



**HAL**  
open science

# Apprentissage automatique pour la prédiction de performances : du sport à la santé

Rayane Elimam

► **To cite this version:**

Rayane Elimam. Apprentissage automatique pour la prédiction de performances : du sport à la santé. Informatique [cs]. IMT - MINES ALES - IMT - Mines Alès Ecole Mines - Télécom, 2024. Français. NNT : 2024EMAL0005 . tel-04805708

**HAL Id: tel-04805708**

**<https://theses.hal.science/tel-04805708v1>**

Submitted on 26 Nov 2024

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

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

**THÈSE POUR OBTENIR LE GRADE DE DOCTEUR  
DE L'INSTITUT MINES-TÉLÉCOM (IMT) -  
ÉCOLE NATIONALE SUPÉRIEURE DES MINES D'ALÈS (IMT MINES ALÈS)**

**En Informatique**

**École doctorale : I2S  
Portée par l'Université de Montpellier**

**Unité de recherche : EuroMov Digital Health in Motion**

**Apprentissage automatique pour la prédiction de performances : du sport à la santé**

**Présentée par Rayane ELIMAM**

**Le 18 Juillet 2024**

**Sous la direction de Jacky MONTMAIN et Stéphane PERREY et l'encadrement de Nicolas SUTTON-CHARANI**

**Devant le jury composé de :**

**Jacques PRIOUX, Professeur, ENS Rennes, M2S**

**Claire TOURNY, Professeure, Université de Rouen, CETAPS**

**Violaine ANTOINE, Maître de conférences HDR, Université Clermont Auvergne, LIMOS**

**Pierre DRUILHET, Professeur, Université Clermont Auvergne, LMBP**

**Jacky MONTMAIN, Professeur, IMT Mines Alès, EuroMov Digital Health in Motion**

**Stéphane PERREY, Professeur, Université de Montpellier, EuroMov Digital Health in Motion**

**Nicolas SUTTON-CHARANI, Maître de conférences, IMT Mines Alès, EuroMov Digital Health in Motion**

**Assia TRIA, Directrice IMT Mines Alès, IMT Mines Alès**

**Examineur**

**Examinatrice**

**Rapporteure**

**Rapporteur**

**Directeur**

**Codirecteur**

**Encadrant**

**Invitée**



## RÉSUMÉ

De nombreux indicateurs de performance existent en sport et en santé (e.g. réhabilitation) qui permettent de caractériser différents critères sportifs et thérapeutiques. Ces différents types de performance dépendent généralement de la charge de travail (à l'entraînement ou en rééducation) subie par les sportifs ou patients. Ces dernières années, beaucoup d'applications de l'apprentissage automatique au sport et à la santé ont été proposées. La prédiction, voir l'explication de performances à partir de données de charges pourrait permettre d'optimiser les entraînements ou les thérapies. Dans ce contexte, la gestion des données manquantes et l'articulation entre les types de charges et les différents indicateurs de performance considérés représentent les 2 problématiques traitées dans ce manuscrit à travers 4 contributions positionnées sur 2 applications : sport et santé.

Les 2 premières contributions du présent travail concernent la gestion des données manquantes par une modélisation incertaine réalisée sur (i) des données de football professionnel fortement incomplètes et (ii) des données COVID-19 bruitées artificiellement. Pour ces 2 contributions, nous avons associé des modèles d'incertitude crédibilistes, *i.e.* basés sur la théorie des fonctions de croyance, à différentes méthodes d'imputation adaptées au contexte chronologique des entraînements/matches et des thérapies. Une fois les données manquantes imputées sous formes de fonctions de croyance, le modèle crédibiliste des  $k$  plus proches voisins adapté à la régression a été utilisé de manière à tirer profit des modèles d'incertitudes incertains associés aux données manquantes.

Dans un contexte de prédiction de performances en match de handball en fonction des charges de travail passées, des modèles de régression multivariés sont utilisés pour prédire simultanément 7 indicateurs de performance athlétiques et techniques. C'est la troisième contribution. La dernière contribution (application en santé) concerne la rééducation de patients post-AVC ayant partiellement perdu l'usage d'un bras. De manière à détecter les patients non-répondants à la thérapie, le problème de la prédiction de différents critères de réhabilitation a permis de réappliquer les différentes contributions de ce manuscrit (imputation crédibiliste de données manquantes et régression multivariés pour la prédiction simultanée de différents indicateurs de performance.

**Mots clés:** *Performance, Apprentissage automatique, Imputation, Régression multivariée, Incertitude, Fonctions de croyances, Sport, Entraînement, Santé, Rééducation.*

## REMERCIEMENTS

Ces remerciements ont beaucoup d'importance pour moi. Cette thèse m'a fait grandir tant humainement que professionnellement. Je ne suis pas arrivé là seul. Donc, je tiens ici à remercier toutes les personnes dont j'ai croisé le chemin. Je commence tout d'abord par remercier mes rapporteurs, **Pr. Pierre DRUILHET** et **Dr. Violaine Antoine** ainsi que mes examinateurs **Pr. Jacques PRIOUX** et **Pr. Claire TOURNY**. L'interaction que j'ai eu avec vous était formidable. un immense merci pour votre bienveillance et aussi pour avoir accepté de faire partie de mon aventure. J'ai énormément apprécié nos échanges et toute la richesse scientifique et humaine qu'ils comportaient. **Monsieur Prioux**, vous avez été le président de mon jury et une personne formidable durant ma thèse qui a contribué à là où j'en suis aujourd'hui. Pour tout cela, un grand merci. Je souhaite aussi tout particulièrement remercier **Madame Assia TRIA** directrice de l'IMT Mines Alès qui m'a fait le grand honneur d'être membre invité de mon jury. **Madame TRIA**, je ne vous remercierai jamais assez pour tout ce que vous avez fait pour moi. Vous m'avez tendu la main au moment où j'en avais besoin, vous avez été et vous continuez à l'être une magnifique source d'inspiration pour moi. Mes remerciements ne suffiront jamais, sachez-le, votre bienveillance et l'empreinte de votre personne sont à jamais en moi, merci infiniment. Je souhaite aussi remercier **Mr. Pierre PERDIGUIER**. Maintenant, je souhaite adresser mes remerciements à ceux qui m'ont guidé et accompagné durant cette thèse. En vérité, je ne sais pas comment parler de mes directeurs de thèse **Jacky MONTMAIN** et **Stéphane PERREY** et de mon encadrant **Nicolas SUTTON-CHARANI**. Pour de vrai, j'ai le sentiment de quitter mon chez-moi tant vous avez été une famille avec moi. **Jacky**, tu n'en es peut-être pas conscient, mais tu n'as pas idée à quel point je te suis reconnaissant. Non seulement pour ton implication sans faille dans ma thèse, mais surtout pour l'humain que tu es. Parmi toutes les personnes que je connais, tu fais probablement partie de ceux qui n'ont pas conscience de à quel point ils sont géniaux. En plus d'être une personne au grand cœur, tu es un scientifique exceptionnel. **Stéphane**, tout comme Jacky, je retiens surtout l'humain que tu es. Tu m'as appris à naviguer dans des terres inconnues pour moi, tout ce qui concerne le sport, la santé, le mouvement. Au début ça me faisait un peu peur, aujourd'hui je suis tombé amoureux de ces domaines d'applications. Tu as aussi toujours été là pour me

rassurer et me redonner confiance. Merci pour tout. **Nicolas**, toi, c'est très spécial, je ne sais pas quels mots sont appropriés. Il me semble tout de même que le plus parlant, c'est de dire que tu es un frère pour moi, ou plutôt un grand frère. Tu m'as tellement appris, tu t'ai tellement investi... tu as été incroyable. Merci infiniment pour tout, merci de m'avoir fait grandir humainement et scientifiquement, merci d'avoir toujours cru en moi. J'espère qu'on aura d'autres aventures scientifiques tous les deux ! Je remercie ta petite famille pour toute la gentillesse dont ils ont fait preuve avec moi, merci à tous. Je vous souhaite tout le meilleur du monde.

Je souhaite remercier tout le laboratoire **EuroMov Digital Health in Motion**. Merci pour votre bienveillance, merci pour les échanges, merci pour les fous rires. Vous avez donné beaucoup de goût et de charme à mon aventure à Montpellier. Je remercie aussi l'école doctorale **I2S** pour leur écoute et leur soutien. Particulièrement, je remercie **Mr, Pascal NOUET** et **Mr, Christophe FIORIO**. Je remercie bien évidemment les membres de comité de suivi de thèse i.e. **Mr, Patrice GUYOT, Mme, Sandra BRINGAY**, pour leur bienveillance et leurs conseils tout au long de ma thèse.

Au cours de cette thèse, j'ai eu aussi à traverser certaines épreuves plus personnelles, et j'ai eu la chance d'avoir des amis à mes côtés e.g. les membres de la **CLIST** i.e. **Ludo, Alilou, Tristan et Alexandre**, du **Bureau 50** i.e. **Khoya Matthieu HAAS, Louis HOGNON, Pauline REBIERE, Christophe LATRILLE, Juliette LOZANO-GOUPIL AKA Kamila, Luc SOUILLA AKA Lucien, Nathan SINSOILLIEZ AKA 47, Theo VELLETAZ AKA Theophile**, ainsi que tous les doctorants. Un merci particulier à **Samar EZZINA, Camille MULLER, Andrii SMYKOVSKIY, Alexandre AKSENOV** ainsi que **Yann desmarais**. Je remercie chaleureusement mes amis de **l'Association des Thésards des Mines D'Alès** qui m'ont aidé à vivre un confinement particulier. Je remercie particulièrement **Jihane BOU-SLIHIM** et **Ali YADDADEN** qui m'ont énormément apporté humainement, mais aussi intellectuellement.

Je remercie particulièrement **Ludovic MOLLE** et **Mohamed ADJEL** pour leur soutien, j'allais dire précieux, mais c'est un euphémisme, dans une partie compliquée de ma vie et de ma thèse. Je n'oublie pas non plus toutes les personnes de **l'IMT Mines Alès** qui ont toujours été présentes quand j'en avais besoin, i.e. **Anne-Catherine DENNI, Edith TEYCHENE, Edwis RICHARD, Julie SYMCOVICZ, Laurie JOURDAS, Keltoum BELAROSSI**. Je demande pardon aux personnes que j'ai oublié de

citer, je vous remercie en espérant que vous m'excuserez. Bien sûr, je remercie mes amis d'enfance **El Banks** avec lesquels j'ai grandi et je prie Dieu pour continuer à grandir avec eux et pour qu'il préserve nos belles amitiés. Un merci tout particulier à mes amis **Salim BOUTLENDJj, Chakib SOUILAMAS, Cerine MOKHTARI** pour leur présence et leur soutien durant toutes ces années i.e. depuis un bon moment ainsi que leur présence précieuse durant ma soutenance de thèse. Cissou, je te remercie pour toute ton aide apportée depuis qu'on est petit notamment le fait d'avoir été parmi beaucoup d'autres choses, un excellent prof particulier au lycée, merci infiniment.

Je tiens à remercier **Loulou** pour sa sincère amitié, et l'épaule solide sur laquelle j'ai pu m'appuyer à maintes reprises. Merci pour tout, pour ta gentillesse et ton dévouement à toute épreuve dont j'espère être à la hauteur. Je remercie la famille **BOUSNINA**, plus particulièrement tata **Nora**, merci infiniment tata, c'est grâce à toi si j'ai pu entendre parler de cette thèse et rencontrer tant de personnes incroyables.

Je remercie toutes les personnes que j'ai côtoyées depuis mon arrivée en France, mes amis de l'Université Paris Cité, **Alaya et Aziz, Zaki, Akram, Mounim, Sliman, Aafaf, Dhoha, Vasanth, Joseph, Sami** et tous les autres. Je remercie mon acolyte de toujours **Imad BOUKHELIFA** pour son dévouement et son amitié sans faille. Je remercie toute ma promotion de Master 2 qui m'a fait vivre une année incroyable, et les personnes qui sont devenues aujourd'hui des valeurs sûres de ma vie, **Adel TALEB, Anis MOKHTARI, Abdou, Sid Ali CHORFI** et toute la team **Cergy et les hommes**. Je remercie aussi mes amies et compagnes de mes amis **Biba, Sonia et Sihem** pour toute leur gentillesse et sympathie depuis qu'on s'est connu à mon égard.

Je remercie **Wafa CHENNI** pour l'amie extraordinaire qu'elle est, pour son soutien à toute épreuve et sa bienveillance. Je n'oublie pas mes enseignants qui ont été une force dans ma vie. Je remercie alors **Béatrice BERNA** ma tata adorée aujourd'hui, **Mr, Mohamed NADIF** qui est sûrement une des personnes qui ont le plus cru en moi et qui m'ont inspiré. Je remercie tous mes enseignants pour la bienveillance avec un remerciement particulier pour **François ROLE** qui a mis toute son énergie dans notre projet de fin d'étude, J'adresse un remerciement particulier à ceux sans qui mon mémoire de Master 2 n'aurait jamais vu le jour i.e. **Mounim MOKHTARI, Aurore NADJEM, Lucie AYRAULT**, merci pour tout.

Je remercie l'équipe du Datalab de **GRTgaz** particulièrement l'équipe Datascience

avec laquelle j'ai eu une riche année d'alternance ou j'ai pu évoluer professionnellement et humainement. Je remercie infiniment **Gregory JARDON, Marketa PICHLOVA-LALLEMENTOVA, Valérie LAFITTE** ainsi que **Olivier LE BIHAN** pour tout ce qu'ils m'ont apporté durant cette année d'alternance.

Je remercie aussi tous mes enseignants de licence de l'Université Paris Cité pour tout ce qu'ils m'ont appris et pour leur bienveillance. J'adresse un remerciement particulier à **Michel SOTO, Neizle DORTA, David JANISZEK David DUBAYLE** sans oublier **Yannick PARCHEMAL** qui est celui par qui tout a commencé en France et surtout qui a bien voulu croire en moi en me donnant ma chance. Tous mes enseignants ont donné d'eux-mêmes et c'est grandement grâce à eux si j'en suis là aujourd'hui.

