Analyse des caractéristiques électriques pour la détection des sujets à risque de mort subite cardiaque

Mariette Dupuy

université de BORDEAUX

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEURE
DE L'UNIVERSITÉ DE BORDEAUX

ECOLE DOCTORALE MATHEMATIQUES ET INFORMATIQUE

Mathématiques appliquées

Par **Mariette DUPUY**

Analyse des caractéristiques électriques pour la détection des sujets à risque de mort subite cardiaque

Sous la direction de : **Marie CHAVENT** et **Rémi DUBOIS**

Soutenue le 14 Janvier 2025

Membres du jury :

Mme. Marie CHAVENT	Professeure	Université de Bordeaux (Talence)	Directrice de thèse
M. Rémi DUBOIS	Professeur	Université de Bordeaux (Pessac)	Co-Directeur
Mme. Florence D'ALCHE-BUC	Professeure	Institut Polytechnique de Paris (Palaiseau)	Présidente du jury
M. Benoit LIQUET	Professeur	Université de Pau et des Pays de l'Adour (PAU)	Rapporteur
M. Julien OSTER	Directeur de Recherche	Inserm (Vandoeuvre les Nancy)	Rapporteur
M. Robin GENUER	Associate professor	ISPED, Université de Bordeaux (Bordeaux)	Examinateur

Membre invité :

M. Michel HAISSAGUERRE	Professeur Universitaire - Praticien Hospitalier	Université de Bordeaux	Invité

Analyse des caractéristiques électriques pour la détection des sujets à risque de mort subite cardiaque

Résumé : La mort subite cardiaque (MSC) représente 30% de la mortalité adulte des pays industrialisés. La majeure partie des MSC est la conséquence d'une arythmie appelée fibrillation ventriculaire, elle-même étant la conséquence d'un muscle cardiaque présentant des anomalies structurelles. Malgré l'existence de thérapies efficaces, la majorité des individus présentant un risque de MSC ne sont pas identifiés de manière préventive à cause de l'absence d'examen disponible. Le développement de marqueurs spécifiques sur des enregistrements électrocardiographiques permettrait une identification et une stratification du risque de MSC. Au cours des six dernières années, l'IHU Liryc a enregistré les signaux électriques à la surface du corps chez plus de 800 individus (sains et pathologiques) à l'aide d'un dispositif haute résolution de 128 électrodes. Des caractéristiques ont été calculées sur ces signaux (durée du signal par électrode, fréquence, fractionnement de l'amplitude, ...). Au total, plus de 1500 caractéristiques électriques sont disponibles par patient. Lors du processus d'acquisition par le système de 128 électrodes en milieu hospitalier, le bruit, ou le mauvais positionnement de certaines électrodes ne permettent pas de calculer les caractéristiques prévues, conduisant ainsi à une base de données incomplète. Cette thèse s'organise autour de deux axes. Nous avons dans un premier temps développé une méthode d'imputation de données manquantes pour répondre au problème des électrodes défaillantes. Puis nous avons développé un score de risque pour la stratification du risque de mort subite. La famille de méthodes la plus souvent utilisée pour gérer les données manquantes est l'imputation : allant d'une simple complétion par la moyenne, à des méthodes par agrégation locale, régressions locales, transport optimal ou encore modification de modèles génératifs. Récemment les Autoencoders (AE) et plus précisément les Denoising AutoEncoder (DAE) ont montré de bonnes performances pour cette tâche. Les AE sont des réseaux de neurones utilisés pour apprendre une représentation des données dans un espace de dimension réduit. Les DAE sont des AE qui ont été proposés pour reconstruire, à partir de données bruitées, les données originales. Nous proposons dans ce travail une nouvelle méthodologie basée sur les DAE appelée modified Denoising AutoEncoder (mDAE) pour permettre l'imputation de données manquantes. Le deuxième axe de recherche de la thèse a consisté à la mise en place d'un score du risque de mort subite cardiaque. Les DAE ont la capacité de modéliser et de reconstruire des données complexes. Nous avons ainsi entrainé des DAE à modéliser la distribution des individus sains sur un sous-groupe sélectionné de caractéristiques électriques. Puis nous avons utilisé ces DAE pour discriminer des patients pathologiques des individus sains en analysant la qualité d'imputation du DAE sur des caractéristiques partiellement masquées. Dans le but de mettre en place un score de risque de la mort subite, nous avons également comparé différentes méthodes de classification.

Mots-clés : Imputation de données manquantes, analyse statistique multiple, apprentissage machine, apprentissage profond, bio-statistiques, bio-informatique

---

Analysis of electrical features for detection of subjects at risk for sudden cardiac death.

Abstract: Sudden cardiac death (SCD) accounts for 30% of adult mortality in industrialized countries. The majority of SCD cases are the result of an arrhythmia called ventricular fibrillation, which itself results from structural abnormalities in the heart muscle. Despite the existence of effective therapies, most individuals at risk for SCD are not identified preventively due to the lack of available testing. Developing specific markers on electrocardiographic recordings would enable the identification and stratification of SCD risk. Over the past six years, the Liryc Institute has recorded surface electrical signals from over 800 individuals (both healthy and pathological) using a high-resolution 128-electrode device. Features were calculated from these signals (signal duration per electrode, frequency, amplitude fractionation, etc.). In total, more than 1,500 electrical features are available per patient. During the acquisition process using the 128-electrode system in a hospital setting, noise or poor positioning of specific electrodes sometimes prevents calculating the intended features, leading to an incomplete database. This thesis is organized around two main axes. First, we developed a method for imputing missing data to address the problem of faulty electrodes. Then, we developed a risk score for the sudden death risk stratification. The most commonly used family of methods for handling missing data is imputation, ranging from simple completion by averaging to local aggregation methods, local regressions, optimal transport, or even modifications of generative models. Recently, Autoencoders (AE) and, more specifically, Denoising AutoEncoders (DAE) have performed well in this task. AEs are neural networks used to learn a representation of data in a reduced-dimensional space. DAEs are AEs that have been proposed to reconstruct original data from noisy data. In this work, we propose a new methodology based on DAEs called the modified Denoising AutoEncoder (mDAE) to allow for the imputation of missing data. The second research axis of the thesis focused on developing a risk score for sudden cardiac death. DAEs can model and reconstruct complex data. We trained DAEs to model the distribution of healthy individuals based on a selected subset of electrical features. Then, we used these DAEs to discriminate pathological patients from healthy individuals by analyzing the imputation quality of the DAE on partially masked features. We also compared different classification methods to establish a risk score for sudden death.

Keywords: Missing data imputation, multiple statistical analysis, machine learning, deep learning, biostatistics, bioinformatics

---

Unité de recherche
UMR xxxx Université, 33000 Bordeaux, France.

A toutes les femmes du monde,

Remerciements

Pendant ces trois années de thèse j'ai été aidée, accompagnée, soutenue, motivée par un grand nombre d'êtres humains que je tiens à remercier.

Je remercie Benoit Liquet et Julien Oster d'avoir accepté de rapporter ma thèse, merci pour votre regard et expertise sur mon travail et merci pour vos retours sur mon manuscrit. Je tiens également à remercier Robin Genuer et Florence d'Alché pour leur présence dans mon jury.

Merci Marie d'avoir été ma directrice durant ces trois années de thèse. Tu m'as transmis ton goût des choses bien faites, de la science précise et rigoureuse. Merci pour nos discussions, le temps passé pour relire mes productions écrites, la confiance que tu m'as donnée quand j'en avais moins. Et aussi merci pour les rires, les blagues, les points mode, ongle et coiffure. Merci pour ta bienveillance.

Merci Rémi d'avoir été mon directeur de thèse. Merci de m'avoir accueillie dans ce lieu d'excellence qu'est le Liryc. Ton point de vue objectif sur les résultats, ta capacité à dégager un axe de recherche, à être sceptique, à te positionner dans le bon cadre scientifique, à mettre en avant nos résultats, à toujours se questionner pour aller plus loin, vérifier qu'on a tout tester m'ont beaucoup aidé dans ma recherche. Merci de toujours savoir comment rendre mes présentations orales meilleures. Et finalement merci de ne m'avoir jamais discriminé à cause de mon genre.

Je te remercie Laura pour l'aide que tu m'as apportée dans les derniers moments stressants de rédaction de manuscrit. Merci pour la bonne humeur que tu apportes à l'équipe.

Je souhaite vous remercier, Pr Haissaguerre pour ce que vous avez créé au Liryc, grâce au projet HELP j'ai pu réaliser cette thèse. J'ai toujours rêvé d'allier mathématiques, informatique et médecine et donc un grand merci de m'avoir permis de réaliser mon rêve. Merci de nous faire partager votre passion intense pour l'électrophysiologie. Merci également d'avoir fait partie de mon jury de thèse en tant que membre invité.

Merci à vous Pr Hocini d'être une telle figure féministe et savante. Merci de faire tout ce que vous faites au sein du Liryc que cela soit au mécénat ou dans votre vie de professeure reconnue pour faire bouger les mentalités. Merci de nous raconter vos débuts, de nous montrer qu'il ne faut pas avoir peur, qu'on a toutes notre place et de nous montrer qu'il est possible d'évoluer dans un milieu masculin.

Je tiens également à remercier ceux qui m'ont fait découvrir le monde de la recherche. Merci Marc pour ce stage d'observation au CNRS. Déjà passionnée, je voulais tout savoir et j'étais émerveillée par ce monde. Merci à toi Rodolphe de m'avoir donnée l'opportunité de faire un stage en L1. J'ai adoré l'équipe, le sujet, c'est là que j'ai vraiment réalisé ma passion pour le mélange de la médecine et des mathématiques. Merci Bernard de m'avoir acceptée au sein du CMI ISI, tu m'as fait rencontré mes plus proches amis et mon copain. Grandir et étudier au sein du CMI ISI a été une expérience formidable. Merci de nous avoir accom-

pagné au début quand tout était vraiment chaotique et merci d'avoir sauvé mon semestre de probabilité à Bristol. Je n'en serai peut-être pas là sinon. Enfin merci à Louis et Marine nos jeunes enseignants auxquels j'ai pu m'identifier en M2, qui m'ont donné la dernière motivation pour me lancer dans cette aventure folle de la thèse.

J'ai passé les trois-quarts de ma thèse au Liryc. Un laboratoire de recherche merveilleux remplis de gens merveilleux. Il est impossible de parler du Liryc sans parler de la gardienne du temple : Amanda. Merci d'être la star que tu es, merci d'être si drôle et attachante. Merci à l'équipe administrative qui est remplie de gens adorables. Merci en particulier à Anne-France, Magalie, Florine, Manon, Ambre. Une attention particulière à ma partenaire Sirène Laure, merci de me faire rire, d'être aussi folle, entière, sincère. Merci également aux équipes scientifiques toujours très bienveillantes. Merci à Nestor, Estelle et Bastien pour leur gentillesse et leur humour. Un merci particulier à l'équipe dont j'ai fait partie pendant ces trois années : l'équipe Signal. Merci à toi Nolwenn de m'avoir accueillie quand je suis arrivée, merci à Nicolas, Amael, Camille, Nicolas, Ayoub d'être de supers collègue. Une petite pensée pour notre Amael national, tu es d'un extrême gentillesse mais par contre tu es tout le temps malade c'est plus possible. Merci à Sébastien d'avoir ralenti ma rédaction de manuscrit. Malgré tout, nos discussions sur l'écologie, l'amour, l'amitié et une tonne de ragots font partie des meilleurs moments de ce dur labeur. Merci à mon inimitable partenaire de bureau Jean-Baptiste. Merci pour tes blagues, merci pour ton soutien, merci pour les BK, tes reds bulls. Merci de m'avoir supporté pendant 2 ans et de ne pas avoir voulu changer de bureau. Au Liryc j'ai également eu l'immense chance de côtoyer de véritable stars du monde de la recherche : Tom et Gabriel. Les polytechniciens sont descendus à Bordeaux, attention ! Merci à toi Gabriel m'avoir fait beaucoup rire et aussi de m'avoir proposé un projet scientifique absolument génial, mais sur lequel je n'ai pas pu te rejoindre. Merci à toi Tom pour toutes les conneries que tu dis, vive Bordeaux et Toulouse beurk. Plus jamais tu chantes dans mon bureau. Finalement, heureusement que tu n'as pas fait toute ta thèse dans mon bureau sinon tu n'aurais pas pu rédiger tes 18 papiers. Déjà 6 mois en coloc dans mon bureau et j'ai faillit te mettre en retard pour la fin de la rédaction de ton manuscrit. Merci à tous les deux, vous êtes des génies merci de m'avoir apporté un peu de connaissance scientifique (toujours en croisade pour éduquer les gueux ces polytechniciens).

Grâce au Liryc j'ai pu rencontrer Sirine, toi qui es maintenant mon amie. Quel bonheur d'avoir une amie avec qui partager ses joies et ses peines. Tu es un véritable rayon de soleil, rempli de gentillesse et de générosité. Ces années à tes côtés au Liryc étaient vraiment extra et ton départ m'a brisé le cœur donc revient vite à Bordeaux. Merci pour ton soutien, ton humour, ton amour des ragots, de la TV réalité, de la pop culture et de la mode.

Que serait ma thèse et ma vie au Liryc sans mentionner mon fidèle acolyte Georges ! Tout nos oppose (nos passe-temps, les décibels de notre voix, et j'en passe) mais on a créée une très belle amitié qui m'est très chère. Merci d'être mon ami. Merci pour ta confiance, merci pour tes cadeaux du Vietnam, merci de toujours me faire goûter ta nourriture que je n'aime pas toujours (oups), merci pour ta patience. Sacré trio qu'on a formé avec ces lunettes 3D. Merci de m'apprendre le calme. Merci d'être venu nous aider aux huîtres. Merci d'être toujours à l'écoute et présent. Et merci d'accepter mes blagues un peu louches...

Grâce à toi j'ai pu rencontrer une bien belle bande de joyeux lurons. "Les gars de la start-up" comme j'aime bien vous appeler. Yannick merci pour tes blagues d'un autre temps (oups), merci pour ton extrême gentillesse, tes conseils. Mikha merci aussi pour tes nombreux conseils, nos discussions sincères, merci de m'avoir redonné confiance sur pleins de

sujets. Merci Ali d'être si drôle, désolée de me moquer de ton accent, t'es le best. Merci pour nos discussions potins, merci à tous les deux pour l'aide que vous m'avez apporté, les conseils que vous m'avez donné, les rires. Boys supporting girls, c'est assez rare pour le faire remarquer.

Je passe maintenant à mes amis de longue date. La plus vieille d'entre toute Capucine. Plus de 15 ans d'amitié et nous voilà toutes les deux en thèse. Merci de faire toujours partie de mes amies les plus proches. Tout le chemin qu'on a parcouru ensemble et j'espère qu'il sera encore très long. Merci d'être mon amie depuis si longtemps et d'être toujours là pour moi.

Merci à mes boys Clément et Maxime. Mes deux meilleurs amis. Vraiment des énergumènes de première classe. Mais merci de me faire tant rire. Merci à vous de m'avoir changé les idées, merci d'être encore à mes côtés aujourd'hui, merci d'être si chiants, merci de me faire améliorer mon argumentaire féministe, merci pour les conversations très peu catholiques, merci de me faire tester ma patience depuis 8 ans. Merci d'être vous. Très hâte de vous voir darons et d'être la marraine de vos enfants.

Merci à toi mon Dorian, le pilier, t'as toujours été là quand j'en avais besoin. Toujours de bon conseils, merci de me tolérer alors que je pense que je te rends fou sur maints sujets. Merci d'être un ami fidèle sur qui je pourrai toujours compter, je suis fière d'être ton amie. Et merci de me faire sans rire parce que toi aussi t'es t'en dis pas mal des bêtises.

Marie-Mathilde, maintenant 9 ans qu'on est amies, 9 ans qui ont rendu notre amitié toujours plus forte et plus belle. Merci d'avoir toujours été à l'écoute sans jamais me juger, merci de me connaître par cœur. Merci de me comprendre et de m'aimer comme je suis. Tellement de passions qu'on partage ensemble, Bernard et le CMI mettraient en avant celle du bavardage. Est-ce qu'on a fait de la boxe pour parler ou pour boxer ? Tellement hâte de nous voir grandir et évoluer.

Gauthier le seul du master à ne pas avoir eu une haine des CMI. Et cela reflète vraiment ta belle personnalité, tu es un humain gentil, bienveillant, drôle aussi parfois je dois l'avouer. Tu t'améliores de plus en plus au niveau de tes tenues et je suis sûre que c'est grâce à mes conseils avisés. Merci de vouloir faire des collaborations avec moi et de croire en ma recherche grâce à toi j'ai presque l'impression d'avoir fait une véritable thèse. Je suis très fière que Pol ait un meilleur ami comme toi et de faire partie de tes amis également. Grâce à toi j'ai pu rencontrer Camille. Camille, tu m'as acceptée dans ton cercle fermé d'amis proche et pour quelqu'un qui n'a pas beaucoup confiance en autrui c'est une belle preuve d'amitié que tu m'as faite. Tu es pour moi un exemple de confiance et de force. Merci pour tes paroles motivantes, pour nos soirées au mada, pour tes anniversaires qui demandent beaucoup trop d'énergie. Merci de nous apprendre des choses à chaque fois qu'on te voit, merci pour les débats qui durent des heures, merci pour ton combat féministe.

Flavie, ma petite caille, la plus jeune de la troupe mais tellement de maturité. Mes dernières années de thèse n'auraient pas été les mêmes sans toi. Merci de m'avoir apporté la lumière dont j'avais besoin, merci de m'avoir soutenue, écoutée. Merci pour ta présence, ta générosité. Merci de faire partie de mes amies les plus proches.

Je tiens maintenant à remercier ma belle-famille. Merci Marie de m'avoir accueilli chez toi, de m'avoir reçu les bras ouverts avec toute la générosité qui te caractérise. Merci de m'avoir fait découvrir le Périgord, merci pour les we de repos, de bonne nourriture et de rire. Merci à Patrick et Sylvie de m'avoir également accueilli sans jamais me juger même quand je mettais un peu le bazar, merci d'avoir accepté mon coté rougnousse. Merci pour les

délicieux repas, les vacances au bord de la piscine, les accras du marché de Saint-Cyprien. Merci à Jean, Alice et leur petit bout Liv, merci pour votre calme, merci de m'avoir accepté dans votre famille et me sentir comme chez moi à chaque fois. Alice tu me fais mourir de rire, Jeannot à chaque fois qu'on se voit je fais tes recettes pendant des mois après. Merci à tous les deux. A toi cher Louis, je te remercie pour le rire que tu m'apportes. Impossible de rester énervée en ta présence et pourtant tu as quand même la capacité de m'énerver en un temps record. Tu es une belle personne, remplie de gentillesse. Merci pour les barbecues, merci pour les soirées, merci pour les we dans le périgord, les Noël et les anniversaires. Merci d'être toi.

Je passe maintenant aux personnes qui me connaissent depuis que je suis bébé : ma famille. Merci tonton, tatie et cousin de m'avoir vue grandir. Merci Bernard d'avoir toujours été là pour nous. Merci à mes mamies, des femmes exceptionnelles, des travailleuses, courageuses. Merci à vous de me donner tout ce que vous me donnez.

Merci Papa et Maman. Tout n'a pas toujours été tout rose, mais on grandit tous ensemble. Merci de nous avoir tout donné. Merci de nous avoir inculqué des belles valeurs. Merci de m'avoir soutenu dans ce choix compliqué et sinueux qu'est le doctorat. Merci pour votre aide, votre générosité. Merci pour votre temps, merci de nous avoir construit un havre de paix. Force et honneur.

Ma sœur, tout ce que je devrai dire sur toi prendrai trop de temps ici. Petite sœur mais remplie de maturité, de sagesse et d'intelligence. Merci de m'avoir montré plein de fois le droit chemin, de m'avoir écoutée, secourue, aidée, aimée. Merci d'être si drôle. T'inquiète je vais essayer maintenant d'arrêter les études et d'avoir un vrai métier.

Pol, il faudrait que je rédige une nouvelle thèse pour exprimer tous les sentiments que j'ai pour toi. En toute sincérité tu es celui qui a sauvé ma thèse. Il n'y a pas de mot assez pur et puissant pour exprimer l'immense gratitude que j'ai envers toi. Tu es l'être humain le plus gentil que je connaisse. Ta personnalité opposée à la mienne m'a apporté tellement de positif pendant ces premières années de notre couple. Je ne parlerai pas de tes défauts, car oui malgré tout tu en as quand même une bonne quantité. Tu sais exactement comment me donner confiance, comment me donner l'énergie et me faire aller de l'avent. Tu m'as interdit de baisser les bras mais tu m'as aussi laissé pleurer pour mieux repartir. Ta culture scientifique mêlée à ton désir de m'aider m'a maintes fois éclairée. Merci de m'aimer comme je suis, merci de m'aider à m'améliorer sans jamais me juger, merci d'accepter ma folie, merci d'accepter mon amour. Merci d'être le plus bel homme sur Terre. Merci pour tout, j'ai énormément hâte de continuer notre vie de docteurs ensemble.

Et finalement un grand merci à Romy (mimi), le plus beau chat du monde.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACP Analyse en Composantes Principales.
AE AutoEncoder.
AUC Area Under the Curve.

DAE Denoising AutoEncoder.
DAVD Dysplasie Arythmogène du Ventricule Droit.

ECG ElectroCardioGramme.
ECG 128 HD ECG 128 électrodes Haute Définition.
ECG 12D ElectroCardioGramme 12 dérivations.
ECG HD ECG à Haute Densité d'électrodes.

FA Forêts Aléatoires.
FV Fibrillation Ventriculaire.
FVI Fibrillation Ventriculaire Idiopathique.

GAN Generative Adversarial Network.

HELP Heterogeneous Electrical tissue Localization Program.

MAR Missing At Random.
MCAR Missing Completely At Random.
mDAE modified Denoising AutoEncoder.
MDB Mean Distance to the Best.
MNAR Missing Not At Random.

OD Oreillette Droite.
OG Oreillette Gauche.

RMSE Root Mean Square Error.

SVM Suppot Vectors Machine.

VAE Variational AutoEncoder.
VD Ventricule Droit.
VG Ventricule Gauche.

WCT Wilson Central Terminal.

Introduction générale

Les motivations

Les maladies cardiovasculaires sont la première cause de mortalité chez les adultes dans le monde (Roth et al., 2020). Dans ce contexte, l'Institut Hospitalo-Universitaire (IHU) Liryc a été créé en 2011, dans le cadre du programme des Investissements d'Avenir dont l'objectif est de dynamiser la recherche médicale et l'innovation. Le Liryc, porté à l'origine par le Pr. Haïssaguerre et désormais par le Pr. Jaïs, est entièrement dédié à la rythmologie et à la prise en charge des maladies cardiovasculaires. Porteur d'une excellence scientifique de grande ampleur, il entend relever le défi de la compréhension, de l'innovation et de la prévention dans ce domaine. Le Liryc regroupe ainsi de multiples spécialités autour de l'électrophysiologie cardiaque, allant de l'étude des cellules cardiaques jusqu'à l'étude du cœur entier et les soins prodigués aux patients. En étudiant les maladies du rythme cardiaque sous plusieurs angles, les équipes du Liryc cherchent à comprendre les mécanismes cardiaques qui engendrent le décès des individus.

Plusieurs outils sont utilisés pour prévenir et diagnostiquer le risque de mort par les maladies cardiovasculaires. Ces outils de mesure de l'activité électrique ont été développés au cours du 20ème siècle. L'électrocardiogramme 12 dérivations (ECG 12D) est l'examen clinique le plus communément utilisé aujourd'hui afin de déceler une anomalie cardiaque en raison de son caractère non-invasif et de sa rapidité de mise en œuvre. Les examens d'imagerie, tels que l'échographie, l'IRM, ou le scanner cardiaque, sont également utilisés afin de détecter des zones structurelles anormales dans le cœur.

Les décès dus aux maladies cardiovasculaires sont, dans plus de la moitié des cas, la conséquence d'arythmies ventriculaires entraînant la mort, telles que les Fibrillations Ventriculaires (FV) (Fishman et al., 2010 et Murakoshi et Aonuma, 2013). La mort causée par une FV est appelée mort subite. La prévention des patients à risque de faire une mort subite, mais qui n'ont jamais présenté d'accident rythmique, est très complexe. Et la majorité des individus mourant d'une mort subite d'origine cardiaque sont des individus qui n'ont pas pu être identifiés assez tôt comme étant à risque et qui n'ont donc pas pu être pris en charge en prévention de l'arrivée d'une FV. Cette absence de prise en charge est due à l'absence d'examen non-invasif assez précis qui permettrait d'identifier les FV.

Au sein de l'IHU Liryc, le Pr Haïssaguerre et l'équipe de traitement du signal dirigée par Rémi DUBOIS mènent des études actives sur l'identification des FV pour la prévention de la mort subite. A l'initiative du Pr Haïssaguerre et pour répondre à la prévention

des morts subites cardiaques, le projet HELP (Heterogeneous Electrical tissue Localization Program) a vu le jour. L'objectif du projet est d'utiliser un ECG haute densité, composé de 128 électrodes afin d'enregistrer le signal cardiaque de manière non-invasive en supposant que l'information enregistrée sera plus fine que celle enregistrée par l'ECG 12D standard et permettra une identification des patients à risque de faire une mort subite.

Au cours des six dernières années, les équipes du projet HELP ont enregistré les signaux électriques à la surface du corps chez plus de 800 individus (contrôles et pathologiques) à l'aide de l'ECG haute définition 128 électrodes. A partir des signaux des 128 électrodes, des caractéristiques électriques directement liées à l'électrophysiologie cardiaque ont été calculées sur ces signaux. Au total, plus de 1500 caractéristiques électriques décrivant l'activité électrique à la surface du corps ont été calculées. L'étude de ces caractéristiques serait la technique idéale pour identifier les sujets à risque de mort subite cardiaque.

Objectifs et contributions

Les travaux et les objectifs de cette thèse s'inscrivent dans l'avancée pour proposer un score de stratification du risque de mort subite grâce à l'étude des caractéristiques électriques.

Après l'enregistrement des signaux sur des individus sains et pathologiques, 14 marqueurs en lien avec l'électrophysiologie cardiaque ont été extraits des signaux. Ce travail est le résultat de thèse de la précédente doctorante Nolwenn TAN (Tan, 2021). L'étude des données nous a confrontés à la présence de données manquantes. En effet, lorsque une électrode n'enregistre pas le signal cardiaque (électrode mal collée), il n'est pas possible d'extraire de données numériques. Ces données manquantes étant présentes chez une très grande partie des patients enregistrés, nous avons décidé de développer une méthode d'imputation des données manquantes. Nous avons réfléchi cette méthode dans un contexte général pour pourvoir l'appliquer à tout nouveau jeu de données tabulaires numériques avec des données manquantes. Nous avons développé une méthode basée sur des Denoising AutoEncoders (DAE). Les DAE ont été proposés pour la première fois par Vincent et al. (2008) pour reconstruire, à partir de données bruitées, les données d'origine. Le DAE fonctionne en corrompant les entrées d'un AutoEncoder (AE). Les DAE ont également été utilisés pour l'imputation de données manquantes (Duan et al., 2014 ; Gondara et Wang, 2018 ; Ryu, M. Kim et H. Kim, 2020). En effet, les DAE, conçus pour reconstruire une sortie non bruitée à partir d'une entrée bruitée, conviennent comme méthode d'imputation quand on considère les valeurs manquantes comme un cas particulier d'entrées bruitées. Cependant les méthodes actuelles qui utilisent des DAE pour imputer les données manquantes proposent seulement de pré-imputer les données manquantes. Ensuite le DAE est entraîné sur le jeu de données pré-imputées pour apprendre la structure des données. En faisant cela, le DAE apprend à reconstruire des données pré-imputées : cela n'est toutefois pas souhaitable puisque ce ne sont pas les bonnes valeurs à apprendre. Pour résoudre la gestion des données pré-imputées d'un DAE dans le cas de l'imputation de données manquantes, nous avons proposé une modification du DAE. Le "modified Denoising AutoEncoder" (mDAE) bénéficie d'une fonction de coût modifiée qui lui permet une meilleure gestion et recons-

truction des données manquantes. Pour étudier l'apport du DAE dans la littérature des méthodes d'imputation de données, nous l'avons comparé à plusieurs méthodes à l'état de l'art et sur plusieurs pourcentages et types de données manquantes. Nous avons également proposé un nouveau critère de comparaison, le Mean Distance to the Best (MDB) qui mesure la performance globale d'une méthode sur tous les jeux de données (pour un pourcentage donné et un mécanisme donné de données manquantes artificielles).

Sur les données cliniques, nous avons mis en place une procédure de comparaison pour choisir la meilleure méthode d'imputation. Les résultats obtenus sur le MDB associés à des temps de calculs très rapides, nous ont confirmé le choix d'utiliser le mDAE sur les données cliniques. Après avoir appliqué le mDAE sur les données numériques extraites des enregistrements de l'ECG 128 électrodes, nous avons utilisé les données imputées et donc sans données manquantes pour mener nos études statistiques. Après avoir fait une première analyse descriptive des données, nous avons appliqué des méthodes de classification binaire pour identifier les patients pathologiques. Au cours de cette étude, nous avons pu étudier les résultats de prédiction obtenus avec notre matrice imputée par rapport à une deuxième matrice de données développée dans un travail de thèse précédent (Tan, 2021). Cette comparaison a permis de montrer l'apport de la nouvelle matrice obtenue grâce à notre méthode d'imputation des données manquantes par rapport à la matrice obtenue par l'agrégation des données d'origine de Nolwenn TAN (Tan, 2021). Pour tenter d'améliorer les résultats de prédiction de pathologie, nous avons proposé une nouvelle méthode de classification binaire basée sur l'utilisation d'un DAE. Plus précisément, nous avons proposé une méthode de détection des patients pathologiques hors de la distribution des patients sains. Pour cela nous avons entraîné un DAE uniquement sur des patients sains. Puis, après avoir masqué une partie des valeurs de nouveaux patients pathologiques et sains, nous avons analysé la différence dans la qualité de reconstruction ces patients. Il y a une différence notable dans la qualité de reconstruction entre les deux groupes. Nous avons utilisé ces erreurs de reconstruction comme critère de classification. Les résultats de courbe ROC obtenus sont proches de ceux obtenus par les méthodes standards de classification binaire. Pour finaliser l'étude de l'identification de patients à risque de mort subite cardiaque, nous avons présenté des résultats préliminaires pour de la classification multi-classes entre patients pathologiques. Ces résultats ne permettent pas de discriminer les pathologies entre elles.

La structure

Ce travail s'organise en 3 chapitres.

Dans le premier chapitre, le contexte clinique dans lequel s'est déroulé cette thèse est décrit. Dans sa première section, la physiologie du cœur puis les diverses techniques d'enregistrement de l'activité électrique cardiaque sont présentées. Dans une deuxième section, sont détaillées les arythmies ventriculaires. Enfin, dans la section 3, le contexte dans lequel les données cliniques ont été enregistrées est évoqué ainsi que la procédure d'extraction des données numériques : c'est à ce stade que nous mettons en avant la présence de données manquantes.

Le deuxième chapitre présente ensuite les solutions apportées pour répondre à la pré-

sence des données manquantes. Sa section 1 expose une première approche de gestion des données manquantes, basée sur l'agrégation des données cliniques qui permet d'obtenir une première matrice de données sans données manquantes (Tan, 2021). La deuxième section présente le mDAE, une méthode d'imputation de données manquantes pour les données numériques tabulaires ; cette méthode est mise en place dans un contexte général d'imputation de données manquantes. Dans la section suivante, nous expliquons comment la méthode, appliquée aux données cliniques, nous permet d'avoir une deuxième matrice de données sans données manquantes.

Enfin, le troisième et dernier chapitre présente les résultats préliminaires de construction du score de risque de mort subite. A cet effet, dans la première section, nous avons mis en concurrence les deux matrices, résultant des réponses à la présence de données manquantes, en utilisant des classifieurs binaires (Support Vectors Machine, Forêts Aléatoires). Dans une deuxième section, nous décrivons notre nouvelle méthode basée sur un DAE pour essayer d'apporter une solution plus fine au problème de classification binaire. Dans la dernière section, nous présentons les résultats préliminaires d'un problème multi-classes visant à discriminer les pathologies entre elles.

Liste des publications

Dupuy, M., Chavent, M., Dubois, R., Autoencoders pour l'imputation de données manquantes, 54èmes Journées de Statistique de la Société Française de Statistique (SFdS), Bruxelles (2023).

Dupuy, M., Chavent, M., Dubois, R., Denoising Autoencoders for The Detection of Patients Out of Distribution of Healthy Individuals. In Computing In Cardiology (Cinc), Karlsruhe (2024).

Dupuy, M., Chavent M., Dubois R., mDAE : modified Denoising AutoEncoder for missing data imputation. arXiv (2024). *En cours de soumission.*

Chapitre 1 : Contexte clinique

Table des matières

Au cours de ce travail de thèse nous analysons et nous étudions des données extraites des enregistrements de l'activité cardiaque du cœur. Dans ce premier chapitre, nous introduisons le contexte clinique dans lequel s'inscrit le manuscrit de thèse. Dans un premier temps, l'anatomie et l'activité électrique du cœur sont expliqués. Les arythmies et les pathologies ventriculaires sont détaillées par la suite. Dans la dernière section du chapitre, les données cliniques étudiées pendant cette thèse sont présentées.

1.1 L'électrophysiologie cardiaque

Cette section est une introduction à l'électrophysiologie cardiaque. La physiologie est la science qui étudie les fonctions et les propriétés des organes et des tissus des êtres vivants. L'électrophysiologie est une partie de la physiologie qui étudie les propriétés électriques des tissus vivants. Elle implique la mesure de différences de tensions ou de courants électriques à différentes échelles biologiques, du canal ionique isolé, jusqu'à des organes entiers, comme le cœur.

1.1.1 La physiologie du cœur

L'anatomie du cœur

Le cœur est un muscle creux situé au centre de la cage thoracique, entre les poumons, au dessus du diaphragme. Son poids moyen chez l'humain est d'environ 250 g pour la femme et 300 g pour l'homme.

A l'échelle macroscopique, cet organe vital est divisé en quatre cavités (voir Figure 1.1). Dans la partie supérieure du cœur, on retrouve l'oreillette droite (OD) et gauche (OG) ; elles reçoivent le sang qui revient au cœur. Dans la partie inférieure, on retrouve le ventricule droit (VD) et gauche (VG) ; ils pompent le sang vers les autres organes du corps. Au milieu du cœur, le septum sépare la partie gauche de la partie droite. Cette paroi permet d'éviter tout échange sanguin entre les deux moitiés du cœur.

A l'échelle microscopique, la paroi extérieure du cœur est divisée en trois couches (voir Figure 1.2). La couche extérieure est l'épicarde, c'est l'enveloppe du cœur. La couche centrale est le myocarde, c'est le muscle cardiaque qui est responsable de la contraction rythmée du cœur. Les performances et les pathologies cardiaques sont, en partie, dépendantes de la qualité de ce tissu. La couche intérieure est l'endocarde.

1. https ://www.chu-rouen.fr/services/cardiologie/valves-cardiaques/fonctionnement-du-coeur/
2. http ://univ.ency-education.com/uploads/1/3/1/0/13102001/histo2an_cardio-paroi_cardiaque2022slimani.pdf

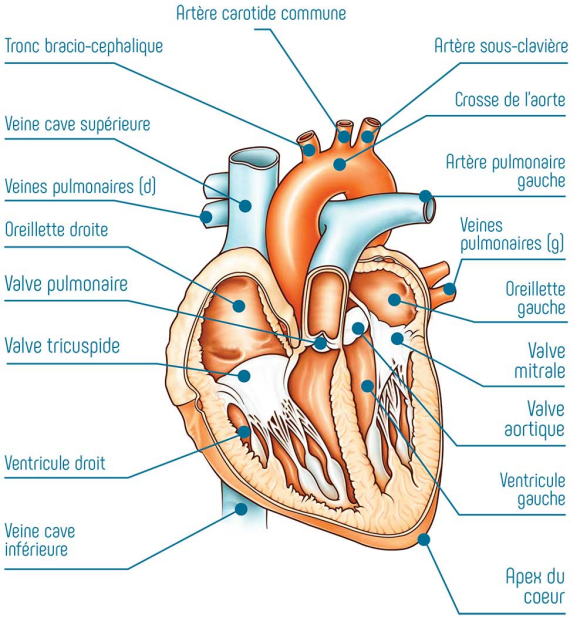


FIGURE 1.1 – **Anatomie macroscopique du cœur**
Extrait du site du Centre Hospitalier Universitaire de Rouen [1]

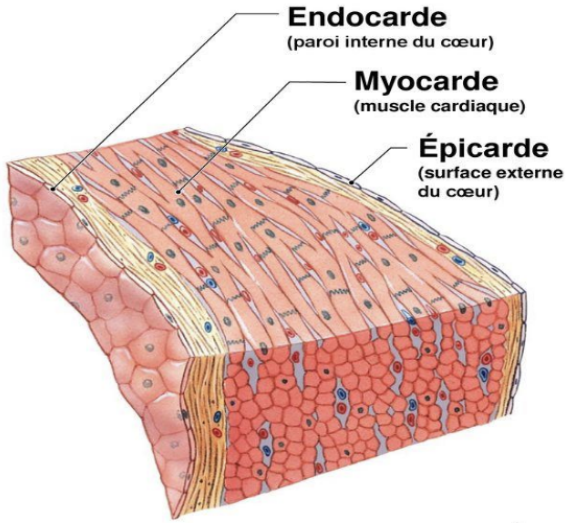


FIGURE 1.2 – **Vue microscopique de la paroi du cœur**
Extrait du cours de Dr A.SLIMANI de l'Université d'Ency [2]

La fonction de double pompe

Le cœur exerce une fonction de double pompe en assurant deux circuits de circulation du sang distincts et simultanés grâce à ses parties droite et gauche qui se contractent simultanément permettant le maintien de la circulation sanguine dans tout le corps à travers les veines et les artères. La circulation du sang dans le cœur fonctionne sous la forme d'un cycle (voir Figure 1.3).

