Dec. 2010), pp. 586–593 (cit. on p. 66).
- [Mannil, 2018] Manoj Mannil, Jochen Von Spiczak, Robert Manka, and Hatem Alkadhi. “Texture Analysis and Machine Learning for Detecting Myocardial Infarction in Noncontrast Low-Dose Computed Tomography: Unveiling the Invisible”. In: *Investigative Radiology* 53.6 (June 2018), pp. 338–343 (cit. on p. 56).
- [McCulloch, 1943] Warren S. McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The Bulletin of Mathematical Biophysics* 5.4 (Dec. 1943), pp. 115–133 (cit. on p. 9).
- [Menze, 2015] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, et al. “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)”. In: *IEEE Transactions on Medical Imaging* 34.10 (Oct. 2015), pp. 1993–2024 (cit. on p. 21).
- [Milecki, 2021a] Leo Milecki, Sylvain Bodard, Jean Michel Correas, Marc Olivier Timsit, and Maria Vakalopoulou. “3D unsupervised kidney graft segmentation based on deep learning and multi-sequence mri”. In: *ISBI*. Vol. 2021-April. IEEE Computer Society, Apr. 2021, pp. 1781–1785 (cit. on pp. 35, 41).
- [Milecki, 2021b] **Leo Milecki**, Sylvain Bodard, Jean Michel Correas, Marc Olivier Timsit, and Maria Vakalopoulou. “3D unsupervised kidney graft segmentation based on deep learning and multi-sequence mri”. In: *ISBI*. Vol. 2021-April. IEEE Computer Society, Apr. 2021, pp. 1781–1785 (cit. on p. 36).

- [Milecki, 2022a] Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean Michel Correas, Marc Olivier Timsit, and Maria Vakalopoulou. “Contrastive Masked Transformers for Forecasting Renal Transplant Function”. In: *MICCAI*. Vol. 13437 LNCS. Springer, 2022, pp. 244–254 (cit. on pp. 35, 63).
- [Milecki, 2022b] **Leo Milecki**, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean Michel Correas, Marc Olivier Timsit, and Maria Vakalopoulou. “Contrastive Masked Transformers for Forecasting Renal Transplant Function”. In: *MICCAI*. Vol. 13437 LNCS. Springer, 2022, pp. 244–254 (cit. on p. 36).
- [Milecki, 2022c] Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. “Constrative Learning for Kidney Transplant Analysis using MRI data and Deep Convolutional Networks”. In: *MIDL Short Paper Track*. 2022 (cit. on pp. 35, 63).
- [Milecki, 2022d] **Leo Milecki**, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. “Constrative Learning for Kidney Transplant Analysis using MRI data and Deep Convolutional Networks”. In: *MIDL Short Paper Track*. 2022 (cit. on p. 36).
- [Milecki, 2023a] Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. “MEDIMP: 3D Medical Images with clinical Prompts from limited tabular data for renal transplantation”. In: *MIDL*. PMLR, Apr. 2023 (cit. on pp. 35, 87).
- [Milecki, 2023b] **Leo Milecki**, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. “MEDIMP: 3D Medical Images with clinical Prompts from limited tabular data for renal transplantation”. In: *MIDL*. PMLR, Apr. 2023 (cit. on p. 36).
- [Misra, 2019] Ishan Misra and Laurens van der Maaten. “Self-Supervised Learning of Pretext-Invariant Representations”. In: *CVPR*. IEEE, Inc., 2019 (cit. on p. 14).
- [Moor, 2023] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. “Foundation models for generalist medical artificial intelligence”. In: *Nature* 616.7956 (Apr. 2023), pp. 259–265 (cit. on p. 22).
- [Moore, 2011] Jason Moore, Xiang He, Shazia Shabir, Rajesh Hanvesakul, David Benavente, et al. “Development and Evaluation of a Composite Risk Score to Predict Kidney Transplant Failure”. In: *American Journal of Kidney Diseases* 57.5 (May 2011), pp. 744–751 (cit. on p. 43).
- [Müller, 2021] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. “Joint Learning of Localized Representations from Medical Images and Reports”. In: *ECCV*. Springer Nature, 2021 (cit. on p. 90).
- [Murugesan, 2023] Balamurali Murugesan, Sukesh Adiga, Bingyuan Liu, Hervé Lombaert, Ismail Ben Ayed, and Jose Dolz. “Trust your neighbours: Penalty-based constraints for model calibration”. In: *MICCAI*. Springer, 2023 (cit. on p. 20).
- [Naesens, 2016] Maarten Naesens, Evelyne Lerut, Marie Paule Emonds, Albert Herelixa, Pieter Evenepoel, et al. “Proteinuria as a Noninvasive Marker for Renal Allograft Histology and Failure: An Observational Cohort Study”. In: *Journal of the American Society of Nephrology* 27.1 (Jan. 2016), pp. 281–292 (cit. on p. 43).
- [Nair, 2010] Vinod Nair and Geoffrey E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *ICML*. JMLR, 2010 (cit. on p. 11).
- [Oord, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation Learning with Contrastive Predictive Coding”. In: (July 2018) (cit. on p. 92).
- [OpenAI, 2023] OpenAI. *Chatgpt: Optimizing language models for dialogue*. Jan. 2023. URL: <https://openai.com/blog/chatgpt/> (cit. on pp. 16, 90, 92, 93).
- [Oquab, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, et al. “DI-NOv2: Learning Robust Visual Features without Supervision”. In: (Apr. 2023) (cit. on p. 17).
- [Orlacchio, 2014] Antonio Orlacchio, Fabrizio Chegai, Costantino Del Giudice, Alessandro Anselmo, Giuseppe Iaria, et al. “Kidney Transplant: Usefulness of Real-Time Elastography (RTE) in the Di-