Je remercie tata **Lamia FETTACHE**, tata **Fouzia BOUMAIZA**, tata **Samia SAHRAOUI**, tonton **Zoheir GOUAREF** et tonton **Nordine KERKEB** paix à son âme, pour tout ce qu'ils ont fait pour moi et tout leur soutien. Je n'oublierai jamais et vous avez mon éternelle gratitude. Je remercie Madame **Flora STIENNE** qui m'a plus qu'aider pour entamer mon histoire en France. Je remercie mes amis *Zaki MAMERI* et **Mohamed SAARI**. Vous êtes des grands frères pour depuis tout petit et très certainement ma vie aurait été bien différente si vous n'aviez pas été là au bon moment. Merci infiniment.

Je remercie tata **Farida ANANI** et **Houda GOUAREF** pour toute leur bienveillance et leur encouragement. Je remercie bien évidemment **Hichem GOUAREF**, merci d'être toujours là pour moi depuis toujours, merci pour toute gentillesse. Je remercie **Tata Yamina** pour avoir fait preuve de tant de douceur et de gentillesse avec moi depuis mon arrivé à Paris et même avant.

Je remercie les enseignants de **l'EXIA CESI Algérie** pour la belle année que j'ai eu avec eux et le soutien qu'ils m'ont apporté, **Izem, Yasmine, Fayçal, Nabil** merci infiniment.

Je tiens à remercier toutes les personnes qui m'ont soutenu de la maternelle au baccalauréat, il y en a tellement, à tout cela j'adresse mon éternelle gratitude. Je remercie particulièrement Tata **Nabila** pour sa gentillesse sans limite depuis ma plus tendre enfance, **Smail MOKHTAR** qui m'a ouvert la porte de chez lui et qui m'a formé sans rien en retour, merci infiniment. Je remercie aussi particulièrement **Tata Yamina** qui a été ma prof particulière durant ma petite enfance et mon adolescence. Je remercie tous mes amis qui ont partagé avec moi ces belles années.

Je remercie la famille **Belabed** pour tout ce qu'ils ont fait pour moi et à tout le soutien qu'ils m'ont apporté. **Tata Mounia** je n'oublierai jamais le jour où tu es venu m'aider à ranger mon petit chez-moi de fortune alors que je venais d'arriver en France, merci pour toute ta bienveillance et ta gentillesse tata. **Skander** toi t'es mon frère quoi qu'il arrive, je ne vais pas m'étaler, tu sais déjà tout, je te remercie pour tout. Je remercie **Meriem BELABED** qui a été une sœur pour moi et une source d'inspiration. J'ai toujours été en admiration face à ton excellence et ta ténacité. Je souhaite avoir la joie de continuer à contempler ton ascension au sommet et pourquoi pas être coauteur un jour **Docteur Belabed**. Je te remercie aussi évidemment pour être à l'origine de ma rencontre avec ma chère femme à travers tes belles initiatives pour enrichir la culture scientifique des Algériens. Je remercie **Anis KARA** pour la personne débordante de bienveillance qu'il est.

Je tiens à marquer un temps d'arrêt, quelques jours après ma soutenance de thèse, **Nawel BELABED** nous a quitté. Je souhaite lui rendre hommage, cette dame est partie jeune. Ton départ nous a tous marqué et rappelé que nous quitterons tous ce monde un jour et que ce qui est important, c'est de profiter de la présence de ceux qui nous sont chers. Le croyant que je suis est convaincu que tu étais un ange incarnée femme et qu'aujourd'hui et pour l'éternité tu es dans un havre de paix. Paix à ton âme Nawel. Nchallah le Seigneur donnera tout le courage et la patience qu'il faut à la famille et aux proches de Nawel pour surmonter cette épreuve et que leurs cœurs soient apaisés.

Je remercie mon cousin **Ahmed Toubdarine**. Hmidou, on a traversé beaucoup de choses ensemble depuis notre arrivée en France. On se connaît plus que bien, je tiens à te remercier pour tout ce que tu as fait pour moi et pour ton soutien inconditionnel. Nchallah, nous pourrions conserver ce lien particulier même si chacun est occupé ailleurs, avec **Skander** bien sûr :)

Je remercie la famille **SZYKULA-SAURIN** i.e. **Alex, Sonja, Ines, Tonton Eric, Tata Irlande**, merci d'avoir été des voisins géniaux et d'avoir contribué à mon épanouissement. Je remercie aussi **Valentine NEGRONI** pour toutes nos longues conversations accompagnées après le travail et l'amie qu'elle a été. J'adresse le même remerciement à **Gladys MULNET**, en plus de ceux pour notre Week-end spéciaux à rien faire. Je remercie infiniment aussi ici tous mes amis de Paris, de Montpellier, d'Algérie et d'ailleurs que je n'ai pas pu citer.

Je remercie tata **Lucie Eva** et **Alain RABUT** pour avoir pris le relais à Montpellier et avoir été des voisins géniaux.

Je tiens à remercier les membre de la famille **FOUFA** qui ont toujours été présents. Merci à tonton **Nabil** et tata **Nabila** pour toute leur bienveillance à mon égard durant toutes ces années. **Rochdi, Mounir**, on se connaît depuis une éternité, je n'ai pas besoin de vous en dire plus, merci d'être des frères pour moi. **Anès**, dans ton cas c'est particulier, tu m'as accueilli à mon arrivée en France alors qu'on se connaissait que depuis peu. Tu m'as aidé à prendre mes repaires, et tant d'autres choses. Je te remercie infiniment toi ainsi que **Martin, Anthony et d'autres**. Vous avez été plus que géniaux avec moi, merci infiniment.

Je remercie la famille **BOUTLENDJ**, tata **Aicha** je n'oublierai jamais comment tu t'es battu pour moi au lycée pour que je sois avec mes amis. Je te remercie toi, tonton **Ali Karim** et **Sabrina** pour m'avoir toujours ouvert les portes de chez vous comme si qu'on était de la même famille. **Salim**, kifkif, tu sais tout, je n'ai pas besoin de rentrer dans les détails. Merci d'avoir toujours été là pour, et de toujours l'être encore, tu es une valeur plus que sûre dans ma vie.

Je remercie aussi la famille **BOURAS**, tonton **Mabrouk** tata **Hacina, Mehdi**, merci pour tous les beaux moments que j'ai vécus à vos côtés, vous m'avez certainement offert une bonne partie de mes plus beaux souvenirs d'adolescence. **Hani**, toi comme pour les autres, tu sais déjà tout. Je te remercie pour tout, merci d'être le **Han** qui répond toujours présent.

Je remercie **Azzedine LAHIANI**. Même si nous nous voyons plus aussi souvent, toi aussi, tu sais à quel point notre lien compte pour moi. Je te remercie pour tout frangin.

Je remercie un ami nouvellement connu, mais qui en peu de temps est devenu un ami précieux. Merci **Farid**.

Je remercie **Farah MEDJDOUB** pour l'amie exceptionnelle qu'elle est et toute ces années de soutien et d'encouragement. Tu sais à quel point je te suis reconnaissant Farah, merci infiniment.

Je remercie la famille **SOUILAMAS**, merci pour tous les bons moments vécus ensemble, pour toute votre bienveillance à mon égard ainsi que pour tous les fous rires qu'on a eus ensemble e.g. un bras de fer sur une table accrochée, ce n'est pas une bonne idée. **Chakib**, pareil, tu sais déjà tout. Je tiens à te remercier pour ta présence et ton

soutien inébranlable ces dernières années et pour tout le reste. Tu es parmi les membres fondateurs de la famille, tu m'as compris.

J'adresse aussi un remerciement à **Riad HADJSADOK**, qui est sûrement la personne qui me met le plus à mal en débat. Comme les autres, je n'ai pas besoin de te dire à quel point je te considère et à quel point je te suis reconnaissant de partager avec moi une belle amitié. Hâte d'assister à ta soutenance ! Je remercie aussi **Mehdi TAIEBEZZRAIMI** pour notre belle amitié, et merci d'avoir toujours cru en moi.

Je remercie la famille **TEBOUDELETTE**. sincèrement, **Tata Fatiha, Tonton Mohamed, Zinou, Rania, Fayçal** j'étais petit, la première fois que je suis venu chez vous, depuis, c'est aussi chez moi. Je me rappelle très bien comment j'avais peur de réveiller tonton quant moi, Fayçal et Zinou faisions des nuits blanches juste à discuter. Merci infiniment pour tout, et Nchallah, nous aurons d'autres occasions de construire de jolis souvenirs. **Fayçal, Zinou**, pareil, vous savez qui vous êtes pour moi et à quel point notre amitié (fraternité) est précieuse pour à mes yeux, merci pour tout. **Tonton Mohamed** affronte aujourd'hui une dure épreuve, mais il n'est pas seul et vous n'êtes pas seuls. Nchallah viendra le jour (très vite nchallah) ou toute cette pénible période appartiendra au passé et deviendra juste un mauvais souvenir.

Je remercie aussi la famille **BABA MOUSSA**, on ne se connaît pas depuis longtemps, mais vous êtes véritablement une bouffée d'air frais. **Hamid** tu sais déjà, merci pour tout, je sais que tu es là en toutes circonstances. Merci **Fouzi**, tu es mon poulain pour l'avenir, merci pour tout Fouzi, tu es mon petit frère ! Je remercie aussi Tonton **Mounir** et tata **Nadia** et **Yasmine**, vous m'avez donné beaucoup de joie en peu de temps et toi Nadiati même si je t'embête tu sais je t'adore ! Je remercie aussi **Fouzi** sans qui ces remerciements seraient illisibles tellement il y avait des fautes.

Je remercie toute la team **BABABOY** i.e. Sarah, Lydia, Ramy, Nassim, Nour, Fella pour toutes nos belles soirées pleines de rires, pour toute votre bienveillance qui a accompagné ma fin de thèse. Mention spéciale aux **Younsi's** Lydia et Sarah, n'oubliez pas que je suis MidYounsi.

Je remercie aussi la famille **BERKANI, Leila, Samia, Rachid, Nacer, Rachida, Mami** paix à son âme, merci d'avoir pour tout l'amour et la bienveillance que vous m'avez donné. **Fifi** toi, c'est à part, je sais à quel point tu m'aimes et tu sais à quel point moi et mes sœurs nous t'aimons, même si nous ne sommes pas liés par le sang, c'est

toi notre tante La fifi. Merci pour tout fifi, Dieu te garde et te préserve nchallah.

Je remercie la famille **FAHEM** i.e. **Nora, Djoudjou, Bahia** pour leur bienveillance et leur soutien. Un énorme merci à **Lila**, tu le sais, toi aussi, tu es un membre à part entière de la famille. Merci pour tout Lila, merci d'être une personne sur laquelle je peux toujours compter, merci d'avoir cru en moi et de m'avoir toujours encouragé, merci d'apporter autant de joie à notre maison.

Certaines personnes que je désires ardemment remercier ne sont malheureusement plus là. J'ai perdu un ami très cher lors de ma première année à Paris, **Oussama Elazhar DJOUDI**. Je souhaite lui rendre hommage à travers ces quelques mots. Depuis ton départ, tu n'as jamais, je dis bien, jamais quitté mes pensées. Je prie le Seigneur pour ton salut. Merci pour l'ami formidable que tu as été, paix à ton âme mon frère. Je rends aussi hommage à **Chahinez MESSLOUB** qui était une jeune femme pleine de vie, nous nous connaissions peu certes, mais je n'ai que les souvenirs des bons moments passés ensemble, paix à ton âme. Je rends aussi hommage à une personne que je n'ai vu qu'une seule, avec qui j'ai eu un échange marquant et qui est parti prématurément, paix à ton âme **Yacine**. En dernier, je souhaite aussi le salut et la paix à l'oncle de ma femme **Farid** qui est décédé durant la pandémie COVID 19, paix à son âme. Je souhaite à tous nos proches perdus d'être en paix.

Je remercie bien évidemment toute ma famille qui a été là depuis toujours i.e. **Mama kheira**, mes oncles, mes tantes (**Amina** la femme de mon oncle **Mourad** ∈ mes tantes) et mes cousins.

Je remercie la famille de mon épouse. j'ai débarqué dans votre vie de manière assez impromptue, vous m'avez tout de même accueilli et fais confiance. **Zaki, tata Fatima, Tonton Abdellah**, vous avez mon éternelle gratitude pour tout, que cela soit pour votre bienveillance ou encore la confiance que vous m'avez faite en m'acceptant malgré vos craintes, mon entrée dans votre famille comme beau-fils. Encore merci. Je remercie mon pays, l'Algérie et toutes les personnes qui m'ont tendu la main au cours de ma vie. Je remercie mon pays de cœur, **La France**, qui m'a accueilli comme si que j'étais Français. Je remercie tout particulièrement toutes les personnes qui ont œuvré pour que j'ai un jour la possibilité d'avoir une vision. Un grand merci à la Clinique de Créteil et au Docteur **Thierry DELAIRE** qui m'ont opéré gracieusement, merci infiniment. Un grand merci aux Médecins qui m'ont suivi à Montpellier. Un énorme merci au Docteur

**Sabah GRAOUCH** à qui je dois très probablement de pouvoir marcher aujourd'hui. Encore merci.

Maintenant, il faut que je remercie ma femme **Rofaida RABAHI**. J'avoue que ce n'est pas simple de décrire en quelques mots comment je te suis reconnaissant. Tu m'as toujours poussé vers l'avant (jusqu'à maintenant). Tu fais tout pour que j'aille vers mes rêves même si je suis têtu comme pas possible. Tu m'as permis de grandir, de mieux me connaître à travers notre amour. Sans oublier qu'ensemble, même si nous avons 2 ans de mariage, nous avons construit des souvenirs incroyables comme si nous en avions 10. Je te remercie ma chérie pour tout et je pèse mes mots quand je dis tout. Ah bien sûr, je dois préciser que tu es ma partenaire particulière qui m'aide à atteindre mes objectifs même les plus fous et que nous avons une complicité qui je pense est très très rare avec  $p$  value  $< 0.001$ . Nchallah, j'assisterai à tous tes accomplissements qui seront grandioses, j'en suis sûr. Merci encore mon bibi.