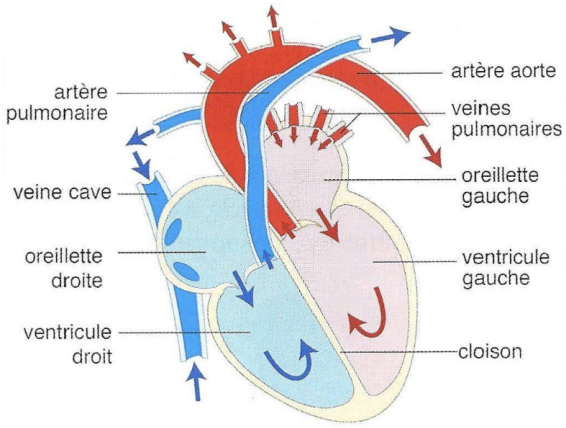


FIGURE 1.3 – **Schéma de la circulation du sang dans le cœur**
Représenté en bleu, le sang pauvre en oxygène, et en rouge, le sang chargé en oxygène.
Extrait du site internet de M.BESSOUD-CAVILLOT [3]

En tant que "double pompe," chaque côté du cœur fonctionne de façon synchrone pour maintenir la circulation sanguine en continu. La phase de diastole correspond à la relaxation du muscle cardiaque et à un remplissage de sang des oreillettes et des ventricules. La phase de systole correspond à la contraction du myocarde et l'expulsion du sang des oreillettes vers les ventricules et des ventricules vers les organes. Mais le côté droit et le côté gauche agissent indépendamment : au cours du cycle cardiaque, la "double pompe" assure la circulation du sang pauvre en oxygène d'une part, et du sang chargé en oxygène d'autre part. La partie droite du cœur recueille le sang pauvre en oxygène (en bleu dans la Figure 1.3) provenant des organes via le système veineux qui converge vers la veine cave inférieure et la veine cave supérieure. Celui-ci est admis dans l'oreillette droite qui le propulse ensuite dans le ventricule droit. La systole ventriculaire permet au sang contenu dans le ventricule droit d'être éjecté dans les poumons via l'artère pulmonaire, où il est oxygéné. Le sang alors chargé en oxygène (en rouge dans la Figure 1.3), est ramené à l'oreillette gauche via les veines pulmonaires. Le sang est transféré dans le ventricule gauche puis éjecté dans l'organisme via l'aorte et le système artériel. Il permet à l'organisme d'obtenir l'oxygène et les nutriments nécessaires à son fonctionnement.

La conduction électrique

La fonction de "double pompe" du cœur est déclenchée par des impulsions électriques générées périodiquement sous contrôle du système nerveux par le nœud sinusal (voir Figure 1.4) situé dans la partie supérieure de l'oreillette droite (MALMIVUO et PLONSEY, 1995).

Chaque impulsion se propage de proche en proche dans les cellules musculaires qui composent le myocarde auriculaire (MATTEUCCI, 1842) induisant ainsi la contraction des oreillettes. L'impulsion arrive alors au nœud auriculo-ventriculaire, seul point de passage électrique entre les oreillettes et les ventricules. De ce point, l'onde électrique emprunte un

3. http ://eric.bessoudcavillot.free.fr/4eme/EPI/4svtPart1TH1Chap1_cours.pdf

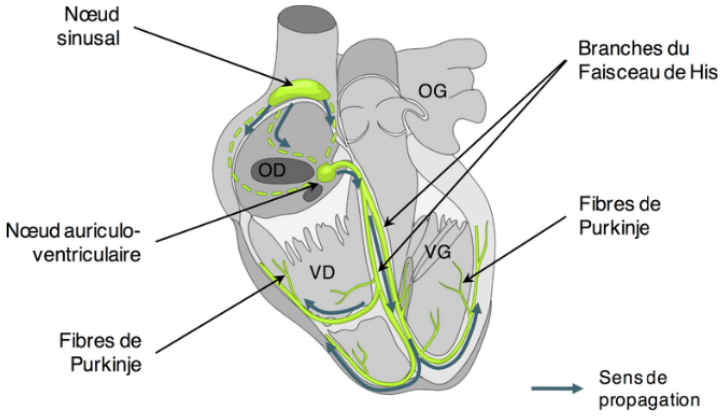


FIGURE 1.4 – **La conduction électrique**
OD : oreillette droite ; OG : oreillette gauche ; VD : ventricule droit ; VG : ventricule gauche. Source : DALLET (2017)

faisceau de conduction rapide appelé Faisceau de His et fibres de Purkinje qui permettent la diffusion rapide de l'influx électrique en de multiples points du myocarde ventriculaire pour donner naissance à une contraction rapide du muscle ventriculaire. Le système nerveux maintient le rythme de ces impulsions, et en conséquence, le rythme cardiaque entre 60 et 85 battements par minute au repos.

Nous venons de voir que la fonction de "double pompe" du cœur fonctionne grâce aux impulsions électriques qui traversent les cellules musculaires cardiaques. Il est possible de visualiser, au cours d'examens médicaux, l'activité électrique du cœur.

1.1.2 Enregistrement de l'activité électrique du cœur

L'activité électrique du cœur peut être étudiée au travers de différents outils médicaux tels que l'électrocardiogramme (ECG), l'Holter ECG, les cathéters, les pacemakers, etc. L'étude de l'activité électrique du cœur permet d'obtenir des informations liées au rythme cardiaque, à la présence ou non de tissus responsables des troubles du rythme ou de la propagation électrique cardiaque. Dans cette thèse, nous nous intéressons aux enregistrements du signal cardiaque par ECG.

L'électrocardiogramme 12 dérivations

L'électrocardiogramme 12 dérivations (ECG 12D) est un système d'enregistrement très utilisé afin de mesurer l'activité électrique du cœur puisqu'il permet d'obtenir une information globale et assez précise de l'activité du cœur.

L'ECG 12D est composé de 10 électrodes positionnées sur la surface du torse et sur des

membres du corps, et grâce auxquelles il est possible d'obtenir une mesure globale de l'activité électrique du cœur. L'ECG permet ainsi de mesurer le rythme cardiaque et de diagnostiquer certains troubles du rythme ou de la propagation électrique cardiaque. L' ECG 12D a été développé par trois scientifiques sur trois époques différentes. Einthoven développe la première partie au début des années 1900 (Einthoven, 1906). Pour cette première partie, il positionne trois électrodes, une sur le bras droit, une sur le bras gauche et une sur la jambe gauche. L'enregistrement de la différence de potentiel électrique entre deux électrodes donne une dérivation. En combinant les différences des trois électrodes, il obtient les **dérivations standards** I, II et III (voir Figure 1.5, en haut à gauche) qui forment le triangle d'Einthoven. On parle ici de dérivations bipolaires car ce sont les différences entre deux électrodes.

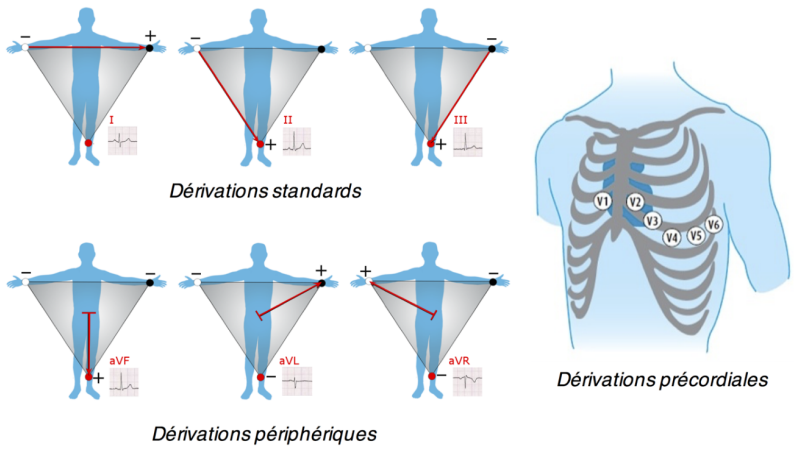


Figure 1.5 – **Principe de l'électrocardiogramme 12 dérivations clinique**
Positionnement des dérivations standards, périphériques et précordiales
Source : Dallet (2017)

En 1934, Wilson propose d'ajouter 6 nouveaux emplacements d'électrodes (Wilson et al., 1934). Ces six électrodes supplémentaires sont positionnées directement sur la face avant du thorax, à proximité du cœur (voir Figure 1.5, à droite). L'ensemble de ces six voies (V1 à V6) forme les **dérivations précordiales**. Wilson mesure alors la différence de potentiel entre une électrode étudiée et une électrode virtuelle de référence située au milieu du cœur appelée "Wilson Central Terminal" (WCT). Le WCT est construit en moyennant les dérivations I, II et III. Il est le centre du triangle d'Einthoven. On parle ici de dérivations précordiales unipolaires car elles sont la différence entre une électrode et une électrode virtuelle. Enfin, dans les années 1940, Goldberger, trouvant les électrodes d'Einthoven trop éloignées les unes des autres, ajoute trois dérivations calculées à partir des dérivations de Einthoven (voir Figure 1.5, en bas à gauche) : aVR, aVL et aVF (Goldberger, 1942). Ce sont les **dérivations périphériques** unipolaires.

Les travaux d'Einthoven, Wilson et Goldberger donnent alors naissance à l'ECG 12D (12 dérivations : I, II, III, aVR, aVL, aVF, V1 à V6) , tel qu'il est pratiqué en clinique encore aujourd'hui pour diagnostiquer et identifier les troubles du rythme cardiaque ou les tissus pathologiques.

Tracé électrique d'un ECG 12D

En rythme sinusal (rythme normal), l'ECG 12D produit le document médical que l'on retrouve dans la Figure 1.6.

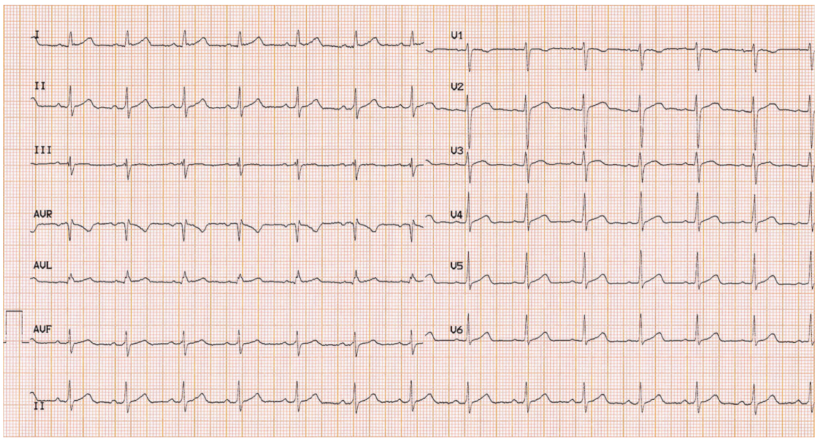


FIGURE 1.6 – **Exemple d'un tracé ECG enregistré lors d'un rythme sinusal**
Extrait du site e-cardiogram [4]

A chaque battement, l'activité électrique qui entraîne la contraction synchrone des oreillettes et des ventricules, laisse apparaître sur l'ECG des formes correspondant à leur activité (voir Figure 1.7) :

- L'onde P : elle correspond à l'activation de la contraction des deux oreillettes. La durée de cette onde indique le temps pris par l'onde électrique pour se propager à travers les oreillettes. La durée de l'onde P pour un individu sain est de l'ordre de 0,08 à 0,1 s pour une amplitude inférieure à 0,25 mV.
- Le complexe QRS : correspond à la propagation de l'onde électrique dans les ventricules. En moyenne, chez une personne saine, le complexe QRS a une durée inférieure à 0,1 secondes et une amplitude comprise entre 1 et 2 mV.
- L'onde T : correspond à la période où les ventricules retrouvent leur état électrique de repos après la contraction. Sa durée est comprise entre 0,2 et 0,25 secondes. Son amplitude ne dépasse pas la moitié de celle du QRS pour un individu sain.

Le rapport (VIRANI et al., 2021) insiste sur l'importance de l'ECG pour une détection précoce des maladies cardiaques. Cependant, certaines limites ont amené à de nouveaux développements.

Amélioration de l'ECG 12 dérivations

Une première limite de l'ECG 12D est sa faible résolution spatiale sur le torse, due à son nombre relativement faible d'électrodes. Une solution pour y répondre est la cartographie

4. https ://www.e-cardiogram.com/rythme-sinusal/
5. https ://www.e-cardiogram.com/complexe-qrs/

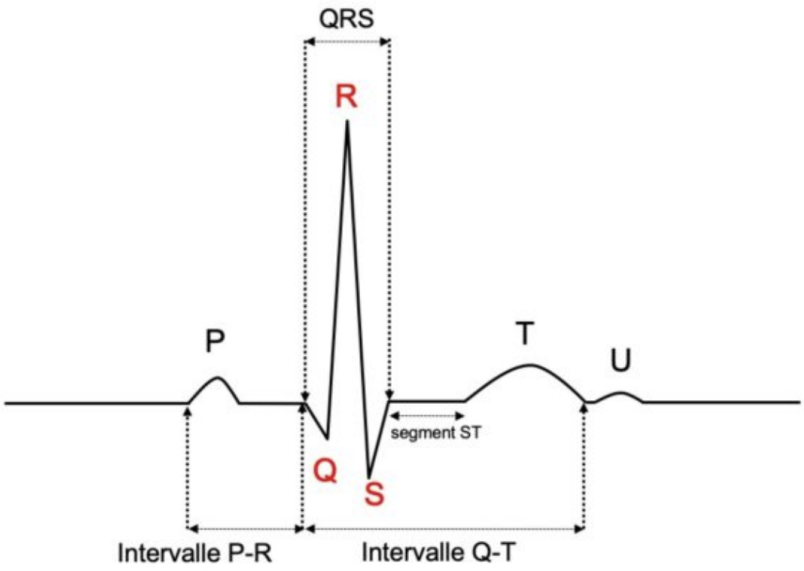


FIGURE 1.7 – **Les ondes du tracé ECG**
Extrait du site e-cardiogram [5]

haute résolution (TACCARDI, DE AMBROGGI et VIGANOTTI, 1976 ; MIRVIS, 2012) qui consiste à positionner entre 20 à 250 électrodes afin d'obtenir une résolution spatiale plus fine de l'activité cardiaque et ainsi accéder à une information plus fine du signal électrique. Le système peut également être doté d'une ceinture de plethysmographie permettant d'enregistrer le signal de respiration. L'appareil haute résolution utilisé pour enregistrer les données de cette thèse est un ECG 128D.

1.2 Arythmies, pathologies ventriculaires et fibrillations ventriculaires

L'ECG 12D, et en particulier sa version améliorée avec un nombre d'électrodes plus grand, permet de surveiller précisément l'activité électrique du cœur. Cette surveillance se déroule lorsque le cœur est en rythme sinusal. Grâce à l'ECG, il est possible d'identifier les altérations du rythme cardiaque. En effet, les altérations de la conduction et l'hétérogénéité électrique du tissu se traduisent directement par une modification des formes d'ondes des signaux de surface. L'objectif des sections suivantes est de définir l'arythmie ventriculaire puis de décrire les pathologies ventriculaires les plus fréquentes, à l'origine des arythmies ventriculaires. Enfin, est présentée la fibrillation ventriculaire : une arythmie létale.

1.2.1 Définition d'une arythmie ventriculaire

L'arythmie cardiaque est une anomalie du rythme ou de la fréquence cardiaque. L'apparition d'une arythmie ventriculaire est liée aux interactions entre différents facteurs. Ces interactions ont été formalisées par le Triangle de Coumel (Coumel, 1987) (voir Figure 1.8). Aux sommets de ce triangle figurent les facteurs nécessaires à la survenue d'une arythmie : le substrat, le facteur déclencheur et le système nerveux autonome.

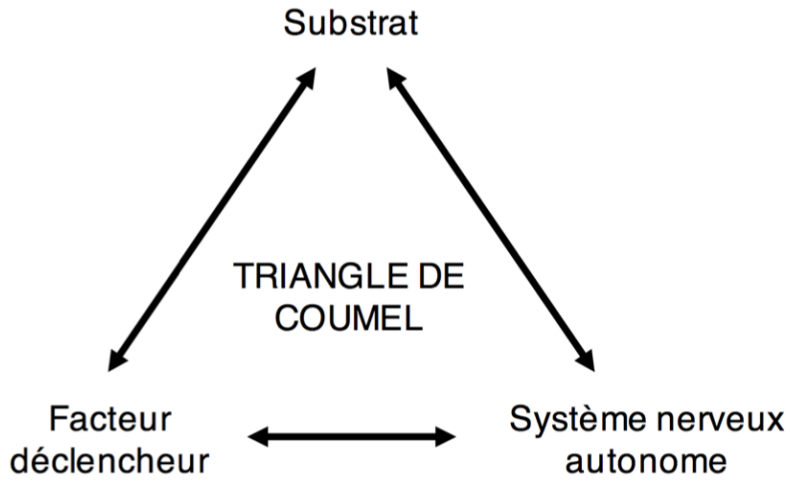


Figure 1.8 – **Triangle de Coumel**
Schéma adapté de Coumel, 1987

Le substrat est le support anatomique et/ou électrophysiologique permettant le maintien de l'arythmie dans le temps. Un premier type de substrat est le substrat fonctionnel. Il est purement électrique et correspond à un dysfonctionnement des canaux ioniques (canaux qui permettent les échanges d'ions pendant la propagation du courant électrique dans le myocarde). Un deuxième type de substrat est le substrat structurel. Il est caractérisé par des zones de fibrose ou de graisse, inertes électriquement. Ces zones créent des perturbations de la propagation électrique dans le tissu myocardique.

Le facteur déclencheur est l'élément électrique susceptible de déclencher l'arythmie par activation du substrat arythmogène. Le facteur déclencheur peut être une extrasystole (battements cardiaques anormaux).

Le système nerveux autonome a la possibilité d'agir sur les deux précédents facteurs. Avec son rôle de modulateur, il peut intensifier ou inhiber la sensibilité du substrat, en particulier fonctionnel, et la conductivité du tissu, favorisant le déclenchement de l'arythmie.

Certaines arythmies peuvent être asymptomatiques mais leur présence ne doit pas être ignorée car elles représentent par exemple un facteur de risque d'accident vasculaire cérébral, d'insuffisance cardiaque ou un risque élevé de mort subite. D'autres arythmies nécessitent une prise en charge médicale immédiate comme certaines tachycardies ventriculaires ou les fibrillations ventriculaires.

1.2.2 Les pathologies ventriculaires

On s'intéresse ici aux pathologies localisées dans les ventricules qui entraînent l'apparition d'arythmies (les pathologies font apparaître les facteurs du Triangle de Coumel décrits précédemment) ; étant des pathologies directement liées au muscle du cœur, le myocarde, elles sont appelées cardiomyopathies. Dans les sections suivantes, nous ne décrirons que les cardiomyopathies qui seront étudiées dans la suite du manuscrit.

Les cardiomyopathies ischémiques

Les cardiomyopathies ischémiques sont le résultat d'une altération structurelle du myocarde liée à une insuffisance d'oxygénation, le plus souvent causée par l'obstruction totale ou partielle d'une artère coronaire (voir Figure 1.9). Ce déficit en oxygène (ischémie) se traduit par une nécrose du tissu myocardique (infarctus du myocarde) qui crée ainsi un substrat favorable aux arythmies ventriculaires.

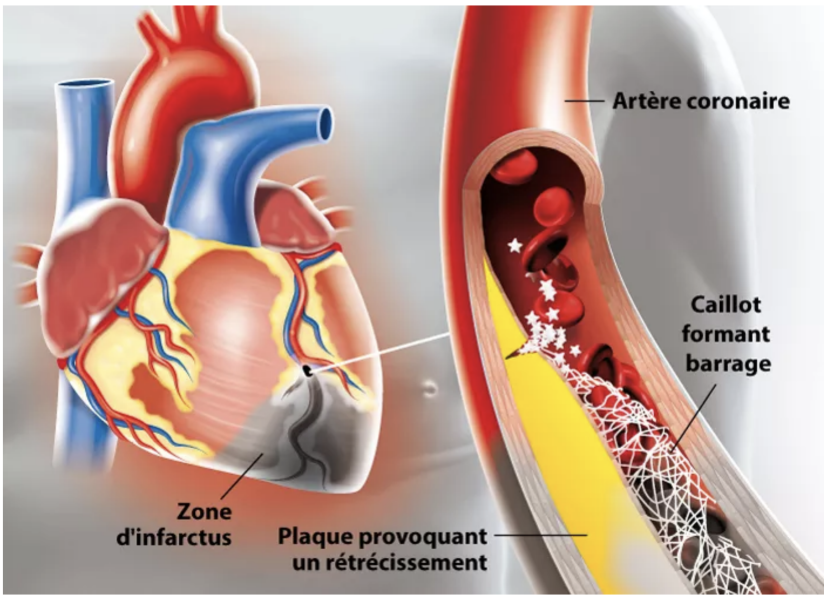


FIGURE 1.9 – **Infarctus du myocarde**
Le déficit en oxygène dans l'artère coronaire provoqué par un caillot crée une nécrose du tissu myocardique.

Les cardiomyopathies non-ischémiques

Les cardiomyopathies non-ischémiques, sont des maladies affectant structurellement le muscle cardiaque, mais non liées à une ischémie. Elles peuvent provoquer également des arythmies. Divers types de cardiomyopathies non-ischémiques sont distingués, notamment les cardiomyopathies dilatées, les cardiomyopathies hypertrophiques et les Dysplasies Arythmogènes du Ventricule Droit (RICHARDSON, 1996).

Les cardiomyopathies dilatées correspondent à une dilatation de la cavité ventriculaire le plus souvent conséquence d'insuffisance cardiaque. En revanche, les cardiomyopathies hypertrophiques surviennent lorsque le muscle cardiaque est trop gros, le plus souvent en conséquence d'anormalités génétiques. La Dysplasie Arythmogène du Ventricule Droit (DAVD) est une maladie génétique qui provoque le remplacement des cellules cardiaques par du tissu fibro-adipeux principalement localisé dans l'épicarde (voir Figure 1.2) pouvant s'étendre jusqu'à l'endocarde (voir Figure 1.2). Ce substrat fibro-adipeux est arythmogène et peut favoriser le déclenchement d'arythmies ventriculaires. Cette cardiomyopathie génétique touche majoritairement les hommes, souvent jeunes, avant 40 ans.

Syndrome de Brugada

Le syndrome de Brugada est une canalopathie (maladie génétique responsable d'une dysfonction des canaux ioniques), qui a longtemps été considérée comme une pathologie d'origine purement électrique. Plus récemment, il a été découvert qu'elle est aussi associée à une pathologie d'origine structurelle (Nademanee et al., 2011).

Les patients affectés par ce syndrome sont le plus souvent asymptomatiques. Le réel enjeu est de détecter chez ces sujets, ceux qui peuvent présenter des syncopes par arythmie ventriculaire, ou un arrêt cardiaque.

1.2.3 Les arythmies peuvent entraîner la mort subite

Les cardiomyopathies sont des terrains favorables et identifiés pouvant entraîner une mort subite. La mort subite est une mort survenant de façon inattendue dans l'heure qui suit les premiers symptômes, généralement causée par une Fibrillation Ventriculaire (FV). La FV est responsable de 50 000 morts subites en France chaque année, soit une mort toutes les 10 minutes.

La fibrillation ventriculaire

La FV correspond à une activité électrique anarchique et désorganisée de la contraction des ventricules (voir Figure 1.10) : le cœur ne réalise plus sa fonction de double pompe. Si aucune intervention (défibrillation, massage cardiaque) n'est réalisée dans les minutes qui suivent la FV, le patient décédera.

Un cas particulier : la Fibrillation Ventriculaire Idiopathique (FVI)

Des morts subites restent "inexpliquées" : elles représentent 14% à 23% des morts subites chez les jeunes de moins de 35 ans (Haissaguerre, Duchateau et al., 2020). Dans le cas où une fibrillation ventriculaire a été authentifiée par un ECG mais que les tests cliniques (tests médicamenteux, examen ECG et examen par imagerie, autopsie) n'ont détecté

FIGURE 1.10 – **Tracé ECG de la fibrillation ventriculaire**

aucune anomalie électrique, structurelle ou métabolique, ces morts subites "inexpliquées" sont appelées FVI.

Cependant, une étude récente (HAISSAGUERRE, HOCINI et al., 2018) menée par les équipes du Liryc, a identifié au cours d'interventions invasives utilisant des cathéters (voir Figure 1.12), des éléments pouvant expliquer ces fibrillations ventriculaires. L'équipe de médecins a localisé des signaux électriques anormaux sur l'endocarde et l'épicarde de patients ayant subi une fibrillation ventriculaire "inexpliquée". La présence de ces signaux (dans 62% des cas de l'étude) met en évidence l'existence de substrat anormal localisé, expliquant ainsi la survenue de l'arythmie ventriculaire. Ces anomalies structurelles étant non décelables par les méthodes d'imagerie standard, l'objectif est de rechercher leur traces électriques sur l'ECG de surface afin d'essayer de prévenir la mort subite cardiaque.

1.3 Le projet HELP : contexte, objectifs, données

L'une des limites de la technique ECG 12D est que les enregistrements ne sont obtenus qu'à partir de 6 **dérivations précordiales** indépendantes, ce qui peut empêcher la détection de potentiels électriques cardiaques de faible amplitude. Un échantillonnage plus étendu de la surface du torse pourrait contribuer à fournir des informations cliniques supplémentaires. A l'initiative du Pr Michel Haissaguerre, le projet HELP (Heterogeneous Electrical tissue Localization Program) a été mis en place. Ce projet vise à examiner la sensibilité supplémentaire de l'ECG en utilisant 128 électrodes à la surface du corps pour mesurer l'activité cardiaque globale ou régionale. On parle alors d'ECG à Haute Densité d'électrodes (ECG HD).

Dans cette section, nous présentons l'origine et les objectifs du projet HELP. Nous expliquons également comment, à partir d'un ECG HD, le signal électrique du cœur a été

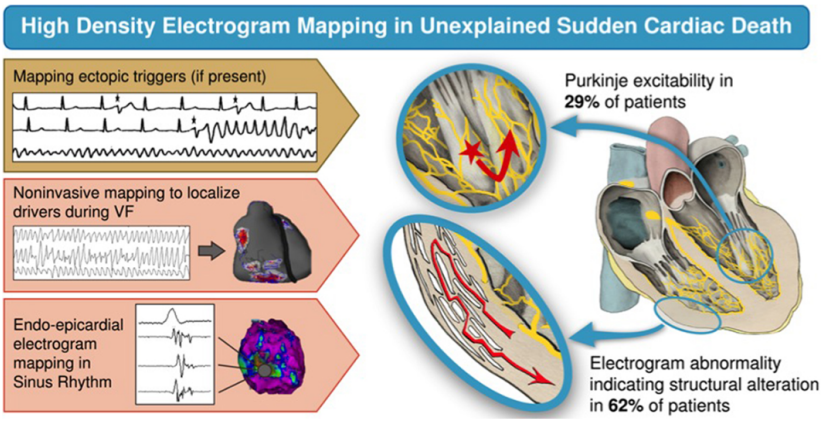


FIGURE 1.11 – **Les mécanismes de la FVI**
Cette figure est extraite de (HAISSAGUERRE, HOCINI et al., 2018). Elle résume les résultats de l'étude incluant 24 patients qui, après réanimation, ont été diagnostiqués avec une FVI. Cette figure montre que pour 62% de ces patients, des zones altérées du myocarde générant des micro-potentiels anormaux, sont observées lors de la cartographie endocardique/épicardique invasive par cathéter. Et, pour 29% de ces patients, aucune anomalie structurelle n'est observée, mais leur FVI pourrait être la conséquence dune anomalie fonctionnelle du Purkinje.

enregistré. Puis quelles informations ont été extraites afin de créer une base de données numériques pour identifier les patients à risque de faire une mort subite.

1.3.1 Présentation du projet HELP

Contexte et objectifs

L'analyse de l'ECG est essentielle au diagnostic clinique des pathologies cardiaques. Cependant, les mesures électriques les plus précises en clinique sont obtenues lors d'explorations électrophysiologiques invasives par introduction de cathéter(s) au contact des parois cardiaques (BUXTON et al., 2006). Les cathéters explorent la surface cardiaque point par point (voir Figure 1.12), ce qui nécessite un temps exploratoire de plusieurs dizaines de minutes. Le caractère invasif de ces explorations comporte un risque, faible mais non nul, de complications graves (hémorragie, tamponnade...), et elles majorent le coût hospitalier. Ces explorations sont en général réalisées au cours d'interventions curatives (ablation) ou cartographiques pour déterminer s'il existe une altération cardiaque responsable d'arythmies graves (fibrillations ventriculaires) ou à haut risque mais jamais en prévention primaire.

L'ECG est une technique non invasive peu coûteuse donnant une vue d'ensemble « macroscopique » du fonctionnement électrique cardiaque. Comme nous l'avons vu dans la section 1.1.2, l'ECG standard utilise 12 dérivations. Les 6 dérivations précordiales V1–V6 sont des enregistrements thoraciques indépendants tandis que les 6 dérivations des membres (DI II III, avR L F) sont dérivées de 2 bipôles d'enregistrement. Le nombre réduit d'élec-

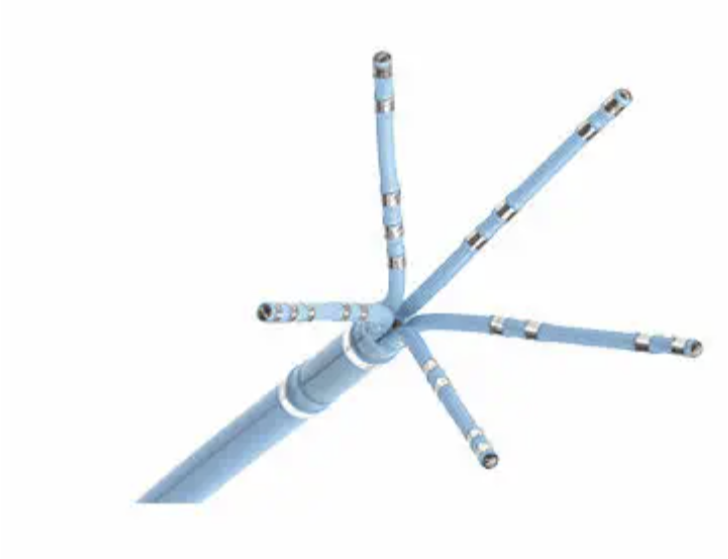


FIGURE 1.12 – **Cathéter de cartographie Biosense webster**
Câble souple, fin (2 à 3 mm de diamètre), en matière plastique, destiné à réaliser un diagnostic ou un traitement à distance, après avoir été introduit par un vaisseau sanguin (artère ou veine). Selon l'objectif recherché, le cathéter pourra mesurer des pressions, injecter du produit de contraste pour opacifier les vaisseaux et cavités cardiaques. Dans le cas du traitement des troubles du rythme, les cathéters ont plusieurs rôles : enregistrer l'activité électrique (normale ou pathologique des oreillettes et des ventricules), déclencher des arythmies et les interrompre, préciser les rapports dans l'espace des cavités cardiaques (cartographie 3D), délivrer une énergie thérapeutique (radiofréquence, cryothérapie) capable d'ablater (détruire) un trouble du rythme. Est retiré en fin d'intervention (matériel à usage unique, non destiné à être implanté).

trodes précordiales peut devenir une limite, dans le cas, par exemple, d'une électrode bruitée/artefactée, qui entraîne une perte significative d'informations.

Les études antérieures ont ainsi démontré les avantages des ECG HD par rapport à l'ECG 12D dans l'objectif de cartographies des arythmies (RAMANATHAN et al., 2004). Cependant, peu d'études ont évalué leur intérêt en rythme sinusal dans un objectif de diagnostic ou de pronostic.

A l'IHU Liryc, des équipes de chercheurs et de cliniciens ont mené des études qui confirment les bénéfices des ECG HD dans l'évaluation de la période où les ventricules retrouvent leur état électrique de repos après la contraction (MEO et al., 2020). Sur la base des travaux antérieurs cliniques et expérimentaux (BURNES et al., 2000), les équipent peuvent avancer que les ECG HD ($\geq$ 64 électrodes) analysés avec les méthodes numériques actuelles devraient pouvoir fournir des informations spatiales et temporelles plus précises et complètes que celles des ECG 12D. Des différences significatives sont anticipées dans l'évaluation des paramètres globaux (QRS) ou des paramètres régionaux, qui augmenteront la sensibilité diagnostique. L'ECG HD devrait mieux percevoir les activités régionales mal visualisées par les 6 électrodes thoraciques standards.

Le projet HELP a pour objectif principal d'évaluer la sensibilité des enregistrements ECG HD dans la mesure des paramètres électriques, en comparaison à un ECG standard. De plus, le projet HELP a pour objectif d'analyser les caractéristiques électriques en rythme sinusal de patients atteints de pathologies ventriculaires.

Les retombées attendues du projet sont un progrès significatif dans l'évaluation des caractéristiques électriques des patients en cardiologie. Une meilleure évaluation pronostique du risque rythmique, chez les patients avec ou sans cardiopathie, est envisageable à plus long terme. Des applications autres que rythmologiques sont hautement anticipées, telle qu'une meilleure détection des anomalies ischémiques à l'état basal ou lors d'une épreuve d'effort (Kania et al., 2019). Enfin, cette méthode d'électrocardiographie à haute densité d'électrodes aura la capacité de se perfectionner grâce aux progrès continus dans l'exploitation numérique associée à l'apprentissage machine.

C'est dans le cadre de cette étude qu'un ECG 128 électrodes haute définition (ECG 128 HD) a été utilisé pour mesurer l'activité électrique du cœur des individus sains ainsi que des patients présentant des pathologies cardiaques, les deux groupes d'individus étant choisis et enregistrés au cours d'un protocole clinique précis.

Le protocole expérimental

Chaque patient est soumis à un enregistrement par ECG 128 HD. Les électrodes sont positionnées sur des bandes contenant entre 10 et 11 électrodes. Ces bandes sont préparées avec une bande adhésive et enduites avec du gel conducteur avant l'installation sur le patient (voir Figure 1.13). La position des bandes est définie en fonction des repères à l'ECG standard (voir Figure 1.5 à droite).

L'ensemble est enfin connecté sur un amplificateur (AD box) lui-même branché à un récepteur USB2 connecté à un ordinateur portable (voir Figure 1.14). Avant l'acquisition des signaux, tous les appareils électriques de la pièce sont débranchés pour éviter les interférences et un test est effectué pour vérifier la qualité de l'enregistrement. L'acquisition des données est réalisée avec le logiciel ActiView version 7.06 (et ultérieures). La durée totale de l'acquisition est de 10 à 30 minutes.

La fréquence d'échantillonnage des potentiels de surface est de 2048 Hz, (à comparer aux 1000 Hz couramment utilisés en clinique). Simultanément à cet enregistrement, une ceinture de pléthysmographie est utilisée afin de mesurer le signal de respiration du patient.

Les patients inclus dans l'étude sont des patients hospitalisés pour arythmie ventriculaire documentée ou suspectée, ou pour trouble de la conduction ventriculaire, dans un contexte ou non d'insuffisance cardiaque. Cette population a été choisie car les limitations de l'ECG 12D y sont particulièrement marquées, notamment dans l'évaluation précise des caractéristiques électriques et du risque rythmique. Ces patients sont donc en mesure de bénéficier d'un maillage plus dense d'électrodes par l'ECG 128 HD pour l'identification des altérations fréquentes et sous-détectées chez ces patients (Fishman et al., 2010).

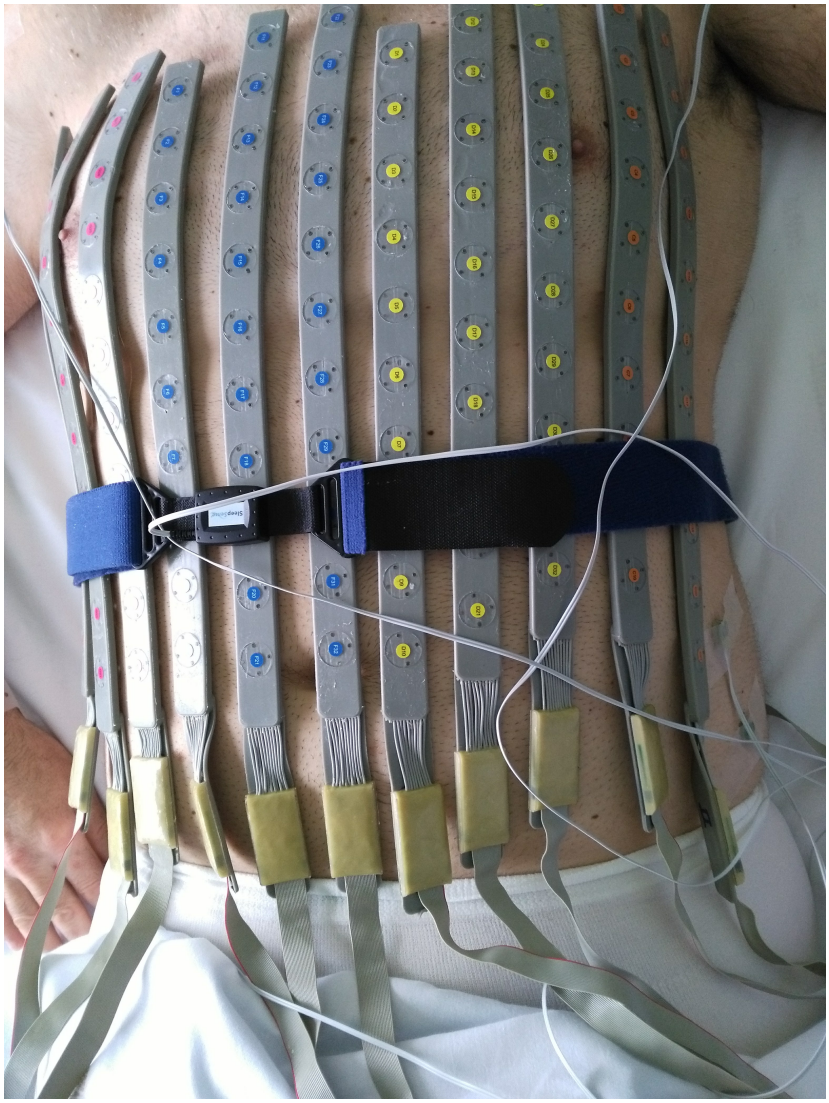


FIGURE 1.13 – **Bandes de l'ECG 128 dérivations**
Les électrodes sont positionnées sur des bandes contenant entre 10 et 11 électrodes. Les bandes sont préparées avec une bande adhésive et enduites avec du gel conducteur.