- agnosis of Graft Interstitial Fibrosis". In: *Ultrasound in Medicine and Biology* 40.11 (Nov. 2014), pp. 2564–2572 (cit. on p. 66).
- [Ouyang, 2021] Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Edith V. Sullivan, Adolf Pfefferbaum, Greg Zaharchuk, and Kilian M. Pohl. "Self-supervised Longitudinal Neighbourhood Embedding". In: *MICCAI*. Vol. 12902 LNCS. Springer, 2021, pp. 80–89 (cit. on p. 33).
- [Ouyang, 2022a] Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Greg Zaharchuk, and Kilian M. Pohl. "Disentangling Normal Aging From Severity of Disease via Weak Supervision on Longitudinal MRI". In: *IEEE Transactions on Medical Imaging* 41.10 (Oct. 2022), pp. 2558–2569 (cit. on pp. 33, 80).
- [Ouyang, 2022b] Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Greg Zaharchuk, and Kilian M. Pohl. "Self-supervised learning of neighborhood embedding for longitudinal MRI". In: *Medical Image Analysis* 82 (Nov. 2022), p. 102571 (cit. on p. 33).
- [Pan, 2023] Mingjie Pan, Yulu Gan, Fangxu Zhou, Jiaming Liu, Ying Zhang, Aimin Wang, Shanghang Zhang, and Dawei Li. "DiffuseIR: Diffusion Models for Isotropic Reconstruction of 3D Microscopic Images". In: *MICCAI*. Springer, 2023, pp. 323–332 (cit. on p. 113).
- [Pascanu, 2013] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training Recurrent Neural Networks". In: *ICML*. JMLR, 2013 (cit. on p. 12).
- [Paszke, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *NeurIPS*. Curran Associates, Inc., 2019 (cit. on pp. 70, 95).
- [Pearson, 1895] Karl Pearson. "Note on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58.347-352 (Dec. 1895), pp. 240–242 (cit. on p. 6).
- [Pearson, 1900] Karl Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *Philosophical Magazine* 50.302 (July 1900), pp. 157–175 (cit. on p. 6).
- [Penedo, 2023] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only". In: (June 2023) (cit. on p. 16).
- [Peng, 2023] Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao, and Kilian M. Pohl. "Generating Realistic Brain MRIs via a Conditional Diffusion Probabilistic Model". In: *MICCAI*. Springer, 2023, pp. 14–24 (cit. on p. 113).
- [Pérez-García, 2021] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". In: *Computer Methods and Programs in Biomedicine* 208 (Sept. 2021), p. 106236 (cit. on pp. 70, 95).
- [Ponticelli, 2012] Claudio Ponticelli and Giorgio Graziani. "Proteinuria after kidney transplantation". In: *Transplant International* 25.9 (Sept. 2012), pp. 909–917 (cit. on p. 28).
- [Radford, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *ICML*. JMLR, 2021 (cit. on pp. 17, 19, 90, 92, 95, 100).
- [Raffel, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21 (2020), pp. 1–67 (cit. on pp. 99, 100).
- [Ramesh, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-Shot Text-to-Image Generation". In: *ICML*. JMLR, 2021 (cit. on p. 18).
- [Ren, 2022] Mengwei Ren, Neel Dey, Martin A. Styner, Kelly Botteron, and Guido Gerig. "Local Spatiotemporal Representation Learning for Longitudinally-consistent Neuroimage Analysis". In: *NeurIPS*. Curran Associates, Inc., 2022 (cit. on pp. 33, 80).