Arrive le moment où je dois remercier ma petite famille. Je ne sais vraiment si j'ai les mots justes pour exprimer toute ma reconnaissant tant l'émotion et les larmes montent en moi. Je commence par remercier mes petites **Aida** et **Lyna**. Je ne vous le dis certainement pas assez, je vous aime de tout mon cœur. Je vous remercie pour tout et surtout pour notre précieuse complicité à l'épreuve du temps. J'ai bientôt 30 ans, et j'arrive encore à être un gamin avec vous, gamin certes, mais grand frère :). **Aida**, je suis fier de toi, et je sais que tu vas t'accomplir dans la vie, j'ai foi en toi. Nchallah chaque jour dans ta nouvelle aventure en France sera un pas de plus vers ton épanouissement. Je t'aime fort Doudi. **Lyna**, tu es encore toute jeune et pourtant capable de tellement de chose (génie du crime aussi). Ah et n'oublie pas que moi, je sais que sous tes airs de Caïd, tu es une boule d'amour et de douceur. Je t'aime fort Layoun. Comme pour ta grande sœur, je sais que ta vie sera une belle aventure dont je suis plus qu'heureux d'être acteur.

Papa AKA **Lyess**, je n'ai pas vraiment les mots, je vais tout de même essayer. Je te remercie pour tout Papa, tout l'amour que tu m'as donné, toute ta bienveillance, toutes les choses que j'ai apprises de toi, tous mes caprices que tu supportais quand j'étais petit comme quand je boudais pour avoir une console. J'aurais beau écrire, je ne saurais jamais te dire merci. Merci pour le Papa exceptionnel que tu es. Merci d'avoir toujours cru en moi, merci de m'avoir inspiré, de m'avoir appris qu'il fallait être bons avec tout

le monde. Tu es un exemple d'humanité Papa, je persiste à penser que tu es unique et que des comme toi ça n'existe pas. Je t'aime fort Papa.

Maman AKA **Samira**, tu sais maman que je ne trouve jamais les mots qui sonnent juste pour t'exprimer ma reconnaissance ou parler de toi. Je peux dire que si je suis aujourd'hui ce que je suis, je te le dois principalement à toi. Merci pour tout maman, merci pour tous les combats que tu as menés pour moi. Merci pour toutes les valeurs que tu m'as inculquées, merci de m'avoir appris que tout est possible et que Dieu est grand. Merci de m'avoir appris à vivre avec mes yeux en les acceptant, non pas en les rejetant. Merci d'avoir fait de chez nous notre QG à moi et mes amis. Merci d'être l'héroïne de toute ma vie. Si un livre était écrit sur ma vie, ça serait toi le héros. J'écris en ayant les larmes aux yeux, je tiens tout de même à te dire aussi que tu dois sûrement être la Maman la plus extraordinaire qui puisse être. Nchallah tout ton futur sera fait que de belles choses et d'accomplissements. Je t'aime fort Maman.

Je souhaite aussi remercier toutes les personnes qui ont été un obstacle pour moi, car peut-être voir très probablement sans eux, je ne serais pas celui que je suis. C'est un remerciement sincère, non pas ironique. Je crois sincèrement que peu de personne pensent à mal, mais plutôt que parfois nos visions divergentes nous mènent à être des obstacles les uns pour les autres. Je demande aussi sincèrement pardon à toutes les personnes que j'ai pu heurter, mal comprendre ou faire du mal au cours de ces années, même si c'était involontaire. Sincèrement, pardon.

Pour terminer, je remercie Dieu, pour les cadeaux qu'il m'a fait, la vie que j'ai menée, la famille que j'ai, les gens que j'ai rencontrés. Sans oublier les yeux que Dieu m'a donnés Contrairement à ce qu'on pourrait penser, mes yeux ont été plus une bénédiction pour moi qu'autre chose. J'espère être un jour digne de tous les miracles que j'ai vécus et tous les bienfaits que j'ai reçus.

Place: Montpellier

Date: 17/09/2024

**Rayane ELIMAM**

# Table des matières

<b>RÉSUMÉ</b>	i
<b>REMERCIEMENTS</b>	iii
<b>TABLE DES FIGURES</b>	xvi
<b>LISTE DES TABLEAUX</b>	xx
<b>GLOSSAIRE</b>	xxii
<b>1 Contexte et objectifs</b>	<b>5</b>
1.1 Contexte général	5
1.2 Concepts et questions de recherche	7
1.2.1 Charge de travail	7
1.2.2 Performance	12
1.3 Relation charge de travail → performance	14
1.3.1 Charge externe < – > Charge interne	14
1.3.2 Charge de travail et valeurs manquantes	18
1.3.3 Charge de travail et performance en match	20
1.4 Conclusion	24
<b>2 Théories et formalisme</b>	<b>28</b>
2.1 Apprentissage supervisé	30
2.1.1 Données	30
2.1.2 Classification	31
2.1.3 Régression	32
2.1.4 Protocoles d'évaluation	33
2.1.5 Algorithmes usuels	34
2.1.6 Régression multsortie	37
2.2 Théorie des fonctions de croyances	40

2.2.1	Motivations . . . . .	40
2.2.2	Formalisme . . . . .	40
2.2.3	K plus proches voisins crédibiliste . . . . .	43
2.3	Imputation de valeurs manquantes . . . . .	44
2.3.1	Imputation incertaine . . . . .	46
<b>3</b>	<b>Modélisation incertaine des valeurs manquantes</b>	<b>48</b>
3.1	Introduction . . . . .	48
3.2	Prédiction de charge interne de joueurs de football . . . . .	48
3.2.1	Problématique et données . . . . .	49
3.2.2	Approche crédibiliste . . . . .	53
3.2.3	Expériences . . . . .	55
3.2.4	Perspectives . . . . .	59
3.3	Imputation crédibiliste pour la prédiction de mortalité quotidienne par COVID 19 . . . . .	60
3.3.1	Formalisme de la tâche prédictive considérée . . . . .	62
3.3.2	Extension de l'EKNN à la régression . . . . .	62
3.3.3	Problématique de l'imputation temporelle . . . . .	63
3.3.4	Méthodes d'imputation . . . . .	63
3.3.5	Modélisation de l'incertitude pour les méthodes d'imputation . . . . .	66
3.3.6	Expérimentation . . . . .	67
3.3.7	Résultats . . . . .	71
3.4	Conclusion . . . . .	74
<b>4</b>	<b>Régression multisortie pour la prédiction de performance en hand- ball féminin</b>	<b>77</b>
4.1	Introduction . . . . .	77
4.2	Prédiction de performances en Handball féminin . . . . .	78
4.2.1	Données et problématique . . . . .	78
4.2.2	Résultats . . . . .	81
4.2.3	Discussion et conclusion . . . . .	89
<b>5</b>	<b>Applications cliniques en rééducation de patients post-AVC</b>	<b>91</b>
5.1	Introduction . . . . .	91

5.2	Analyses préliminaires . . . . .	93
5.2.1	Description des données . . . . .	93
5.2.2	Analyse de corrélation . . . . .	98
5.2.3	Trajectoires de performance et de charge de travail du jeu Lady-BugPre . . . . .	99
5.3	Prédiction du RPE . . . . .	101
5.3.1	Problématique . . . . .	101
5.3.2	Expérimentation . . . . .	102
5.3.3	Imputation crédibiliste . . . . .	103
5.3.4	Résultats . . . . .	103
5.4	Prédiction de performances aux prochaines sessions . . . . .	106
5.4.1	Expérimentation . . . . .	107
5.4.2	Résultats . . . . .	108
5.5	Discussion et Conclusion . . . . .	112
	<b>BIBLIOGRAPHIE . . . . .</b>	<b>117</b>
	<b>LISTE DES PUBLICATIONS . . . . .</b>	<b>136</b>

# Liste des Figures

1.1	<b>Représentation de la relation entre la charge de travail avec des indicateurs de performance</b> . . . . .	7
1.2	<b>Relation entre les différentes charges dans le processus d'entraînement</b>	9
1.3	<b>Exemples d'indicateurs de charges internes et externes</b> . . . . .	10
1.4	<b>Schéma représentatif de la relation charge de travail → performances dans un contexte de sport collectif</b> . . . . .	25
2.1	<b>Algorithme des <math>k</math> plus proches voisins : Nous voulons déterminer la charge de travail d'une activité, la nouvelle observation doit être assignée à un label <math>y_i \in \Omega^Y = \{RPE\_faible, RPE\_moyen, RPE\_fort\}</math></b> <b>Le point rouge est la nouvelle observation à classifier, les points verts appartiennent à la classe RPE faible, les points violets à la classe RPE moyen et les points bleus à classe RPE fort. Les <math>k=5</math> plus proches voisins de la nouvelle observation sont liés à elle par un segment sur la figure. Le vote majoritaire assignera donc la classe RPE fort à la nouvelle observation</b> . . . . .	35
2.2	<b>Algorithme Arbre de décision CART : Nous voulons déterminer si un athlète va se blesser et si oui ou non il va se blesser durant une activité. Les critères sont l'âge, le nombre d'accélération, la charge de travail, si l'athlète a déjà été blessé dans le passé et la qualité du sommeil <math>x_i = (age, past\_injuries, sleep\_quality_i, training\_load_i, number\_accelerations_i)</math></b> <b>les vecteurs d'attributs sont associés à des labels <math>y_i \in \Omega^Y = \{no, shoulder, knee, ankle, sh\}</math></b>	
3.1	<b>Distribution du RPE</b> . . . . .	52
3.2	<b>Imputation chronologique des données</b> . . . . .	63
3.3	<b>Imputation par méthode LOCF</b> . . . . .	64
3.4	<b>Imputation par Mméthode CMA</b> . . . . .	64

3.5	imputation par méthode <i>TEKNN</i> . . . . .	66
3.6	<b>Schéma d'imputation crédibiliste</b> . . . . .	67
3.7	<b>Évolution des nouveaux décès et des nouveaux cas</b> . . . . .	68
3.8	<b>Données brutes : Évolution du nombre de nouveaux morts quotidien</b>	70
3.9	<b>Processus expérimental</b> . . . . .	70
3.10	<b>Erreurs d'imputation</b> . . . . .	72
3.11	<b>Résultats prédictifs sur des données imputées avec la méthode d'imputation</b> <b><i>TEKNN</i>, hyperparamètres: <math>K = 1, q = 4, v = 0</math></b> . . . . .	73
3.12	<b>Résultats prédictifs sur des données imputées avec la méthode d'imputation</b> <b><i>TEKNN</i>, hyperparamètres: <math>K = 1, q = 3, v = 0.3</math></b> . . . . .	74
3.13	<b>Performances prédictives par rapport aux niveaux de bruit pour les</b> <b>données imputées avec la méthode d'imputation CMA : <math>K = 20 : q = 4</math></b>	75
3.14	<b>Performances prédictives par rapport aux niveaux de bruit pour les</b> <b>données imputées avec la méthode d'imputation LOCF : <math>K = 10 : q = 4</math></b> . . . . .	76
3.15	<b>Performances prédictives par rapport aux niveaux de bruit pour les</b> <b>données imputées avec la méthode d'imputation <i>TEKNN</i> : <math>K = 1 : q = 4</math></b> . . . . .	76
4.1	<b>Évaluation chronologique : À chaque itération, les matchs précédents</b> <b>(en rouge) sont utilisés comme données d'apprentissage pour la prédiction</b> <b>des performances au match suivant (en bleu). La couleur grise</b> <b>représente la partie des matchs suivants qui n'est pas utilisée pour</b> <b>l'apprentissage. Les nombres situés sous les barres chronologiques</b> <b>correspondent aux indices des matchs</b> . . . . .	81
4.2	<b>Boxplots caractérisant les distributions des labels de performances</b> <b>: Le premier, second (<i>i.e.</i> médiane) et troisième quartiles sont re-</b> <b>spectivement représentés par les boîtes basse, milieu et haute</b> . . . . .	84

4.3	<b>Dynamiques des 7 labels de performances durant la 1<sup>ère</sup> partie de la saison : Chaque point représente la valeur moyenne d'un indicateur de performance (code couleur) pour une joueuse donnée à un match donné. Quand plusieurs joueuses ont les mêmes valeurs de performance (Pour un indicateur donné à un match donné) Les points correspondants sont fusionnés et la taille des points représente le nombre de points fusionnés. Des courbes lissées sont proposées (régression locale) avec un intervalle de confiance à 95%</b> . . . . .	85
4.4	<b>Dynamiques des 7 labels de performances durant la 2<sup>ème</sup> partie de la saison : Chaque point représente la valeur moyenne d'un indicateur de performance (code couleur) pour une joueuse donnée à un match donné. Quand plusieurs joueuses ont les mêmes valeurs de performances (pour un indicateur donné à un match donné), les points correspondants sont fusionnés et la taille des points représente le nombre de points fusionnés. Des courbes lissées sont proposées (régression locale) avec un intervalle de confiance à 95%</b> . . . . .	86
4.5	<b>Carte de chaleur des corrélations entre les indicateurs de charge de travail (en noir) et les indicateurs de performance en jeu (en rouge) : Les corrélations négatives sont représentées en bleu et les corrélations positives en rouge, l'intensité des couleurs représentant les valeurs absolues des corrélations. Les corrélations non significatives sont barrées</b> . . . . .	87
5.1	<b>Taux de valeurs manquantes par jeu/questionnaire en %</b> . . . . .	94
5.2	<b>Évolution des scores moyens à LadyBug</b> . . . . .	95
5.3	<b>Évolution des temps de jeu à LadyBug</b> . . . . .	96
5.4	<b>Évolution des path-ratios à LadyBug</b> . . . . .	96
5.5	<b>Évolution RPE</b> . . . . .	97
5.6	<b>Évolution notes</b> . . . . .	97
5.7	<b>Corrélations de Pearson entre indicateurs de performance et de charge de rééducation au jeu LadyBug</b> . . . . .	98
5.8	<b>Évolution score LadyBugPre par patient</b> . . . . .	99
5.9	<b>Évolution durées LadyBugPre par patient</b> . . . . .	100

