

Sujet de thèse- PhD thesis

Détection non-supervisée d'anomalies dans des études longitudinales et multi-modales d'imagerie médicale

Unité de recherche : Laboratoire CREATIS

Directeur de thèse : M. DUCHATEAU Nicolas (MCU UCBL, HDR)
nicolas.duchateau@creatis.insa-lyon.fr

Co-directeur/trice de thèse : Mme. LARTIZIEN Carole (DR CNRS)
carole.lartizien@creatis.insa-lyon.fr

Domaine et contexte scientifiques

La détection d'anomalies en imagerie médicale est une tâche complexe, en particulier dans le contexte d'anomalies non annotées et difficiles à détecter même pour des experts. En neuroimagerie, c'est notamment le cas des anomalies liées aux maladies neurodégénératives (principalement les maladies de Parkinson et d'Alzheimer), qui peuvent concerner de petites régions difficilement visibles sur les images d'IRM du cerveau. En imagerie cardiaque, c'est également le cas des changements subtils au niveau de biomarqueurs de la fonction cardiaque extraits des images (principalement la forme et la déformation du myocarde, et les flux à travers les valves), également difficiles à détecter même pour des experts.

L'analyse automatique de ces données par apprentissage supervisé est limitée par la difficulté à constituer de grandes bases de données représentatives de la variabilité des motifs pathologiques. Les méthodes d'**apprentissage non-supervisé** offrent une alternative idéale, via une formulation visant à identifier les différences par rapport à un modèle de référence appris pour des données normales [LAG-24,CAI-24].

Les données étudiées étant complexes et de haute dimension, il est particulièrement intéressant de réaliser ce processus via une représentation simplifiée caractéristique des sujets sains (un espace latent de faible dimension pour la population normative), estimé par apprentissage de représentation [BEN-13]. Les anomalies sont alors détectées en mesurant l'écart d'un sujet pathologique à la distribution normative dans cet espace latent. Malgré de bonnes performances dans différentes applications cliniques, cette approche reste insuffisante pour la détection de lésions "subtiles", par exemple dans les phases précoces de développement de maladies.

Objectifs de la thèse

Ce projet vise à développer des algorithmes de **détection d'anomalies performants, fiables, et généralisables** à différents domaines cliniques, ce que nous souhaitons démontrer à la fois en neuroimagerie et imagerie cardiaque.

Plus spécifiquement, nous ciblons à la fois l'intégration de **plusieurs observations à différents instants du suivi longitudinal** d'un patient (des mesures répétées au cours du suivi d'un patient dans le temps [HED-06]), et de

données de différents types (différentes modalités et/ou différents biomarqueurs extraits des images). Nous disposerons ainsi à la fois d'une information sur la progression de la pathologie via l'évolution des anomalies potentielles (nombre, taille, et localisation), ainsi que d'une plus grande robustesse vis-à-vis des détections incertaines ou non visibles à un stade précoce.