- [Ricci Lara, 2022] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. "Addressing fairness in artificial intelligence for medical imaging". In: *Nature Communications* 13.1 (Aug. 2022), pp. 1–6 (cit. on pp. 20, 113).
- [Rizzo, 2018] Stefania Rizzo, Francesca Botta, Sara Raimondi, Daniela Origgi, Cristiana Fanciullo, Alessio Giuseppe Morganti, and Massimo Bellomi. "Radiomics: the facts and the challenges of image analysis". In: *European Radiology Experimental* 2.1 (Dec. 2018) (cit. on p. 47).
- [Rombach, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *CVPR*. IEEE, Inc., 2022 (cit. on p. 18).
- [Ronneberger, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *MICCAI*. Vol. 9351. Springer, 2015, pp. 234–241 (cit. on p. 19).
- [Rosenblatt, 1958] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6 (Nov. 1958), pp. 386–408 (cit. on p. 9).
- [Roufousse, 2018] Candice Roufousse, Naomi Simmonds, Marian Clahsen-Van Groningen, Mark Haas, Kammi J. Henriksen, et al. "A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology". In: *Transplantation* 102.11 (Nov. 2018), pp. 1795–1814 (cit. on pp. 25, 29).
- [Rumelhart, 1987] David E. Rumelhart and James L. McClelland. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. 1987, pp. 318–362 (cit. on p. 12).
- [Sahasrabudhe, 2021] Mihir Sahasrabudhe, Pierre Sujobert, Evangelia I. Zacharaki, Eugenie Maurin, Beatrice Grange, Laurent Jallades, Nikos Paragios, and Maria Vakalopoulou. "Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis". In: *IEEE Journal of Biomedical and Health Informatics* 25.6 (June 2021), pp. 2125–2136 (cit. on p. 90).
- [Sammut, 2021] Stephen John Sammut, Mireia Crispin-Ortuzar, Suet Feung Chin, Elena Provenzano, Helen A. Bardwell, et al. "Multi-omic machine learning predictor of breast cancer therapy response". In: *Nature* 601.7894 (Dec. 2021), pp. 623–629 (cit. on p. 24).
- [Sandmair, 2016] Martin Sandmair, Matthias Hammon, Hannes Seuss, Ragnar Theis, Michael Uder, and Rolf Janka. "Semiautomatic segmentation of the kidney in magnetic resonance images using unimodal thresholding". In: *BMC Research Notes* 9.1 (Dec. 2016), p. 489 (cit. on p. 47).
- [Schuhmann, 2021] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs". In: *NeurIPS Workshop on Data Centric AI*. Curran Associates, Inc., 2021 (cit. on p. 17).
- [Schwenger, 2006] Vedat Schwenger, Ulrich Paul Hinkel, Anna Maria Nahm, Christian Morath, and Martin Zeier. "Real-time contrast-enhanced sonography in renal transplant recipients". In: *Clinical Transplantation* 20.SUPPL. 17 (Dec. 2006), pp. 51–54 (cit. on p. 31).
- [Selvaraju, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *ICCV*. IEEE, Inc., Oct. 2017 (cit. on pp. 20, 77).
- [Serón, 2007] D. Serón and F. Moreso. "Protocol biopsies in renal transplantation: Prognostic value of structural monitoring". In: *Kidney International* 72.6 (Sept. 2007), pp. 690–697 (cit. on p. 29).
- [Shabir, 2014] Shazia Shabir, Jean Michel Halimi, Aravind Cherukuri, Simon Ball, Charles Ferro, et al. "Predicting 5-Year Risk of Kidney Transplant Failure: A Prediction Instrument Using Data Available at 1 Year Posttransplantation". In: *American Journal of Kidney Diseases* 63.4 (Apr. 2014), pp. 643–651 (cit. on p. 43).
- [Sharfuddin, 2014] Asif Sharfuddin. "Renal relevant radiology: Imaging in kidney transplantation". In: *Clinical Journal of the American Society of Nephrology* 9.2 (2014), pp. 416–429 (cit. on pp. 23, 66, 89).