5.10	<b>Évolution PathRatio LadyBugPre par patient</b>	100
5.11	<b>Évolution RPE par les patients</b>	100
5.12	<b>Évaluation chronologique (jeu de données non bruité)</b>	104
5.13	<b>Résultats prédictifs pour l'imputation par moyenne</b>	105
5.14	<b>Résultats prédictifs pour l'imputation par la méthode LOCF</b>	105
5.15	<b>Évaluation chronologique par modèle prédictif</b>	109
5.16	<b>Évaluation chronologique après suppression des données du patient "P10"</b>	111
5.17	<b>Evaluation chronologique du modèle FA-MULTI par indicateur après suppression des données du patient "P10"</b>	111

# Liste des Tableaux

2.1	<b>Matrice de confusion</b> . . . . .	32
3.1	<b>Distributions des variables d'entraînement</b> . . . . .	51
3.2	<b>Moyenne et écart-type des attributs personnels</b> . . . . .	52
3.3	<b>Comparaison des performance prédictives de la baseline et du <i>KNN</i> associées à différents modèles d'imputation avec labels précis pour différentes tailles d'historique <math>q</math></b> . . . . .	57
3.4	<b>Comparaison des performance prédictives de la baseline et de l'<i>EKNN</i> associées à différents modèles d'imputation avec labels précis pour différentes tailles d'historique <math>q</math></b> . . . . .	58
3.5	<b>Comparaison des performance prédictives de l'<i>EKNN</i> en fonction du modèle d'imputation incertaine considéré pour différentes tailles d'historique <math>q</math></b> . . . . .	58
4.1	<b>Moyenne (<math>\pm</math> écart-type) comptage de mesures dérivées d'accéléromètre à différents niveaux d'intensité</b> . . . . .	82
4.2	<b>Valeurs (<math>\pm</math> écart-type) des indicateurs propriétaire et du temps de jeu</b>	82
4.3	<b>Moyenne (<math>\pm</math> écart-type) comptage d'actions en jeu catégorisées en attack / defence et positive (+) / negative (-)</b> . . . . .	83
4.4	<b>Taux de rareté des labels de performances (3 athlétiques et 4 techniques)</b> . . . . .	84
4.5	<b>Performances prédictives issues de l'évaluation chronologique</b> . . . .	87
4.6	<b>Temps de calcul pour tous les modèles</b> . . . . .	88
5.1	<b>Statistiques descriptives : des indicateurs des jeux LadyBug (Score, Temps, PathRatio) et des questionnaires Pre et Post (RPE, Fatigue-Gen, FatigueBras, Douleur)</b> . . . . .	95

5.2	<b>RMSE moyen <math>\pm</math> écart-type par approche pour la prédiction du RPE (jeu de données non bruité)</b>	104
5.3	<b>aRMSE moyen <math>\pm</math> écart-type par approche pour les prédictions emboîtées de performances de rééducation</b>	108
5.4	<b>RMSE moyen <math>\pm</math> écart-type par modèle prédictif et par indicateur de performance de rééducation</b>	110
5.5	<b>aRMSE moyen <math>\pm</math> écart-type par modèle prédictif après suppression des données du patient "P10"</b>	110
5.6	<b>RMSE moyen <math>\pm</math> écart-type par modèle prédictif et par indicateur de performance de rééducation après suppression des données du patient P10</b>	112



## Introduction

### Le concept de performance

Depuis la nuit des temps, l'homme cherche à améliorer ses performances, que ce soit dans la chasse, la cueillette, la construction de son habitat ou encore dans les activités physiques. Cette quête de performance est intrinsèquement liée à notre nature et à notre volonté de survivre et de prospérer dans un monde en constante évolution.

Le concept de performance est polysémique et peut être défini de différentes manières. En sport collectif, la performance est généralement définie comme le résultat obtenu par une équipe *e.g.* la victoire lors d'une compétition. Pour un athlète donné, nous parlons de performance individuelle et celle-ci est dépendante du sport concerné. Par exemple, dans les sports chronométrés l'indicateur de performance d'un athlète sera classiquement le temps réalisé. Dans les sports collectifs, la performance individuelle est encore dépendante d'autres facteurs liés en particulier aux interactions avec les autres joueurs de l'équipe, *e.g.* en football le nombre de but marqués par l'avant-centre dépend du nombre de fois que ses coéquipiers l'auront servi durant le match, la saison, etc.

### Intelligence artificielle et Sport

L'intelligence artificielle (IA) a connu un essor fulgurant ces dernières années et ses applications se multiplient dans de nombreux domaines. Elle est utilisée dans l'aide à l'optimisation des performances dans une grande variété de domaines.

Dans l'industrie, l'IA est utilisée pour améliorer la production, la maintenance et la qualité des produits [1, 2, 3]. Dans le domaine de l'énergie, elle permet de mieux gérer la consommation et de réduire les émissions de gaz à effet de serre [4, 5]. En médecine, l'IA a un grand panel d'applications dont l'amélioration du diagnostic des maladies en particulier à travers l'analyse automatique d'images (scanner, IRM, etc.), le développement de nouveaux médicaments et la planification d'interventions chirurgicales [6]. De prime abord, ces usages ne se ressemblent pas, mais ils contribuent tous à aider l'homme à mieux faire une tâche, mieux *performer*.

## Applications de l'IA dans le sport

L'IA a également connu un grand intérêt de la part des acteurs du monde sportif pour l'aide à l'optimisation des performances à travers de nombreux usages lors des dernières années. Les outils utilisés en intelligence artificielle, et tout particulièrement l'apprentissage automatique (auquel se résume pour beaucoup l'intelligence artificielle du 21<sup>ème</sup> siècle), permettent de faciliter des processus complexes tels que l'analyse des performances des athlètes enregistrées en masse pour identifier les points d'amélioration, l'aide au développement de programmes d'entraînement personnalisés et optimisés grâce à une meilleure compréhension de la réponse physiologique des athlètes à l'activité sportive [7]. Une autre voie d'optimisation des performances en sport porte sur la prévention des blessures et l'amélioration de la récupération pour les athlètes [8]. Ces quelques éléments ne sont qu'une infime partie des usages faits de l'IA dans le monde du sport comme nous le verrons dans ce manuscrit.

Dans ces travaux de thèse, nous nous intéressons aux relations entre les différentes charges de travail que subissent les athlètes durant leurs activités physiques à l'entraînement, nous distinguons à cet effet l'effort produit et la difficulté ou l'intensité ressentie à l'exercice, mais aussi aux relations qui lient ces charges à la performance individuelle de l'athlète. Le premier chapitre de ce manuscrit fait un état des lieux de la littérature scientifique sur le traitement de ces problématiques. Le deuxième chapitre de ce manuscrit quant à lui introduira les différents outils mathématiques utilisés dans ces travaux de thèse pour gérer les flux de données relatives aux charges et performances d'un athlète et affectées par des incertitudes de différentes natures.

Une problématique commune aux différentes applications de l'intelligence artificielle est la nécessité de disposer de données, or dans le sport, certaines données sont renseignées par le biais d'une intervention humaine ce qui entraîne une problématique de gestion des *valeurs manquantes* lors de la collecte des données nécessaires à l'apprentissage des modèles d'IA. Par exemple, la quantification de la charge de travail ressentie par un athlète nécessite que celui-ci remplisse un formulaire à l'issue de son entraînement... cette contrainte est à l'origine de nombreuses valeurs manquantes dans le suivi de l'athlète. Nous proposons dans le troisième chapitre de ce manuscrit une approche pour répondre à cette problématique à l'aide d'outils mathématique basés sur la *théorie des fonctions de croyances*. Notre approche se distingue par le fait qu'elle permet d'associer un niveau d'incertitude aux valeurs de remplacement des valeurs manquantes (valeurs imputées). Notre approche est appliquée à deux différents jeux de données, un sur le sport, et le second sur la pandémie COVID-19.

L'activité physique étant au coeur du sport, il est envisageable de transposer les travaux utilisant l'IA dans le domaine du sport au monde de la rééducation pour des patients atteints de pathologies affectant leurs capacités physiques, e.g les patients qui

ont une sclérose en plaque, qui ont subi un accident vasculaire cérébral etc. La notion de performance accompagne le processus de rééducation. Même si les enjeux sont bien différents, les objectifs des professionnels du monde sportif et ceux de la rééducation physique se rejoignent dans la volonté d'améliorer les performances des patients/athlètes. C'est de cette transposition que nous traiterons au chapitre cinq de ce manuscrit.

## Sport et rééducation physique

Le sport et la rééducation partagent de nombreux points communs. Dans les deux cas, les thérapeutes/préparateurs physique cherchent à améliorer les performances des individus à travers des programmes de rééducation/entraînement où l'intelligence artificielle peut être utilisée pour aider à la personnalisation des programmes de rééducation/entraînement en fonction des individus et des résultats observés au cours du temps. Ainsi, l'IA permet-elle aux experts de mieux comprendre la réponse des patients/athlètes à leurs programmes d'entraînement ? Par exemple, lors d'un entraînement sportif ou d'une session de rééducation, les individus participent à des activités physiques dans le but d'améliorer une ou plusieurs fonctions. Pour un sportif, l'objectif peut être de devenir plus rapide en faisant régulièrement des activités qui vont exacerber cette dimension de la performance ; l'objectif pour un patient en rééducation peut être le même, il doit par exemple retrouver la dextérité d'un mouvement pour son bras parétique. L'intelligence artificielle propose des approches qui permettent une meilleure compréhension de la réponse des individus aux activités physiques auxquelles ils sont soumis. Les chapitres 4 et 5 de ce manuscrit traiteront des relations entre les charges de travail subies lors d'activités physiques et la performance individuelle chez l'athlète comme chez le patient. Ces relations sont modélisées à travers l'usage d'algorithmes d'apprentissage automatique.

## Apprentissage automatique

L'apprentissage automatique est une branche de l'IA qui permet aux ordinateurs d'apprendre à partir de données. L'apprentissage automatique supervisé est un type d'apprentissage automatique dans lequel les données d'apprentissage sont labellisées avec la réponse souhaitée, *i.e.* que l'on dispose des associations entre les variables observées (*e.g.* une durée passée à courir) et la présence ou l'absence d'une blessure (le label) sur un jeu de données, dit jeu d'apprentissage.

Dans le contexte de l'optimisation de la performance individuelle des athlètes, l'apprentissage automatique peut être utilisé pour mieux comprendre la performance des athlètes et sa relation avec les différents facteurs qui l'impactent. Ce questionnement relatif au sport

peut être transposé à la rééducation pour mieux comprendre comment rétablir au mieux les capacités perdues des patients à travers des programmes adaptés.

## Plan de la thèse

Le premier chapitre de cette thèse présente un état de l'art sur les problématiques liées à la performance sportive. Le deuxième chapitre introduit l'apprentissage automatique, la théorie des fonctions de croyances pour gérer les différentes incertitudes qui affectent les données sportives et l'imputation des valeurs manquantes. Dans le troisième chapitre, nous exposons des travaux sur l'imputation des valeurs manquantes et la modélisation de l'incertitude relative aux valeurs imputées. Le chapitre 4 est dédié aux travaux sur la prédiction de performances dans le cadre du Handball féminin d'élite. Le dernier chapitre propose une application pour prédire la réponse de patients post-AVC à un programme de rééducation en réalité virtuelle.

## Chapitre

### 1

## Contexte et objectifs

### 1.1 Contexte général

Aujourd'hui, les organisations sportives, médicales et autres ont accès à de plus en plus d'équipements leur permettant entre autres de collecter et réaliser un suivi de différents paramètres propres à l'activité physique et aux mouvements des sportifs [9]. Nous pouvons citer les systèmes de positionnement par satellite ou Global Positioning System en anglais (GPS) qui permettent de mesurer plusieurs indicateurs, *e.g.* la distance totale parcourue, la vitesse moyenne ou encore la distance parcourue dans une zone de vitesse spécifique. Nous pouvons aussi évoquer les centrales inertielle ou Inertial Measurement Unit en anglais (IMU) qui permettent d'obtenir un grand nombre d'indicateurs relatifs aux accélérations et aux impacts notamment. Ces deux exemples ne sont qu'une très petite partie de la pléthore d'appareils qui existent pour mesurer un grand nombre de paramètres/indicateurs liés aux activités physiques et sportives. Dans ce contexte, une des applications possibles est de pouvoir suivre et comprendre l'évolution des performances, performances qui peuvent s'évaluer au regard des objectifs arrêtés par les entraîneurs et/ou préparateurs physiques (*i.e.*, favoriser le développement des capacités physiques puis leur maintien pendant une saison sportive, tout en minimisant les risques de blessure,...) [10]. Les équipements évoqués ci-dessus permettent de mesurer de nombreux indicateurs associés aux qualités physiques, mais d'autres mesures y sont couramment associées : il s'agit des réponses physiologiques ou encore psychologiques [11]. Leur utilité première est de quantifier différentes caractéristiques relatives aux actions motrices, *e.g.* les efforts déployés par un athlète durant une activité (entraînement, match,...). Beaucoup d'applications ont vu le jour, en particulier dans le sport de haut niveau avec pour principaux objectifs la prévention des blessures associée ou non à l'optimisation des performances physiques et techniques [7, 10].