FIGURE 1.14 – **Système d'enregistrement non-invasif des signaux cardiaques.**

1.3.2 Pré-traitement du signal cardiaque

Le fichier obtenu à la fin de l'enregistrement ECG 128 HD comporte 128 voies (1 pour chaque électrodes), chacune retraçant l'activité électrique du cœur pendant la période d'acquisition. Pour détecter l'activité anormale dans un ECG il faut diminuer au maximum le bruit contenu dans le signal enregistré. Pour faire ce pré-traitement des signaux, la technique "SAECG" (Tan, 2021) a été utilisée, que nous présentons dans la section suivante.

De l'enregistrement au battement moyenné

La technique SAECG pour "Signal Averaged ECG" est une méthode de moyennage des battements du signal cardiaque. Le principe est de moyenner l'ensemble des battements cardiaques équivalents afin de réduire les sources de bruit telles que le bruit induit par la respiration ou le bruit associé à la variabilité physiologique (Tan et al., 2019).

Pour mettre en place la technique "SAECG", il faut tout d'abord, parmi les battements constituant le signal enregistré par une des 128 électrodes, choisir un battement cardiaque comme modèle. Ce choix est fait qualitativement en observant les battements de toutes les voies. Ce battement est appelé "template virtuel".

Ensuite, une Analyse en Composantes Principales (ACP) (Hotelling, 1933) est réalisée. L'ACP est une méthode statistique souvent utilisée pour réduire la dimension d'un jeu de données tout en conservant le maximum de variance du nuage de points projetés. Cette méthode repose sur la génération de nouvelles variables, appelées composantes principales (CP), représentant des axes ou directions dans l'espace des données. La première composante principale est définie comme la direction maximisant la variance des données projetées. La deuxième composante principale, quant à elle, correspond à la direction orthogonale à la première tout en maximisant la variance résiduelle des données et ainsi de suite. En gardant seulement les premières composantes principales, les variables initiales potentiellement corrélées sont alors transformées, via une projection, en un nombre plus petit de nouvelles variables indépendantes. Les composantes principales associées aux petites variances, composantes ayant un rapport signal sur bruit plus faible, sont alors éliminées.

L'ACP est réalisée sur les 128 voies, mathématiquement sur la matrice $X \in \mathbb{R}^{T\times 128}$, où T est la durée de l'enregistrement. Appliquée ici, l'ACP permet de synthétiser l'information des voies, en éliminant le bruit. Pour la méthode SAECG, les CPs seront utilisées comme des voies "virtuelles". La voie "virtuelle" avec la plus grande variance est celle qui représente au mieux les signaux. Cette voie "virtuelle" est la première CP de l'ACP. Pour chaque battement de la voie virtuelle, la corrélation entre le battement et le "template virtuel" est calculée. Si cette corrélation est supérieure à un seuil, fixé à 0,9, la position du battement dans la voie virtuelle est conservée pour l'alignement. Les positions conservées servent à extraire les battements dans les 128 voies d'origine (avant ACP). Les battements extraits sont alors alignés puis moyennés (Figure 1.15).

Nous disposons à ce stade de 128 signaux moyennés (un pour chaque électrode). Dans la suite du manuscrit, nous représentons les données sous forme d'un cube (voir Figure 1.16) ; chaque dimension du cube représente les informations suivantes :

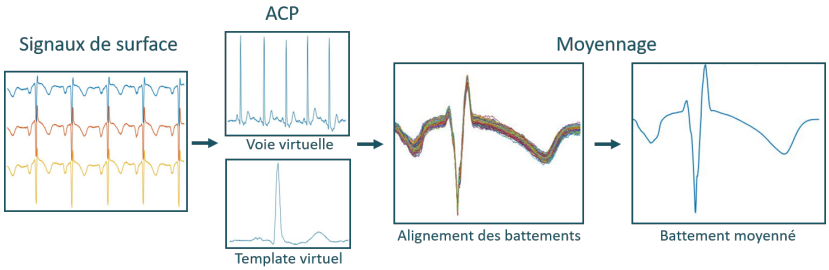


FIGURE 1.15 – **Méthode de moyennage du signal d'un électrocardiogramme**
Technique du SAECG (Signal Averaging ECG). ACP : Analyse en composantes principales
Source : TAN, 2021

— le nombre de patients qui ont effectué un enregistrement avec l'ECG 128 HD (dans le dernier Chapitre, nous détaillerons précisément quels patients sont utilisés pour présenter les résultats dans ce manuscrit) : **dimension N**

— le nombre de signaux par patient ; dans le meilleur des cas, les 128 ont enregistré un signal, mais il peut y avoir un nombre de signaux inférieur si des électrodes sont défectueuses ou trop bruitées : **dimension L**

— la durée d'un battement moyenné : **dimension temps**

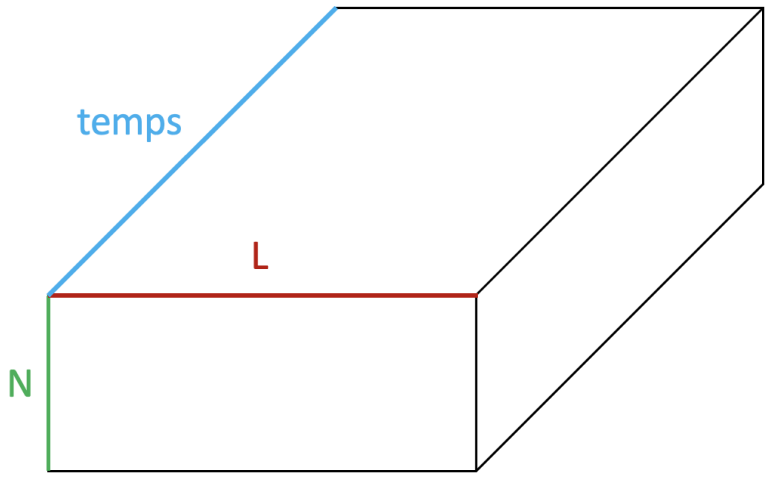


FIGURE 1.16 – **Représentation des données cardiaques sous forme de cube**
Chaque dimension du cube représente une information des données : la dimension N donne le nombre de patients enregistrés, la dimension L donne le nombre de signaux considérés et la dimension "temps" correspond à la durée d'un battement moyenné.

Calcul des marqueurs

En traitement du signal, il est commun d'extraire des marqueurs du signal brut afin d'extraire une information plus ciblée. Dans le travail de thèse de Nolwenn TAN (Tan, 2021), 14 marqueurs liés à l'électrophysiologie cardiaque ont été calculés sur chaque battement moyenné. Dans sa thèse, elle y décrit les processus de calcul de ces marqueurs (tous les marqueurs sont décrits en détails dans son manuscrit de thèse Tan, 2021).

Dans ces marqueurs, il est possible de retrouver des techniques de filtrage (par exemple des marqueurs qui permettent une transformation fréquentielle du signal). Il y a également des marqueurs choisis pour identifier la signature électrique des substrats structurels arythmogènes dans les signaux de surface. Cette identification se fait grâce à des calculs de différence en amplitude, du plus haut pic et du plus bas pic, de durée du complexe QRS, de variabilité des durées des QRS ou encore grâce des paramètres qui analysent la forme du signal (asymétrie et étalement du signal).

Avec ce calcul de marqueurs, la dimension du cube définie par le temps du battement moyenné est maintenant réduite à la dimension $M = 14$, ce qui nous donne le cube présenté en Figure 1.17.

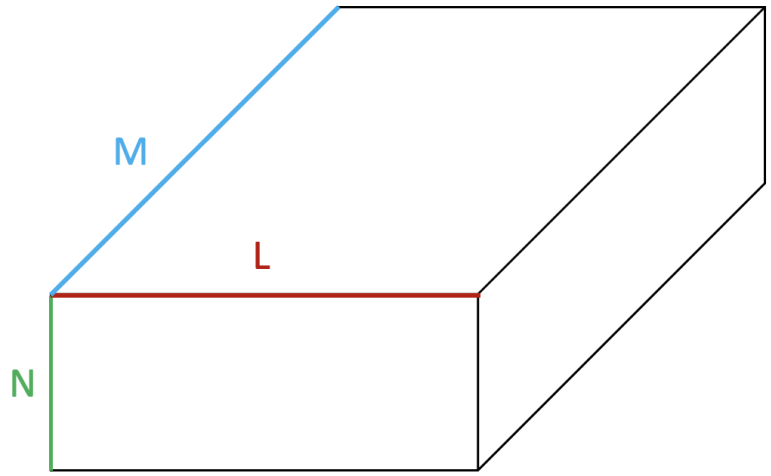


Figure 1.17 – **Cube après extraction de 14 marqueurs sur les battements moyennés** Chaque dimension du cube représente une information des données : la hauteur N donne le nombre de patients enregistrés, la largeur L donne le nombre de signaux considérés et la profondeur correspond au nombre de marqueurs calculés sur le battement moyenné.

Les 14 marqueurs sont calculés sur différents types de signaux. L'avantage d'analyser différents types de signaux est de mieux comprendre l'activité électrique cardiaque et d'accéder à une information électrique qui, parfois invisible sur un signal d'intérêt, apparaît sur un autre.

Les différents types de signaux

Les premiers signaux présentés sont appelés signaux unipolaires. Ils sont calculés par la différence entre l'électrode de référence (WCT de la Figure 1.5) et l'électrode d'intérêt. Cette électrode de référence, de par son placement, est lointaine de certaines électrodes. Ainsi les signaux unipolaires peuvent enregistrer une activité lointaine puisque l'unipolaire mesure la différence de potentiel entre deux électrodes potentiellement éloignées. Les signaux unipolaires enregistrent donc un fonctionnement assez large du cœur.

Pour avoir une meilleure spécificité dans l'espace, des différences de potentiels locales sont utilisées pour pouvoir analyser l'activité électrique locale du cœur, en soustrayant deux électrodes proches l'une de l'autre. Un signal bipolaire (vertical ou horizontal) possède un meilleur rapport "signal sur bruit" que les signaux unipolaires et permet ainsi de mieux amplifier l'activité locale du cœur.

Les signaux laplaciens mettent également en évidence l'information locale d'une activité électrique (McFarland et al., 1997). Comme les signaux bipolaires qui amplifient l'activité locale, les signaux laplaciens amplifient encore plus ces signaux locaux car ils sont calculés sur 9 électrodes au lieu de 2. Ainsi les composantes horizontales/verticales/angulaires du signal sont toutes conservées. Au contraire des signaux bipolaires qui peuvent manquer l'activité qui se situe dans l'une de ces trois directions (c'est pourquoi nous étudions les différences verticales et horizontales). Le problème des laplaciens est qu'en amplifiant l'activité locale ils amplifient également le bruit, ce qui signifie qu'il est parfois difficile de distinguer le signal réel du bruit.

Ces quatre types de signaux sont présentés dans la Figure 1.18). Il y a un signal unipolaire par électrode, donc 128 signaux unipolaires si toutes les électrodes ont enregistré le signal. Les signaux bipolaires verticaux correspondent à la soustraction entre les potentiels de deux électrodes adjacentes, dans le plan vertical, comme par exemple le potentiel de l'électrode 57 et le potentiel de l'électrode 58. Au total, 99 signaux bipolaires verticaux sont créés. Les signaux bipolaires horizontaux correspondent à la soustraction entre les potentiels de deux électrodes adjacentes, dans le plan horizontal, comme par exemple le potentiel de l'électrode 57 et le potentiel de l'électrode 68. Au total, 99 signaux bipolaires horizontaux sont créés. Un signal laplacien correspond à la soustraction entre le potentiel d'une électrode centrale et la moyenne des potentiels de ses 8 électrodes voisines. Au total, 72 signaux laplaciens sont créés.

Avec ces quatre types de signaux, nous pouvons calculer les mêmes marqueurs sur chacun de ces signaux ; 4 cubes de données sont ainsi obtenus : un cube pour chaque type de signal. Une différence est à noter dans la dimension L de ces cubes (voir Figure 1.19).

1.3.3 La présence de données manquantes

Comme nous l'avons vu, un volet du projet HELP consiste en l'acquisition de données en clinique, en milieu hospitalier, avec les contraintes et aléas techniques qui en résultent. Un problème fréquemment rencontré est lié à la qualité des signaux mesurés par les électrodes. Si une électrode ne fonctionne pas ou si elle est mal collée, alors le signal correspondant

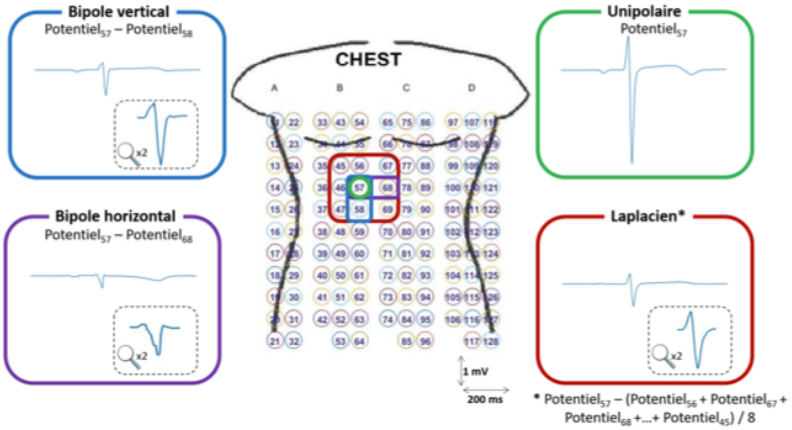


FIGURE 1.18 – **Schéma résumant le calcul de 4 types de signaux.**
Source : TAN, 2021

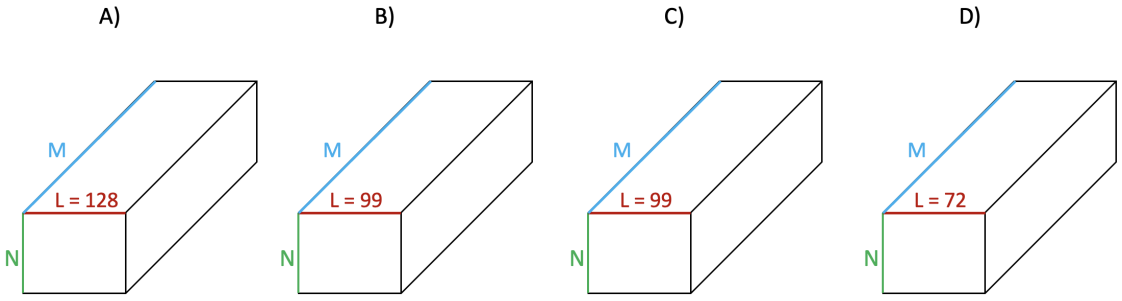


FIGURE 1.19 – **Les quatre cubes de données après extraction des 14 marqueurs**
A) Cube avec signaux unipolaires B) Cube avec signaux bipolaires verticaux C) Cube avec signaux bipolaires horizontaux D) Cube avec signaux laplaciens

n'est pas enregistré ou pas exploitable. Dans ce cas, il n'est pas possible de calculer le battement moyenné à l'électrode considérée, ni les signaux bipolaires et laplaciens utilisant cette électrode. En conséquence, les 14 marqueurs correspondants sont manquants. Cette absence de signal, et donc l'absence des 14 marqueurs, est illustrée dans les cubes de la Figure 1.20 par des lignes pointillées. Prenons l'exemple du cube unipolaire (cube A) dans la Figure 1.20). Si pour le patient $n \in N$, l'électrode $e \in L$ n'a pas enregistré de signal, alors une ligne de coordonnées (n, e) est manquante. Cette ligne se répercute dans les trois autres cubes car l'électrode e est utilisée pour les signaux bipolaires et laplaciens.

Ainsi la base de données cliniques pré-traitées du projet HELP contient des données manquantes. Pour la suite du travail, nous nommerons ces données, enregistrées par l'ECG 128 HD et contenant des données manquantes les **cubes d'origine**. Nous pourrons spécifier lequel des 4 cubes d'origine est considéré en précisant le type de signal. Par exemple, pour évoquer le cube A) de la Figure 1.20 nous utiliserons : **cube unipolaire.**

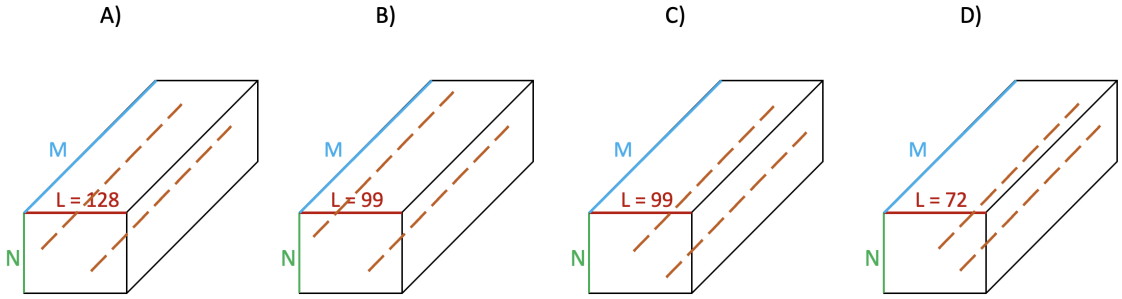


FIGURE 1.20 – **Les 4 cubes de données d'origine associés aux 4 types de signaux avec illustration du signal manquant**

Dans ces cubes sont représentés les signaux manquants à l'aide de lignes pointillées. Quand le signal est manquant, les marqueurs ne peuvent pas être calculés pour le patient considéré. La dimension M du cube est manquante.

1.4 Conclusion

Ce premier chapitre a abordé les bases de l'électrophysiologie cardiaque ; les cardiopathies évoquées dans ce manuscrit ont également été décrites. Nous y avons enfin partagé les grandes lignes du projet HELP, porté par le Pr Michel Haissaguerre ; rappelons que l'objectif du projet HELP est de tester la capacité d'un ECG HD à détecter des arythmies graves pouvant générer une mort subite. Grâce aux enregistrements des patients générés dans le cadre de ce projet, nous avons à notre disposition des données que nous avons présentées, dans la dernière section, sous forme de cubes ; ces cubes d'origine contiennent des données manquantes dues au mauvais fonctionnement de certaines électrodes.

Dans la section suivante, nous présenterons deux méthodes qui permettent de répondre à la présence de données manquantes dans ces données cliniques. Dans un premier temps, nous reviendrons sur le travail fait au cours de la thèse du Dr TAN (TAN, 2021) qui propose une méthode d'agrégation des données. Dans un second temps, nous décrirons la méthode de gestion de données manquantes que nous avons développée au cours de cette thèse : une méthode d'imputation des données manquantes.

Chapitre 2 : Répondre à la présence de données manquantes

Table des matières

Le premier chapitre a introduit la présence de données manquantes dans les données cliniques. Ce chapitre décrit deux méthodes qui permettent de répondre à cette présence des données manquantes. Dans un premier temps, est exposée la méthode proposée par Nolwenn TAN au cours de sa thèse (Tan, 2021), basée sur une technique d'agrégation de données. Dans un second temps nous détaillons la méthode nommée mDAE, développée spécifiquement au cours de cette thèse, qui consiste à imputer les données manquantes par un AE. La méthode mDAE est posée dans un cadre général d'imputation de données manquantes, dans des données tabulaires numériques ; elle s'accompagne de la proposition d'une méthodologie de choix de la meilleure méthode d'imputation pour des nouvelles données possédant des données manquantes. Enfin, nous appliquons la méthode mDAE aux données cliniques de ce projet.

2.1 Agrégation des données

Rappels La base de données peut être représentée sous la forme de 4 **cubes d'origine** (voir Figure 1.20) ; un cube par type de signal (unipolaire, bipolaire vertical, bipolaire horizontal, laplacien) extrait de l'enregistrement ECG 128 HD. Ces 4 cubes diffèrent par le nombre de signaux (dimension L du cube) utilisés pour calculer chaque type de signal. $L = 128$ pour les signaux unipolaires, $L = 99$ pour les signaux bipolaires verticaux et horizontaux, $L = 72$ pour les signaux laplaciens. Nous avons également observé la présence de signaux manquants, liés au mauvais fonctionnement d'une électrode lors de l'enregistrement clinique. Ces signaux manquants rendent impossible le calcul des $M = 14$ marqueurs d'électrophysiologie pour le signal manquant ce qui se traduit par une ligne de données manquantes à l'emplacement (n, l) dans chaque cube pour le patient n dont l'électrode l n'a pas été enregistrée (ligne pointillée rouge dans la Figure 1.20).

Pour compléter les lignes de données manquantes des cubes d'origine, le travail proposé dans le cadre de la thèse de Nolwenn TAN (Tan, 2021) repose sur une agrégation spatiale des électrodes sur le torse. Ce travail sur les signaux a pour but d'extraire le plus finement possible les informations de l'ECG grâce à un découpage en zones du torse, puis un calcul de distributions empiriques et enfin un calcul de paramètres statistiques.

2.1.1 Découpage en zones

La première étape de la procédure d'agrégation des données est le découpage en zones (suivant la dimension L) de chacun des cubes. La localisation des substrats anormaux pour caractériser certaines pathologies cardiaques a conduit à la division du torse en plusieurs zones d'analyse.

Dans cette première étape, chacun des 4 cubes associé à un type de signal est divisé selon 5 zones. Ces zones correspondent à des zones du torse (voir Figure 2.1) : la zone 1 correspond au torse entier ; les zones de 2 à 5 correspondent à 4 zones du torse, découpées comme dans la Figure 2.1 à droite : supérieure droite et gauche, inférieure droite et gauche.

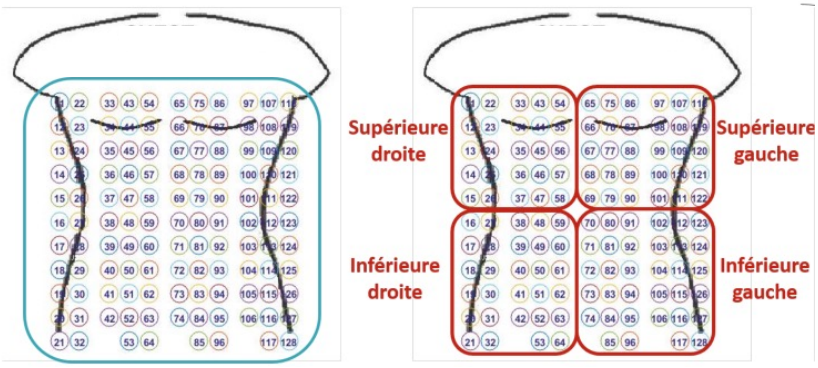


FIGURE 2.1 – **Présentation du découpage en 5 zones du torse**

De cette manière, si l'on considère le **cube unipolaire**, il est transformé en 2 nouveaux cubes. Le premier nouveau cube (à gauche sur la Figure 2.2) représente le cube avec les signaux de tout le torse (zone 1, on considère les 128 signaux ($L = 128$) dans cette zone. Le deuxième nouveau cube (à droite sur la Figure 2.2) représente le cube avec les 4 autres zones du torse (zones 2 à 5). La zone 2 couvre 30 signaux. La zone 3 couvre 34 signaux. La zone 4 couvre 25 signaux. Enfin, la zone 5 couvre 29 signaux. Cette procédure est effectuée sur les 4 **cubes**. Ce découpage en régions permet d'analyser les données sous forme de distribution.

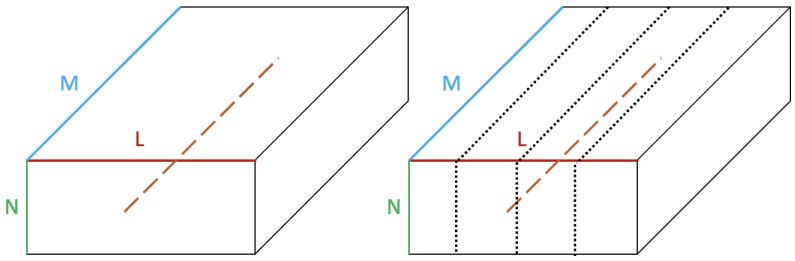


FIGURE 2.2 – **Deux nouveaux cubes unipolaires représentant les 5 zones du torse**
Le cube de gauche représente la zone du torse entier, il comprend tous les signaux, $L = 128$. Le cube de droite est découpé selon 4 autres zones du torse (supérieure droite et gauche, inférieure droite et gauche), avec respectivement $L = 25$, $L = 30$, $L = 29$, $L = 34$. La ligne pointillée rouge représente la dimension M manquante due au mauvais fonctionnement d'une électrode.

2.1.2 Caractérisation sous forme de distribution

Les deux cubes de la Figure 2.2 sont décrits par une dimension $M = 14$, qui correspond aux 14 marqueurs extraits du signal moyenné. Pour plus de clarté, dans les prochaines sections, on fusionne les deux nouveaux cubes représentant les 5 zones de découpage en un seul cube (voir Figure 2.3). A partir du découpage en région des **cubes**, pour un marqueur fixé m d'un patient n, une distribution empirique le long des signaux de chaque zone est

calculée. Un histogramme représente la distribution des valeurs du marqueur m pour le patient n sur sa zone spatiale étudiée. Si l'électrode $l \in L$ n'a pas enregistré le signal cardiaque et que cette électrode fait partie de la zone 3, la distribution ne sera pas calculée sur 30 électrodes, elle ne le sera que sur 29. Cette procédure est effectuée sur les 4 **cubes**. Grâce au calcul des distributions empiriques, il a été permis de s'affranchir de la présence de signal manquant. Dans le but de décrire ces distributions numériquement, des paramètres statistiques sont calculés.

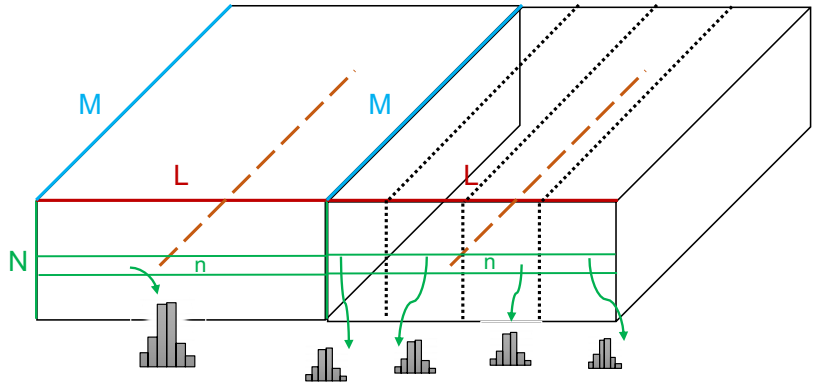


FIGURE 2.3 – **Cube unipolaire avec calcul des distributions**
La ligne pointillée rouge représente le signal électrique absent, dû au mauvais fonctionnement d'une électrode. Une distribution est calculée pour chaque marqueur, de chaque patient, de chaque région.

2.1.3 Calcul des paramètres statistiques

Nous avons vu précédemment que chaque marqueur m de chaque patient n est décrit par une distribution sur un certain nombre de signaux (selon la zone du torse considérée). La dernière étape de cette agrégation du cube de données proposée par Nolwenn TAN (TAN, 2021) consiste à décrire numériquement chaque histogramme par les 6 paramètres statistiques suivants :

- la moyenne
- l'écart-type
- la médiane
- l'interquartile
- la valeur du quantile 0.05
- la valeur du quantile 0.95

En calculant les paramètres statistiques sur les distributions empiriques, les cubes de données sont de la forme présentée dans la Figure 2.4. Chaque marqueur m de chaque patient n est maintenant décrit par $p = 6$ paramètres statistiques. Cette procédure est effectuée sur les 4 cubes associés aux 4 signaux d'intérêt.

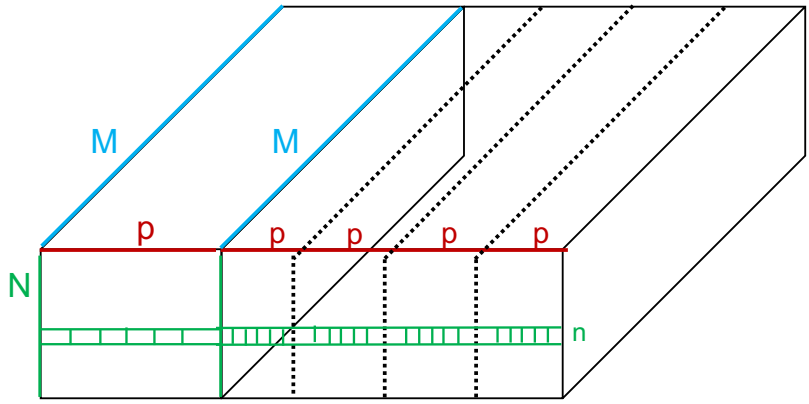


FIGURE 2.4 – **Cube unipolaire avec calcul de $p = 6$ paramètres statistiques**
Les 6 paramètres statistiques sont calculés sur les distributions de chaque marqueur, de chaque patient, de chaque région.

En agrégeant les données spatialement, puis en calculant des paramètres statistiques sur des distributions de marqueurs, il a été permis de s'affranchir de la présence de signaux manquants. La description sous forme de cube des données nous a permis de mieux appréhender toutes les informations contenues dans les données cliniques. Cette représentation en 3 dimensions nous permet de facilement expliquer le passage en une représentation en 2 dimensions des données, qui est la représentation commune des données utilisées en apprentissage statistique.

2.1.4 Représentation matricielle du cube des données agrégées

Dans cette section, nous décrivons le passage des données d'une représentation en 3 dimensions à une représentation en 2 dimensions.

Un **cube** est caractérisé par 3 dimensions : la dimension N donnant le nombre de patients, la dimension M donnant le nombre de marqueurs et la dernière dimension représentant les paramètres statistiques calculés sur 5 régions. Un patient est donc décrit par **14 marqueurs**, chaque marqueur étant décrit par **6 paramètres statistiques calculés sur 5 régions**. Sous forme de matrice, cela donne $14 \times 5 \times 6 = 420$ variables. Les 4 matrices associées aux 4 types de signaux ont donc chacune 420 variables. Dans la Figure 2.5, nous présentons la forme que prend cette représentation matricielle pour un patient.

En présentant les données comme dans la Figure 2.5, chaque variable est le résumé de trois informations : la région, le marqueur, le paramètre statistique. Par exemple, la première variable est le premier paramètre statistique calculé sur le premier marqueur de la première région.

Dans la suite du manuscrit, nous appellerons cette matrice provenant des données agrégées et sans données manquantes, la **la matrice agrégée**. Il y a une **matrice agrégée** par type de signal (par exemple on parle de la matrice agrégée unipolaire). Les 4 matrices ont chacune 420 variables. Le nombre de ligne est à adapter en fonction des patients que l'on

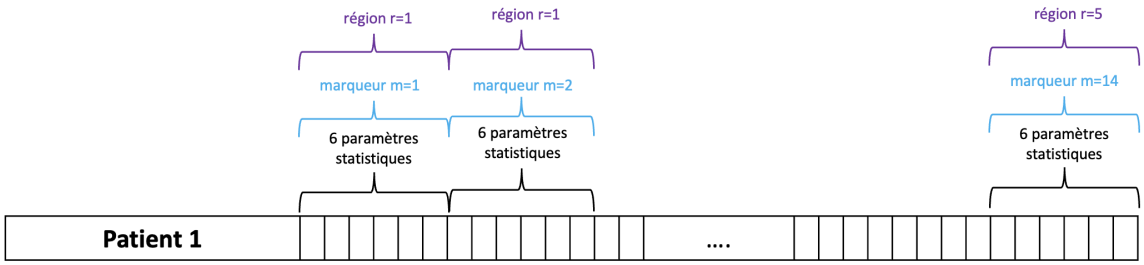


FIGURE 2.5 – **Représentation matricielle d'un cube de données agrégées pour un patient**

souhaite étudier (cette question sera développée dans le Chapitre 3).

Si, dans cette section, il a été possible de s'affranchir des données manquantes en recodant les **cubes d'origine**, nous avons voulu, dans un deuxième temps, utiliser directement ces cubes en gérant les données manquantes de manière directe. Pour résoudre le problème des données manquantes, il n'est pas possible d'enlever des marqueurs sinon les patients ne seraient pas décrits par le même nombre de marqueurs. Il n'est pas non plus possible de supprimer des patients ayant des signaux manquants car presque tous les patients ont au moins une électrode qui n'a pas enregistré de signal. La seule issue afin de travailler sur des **cubes d'origine** est l'imputation des données manquantes.

Dans la section suivante, nous présentons le travail qui a été réalisé au cours de cette thèse pour proposer une nouvelle méthode d'imputation de données manquantes, le mDAE (modified Denoising AutoEncoder). Nous avons développé cette méthode pour répondre à la problématique de l'imputation des données manquantes dans un contexte général. Nous proposons également une méthodologie permettant le choix de la méthode d'imputation la plus adaptée à de nouvelles données tabulaires numériques ayant des données manquantes. Ce travail fait l'objet d'un article actuellement dans un processus de soumission (Mariette DUPUY, Marie CHAVENT et Remi DUBOIS, 2024).

2.2 mDAE : une méthode pour l'imputation des données manquantes

Avec l'augmentation rapide de la collecte de données, la présence de données manquantes constitue un défi majeur dans divers domaines. Les données peuvent être manquantes pour plusieurs raisons. Par exemple, comme dans notre cas, quand la réalité du terrain empêche de réaliser les enregistrements de manière optimale. Il est généralement nécessaire de gérer les données incomplètes avant d'appliquer des méthodes d'apprentissage supervisé. Il y a plusieurs possibilités pour gérer les données manquantes. Une des possibilité est la suppression des lignes ou des colonnes contenant des données manquantes. Cependant, supprimer les lignes ou les colonnes contenant des données manquantes, entraîne une perte considérable d'informations notamment lorsque les données manquantes

sont réparties à plusieurs endroits dans le jeu de données (comme dans nos cubes de données). La deuxième possibilité est l'**imputation des données manquantes**. Elle consiste à remplir les données manquantes avec des valeurs estimées à partir des données présentes.

L'imputation de données est un domaine de recherche actif (Van Buuren, 2018 ; Little et Rubin, 2019), avec plus de 150 implémentations de méthodes disponibles selon Mayer et al. (2021). Cette section se concentre sur les méthodes d'imputation de l'état de l'art. Ces méthodes appartiennent à trois catégories : les méthodes d'apprentissage machine standard, les méthodes d'apprentissage profond et les méthodes de transport optimal. Les méthodes basées sur l'apprentissage machine incluent, entre autres, les k-plus proches voisins (Troyanskaya et al., 2001), la complétion de matrices via SVD à seuil progressif (Mazumder, Hastie et Tibshirani, 2010), l'imputation multivariée par équations en chaîne (Van Buuren et Groothuis-Oudshoorn, 2011) ou MissForest (Stekhoven et Bühlmann, 2012). Les méthodes basées sur l'apprentissage profond incluent, entre autres, les GAN (Generative Adversarial Network) (Goodfellow et al., 2014 ; Yoon, Jordon et Schaar, 2018), les VAE (Variational AutoEncoders) (Kingma et Welling, 2013 ; Ivanov, Figurnov et Vetrov, 2018 ; Mattei et Frellsen, 2019 ; Peis, Ma et Hernández-Lobato, 2022), et les méthodes basées sur des DAE (Denoising AutoEncoders) (voir par exemple la revue de Pereira et al., 2020). On peut également mentionner les travaux récents de Muzellec et al. (2020) et Zhao et al. (2023) basés sur le transport optimal.

Nous proposons dans cette section le "modified Denoising AutoEncodeur" (mDAE), un algorithme basé sur les AE et plus particulièrement les DAE. Cette méthode sera alors comparée sur des bases de données standard aux méthodes de l'état de l'art.

2.2.1 Présentation des AutoEncoders et des Denoising AutoEncoders

Les AutoEncoders

Les AE (Bengio et al., 2009) sont des réseaux de neurones artificiels utilisés pour apprendre une représentation efficace des données via une fonction d'encodage et recréer ces données via une fonction de décodage. Dans ce travail, nous nous concentrons sur le cas particulier des données tabulaires numériques. Pour les données tabulaires, l'entrée d'AE est un ensemble de n observations $\mathbf{x}_1, ..., \mathbf{x}_n$ dans $\mathbb{R}^p$ qui forment les lignes d'une matrice de données $\mathbf{X} = (x_{ij})$, de dimension $n \times p$, où p est le nombre de variables. Nous supposons que ces données ont été standardisées de manière à ce que chaque variable p ait une moyenne nulle et une variance unitaire. Cette standardisation est plus appropriée ici que la normalisation des variables entre 0 et 1, comme cela est souvent fait lors de l'utilisation d'AE.

Supposons un AE "simple", la fonction d'encodage f_θ d'un AE simple (voir Figure 2.6) transforme une entrée $\mathbf{x}_i \in \mathbb{R}^p$ en un vecteur latent $\mathbf{y}_i \in \mathbb{R}^q$:

$$\mathbf{y}_i = f_\theta(\mathbf{x}_i) = s(\mathbf{W}\mathbf{x}_i + \mathbf{b}), \tag{2.1}$$

où $\mathbf{W} \in \mathbb{R}^{q \times p}$ est une matrice de poids, $\mathbf{b} \in \mathbb{R}^p$ est un vecteur de biais et s est une fonc-

tion d'activation (par exemple, ReLU ou sigmoïde). La fonction de décodage $g_{\theta'}$ transforme ensuite le vecteur latent $\mathbf{y}_i \in \mathbb{R}^q$ en une sortie $\mathbf{z}_i \in \mathbb{R}^p$:

$$\mathbf{z}_i = g_{\theta\prime}(\mathbf{y}_i) = s(\mathbf{W}'\mathbf{y}_i + \mathbf{b}'), \tag{2.2}$$

où $\mathbf{W}' \in \mathbb{R}^{p\times q}$ et $\mathbf{b}' \in \mathbb{R}^q$. Ici, la fonction d'activation s dans la couche de sortie doit être la fonction identité, puisque nous essayons de reconstruire des entrées qui prennent leurs valeurs dans $\mathbb{R}$. En effet, la fonction d'activation sigmoïde (resp. ReLu) donne des valeurs de sortie entre 0 et 1 (resp. des valeurs positives), ce qui n'est pas approprié ici.