- [Shehata, 2020] Mohamed Shehata, Mohammed Ghazal, Hadil Abu Khalifeh, Ashraf Khalil, Ahmed Shalaby, Amy C. Dwyer, Ashraf M. Bakr, Robert Keynton, and Ayman El-Baz. "A Deep Learning-Based Cad System for Renal Allograft Assessment: Diffusion, Bold, and Clinical Biomarkers". In: *ICIP*. Vol. 2020-October. IEEE, Inc., 2020, pp. 355–359 (cit. on pp. 57, 66, 81).
- [Shehata, 2016] Mohamed Shehata, Fahmi Khalifa, Elizabeth Hollis, Ahmed Soliman, Ehsan Hosseini-Asl, Mohamed Abou El-Ghar, Maryam El-Baz, Amy C. Dwyer, Ayman El-Baz, and Robert Keynton. "A new non-invasive approach for early classification of renal rejection types using diffusion-weighted MRI". In: *ICIP*. Vol. 2016-August. IEEE, Inc., 2016, pp. 136–140 (cit. on p. 57).
- [Shehata, 2018] Mohamed Shehata, Ali Mahmoud, Ahmed Soliman, Fahmi Khalifa, Mohammed Ghazal, Mohamed Abou El-Ghar, Mousen El-Melegy, and Ayman El-Baz. "3D kidney segmentation from abdominal diffusion MRI using an appearance-guided deformable boundary". In: *PLOS ONE* 13.7 (July 2018), e0200082 (cit. on p. 48).
- [Shen, 2023] Chenyu Shen, Ziyuan Yang, and Yi Zhang. "PET Image Denoising with Score-Based Diffusion Probabilistic Models". In: *MICCAI*. Springer, 2023, pp. 270–278 (cit. on p. 113).
- [Shorten, 2019] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6.1 (Dec. 2019), pp. 1–48 (cit. on p. 13).
- [Simard, 2003] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. "Best practices for convolutional neural networks applied to visual document analysis". In: *ICDAR*. Vol. 2003-January. IEEE, Inc., 2003, pp. 958–963 (cit. on p. 13).
- [Simonyan, 2015] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR*. Curran Associates, Inc., 2015 (cit. on p. 12).
- [Singhal, 2023a] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, et al. "Large language models encode clinical knowledge". In: *Nature* (July 2023), pp. 1–9 (cit. on pp. 22, 111).
- [Singhal, 2023b] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, et al. "Towards Expert-Level Medical Question Answering with Large Language Models". In: (May 2023) (cit. on p. 22).
- [Smaira, 2020] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. "A Short Note on the Kinetics-700-2020 Human Action Dataset". In: (Oct. 2020) (cit. on p. 73).
- [Sohl-Dickstein, 2015] Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, Surya Ganguli, and Sngani@stanford Edu. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *ICML*. JMLR, 2015, pp. 2256–2265 (cit. on pp. 10, 113).
- [Song, 2019] Yang Song and Stefano Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution". In: *NeurIPS*. Curran Associates, Inc., 2019 (cit. on pp. 10, 113).
- [Sowrirajan, 2020] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. "MoCo-CXR: MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models". In: *MIDL*. Oct. 2020 (cit. on p. 89).
- [Springenberg, 2015] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. "Striving for Simplicity: The All Convolutional Net". In: *ICLR*. Curran Associates, Inc., 2015 (cit. on p. 20).
- [Stanton, 2001] Jeffrey M. Stanton. "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors". In: *Journal of Statistics Education* 9.3 (Jan. 2001) (cit. on p. 7).
- [Su, 2021] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. "RoFormer: Enhanced Transformer with Rotary Position Embedding". In: (Apr. 2021) (cit. on p. 111).
- [Sun, 2018] Roger Sun, Elaine Johanna Limkin, Maria Vakalopoulou, Laurent Dercle, Stéphane Champiat, et al. "A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study". In: *The Lancet Oncology* 19.9 (Sept. 2018), pp. 1180–1191 (cit. on p. 24).
- [Suthanthiran, 1994] Manikkam Suthanthiran and Terry B. Strom. "Renal Transplantation". In: *New England Journal of Medicine* 331.6 (Aug. 1994), pp. 365–376 (cit. on pp. 23, 65, 89).