Durant une activité sportive, un athlète produit ce qui est dénommé communément une charge (au sens de dose) de travail, définie par le produit d'un volume de travail musculaire (*e.g.* nombre de répétitions) et une intensité de travail (*e.g.* pourcentage du niveau de force maximale ou de la vitesse de sprint) associant des réponses psychophys-

iologiques (*e.g.* consommation d'oxygène, fréquence cardiaque). La relation entre charge de travail et performance physique suscite depuis plusieurs décennies l'attention des organisations sportives, des structures scientifiques et médicales, donnant lieu à un large corpus de connaissances avec des finalités variées [12, 13, 14]. Cependant, comparativement aux sports individuels où l'endurance, la puissance et le rendement ont une sémantique communément admise (cyclisme, course à pied, natation), cette relation a été moins explorée dans les sports collectifs. En effet, elle demeure plus difficile à évaluer du fait de la présence d'interactions au sein de l'équipe et avec l'adversaire. Dans un sport collectif, plusieurs facteurs comme les interactions entre les joueurs d'une même équipe, le comportement de l'adversaire ou encore la stratégie de l'équipe sont autant d'obstacles à l'étude de la relation charge de travail - performance [15, 16, 17].

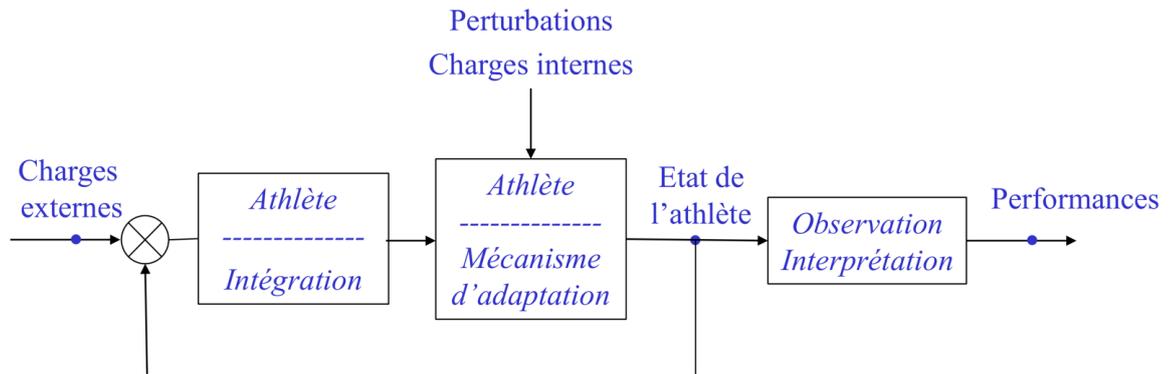
Les chercheurs en sciences du sport ont étudié plusieurs indicateurs de performances dans la littérature : athlétiques (*e.g.* la vitesse de course, le nombre d'accélération, le nombre de changements de direction, etc.), techniques (*e.g.* le nombre de buts inscrits, de passes réussies, etc.) ou encore subjectifs (*e.g.* la performance perçue par le coach). Hormis la victoire, la défaite ou le record, il n'est pas simple d'identifier un seul critère susceptible d'expliquer la performance individuelle. Cela est encore plus difficile dans un contexte où les individus sont interdépendants. Une majorité d'études évaluent l'impact des charges de travail sur les performances athlétiques des sportifs par le biais de tests spécifiques (*i.e.* qui évaluent une capacité physique comme la force, l'endurance, la vitesse) en cours de saison [18], ou en définissant un score de performance qui combine plusieurs indicateurs [19, 20]. Cependant, à ce jour, il n'existe que très peu d'études explorant l'effet des charges de travail à la fois sur les performances techniques et sur les performances athlétiques en condition de match dans les sports collectifs [21, 22, 20, 23, 24, 25, 26, 27, 28]. Par conséquent, l'objectif premier de cette thèse a été de caractériser la relation entre les charges de travail et les indicateurs de performances athlétiques et techniques en sports collectifs, afin de pouvoir la proposer au staff comme outil dans l'ajustement des contenus d'entraînement. Pour ce faire, nous avons cherché à construire un modèle basé sur l'apprentissage automatique qui met en relation la charge de travail et la performance individuelle des joueurs d'une même équipe. Ainsi, la connaissance de cette relation devrait permettre de concevoir et d'aider à améliorer un programme d'entraînement, une fois que l'on aura su expliquer l'influence des paramètres de cet entraînement sur le sens de variation des performances sportives. En pratique, les travaux menés dans cette thèse ont pour finalité l'aide à l'optimisation des performances des joueurs en sports collectifs.

Hormis l'utilisation de capteurs permettant la collecte d'indicateurs objectifs reliés aux actions motrices produites, des questionnaires sont classiquement utilisés en complément, afin de recueillir des indicateurs subjectifs comme l'évaluation du bien être ou de l'effort perçu lors de l'exercice [29, 30, 12, 31, 32]. La charge de travail apparaît donc

multidimensionnelle tout comme peut l'être la performance.

## 1.2 Concepts et questions de recherche

Plusieurs questions de recherche gravitent autour de la performance figure 1.1. La performance sportive est influencée par plusieurs facteurs dont la charge de travail, la qualité du sommeil, les qualités physiques [18], l'état psychologique [33], et l'état de forme physique [34]. D'autres facteurs indépendants de l'athlète peuvent aussi impacter la performance comme les conditions climatiques [35], le comportement de l'opposant, etc. Dans la mesure où un des objectifs de ce travail de thèse est d'explorer la relation entre la charge de travail et la performance, le schéma bloc qui suit essaie d'en fournir une représentation synthétique. La performance y apparaît comme l'effet observable de la réponse de l'athlète à des sollicitations externes, réponse qui sera modulée par l'état de forme de l'athlète.



**Figure 1.1. Représentation de la relation entre la charge de travail avec des indicateurs de performance**

### 1.2.1 Charge de travail

La charge de travail engendre des adaptations à différents niveaux chez l'athlète, qui auront des répercussions sur ses capacités physiques, ses performances athlétiques et sa sensibilité aux blessures, etc [36, 10, 37, 38]. Une explication simple de l'exercice physique peut être réduite à une relation dose-réponse. Dans cette relation, la dose peut s'assimiler à un volume de travail *i.e.* le stimulus (*e.g.* distance parcourue) subi par l'athlète, et la réponse à cette sollicitation correspond à l'adaptation psychophysiologique au niveau interne chez l'athlète [37]. Dit autrement, la charge de travail peut être définie comme la quantité d'adaptation interne subie durant un exercice physique pour une période de temps donnée [36]. Dans le milieu sportif, un athlète suit un entraînement régulier qui induit des adaptations au niveau de plusieurs sous-systèmes bi-

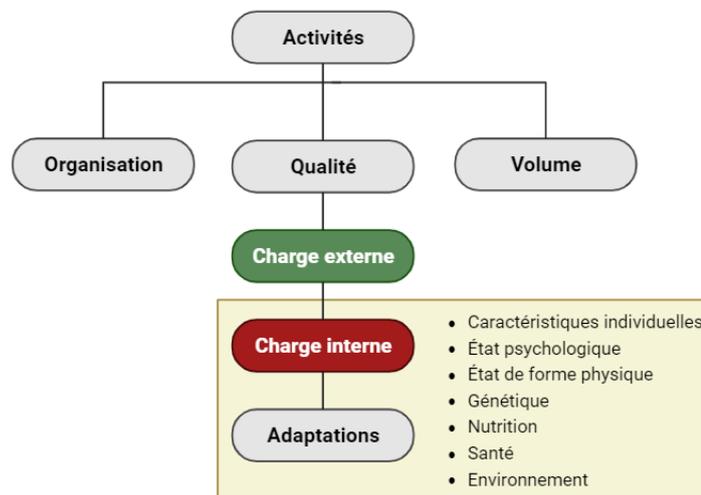
ologiques inter-connectés (musculaire, cardio-respiratoire, nerveux) [12, 39]. Les adaptations qui résultent des entraînements et des événements compétitifs sont associées aux variations de la performance, déterminées par des dynamiques différentes post-exercice de deux états : la fatigue et la condition physique (fitness). Il existe une classification des charges, telle que formulée par [40]. Cette classification permet de distinguer les charges externes *i.e.* les charges produites par l'athlète durant ses activités physiques (*e.g.* distance parcourue, durée de l'activité,...) et les charges internes à savoir les indicateurs relatifs à la réponse au niveau de l'athlète (*e.g.* consommation d'oxygène, fréquence cardiaque moyenne,...). L'adaptation propre à chaque individu est fonction de la charge interne, *i.e.* la somme des réponses psychophysologiques aux efforts produits [11].

#### 1.2.1.1 Charges externe et interne

La classification en charges externe et interne permet de distinguer les types de charge que peut subir un athlète durant un exercice physique. De nombreux indicateurs de charge sont mesurés grâce à l'apparition de nouveaux dispositifs/capteurs portables comme les GPS et les centrales inertielles ou IMU, ou encore les cardiofréquencemètres [41]. Un GPS est à l'origine un système américain de navigation et de localisation par satellite, et constitue par extension un appareil de géolocalisation par satellites désigné par GNSS. Une IMU est un dispositif électronique capable d'intégrer les mouvements d'un mobile, qui mesure et rapporte l'accélération, l'orientation, les taux angulaires et autres forces gravitationnelles. Il est composé de 3 accéléromètres, de 3 gyroscopes et de 3 magnétomètres. Un cardiofréquencemètre est un appareil de surveillance qui permet de mesurer/afficher la fréquence cardiaque en temps réel ou d'enregistrer la fréquence cardiaque pour une analyse ultérieure. Comme évoqué précédemment, la charge externe correspond aux efforts mécaniques des actions motrices produits par un athlète durant une activité physique [11]. Cette charge est de mieux en mieux quantifiée grâce aux nouvelles technologies portables cités ci-dessus (GPS couplé à une IMU). Plusieurs études en sports collectifs ont exploré les liens entre les charges externes et les risques de blessure sans contact [32, 42, 43], mais aussi avec la charge interne [44, 45, 46, 47, 48].

Rappelons que la charge interne est la somme des réponses psychologiques et physiologiques que subit un athlète pendant et à la suite d'un exercice physique. C'est en quelque sorte la réponse de l'organisme aux charges externes. La charge interne apparaît comme le principal facteur des adaptations de l'athlète durant une activité physique, et possède donc un impact fort sur l'état de forme (fitness) et la fatigue générée durant l'exercice physique [37, 11, 10, 38]. La charge interne permet de rendre compte de l'intensité de l'effort de l'athlète pendant une activité. Elle dépend de

facteurs physiologiques, psychologiques et environnementaux. En quantifiant la charge interne, les préparateurs physiques et entraîneurs cherchent à adapter l'entraînement aux besoins et capacités de chaque athlète, en évitant la surcharge ou la sous-charge qui peuvent conduire à augmenter le risque de blessure. Il existe plusieurs méthodes pour mesurer la charge interne, comme la fréquence cardiaque, la consommation d'oxygène, le lactate sanguin, etc. Cependant, l'échelle de perception de la difficulté de l'effort (ou Rating of Perceived Exertion : RPE) [49], bien que subjective, est la plus communément utilisée, pour son faible coût, sa simplicité, et le fait qu'elle permette d'avoir une appréciation globale de la charge interne qui dépend de plusieurs sous-systèmes (cardiaque, musculaire,...); la validité de cette correspondance ayant été appréciée par [50]. Beaucoup d'études ont établi la validité de l'utilisation d'indicateurs subjectifs pour la quantification de la charge interne, mais force est de constater que le RPE et ses dérivées demeurent les indicateurs de charge interne les plus étudiés [51, 52, 53, 54]. Les deux types de charges (externe et interne) possèdent leurs avantages et leurs limites [40]. La figure 1.2 montre un schéma représentatif d'un processus d'entraînement en soulignant le lien entre la charge externe et la charge interne et in fine les adaptations. Afin d'avoir une planification optimale des entraînements sportifs, la littérature propose un certain nombre d'études qui ont cherché à produire le meilleur modèle pour la prédiction de la charge interne en fonction des charges externes et des charges internes antérieures lors d'une activité en sports collectifs en prenant en compte les entraînements et les matchs [45, 46, 55, 56, 47, 57, 44].

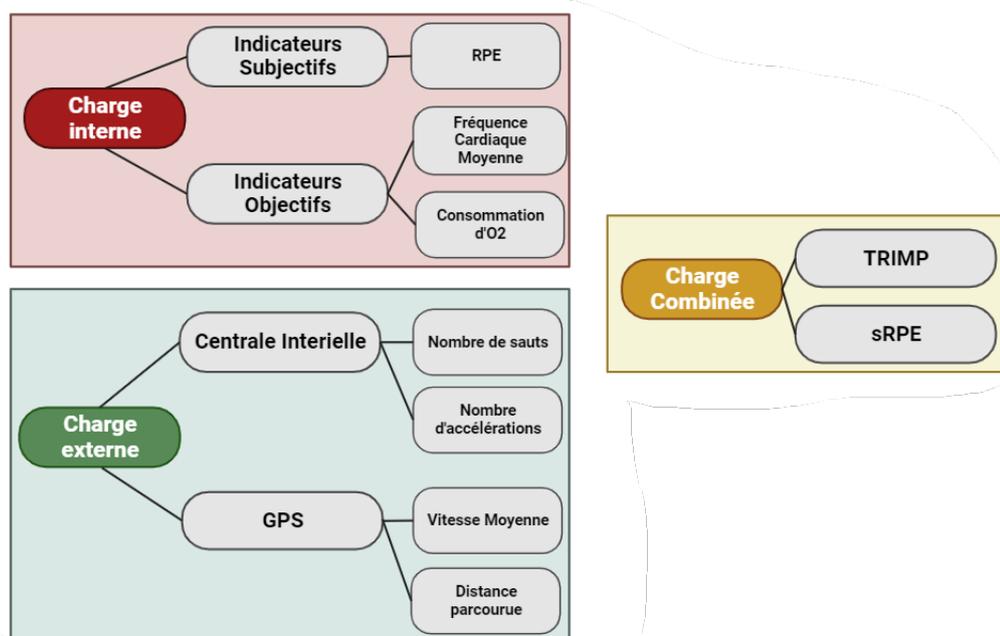


**Figure 1.2. Relation entre les différentes charges dans le processus d'entraînement**

Plusieurs indicateurs de charge interne et externe existent. La figure 1.3 montre quelques exemples d'indicateurs de charge, mais aucun ne pourra couvrir toutes les dimensions qui composent la charge de travail que subit l'athlète durant l'entraînement

ou le match. La quantification de la charge de travail a suscité et suscite toujours un fort intérêt dans les sciences du sport [30, 40, 11, 10]. [11] ont écrit, 15 ans après les avoir introduits, un premier commentaire sur les concepts de charge externe et interne. Ces concepts sont devenus très répandus dans la littérature des sciences du sport. Dans ce commentaire, les auteurs ont clarifié les définitions de charge interne et externe et également donné plus d'éléments en faveur de la pertinence de ces concepts dans le cadre du suivi des athlètes. Les auteurs ont souligné l'importance d'utiliser la charge interne lors du suivi des athlètes et ont discuté les limites de l'utilisation exclusive de la charge externe. Un second travail similaire [58] a été récemment proposé pour répondre aux critiques faites [59] au concept de charge de travail. La charge interne est recommandée comme indicateur principal car c'est cette charge qui module l'adaptation dans l'organisme, puisqu'une même charge externe n'aura pas la même réponse interne d'un athlète à l'autre. La charge interne (en tant que réponse) dépend de nombreux facteurs. Elle est fonction par exemple de variables de l'état psychologique, la santé, la qualité du sommeil, ou encore d'autres constantes comme la génétique d'un athlète, les conditions environnementales.