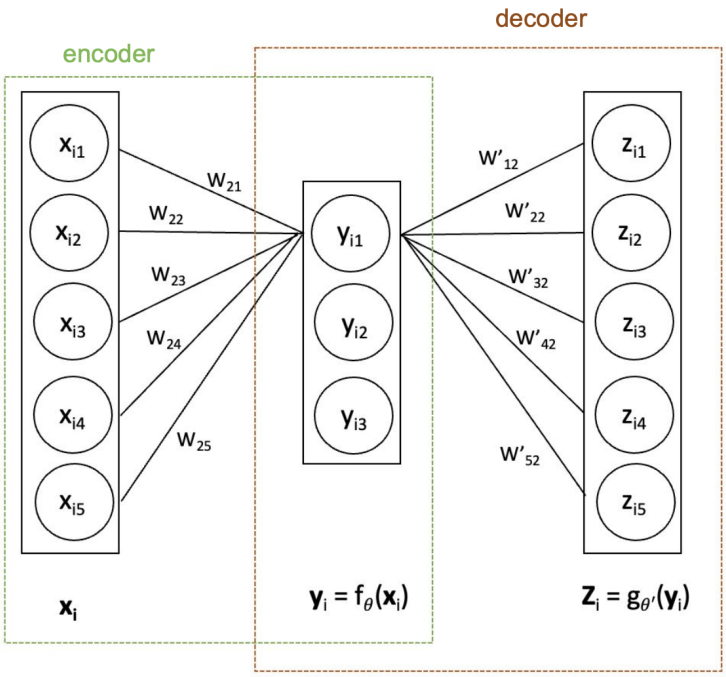


FIGURE 2.6 – **Schéma d'un AutoEncoder simple**

En général, les AE ont plus d'une couche cachée et les paramètres $\theta = (\mathbf{W}_1, ..., \mathbf{W}_K, \mathbf{b}_1, ..., \mathbf{b}_K)$ de l'encodeur et $\theta' = (\mathbf{W}'_1, ..., \mathbf{W}'_K, \mathbf{b}'_1, ..., \mathbf{b}'_K)$ du décodeur sont ceux qui minimisent une fonction de coût. Pour les données numériques standardisées, la fonction de coût, $\mathcal{L}_{AE}$, qui calcule l'erreur quadratique de reconstruction est souvent utilisée. Elle se base souvent sur la norme L_2 et est définie par :

$$\mathcal{L}_{AE} = \sum_{i=1}^{n} \|\mathbf{x}_i - (g_{\theta'} \circ f_\theta)(\mathbf{x}_i)\|_2^2 = \|\mathbf{X} - \mathbf{Z}\|_F^2, \tag{2.3}$$

où $\|\mathbf{X} - \mathbf{Z}\|_F$ est la norme de Frobenius entre la matrice de données $\mathbf{X}$ et sa matrice reconstruite $\mathbf{Z}$. Notez que ce critère favoriserait la reconstruction des variables (colonnes de $\mathbf{X}$) avec une variance élevée, c'est pourquoi la matrice de données $\mathbf{X}$ est standardisée.

Les Denoising AutoEncoders

Les DAE (VINCENT et al., 2008) sont des AE proposés pour l'extraction de variables robustes à un bruit ajouté artificiellement, en apprentissage profond VINCENT et al. (2008).

Pour ce faire, un AE est entraîné à reproduire les données originales à partir des données volontairement bruitées. Cette corruption peut être, par exemple, définie comme un bruit où chaque observation $\mathbf{x}_i$ est corrompue en mettant aléatoirement une proportion μ de ses valeurs à zéro. Soit $N(\mathbf{x}_i)$ le vecteur $\mathbf{x}_i$ bruité. La fonction de coût $\mathcal{L}_{DAE}$ est ici légèrement différente de l'équation (2.3) car elle compare l'entrée $\mathbf{x}_i$ avec la sortie $\mathbf{z}_i = (g_{\theta'} \circ f_\theta)(N(\mathbf{x}_i))$ obtenue avec des observations bruitées $N(\mathbf{x}_i)$ (voir Figure 2.7). La fonction de coût $\mathcal{L}_{DAE}$ (2.3) minimisée par un DAE est alors :

$$\mathcal{L}_{DAE} = \sum_{i=1}^{n} \|\mathbf{x}_i - (g_{\theta'} \circ f_\theta)(N_\mu(\mathbf{x}_i))\|^2 = \|\mathbf{X} - \mathbf{Z}\|_F^2 \tag{2.4}$$

La proportion μ du bruit est un hyperparamètre du modèle qui peut être optimisé.

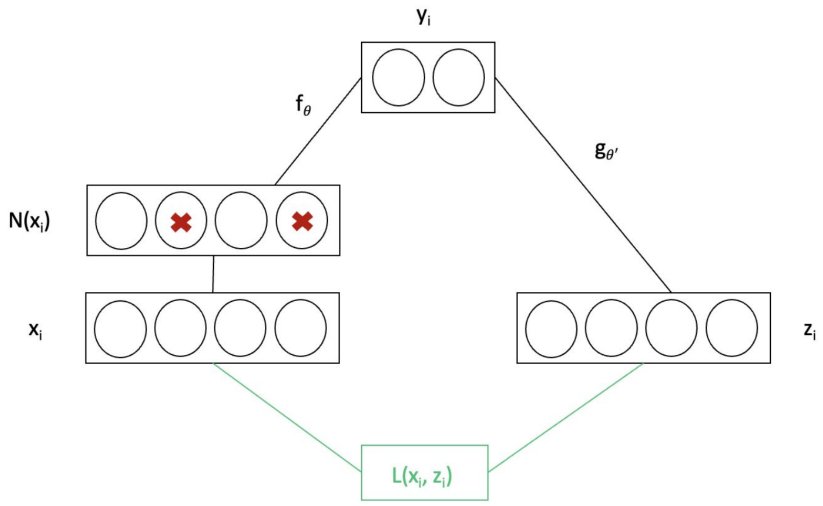


FIGURE 2.7 – **Schéma d'un Denoising AutoEncoder.**
Les croix rouges représentent les valeurs dans $N(x_i)$ aléatoirement mises à 0.

2.2.2 Adaptation de la fonction de coût du DAE pour l'imputation des données manquantes

Les DAE ont également été utilisés pour l'imputation de données manquantes (voir par exemple la revue de PEREIRA et al., 2020). En effet, les DAE ayant été définis pour reconstruire des données bruitées, ils peuvent être adaptés à la reconstruction de données manquantes, ces dernières étant alors considérées comme du bruit.

Comme le souligne PEREIRA et al., 2020 dans son article de synthèse, la méthode classique des travaux utilisant les DAE pour imputer les données manquantes consiste à remplacer chaque donnée manquante par la moyenne de cette variable calculée sur l'ensemble des observations pour lesquelles la variable est disponible.

Soit $\mathbf{X}$ la matrice de données incomplètes (i.e la matrice de données avec des valeurs manquantes) standardisée. La pré-imputation de $\mathbf{X}$ par la moyenne de chaque variable

consiste à remplacer les valeurs manquantes par 0 puisque les données sont standardisées La matrice de données pré-imputée $\tilde{\mathbf{X}}$ s'écrit alors comme la projection de $\mathbf{X}$ sur les entrées présentes :

$$\tilde{\mathbf{X}} = P_\Omega(\mathbf{X}) = \begin{cases} x_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{if } (i,j) \notin \Omega. \end{cases} \tag{2.5}$$

où Ω est l'ensemble des indices $(i,j) \in \{1,...,n\} \times \{1,...,p\}$ où les valeurs x_{ij} sont présentes. Le DAE est ensuite entraîné à reconstruire la matrice de données pré-imputée $\tilde{\mathbf{X}}$ en minimisant la fonction de coût (2.4) :

$$\mathcal{L}_{DAE} = \sum_{i=1}^{n} \|\tilde{\mathbf{x}}_i - (g_{\theta'} \circ f_\theta)(N(\mathbf{x}_i))\|^2 = \|P_\Omega(\mathbf{X}) - \mathbf{Z}\|_F^2, \tag{2.6}$$

où $\mathbf{Z}$ est la reconstruction de la matrice pré-imputée $\tilde{\mathbf{X}}$. Les valeurs manquantes dans $\mathbf{X}$ sont remplacées par celles reconstruites dans $\mathbf{Z}$ (Figure 2.8) et la matrice des données imputées est :

$$\hat{\mathbf{X}} = P_\Omega(\mathbf{X}) + P_{\Omega^\perp}(\mathbf{Z}), \tag{2.7}$$

où $\Omega^\perp$ est l'ensemble des indices $(i,j) \in \{1,...,n\} \times \{1,...,p\}$ où x_{ij} est manquant.

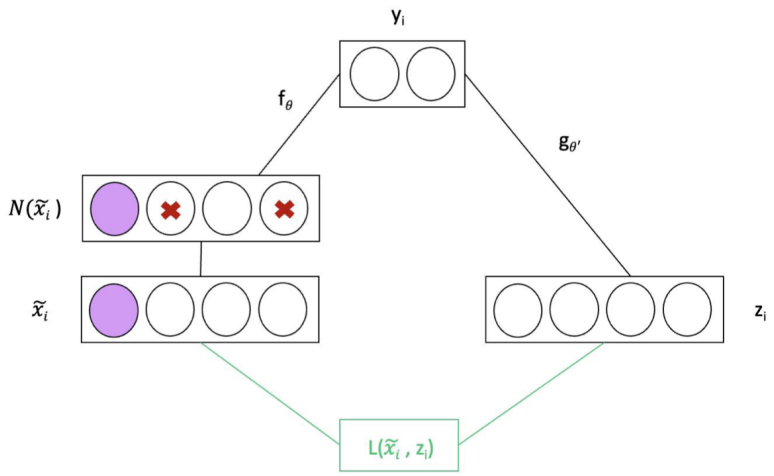


FIGURE 2.8 – **Schéma d'un DAE directement appliqué sur des données pré-imputées.** Les points violets dans x_i représentent les données manquantes mises à 0. Les croix rouges représentent les valeurs dans $N(x_i)$ aléatoirement mises à 0.

Si l'utilisation d'une matrice pré-imputée $\tilde{\mathbf{X}}$ résout le problème de la fonction de coût qui ne peut pas gérer la présence de valeurs manquantes, la minimisation de la fonction de coût (2.6) entraîne le DAE à reconstruire des zéros aux emplacements des valeurs manquantes, ce qui n'est pas pertinent. Notre proposition est d'appliquer un DAE à la matrice de données pré-imputée, comme dans les travaux précédents, mais aussi de modifier l'erreur de reconstruction (2.6) pour ne pas estimer le coût de reconstruction sur les données manquantes (voir la Figure 2.9). Cette méthode, appelée par la suite mDAE, applique un DAE sur des données standardisées et pré-imputées, et optimise le réseau avec la fonction de coût suivante :

$$\mathcal{L}_{mDAE} = \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{Z})\|_F^2, \tag{2.8}$$

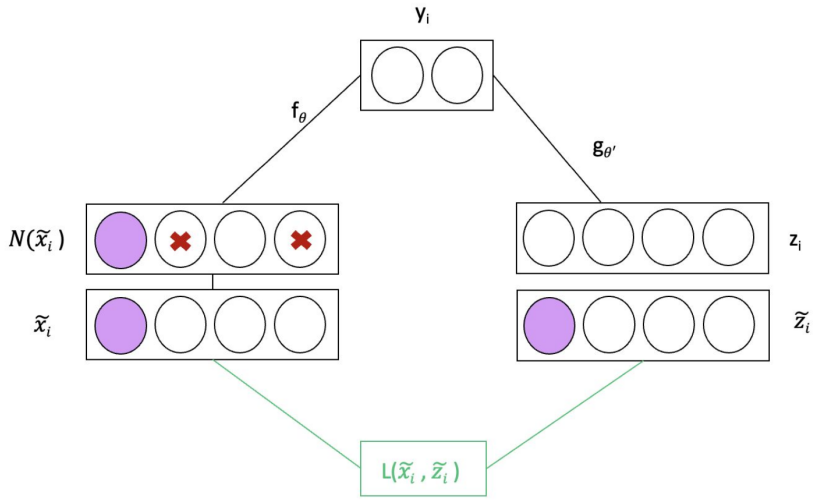


FIGURE 2.9 – **Schéma du modified Denoising AutoEncoder.**
Les points violets dans x_i représentent les données manquantes mises à 0. Les croix rouges représentent les valeurs dans $N(x_i)$ aléatoirement mises à 0.

Dans la section 2.2.3, nous montrerons grâce à une étude des hyperparamètres que l'utilisation de cette fonction de coût modifiée dans la méthodologie mDAE permet une meilleure reconstruction des valeurs manquantes par rapport à l'utilisation de la fonction de coût non modifiée sur des données pré-imputées. Cela nous permet de montrer ainsi la contribution de la méthode mDAE par rapport aux méthodes d'imputation précédentes basées sur les DAE.

Choix de l'hyperparamètre μ

L'hyperparamètre μ de la méthodologie mDAE pour l'imputation de données manquantes est la proportion de zéros utilisée pour corrompre les données avec du bruit (croix rouges dans $N(\tilde{\mathbf{x}}_i)$ dans la Figure 2.9). Cet hyperparamètre peut être choisi aléatoirement dans une grille de valeurs μ dans l'intervalle $[0, 1]$. Alternativement, il peut être choisi dans cette même grille par une procédure optimisée afin de minimiser l'erreur de reconstruction des valeurs manquantes.

Ainsi, les valeurs présentes (non-manquantes) des données d'origine sont réparties en deux ensembles : un ensemble d'entraînement pour optimiser l'hyperparamètre μ et un ensemble de validation pour estimer l'erreur de reconstruction des valeurs manquantes. Soit $V \subset \Omega$ le sous-ensemble d'indices (i, j) de l'ensemble de validation, tiré aléatoirement dans l'ensemble des valeurs présentes Ω. Pour chaque valeur de μ dans la grille, l'erreur de reconstruction des valeurs manquantes est estimée en utilisant la procédure suivante :

1. Les paramètres du mDAE sont appris sur l'ensemble d'entraînement $\Omega \setminus V$ par minimisation de la fonction de coût :

$$\mathcal{L}_{mDAE} = \|P_{\Omega\setminus V}(\mathbf{X}) - P_{\Omega\setminus V}(\mathbf{Z})\|_F^2, \tag{2.9}$$

   où $\Omega \setminus V$ est l'ensemble des entrées observées sans celles tirées aléatoirement pour la validation.
2. L'erreur quadratique moyenne (MSE) de la reconstruction des valeurs manquantes est estimée sur l'ensemble de validation par :

$$MSE_{val} = \frac{1}{|V|}\|P_V(\mathbf{X}) - P_V(\mathbf{Z})\|_F^2. \tag{2.10}$$

   où $\mathbf{Z}$ est la matrice reconstruite avec le mDAE appris sur l'ensemble d'entraînement $\Omega \setminus V$ et $|V|$ est le cardinal de l'ensemble de validation.

Les deux étapes précédentes sont répétées B fois (pour les B tirages de valeurs manquantes) et la moyenne des erreurs de reconstruction des valeurs manquantes est calculée pour obtenir une estimation plus robuste.

Choix de la structure

Deux familles de structures sont connues dans la famille des AE. Les réseaux undercomplete, où la dimension de l'espace latent (voir Equation 2.1 et Figure 2.6) est plus petite que la couche d'entrée, et les réseaux overcomplete, où la dimension de l'espace latent est de plus grande dimension que la couche d'entrée. Si la structure overcomplete n'est souvent pas pertinente pour l'objectif recherché avec des AE, il a été montré que les structures overcomplete donnent de bons résultats avec les DAE (Pereira et al., 2020). Pour optimiser l'architecture nous proposons une grille de 6 structures simples (deux undercomplete et quatre overcomplete) pour choisir la "meilleure" structure lors de l'utilisation de la méthode mDAE (voir Figure 2.10). Pour chaque structure de cette grille, l'erreur de reconstruction des valeurs manquantes est estimée sur les données de validation, en utilisant la même procédure que pour la sélection de l'hyperparamètre μ (voir section 2.2.2).

Idéalement, l'hyperparamètre μ et la structure devraient être choisis simultanément en considérant de manière exhaustive toutes les combinaisons possibles.

2.2.3 Étude numérique

La première partie de cette étude numérique concerne l'étude des hyperparamètres de la méthode mDAE. Plus précisément, une étude des hyperparamètres est menée pour vérifier la pertinence des choix faits pour construire cette méthode. En outre, Peis, Ma et Hernández-Lobato, 2022 soulignent que la grande majorité des méthodes utilisant des DAE ne présentent pas de justifications quant aux décisions prises pour choisir la structure du DAE et les hyperparamètres. Seulement quelques exceptions utilisent des approches de type "grid-search". Suivant les recommandations de Pereira et al., 2020, nous proposons

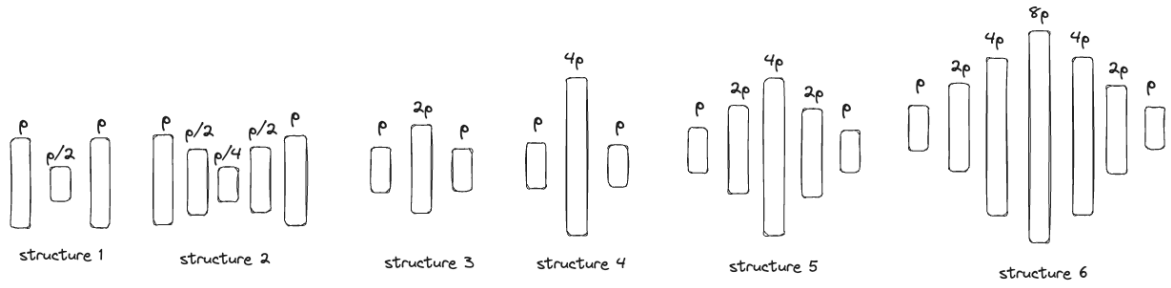


FIGURE 2.10 – **Une grille de six structures simples où p est la dimension de la couche d'entrée.**

une méthodologie générale et reproductible de recherche du choix des hyperparamètres et de la structure. De plus, l'étude des hyperparamètres fournit des recommandations pour la sélection de la structure et des hyperparamètres lorsque l'approche par grille de recherche est trop coûteuse en termes de temps de calcul. L'étude des hyperparamètres permet également de vérifier la pertinence des composants de la méthode mDAE (choix de l'hyperparamètre par validation croisée, structure du réseau). L'importance de chaque composant est évaluée à l'aide de la Root Mean Square Error (RMSE) de la reconstruction des valeurs manquantes artificiellement ajoutées à 7 ensembles de données provenant de l'UCI Machine Learning Repository.

Après cette étude des hyperparamètres, la méthode mDAE est comparée à huit autres méthodes d'imputation (quatre basées sur l'apprentissage machine et quatre basées sur l'apprentissage profond et le transport optimal) ainsi qu'à la méthode de référence de l'imputation par la moyenne. Les dix méthodes d'imputation sont comparées en utilisant de nouveau le RMSE de reconstruction des valeurs manquantes ajoutées artificiellement aux sept ensembles de données utilisés dans l'étude des hyperparamètres.

Toutes les comparaisons sont effectuées à l'aide de 7 jeux de données tabulaires complets (sans valeurs manquantes) choisis parmi les 23 jeux de données du UCI Machine Learning Repository récemment utilisés par MUZELLEC et al., 2020 pour comparer les méthodes d'imputation. Ces 7 jeux de données (voir le tableau 2.1) ont été choisis pour être tous numériques (car la méthode mDAE ne convient qu'aux valeurs manquantes numériques), de tailles différentes et pas trop nombreux (pour éviter de trop longs temps de calculs). À notre connaissance, il n'existe pas un groupe de jeux de données de référence pour l'imputation des données manquantes. Les 26 articles étudiés dans l'article de synthèse de PEREIRA et al., 2020 utilisent presque tous des groupes de jeux de données différents.

Pour évaluer chaque méthode d'imputation, une certaine proportion de chaque jeu de données est d'abord artificiellement remplacée par des valeurs manquantes. Les valeurs manquantes artificielles sont tirées en utilisant soit le mécanisme MAR (Missing At Random), soit le mécanisme MCAR (Missing Completely At Random) ou le mécanisme MNAR (Missing Not At Random) (voir par exemple RUBIN, 1976). Il est à noter que les valeurs manquantes MCAR et MAR ont été générées en utilisant un modèle de masquage logistique tel qu'implémenté dans le dépôt GitHub de MUZELLEC, s. d. Ensuite, pour un masque donné $\Omega^{\perp}$ de valeurs manquantes artificielles, la performance de la méthode est évaluée en utilisant

le RMSE entre la matrice de données initiale $\mathbf{X}$ et la matrice de données reconstruite $\mathbf{Z}$ sur $\Omega^{\perp}$:

$$RMSE = \sqrt{\frac{1}{|\Omega^{\perp}|}\|P_{\Omega^{\perp}}(\mathbf{X}) - P_{\Omega^{\perp}}(\mathbf{Z})\|_F^2}, \tag{2.11}$$

où $|\Omega^{\perp}|$ est le nombre de valeurs manquantes artificielles. Pour obtenir des résultats plus robustes, le processus est répété B fois avec B ensembles de valeurs manquantes artificielles tirées aléatoirement en utilisant l'un des trois mécanismes de génération. Enfin, une méthode est évaluée par la moyenne et l'écart-type des B valeurs de RMSE obtenues avec une certaine proportion de données manquantes artificielles et un certain mécanisme de valeurs manquantes (MAR, MCAR ou MNAR). Il est à noter que tous les jeux de données sont standardisés (c'est-à-dire que toutes les caractéristiques sont centrées et mises à l'échelle pour avoir une variance unitaire) avant de procéder aux expériences, afin de donner le même poids à toutes les caractéristiques dans les analyses. Tous les résultats présentés dans cette section sont reproductibles à l'aide du code Python, qui sera disponible sur GitHub.

Noms	**Abréviations**	**Lignes**	**Colonnes**
Breast cancer diagnostic	breast	509	30
Connectionist bench sonar	sonar	208	60
Ionosphere	iono	351	34
Blood transfusion	blood	748	4
Seeds	seeds	210	7
Climate model crashes	climate	540	18
Wine quality red	wine	1599	10

TABLE 2.1 – **Les 7 jeux de données utilisés dans l'étude numérique**

Etude des hyperparamètres de la méthode mDAE

Lors de l'étude des hyperparamètres de la méthode mDAE, on évalue l'importance des différents hyperparamètres d'une méthode, en comparant les résultats obtenus avec et sans cet hyperparamètre. Ici, les hyperparamètres suivants sont étudiés :

- l'utilisation de la fonction de coût modifiée (2.8) plutôt que la fonction de coût standard L_2 (2.6),
- l'utilisation d'une valeur optimisée de l'hyperparamètre μ (comme décrit dans la section 2.2.2) plutôt qu'une valeur choisie aléatoirement dans $[0, 1]$,
- l'utilisation d'une structure overcomplete (la 5ème structure dans la Figure 2.10) plutôt qu'une structure sous-paramétrée (la 2ème structure dans la Figure 2.10).

La Table 2.2 montre les résultats de l'étude des hyperparamètres pour les sept jeux de données et 20% de valeurs manquantes artificielles MCAR. La valeur moyenne sur les B ensembles de valeurs manquantes artificielles ($\pm$ l'écart type) du RMSE de reconstruction des valeurs manquantes artificielles est calculée pour chaque ensemble de données avec la méthode mDAE, avec la méthode privée de sa fonction de coût modifiée (c'est-à-dire avec une fonction de coût L_2 standard), avec la méthode privée de son choix optimisé de μ (c'est-à-dire avec un choix aléatoire), avec la méthode privée de sa structure overcomplete (c'est-à-dire avec une structure sous-paramétré). Chaque fois, la coût de qualité de l'imputation (c'est-à-dire l'augmentation du RMSE moyen) est mesurée entre le mDAE sans l'un des trois composants (la coût modifiée, un choix optimisé de μ ou une structure overcomplete) et le mDAE complet. Par exemple, pour le jeu de données sur le cancer du sein, l'utilisation de la coût standard L_2 augmente le RMSE moyen de $46.99\% = \frac{0.685-0.466}{0.466}$.

Method	**breast**	**climate**	**sonar**	**iono**	**seeds**	**wine**	**blood**
mDAE	**0.466 ± 0.016**	1.007 ± 0.007	**0.656 ± 0.007**	**0.776 ± 0.018**	**0.496 ± 0.022**	**0.790 ± 0.030**	**0.701 ± 0.059**
mDAE w/o	0.685 ± 0.036	**1.005 ± 0.008**	0.988 ± 0.013	0.808 ± 0.020	0.587 ± 0.028	0.828 ± 0.034	0.755 ± 0.058
modified loss	(46.996%)	(-0.199%)	(50.610%)	(4.124%)	(18.347%)	(4.810%)	(7.703%)
mDAE w/o	0.501 ± 0.043	1.030 ± 0.013	0.682 ± 0.049	0.802 ± 0.039	0.514 ± 0.054	0.853 ± 0.033	0.710 ± 0.055
optimal μ	(7.511%)	(2.284%)	(3.963%)	(3.351%)	(3.629%)	(7.975%)	(1.284%)
mDAE w/o	0.500 ± 0.011	1.147 ± 0.013	0.699 ± 0.008	0.808 ± 0.025	0.671 ± 0.209	0.932 ± 0.045	0.960 ± 0.140
overcomplete	(7.296%)	(13.903%)	(6.555%)	(4.124%)	(35.282%)	(17.975%)	(36.947%)

TABLE 2.2 – **RMSE moyen de reconstruction ($\pm$ l'écart-type) pour $B = 8$ tirages aléatoires de 20% de valeurs manquantes artificielles MCAR.**
Première ligne : résultats de la méthode mDAE (avec la fonction de coût modifiée, le choix optimal de l'hyperparamètre μ et avec une structure over-complète). Deuxième ligne : résultats sans (w/o) le coût modifié (avec le coût standard L_2 à la place). Troisième ligne : résultats sans (w/o) le choix optimal de μ (avec un choix aléatoire de μ à la place). Quatrième ligne : résultats sans (w/o) structure surcomplète (avec une structure sous-complète à la place). Les résultats entre parenthèses sont le taux de croissance de la RMSE moyenne lorsque la composante considérée est supprimée.

La première ligne du Tableau 2.2 montre que la méthodologie mDAE avec ses trois composants (fonction de coût modifiée, choix optimisé de μ et structure overcomplete 5 de la Figure 2.10) reconstruit toujours mieux les données manquantes, sauf pour les données "climate", pour lesquelles la modification de la fonction de coût n'améliore pas les résultats. Il convient de noter que ce dernier résultat est cohérent avec ceux obtenus par MUZELLEC, s. d. qui ont trouvé, pour le jeu de données "climate" (et 30 % de MCAR), que les 5 méthodes d'imputation comparées dans leur article ne donnaient pas de meilleurs résultats (en termes de RMSE) l'imputation par la moyenne.

La deuxième ligne du Tableau 2.2 montre l'amélioration (en termes de RMSE) lorsque l'on utilise la fonction de coût modifiée plutôt que d'utiliser simplement un DAE sur des données pré-imputées (comme dans les travaux précédents). La non-utilisation de la fonc-

tion de coût modifiée augmente le RMSE pour les jeux de données "breast" et "seeds" jusqu'à 50 %, ce qui montre la contribution de la méthodologie mDAE.

La troisième ligne montre que l'utilisation d'une valeur aléatoire pour l'hyperparamètre μ plutôt que d'une valeur optimisée détériore la qualité de l'imputation pour tous les jeux de données, mais dans une moindre mesure (entre 1 et 8 % d'augmentation du RMSE). Le gain obtenu en choisissant le meilleur μ dans une grille plutôt que de manière aléatoire dans $[0, 1]$ est insignifiant. Ce résultat est important, car il permet à l'utilisateur de choisir aléatoirement l'hyperparamètre μ dans $[0, 1]$ pour économiser du temps de calcul lorsque cela est nécessaire.

La quatrième ligne montre que l'utilisation d'une structure undercomplete plutôt que overcomplete augmente clairement le RMSE d'environ 35 % pour 2 des 7 jeux de données. Le choix de la structure est une question centrale lors de l'utilisation de DAE. Ce résultat peut donc être utilisé pour recommander le choix d'une structure overcomplete et éviter la recherche d'une structure optimale dans une grille. La Figure 2.11 complète les résultats du Tableau 2.2 en examinant les résultats pour les 6 structures différentes de la Figure 2.10. Elle montre qu'ici, les deux structures undercomplete donnent toujours de moins bons résultats que les quatre structures overcomplete.

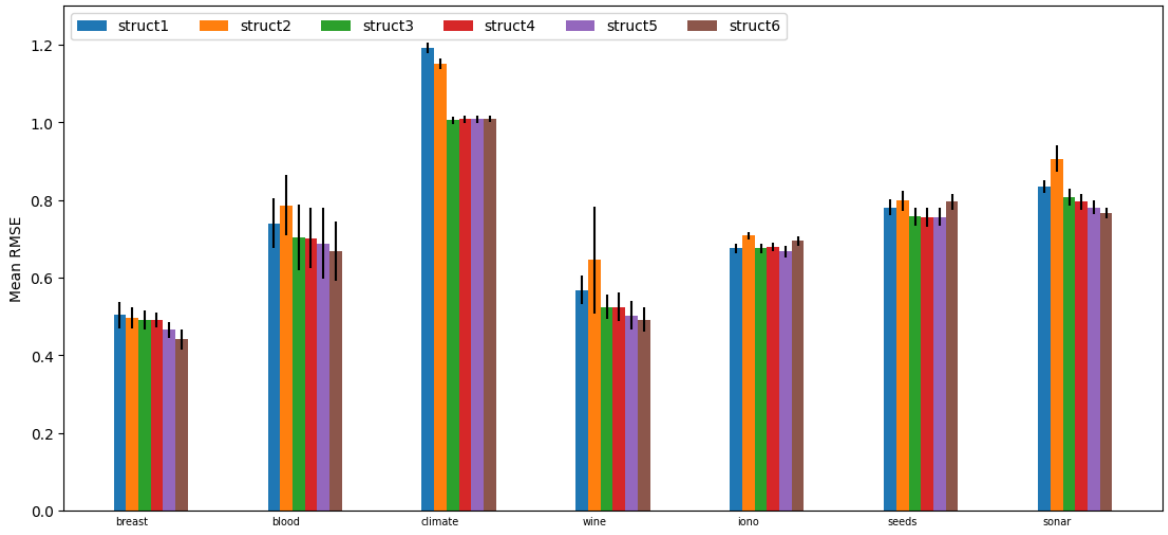


FIGURE 2.11 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 12$ tirages aléatoires de 20% de valeurs manquantes artificielles MCAR et 6 structures différentes de mDAE.**

Les deux premières structures sont undercomplete, les quatre dernières sont overcomplete).

Les résultats supplémentaires de cette étude des hyperparamètres, y compris ceux pour d'autres types de valeurs manquantes artificielles (MAR et MNAR) et d'autres proportions de valeurs manquantes artificielles (20% et 40%), sont fournis en Annexe A. Ces conclusions renforcent l'importance de la modification de la fonction de coût, du choix d'une structure overcomplete, et indiquent l'importance relativement moindre de la recherche par grille par rapport à la sélection aléatoire de l'hyperparamètre μ. Finalement, cette étude a deux

objectifs. Premièrement, démontrer les avantages de l'utilisation de la fonction de coût modifiée. Deuxièmement, il s'agit de montrer l'avantage de choisir une structure "overcomplete" pour le DAE et de vérifier l'avantage de choisir l'hyperparamètre par optimisation dans une grille.

Comparaisons avec les méthodes à l'état de l'art

Cette section compare la méthode mDAE à quatre méthodes relativement classiques de machine learning et à quatre méthodes plus récentes (voir le tableau 2.3) d'apprentissage profond et de transport optimal.

La méthode KNN ((Troyanskaya et al., 2001)) remplace les valeurs manquantes par une moyenne pondérée des k plus proches voisins. La méthode SoftImpute (Mazumder, Hastie et Tibshirani, 2010) utilise une méthode basée sur une SVD itérative à seuils progressifs. Deux méthodes itératives à équations enchaînées (Van Buuren et Groothuis-Oudshoorn, 2011) imputent les données manquantes d'une colonne en utilisant des modèles basés sur les colonnes voisines (elles-mêmes avec des données manquantes) : la méthode missForest (Stekhoven et Bühlmann, 2012) est basée sur les forêts aléatoires et la méthode BayesianRidge est basée sur la régression ridge, pour estimer à chaque étape les fonctions de régression. Les quatre autres méthodes (plus récentes) du tableau 2.3 sont GAIN (Yoon, Jordon et Schaar, 2018) qui est une adaptation des Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) pour imputer les données manquantes, MIWAE Mattei et Frellsen, 2019 qui est une adaptation des Variational AutoEncoders (VAE) (Kingma et Welling, 2013), et deux méthodes utilisant le transport optimal : l'algorithme appelé Batch Sinkhorn Imputation proposé par Muzellec et al. (2020), et la méthode TDM proposée par Zhao et al. (2023).

Names	**Abreviations**
k-nearest neighbors[1]	knn
SoftImpute[2]	si
missForest[3]	rf
BayesianRidge[3]	br
Generative Adversarial Imputation Network[4]	gain
Missing Data Importance Weighted Autoencoders[5]	miwae
Batch Sinkhorn Imputation[2]	skh
Transformed Distribution Matching for missing value imputation[6]	tdm

Table 2.3 – **Méthodes utilisées dans l'étude numérique**

1. Available in the class KNNImputer, https ://scikit-learn.org/stable/api/sklearn.impute.html
2. https ://github.com/BorisMuzellec/MissingDataOT
3. Available in the class IterativeImputer, https ://scikit-learn.org/stable/api/sklearn.impute.html
4. https ://github.com/jsyoon0823/GAIN
5. https ://github.com/pamattei/miwae
6. https ://github.com/hezgit/TDM

Pour KNN et SoftImpute, les hyperparamètres sont sélectionnés par validation croisée. Comme indiqué dans les implémentations utilisées pour les deux méthodes d'équations enchaînées, les hyperparamètres des régressions de la méthode de bayésien ridge sont estimés lors de l'optimisation du modèle. Les hyperparamètres des forêts aléatoires sont 100 arbres, et toutes les variables sont prises en compte lors de la recherche du meilleur "split" (i.e bagged trees). Les hyperparamètres recommandés dans les articles et les implémentations correspondantes sont utilisés pour les quatre dernières méthodes.

Pour la méthode mDAE, les paramètres étudiés dans la section 2.2.3 (choix de μ par validation croisée et de la structure overcomplete 5 de la figure 2.10) sont utilisés. Des paramètres plus justes pour la méthode mDAE auraient consisté à choisir μ et la structure par validation croisée sur toutes les combinaisons de paramètres possibles. Cette approche n'a pas été adoptée pour des raisons de temps de calcul dans cette étude numérique.

Avec ces paramètres, les huit méthodes du tableau 2.3, ainsi que la méthode mDAE et la méthode d'imputation par la moyenne, sont comparées dans la Figure 2.12 sur les 7 jeux de données et 20% des valeurs manquantes artificielles MCAR. La valeur moyenne ($\pm$ l'écart type) du RMSE de reconstruction des valeurs manquantes artificielles est représentée pour chaque ensemble de données et chaque méthode.

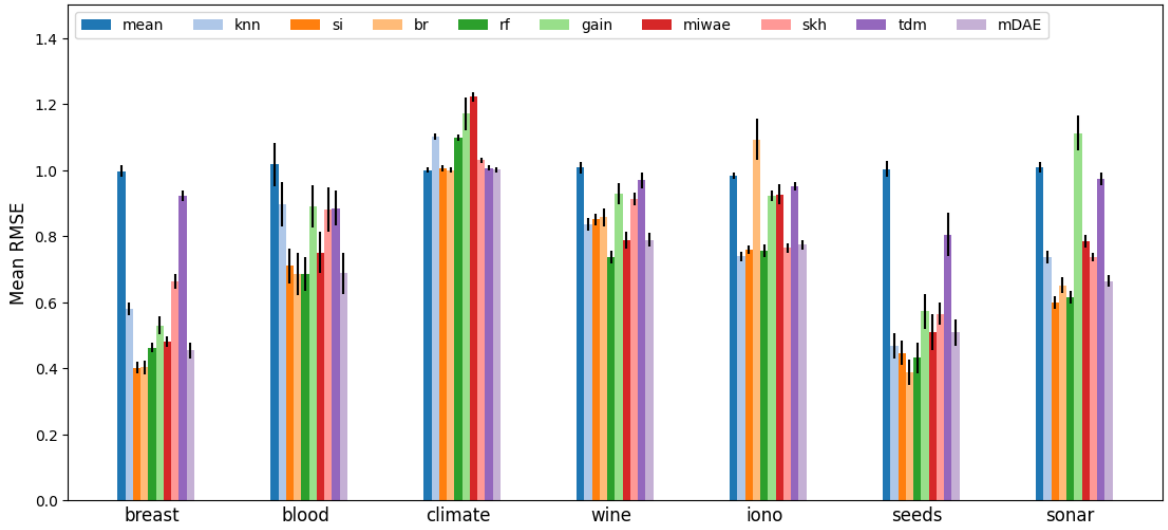


FIGURE 2.12 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 12$ tirages aléatoires de 20% de valeurs manquantes artificielles MCAR.**

Nous remarquons dans la Figure 2.12 que certaines méthodes comme SoftImpute (SI), missForest (RF) ou mDAE fonctionnent raisonnablement bien sur tous les ensembles de données (aucun jeu de données où la valeur RMSE est beaucoup plus mauvaise que les autres). On peut également noter que la méthode mDAE donne des résultats meilleurs ou équivalents sur les 7 ensembles de données que les quatre méthodes basées sur les réseaux neuronaux et le transport optimal (gain, miwae, skh et tdm).