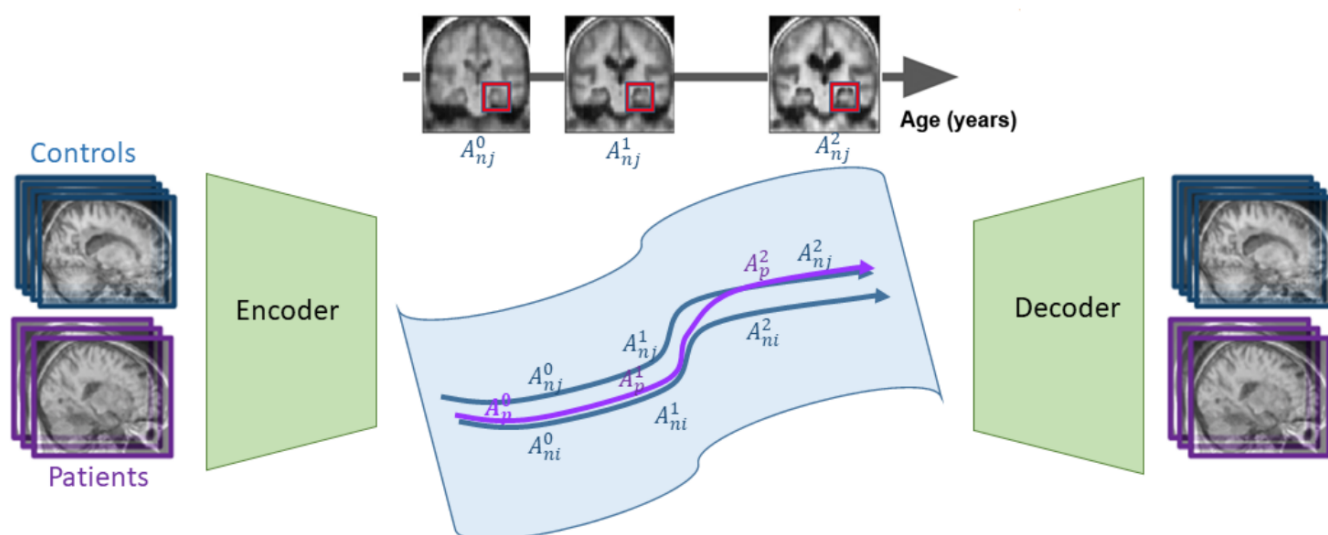


Fig.1 : Vue d'ensemble de l'approche visée dans ce projet, qui consiste à modéliser l'évolution (la trajectoire) des sujets étudiés, via une représentation simplifiée (espace latent de faible dimension qui encode les données de haute dimension comme les images), puis de l'exploiter pour détecter les anomalies de l'évolution d'un patient par rapport à l'évolution d'un groupe de sujets contrôles.

Verrous scientifiques

La plupart des modèles de détection d'anomalies n'exploitent pas l'information contenue dans les données d'imagerie d'un patient acquises soit lors d'examens dynamiques (séries temporelles) soit à différents points temporels au cours du suivi (série longitudinale). **Modéliser l'évolution de données longitudinales** pose plusieurs défis. D'abord, sur la conception de modèles pertinents pour considérer l'ensemble des données et des caractéristiques de la pathologie pour répondre à la question clinique. Cela nécessite la prise en compte du lien entre les données acquises à différents instants du suivi, qui plus est avec des données parfois manquantes [COU-19].

Par ailleurs, analyser une pathologie à partir d'un seul point de vue (un seul type de descripteur issu d'imagerie) est très limitant en termes de performance et de robustesse. En clinique, plusieurs acquisitions et mesures sont systématiquement réalisées, et complétées par des données autres que l'imagerie. L'enjeu pour les méthodes computationnelles est d'arriver à **intégrer dans l'analyse plusieurs descripteurs complexes et de types hétérogènes**, qui ne peuvent pas être concaténés ou fusionnés par des méthodes standards.

L'apprentissage de représentation offre un cadre approprié pour adresser ces deux verrous méthodologiques, et les exploiter pour la détection d'anomalies. Néanmoins, il est encore peu exploré vis-à-vis de la dépendance ou complémentarité potentielles entre les différents types de données et les différents instants et échelles auxquelles elles sont étudiées, ce que nous visons directement dans ce projet.

Un autre verrou scientifique du projet concerne l'architecture du modèle de détection d'anomalies et son couplage avec la représentation apprise. Nous nous appuyons sur les travaux réalisés dans l'équipe ainsi que sur une

analyse approfondie de l'état-de-l'art récent dans ce domaine, afin de sélectionner les approches les plus pertinentes (modèles discriminants dans l'espace latent, modèles génératifs de diffusion ou Bayésien) et d'étudier la meilleure stratégie de couplage à la représentation apprise (e.g. entraînement joint ou découplé).

Programme de recherche et démarche scientifique proposés

La thèse comportera une partie plus **méthodologique**, qui visera à d'abord généraliser la détection d'anomalies dans le cadre de données longitudinales, puis multi-modales. La détection d'anomalies s'appuiera sur les approches mono-instant / mono-modales développées au laboratoire [ALA-20,PIN-23]. Elle sera consolidée par l'état-de-l'art en apprentissage de représentation : auto-encodeurs variationnels avec des contraintes pour la cohérence de l'espace latent (prenant en compte des variables confondantes). Nous explorerons également d'autres pistes basées sur des modèles de diffusion pour une reconstruction réaliste depuis l'espace latent (facilitant les interprétations cliniques). L'analyse de l'évolution longitudinale d'un sujet nécessitera de modéliser des trajectoires de ces données dans l'espace latent [SAU-22,PUG-24], robuste en présence de données manquantes et échantillonnées à des points temporels variables. Nous nous appuierons sur des travaux préliminaires prometteurs basés sur des modèles à effets mixte, qu'il conviendra d'adapter à la tâche de détection d'anomalie. L'analyse multi-modale se placera dans le cadre des techniques de fusion, avec une représentation latente mélangeant l'apport de plusieurs descripteurs, qui devra prendre en compte la complémentarité des données, notamment de façon progressive, un aspect récemment étudié par les encadrants [BER-22,FRE-24].