En résumé, les charges externe et interne peuvent être utilisées pour mieux comprendre le lien entre la charge de travail et la réponse des athlètes à leurs activités de manière individualisée. Ces concepts et les outils de suivi lorsqu'ils sont bien utilisés dans le contexte sportif, aident les entraîneurs et les préparateurs physiques à mieux contrôler et planifier l'entraînement.



**Figure 1.3. Exemples d'indicateurs de charges internes et externes**

### 1.2.1.2 Quantification de la charge de travail

Historiquement, les entraîneurs se basaient sur leurs années d'expérience pour planifier les entraînements. Aujourd'hui, les entraîneurs se tournent vers des méthodes issues du monde scientifique pour l'optimisation des programmes d'entraînement. Dans un film de 2011, *Le stratège* de Clint Eastwood, un sélectionneur d'une équipe de base-ball professionnelle aux Etats-Unis va s'appuyer sur des théories statistiques et engager un économiste amateur issu de l'Université de Yale, pour qu'ensemble, ils reconsidèrent la valeur de chaque joueur sur la base de statistiques, et réunissent une surprenante équipe composée de joueurs dits de seconde zone du base-ball professionnel en 2002 avant de connaître à nouveau le succès.

En 1991 les auteurs de [30] évoquaient le recours aux questionnaires et cahiers d'entraînement, et à un suivi physiologique ainsi qu'à l'observation directe (i.e suivi d'indicateurs directs comme la distance parcourue) par l'entraîneur pour le suivi des athlètes. L'utilisation d'échelles subjectives comme le RPE a été décrite comme étant la méthode la plus économique et la plus englobante. Mais ces échelles subjectives souffrent de problèmes relatifs à leur mauvaise usage, les réponses peuvent être mal renseignées par les athlètes de manière intentionnelle ou non intentionnelle. Ces faiblesses peuvent nuire à la précision relative de ces échelles. Cela peut s'expliquer par le fait que la réponse interne implique différentes dimensions (physiologique, psychologique, neuromusculaire,...) alors que les échelles donnent une note unique portant sur un ressenti global (bien être, difficulté, douleur,...). Malgré leurs faiblesses et après plus de 30 ans, les échelles subjectives sont toujours d'usage pour quantifier la charge interne et restent parmi les méthodes les plus efficaces. Les cahiers d'entraînement avaient été considérés plus fiables, mais leur inconvénient principal était la difficulté du traitement des données qu'ils génèrent. Le suivi physiologique avec la consommation d'oxygène, la fréquence cardiaque, la concentration de lactate dans le sang, etc. ont pour avantage de fournir des mesures objectives de l'intensité de l'activité. L'observation directe (charge externe uniquement) quant à elle permet un bon suivi, mais est moins viable pour un suivi longitudinal car elle ne prend pas en compte des indicateurs représentatifs de la charge interne pour chaque athlète. Cependant, ces méthodes avaient été jugées peu pratiques pour une utilisation à long terme. Les auteurs de [12] ont abordé le sujet plus tard, on y retrouve les échelles ou questionnaires, les cahiers d'entraînement, les marqueurs physiologiques et l'observation directe de charges externes et de nouveaux indicateurs de charge de travail. On retrouve par exemple parmi les indicateurs de charges de travail les plus connus comme le *Training Impulse* (TRIMP) [60, 61], qui est un indicateur utilisant des mesures de la charge interne (e.g. fréquence cardiaque). Les auteurs ont souligné que l'influence des adaptations physiologiques sur la performance était quantifiée avec peu de précision ce qui avait aussi été mentionné dans [30]. Cependant plus

récemment, de nouvelles technologies portables (IMU, GPS,...) sont apparues pour permettre à très court terme de suivre l'activité des athlètes de manière plus fine et de calculer un grand nombre d'indicateurs de charges externes en quasi temps réel. Grâce aux nouveaux équipements portables des indicateurs de charge de travail propriétaires sont apparus. On peut citer le PlayerLoad, qui correspond à la norme de la dérivée de l'accélération [62, 63, 64], ou encore le RHIE (Repeat High Intensity Efforts), qui est un indicateur de l'adéquation entre le repos (entre chaque effort de haute intensité) et le niveau de l'effort produit [65, 66]. Jusqu'à aujourd'hui, il n'existe pas d'indicateur de charge de travail qui fasse autorité, même si nous pouvons observer que le *Session Rating of Perceived Exertion* (sRPE) qui est le produit de la durée de l'activité (volume) par le RPE (intensité) demeure l'indicateur le plus communément utilisé [67]. Les entraîneurs et préparateurs physique cherchent aussi à répartir au mieux la charge de travail des athlètes selon le contexte (*e.g.* compétition, repos, présaison,...). Pour apprécier la répartition de la charge de travail d'un athlète, la quantification des charges de travail précédentes est la plus souvent réalisée sous la forme de moyennes faites sur des horizons courts (*e.g.* 7 jours précédents) qu'on nomme *charges aiguës* et sur des horizons longs (*e.g.* 21 jours précédents) qu'on nomme *charges chroniques*. Un autre indicateur existe. Il s'agit du *rapport charge aiguë/charge chronique* qui permet d'apprécier globalement la distribution de la charge de travail d'un athlète *i.e.* observer si sa charge de travail a été répartie sur une longue ou courte période.

### 1.2.2 Performance

L'objectif de tout entraîneur ou préparateur physique est d'améliorer une ou plusieurs capacités chez leurs athlètes tout en minimisant les risques de blessures. Plus particulièrement dans le contexte des sports collectifs, le but premier est la victoire qui elle dépend des performances individuelles des joueurs de l'équipe [12]. La performance quel que soit le sport implique plusieurs aspects, mais il est possible d'en sélectionner un seul comme par exemple dans les sports individuels chronométrés. La performance individuelle dans les sports collectifs est plus complexe. Il est très souvent attendu de l'athlète qu'il performe sur différents critères. Il faut distinguer deux principaux types de performances individuelles en sport collectif :

- Les performances athlétiques
- Les performances techniques

Ces deux types de performances recouvrent eux-mêmes un grand nombre de critères chacun. Ainsi ne retenir qu'un seul indicateur pour la performance n'est pas réaliste. Exceller dans son sport est l'objectif de tout athlète, les joueurs doivent se déplacer toujours plus vite, démontrer toujours plus de capacités techniques, tactiques, et athlétiques

[68]. La performance individuelle en sports collectifs englobe plusieurs aspects qui sont très liés au sport concerné. Dans une première approche, l'enjeu est de mieux comprendre la relation entre charge de travail et performance individuelle pour faire une planification optimisée des entraînements afin de maximiser la performance individuelle et par suite la performance collective.

La méthode d'évaluation de la performance individuelle la plus répandue en sport collectif repose sur l'utilisation de tests physiques [69], d'autant plus que nous sommes dans un contexte où l'information est omniprésente et accessible par l'intermédiaire de nouvelles technologies simplifiant la mise en place de tests [70]. Ces tests peuvent être répartis en deux catégories : des tests généraux et des tests spécifiques. Les tests généraux ont leur importance car ils permettent d'étudier les qualités physiques telles que la force, la puissance, l'endurance, la flexibilité, etc. L'objectif de ces tests étant de s'assurer de l'acquisition des niveaux athlétiques adaptés pour un joueur. Les tests généraux permettent d'une part d'évaluer les capacités physiques de l'athlète, et d'autre part de vérifier le taux de risque de blessures. Lorsqu'une qualité physique est insuffisante, le risque d'apparition de problèmes de surcharge et de mauvaise adaptation de l'organisme augmente favorisant l'émergence de blessures [71]. Les tests spécifiques, sont davantage relatifs aux besoins d'un athlète de manière ponctuelle. Ces tests sont spécifiques à des compétences *e.g.* aptitudes techniques, et ont un intérêt tout aussi important que celui des tests généraux. La distinction tests généraux et tests spécifiques était surtout orientée par le lieu de la réalisation du test, à savoir sur le terrain, dans le gymnase ou en contexte de laboratoire. Aujourd'hui, grâce aux technologies existantes comme les GPS portables, cette distinction devient obsolète. Certains tests nécessitent toujours la mise en oeuvre en laboratoire. Cela dit, les tests sur terrain sont moins coûteux et garde un avantage économique sur les tests en laboratoire [69].

Une autre méthode d'évaluation de la performance en sports collectifs s'appuie sur les actions réalisées (performances techniques) durant un match [72, 21, 20] ou encore d'autres indicateurs de performance tournés vers les performances athlétiques en match [22]. Nous allons dans ce manuscrit étudier conjointement les deux types de performances *i.e.* athlétiques et techniques. La performance individuelle en match est la plus complexe à évaluer en raison de la multiplicité des actions techniques réalisées en match ainsi que celle des mouvements athlétiques produits. Dans nos travaux la dimension technique est évaluée dans son aspect positif (actions positives *e.g.* nombre de duels gagnés) et son aspect négatif (actions négatives *e.g.* nombre de duels perdus). La dimension athlétique quant à elle, est évaluée par le nombre et le niveau d'accélération effectuées en match qui ont des liens avec les actions techniques [73, 74, 75, 76, 77](voir 4).

### 1.3 Relation charge de travail → performance

Dans cette section nous allons aborder plusieurs aspects de la relation charge de travail → performance. Nous allons tout d'abord aborder les travaux qui ont étudié la relation charge externe < – > charge interne. Nous poursuivrons sur le traitement des valeurs manquantes dans la quantification de la charge de travail. En effet, le suivi de la charge de travail a souvent recours à des échelles d'évaluation subjective comme le RPE pour mesurer la charge de travail interne, mais cette mesure étant dépendante du bon vouloir de l'athlète, la présence de valeurs manquantes est une problématique à considérer dans l'analyse des données. Nous aborderons enfin les différents travaux menés pour comprendre l'influence de la charge de travail sur la performance.

#### 1.3.1 Charge externe < – > Charge interne

Cette dernière décennie a vu beaucoup d'études menées dans le but de proposer une approche pour prédire la charge interne (généralement le RPE ou le sRPE) en amont des activités imposées à l'athlète et pouvoir comprendre ce qui l'impacte [45, 48, 47, 44]. En soit, les différentes études utilisent des modèles d'apprentissage automatique. Les entrées des modèles varient d'une étude à l'autre, *i.e.* d'une étude à l'autre les entrées des modèles varient en termes d'indicateurs, *e.g.* certaines n'utilisent que des charges externes, d'autres utilisent des indicateurs de charge interne comme la fréquence cardiaque.

Notons la charge externe (external load) lors d'une session  $i$   $el_i$ , et la charge interne (internal load) lors de la même session  $il_i$ . La charge externe est souvent relative à plusieurs indicateurs  $el_i == \{el_i^1, \dots, el_i^P\}$ . Il sera donc question de plusieurs indicateurs de charges externes en entrée des modèles utilisés dans les différentes études que nous allons aborder. Le but des études qui cherchent à prédire la charge interne est de construire une fonction  $f$  qui permette de faire le lien entre les entrées du modèle  $el^j$  et sa sortie  $il$  :

$$il_i \leftarrow f(el_i^1, \dots, el_i^P) \quad (1.1)$$

ou encore si on explicite la dynamique du modèle :

$$il_i \leftarrow f(el_i^1, \dots, el_i^P, il_{i-1}, il_{i-2}, \dots, il_{i-Q}) \quad (1.2)$$

Les objectifs principaux de ces modèles sont i) pouvoir prédire la charge interne à partir des charges externes, ii) d'analyser les modèles pour explorer l'influence des différentes charges externes sur la charge interne et enfin, iii) pouvoir utiliser ces modèles pour mieux planifier les entraînements. La charge interne, est le plus souvent assimilée

au RPE ou sRPE. Les mesures du pouvoir prédictif des modèles dans les différents travaux sont la racine de l'erreur quadratique moyenne pour la plupart et la moyenne des erreurs absolues pour d'autres. Dans les deux cas, plus l'erreur est proche de 0, meilleures sont les prédictions du modèle. L'apprentissage automatique et plusieurs des modèles utilisés sont présentés dans la section 2, où tous les modèles ont pour but de construire une fonction qui mette en relation les différentes entrées (*e.g.* les charges externes) avec la variable de sortie du modèle (*e.g.* la charge interne).