Mais aucune méthode n'est toujours gagnante. Afin de mesurer les performances globales d'une méthode sur plusieurs jeux de données, nous proposons d'utiliser une nouvelle

métrique appelée Mean Distance to the Best (MDB). Cette métrique mesure la performance globale d'une méthode sur tous les jeux de données (pour un pourcentage donné et un mécanisme donné de données manquantes artificielles). Ce critère est défini comme la moyenne (sur les jeux de données) des distances entre le RMSE de la méthode considérée et la RMSE de la meilleure méthode. Il est égal à 0 si le RMSE de la méthode est le meilleur pour tous les jeux de données, et il augmente si le RMSE de la méthode considérée est éloigné du RMSE de la meilleure méthode, en moyenne, sur les ensembles de données. Si I représente le nombre d'ensembles de données et J le nombre de méthodes, le MDB d'une méthode j est définie comme suit :

$$MDB(j) = \frac{1}{I}\sum_{i=1}^{I}\left(R_{ij} - \min_{\ell=1\ldots J} R_{i\ell}\right) \tag{2.12}$$

où R_{ij} est le RMSE obtenu avec la méthode j sur l'ensemble de données i. $MDB(j)$ s'interprète comme la moyenne (sur les ensembles de données) des distances entre le RMSE de la méthode j et le RMSE de la meilleure méthode. Elle est égale à 0 si la méthode j est la meilleure pour tous les ensembles de données. Elle augmente si la qualité de la méthode j est éloignée de la qualité de la meilleure méthode, en moyenne sur les jeux de données.

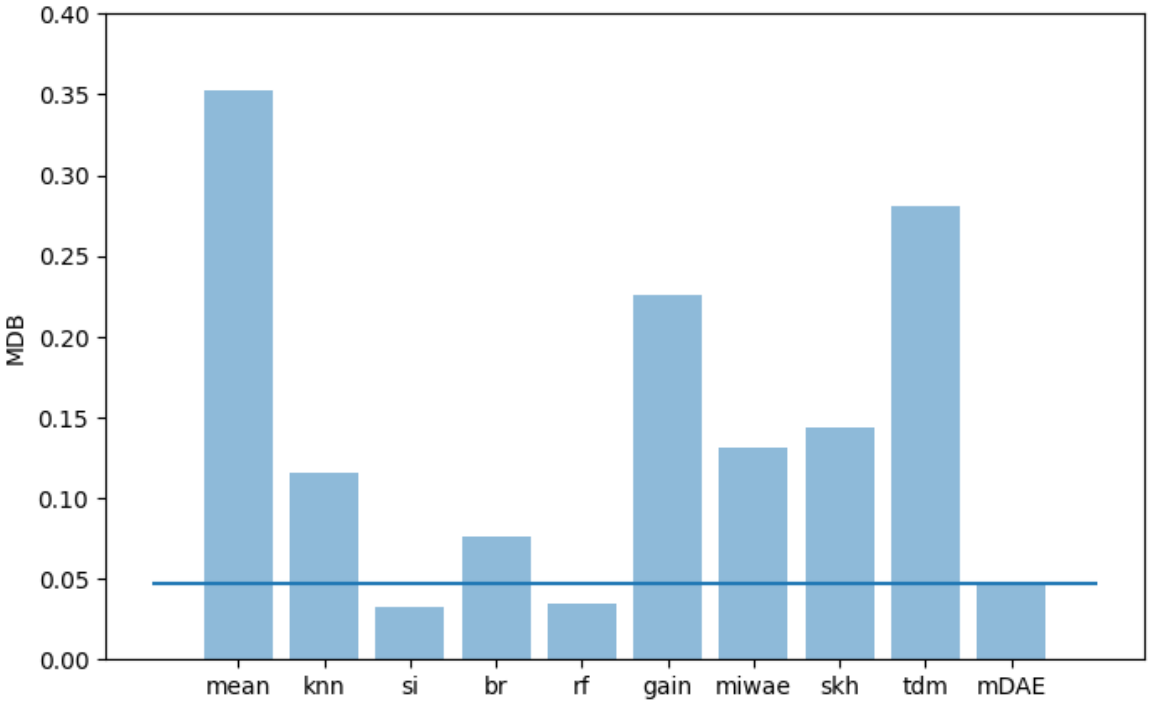


FIGURE 2.13 – **Distance Moyenne au Meilleur (MDB) obtenu avec 20% de valeurs manquantes artificielles MCAR.**

La figure 2.13 montre le MDB obtenu avec 20% de valeurs manquantes MCAR artificielles. Cette figure montre que les deux meilleures méthodes selon ce critère sont SoftImpute (si) et missForest (rf). La méthode mDAE est la troisième meilleure méthode. Les

figures 47, 48 et 49 de l'annexe B montrent les résultats avec 40 % de valeurs manquantes MCAR artificielles, et avec 20 % ou 40 % de valeurs manquantes MAR et MNAR. Avec ces différentes proportions et types de données manquantes, les trois premières restent SoftImpute, missForest et mDAE. SofImpute est toujours en première place, à égalité une fois (40 % MAR) avec mDAE. Les méthodes mDAE et missForest sont généralement l'une ou l'autre en deuxième et troisième position.

Ces résultats confirment la bonne performance des méthodes d'imputation classiques (SoftImpute et missForest) par rapport aux méthodes plus récentes basées sur les réseaux de neurones et le transport optimal (gain, miwae, skh et tdm), ainsi que la bonne performance (comparable à missForest) de la méthode mDAE basée sur les DAE.

Robustesse de la méthode mDAE

Pour étudier la robustesse de la méthode mDAE, nous testons la capacité de la méthode sur une grille de pourcentages de données manquantes. Dans la Figure 2.14, on peut analyser les valeurs moyennes de RMSE pour des proportions de données manquantes allant de 10% à 90%. Cette figure montre que la méthode mDAE reste relativement robuste jusqu'à 70% de données manquantes.

2.2.4 Conclusion

Dans cette section nous proposons une méthode d'imputation des données manquantes, basée sur un DAE, ainsi qu'une procédure de choix des hyperparamètres (la proportion de bruit μ et la structure du réseau). Une étude des hyperparamètres de cette méthode a été réalisée avec différents jeux de données, différents types et proportions de données manquantes. Elle a montré l'amélioration relativement faible des résultats lorsque l'hyperparamètre μ est choisi par validation croisée plutôt que de manière aléatoire. Au contraire, l'utilisation d'un réseau overcomplete plutôt que undercomplete semble appropriée. Nous montrons également dans cette étude que l'utilisation d'une fonction de coût modifiée dans la méthode mDAE permet une meilleure reconstruction des valeurs manquantes que l'utilisation de la fonction de coût non modifiée sur des données pré-imputées. Ceci permet de montrer la contribution de la méthode mDAE par rapport aux précédentes méthodes d'imputation basées sur les DAE.

Puis une étude numérique a comparé la méthode mDAE proposée avec huit autres méthodes d'imputation de valeurs manquantes standards ou récentes. Les résultats ont montré le bon comportement de SofImpute, mDAE et missForest. Un nouveau critère appelé MDB a été utilisé pour comparer globalement les méthodes sur tous les ensembles de données considérés et pour les classer. Si la méthode mDAE proposée donne parfois le meilleur score RMSE (pour un jeu de données donné), ce n'est pas le cas pour tous les ensembles de données. Si l'on considère l'ensemble des jeux de données, le critère MDB classe (pour toutes les configurations considérées dans cette étude numérique) trois méthodes basées sur de l'apprentissage machine et la méthodologie mDAE dans les 4 premières positions. Plus précisément, la méthode mDAE est généralement placée deuxième ou troisième pour

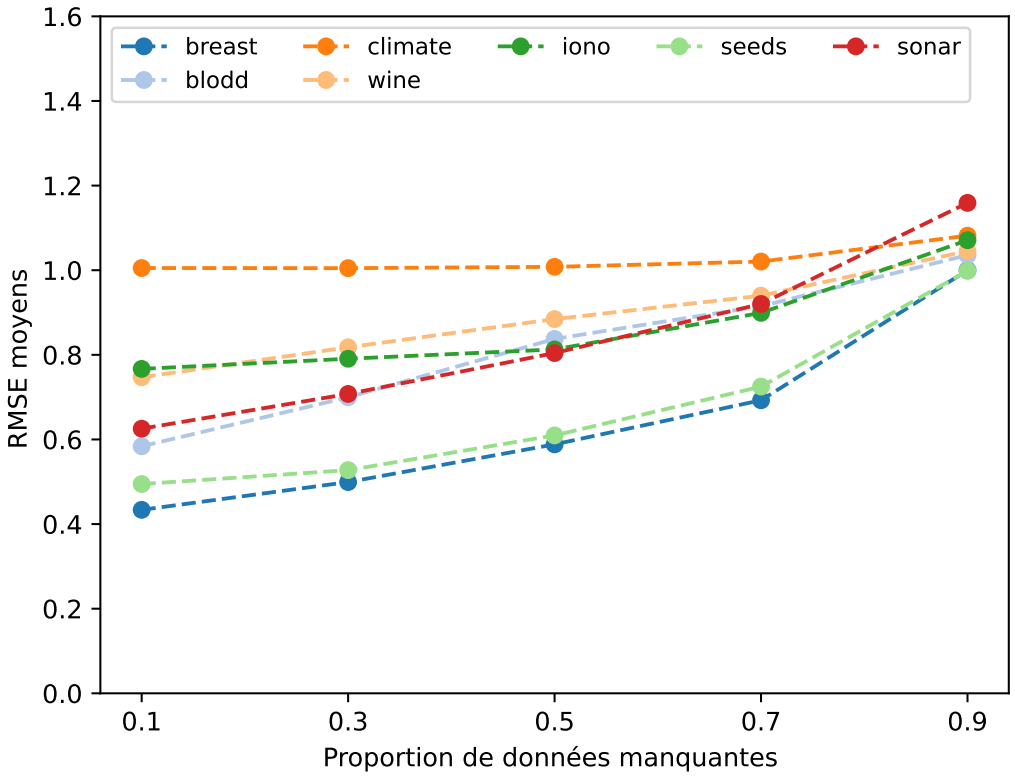


FIGURE 2.14 – **RMSE moyens de la méthode mDAE pour des pourcentages de données manquantes allant de 10% à 90%.**
Le mDAE a été appliqué sur les sept jeux de données.

ce critère (en alternance avec la méthode missForest), la méthode SoftImpute étant toujours classée première. Les quatre méthodes les plus récentes basées sur l'apprentissage profond et le transport optimal se sont systématiquement retrouvées dans les quatre dernières positions pour tous les types et proportions de valeurs manquantes artificielles. Ces méthodes donneraient peut-être de meilleurs résultats avec des données de traitement d'images ou de langage naturel. Le code Python de cette comparaison numérique sera mis à disposition sur GitHub afin qu'il puisse être reproduit avec d'autres jeux de données ou complété par d'autres méthodes.

2.3 Application de la méthode mDAE aux données clinique

Dans la section précédente, nous avons présenté la nouvelle méthode d'imputation de données manquantes que nous avons développée au cours de cette thèse (Mariette DUPUY, Marie CHAVENT et Remi DUBOIS, 2024). En effet, à la fin du Chapitre 1, nous avions montré

que les données cliniques que nous utilisons au cours de cette thèse comportaient des données manquantes ; ces données manquantes surviennent quand une électrode n'enregistre pas le signal cardiaque. Dans la Figure 1.20 nous avions présenté les **cubes d'origine** avec les données manquantes. Dans cette troisième section, nous présentons une étude des hyperparamètres et nous appliquons la méthodologie de comparaison des méthodes à l'état de l'art à la méthode mDAE sur les données cliniques. Notre méthode fonctionnant sur des données matricielles, nous allons présenter le passage de 3 dimensions à 2 dimensions des **cubes** d'origine de la Figure 1.20.

2.3.1 Représentation matricielle du cube des données d'origine

Pour appliquer les méthodes d'imputation de données manquantes présentées dans la section précédente, nous devons expliciter le passage des cubes de données d'origine (Figure 1.20) (4 cubes avec des données manquantes, pour chaque type de signal d'intérêt) en 3 dimensions à une matrice de données en 2 dimensions.

Par exemple, si nous considérons le **cube unipolaire** (voir A) dans la Figure 1.20 pour rappel), pour lequel le nombre de signaux L vaut 128 et le nombre de marqueurs M vaut 14. Alors, pour chaque patient, sur chaque signal unipolaire, 14 marqueurs liés à l'électrophysiologie sont calculés (voir section 1.3.2). Un patient est décrit par $128 \times 14 = 1792$ variables dans le cas du signal unipolaire. On obtient la matrice, pour un patient, en transformant le cube d'origine unipolaire comme présenté dans la Figure 2.15.

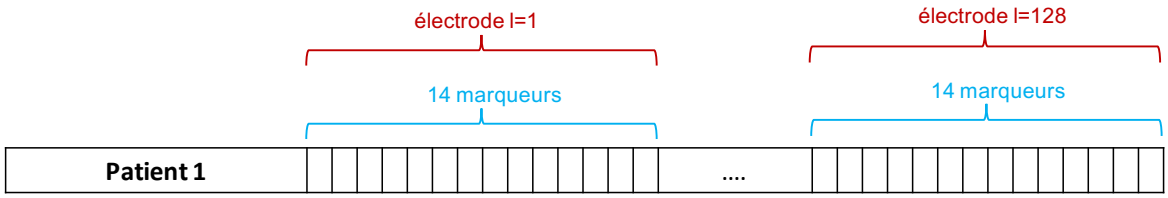


FIGURE 2.15 – **Matrice issue du cube d'origine unipolaire sans donnée manquante** Chaque patient est décrit par 128 signaux unipolaires. Sur chaque signal unipolaire ont été extraits 14 marqueurs.

Cependant, dans le **cube unipolaire**, nous avons vu que nous pouvions trouver des signaux manquants si une électrode était défaillante. La présence d'un signal manquant rendant impossible le calcul des 14 marqueurs, c'est la dimension M entière qui est manquante. Si on considère par exemple que l'électrode 128 n'a pas enregistré de signal cardiaque, la matrice de données avec les données manquantes a la forme présentée dans la Figure 2.16. Avec la Figure 2.16, on peut visualiser les données manquantes sous forme de blocs de taille 14. Dans la suite du manuscrit, nous appellerons la matrice issue des **cubes d'origine** avec des blocs de données manquantes : la **matrice blocs**. Il y a une **matrice blocs** calculée pour chaque type de signaux. Il y a donc 4 **matrices blocs**.

Dans la Figure 2.17, nous avons visualisé la présence des données manquantes en blocs pour la **matrice blocs unipolaire**. Les lignes de cette matrice représentent les patients

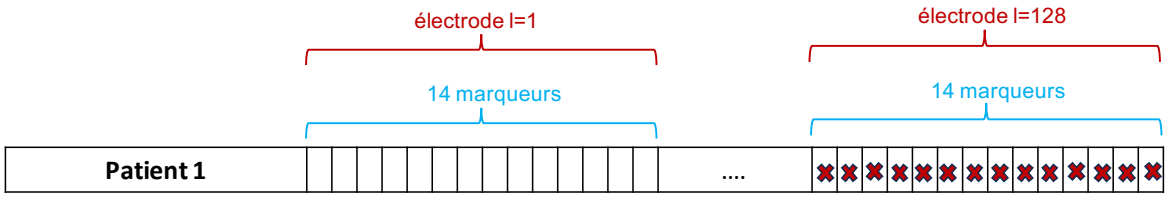


FIGURE 2.16 – **Matrice blocs unipolaire avec affichage des données manquantes**
Chaque patient est décrit par 128 signaux unipolaires. Sur chaque signal unipolaire ont été extraits 14 marqueurs. Si une électrode n'enregistre pas le signal, les 14 marqueurs ne sont pas calculés. Il y a alors 14 valeurs manquantes sous la forme d'un bloc, illustrées par des croix rouges sur la figure.

enregistrés et les colonnes représentent les variables telles qu'elles sont présentées dans la **matrice blocs** de la Figure 2.16. Les données manquantes sont représentées par la couleur verte. Dans cette visualisation on peut voir des grands blocs de données manquantes sur une ligne (encerclés en blanc) : pour ce patient plusieurs électrodes voisines ont cessé de fonctionner, donc les blocs de données manquantes de taille 14 sont côte à côte. Dans la Figure 2.17, on trouve également des colonnes de données manquantes (encerclées en jaune) : pour plusieurs patients, la même électrode n'a pas bien enregistré le signal.

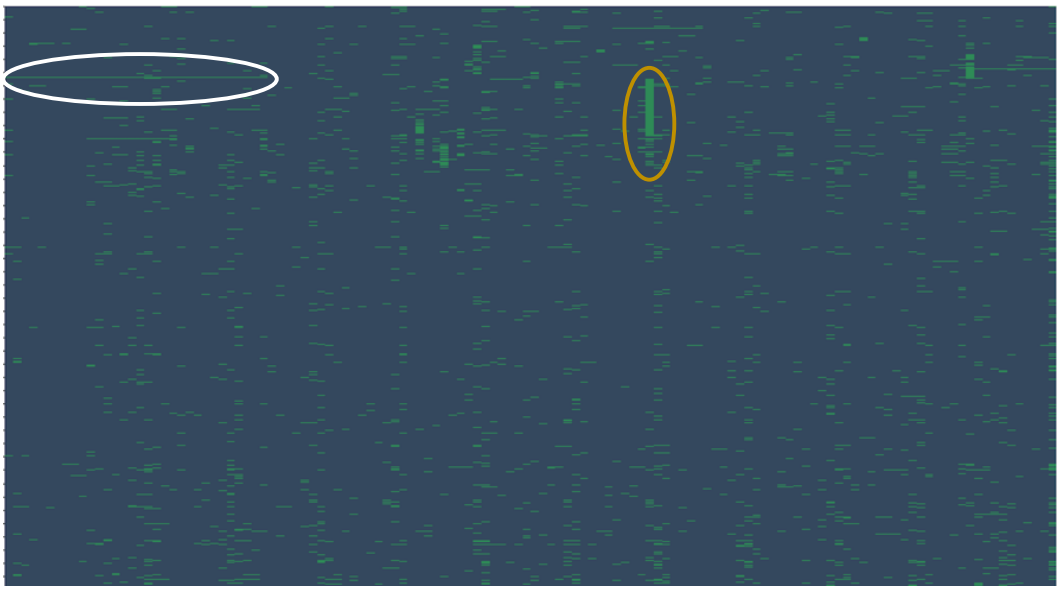


FIGURE 2.17 – **Visualisation des blocs de données manquantes dans la matrice blocs unipolaire**
Les lignes de cette matrice représentent les patients et les colonnes représentent les variables. Les données manquantes sont représentées par la couleur verte.

Dans cette **matrice blocs unipolaire**, il y a 2.8% de données manquantes. Il y a environ 3.5% de données manquantes dans la **matrice blocs bipolaire verticale**, 3.6% dans la **matrice blocs bipolaire horizontale** et 4.9% dans la **matrice blocs laplaciens**.

2.3.2 Imputation de la matrice blocs

Pour mettre en place la stratégie d'imputation des 4 **matrices blocs**, nous avons suivi la méthodologie d'évaluation et de comparaison présentée dans la section précédente (Section 2.2.3).

Dans un premier temps, comme dans la section précédente, nous évaluons les hyperparamètres de la méthode mDAE afin de choisir les bons hyperparamètres adaptés aux données cliniques. Cette évaluation concerne notamment le choix de la structure du DAE, le choix de la forme du bruit et le choix μ (voir section 2.2.2 pour un rappel sur le choix de μ). Dans un deuxième temps, nous avons comparé la méthode mDAE avec les deux méthodes ayant les plus petites valeurs du critère MDB (résultats à retrouver dans la Figure 2.13) de l'étude que nous avons menée dans la section précédente : les méthodes SoftImpute et missForest.

Pour cette étude numérique, 5% des données de chacune des 4 **matrices blocs** sont remplacées par des données manquantes artificielles. Nous avons choisi 5% de données manquantes pour être cohérents avec la quantité réelle de données manquantes que l'on retrouve dans les données cliniques. Pour respecter le cadre des données cliniques, les données manquantes sont positionnées par électrode (les 14 marqueurs d'une électrode sont artificiellement manquants). L'indice de la première colonne du bloc et de la ligne sera choisi aléatoirement tout en respectant que le bloc soit sur une électrode entière. Pour un masque $\Omega^{\perp}$ de données manquantes artificielles, les performances des méthodes sont évaluées à l'aide du RMSE (Eq 2.8) entre la matrice d'origine et la matrice reconstruite sur $\Omega^{\perp}$. Pour obtenir des résultats plus robustes, le processus est répété B fois avec B tirages de blocs de taille 14 de données manquantes. Chaque méthode est évaluée par la moyenne et l'écart-type des B valeurs de RMSE obtenues avec 5% de données manquantes artificielles en bloc.

Etude des hyperparamètres de la méthode mDAE pour les données cliniques

Nous étudions dans cette section, l'influence des différents hyperparamètres sur ces nouvelles données cliniques. L'étude mise en place analyse les effets, sur la qualité de l'imputation, de 3 hyperparamètres de la méthode mDAE : la forme du bruit, la méthode pour choisir la proportion de bruit μ et la structure du DAE.

Comme dans la section sur le choix de la structure de la méthode mDAE (voir section 2.2.2 pour un rappel), nous avons étudié la capacité des 6 structures présentées dans la Figure 2.10 pour imputer les données manquantes.

De plus, dans la section 2.2.2, nous avions présenté le bruit ajouté dans le réseau pour corrompre les données comme des valeurs choisies aléatoirement et mises à 0. (voir 2.2.1). Notons ce **bruit aléatoire**. Cette forme de choix aléatoire de bruit peut être questionnée pour répondre à l'imputation des données manquantes dans ce contexte clinique. En effet, dans les **matrices blocs** que nous devons imputer, la forme des données manquantes est spéciale : des blocs de taille 14. Nous testons donc l'efficacité d'une adaptation de la forme du bruit aux données manquantes présentes à l'origine dans le jeu de données. Ce bruit que

l'on nomme **bruit en blocs** est à l'image des données manquantes : des blocs de valeur 0 et de taille 14. La position (indice de la ligne et de la première colonne) de chaque bloc est choisie aléatoirement. Chaque bloc du bruit correspond exactement aux 14 marqueurs d'une électrode.

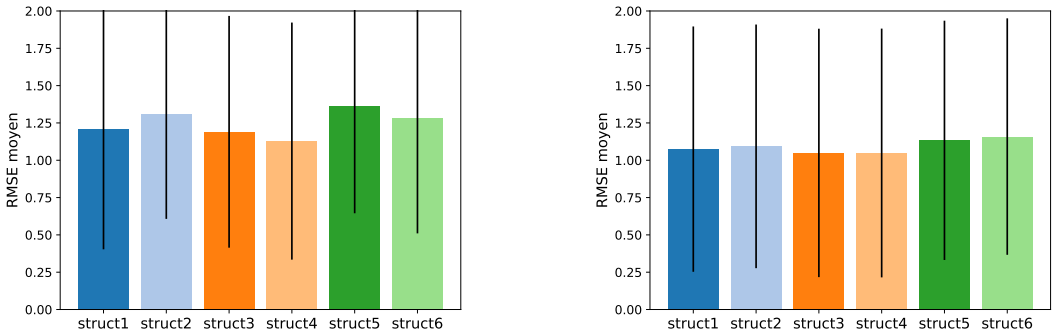
Enfin, nous étudions également deux méthodes de choix de la proportion de bruit μ à ajouter au jeu de données, comme dans la section 2.2.2.

Pour résumer, nous regardons les résultats des 4 expérimentations suivantes :

— première expérimentation : étude des 6 structures avec le bruit aléatoire et le choix optimisé de μ
— deuxième expérimentation : étude des 6 structures avec le bruit aléatoire et le choix aléatoire de μ
— troisième expérimentation : étude des 6 structures avec le bruit bloc et le choix optimisé de μ
— quatrième expérimentation : étude des 6 structures avec le bruit bloc et le choix aléatoire de μ

Il est important de préciser que les 4 expérimentations sont comparables car elles ont toutes été testées sur les mêmes 8 tirages de données manquantes artificielles en bloc.

La Figure 2.18 montre les résultats des première et deuxième expérimentations. Pour chaque expérimentation, nous affichons les RMSE moyens de la reconstruction des valeurs manquantes artificielles pour $B = 8$ tirages de données manquantes. La procédure d'évaluation est la même que celle décrite dans la section 2.2.2.



(a) Bruit aléatoire avec choix aléatoire de la proportion μ

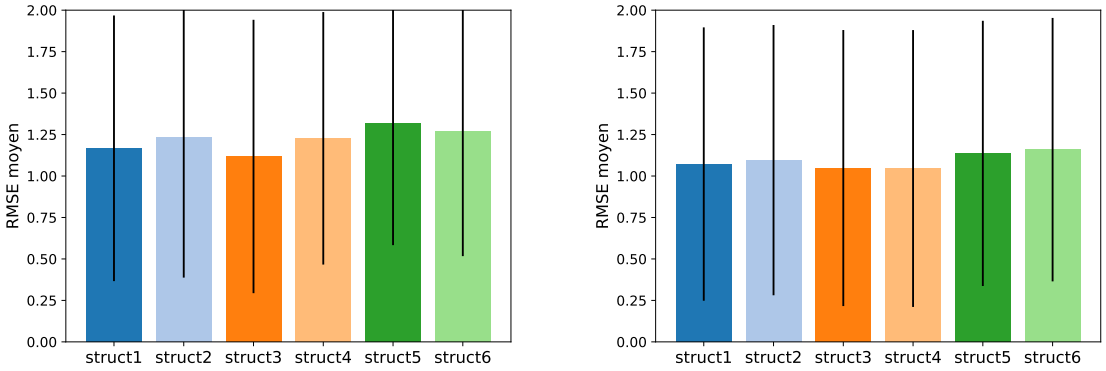
(b) Bruit aléatoire avec choix optimisé de la proportion μ

FIGURE 2.18 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires de 5% de valeurs manquantes artificielles en blocs de taille 14**

Dans la Figure 2.18a, les valeurs des RMSE moyens les plus petites sont obtenues avec les structures 1,2 et 3. Les plus grandes valeurs de RMSE sont obtenues pour les structures 5 et 6. Il semble que, pour les deux catégories de structures, les structures qui donnent les meilleurs résultats sont les structures les plus petites en termes de nombre de couches et de

neurones. Pour le choix aléatoire de la proportion μ avec du bruit aléatoire, la structure qui donne les meilleurs RMSE moyens est la structure 4 mais les structures ont des valeurs de RMSE assez équivalentes. Pour la Figure 2.18b, les différences de RMSE moyens sont moins nettes. Cependant on peut voir que les deux plus grosses structures (struct 5 et 6) et les deux structures undercomplete donnent les RMSE les plus élevés. Pour un même type de structure de DAE et pour globalement toutes les structures, choisir μ de manière optimisée, donne des meilleurs résultats que de le choisir aléatoirement.

La Figure 2.19 présente les RMSE moyens de l'impact de différentes structures pour le bruit sous forme de blocs. Dans la Figure 2.19a, on peut également voir que les structures 3 et 4 sont les structures qui donnent les plus petits RMSE moyens, suivies de près par la structure 1. Les résultats de la Figure 2.19b montrent des RMSE assez similaires, avec des valeurs plus élevées encore une fois pour les deux plus grosses structures : les structures 5 et 6. La Figure 2.19b affiche globalement des meilleurs résultats de qualité de l'imputation et également structure à structure. Pour le bruit sous forme de blocs, le meilleur moyen de choisir la proportion de bruit μ est la méthode optimisée.



(a) Bruit sous forme de blocs avec choix aléatoire de la proportion μ

(b) Bruit sous forme de blocs avec choix optimisé de la proportion μ

FIGURE 2.19 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires de 5% de valeurs manquantes artificielles en blocs de taille 14.**

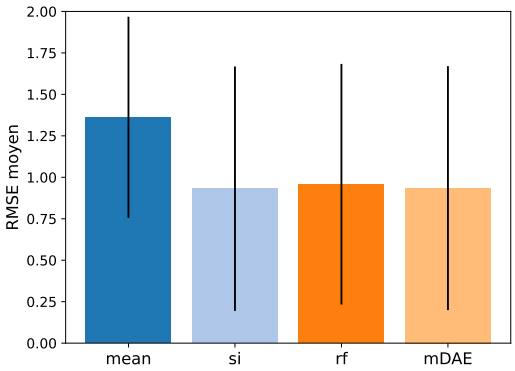
Finalement, comme chaque expérimentation a été effectuée sur les mêmes tirages de données manquantes, il est possible de comparer les 4 graphiques entre eux. La méthode de choix optimisé de μ est la méthode qui donne de meilleurs résultats (même si la différence n'est pas très significative). Cependant le choix de la forme du bruit n'apporte pas tant de différence dans la qualité de l'imputation. Pour rester dans la lignée de la section précédente, nous avons décidé d'utiliser un bruit aléatoire pour la suite de l'étude numérique. Dans ce cas, les structures qui donnent les meilleurs résultats sont les structures 3 et 4. Nous avons choisi d'utiliser la structure 3 pour le reste de l'étude numérique.

Maintenant que les choix des 3 hyperparamètres (bruit aléatoire, choix optimisé de μ, structure 3) sont faits, nous allons comparer la méthode mDAE à des méthodes d'imputation à l'état de l'art parmi celles que nous avons déjà étudiées dans la section 2.2.3.

Comparaison des méthodes d'imputation sur les 4 matrices blocs

Pour évaluer la qualité de performances de notre méthode mDAE, nous la comparons à deux méthodes à l'état de l'art des méthodes d'imputation de données manquantes (voir section 2.2.3). Nous avons choisi de comparer le mDAE aux deux méthodes qui ont montré le meilleur comportement sur les jeux de données UCI : SoftImpute (si) et missForest (rf) dans la section précédente (pour rappel dans la Figure 2.13 si et rf ont les plus petites valeurs du critère MDB). Nous avons également ajouté la méthode d'imputation par la moyenne.

Dans cette partie, nous présentons les résultats des méthodes d'imputation appliquées aux 4 **matrice blocs**. Les hyperparamètres de SoftImpute et missForest sont choisis de la même manière que dans la section précédente (voir Section 2.2.3). Nous avons, comme précédemment, introduit artificiellement 5% de données manquantes sous forme de blocs de taille 14, pour rappeler l'effet d'une électrode manquante. Dans les Figures 2.20a, 2.21a, 2.22a, 2.23a, la valeur moyenne ($\pm$ l'écart type) du RMSE de reconstruction des valeurs manquantes artificielles est représentée pour chaque méthode pour $B = 8$ tirages de données manquantes. De plus, dans les Tableaux 3.26, 2.21b, 2.23b, 2.22b, nous affichons les temps d'exécution de chacune des méthodes. Chaque valeur représente le temps moyen d'exécution de la méthode pour une tirage de données manquantes artificielles.



(a) RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 12$ tirages aléatoires de 5% de valeurs manquantes artificielles en blocs de taille 14.

mean	si	rf	mDAE
1.5 $\pm$ 1 sec	43 $\pm$ 12 min	2 $\pm$ 0,4 h	48 $\pm$ 20 sec

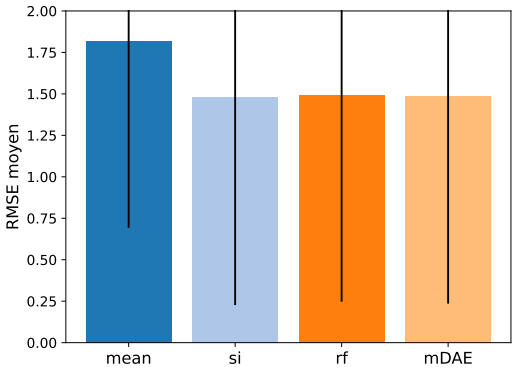
(b) Temps d'exécution pour un tirage de données manquantes

FIGURE 2.20 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires et temps d'exécution associés à l'application des méthodes sur la matrice blocs unipolaire.**

Dans la Figure 2.20a, on peut voir que les résultats des trois méthodes à l'état de l'art sont très similaires. La méthode rf est très légèrement moins performante. Les méthodes si et mDAE ont des valeurs de RMSE moyens très proches. Le véritable écart entre les méthodes, et ce qui permet d'appuyer notre choix d'utiliser la méthode mDAE pour la suite, réside dans les temps de calculs des méthodes d'imputation, présentés dans la Figure 2.20b. Alors que la méthode mDAE a besoin d'une minute pour obtenir des résultats, la méthode si nécessite

plus de 20 minutes. A résultats équivalents, le choix de la méthode la plus rapide devient le meilleur choix. Nous avons ensuite fait la même étude pour les trois autres **matrices blocs**.

Dans les trois Figures 2.21, 2.22, 2.23, nous présentons les RMSE de reconstructions moyens de la méthode mDAE ainsi que ceux des méthodes concurrentes pour $B = 8$ tirages. De plus, nous présentons les temps de calcul des 4 méthodes pour un tirage de données manquantes.



(a) RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires de 5% de valeurs manquantes artificielles en blocs de taille 14.

mean	si	rf	mDAE
0,5 $\pm$ 0,6 ms	23 $\pm$ 2 min	40 $\pm$ 1 min	50 $\pm$ 35 sec

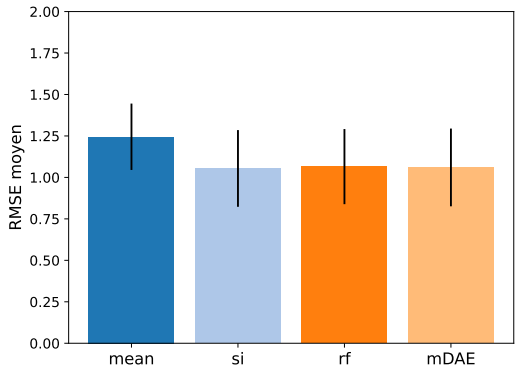
(b) Temps d'exécution pour un tirage de données manquantes

FIGURE 2.21 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires et temps d'exécution associés à l'application des méthodes sur la matrice blocs laplaciens.**

Dans les trois Figures 2.21a, 2.22a, 2.23a, nous retrouvons des résultats semblables que dans la Figure 2.20a. La méthode de l'imputation par la moyenne donne le RMSE le plus élevé, la méthode rf donne un RMSE légèrement plus élevé que les méthodes si et mDAE. Globalement, les résultats sur les matrices blocs bipolaires verticaux et horizontaux se rapprochent des valeurs de RMSE de la matrice blocs unipolaires, mais ceux sur la matrice bloc laplaciens sont plus élevés. Par exemple, le RMSE moyen de la méthode mDAE pour la matrice blocs unipolaire vaut environ 0.8 alors qu'il vaut environ 1.5 pour la matrice blocs laplaciens. Les temps d'exécution des 4 méthodes d'imputation pour les trois matrices (voir Tableaux 2.21b, 2.22b, 2.23b) confirment la rapidité de la méthode mDAE.

Les hyperparamètres ont été étudiés et choisis. Les méthodes d'imputation de données manquantes ont été mises en concurrence. Nous pouvons maintenant imputer les 4 **matrices blocs** par la méthode mDAE avec un bruit aléatoire, une structure overcomplete 3 (à retrouver dans 2.10) et un choix de μ optimisé. A la fin du processus d'imputation, les valeurs manquantes dans les **matrices blocs** $\mathbf{X}$ sont remplacées par celles reconstruites dans la matrice de sortie de la méthode mDAE $\mathbf{Z}$. La matrice des données imputées est alors :

$$\hat{\mathbf{X}} = P_{\Omega}(\mathbf{X}) + P_{\Omega^{\perp}}(\mathbf{Z}), \tag{2.13}$$

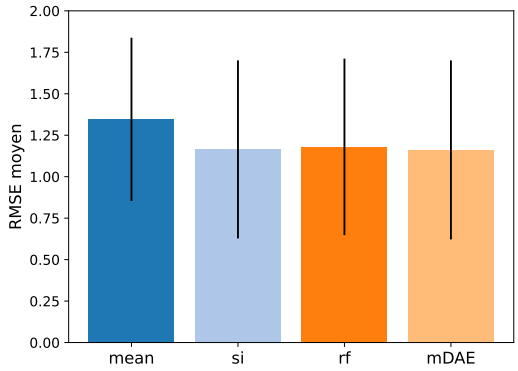


(a) RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires de 5% de valeurs manquantes artificielles en blocs de taille 14.

mean	si	rf	mDAE
0,4 $\pm$ 0,02 sec	24 $\pm$ 3 min	77 $\pm$ 2 min	28 $\pm$ 6 sec

(b) Temps d'exécution pour un tirage de données manquantes

FIGURE 2.22 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires et temps d'exécution associés à l'application des méthodes sur la matrice blocs bipolaires verticaux.**



(a) RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires de 5% de valeurs manquantes artificielles en blocs de taille 14.

mean	si	rf	mDAE
0,8 $\pm$ 0,09 sec	25 $\pm$ 3 min	78 $\pm$ 1 min	30 $\pm$ 8 sec

(b) Temps d'exécution pour un tirage de données manquantes

FIGURE 2.23 – **RMSE moyen de la reconstruction ($\pm$ l'écart type) pour $B = 8$ tirages aléatoires et temps d'exécution associés à l'application des méthodes sur la matrice blocs bipolaires horizontaux.**

où $\Omega^{\perp}$ est l'ensemble des indices $(i, j) \in \{1, ..., n\} \times \{1, ..., p\}$ où $\mathbf{x}_{ij}$ est manquant et P est défini dans l'équation 2.5.