La thèse comportera également une partie plus **applicative**, qui ciblera principalement les cohortes cliniques existantes suivantes :

- En **imagerie cardiaque**, la cohorte ADOL [FAB-20] est constituée de 56 sujets avec retard de croissance intra-utérin, avec suivi longitudinal dans un protocole d'effort avec acquisition toutes les 2 minutes. Les données seront des descripteurs de la fonction cardiaque extraits d'images d'échocardiographie 2D (B-mode et Doppler), caractérisant la forme, le mouvement, et la déformation locale du myocarde et les flux des valves, obtenus avec les outils de recherche et commerciaux déjà disponibles au sein du laboratoire et des hôpitaux collaborateurs.

- En **neuroimagerie**, nous proposons de travailler sur différentes bases de données publiques:

- La base de données nord-américaine ADNI (<http://adni.loni.usc.edu>) agrège les données cliniques, biologiques d'imagerie multimodales (TEP/IRM, CT) de plus de 2500 participants suivis au cours du temps. Pour cette étude, nous pourrions nous appuyer sur deux cohortes. La première sera constituée de sujets ne présentant pas de troubles cognitifs pendant toute la durée de l'étude et donc considérés comme 'cognitively normal' (CN). Cette cohorte (>300 sujets) permettra de modéliser le processus de vieillissement normal du cerveau. La seconde cohorte (> 300 sujets) sera constituée de sujets diagnostiqués avec la maladie d'Alzheimer à un point de mesure au moins (qui peut ne pas être le premier point de suivi longitudinal). Le nombre d'échantillons temporels pour ces données est de l'ordre de 3 à 7 par sujet de la base.

- La base de données américaine PPMI (<https://www.ppmi-info.org/>) constituée de patients atteints de la maladie de Parkinson. Cette base, comme ADNI inclut des examens longitudinaux regroupant de l'imagerie et des données cliniques de sujets sains et patients à différents stades la maladie, dont certains initialement en phase précoce au cours de laquelle le diagnostic de la pathologie est très difficile à réaliser. Le début de la thèse inclura une exploration de cette cohorte pour déterminer le nombre et type de données pertinents.

- la base de données anglaise UK Biobank agrège les données de santé (incluant des données d'imagerie ainsi que des données cliniques et génétiques) de plusieurs centaines de milliers de sujets suivis au cours du temps

Selon l'avancement, la généralisation des développements réalisés durant la thèse sera évaluée sur d'autres cohortes cliniques. Pour l'application cardiaque, nous pourrions envisager de travailler sur une cohorte constituée d'échocardiographies 2D réalisées chez des sportifs dans le cadre d'un protocole d'effort, pour la détection d'anomalies pouvant favoriser les arythmies. Cette cohorte, acquise et en cours d'annotation à l'Hospital Clínic de Barcelone (Espagne), via M. Sanz de la Garza (cardiologue, spécialiste des athlètes) fait l'objet d'un accord de consortium en cours de rédaction avec CREATIS. Pour la neuroimagerie, nous prévoyons d'identifier des bases de données longitudinales pertinentes (détection précoce d'anomalies) au sein de notre réseau de collaborateurs neuroscientifiques et engagerons des démarches réglementaires pour pouvoir accéder à ces bases.

Encadrement et environnement scientifique

Le candidat ou la candidate intégrera l'équipe de recherche MYRIAD du laboratoire CREATIS, spécialisée dans le développement de méthode par apprentissage profond pour le traitement et l'analyse d'images médicales. Il ou elle sera co-encadré.e par Nicolas Duchateau, MCU UCBL, HDR et Carole Lartizien, DR CNRS, tous les deux spécialistes en méthode par apprentissage profond pour l'analyse de données multimodales cardiaques et cérébrales, respectivement.

Le doctorant bénéficiera également du support des projets R&D en cours pour ses besoins matériels (matériel informatique, accès au cluster du laboratoire et centres de calcul partenaires, plateformes de gestion des données du laboratoire et de ses partenaires) et pour ses missions, ainsi que de l'environnement scientifique stimulant de l'équipe MYRIAD et du laboratoire CREATIS.