En 2017 l'étude de [45] a abordé la prédiction du RPE, sur la base d'un suivi de 41 joueurs professionnels de Football australien en match et en entraînement sur une période de 27 semaines. Ils ont comparé deux types de modélisations : une modélisation de groupe (*i.e.* un modèle pour tous les joueurs) et une modélisation individualisée (*i.e.* un modèle par joueur). Deux modèles ont été comparés : le réseau de neurones artificiels et les équations d'estimation généralisées. Le réseau de neurones artificiels est un algorithme d'apprentissage automatique supervisé inspirés du fonctionnement du cerveau humain. Il est constitué de couches de neurones artificiels connectés d'une couche à l'autre. Chaque neurone reçoit des entrées, les combine avec des poids, et produit une sortie selon une fonction d'activation. L'algorithme apprend les paramètres du modèle (poids) à partir de données d'apprentissage. Les équations d'estimations généralisées sont une méthode qui permet de modéliser des données corrélées, comme des mesures répétées. Elles étendent les modèles linéaires généralisés en tenant compte de la structure de corrélation entre les observations. La méthode se base sur la résolution d'une équation qui implique plusieurs paramètres dont la moyenne de la variable de sortie ainsi que d'autres paramètres relatifs aux variables en entrée du modèle.

Les résultats ont montré une supériorité des réseaux de neurones artificiels sur les équations d'estimation généralisée, et les meilleures performances sont obtenues avec la modélisation individualisée (erreur :  $1.24 \pm 0.41$ ). L'analyse des deux types de modèle et des deux types de modélisation a montré que les indicateurs de charge externe les plus déterminants pour la prédiction de la charge interne (RPE) étaient la distance totale parcourue lors de la session, la distance parcourue à haute vitesse et le rapport entre la distance parcourue à haute vitesse et la durée de la session. Les auteurs de [56] ont mené une étude similaire sur 38 joueurs de Football professionnels qui ont été suivis durant 2 saisons entières. Les réseaux de neurones ont été comparés à la régression LASSO, dans deux types de modélisation, de groupe et individualisée, comme pour [45]. La régression LASSO est une variante de la régression linéaire qui vise à minimiser le surapprentissage et qui intègre une contrainte qui met l'effet de certaines variables en entrée du modèle à 0. Un large éventail d'indicateurs de charges externes obtenus par GPS et accéléromètres a été utilisé en entrée des modèles. Les résultats obtenus ont montré une supériorité de la régression LASSO sur les réseaux de neurones artificiels. Contrairement à [45], les meilleures performances prédictives ont été obtenues avec la

modélisation de groupe. L'analyse des modèles a montré que les décélérations étaient des indicateurs de charge externe déterminants pour la prédiction du RPE.

Nous pouvons aussi évoquer les auteurs de [48] qui avaient proposé un travail similaire à celui de [45, 55], à la différence près que les modèles pour la prédiction du RPE avaient en entrée des indicateurs de charges externes (données GPS) et des indicateurs de charges internes (données fréquence cardiaque). Ils prenaient explicitement en compte le caractère dynamique du modèle. Dans cette étude, plusieurs modèles ont été comparés. L'algorithme forêt aléatoire a réalisé les meilleures performances prédictives (erreur :  $0.96 \pm 0.08$ ). Une sélection de variables a permis d'avoir un modèle simplifié qui utilise en entrée seulement la distance totale parcourue entre 18 km/h et 24 km/h et le produit de la distance totale parcourue avec la vitesse moyenne. Le modèle simplifié a obtenu un pouvoir prédictif légèrement inférieur (erreur :  $1.09 \pm 0.05$ ). Ces résultats suggèrent que les indicateurs de vitesse et de distance parcourue sont les plus importants pour prédire le RPE. L'algorithme forêt aléatoire est un algorithme d'apprentissage automatique qui se base sur un autre algorithme, l'algorithme arbre de décision, ce dernier construit des règles de décisions à partir des variables en entrée du modèle en définissant des seuils. L'algorithme forêt aléatoire consiste à faire plusieurs arbres de décision basés sur différents sous ensembles des données d'apprentissage.

Les travaux de [48, 45, 56] ont pu montrer l'intérêt de l'usage des techniques d'apprentissage automatique tant pour la prédiction de la charge interne que pour la compréhension de l'influence des différentes charges externes (et internes *e.g.* fréquence cardiaque [48]) sur les charges internes (RPE renseignés par les athlètes). D'autres études se sont penchées sur la question de l'influence des charges de travail externes et internes antérieures à une session  $i$  sur la prédiction du RPE d'un athlète lors de cette session  $i$ , les charges antérieures sont sous forme de charge aïgue, charge chronique, et rapport de charge aïgue/chronique. Dans l'étude de [47] l'objectif était la prédiction du RPE et du sRPE en fonction des charges aiguës, charges chroniques et des rapports des charges aiguës/chroniques en faisant usage de modèle d'apprentissage automatique. Les participants de l'étude étaient 22 joueurs de football professionnels qui ont été suivis lors de 160 sessions d'entraînement et 35 matchs sur une saison. Les charges de travail externes ont été collectées grâce à des appareils GPS. Une sélection de variables a été faite pour déterminer les indicateurs de charges de travail qui avaient la plus grande influence sur le RPE et le sRPE. Les résultats ont montré que les charges de travail aiguës (7 jours précédents) ont un effet important sur le RPE et le sRPE. D'autre part, l'analyse des prédictions a montré une plus grande précision pour la prédiction de valeurs moyennes de RPE et de sRPE par rapport aux valeurs extrêmes de RPE et de sRPE (*e.g.* les valeurs moyennes de RPE sont entre 4-7, les fortes valeurs extrêmes sont supérieures à 7, les faibles valeurs extrêmes sont inférieures à 4). Dans la même optique, l'étude de [44] avait pour objectif d'évaluer la capacité à prédire le RPE en fonction des

charges externes et internes antérieures. Pour ce faire, 38 joueurs de Football professionnels ont été suivis durant une saison, les charges externes (GPS et accéléromètres) et internes (RPE) ont été collectées durant 151 entraînements et 44 matchs. Plusieurs modèles d'apprentissage ont été comparés en termes d'erreur (racine de l'erreur quadratique moyenne) à travers une évaluation chronologique ou à chaque itération, le RPE de chaque joueur de la prochaine session était prédit. La position de jeu des joueurs a été ajoutée aux entrées des modèles. Les résultats de [44] ont montré une supériorité des modèles de forêt aléatoire (erreur : 1.1) et XGBoost (erreur : 1) sur les autres modèles utilisés, l'algorithme XGBoost est aussi basé sur l'utilisation d'un ensemble d'arbres de décision. Après analyse des modèles, un effet mémoire du RPE a été mis en évidence. En effet, les RPE renseignés sur le mois écoulé étaient des indicateurs déterminants pour la prédiction du RPE à la prochaine session, ces résultats méritent une attention particulière car ils indiquent que la charge interne (RPE) doit être considérée comme un processus dynamique.

Ces études fournissent des preuves en faveur de l'utilité de l'apprentissage automatique aux entraîneurs et préparateurs physique, notamment pour comprendre l'influence des différentes charges de travail sur les joueurs [48, 45, 56]. Ces approches permettent aussi une meilleure planification des entraînements et matchs grâce à la prédiction des charges internes en amont des sessions [47, 44]. Un second constat est l'efficacité des modèles de type arbre (forêt aléatoire, XGBoost, arbre de décision) dans la prédiction du RPE. Ces modèles sont d'autant plus intéressants qu'ils permettent une interprétation des résultats plus intuitive à travers des règles de décision. Nous avons détaillé ici quelques études qui traitent de la relation charge externe  $\leftarrow$  charge interne, ainsi que de la relation charge antérieure  $\rightarrow$  charge interne. D'autres études ont été menées ces 5 dernières années et les résultats de ces travaux similaires convergent vers des conclusions semblables [46, 57, 78, 79, 80].

Dans le cadre de cette thèse, nous avons voulu proposer une approche pour la prédiction de la charge interne (RPE) adaptée à la planification des entraînements : nous proposons de mettre en entrée des modèles d'apprentissage automatique d'une part, les plannings d'entraînement qui nous permettent d'intégrer l'historique des charges externes que subit un athlète durant sa préparation et d'autre part, un historique des charges internes (e.g, le RPE à des dates antérieures) (voir section 3.2). Nous faisons donc le choix d'une prédiction du RPE basé sur un modèle dynamique *i.e.* un modèle d'apprentissage automatique qui prend en entrée les données passées relatives aux charges externe (plannings d'entraînement) et interne (RPE).

### 1.3.2 Charge de travail et valeurs manquantes

Le suivi de la charge de travail est un élément clé pour différents objectifs du staff de clubs sportifs, pour la prévention des risques de blessures en particulier [43], mais aussi pour la planification des entraînements, la gestion de la fatigue des athlètes [81, 82, 83] ou encore l'optimisation des performances de leurs athlètes [14]. Les entraîneurs et préparateurs physiques utilisent couramment des indicateurs subjectifs comme le RPE ou mixtes comme le sRPE. Les indicateurs subjectifs étant soumis à la bonne volonté et à la discipline des athlètes, supposés remplir les questionnaires en fin d'entraînement, présentent des historiques affectés par de nombreuses valeurs manquantes. Les valeurs manquantes nuisent au suivi de la charge de travail, elles affectent la fiabilité de ce suivi, notamment le calcul des charges aiguës, chroniques et des rapports charges aiguës/chroniques. Par ailleurs, dans le cadre des modèles dynamiques que nous souhaitons établir, ces valeurs manquantes rendent impossible l'utilisation d'historiques des données pour la prédiction [32, 84, 60, 85]. Il nous faut donc proposer un pré-traitement pour disposer d'historiques des charges (internes en particulier) qui soient complets et utilisables pour nos modèles de prédiction.

La solution à cette problématique est l'imputation des données manquantes, qui consiste à faire une estimation des valeurs manquantes. La plus élémentaire des techniques d'imputation est l'imputation par la moyenne de groupe : il s'agit de remplacer les valeurs manquantes par la moyenne des charges de travail des autres athlètes présents à la même session. Les techniques reposant sur des modèles d'apprentissage automatique proposent des alternatives plus fiables en permettant de prédire les charges de travail manquantes en fonction de variables contextuelles relatives à la session et à l'athlète. Quelques études ont abordé cette question en sport collectif [86, 87, 88].

Les travaux de [86] avaient pour but d'examiner l'impact des données de charges de travail manquantes dans les sports collectifs et de proposer une méthode pratique et efficace d'imputation des valeurs manquantes. Les charges de travail (sRPE) de 10 footballeurs professionnels ont été collectées au cours d'une saison de 32 semaines. Les données ont été bruitées aléatoirement dans une fourchette de 5 à 50% avec un pas d'augmentation de 5% et les données ont été imputées à l'aide de 12 méthodes :

- Imputation par la moyenne de groupe à la même session
- Imputation par la médiane de groupe à la même session
- Imputation par la moyenne de l'athlète aux sessions précédentes
- Imputation par la médiane de l'athlète aux sessions précédentes
- Imputation par la moyenne mobile de l'athlète calculée sur les 4 dernières semaines

- Imputation par la médiane mobile de l'athlète calculée sur les 4 dernières semaines
- Imputation par la moyenne de l'athlète calculée sur les sessions précédentes tombées le même jour de la semaine
- Imputation par la médiane de l'athlète calculée sur les sessions précédentes tombées le même jour de la semaine
- Imputation par la moyenne entre la moyenne de l'athlète aux sessions précédentes et la moyenne de groupe à la session manquante
- Imputation par la moyenne entre la médiane de l'athlète aux sessions précédentes et la médiane de groupe à la session manquante
- Imputation par la moyenne entre la moyenne mobile de l'athlète calculée sur les 4 dernières semaines et la moyenne de groupe à la session manquante
- Imputation par la moyenne entre la médiane mobile de l'athlète calculée sur les 4 dernières semaines et la médiane de groupe à la session manquante.

Les performances ont été mesurées à l'aide de la racine de l'erreur quadratique normalisée et de la moyenne des erreurs absolues. La méthode la mieux adaptée à tous les niveaux de valeurs manquantes était la moyenne de groupe, qui reste la méthode la plus courante et la plus simple. Les auteurs de [88] ont critiqué l'imputation par la moyenne de groupe en postulant que le RPE reflète le ressenti d'un athlète particulier par rapport à la difficulté d'un effort produit. Dans ce cas, l'imputation de cette valeur par la moyenne de groupe peut en effet sembler inappropriée, en sachant que chaque athlète a une réponse propre à un effort donné. Plusieurs méthodes ont été évaluées pour imputer le RPE sur 987 jeux de données collectés à partir des compétitions féminines de rugby. L'imputation par la moyenne de groupe, les algorithmes d'apprentissage automatique k plus proches voisins, forêt aléatoire, machine à vecteurs de support, réseaux de neurones artificiels, régression linéaire, etc. ont été évalués à différents niveaux de données manquantes. Les modèles d'apprentissage automatique ont tous obtenu des résultats comparables en termes de qualité d'imputation et la forêt aléatoire était la meilleure méthode d'imputation. L'imputation par la moyenne de groupe était en revanche la méthode la moins performante. Cela dit, la qualité des imputations était faible pour toutes les méthodes, ce qui démontrait selon les auteurs que l'imputation du RPE nécessite l'emploi d'approches plus sophistiquées. Les auteurs de [88] ont déconseillé l'imputation par la moyenne de groupe et ont recommandé aux praticiens d'examiner attentivement toute méthode d'imputation relative aux charges de travail manquantes avant de les exploiter. Les auteurs de [87] ont aussi fait une tentative sur la question

des valeurs manquantes pour le RPE. Les objectifs étaient d'identifier les meilleures méthodes d'imputation simple et d'examiner les tendances de la charge de travail à l'aide de l'imputation multiple, *i.e.* les écarts entre les valeurs réelles et les valeurs imputées avec plusieurs méthodes. 93 joueurs de basket-ball ont participé à l'étude, les charges externes et le RPE ont été collectés au cours d'une saison sportive. Les valeurs manquantes ont été générées artificiellement et imputées à l'aide de méthode d'imputation simple, *e.g.* moyenne de groupe et méthode d'imputation par apprentissage automatique. L'imputation par la moyenne de groupe a été surpassée en termes de qualité d'imputation par une méthode d'imputation par apprentissage qui utilisait des informations sur le contexte de la session et de l'athlète.