2.4 Conclusion

Dans le Chapitre 2, nous avons présenté deux méthodes pour répondre à la présence des données manquantes dans les données cliniques. Les points à retenir à la fin de ce chapitre sont les suivants :

- nous avons présenté deux réponses au problème de la présence des données manquantes dans les 4 **cubes d'origine** (voir Figure 1.20) : la première réponse est une méthode présentée dans la thèse de Nolwenn Tan (Tan, 2021) qui utilise une technique d'agrégation de données ; la seconde méthode est l'application de notre méthode mDAE sur les données cliniques ;
- nous proposons la première contribution de cette thèse (Mariette Dupuy, Marie Chavent et Remi Dubois, 2024) dans laquelle nous étudions une nouvelle méthode d'imputation de données manquantes appelée mDAE ; nous proposons également une méthodologie de comparaison avec les méthodes à l'état de l'art de l'imputation de données manquantes, avec l'utilisation d'un nouveau critère le "Mean Distance to the Best" ;
- à la fin de ce chapitre, nous avons ainsi à notre disposition deux matrices de données complètes : les **matrices agrégées** et les **matrices blocs**.

Dans le chapitre qui suit, le dernier de ce manuscrit, nous allons utiliser les deux matrices de données complètes : **matrices agrégées** et **matrices blocs** afin de proposer une étude permettant de discriminer les patients à risque de faire une mort subite.

Chapitre 3 : Etude statistique et prédictions de pathologies cardiaques

Table des matières

Ce chapitre présente l'énoncé du problème que nous voulons résoudre. Puis nous rappelons les outils nécessaire pour mener notre étude statistique. Dans la section 3 nous appliquons l'étude statistique entre les deux matrices présentées dans le chapitre 2. Dans la quatrième section nous présentons une méthode de classification basée sur un DAE appris uniquement sur des patients sains. Nous terminons ce chapitre sur l'étude d'un problème de classification multi-classe entre pathologies.

3.1 Énoncé du problème

Le projet HELP dans lequel s'inscrit cette thèse, a pour objectif d'étudier la capacité d'un ECG 128 HD à faire un diagnostic de risque de mort subite à partir de d'enregistrement en rythme sinusal. Dans une étude (Haissaguerre, Hocini et al., 2018) menée par les équipes du LIRYC, les cliniciens ont identifié au cours d'interventions invasives, des éléments pouvant expliquer ces fibrillations ventriculaires. L'équipe a notamment localisé des signaux électriques anormaux sur l'endocarde et l'épicarde de patients ayant subit une FVI. La présence de ces signaux (dans 62% des cas de l'étude) met en évidence l'existence de substrat anormal localisé, expliquant ainsi la survenue de l'arythmie ventriculaire. Ces anomalies structurelles étant non décelables par les méthodes d'imagerie standards, l'objectif est de rechercher leurs traces électriques sur l'ECG 128 HD. Cette recherche de la trace électrique sur l'ECG 128 HD est une idée novatrice. Il n'existe, à notre connaissance, pas d'autres laboratoires de recherche pluridisciplinaires essayant de répondre à la prévention de la mort subite par l'analyse des caractéristiques électriques d'un ECG 128 HD. Nous allons, dans ce Chapitre 3 présenter les résultats que nous avons obtenus par analyse de caractéristiques issues de ces données pour la prédiction des individus à risque de mort subite. Les résultats présentés portent sur des populations de patients identifiés comme malades, qui permettent d'évaluer les performances des méthodes proposées.

3.2 Présentation des données, des méthodes de prédiction et des outils statistiques

3.2.1 Les variables

Dans ce chapitre nous travaillons avec les deux matrices de données complètes décrites dans le Chapitre 2. La première matrice que nous avons présentée dans la section 2.1 est notée la **matrice agrégée** résultant de travail de Nolwenn TAN (Tan, 2021). Pour rappel la configuration des variables pour un patient est présentée dans la Figure 3.24 pour les signaux unipolaires qui est composée de $14 \times 5 \times 6 = 420$ variables par individu.

Les trois autres **matrices agrégées**, associées aux trois autres types de signaux (bipolaires horizontaux, bipolaires verticaux et laplaciens) ont également 420 variables. Ici, nous utilisons la matrice concaténée composée des 4 **matrices agrégées** et possédant donc $4 \times 420 = 1780$ variables, on note cette matrice la **matrice agrégée entière**. C'est la base

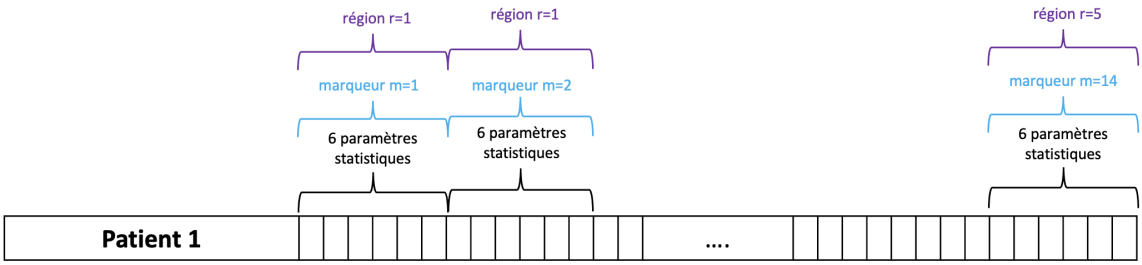


FIGURE 3.24 – **Matrice agrégée unipolaire**
Chaque patient est décrit par 6 paramètres statistiques résumant 14 marqueurs calculés sur 5 région du torse.

de données qui est à ce jour utilisée par les équipes d'ingénieurs du Liryc qui développent de nouveaux marqueurs.

La deuxième matrice que nous allons étudier pour discriminer différentes pathologies est la **matrice blocs**, cette matrice est la version matricielle des cubes d'origine (voir Figure 1.20) dont nous avons imputé les données manquantes avec notre méthode mDAE présentée dans la section 2.2.Cette matrice a la forme présentée dans la Figure 3.25.

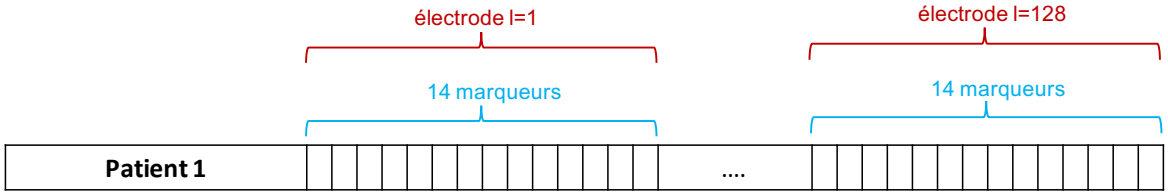


FIGURE 3.25 – **Matrice blocs unipolaire**
Chaque patient est décrit par 128 signaux unipolaires enregistrés par 128 électrodes. Sur chaque signal unipolaire ont été extraits 14 marqueurs.

La **matrice blocs unipolaire** est composée de $128 \times 14 = 1792$ variables correspondant aux 14 marqueurs (voir Section 1.3.2) calculés sur les 128 signaux mesurés par les électrodes posées sur le torse du patient. Pour la **matrice bloc laplaciens** il y a $72 \times 14 = 1008$ variables (voir Section 1.3.2). Pour les **matrices blocs bipolaires verticale et horizontale** il y a $99 \times 14 = 1386$ variables. Ce qui donne au total en agrégeant les 4 **matrices blocs** une matrice de 5572 variables. Dans les prochaines sections nous travaillerons donc également sur une matrice concaténée que nous appellerons **matrice blocs entière** possédant 5572 variables.

3.2.2 Les individus

Le travail présenté ici porte sur deux groupes d'individus : les **individus sains**, qui représentent le groupe contrôle de cette étude, ils sont au nombre de 80. Le protocole d'ac-

Sains	Brugada	DAVD	FVI
80	49	36	45
80	130		
210			

FIGURE 3.26 – **Tableau récapitulatif du nombre d'individus par groupe**

quisition de ces patients a été mis en œuvre à l'IHU Liryc, sous la référence 'ECG-HD'. Les individus de ce groupe doivent respecter des critères d'inclusion, ils doivent avoir plus de 18 ans et sans pathologie cardiaque connue. Dans le cadre de ce même protocole, l'acquisition du groupe de patients pathologiques a été réalisé au service de cardiologie du CHU du Bordeaux, le plus souvent pour une première prise en charge ou le suivi d'une arythmie ventriculaire diagnostiquée (FV, mort subite) ou suspectée (syncope, histoire familiale de mort subite, anomalie ECG). Parmi les pathologies répertoriées dans cette étude clinique, trois pathologies spécifiques ont été étudiées ici : le syndrome de Brugada (voir Section 1.2.2), les DAVD (voir Section 1.2.2), et les FVI (voir Section 1.2.3). La proportion des individus de chaque groupe est détaillée dans le Tableau 3.26.

Il est nécessaire de noter que nous sommes dans un contexte de grande dimension avec peu d'observations. Le nombre de variables de la **matrice agrégée entière** vaut 1780 et le nombre de variables de la **matrice blocs entière** vaut 5572. Le nombre de patients est 210.

3.2.3 Les méthodes de classification

L'idée ici n'est pas de donner un cours exhaustif sur les Support Vector Machine (SVM) et les Forêts Aléatoires (FA) mais plutôt de donner le matériel nécessaire au lecteur non expert en apprentissage statistique pour comprendre le type de frontières de décision générées par ces méthodes et l'influence qu'ont leurs paramètres sur celles-ci. Afin de définir les SVM et les FA, nous considérons les notations suivantes. Nous notons par $\mathcal{D} = \{(x_i, y_i)\}_{i=1,\dots,n}$ un ensemble d'apprentissage ou jeu de données où $x_i \in \mathbb{R}^p$ est l'individu i et $y_i \in \{-1, 1\}$ est le label associé. Nous considérons le cadre binaire ici mais les définitions peuvent être étendues au cadre classification multi-classes.

Les SVM

La méthode des SVM est une méthode d'apprentissage supervisé discriminative qui consiste à trouver l'hyperplan qui sépare au mieux les différentes classes dans l'espace des variables (CORTES, 1995 ; PLATT et al., 1999). La règle de classification du SVM est donnée pour tout $x \in \mathbb{R}^p$ par :

$$F(x) = \text{sign}(\langle w, x\rangle + b), \tag{3.14}$$

où sign est la fonction signe, i.e., $\text{sign}(x) = 1$ si $x \geq 0$ et $sign(x) = -1$ sinon. Cette méthode fait donc l'hypothèse qu'un tel hyperplan existe, hypothèse étant plus ou moins raisonnable en fonction des variables et du problème de classification. Néanmoins, il arrive souvent que

les deux classes ne soient pas linéairement séparable (bruit dans les données, problème de classification compliqué, ...) et donc chercher l'hyperplan séparateur qui maximise la marge (distance entre l'hyperplan et les deux points de classe différentes les plus proches de l'hyperplan, ces points sont appelés les vecteurs support) peut conduire à des problèmes de convergence ou à l'obtention d'une mauvaise solution (hyperplan séparateur sous optimal).

Jusqu'ici la frontière de décision obtenue est un hyperplan dans l'espace des variables d'origine. Cependant, les données ne sont pas toujours linéairement séparables. Il est possible d'introduire une transformation non linéaire des données pour ensuite pouvoir séparer linéairement les données dans ce nouvel espace.

L'introduction d'un noyau K (appelé "Kernel trick" dans la littérature) permet le calcul de l'hyperplan séparateur dans un autre espace de variables potentiellement de grande dimension sans calculer explicitement les coordonnées des individus dans cet nouvel espace. Le fonction de classification pour un SVM à noyau s'écrit :

$$F(x) = \text{sign}\left(b + \sum_{x_i:VS} \alpha_i y i K(x, x_i)\right), \tag{3.15}$$

où la somme est faite sur les individus qui sont vecteurs de support et α_i est un multiplicateur de Lagrange qui vérifie la relation pour tout i, $0 \leq \alpha_i \leq C$. En pratique, les deux paramètres à calibrer sont donc :

- le coefficient de régularisation C, qui permet l'ajustement ou l'adaptation de la marge à des données mal classées ;
- le type de noyau K utilisé, qui peut être polynomial, gaussien, ou autre.

Les forêts aléatoires

La méthode des FA (Breiman, 2001) repose sur la création d'un ensemble d'arbres de décision, appelés aussi arbres CART (Breiman, 2017), et l'agrégation de leurs prédictions pour obtenir un modèle plus robuste et plus précis. La création de cet ensemble d'arbres de décision se décompose en deux phases. La première consiste à tirer q échantillons de bootstrap $\mathcal{L}^n_{\Theta_1}, \ldots, \mathcal{L}^n_{\Theta_q}$ (Breiman, 1996). Un échantillon bootstrap est un sous-ensemble de données obtenu en tirant aléatoirement des observations avec remise à partir d'un ensemble de données initial. Chaque arbre de décision de la FA sera construit sur un échantillon bootstrap différent. La construction des arbres de décision pour une FA suit une variante aléatoire qui est différente de la construction usuelle des arbres de décision. Plutôt que de mettre en compétition l'ensemble des variables à chaque nœud d'un arbre, $mtry$ variables sont tirés uniformément et sans remise parmi l'ensemble des variables. Le meilleur découpage est déterminé uniquement parmi ces $mtry$ variables. Le paramètre $mtry$ est fixé au départ et est le même pour tous les arbres de décision de la FA. L'apprentissage des arbres de décision sur des échantillons bootstrap différents et le tirage de $mtry$ variables permet d'obtenir une diversité dans la procédure de prédiction des arbres de décision obtenus.

Les deux paramètres à optimiser sont :

- le nombre d'arbres de décision constituant la FA, souvent noté $ntree$;

— le nombre de variables tirées aléatoirement et mises en compétition à chaque découpage de noeud, noté $mtry$, et souvent défini par défaut comme $mtry = \sqrt{p}$.

Genuer, Poggi et Tuleau (2008) ont montré que le paramètre $mtry$ est celui des deux qui a le plus d'impact sur les performances de la FA obtenue et qu'augmenter le nombres d'arbres de décision constituant la FA ne permettait pas d'obtenir de meilleure performance au délà d'un seuil.

3.2.4 Les outils d'analyse statistique

Nous présentons les outils statistiques que nous utiliserons dans les sections suivantes afin de présenter nos résultats.

La courbe ROC (Receiver Operating Characteristic)

La courbe ROC permet de visualiser le compromis entre le taux de vrais positifs (sensibilité) et le taux de faux positifs (1 - spécificité) pour différents seuils de décision du modèle. Pour rappel,

$$\text{sensibilité} = \frac{VP}{VP + FN} \tag{3.16}$$

où VP représente le taux de vrais positifs et FN le taux de faux négatifs. On a également,

$$\text{spécificité} = \frac{VN}{VN + FP} \tag{3.17}$$

où VN est le taux de vrais négatifs et FP le taux de faux positifs.

Dans la construction de la courbe ROC, l'axe des ordonnées représente la sensibilité. L'axe des abscisses représente (1 - spécificité). La courbe est tracée en variant le seuil de décision du modèle de classification.

L'AUC (Area Under the Curve)

L'AUC est une métrique utilisée pour évaluer les performances des modèles de classification binaire. Elle représente l'aire sous la courbe ROC. L'AUC fournit une mesure synthétique de la capacité du modèle à différencier les classes positives des classes négatives.

L'AUC s'interprète de la manière suivante :

— AUC = 1 : Le modèle est parfait, classant correctement tous les échantillons.

— AUC = 0.5 : Le modèle ne fait pas mieux qu'un choix aléatoire.

L'AUC est un bon outil d'analyse de résultats car elle résume la qualité de la classification sur tous les seuils possibles de décision. Une AUC élevée indique que le modèle a une forte capacité de discrimination, détectant efficacement les échantillons positifs sans générer un excès de faux positifs.

La matrice de confusion

Une matrice de confusion est un tableau de contingence qui résume les performances d'un modèle de classification en comparant les prédictions du modèle avec les classes réelles. Pour un problème avec k classes, la matrice de confusion devient une matrice kk. La diagonale de cette matrice contient les prédictions correctes pour chaque classe

Dans la section suivante nous allons présenter les résultats de l'analyse descriptive des deux matrices de données complètes. Nous allons également présenter les résultats obtenus avec les deux classifieurs binaires (SVM et FA) pour prédire les patients pathologiques.

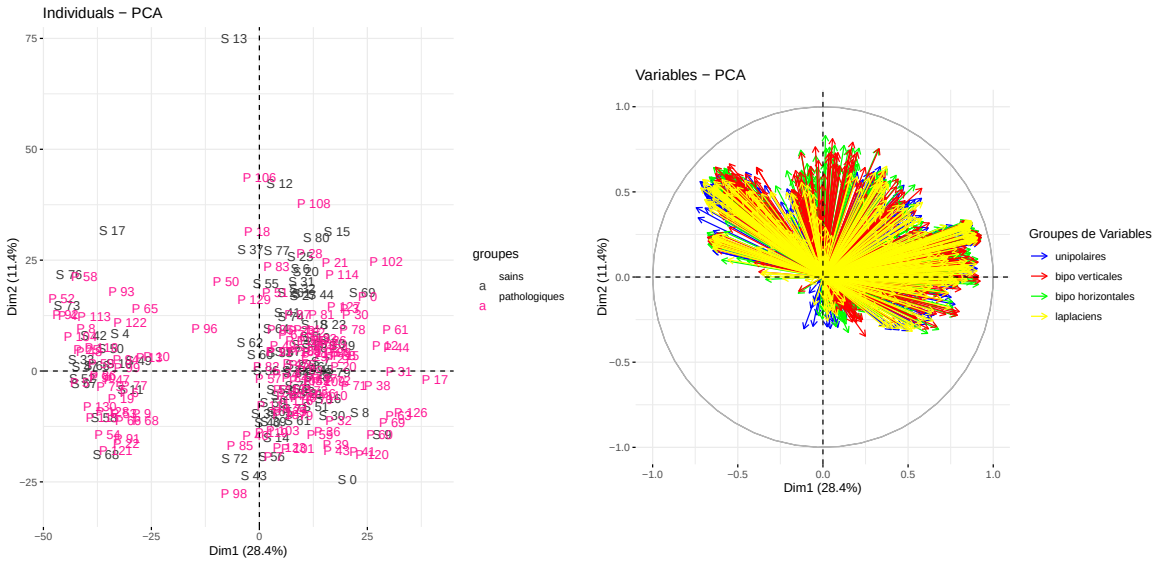
3.3 Discriminer les patients sains des patients pathologiques

En plus d'analyser les performances de classification obtenues par deux méthodes de classification (SVM et FA) nous allons nous attacher à analyser l'apport de la procédure d'imputation que nous avons mise en place dans le Chapitre 2. Dans cette section nous étudierons des résultats à la fois sur la **matrice agrégée entière** définie dans la thèse de Nolwenn TAN (Tan, 2021) et rappelée dans la Figure 3.24 et sur la **matrice blocs entière** présentée dans la Figure 3.25. Nous travaillerons avec les 80 patients sains et les 130 patients pathologiques. Il y a donc au total $80 + 130 = 210$ patients dans les matrices. La **matrice agrégée entière** est de dimension de dimension 210×1780 et la **matrice blocs entière** est de dimension 210×5572. Pour les deux matrices nous sommes dans un cas de grande dimension où le nombre d'individus est bien moins grand que le nombre de variables, surtout dans la **matrice blocs entière**.

3.3.1 Analyse descriptive individus sains versus patients pathologiques

Pour avoir des premières informations sur les données que nous cherchons à discriminer, l'analyse descriptive est une étape essentielle pour la résolution d'un problème de classification. Pour cette analyse descriptive nous pouvons choisir une méthode de statistique descriptive multivariée telle que l'ACP. L'ACP (voir section 1.15 pour un rappel sur le fonctionnement de l'ACP) permet de savoir quels individus se ressemblent et quelles variables sont liées. Nous allons utiliser les visualisations de l'ACP pour étudier les distances entre les individus et des corrélations entre les variables. La Figure 3.27 permet de visualiser les résultats graphiques de l'ACP normée pour la **matrice agrégée entière**.

La Figure 3.27a affiche en gris les patients sains et en rose les patients pathologiques. Ce plan explique presque 40% de l'inertie. Nous pouvons donc dire que ce plan de projection conserve bien les distances entre les individus ce qui nous permet de faire une analyse proche de la réalité des patients. On voit clairement deux groupes d'individus qui se séparent très bien par rapport à l'axe 1 mais qui ne sont pas les patients sains versus les



(a) Graphique des individus sur la matrice agrégée entière dans le plan factoriel 1-2

(b) Cercle des corrélations sur la matrice agrégée entière dans le plan factoriel 1-2

FIGURE 3.27 – **Analyse en Composantes Principales appliquée sur la matrice agrégée entière**

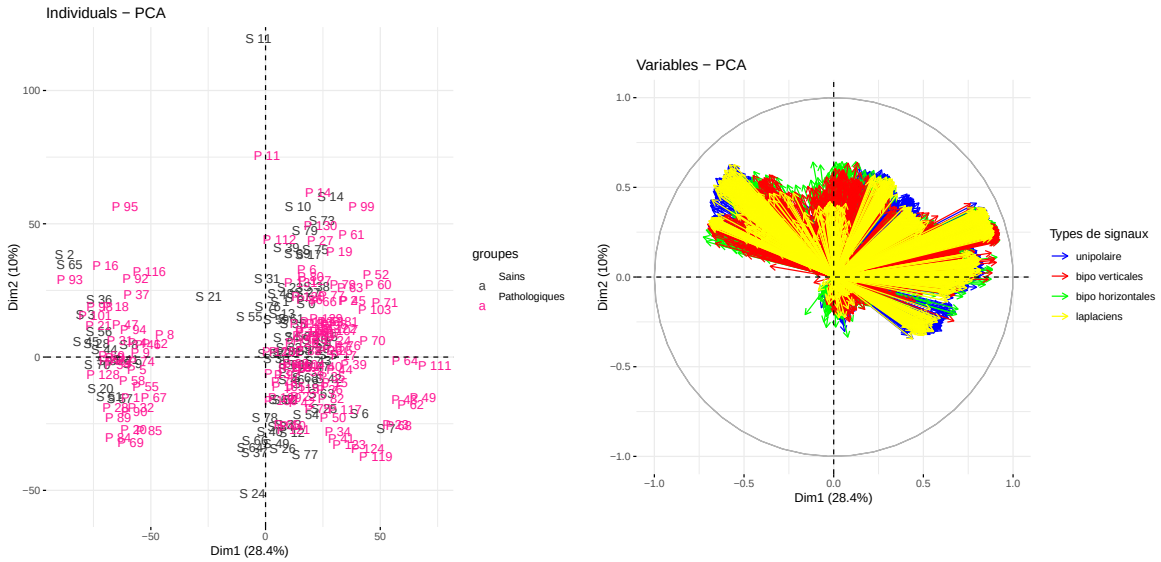
patients pathologiques. Les axes qui portent la variance du nuage de points ne sont pas ceux qui permettront la classification. Certains patients comme les patients S0, S13, S17 sont très éloignés des nuages de points. De plus, la majorité des patients sains et pathologiques sont mélangés.

Sur le cercle des corrélations (Figure 3.27b) nous avons coloré les variables en fonction de leur appartenance aux sous-matrices associées aux signaux (matrice agrégée unipolaire en bleu, matrice agrégée bipolaire verticale en rouge, matrice agrégée bipolaire horizontale en verte, matrice agrégée laplaciens en jaune). Nous pouvons voir sur cette figure que :

— la majorité des variables sont bien projetées (flèches proches du cercle ce qui indique une forte corrélation entre les variables et les composantes principales), mise à part un groupe de variables dans la partie inférieure de l'axe 2 ;
— une superposition des variables appartenant aux sous-matrices ;
— pour chaque sous-matrice les variables forment des groupes de variables très corrélés entre elles ;
— de part et d'autre de l'axe horizontal, deux groupes de variables sont mieux projetés que les autres et corrélés à l'axe 1. Ces variables séparent les deux groupes patients de la Figure 3.27a.

La Figure 3.28 permet de visualiser les résultats graphiques de l'ACP normée pour la **matrice blocs entière**. Ce plan explique la même quantité d'inertie que l'ACP normée sur la matrice agrégée entière (presque 40%).

La Figure 3.28a affiche en gris les patients sains et en rose les patients pathologiques.



(a) Graphique des individus sur la matrice blocs entière dans le plan factoriel 1-2

(b) Cercle des corrélations sur la matrice bloc entière dans le plan factoriel 1-2

FIGURE 3.28 – **Analyse en Composantes Principales appliquée sur la matrice bloc entière**

Sur cette Figure, les patients sains ne sont plus exactement au milieux des patients pathologiques. Il est possible de discerner une légère séparation entre les patients sains et les patients pathologiques. Certains patients comme les patients S11, S21, S24, M11, M95 sont assez éloignés des nuages de points. On retrouve également les deux nuages de points bien séparés.

Sur le cercle des corrélations (Figure 3.28b) nous avons coloré les variables de la même manière que précédemment. Nous pouvons voir sur cette figure que :

- les variables sont bien projetées à l'exception d'un groupe dans les valeurs négatives de l'axe 2 ;
- pour cette matrice de données encore, il y a une superposition des variables appartenant aux sous-matrices relatives aux types de signaux ;
- pour une sous-matrice les variables forment des groupes de variables très corrélées entre elles ;
- certains groupes de variables très corrélés à l'axe 1 entraînent une séparation nette des individus.

Dans la Figure 3.29 nous visualisons le cercle des corrélations pour une sous-matrice : la **matrice blocs unipolaire**. Sur ce cercle des corrélations nous avons coloré les variables en fonction du marqueur d'électrophysiologie qu'elle représente (voir section 1.3.2 et voir Figure 3.25 pour la configuration de la matrice).

Nous pouvons voir que les variables très corrélées entre elles sont en fait les variables correspondants aux mêmes marqueurs. Certaines de ces variables sont très corrélées à l'axe

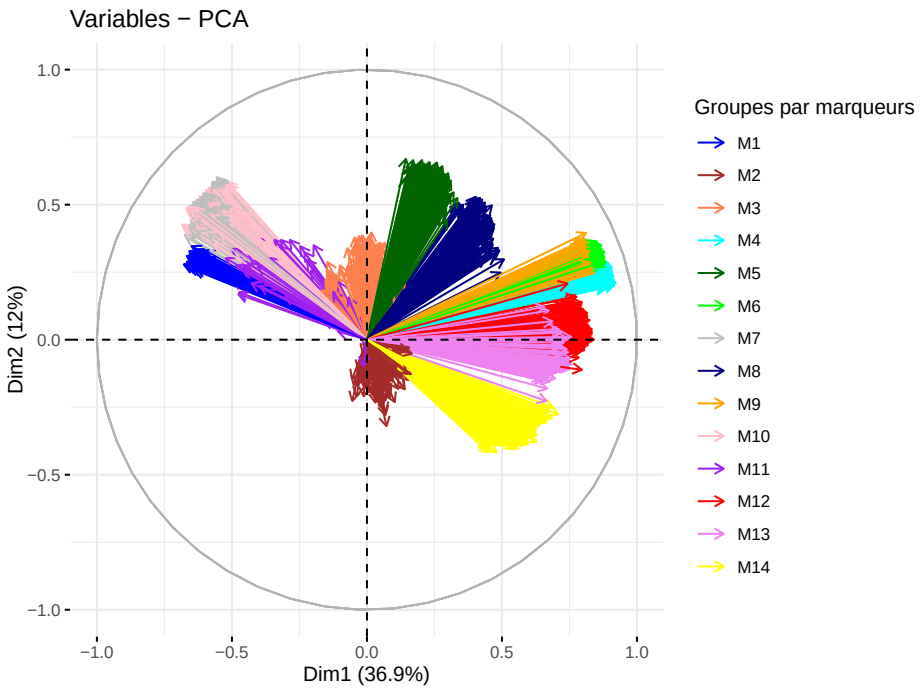


FIGURE 3.29 – **Cercle des corrélations sur la matrice bloc unipolaire dans le plan factoriel 1-2**
Chaque couleur représente un marqueur d'électrophysiologie. Il y a 14 marqueurs étudiés dans cette base de données.

1, ce sont donc ces variables qui créent les deux groupes d'individus dans les graphiques d'individus vus précédemment. Les groupes de variables les plus corrélés positivement à l'axe 1 sont les groupes des marqueurs M4, M6, M9, M12. Les groupes de variables les plus corrélés négativement à l'axe 1 sont les groupes des marqueurs M1, M7, M10. Les individus ayant des valeurs supérieures à la moyenne dans les marqueurs du premier groupe sont les individus dans le groupe de droite tandis que les individus ayant des valeurs plus grandes que la moyenne dans le deuxième groupe de marqueurs seront dans le groupe de gauche dans les graphiques des individus (voir Figure 3.28a).

3.3.2 Les résultats du problème de classification binaire sains versus pathologies

Dans cette sous section nous présentons les résultats du problème de classification binaire : individus sains (classe 0) versus patients pathologiques (classe 1). Pour répondre à

ce problème de classification binaire nous avons utilisé un découpage 3-folds stratifié (la distribution des classes est respectée dans chaque fold) que nous avons répété $B = 30$ fois. La calibration des paramètres K et C des SVM est faite avec la fonction *GridSearchCV* de *scikit learn*. La fonction choisit, au cours d'un processus de validation croisée, la meilleure combinaison entre deux types de noyaux (linéaire et gaussien) et quatre valeurs de C ([1,5, 10,20]). Pour la méthode FA le paramètres *mtry* est défini par défaut et vaut $\sqrt{p}$ avec p le nombre de variables.

Dans le Tableau 3.4 nous présentons les résultats d'AUC moyens pour les deux méthodes de classification présentées précédemment (FA et SVM). Ces deux méthodes ont été appliquées sur la **matrice agrégée entière** et sur la **matrice blocs entière**.

matrice agrégée entière		matrice blocs entière	
SVM	FA	SVM	FA
0.54 ± 0.05	0.65 ± 0.01	0.82 ± 0.01	0.81 ± 0.02

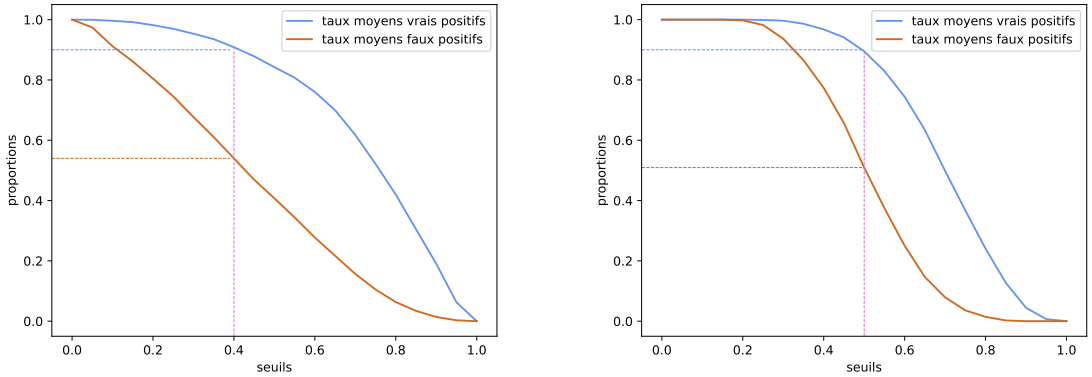
TABLE 3.4 – **AUC moyens sur 30 tirages (moyenne $\pm$ écart-type)**

Dans ce Tableau 3.4 nous pouvons regarder les résultats pour chaque méthode de classification. Pour la méthode SVM appliquée à la **matrice agrégée entière** nous obtenons une AUC moyenne de 0.54, sur la **matrice blocs entière** l'AUC moyenne vaut 0.82. En ce qui concerne la méthode FA la différence d'AUC obtenus avec les deux matrices est également grande : 0.65 pour la **matrice agrégée entière** et 0.81 pour la **matrice blocs entière**. Ces résultats sont très importants pour le projet HELP. En effet, pour ce problème de classification résolu avec les deux classifieurs binaires présentés, les résultats sont meilleurs en utilisant la **matrice blocs entière**. L'imputation avec la méthode mDAE (qui donne la **matrice blocs entière**) permet d'obtenir de meilleurs résultats de discrimination entre les patients sains et les patients pathologique que la méthode d'agrégation utilisée précédemment dans la thèse de Nolwenn (TAN, 2021) (qui donne la **matrice agrégée entière**). Cela signifie sûrement que l'information utile pour la prédiction de patients pathologiques a été perdue au cours du processus d'agrégation des données de Nolwenn TAN. En ce qui concerne les performances des deux méthodes de classification il n'y a pas de différence significative dans les résultats.

Les résultats qui suivent vont nous permettre d'analyser, pour chaque méthode de classification, les seuils qui nous permettront de favoriser plus ou moins les faux positifs et les vrais positifs. Dans la Figure 3.30 on retrouve les courbes moyennes sur 30 tirages des taux de vrais positifs et faux positifs, en fonction d'une grille de seuil avec des valeurs entre 0 et 1 avec des pas de 0.05, pour la méthode SVM (Figure 3.30a) et la méthode FA (Figure 3.30b).

Nous voulons une proportion de vrais positifs élevée car nous préférons sur-évaluer le risque d'un patient de faire une mort subite. Pour avoir une proportion de 0.9 de vrais positifs avec la méthode SVM (Figure 3.30a) le seuil doit être défini à 0.4. En contre-partie avec ce seuil, le taux moyen de faux positifs vaut 0.55 environ. Dans la même logique, c'est un seuil à 0.5 qui donne une proportion de 0.9 de vrais positifs pour la méthode FA (voir Figure 3.30b). Ce seuil donne un taux de faux positifs d'environ 0.5.

Pour la suite des résultats nous avons choisi un seuil de 0.5 afin d'étudier la précision



(a) Courbes des taux moyens de vrais et faux positifs pour la méthode SVM.

(b) Courbes des taux moyens de vrais et faux positifs pour la méthode FA.

FIGURE 3.30 – **Courbes des taux moyens de vrais et faux positifs pour les méthodes SVM et FA**

La ligne pointillée bleue donne la valeur du seuil nécessaire pour obtenir un taux moyen de 0.9 de vrais positifs. Ce seuil donne la proportion associée de faux positifs.

(taux de bonnes classifications) des méthodes sur les deux matrices de données. Il est alors possible d'analyser les prédictions faites par les deux méthodes sur les ensembles tests. Le premier résultat que nous analysons est l'ensemble des boîtes à moustaches calculées pour les deux méthodes de classification et sur les deux matrices étudiées.

Pour chacune des deux méthodes, la différence de précision entre la **matrice agrégée entière** et la **matrice blocs entière** est assez importante. Pour la méthode SVM on obtient une précision moyenne autour de 0.62 sur la **matrice agrégée entière** et de 0.75 sur la **matrice blocs entière**. Pour la méthode FA on obtient une précision moyenne autour de 0.64 sur la **matrice agrégée entière** et de 0.74 sur la **matrice blocs entière**. Ces résultats confortent l'idée selon laquelle la **matrice blocs entière** à plus d'information pour résoudre le problème de classification que la **matrice agrégée entière**. Sur cette matrice, les résultats des deux méthodes de classification ne sont pas significativement différents, la méthode SVM semble donner des résultats légèrement plus stables.

Nous allons maintenant étudier plus finement quels sont les individus qui sont mal classés par les deux méthodes sur la **matrice bloc entière**. Pour ce faire nous avons calculé le pourcentage de mauvais classement des individus sains (Figure 3.32) et des patients pathologiques (Figure 3.33) pour la méthode SVM.

Il y a 56 individus sains mal classés et 47 patients pathologiques mal classés par la méthode SVM. Il y a plus d'individus sains plus souvent mal classés que de patients pathologiques. Sur 30 tirages, 11 individus sains sont 100% du temps mal classés. Chez les patients pathologiques c'est le cas pour 4 patients. Globalement, le SVM fait moins d'erreurs sur les patients pathologiques que sur les individus sains.

Nous pouvons comparer ces graphiques avec ceux obtenus par la méthode FA (Figures 3.34 et 3.35). Nous voyons la même tendance de diminution rapide des pourcentages de

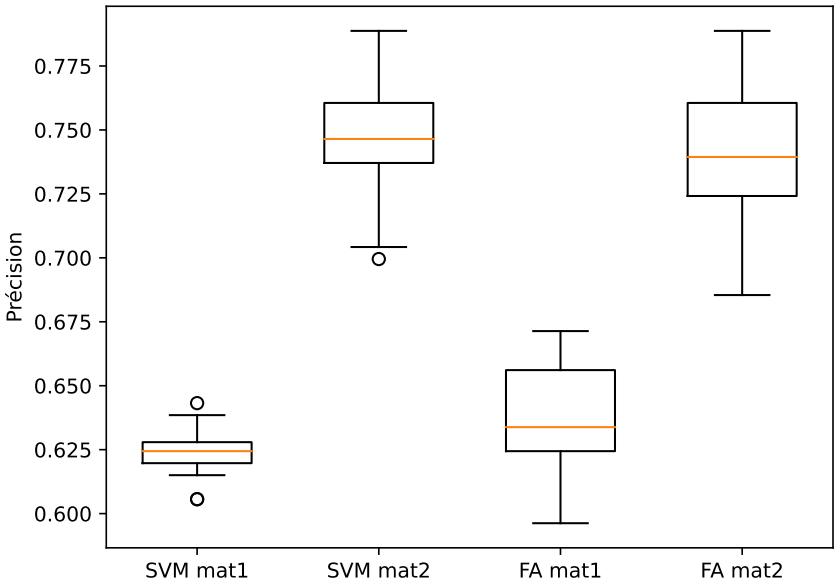


FIGURE 3.31 – **Boîtes à moustaches de la précision calculée sur 30 tirages des méthodes SVM et FA appliquées sur deux matrices de données (matrice agrégée entière et matrice blocs entière)**

mat1 représente la notation de **matrice agrégée entière** et mat2 représente la matrice blocs entière.

mauvais classement.

Il y a 70 individus sains mal classés et 43 patients pathologiques mal classés par la méthode FA. Il y a encore une fois plus d'individus sains plus souvent mal classés que de patients pathologiques. Sur 30 tirages, 14 individus sains sont 100% du temps mal classés. Chez les patients pathologiques c'est le cas pour 1 patients. Globalement, la méthode FA fait moins d'erreurs sur les patients pathologiques que sur les individus sains.

Certains patients sont toujours mal classés pour les deux méthodes. En effet, les deux méthodes ont 56 individus sains et 21 patients malades mal classés en commun.

Nous proposons de visualiser les individus mal classés par la méthode SVM et par la méthode FA appliquées sur **matrice blocs entière** sur le graphique des individus de l'ACP normée en colorant en rouge tous les patients mal classés et en vert tous les patients bien classés.