Collaborations

Concernant les applications en cardiologie, ce projet s'appuiera sur une collaboration établie avec le **Fetal Medicine Research Center**, Barcelone (Espagne), porté par F. Crispi (obstétricienne et experte clinique renommée en cardiologie fœtale),

- Concernant les applications en neurologie, nous souhaitons initier ou renforcer des collaborations avec des centres de recherche en neurosciences, par exemple l'ICM (Paris), Neurospin (Saclay), ou le GIN (Grenoble), en complément des bases de données publiques (ADNI et PPMI) sur lesquelles seront initiées les recherches.

- Sur le versant méthodologique, nous envisagerons de collaborer ponctuellement avec des équipes de recherche travaillant sur la détection d'anomalies et/ou l'analyse de données longitudinales. Les encadrants de thèse s'appuieront alors sur leur réseau de collaborateurs en fonction de l'avancement et des besoins du projet.

Profil du candidat recherché:

Le candidat ou la candidate devra avoir une bonne connaissance théorique en apprentissage statistique ou en analyse d'images et de l'expérience dans les deux domaines. Un bon niveau en programmation est attendu. De l'expérience avec des bibliothèques de machine learning (Tensorflow, Pytorch ou Scikit learn) et les systèmes d'exploitation sous noyau Linux serait appréciée. Nous sommes à la recherche d'un candidat ou d'une candidate enthousiaste et autonome, motivé et intéressé par la recherche pluridisciplinaire.

Financement

Ce sujet de thèse est ouvert au concours de l'école doctorale EEA pour un financement de thèse ministériel. Le dépôt des dossiers de candidatures seront remontés par le laboratoire fin mai 2025 pour une audition début juin. Nicolas Duchateau et Carole Lartizien participeront activement à la préparation de l'audition avec le ou la candidate retenu.e.

Contact

Merci d'envoyer votre CV accompagné d'une lettre de motivation et des bulletins de notes les plus récents à nicolas.duchateau@creatis.insa-lyon.fr et carole.lartizien@creatis.insa-lyon.fr.

Références bibliographiques sur le sujet de thèse:

- [ALA-20] Alaverdyan Z, Jung J, Bouet R, Lartizien, C. Regularized Siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion

screening. *Med Image Anal.* 2020;60:101618.

- [BEN-13] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* 2013;35:1798-828.
- [BER-22] Bernardino G, Jonsson A, Loncaric F, Martí Castellote PM, Sitges M, Clarysse P, Duchateau N. Reinforcement learning for active modality selection during diagnosis. *Proc. MICCAI, LNCS.* 2022;13431:592-601.
- [CAI-24] Cai Y, Zhang W, Chen H, Cheng KT. MediAnomaly: A comparative study of anomaly detection in medical images. *Arxiv* 2024.
- [COU-19] Couronne R, Vidailhet M, Corvol JC, Lehericy S, Durrleman S. Learning disease progression models with longitudinal data and missing values. *Proc. ISBI.* 2019:1033-7.
- [FAB:20] Fabijanovic D, Loncaric F, Sarvari S, Velve K, Rodriguez-Lopez M, Sepulveda-Martinez A, Blanco I, Cikes M, Sitges M, Gratacos E, Bijmens B, Crispi F. Cardiac remodeling in a fetal growth restriction cohort – a follow-up study from preadolescence into adolescence. *Eur Heart J.* 2020;41:ehaa946.3205.
- [FRE-25] Freiche B, Bernardino G, Deleat-Besson R, Clarysse P, Duchateau N. Hierarchical data integration with Gaussian processes: application to the characterization of cardiac ischemia-reperfusion patterns. *IEEE Trans Med Imaging.* 2025. In press.
- [HED-06] Hedeker D & Gibbons RD. *Longitudinal data analysis.* John Wiley & Sons, Inc, New Jersey. 2006.
- [LAG-24] Lagogiannis I, Meissen F, Kaissis G, Rueckert D. Unsupervised pathology detection: a deep dive into the state of the art. *IEEE Trans Med Imaging.* 2024;43:241-52.
- [PIN-23] Pinon N, Oudoumanessah G, Trombetta R, Dojat M, Forbes F, Lartizien C. Brain subtle anomaly detection based on auto-encoders latent space analysis : Application to de novo Parkinson patients. *Proc. ISBI.* 2023:1-5.
- [PUG-24] Puglisi L, Alexander DC, Ravi D. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. *Proc. MICCAI, LNCS.* 2024;15002:173-83.
- [SAU-22] Sauty B, Durrleman S. Progression models for imaging data with longitudinal variational auto encoders. *Proc. MICCAI, LNCS.* 2022;13431:3-13.