- [Szegedy, 2014] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going Deeper with Convolutions”. In: *CVPR*. IEEE, Inc., 2014 (cit. on p. 12).
- [Taleb, 2020] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. “3D Self-Supervised Methods for Medical Imaging”. In: *NeurIPS*. Curran Associates, Inc., 2020 (cit. on pp. 65, 89).
- [Tan, 2019] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *ICML*. JMLR, 2019 (cit. on p. 12).
- [Thoppilan, 2022] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, et al. “LaMDA: Language Models for Dialog Applications”. In: (Jan. 2022) (cit. on p. 16).
- [Tibshirani, 1996] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society* 58.1 (1996), pp. 267–288 (cit. on p. 6).
- [Touvron, 2020] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. “Training data-efficient image transformers & distillation through attention”. In: *ICML*. JMLR, 2020 (cit. on p. 12).
- [Touvron, 2023a] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, et al. “LLaMA: Open and Efficient Foundation Language Models”. In: (Feb. 2023) (cit. on p. 16).
- [Touvron, 2023b] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. “Llama 2: Open Foundation and Fine-Tuned Chat Models”. In: (July 2023) (cit. on p. 16).
- [Truchot, 2023] Agathe Truchot, Marc Raynaud, Nassim Kamar, Maarten Naesens, Christophe Legendre, et al. “Machine learning does not outperform traditional statistical modelling for kidney allograft failure prediction”. In: *Kidney International* 103.5 (May 2023), pp. 936–948 (cit. on pp. 43, 82).
- [Tu, 2023] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, et al. “Towards Generalist Biomedical AI”. In: (July 2023) (cit. on p. 22).
- [Tunstall, 2022] Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural Language Processing with Transformers*. 1. O’Reilly Media, 2022 (cit. on p. 15).
- [Van Griethuysen, 2017] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer Research* 77.21 (Nov. 2017), e104–e107 (cit. on pp. 56, 58, 73, 77).
- [Van Waerebeke, 2022] Martin Van Waerebeke, Gregory Lodygensky, and Jose Dolz. “On the pitfalls of entropy-based uncertainty for multi-class semi-supervised segmentation”. In: *MICCAI Workshop UNSURE*. Springer, 2022 (cit. on p. 20).
- [Varoquaux, 2023] Gael Varoquaux and Olivier Colliot. “Evaluating Machine Learning Models and Their Diagnostic Value”. In: *Machine Learning for Brain Disorders*. Humana, New York, NY, 2023, pp. 601–630 (cit. on p. 8).
- [Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need”. In: *NeurIPS*. Curran Associates, Inc., 2017 (cit. on pp. 12, 15, 16, 65, 69, 100).
- [Wang, 2015a] C.J. Wang, J.B. Wetmore, G.S. Crary, and B.L. Kasiske. “The Donor Kidney Biopsy and Its Implications in Predicting Graft Outcomes: A Systematic Review”. In: *American Journal of Transplantation* 15.7 (July 2015), pp. 1903–1914 (cit. on p. 27).
- [Wang, 2023a] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, et al. “Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks”. In: *CVPR*. IEEE, Inc., 2023 (cit. on p. 18).
- [Wang, 2015b] Xiaolong Wang and Abhinav Gupta. “Unsupervised Learning of Visual Representations using Videos”. In: *ICCV*. IEEE, Inc., May 2015 (cit. on p. 14).