Nous pouvons retrouver visuellement que les deux méthodes de classification classent mal les mêmes ensembles de patients sains et pathologiques. Dans les deux graphiques nous observons la même dichotomie droite/gauche (mal classés/bien classés) et cela au sein des deux nuages de points.

Dans cette partie nous avons étudié une méthodologie de classification comparant deux méthodes de classification basées sur des algorithmes d'apprentissage machine. Nous allons

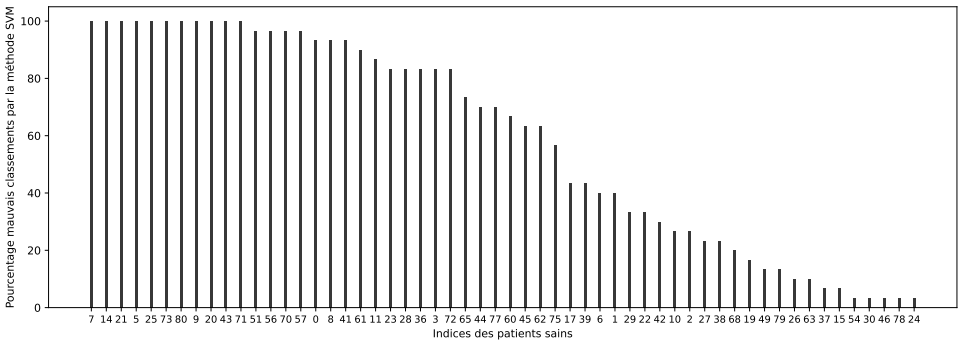


FIGURE 3.32 – **Proportion des individus sains mal classés par la méthode SVM**

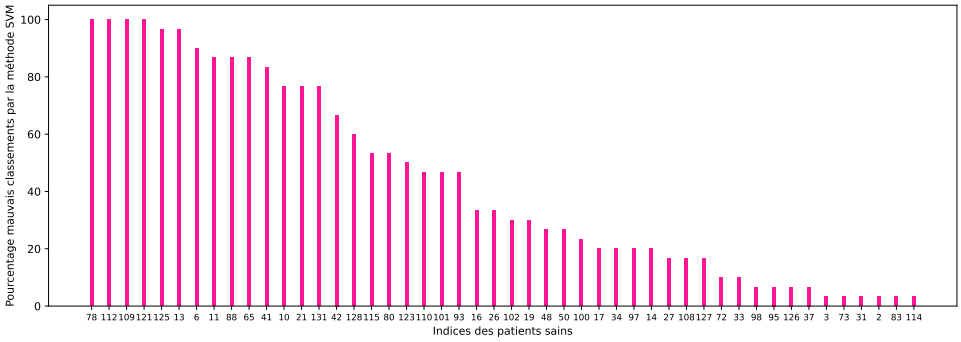


FIGURE 3.33 – **Proportion des individus pathologiques mal classés par la méthode SVM**

dans la section suivante ajouter à notre étude l'analyse des performances d'un réseau de neurones.

3.3.3 Utilisation d'un réseau de neurone

Nous avons décidé d'explorer les capacités d'un réseau de neurones à trouver une solution non linéaire à notre problème de classification. Les réseaux de neurones ont de très bonnes performances sur les images. Dans ces bases de données le nombre de lignes est bien plus grand que le nombre de colonnes. Ce qui rend idéale l'utilisation des réseaux de neurones qui ont de nombreux paramètres à optimiser (voir Section 2.6). Dans notre cas le nombre de lignes est bien inférieur au nombre de colonnes ce qui rendra difficile l'optimisation du réseau. Si le réseau n'arrive pas à optimiser ses poids alors il fera du sur-apprentissage. Le sur-apprentissage survient lorsque le réseau réussi à apprendre parfaitement les données d'entraînement mais n'arrive pas à s'adapter aux données tests. La fonction de coût sur les données d'apprentissage donnent alors de très bons résultats tandis que la fonction de coût sur les données test montrent les très mauvaises performances du réseau pour s'adapter à de nouvelles données.

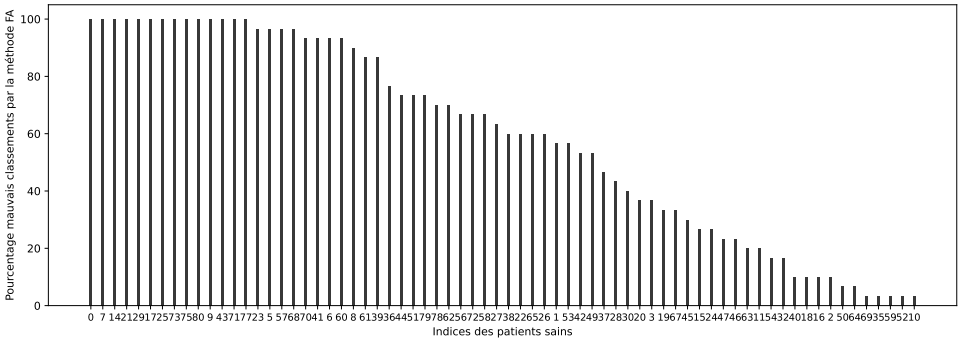


FIGURE 3.34 – **Proportion des individus sains mal classés par la méthode FA**

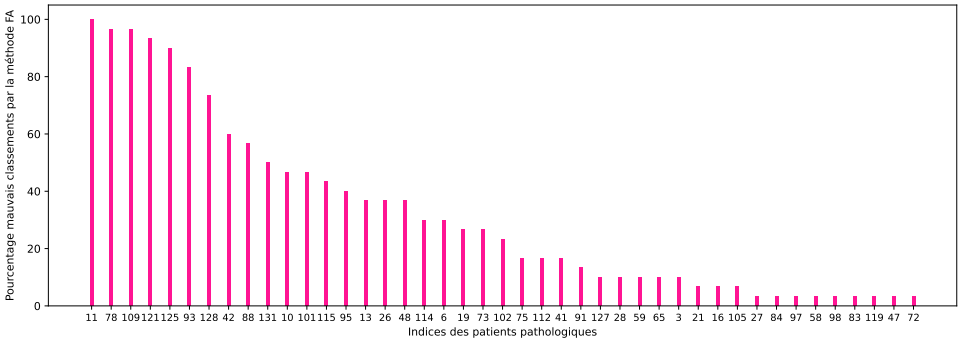
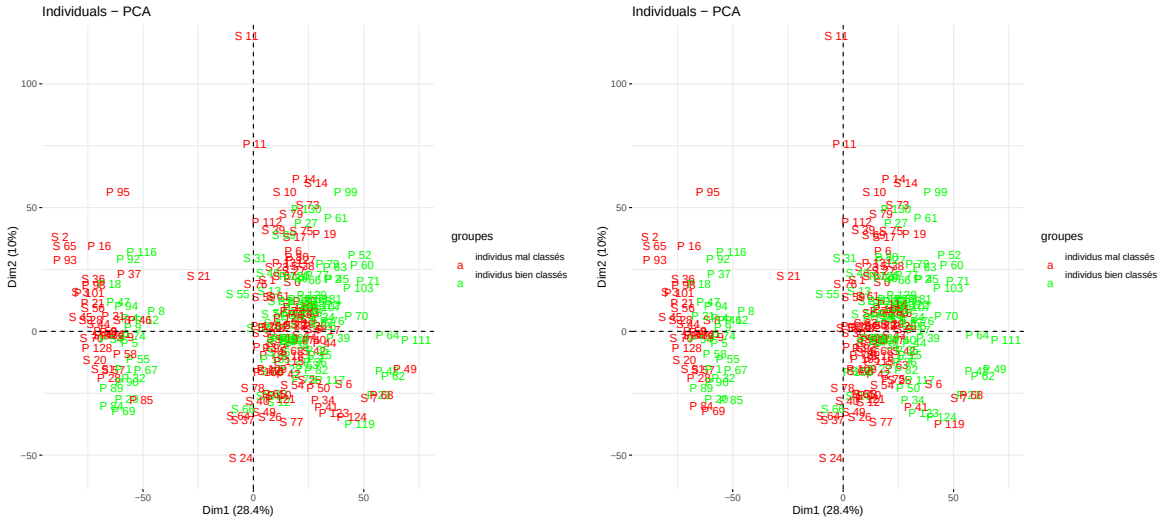


FIGURE 3.35 – **Proportion des individus malades mal classés par la méthode FA**

Pour éviter le sur-apprentissage, la première solution est de tester des structures de réseaux de neurones à 1 couche, ce qui permet d'avoir moins de paramètres à optimiser. Ensuite, il est possible d'ajouter une technique de régularisation appelée "weight decay" (KROGH et HERTZ, 1991) qui consiste à pénaliser les valeurs élevées des poids dans la fonction de coût du modèle, ce qui favorise un modèle plus simple et moins sujet à capturer le bruit des données d'entraînement. Dans la pratique, le "weight decay" est implémenté en ajoutant un terme de régularisation L2 à la fonction de perte. En pénalisant les poids élevés, le weight decay limite la complexité du modèle, favorisant des solutions avec des poids de petite amplitude. Cela évite que le modèle ne s'ajuste trop aux fluctuations de l'ensemble d'entraînement (réduisant ainsi le sur-apprentissage) et améliore sa capacité de généralisation. La dernière méthode que nous avons ajouté pour prévenir le sur-apprentissage est la méthode de "dropout" (SRIVASTAVA et al., 2014). Le "dropout" consiste à désactiver aléatoirement un certain pourcentage de neurones dans le réseau pendant chaque itération d'entraînement. Supposons qu'un réseau comporte n neurones dans une couche donnée, et que le taux de "dropout" est de d. À chaque itération d'entraînement, chaque neurone a une probabilité d d'être ignoré (ou mis à zéro). En conséquence, les sous-réseaux obtenus à chaque itération sont légèrement différents, forçant le réseau global à apprendre des représentations plus robustes et moins dépendantes d'une combinaison précise de neurones.

La première étape de cette étude réside dans le choix d'une bonne structure pour le réseau. Pour cela nous avons testé 5 architectures de réseau (voir Figure 3.37) dont nous avons



(a) En rouge les patients mal classés au moins une fois par la méthode de classification SVM et en vert les patients bien classés.

(b) En rouge les patients mal classés au moins une fois par la méthode de classification FA et en vert les patients bien classés.

FIGURE 3.36 – **Graphique des individus sur la matrice blocs entière dans le plan factoriel 1-2**

S pour individus sains et P pour patients pathologiques

affiché les boîtes à moustache des valeurs de précision dans la Figure 3.38. Ces résultats ont été obtenus en sur 30 tirages de 5-folds.

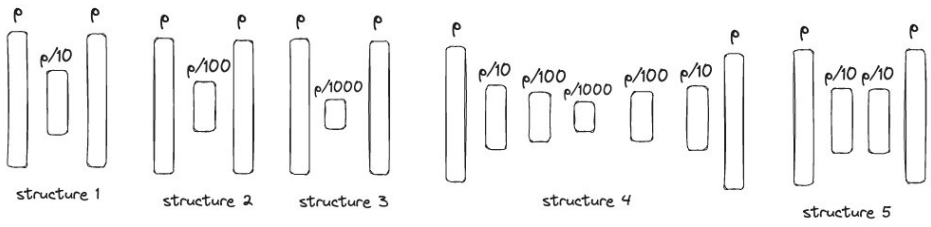


FIGURE 3.37 – **Schéma des structures appliqués pour répondre au problème de classification**

Nous pouvons voir que les résultats ici sont mauvais. Pour toutes les structures les valeurs médianes sont égales environ à 0.5. Si on regarde également les fonctions de coût (les fonctions qui permettent d'optimiser les paramètres du réseau) sur les ensembles d'entraînement (en rouge dans la Figure 3.39) et de validation (en bleu) sur un tirage, nous voyons que le réseau est en sur-apprentissage et cela malgré les outils que nous avons utilisé pour l'éviter.

Nous avons voulu tester la capacité des réseaux de neurones pour résoudre le problème de classification binaire. Dans l'état actuel de la base de données, les réseaux de neurones ne sont pas le bon outil pour répondre à l'identification de patients à risque de mort subite.

Si les réseaux de neurones n'ont pas donné de bons résultats, les SVM et les FA sur la

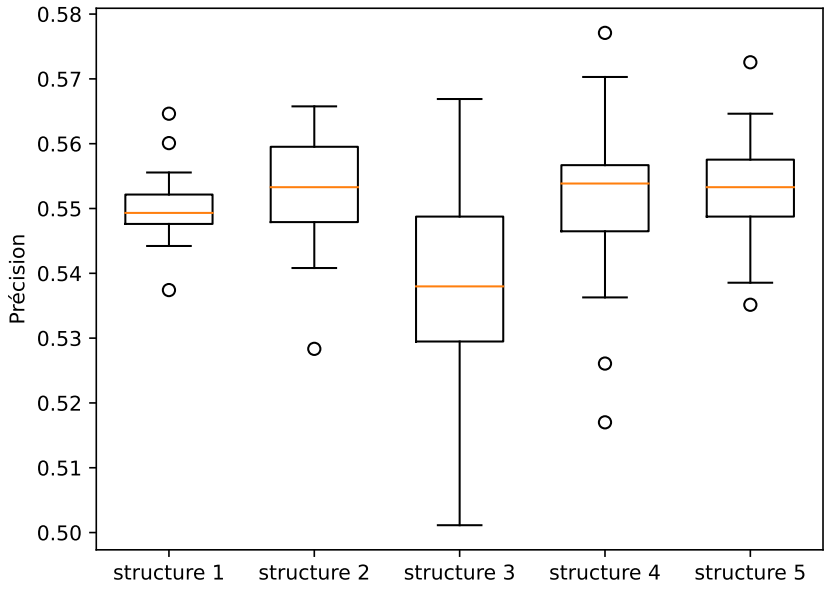


FIGURE 3.38 – **Boîtes à moustache des valeurs de précision des 5 structures de réseaux de neurones**

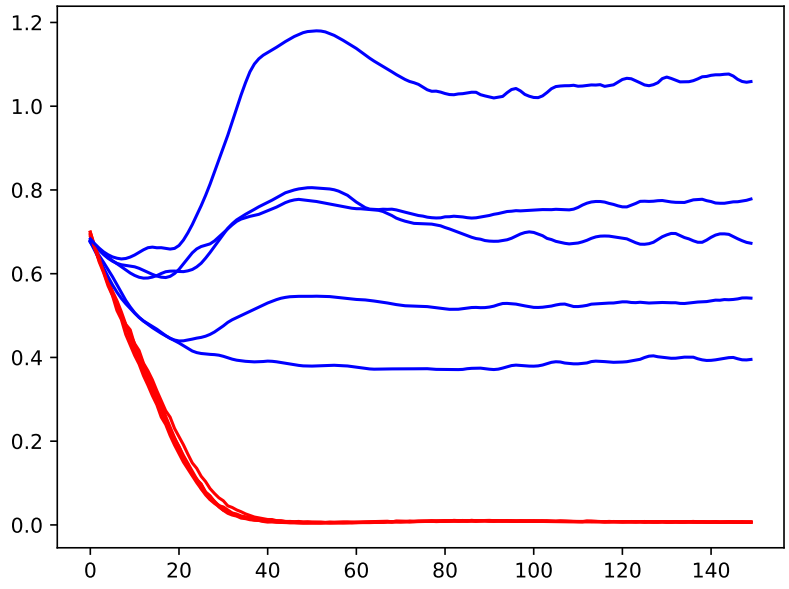


FIGURE 3.39 – **Fonctions de coût pour les données d'entraînement en rouge et de validation en bleu**

Chaque courbe correspond à un découpage.

matrice blocs entière ont donné des résultats assez satisfaisant permettant de déterminer si un nouvel individu est sain ou pathologique. Pour approfondir l'étude de la résolution du problème binaire sains versus pathologiques nous proposons dans le section suivante une nouvelle méthode de classification basée sur les DAE.

3.4 Détection de patients pathologiques à partir de la distribution d'individus sains

Dans cette section nous proposons un classifieur binaire, développé à partir d'un DAE, pour discriminer les individus sains des patients pathologiques. Cette méthode et les résultats associés ont fait l'objet d'un article publié à la conférence Computing in Cardiology 2024 (Dupuy, Chavent et Dubois, 2024).

3.4.1 Méthode générale

Cette section détaille la méthode utilisée pour détecter les individus qui se situent en dehors de la distribution des individus sains. Cette méthode se décompose en plusieurs étapes.

Pour répondre à la problématique de grande dimension que nous avons mis en avant dans la section précédente, la première étape de la méthode est une sélection de variables effectuée en utilisant l'algorithme de Gram-Schmidt (Stoppiglia et al., 2003). La technique d'orthogonalisation de Gram-Schmidt est appliquée afin de donner un ordre de pertinence linéaire aux variables, de la plus discriminante à la moins discriminante, en s'affranchissant de la redondance entre les variables. Les principes de cette technique sont les suivants :

- un vecteur de labels, contenant des 0 pour les patients sains et 1 pour les patients pathologiques est créé ;
- après avoir normalisé les variables d'entrée et de sortie, la corrélation de chaque variable avec le vecteur réponse est étudiée (l'angle entre la variable descriptive et la variable réponse) dans l'espace des N individus. La variable avec la plus grande corrélation est sélectionnée (celle qui a le plus petit angle). Il reste $p-1$ variables à classer où p est le nombre de variables ;
- Les variables descriptives et la variable réponse sont projetées dans le sous-espace orthogonal au premier descripteur sélectionné de dimension N-1, afin d'enlever de chaque variable l'information portée par la variable sélectionnée. Les formules de projection utilisées sont les suivantes :

$$\forall i \in \{1, ..., 1680\} \setminus \{i_0\}\,, \vec{D}_i^1 = \vec{D}_i^0 - \langle \vec{D}_i^0 | \vec{D}_{i_0}^0 \rangle \vec{D}_{i_0}^0 \text{ et } \vec{S}^1 = \vec{S}^0 - \langle \vec{S}^0 | \vec{D}_{i_0}^0 \rangle \vec{D}_{i_0}^0$$

  où i_0 représente l'indice du descripteur sélectionné qui va servir à la projection, $\vec{D}_i^j$ représente le descripteur i projeté dans le sous-espace j et $\vec{S}^j$ la variable réponse projeté dans le sous-espace j. La seconde variable descriptive sélectionnée est celle affichant la corrélation la plus élevée dans ce sous-espace orthogonal ;

— le processus est répété de façon itérative jusqu'à ce que le nombre maximal de variables possible à classer soit atteint. L'algorithme peut classer jusqu'à $N - 1$ variables.

Cette stratégie de sélection de variables appliquée à notre problématique permet de sélectionner un sous ensemble G de variables pertinentes pour le problème de classification binaire considéré qui constitueront l'espace d'entrée du DAE utilisé.

L'étape suivante consiste à entraîner un DAE uniquement sur les individus sain. Les propriétés des DAE font que le modèle entraîné est capable de reconstruire de manière robuste, les vecteurs d'entrées des individus sains même en présence de bruit sur certaines de leurs variables (voir section 2.2.1). Ensuite, le DAE entraîné est appliqué sur les individus sains tests et les patients pathologiques tests qui sont partiellement masqués (une partie m des variables est mise à 0) ; ce qui équivaut à considérer ces variables comme bruitées. Les capacités du DAE étant de débruiter les données, la sortie du DAE permet l'obtention de valeurs reconstruites. Le DAE ayant été entraîné sur des patients sains uniquement, l'hypothèse est que la reconstruction sera de bonne qualité pour un individu sain, et de mauvaise qualité pour un patient malade. Ainsi la Mean Absolute Error (MAE), est calculé sur les valeurs masquées pour évaluer les performances du DAE dans la reconstruction de ces entrées. La Figure 3.40 résume la procédure de reconstruction des variables masquées chez les individus sains tests et les patients pathologiques tests.

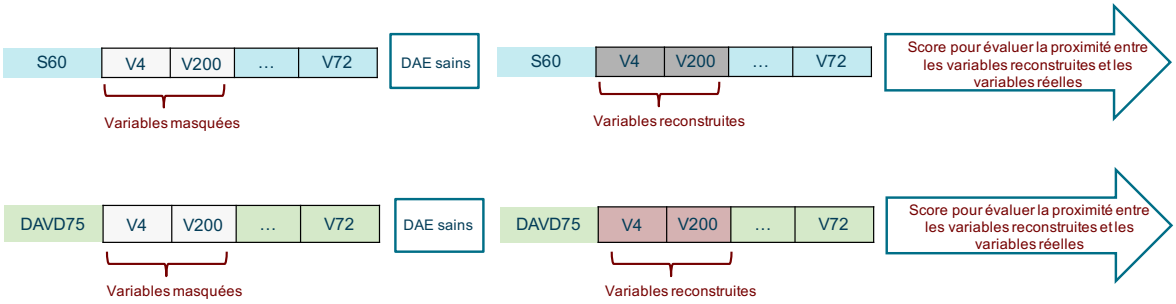


FIGURE 3.40 – **Procédure de reconstruction des variables masquées chez les patients sains et chez les patients pathologiques.**
Ici le schéma est présenté pour le problème de classification sains versus DAVD. En bleu les patients tests sains, en vert les patients tests pathologiques. Pour chaque patient une partie de ses variables est masquées.

La qualité de reconstruction des m valeurs masquées d'un patient est évaluée en calculant la différence entre les valeurs reconstruites et les valeurs réelles. En pratique, la qualité de la reconstruction des variables masquées d'un patient est calculée par le MAE entre la matrice des données de départ $\mathbf{X}$ et la matrice reconstruite $\mathbf{Z}$ seulement sur les indices des variables masquées en entrée :

$$MAE = \frac{1}{|\Omega^{\perp}|} \|P_{\Omega^{\perp}}(\mathbf{X}) - P_{\Omega^{\perp}}(\mathbf{Z})\|_F^2, \tag{3.18}$$

où $\Omega_i^{\perp}$ représente les positions des variables mises à 0 pour chaque patient i et $|\Omega^{\perp}|$ est le nombre de valeurs masquées pour ce patient i et P est défini dans l'équation 2.5.

Les résultats sont calculés en utilisant une procédure de validation croisée k-folds. L'utilisation de la procédure k-folds nous permet une évaluation plus robuste du critère MAE.

3.4.2 Résultats

Dans cette section, nous présentons les résultats de l'évaluation de notre méthode. Nous appliquons notre méthode sur deux pathologies : les patients DAVD et les patients FVI sur la **matrice agrégée entière**. Nous choisissons une procédure de validation croisée 5-folds, ce qui signifie que chaque patient se trouve une fois dans un fold test. Dans les résultats qui suivent on concatène les résultats des 5 folds test, dans ce vecteur on retrouve tous les patients de la base de données évalués dans le fold test concaténé. Nous avons choisi empiriquement, après l'analyse de différentes expérimentations, $G = 20$ variables sélectionnées par Gram-Schmidt et $m = 0.2$ la proportion de variables masquées.

Nous étudions deux problèmes binaires :

— individus sains versus patients DAVD
— individus sains versus patients FVI

Sélection de variables

Dans le tableau 3.5, nous présentons les cinq variables les plus pertinentes retournées par Gram-Schmidt pour chaque problème. Comme nous avons travaillé avec la **matrice agrégée entière**, les noms de variables respectent un certain code : chaque variable est composée de quatre informations : région, type de signal, paramètre statistique, marqueurs. Les noms des variables les plus pertinentes sont présentés de cette manière : marqueur_signal_region_statistiq

DAVD
QRS Wavelets kurtosis _ laplacien _ haut-droit _ percentile 5%
QRS Wavelets raz _ laplacien _ global _ moyenne
QRS Wavelets rms _ laplacien _ haut-droit _ percentile 5%
QRS Peak to peak voltage _ laplacien _ haut-droit _ percentile 95%
QRS Wavelets rms _ bipolaire-horizontal _ global _ percentile 5%
FVI
QRS Peak to peak voltage _ laplacien _ global _ percentile 5%
QRS Wavelets rms _ laplacien _ haut-droit _ percentile 5%
QRS Wavelets raz _ laplacien _ global _ moyenne
QRS Wavelets rms _ laplacien _ haut-droit _ percentile 5%
QRS Peak to peak voltage _ bipolaire-vertical _ haut-gauche _ percentile 5%

TABLE 3.5 – **Noms de cinq des vingt variables les plus pertinentes sélectionnées avec l'algorithme de Gram-Schmidt**
Pour chaque problème binaire les noms des variables les plus pertinentes sont présentés de cette manière : marqueur_signal_region_statistique.

Pour la pathologie DAVD, les variables sélectionnées sont principalement des variables associées à la fragmentation du signal calculées par une évaluation des hautes fréquences ("wavelets"), et également une variable relative à l'amplitude ("peak to peak voltage"). En outre, les variables sélectionnées sont celles qui décrivent les signaux qui se situent principalement dans la zone supérieure droite ou sur le torse entier ce qui est cohérent avec les informations sur la pathologie DAVD 1.2.2 qui est une maladie impliquant la base du ventricule droit avec une infiltration graisseuse qui provoque une fragmentation du signal endocavitaire. En fonction de l'état d'avancement de la pathologie, celle-ci peut conduire à des complications entraînant l'altération de la taille du ventricule, dans ces cas il y aura un impact sur l'amplitude du signal .

Pour le groupe FVI 1.2.3, nous savons que le signal électrique endocavitaire (à l'intérieur du cœur) est fragmenté localement dans le ventricule. Étant donné que ce phénomène peut se produire n'importe où dans le cœur, il n'est pas surprenant d'obtenir des variables résumant les zones droites, gauches et torse entier. Le fait d'avoir des variables basées sur la fragmentation n'est également pas surprenant dans ce cas.

En ce qui concerne les variables qui sont presque toutes sur les signaux laplaciens et une bipolaire, cela pourrait être dû à l'avantage que nous avons noté dans la section 1.18, à savoir que ces signaux amplifient vraiment les signaux locaux et qu'ils semblent être les plus pertinents pour les deux pathologies pour capturer les différences avec les patients sains. Il est également intéressant de noter que ce n'est pas seulement la moyenne de ces variables qui font la différence, mais plutôt les valeurs extrêmes (montrant ainsi que ces patients sont plus aberrants par rapport à la distribution normale de la population saine).

Histogrammes de l'erreur de reconstruction des individus sains et pathologiques.

Les résultats que nous présentons ont été obtenus en utilisant la **matrice agrégée entière**. Le DAE a été entraîné sur 500 epochs, avec un "learning rate" de 10^{-3}.

Pour chaque problème binaire, nous avons calculé la valeur des MAE pour les patients test. L'histogramme de ces erreurs MAE tests lissé avec un noyau gaussien est présenté dans les Figures 3.41 et 3.42.

Comme attendu, il apparaît ici que les valeurs de reconstruction sont meilleures chez les individus sains que chez les patients pathologiques pour les deux problèmes binaires considérés. Pour le DAE entraîné avec les variables sélectionnées pour la pathologie DAVD, la médiane du MAE des variables reconstruites pour le groupe sain est de $0,66[0,36 - 0,95]$ ([1er, 3ème quartile]) contre $1,58[0,52 - 2,02]$ pour le groupe DAVD. Un test de Kolmogorov-Smirnov (Hodges Jr, 1958) sur les deux séries de MAE (MAE des individus sains test et MAE des patients pathologiques test) nous indique une p-valeur plus petite que le seuil 0.05, donc nous pouvons rejeter l'hypothèse nulle. Les deux histogrammes n'ont pas les mêmes lois de probabilité.

Pour le DAE entraîné avec les variables sélectionnées pour la pathologie FVI, la médiane du MAE des caractéristiques reconstruites est de $0,53[0,26-0,82]$ contre $1,32[0,42-2,17]$ pour le groupe IVF. Le test de Kolmogorov-Smirnov donne également une p-valeur plus

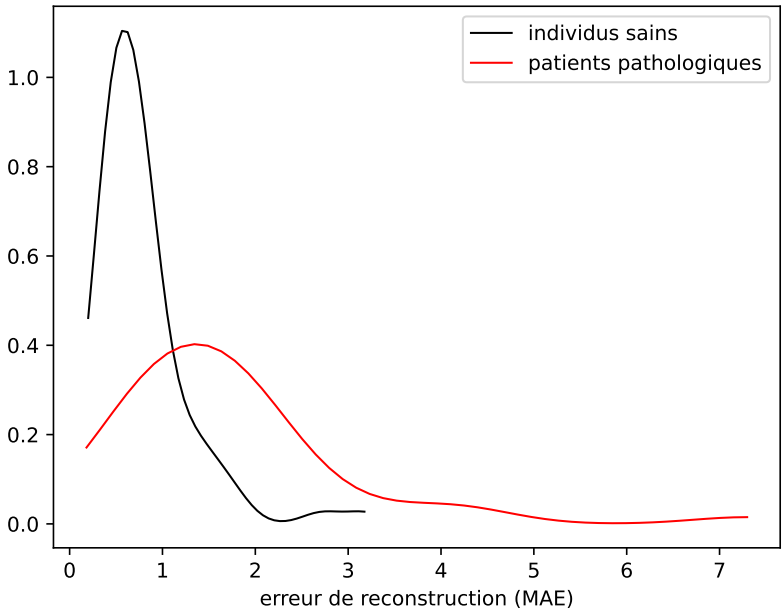


FIGURE 3.41 – **Histogramme lissé par un noyau gaussien de l'erreur de reconstruction (MAE) sur les variables mises à 0 pour le problème discriminant individus sains versus DAVD**

petite que le seuil 0.05, donc nous pouvons rejeter l'hypothèse nulle.

Pour chaque problème discriminant, nous observons que l'utilisation d'un DAE entraîné sur une population saine réduite aux variables les plus pertinentes aboutit à deux distributions de l'erreur de reconstruction distinctes. Il y a une différence notable dans la qualité de reconstruction, valeur qui peut être utilisée comme critère de classification.

Classification

Pour utiliser le critère MAE test de chaque patient pour faire de la classification binaire il suffit d'utiliser le critère de décision suivante : si MAE $\leq$ à un seuil s_i le patient est classé comme sain sinon il est classé comme pathologique. Ce critère de décision est appliqué sur une grille de seuils uniformes. La qualité du classifieur obtenu est évaluée par un analyse ROC. Nous avons avons ensuite comparé notre nouveau classifieur aux deux méthodes présentées dans la section 3.2.3 : les Forêts Aléatoires (BREIMAN, 2001) et les Support Vector Machine (PLATT et al., 1999). La méthode SVM a été appliquée aux mêmes ensembles de données réduits avec les variables sélectionnées par Gram-Schimdt que la méthode DAE. Pour les deux méthodes concurrentes, nous utilisons les paramètres par défaut (C et K pour les SVM et *mtry* pour les FA) disponibles dans les fonctions de scikit-learn [6]. La Figure

6. https ://scikit-learn.org/stable/supervised_learning.html

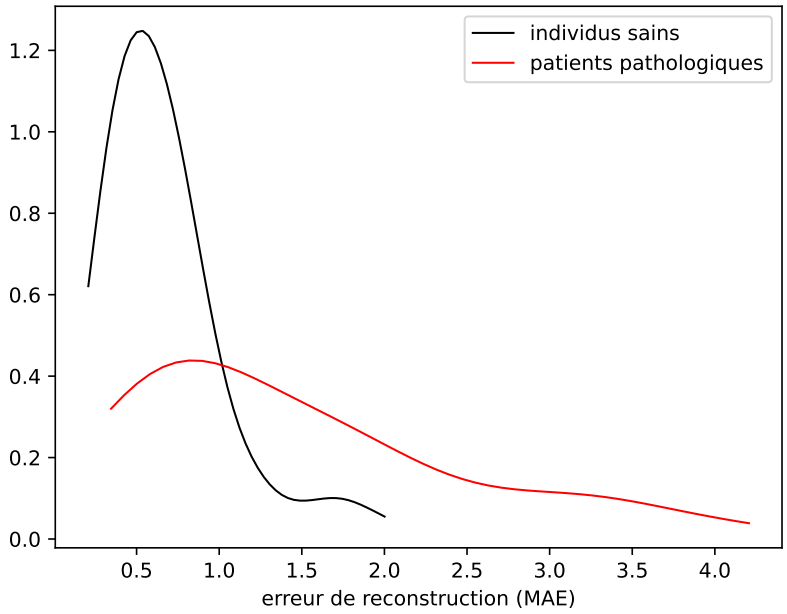


FIGURE 3.42 – **Histogramme lissé par un noyau gaussien de l'erreur de reconstruction (MAE) sur les variables mises à 0 pour le problème discriminant individus sains versus FVI**

3.43 montre les résultats pour la pathologie DAVD. L'AUC de la méthode DAE est de $0,81$, celle de la méthode SVM est de $0,86$ et celle des FA est de $0,90$. La méthode FA, dans ces cas, obtient de meilleurs résultats pour la discrimination entre les individus sains et ceux présentant une pathologie.

La Figure 3.44 présente les résultats pour la pathologie FVI. L'AUC de la méthode DAE est de $0,82$. L'AUC de la méthode SVM est de $0,87$ et l'AUC des FA est de $0,93$. Les résultats pour cette pathologie sont encore meilleurs, les FA restant la meilleure méthode pour ce problème de classification.

Dans cette section, nous avons présenté une méthodologie basée sur les DAE pour détecter les patients qui se situent en dehors de la distribution des individus sains. En intégrant des techniques de réduction en dimension avec Gram-Schimdt et en utilisant des DAE, nous avons développé une nouvelle stratégie pour distinguer les individus sains de ceux présentant deux pathologies spécifiques : les patients DAVD et les patients FVI. Les résultats obtenus sont prometteurs car proches des résultats obtenus par les méthodes SVM et FA.

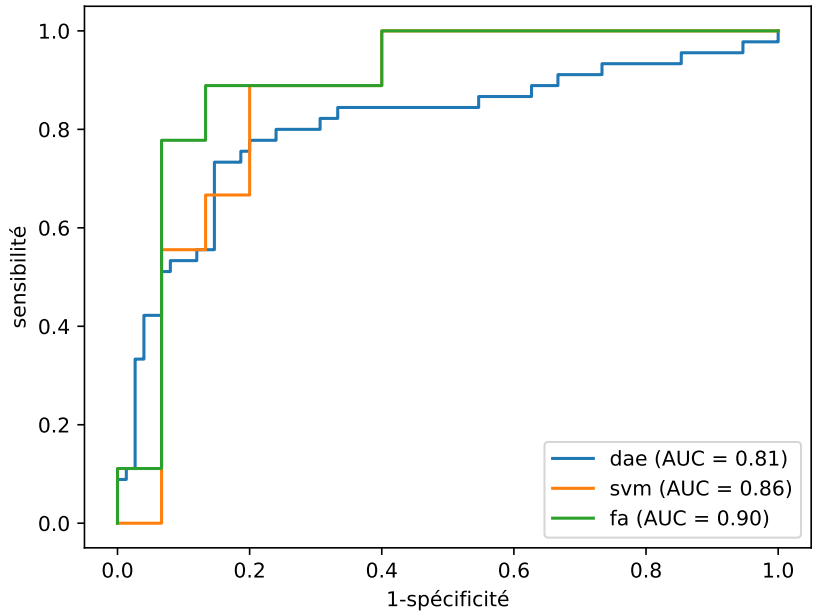


FIGURE 3.43 – **Aire sous la courbe pour la population DAVD par rapport au problème de classification des individus sains.**

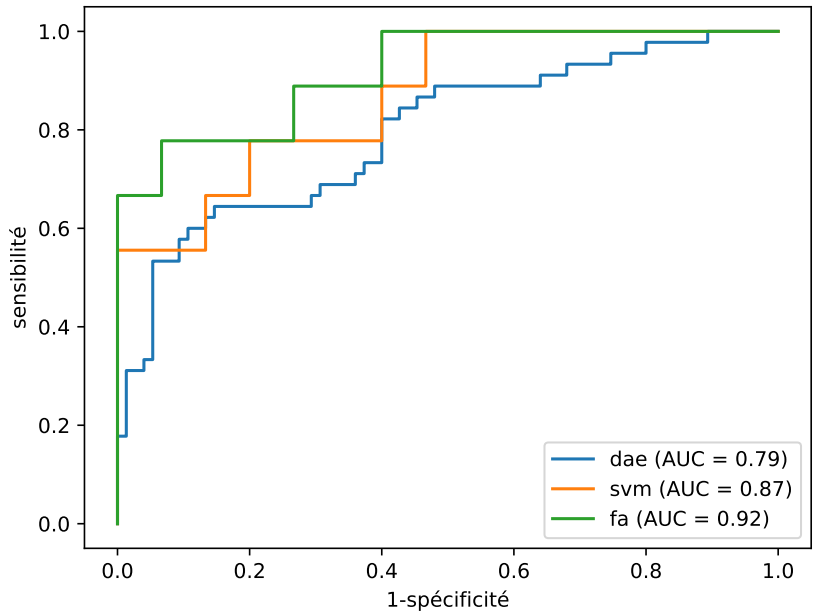


FIGURE 3.44 – **Aire sous la courbe pour la population FVI par rapport au problème de classification des individus sains.**

3.5 Classification multi-classes entre pathologies

Les résultats présentés dans cette section sont l'ébauche d'une étude de classification multi-classes entre pathologies.

3.5.1 Analyse descriptive

Pour l'analyse descriptive du problème multi-classe nous utilisons également l'ACP en affichant seulement les patients pathologiques, colorés par leur pathologie.