- [Wang, 2023b] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. "Foundation Model for Endoscopy Video Analysis via Large-Scale Self-supervised Pre-train". In: *MICCAI*. Springer, 2023, pp. 101–111 (cit. on p. 112).
- [Wei, 2022a] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, et al. "Emergent Abilities of Large Language Models". In: (June 2022) (cit. on p. 16).
- [Wei, 2022b] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *NeurIPS*. Curran Associates, Inc., 2022 (cit. on p. 111).
- [Weitz, 2023] Philippe Weitz, Masi Valkonen, Leslie Solorzano, Circe Carr, Kimmo Kartasalo, et al. "The ACROBAT 2022 Challenge: Automatic Registration Of Breast Cancer Tissue". In: (May 2023) (cit. on p. 21).
- [Will, 2014] Susanne Will, Petros Martirosian, Christian Würslin, and Fritz Schick. "Automated segmentation and volumetric analysis of renal cortex, medulla, and pelvis based on non-contrast-enhanced T1- and T2-weighted MR images". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 27.5 (Oct. 2014), pp. 445–454 (cit. on p. 54).
- [Wu, 2023] Junde Wu, Yu Zhang, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, and Yueming Jin. "Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation". In: (Apr. 2023) (cit. on p. 111).
- [Wu, 2020] Richard Wu, Aoqian Zhang, Ihab F Ilyas, and Theodoros Rekatsinas. "Attention-based Learning for Missing Data Imputation in HoloClean". In: *MLSys*. 2020 (cit. on p. 66).
- [Xia, 2021] Yan Xia, Le Zhang, Nishant Ravikumar, Rahman Attar, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, and Alejandro F. Frangi. "Recovering from missing data in population imaging – Cardiac MR image imputation via conditional generative adversarial nets". In: *Medical Image Analysis* 67 (Jan. 2021), p. 101812 (cit. on p. 66).
- [Xie, 2016] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks". In: *CVPR*. IEEE, Inc., 2016 (cit. on p. 12).
- [Xie, 2021a] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. "UniMiSS: Universal Medical Self-Supervised Learning via Breaking Dimensionality Barrier". In: *ECCV*. Springer Nature, 2021 (cit. on p. 20).
- [Xie, 2021b] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. "SimMIM: A Simple Framework for Masked Image Modeling". In: *CVPR*. IEEE, Inc., 2021 (cit. on p. 15).
- [Yaeger, 1996] Larry Yaeger, Richard Lyon, and Brandyn Webb. "Effective Training of a Neural Network Character Classifier for Word Recognition". In: *NeurIPS*. Curran Associates, Inc., 1996 (cit. on p. 13).
- [Yang, 2016] Xin Yang, Hung Le Minh, Kwang-Ting (Tim) Cheng, Kyung Hyun Sung, and Wenyu Liu. "Renal compartment segmentation in DCE-MRI images". In: *Medical Image Analysis* 32 (Aug. 2016), pp. 269–280 (cit. on pp. 25, 47).
- [Yang, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *NeurIPS*. Curran Associates, Inc., 2019 (cit. on p. 15).
- [Yi, 2022] Zhengzi Yi, Fadi Salem, Madhav C. Menon, Karen Keung, Caixia Xi, et al. "Deep learning identified pathological abnormalities predictive of graft loss in kidney transplant biopsies". In: *Kidney International* 101.2 (Feb. 2022), pp. 288–298 (cit. on p. 82).
- [Yilmaz, 2003] Serdar Yilmaz, Steven Tomlanovich, Timothy Mathew, Eero Taskinen, Timo Paavonen, Mercı Navarro, Eleanor Ramos, Leon Hooftman, and Pekka Häyry. "Protocol core needle biopsy and histologic Chronic Allograft Damage Index (CADI) as surrogate end point for long-term graft survival in multicenter studies". In: *Journal of the American Society of Nephrology* 14.3 (Mar. 2003), pp. 773–779 (cit. on p. 43).
- [Yoo, 2023] Daniel Yoo, Valentin Goutaudier, Gillian Divard, Juliette Gueguen, Brad C. Astor, et al. "An automated histological classification system for precision diagnostics of kidney allografts". In: *Nature Medicine* 29.5 (May 2023), pp. 1211–1220 (cit. on pp. 25, 82).

- [Yoon, 2023] Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. "SADM: Sequence-Aware Diffusion Model for Longitudinal Medical Image Generation". In: *IPMI*. Springer, 2023 (cit. on p. 113).
- [Yoon, 2018] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "GAIN: Missing Data Imputation using Generative Adversarial Nets". In: *ICML*. JMLR, 2018 (cit. on p. 66).
- [Yu, 2013] Chun Yan Yu, Wei Shi Zhang, Ying Ying Yu, and Ying Li. "A novel active contour model for image segmentation using distance regularization term". In: *Computers & Mathematics with Applications* 65.11 (July 2013), pp. 1746–1759 (cit. on p. 47).
- [Yu, 2012] Chun-yan Yu and Ying Li. "A Watershed Method for MR Renography Segmentation". In: *ICBEB*. IEEE, Inc., 2012, pp. 700–703 (cit. on pp. 47, 54).
- [Zbontar, 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *ICML*. JMLR, 2021 (cit. on p. 15).
- [Zhai, 2022] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. "Scaling Vision Transformers". In: *CVPR*. IEEE, Inc., 2022 (cit. on p. 12).
- [Zhang, 2016] Richard Zhang, Phillip Isola, and Alexei A. Efros. "Colorful Image Colorization". In: *ECCV*. Springer Nature, Mar. 2016 (cit. on p. 14).
- [Zhang, 2020] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: (Oct. 2020) (cit. on pp. 19, 21, 90, 100).
- [Zhao, 2021] Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M. Pohl. "Longitudinal self-supervised learning". In: *Medical Image Analysis* 71 (July 2021), p. 102051 (cit. on p. 33).
- [Zheng, 2018] Alice Zheng and Amanda Casari. *Feature Engineering in Machine Learning*. O'Reilly Media, 2018 (cit. on p. 5).
- [Zhou, 2023] Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, et al. "A foundation model for generalizable disease detection from retinal images". In: *Nature* 622.7981 (Sept. 2023), pp. 156–163 (cit. on p. 22).
- [Zhu, 2022] Lili Zhu, Renjun Huang, Ming Li, Qingmin Fan, Xiaojun Zhao, Xiaofeng Wu, and Fenglin Dong. "Machine Learning-Based Ultrasound Radiomics for Evaluating the Function of Transplanted Kidneys". In: *Ultrasound in Medicine & Biology* 48.8 (Aug. 2022), pp. 1441–1452 (cit. on p. 57).
- [Zou, 2018] James Zou and Londa Schiebinger. "AI can be sexist and racist — it's time to make it fair". In: *Nature* 559.7714 (July 2018), pp. 324–326 (cit. on pp. 20, 113).

**Titre:** Apprentissage de représentations de données multimodales et longitudinales pour le suivi de la transplantation rénale

**Mots clés:** apprentissage profond, apprentissage auto-supervisé, apprentissage multimodal, données longitudinales, imagerie médicale, transplantation rénale.

**Résumé vulgarisé:**

La transplantation rénale est une procédure vitale pour les patients atteints d'insuffisance rénale en phase terminale, mais son succès à long terme repose sur le suivi continu du rein transplanté. Cette thèse explore l'intégration d'algorithmes d'intelligence artificielle (IA) dans ce suivi, en particulier dans l'analyse des données d'imagerie médicale, afin d'améliorer l'efficacité et la précision du diagnostic des complications, des phénomènes de rejet et de l'évaluation de la fonction du greffon. L'étude englobe le développement de nouvelles méthodologies d'apprentissage profond et d'apprentissage auto-supervisé. En combinant des techniques d'IA avancées avec l'imagerie médicale, cette recherche contribue à la détection précoce des problèmes post-transplantations, ainsi qu'à l'avancement du domaine de la transplantation rénale. Nos conclusions soutiennent la pertinence de l'intégration de l'IA dans la pratique clinique, au bénéfice des patients et des experts médicaux.

**Title:** Representation Learning of Multimodal and Longitudinal data for Renal Transplantation monitoring

**Keywords:** deep learning, self-supervised learning, multimodal learning, longitudinal data, medical imaging, renal transplantation.

**Popularized abstract:**

Renal transplantation is a life-saving procedure for patients with end-stage renal disease, but its long-term success relies on continuous monitoring of the transplanted kidney's health. This thesis explores the integration of artificial intelligence (AI) algorithms into renal transplantation monitoring, particularly in the analysis of medical imaging data. Leveraging the power of AI, this thesis aims to enhance the efficiency and accuracy of diagnosing complications, rejection, and graft function. The study encompasses the development of novel deep learning methodologies to analyze medical images, focusing on MRI scans. By fusing advanced AI techniques with medical imaging, this research contributes to the early detection of issues, improving patient outcomes, and ultimately advancing the field of renal transplantation. Our findings support the relevance of integrating AI and medical imaging into clinical practice to benefit patients and medical experts.