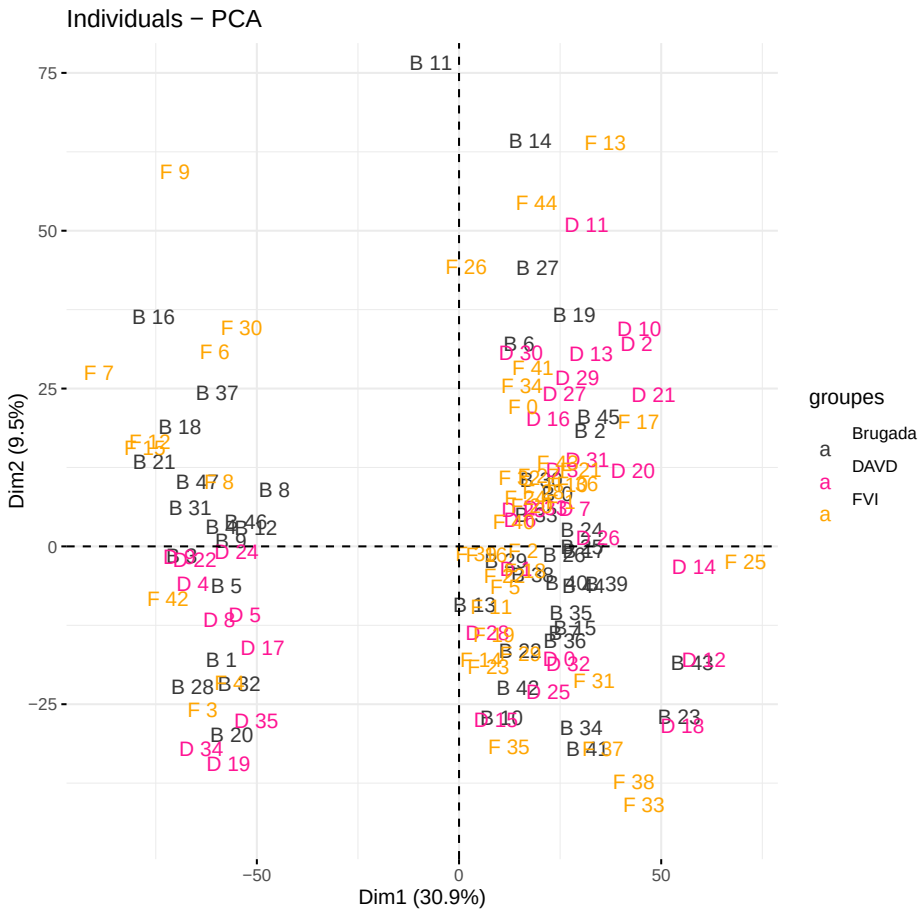


FIGURE 3.45 – **Graphique des individus sur la matrice blocs entière dans le plan factoriel 1-2**

Dans ce graphique nous pouvons voir que le problème à résoudre sera encore plus complexe que le problème de classification binaire précédent. Tous les patients sont mélangés sur ce premier plan factoriel.

3.5.2 Méthode

Dans cette section nous allons explorer la possibilité de déterminer la pathologie d'un nouveau patient déjà identifié comme étant pathologique. Nous pouvons retrouver les proportions de patients de chaque pathologie dans le Tableau 3.26 (48 patients Brugada, 36 patients DAVD et 45 patients FVI). Pour un problème multi-classe, le SVM utilise l'approche "one versus one". Un modèle binaire est entraîné pour chaque paire unique de classes. Chaque modèle apprend à distinguer une paire de classes et ignore les autres. Lorsqu'une nouvelle donnée est introduite pour prédiction, chaque classifieur binaire "one versus one" prédit à quelle classe parmi sa paire appartient cette donnée. Ensuite, une "élection" par vote majoritaire est effectuée : la classe qui reçoit le plus de votes parmi les classifieurs binaires est attribuée comme prédiction finale pour cette donnée. Dans le cas d'une égalité de votes entre classes (ce qui peut arriver si les données sont proches de la frontière de décision), la fonction développée dans *scikit-learn* utilise une règle arbitraire pour choisir une classe, ou attribue une classe de manière aléatoire. Pour résoudre ce problème nous avons calculé les résultats sur 30 tirages de découpage 3-folds.

3.5.3 Résultats

Pour analyser la capacité de classification du SVM dans ce problème multi-classes nous avons utilisé une matrice de confusion entre les patients Brugada, DAVD et FVI. Cette matrice de confusion moyenne est présentée dans la Figure 3.46 :

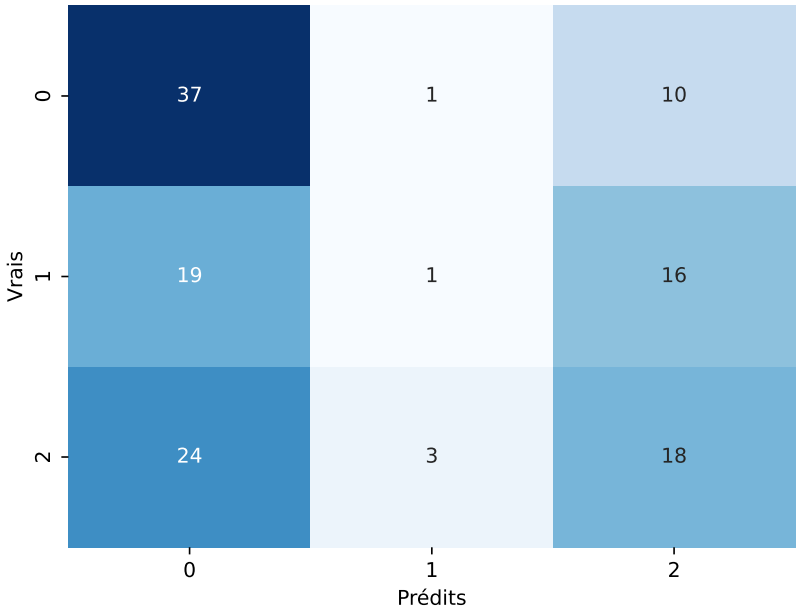


FIGURE 3.46 – **Matrice de confusion moyenne sur 30 tirages de découpage 3-folds**
0 : syndrome de Brugada ; 1 : DAVD, 2 : FVI

La classe 0 qui correspond aux patients Brugada est la classe la mieux prédite (37 patients

en moyenne sur 48). Sur les 45 patients FVI 18 sont correctement prédits. Seulement 1 patient DAVD est correctement prédit. En général les patients mal classés sont soit prédits comme des patients Brugada, soit comme des patients FVI. Si l'on s'appuie sur l'analyse du graphique des individus de l'ACP en Figure 3.45, on remarque que les patients DAVD sont essentiellement mélangés aux autres pathologies. Pour les patients Brugada ou FVI on peut voir que certains patients sont plus éloignés des groupes ce qui les rend peut-être plus facilement identifiables par l'algorithme.

3.6 Conclusion

Le dernier chapitre de ce manuscrit de thèse avait pour but de présenter des résultats préliminaires sur la possibilité de discriminer des pathologies cardiaques avec des données issues d'un ECG 128 HD. Les points à retenir à la fin de ce chapitre sont les suivants :

- nous avons détaillé une étude statistique permettant d'aborder le problème de classification binaire individus sains versus patients pathologiques. Cette étude a été faite sur les deux matrices présentées dans le Chapitre 2 : la matrice agrégée entière et la matrice blocs entière. Les résultats obtenus sur la matrice blocs entière avec les méthodes SVM et FA sont prometteurs ;
- nous avons proposé une nouvelle méthode de classification basée sur un DAE et la reconstruction de patients dont une partie de leurs variables a été masquée (Dupuy, Chavent et Dubois, 2024). Cette méthode obtient des résultats de classification équivalent aux méthodes SVM et FA ;
- le problème de classification binaire donnant des premiers résultats plutôt satisfaisants nous avons étendu nos analyses au problème de classification multi-classes. Dans ce cas, les méthodes d'apprentissage machine ne permettent pas pour le moment d'identifier les patients à risque de faire une mort subite.

Conclusion et perspectives

Conclusion

Ce travail de thèse s'inscrit dans le cadre du projet HELP, visant à améliorer l'identification et la prévention des patients à risque de faire une mort subite cardiaque. Ce projet repose sur l'étude des signaux enregistrés par un ECG 128D. Nous avons présenté dans le Chapitre 1 le contexte clinique dans lequel s'inscrit cette thèse. La première section était une introduction aux bases de l'électrophysiologie cardiaque. Puis nous avons décrit, dans une deuxième section, les principales arythmies ventriculaires ainsi que les pathologies étudiées dans le manuscrit. Après la description du projet HELP et des données qui ont été enregistrées au cours d'un ECG 128 HD, nous avons révélé la présence d'un grande nombre de données manquantes. Cette grande quantité de données manquantes limite l'utilisation d'algorithmes d'apprentissage statistique pour la prédiction de mort subite. Pour cela nous avons décidé de gérer les données manquantes.

Dans le deuxième chapitre de la thèse, nous présentons deux méthodes de gestion des données manquantes. La première méthode est apportée par Nolwenn TAN au cours de sa thèse (Tan, 2021). Nolwenn TAN avait proposé une méthode d'agrégation des données d'origine extraites des signaux cardiaques. Nous avons proposé comme seconde solution, une méthode d'imputation des données manquantes : cette méthode appelée le mDAE est basée sur une modification du DAE afin de l'adapter à la présence de données manquantes. En plus de la méthode d'imputation applicable dans un contexte général sur des données tabulaires numériques, nous avons proposé une méthodologie de choix de la meilleure méthode d'imputation adaptée à un nouveau jeu de données. Cette méthodologie repose notamment sur la définition d'une nouvelle métrique appelée Mean Distance to the Best (MDB), pour évaluer et orienter le choix de la méthode la plus adaptée à de nouveaux ensembles de données. Dans la dernière section du Chapitre 2, nous avons appliqué le mDAE aux données cliniques en choisissant les hyperparamètres adaptés à ces données.

Dans le Chapitre 3, nous avons, dans un premier temps, proposé une étude du problème de classification binaire individus sains versus patients pathologiques. Au cours de cette étude, nous nous attachons à comparer les résultats sur la matrice obtenue par imputation des données d'origine avec notre méthode mDAE, avec la matrice obtenue par agrégation des données d'origine de la thèse de Nolwenn TAN. Nous avons utilisé l'ACP pour faire une analyse descriptive des données puis nous avons appliqué deux classifieurs binaires (SVM et FA) dont nous avons analysé les performances. Nous concluons que la matrice sans don-

née manquante obtenue par notre méthode d'imputation donne de meilleurs résultats de classification binaire. Notre contribution apportée par la méthode mDAE apporte un réel bénéfice pour l'utilisation des données cliniques dans le projet HELP. Pour approfondir la recherche d'une méthode pouvant résoudre le problème de classification binaire sains versus pathologiques nous avons proposé notre seconde contribution. Nous avons utilisé une méthode basée sur un DAE qui a la capacité de reconstruire des données bruitées. Ce DAE a été entraîné à apprendre la structure de patients sains. Puis nous avons utilisé ce DAE pour reconstruire des nouveaux patients pathologiques en se basant sur l'hypothèse qu'ils seraient moins bien reconstruits et donc qu'il serait possible d'identifier les patients pathologiques par rapport à des individus sains bien reconstruits. Cette méthode donne des résultats (distributions des erreurs de reconstruction différentes entre individus sains et patients pathologiques, valeurs d'AUC autour de 0.8) plutôt prometteurs. Dans la dernière section du Chapitre 3, nous proposons des résultats préliminaires sur le problème de classification multi-classes où chaque classe est une pathologie. En utilisant un classifieur SVM, les résultats obtenus ne permettent pas de discriminer les pathologies entre elles. Finalement à la fin de ce chapitre, nous avons conclu que, en l'état actuel des données extraites de l'ECG 128 HD, discriminer les patients sains des patients pathologiques donne des plutôt bons résultats. Cependant le défi reste de trouver une méthode qui réussit à bien discriminer les pathologies entre elles.

Un projet innovant et encore en phase d'exploration

Comme nous l'avons présenté dans ce manuscrit, ce travail de thèse a été le fruit d'un travail très exploratoire. Nous avons travaillé sur des données cliniques qui, pour plusieurs raisons, nous ont confronté à des défis majeurs. Cette thèse a commencé à un moment où le projet HELP a pris une nouvelle ampleur, de nouveaux chercheurs ont été engagés, de nouveaux objectifs ont été fixés. L'utilisation de la base de données elle-même a été un défi car c'est une base en constant changement. Les patients choisis et les pathologies à étudier ont toujours été en discussion entre les cliniciens responsables du projet. La qualité des données due aux enregistrements en milieu hospitalier a posé quelques difficultés également. La première difficulté étant la présence de données manquantes à laquelle nous avons répondu en proposant notre première contribution, la méthode mDAE. La seconde étant l'incertitude de la présence des informations pour prévenir la mort subite dans les données. Les variables extraites et la méthode pour les extraire ont été en constante réflexion durant ma thèse par les équipes d'ingénieurs travaillant dessus. Il n'y a donc jamais vraiment eu de base de données dont l'information extraite était sûre. Finalement comme le projet HELP est un projet unique et innovant il était difficile de trouver le bon point de départ. Ainsi avons-nous proposé dans ce manuscrit un travail qui apporte des résultats préliminaires à l'immense défi qu'est la prédiction de la mort subite. Notre travail soulève de nombreuses pistes que pourront suivre les équipes du projet HELP.

Perspectives

Dans cette section nous précisons les pistes de réflexion que peuvent soulever les travaux présentés dans ce manuscrit de thèse.

La méthode mDAE Nous avons développé cette méthode pour qu'elle fonctionne sur des données numériques tabulaires ; il serait intéressant de réfléchir à une adaptation pour d'autres types de données telles que des données mixtes (catégorielles et numériques). Pour comparer les capacités de reconstruction des méthodes, nous avons utilisé des jeux de données UCI qui sont de petite taille ; une comparaison sur des jeux de plus grande dimension (plus de variables et plus d'individus) serait à mener. Enfin nous avons développé le mDAE en bruitant les données par des valeurs choisies aléatoirement et mises à 0, mais il existe plusieurs techniques de bruitage qui pourraient être comparées.

L'imputation des données cliniques Avec la méthode mDAE nous avons imputé les données cliniques par sous-matrices associées aux signaux d'intérêt (unipolaires, bipolaires verticaux, bipolaires horizontaux, laplaciens). Nous aurions pu choisir d'imputer directement la concaténation des sous-matrices. Il aurait également pu être intéressant de regarder l'apport d'une imputation par catégorie d'individus (contrôles, différentes pathologies) : chaque pathologie et les patients contrôles seraient alors imputés par un mDAE adapté.

Détection des patients hors de la distribution des individus sains Dans le Chapitre 3, nous proposons une méthode basée sur les DAE qui sont utilisés pour apprendre la structure de patients sains puis pour reconstruire de nouveaux individus. Nous avons regardé la différence des distributions des erreurs de reconstruction de ces individus dans lesquelles nous avons mis en avant une différence. Nous proposons de réduire en dimension les jeux de données cliniques sur lesquels nous appliquons le DAE avec la méthode Gram-Schmidt : il serait intéressant de confronter d'autres techniques de sélection de variables. Dans cette méthode nous avons choisi empiriquement la proportion de variables à masquer mais ce choix pourrait être fait automatiquement avec la proposition d'une grille de proportions. Tout comme la structure du DAE que nous utilisons pourrait être confrontée à d'autres structures. Le DAE réussit bien à apprendre la structure des individus sains et donc la différence de reconstruction avec des patients pathologiques est bien visible. Il faudrait poursuivre l'étude de l'efficacité de cette méthode en l'appliquant également sur la matrice blocs entière et en étudiant la possibilité de discriminer les pathologies entre elles.

Problème de classification binaire Dans le Chapitre 3, nous présentons également une comparaison de résultats de classification binaires pour discriminer les individus sains des patients pathologiques. Cette étude nous a permis de confirmer l'apport de notre méthode mDAE sur les données cliniques. Nous montrons de bons résultats de classification binaire mais plusieurs pistes peuvent encore être étudiées. Il faudrait déjà, comme nous le mentionnions plus haut, avoir une base de données plus stable pour avoir plus de recul sur l'origine des erreurs de classification. Il faudrait reprendre avec un clinicien tous les patients mal

classés pour chercher si la raison du mauvais classement réside plutôt dans une mauvaise qualité d'enregistrement du signal ou dans une mauvaise labellisation que dans une erreur de l'algorithme d'apprentissage. Il serait également intéressant d'approfondir l'analyse descriptive qui pourra nous donner plus d'informations sur les liens entre les variables. Puis il s'agirait d'associer cette analyse descriptive à de la sélection de variables pour déterminer les variables les plus pertinentes pour répondre au problème de classification. Ce projet étant un projet en constante évolution et porté par des cliniciens, obtenir des indices d'explicabilité du problème binaire serait une véritable plus-value.

Problème de classification multi-classe Dans la dernière section du chapitre 3, nous avons proposé une première étude des résultats que donne l'étude de la discrimination entre pathologies. Les résultats sont mauvais ; encore une fois, il faudrait investiguer pour savoir si c'est plutôt une question de mauvaise méthodologie ou de manque d'informations dans la base de données. Nous le rappellons, résoudre la mort subite par un ECG 128 HD n'a jamais été fait, donc personne ne sait qu'elle est la bonne information qui permettra de répondre à cette problématique challengeant les médecins depuis plus de 50 ans.

Bibliographie

Roth, Gregory A et al. (2020). "Global burden of cardiovascular diseases and risk factors, 1990–2019 : update from the GBD 2019 study". In : *Journal of the American college of cardiology* 76.25, p. 2982-3021.

Fishman, Glenn I et al. (2010). "Sudden cardiac death prediction and prevention : report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society Workshop". In : *Circulation* 122.22, p. 2335-2348.

Murakoshi, Nobuyuki et Kazutaka Aonuma (2013). "Epidemiology of arrhythmias and sudden cardiac death in Asia". In : *Circulation Journal* 77.10, p. 2419-2431.

Tan, Nolwenn (2021). "Caractérisation et détection des micro-potentiels anormaux associés aux arythmies ventriculaires létales, par analyse des signaux électrocardiographiques haute-résolution enregistrés à la surface du torse. (Characterization and detection of abnormalpotentials from severe ventricular arythmias,by analyzing high-resolution body surfacesignals)". Thèse de doct. University of Bordeaux, France. url : https://tel.archives-ouvertes.fr/tel-03214939.

Vincent, Pascal et al. (2008). "Extracting and composing robust features with denoising autoencoders". In : *Proceedings of the 25th international conference on Machine learning*, p. 1096-1103.

Duan, Yanjie et al. (2014). "A deep learning based approach for traffic data imputation". In : *17th International IEEE conference on intelligent transportation systems (ITSC)*. IEEE, p. 912-917.

Gondara, Lovedeep et Ke Wang (2018). "Mida : Multiple imputation using denoising autoencoders". In : *Advances in Knowledge Discovery and Data Mining : 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*. Springer, p. 260-272.

Ryu, Seunghyoung, Minsoo Kim et Hongseok Kim (2020). "Denoising autoencoder-based missing value imputation for smart meters". In : *IEEE Access* 8, p. 40656-40666.

Malmivuo, Jaakko et Robert Plonsey (1995). *Bioelectromagnetism : principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, USA.

Dallet, Corentin (nov. 2017). "Caractérisation locale de la propagation de l'onde d'activation cardiaque pour l'aide au diagnostic des tachycardies atriales et ventriculaires : application à l'imagerie électrocardiographique non-invasive". Theses. Université de Bordeaux. url : https://theses.hal.science/tel-01941399.

Matteucci, C. (1842). “Sur un phenomene physiologique produit par les muscles en contraction”. In : *Ann Chim Phys.* 6, p. 339-341. URL : https://cir.nii.ac.jp/crid/1572543025237070464.

Einthoven, Willem (1906). “The telecardiogramme”. In : *Arch Int Physiol* 4, p. 132-141.

Wilson, Frank N et al. (1934). “Electrocardiograms that represent the potential variations of a single electrode”. In : *American Heart Journal* 9.4, p. 447-458.

Goldberger, Emanuel (1942). “The aVL, aVR, and aVF leads : a simplification of standard lead electrocardiography”. In : *American Heart Journal* 24.3, p. 378-396.

Virani, Salim S. et al. (2021). “Heart Disease and Stroke Statistics—2021 Update”. In : *Circulation* 143.8, e254-e743. DOI : 10.1161/CIR.0000000000000950. eprint : https://www.ahajournals.org/doi/pdf/10.1161/CIR.0000000000000950. URL : https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000950.

Taccardi, B, L De Ambroggi et C Viganotti (1976). “Body surface mapping of heart potentials”. In : *The theoretical basis of electrocardiology*, p. 436-466.

Mirvis, David M (2012). *Body surface electrocardiographic mapping*. T. 82. Springer Science & Business Media.

Coumel, Ph (1987). “The management of clinical arrhythmias. An overview on invasive versus non-invasive electrophysiology”. In : *European heart journal* 8.2, p. 92-99.

Richardson, P (1996). “Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the definition and classification of cardiomyopathies”. In : *Circulation* 93, p. 841-842.

Nademanee, Koonlawee et al. (2011). “Prevention of ventricular fibrillation episodes in Brugada syndrome by catheter ablation over the anterior right ventricular outflow tract epicardium”. In : *Circulation* 123.12, p. 1270-1279.

Haissaguerre, Michel, Josselin Duchateau et al. (2020). “Idiopathic ventricular fibrillation : role of Purkinje system and microstructural myocardial abnormalities”. In : *Clinical Electrophysiology* 6.6, p. 591-608.

Haissaguerre, Michel, Meleze Hocini et al. (2018). “Localized structural alterations underlying a subset of unexplained sudden cardiac death”. In : *Circulation : Arrhythmia and Electrophysiology* 11.7, e006120.

Buxton, Alfred E et al. (2006). “ACC/AHA/HRS 2006 key data elements and definitions for electrophysiological studies and procedures : a report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (ACC/AHA/HRS Writing Committee to Develop Data Standards on Electrophysiology)”. In : *Journal of the American College of Cardiology* 48.11, p. 2360-2396.

Ramanathan, Charulatha et al. (2004). “Noninvasive electrocardiographic imaging for cardiac electrophysiology and arrhythmia”. In : *Nature medicine* 10.4, p. 422-428.

Meo, Marianna et al. (2020). “Body surface mapping of ventricular repolarization heterogeneity : an ex-vivo multiparameter study”. In : *Frontiers in physiology* 11, p. 933.

Burnes, John E et al. (2000). “Noninvasive ECG imaging of electrophysiologically abnormal substrates in infarcted hearts : A model study”. In : *Circulation* 101.5, p. 533-540.

Kania, Michał et al. (2019). "High-resolution body surface potential mapping in exercise assessment of ischemic heart disease". In : *Annals of Biomedical Engineering* 47, p. 1300-1313.

Tan, Nolwenn et al. (2019). "Analysis of signal-averaged electrocardiogram performance for body surface recordings". In : *2019 Computing in Cardiology (CinC)*. IEEE, Page-1.

Hotelling, Harold (1933). "Analysis of a complex of statistical variables into principal components." In : *Journal of educational psychology* 24.6, p. 417.

McFarland, Dennis J et al. (1997). "Spatial filter selection for EEG-based communication". In : *Electroencephalography and clinical Neurophysiology* 103.3, p. 386-394.

Dupuy, Mariette, Marie Chavent et Remi Dubois (2024). *mDAE : modified Denoising AutoEncoder for missing data imputation*. arXiv : 2411.12847 [cs.LG]. url : https://arxiv.org/abs/2411.12847.

Van Buuren, Stef (2018). *Flexible imputation of missing data*. CRC press.

Little, Roderick JA et Donald B Rubin (2019). *Statistical analysis with missing data*. T. 793. John Wiley & Sons.

Mayer, Imke et al. (2021). *R-miss-tastic : a unified platform for missing values methods and workflows*. arXiv : 1908.04822 [stat.ME].

Troyanskaya, Olga et al. (2001). "Missing value estimation methods for DNA microarrays". In : *Bioinformatics* 17.6, p. 520-525.

Mazumder, Rahul, Trevor Hastie et Robert Tibshirani (2010). "Spectral regularization algorithms for learning large incomplete matrices". In : *The Journal of Machine Learning Research* 11, p. 2287-2322.

Van Buuren, Stef et Karin Groothuis-Oudshoorn (2011). "mice : Multivariate imputation by chained equations in R". In : *Journal of statistical software* 45, p. 1-67.

Stekhoven, Daniel J et Peter Bühlmann (2012). "MissForest—non-parametric missing value imputation for mixed-type data". In : *Bioinformatics* 28.1, p. 112-118.

Goodfellow, Ian et al. (2014). "Generative adversarial nets". In : *Advances in neural information processing systems* 27.

Yoon, Jinsung, James Jordon et Mihaela Schaar (2018). "Gain : Missing data imputation using generative adversarial nets". In : *International conference on machine learning*. PMLR, p. 5689-5698.

Kingma, Diederik P et Max Welling (2013). "Auto-encoding variational bayes". In : *arXiv preprint arXiv :1312.6114*.

Ivanov, Oleg, Michael Figurnov et Dmitry Vetrov (2018). "Variational autoencoder with arbitrary conditioning". In : *arXiv preprint arXiv :1806.02382*.

Mattei, Pierre-Alexandre et Jes Frellsen (2019). "MIWAE : Deep generative modelling and imputation of incomplete data sets". In : *International conference on machine learning*. PMLR, p. 4413-4423.

Peis, Ignacio, Chao Ma et José Miguel Hernández-Lobato (2022). "Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo". In : *Advances in Neural Information Processing Systems* 35, p. 35839-35851.

Pereira, Ricardo Cardoso et al. (2020). "Reviewing autoencoders for missing data imputation : Technical trends, applications and outcomes". In : *Journal of Artificial Intelligence Research* 69, p. 1255-1285.

Muzellec, Boris et al. (2020). "Missing data imputation using optimal transport". In : *International Conference on Machine Learning*. PMLR, p. 7130-7140.

Zhao, He et al. (2023). "Transformed distribution matching for missing value imputation". In : *International Conference on Machine Learning*. PMLR, p. 42159-42186.

Bengio, Yoshua et al. (2009). "Learning deep architectures for AI". In : *Foundations and trends® in Machine Learning* 2.1, p. 1-127.

Rubin, Donald B (1976). "Inference and missing data". In : *Biometrika* 63.3, p. 581-592.

Muzellec, Boris (s. d.). *MissingDataOT*. url : `https://github.com/BorisMuzellec/MissingDataOT`.

Cortes, Corinna (1995). "Support-Vector Networks". In : *Machine Learning*.

Platt, John et al. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In : *Advances in Large Margin Classifiers* 10.3, p. 61-74.

Breiman, Leo (2001). "Random Forests". In : *Machine learning* 45, p. 5-32.

— (2017). *Classification and regression trees*. Routledge.

— (1996). "Bagging predictors". In : *Machine learning* 24, p. 123-140.

Genuer, Robin, Jean-Michel Poggi et Christine Tuleau (2008). "Random Forests : some methodological insights". In : *arXiv preprint arXiv :0811.3619*.

Krogh, Anders et John Hertz (1991). "A simple weight decay can improve generalization". In : *Advances in neural information processing systems* 4.

Srivastava, Nitish et al. (2014). "Dropout : a simple way to prevent neural networks from overfitting". In : *The journal of machine learning research* 15.1, p. 1929-1958.

Dupuy, M, M Chavent et R Dubois (2024). "Denoising Autoencoders for The Detection of Patients Out of Distribution of Healthy Individuals". In : *Computing In Cardiology*.

Stoppiglia, Hervé et al. (mars 2003). "Ranking a Random Feature for Variable and Feature Selection". In : *J. Mach. Learn. Res.* 3.null, p. 1399-1414. issn : 1532-4435.

Hodges Jr, JL (1958). "The significance probability of the Smirnov two-sample test". In : *Arkiv för matematik* 3.5, p. 469-486.

Appendix

A Annexe

Method	breast	climate	sonar	iono	seeds	wine	blood
mDAE	**0.535** ± **0.012**	**1.005** ± **0.009**	**0.735** ± **0.011**	**0.793** ± **0.012**	**0.537** ± **0.026**	**0.862** ± **0.029**	0.761 ± 0.053
mDAE w/o modified loss	0.829 ± 0.021 (54.953%)	1.006 ± 0.008 (0.100%)	1.007 ± 0.007 (37.007%)	0.880 ± 0.013 (10.971%)	0.741 ± 0.029 (37.989%)	0.927 ± 0.027 (7.541%)	0.844 ± 0.052 (10.907%)
mDAE w/o optimal μ	0.538 ± 0.028 (0.561%)	1.023 ± 0.018 (1.791%)	0.764 ± 0.041 (3.946%)	0.832 ± 0.035 (4.918%)	0.563 ± 0.052 (4.842%)	0.885 ± 0.055 (2.668%)	**0.746** ± **0.054** (-1.971%)
mDAE w/o overcomplete	0.548 ± 0.014 (2.430%)	1.159 ± 0.014 (15.323%)	0.774 ± 0.013 (5.306%)	0.845 ± 0.016 (6.557%)	0.756 ± 0.203 (40.782%)	0.959 ± 0.028 (11.253%)	0.849 ± 0.106 (11.564%)

TABLE 6 – **RMSE moyen de la reconstruction (± l'écart type) pour $B = 8$ tirages aléatoires de 40% de valeurs manquantes artificielles MCAR.**

Method	breast	climate	sonar	iono	seeds	wine	blood
mDAE	0.484 ± 0.039	1.009 ± 0.012	**0.657** ± **0.033**	**0.834** ± **0.029**	**0.468** ± **0.057**	**0.829** ± **0.042**	**0.613** ± **0.187**
mDAE w/o modified loss	0.812 ± 0.066 (67.769%)	**1.005** ± **0.011** (-0.396%)	0.978 ± 0.026 (48.858%)	0.898 ± 0.028 (7.674%)	0.682 ± 0.088 (45.726%)	0.892 ± 0.061 (7.600%)	0.839 ± 0.299 (36.868%)
mDAE w/o optimal μ	**0.482** ± **0.038** (-0.413%)	1.033 ± 0.015 (2.379%)	0.686 ± 0.042 (4.414%)	0.880 ± 0.055 (5.516%)	0.485 ± 0.071 (3.632%)	0.888 ± 0.090 (7.117%)	0.637 ± 0.225 (3.915%)
mDAE w/o overcomplete	0.521 ± 0.035 (7.645%)	1.161 ± 0.013 (15.064%)	0.716 ± 0.030 (8.980%)	0.899 ± 0.046 (7.794%)	0.830 ± 0.294 (77.350%)	0.974 ± 0.065 (17.491%)	0.967 ± 0.345 (57.749%)

TABLE 7 – **RMSE moyen de la reconstruction (± l'écart type) pour $B = 8$ tirages aléatoires de 20% des valeurs manquantes artificielles de MAR.**

Method	breast	climate	sonar	iono	seeds	wine	blood
mDAE	**0.510 ± 0.032**	**1.005 ± 0.006**	**0.718 ± 0.017**	**0.804 ± 0.018**	**0.511 ± 0.040**	**0.830 ± 0.021**	**0.658 ± 0.090**
mDAE w/o modified loss	0.854 ± 0.045 (67.451%)	1.005 ± 0.008 (0.000%)	1.000 ± 0.024 (39.276%)	0.891 ± 0.025 (10.821%)	0.821 ± 0.052 (60.665%)	0.916 ± 0.027 (10.361%)	0.893 ± 0.155 (35.714%)
mDAE w/o optimal μ	0.546 ± 0.059 (7.059%)	1.033 ± 0.021 (2.786%)	0.766 ± 0.033 (6.685%)	0.846 ± 0.046 (5.224%)	0.537 ± 0.052 (5.088%)	0.925 ± 0.047 (11.446%)	0.668 ± 0.099 (1.520%)
mDAE w/o overcomplete	0.526 ± 0.023 (3.137%)	1.149 ± 0.015 (14.328%)	0.780 ± 0.024 (8.635%)	0.868 ± 0.028 (7.960%)	0.743 ± 0.230 (45.401%)	1.018 ± 0.140 (22.651%)	0.989 ± 0.232 (50.304%)

TABLE 8 – **RMSE moyen de la reconstruction (± l'écart type) pour $B = 8$ tirages aléatoires de 40% des valeurs manquantes artificielles de MAR.**

Method	breast	climate	sonar	iono	seeds	wine	blood
mDAE	**0.486 ± 0.029**	1.001 ± 0.006	**0.684 ± 0.013**	**0.829 ± 0.027**	**0.503 ± 0.032**	**0.805 ± 0.033**	**0.738 ± 0.176**
mDAE w/o modified loss	0.795 ± 0.048 (63.580%)	**1.000 ± 0.008** (-0.100%)	1.005 ± 0.019 (46.930%)	0.890 ± 0.033 (7.358%)	0.682 ± 0.084 (35.586%)	0.864 ± 0.043 (7.329%)	0.943 ± 0.248 (27.778%)
mDAE w/o optimal μ	0.518 ± 0.050 (6.584%)	1.024 ± 0.011 (2.298%)	0.698 ± 0.030 (2.047%)	0.836 ± 0.021 (0.844%)	0.531 ± 0.025 (5.567%)	0.839 ± 0.076 (4.224%)	0.772 ± 0.213 (4.607%)
mDAE w/o overcomplete	0.521 ± 0.025 (7.202%)	1.156 ± 0.016 (15.485%)	0.742 ± 0.028 (8.480%)	0.893 ± 0.055 (7.720%)	0.686 ± 0.229 (36.382%)	0.965 ± 0.091 (19.876%)	0.950 ± 0.288 (28.726%)

TABLE 9 – **RMSE moyen de la reconstruction (± l'écart type) pour $B = 8$ tirages aléatoires de 20% des valeurs manquantes artificielles de MNAR.**

Method	breast	climate	sonar	iono	seeds	wine	blood
mDAE	**0.543 ± 0.030**	**1.005 ± 0.006**	**0.738 ± 0.018**	**0.798 ± 0.018**	**0.524 ± 0.031**	**0.873 ± 0.042**	**0.737 ± 0.102**
mDAE w/o modified loss	0.866 ± 0.043 (59.484%)	1.007 ± 0.007 (0.199%)	0.998 ± 0.016 (35.230%)	0.893 ± 0.011 (11.905%)	0.800 ± 0.039 (52.672%)	0.950 ± 0.047 (8.820%)	0.885 ± 0.109 (20.081%)
mDAE w/o optimal μ	0.574 ± 0.048 (5.709%)	1.016 ± 0.012 (1.095%)	0.750 ± 0.039 (1.626%)	0.830 ± 0.035 (4.010%)	0.565 ± 0.051 (7.824%)	0.922 ± 0.039 (5.613%)	0.750 ± 0.115 (1.764%)
mDAE w/o overcomplete	0.559 ± 0.024 (2.947%)	1.158 ± 0.013 (15.224%)	0.796 ± 0.025 (7.859%)	0.859 ± 0.020 (7.644%)	0.827 ± 0.236 (57.824%)	1.000 ± 0.034 (14.548%)	0.927 ± 0.082 (25.780%)

TABLE 10 – **RMSE moyen de la reconstruction (± l'écart type) pour $B = 8$ tirages aléatoires de 40% de MNAR valeurs manquantes artificielles.**

B Annexe

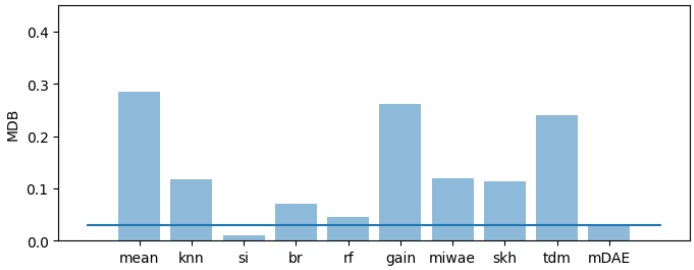
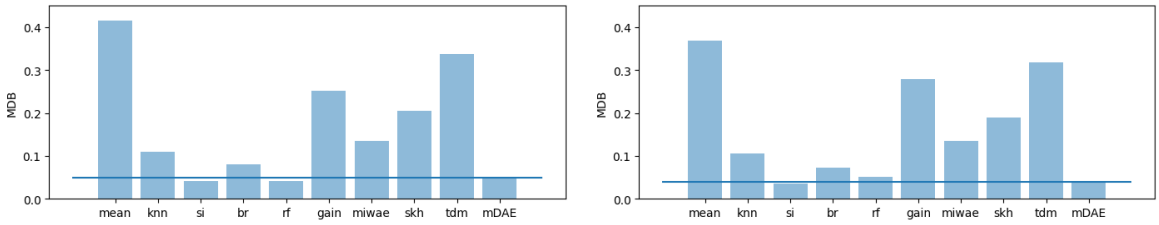


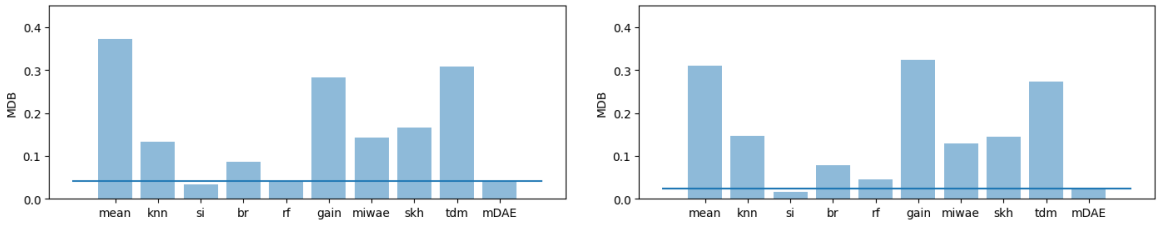
Figure 47 – **Mean Distance to the Best (MDB) obtenu avec 40% de données manquantes MCAR**



(a) MAR 20%

(b) MAR 40%

Figure 48 – **Mean Distance to the Best (MDB) obtenu avec des données manquantes MAR**



(a) MNAR 20%

(b) MNAR 40%

Figure 49 – **Mean Distance to the Best (MDB) obtenu avec des données manquantes MNAR**

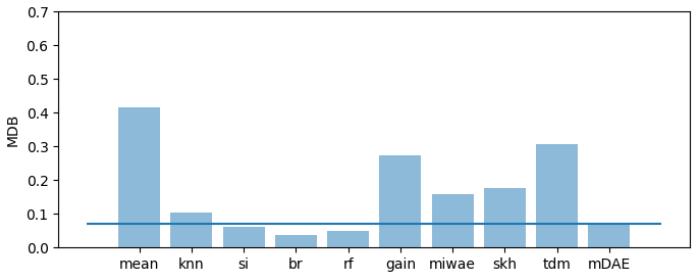


FIGURE 50 – **Mean Distance to the Best (MDB) obtenu avec 10% de données manquantes MCAR**