Nous constatons que les résultats des études précédemment rapportées sont assez hétérogènes, voire contradictoires. La moyenne de groupe qui considère tous les athlètes au même niveau (*i.e.* même état de forme, même capacités physiques etc) nous semble à proscrire sachant que chaque athlète a une réponse (charge interne) propre aux stimuli de sa session (charge externe). Les charges de travail les plus fréquemment utilisées dans la littérature sont le RPE (charge interne) et le sRPE (indicateur mixte). Cependant, il est à noter que l'imputation à l'aide de méthodes d'apprentissage automatique semble fournir des résultats prometteurs, ce qui doit conduire à nous interroger sur les indicateurs contextuels de la session et de l'athlète les plus pertinents à mettre en entrée des modèles. Dans nos travaux de thèse, nous avons expérimenté ces méthodes (moyenne de groupe, apprentissage automatique) ainsi qu'une nouvelle méthode qui présuppose l'effet mémoire du RPE à travers une approche qui tient compte de la qualité d'imputation des valeurs manquantes et qui cherche à modéliser l'incertitude associée à celles-ci (voir 3).

### 1.3.3 Charge de travail et performance en match

La relation entre charge de travail et performance est au coeur de nombreux travaux [84, 89, 90, 23] qui ont pour ambition de construire des modèles mathématiques qui reflètent cette relation. Parmi les premiers modèles proposés dans la littérature, on retrouve le modèle fitness-fatigue de Banister [60]. Ce modèle a pour hypothèse principale qu'une session d'activité physique engendre à la fois un gain (fitness ou état de forme) et une dégradation (fatigue) chez l'athlète, les niveaux de fatigue et forme caractérisent l'état de celui-ci. Le modèle a été pensé pour un type de performance, la performance athlétique :

$$p_t = p^* + k_1 \sum_{i=1}^{t-1} w_i e^{-\frac{1}{\tau_1}(t-i)} - k_2 \sum_{i=1}^{t-1} w_i e^{-\frac{1}{\tau_2}(t-i)} \quad (1.3)$$

Le modèle (1.3) modélise la performance à un instant  $t$  comme la différence normalisée entre la fatigue et la forme (ou fitness) en réponse à des charges d'entraînement.

Les constantes  $\tau_1$  et  $\tau_2$  sont des constantes de temps, significativement différentes, ce qui souligne que fatigue et fitness obéissent à des dynamiques bien distinctes en accord avec l'interprétation physiologique de Banister,  $k_1$  et  $k_2$  sont des constantes de gain avec une sémantique de dimensionnement ou de commensurabilité. Ces dernières décennies, plusieurs variantes du modèle ont été proposées [91, 89, 92, 93, 94, 90, 95]. De nombreuses études applicatives [96] et des revues très exhaustives ont encore été récemment publiées [97]. Les études ultérieures au travail de Banister se sont trop souvent concentrées sur les équations de récurrence du modèle Fitness-Fatigue seul, en introduisant des retards purs ou une interaction entre la fatigue et la condition physique pour mieux approcher le comportement observé. Le modèle Fitness-fatigue aurait pu répondre à certaines questions pratiques fondamentales tel que i) comment définir le meilleur entraînement pour réaliser une performance sans épuiser l'athlète, ii) comment réaliser une performance en un temps limité iii) comment évaluer de manière fiable la fatigue et la condition physique de l'athlète v) comment prévenir le risque de blessure ou tout comportement anormal. Néanmoins, les chercheurs semblent s'être uniquement concentrés sur la validation et se sont enfermés dans le carcan des équations de récurrence du modèle physiologique proposé initialement par Banister [84]. Peu de ces études ont finalement pris le recul nécessaire pour replacer ce modèle de 1975 dans le formalisme de la représentation d'état. C'est sans doute une tendance regrettable car en abandonnant une partie de l'interprétation physiologique du modèle initial, des modèles plus riches auraient pu être construits, certes au détriment de l'interprétation physiologique, mais plus précis et néanmoins interprétables mathématiquement [12]. Une évolution possible du modèle est de coupler le modèle fitness-fatigue à un modèle d'apprentissage automatique. Cela a été proposé dans [98], les auteurs ont aussi souligné que l'utilisation d'un indicateur de charge de travail unique en entrée du modèle fitness-fatigue est une restriction forte et proposent de s'orienter vers des modèles multivariés, exploitant plusieurs indicateurs de charge de travail en entrées. Cela permettrait de réduire le surapprentissage, de mieux apprécier la charge de travail grâce aux différents indicateurs, et de préserver l'interprétation physiologique des fonctions de transfert fitness-fatigue. Une autre critique qui peut être faite au modèle fitness-fatigue est qu'il aborde d'une part seulement la performance athlétique, et d'autre part, les applications du modèle ont été faites dans des contextes assez spécifiques où l'athlète est évalué en laboratoire ou sur terrain, sur une activité physique bien normée. Cela rend le modèle Fitness-Fatigue peu applicable aux sports collectifs ou tout autre sport où il existe une interaction entre les athlètes qui peut altérer leurs performances, ou encore d'autres facteurs contextuels qui affecteraient les sorties du modèle.

Aujourd'hui, les chercheurs des sciences du sport préfèrent les modèles dérivés de l'apprentissage automatique pour mieux coller aux observations des performances sportives [7]. Cependant, ces modèles étaient jusqu'ici difficiles à interpréter et ne

se prêtaient donc pas facilement à la définition d'un entraînement adaptatif sensible au contexte ou à l'état de forme de l'athlète. Certains modèles comme les arbres de décision se donnent pour objet de rendre interprétables leurs décisions. De plus, les modèles d'apprentissage automatique nécessitent de grandes quantités de données avant de devenir des outils de simulation efficaces, condition sine qua non qui jusqu'à il y a peu n'était pas satisfaite. Nous avons orienté les travaux de thèse vers la performance individuelle dans les sports collectifs, qui est plus complexe que la performance en sport individuel, cela étant dû aux nombreux facteurs d'interaction qui influencent la performance individuelle en sport collectif. Dans les sports collectifs, les entraîneurs et les préparateurs physiques effectuent généralement des tests sur le terrain pendant la saison afin d'évaluer l'influence de l'entraînement sur les capacités physiques des athlètes en vue des prochains matchs. Il existe des études telles que [23], dans lesquelles les auteurs ont étudié l'influence de la charge de travail (durée et RPE) sur les performances athlétiques évaluées à l'aide de test comme le *Interval Shuttle Run Test* [99] (ISRT) pour les joueurs de football de haut niveau. Les auteurs de [23] ont montré, à l'aide de modèles à effet aléatoire, que la durée passée en entraînement et en match avait une influence positive sur les performances de l'ISRT. Les modèles à effet aléatoire se distinguent des autres modèles comme la régression linéaire par le fait que les paramètres des modèles varient car ils prennent en compte les groupes *e.g.* l'athlète, la position de jeu, et les paramètres sont donc différents d'un groupe à l'autre. Peu d'études se concentrent sur la prédiction des performances en cours de match. Par exemple, les études [20, 21] et [25] se sont concentrées sur la performance technique en match. Les auteurs de [21] ont effectué une analyse de corrélation sur un ensemble de données couvrant 7 saisons consécutives (2007 - 2014) d'une équipe de basket-ball professionnelle et affirment qu'il n'y a pas de relation significative entre la charge de travail (durée) et les performances en match (nombre d'actions de jeu positives réalisées). D'autre part, les auteurs de [20] ont proposé un indicateur de charge de travail mixte qui combine des indicateurs subjectifs (bien-être perçu) et objectifs (mesures GPS). Leur analyse à l'aide de modèles quadratiques mixtes a montré une relation significative avec la performance technique en match, déterminée par un indicateur propriétaire basé sur les actions en match. Les modèles quadratiques mixtes sont des modèles qui impliquent la résolution d'une équation quadratique qui identifie la relation l'entrée et la sortie du modèle. Plusieurs types d'indicateurs de performance sont décrits dans la littérature des sciences du sport : les indicateurs athlétiques (vitesse de course moyenne et maximale, nombre et intensité des accélérations), les indicateurs techniques (nombre de buts, de passes décisives) ou les indicateurs de performance de l'équipe (classement, résultat du match), etc. Dans [25], les auteurs ont étudié la performance de l'équipe en termes d'actions en match dans le rugby d'élite. Grâce à une analyse de modèle linéaire mixte, les auteurs montrent que les performances les plus pertinentes en match

sont les actions défensives en raison de leur grande influence sur la performance collective. La performance collective a été identifiée à partir des résultats des matchs (victoire vs défaite), auxquels a été ajouté un autre type de classement (Britannic Ranking) qui prend en compte l'influence de l'emplacement des matchs sur les résultats (extérieur vs à domicile). D'autres résultats intéressants sont les relations observées entre la charge de travail et la performance individuelle démontrée par une analyse de modèles à seuils, où les auteurs ont souligné l'effet positif du nombre d'impacts corporels chroniques sur la capacité à gagner des duels 1 contre 1 pour les joueurs. Cette étude s'est concentrée sur les performances techniques et n'a pas abordé la relation entre la charge de travail et les performances athlétiques. La performance athlétique a été étudiée dans [22], où les auteurs ont évalué l'influence de la charge de travail hebdomadaire passée (distance totale parcourue, distance parcourue en marchant/courant à différentes intensités, nombre d'accélération/décélération à différentes intensités) sur les performances athlétiques futures (mêmes indicateurs que les charges de travail). Les principaux résultats de [22] sont i) des corrélations linéaires négatives de la plupart des indicateurs de charge de travail passés (sur une semaine) avec les performances athlétiques du prochain match, ii) l'impact positif du nombre de sessions d'entraînement sur les résultats du jeu mis en évidence par une analyse de régression logistique (la régression logistique est un algorithme d'apprentissage automatique qui permet de mettre en relation des entrées et une variable de sortie binaire). Les auteurs ont montré que les chances d'obtenir des résultats positifs étaient plus grandes pour les semaines comportant moins de sessions d'entraînement. Néanmoins, il reste difficile d'identifier les critères qui déterminent la performance individuelle ou la charge de travail, en particulier dans un contexte où les joueurs sont interdépendants.

Ces travaux nous ont motivés à revoir la relation entre charge de travail et performances individuelles en match tout en abordant les deux grands types de performances individuelles, athlétique et technique (actions réalisées en match). Cet objectif a justifié l'utilisation de modèles d'apprentissage automatique dits multisorties qui permettent de tenir compte des contraintes entre les différents indicateurs de performance interdépendants (voir 4). Nous avons considéré pour la charge de travail une multitude d'indicateurs de charges externes à différents niveaux d'intensité. Ce travail avait pour but d'une part d'examiner la relation entre charges externes antérieures et performances individuelles en match, mais aussi la relation de cette dernière avec des facteurs contextuels comme le lieu du match, la position de jeu ou encore le type de compétition. Une autre motivation a été qu'à ce jour, peu d'études ont examiné l'influence d'autres facteurs sur la performance technique (basée sur les actions en match) ou sur la performance athlétique dans les sports collectifs. [21, 22, 20, 24, 25, 26, 27, 28].

## 1.4 Conclusion

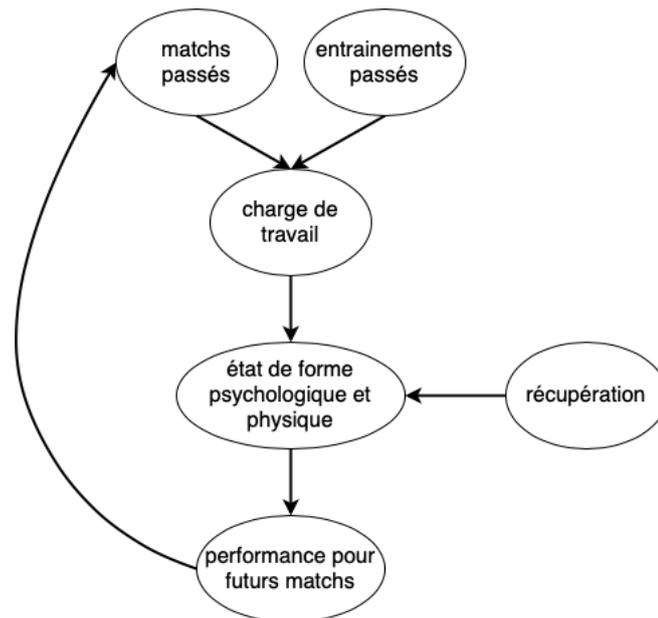
Nous avons constaté que dans les activités physiques et sportives, la charge de travail est un élément central qui impacte les performances et les risques de blessure. Nous avons aussi observé que dans les sports dits collectifs, il y a peu d'études qui tentent d'examiner la relation entre la charge de travail et la performance individuelle en match. Les connaissances actuelles suggèrent qu'il existe une relation entre la charge de travail (externe et interne) et la performance individuelle athlétique en match [14, 22]. Les auteurs ayant abordé la question de la relation avec la performance individuelle technique n'observent pas de relations significatives avec la charge de travail. Pour autant, certaines études comme celles de [25] et [20] montrent des résultats qui encouragent la poursuite des travaux sur ce type de performance.

Nous avons aussi souligné qu'il existe plusieurs indicateurs de performance qu'elle soit athlétique ou technique. En ce qui concerne la charge de travail, cette dernière regroupe deux types de charges liées par une relation causale [11]. La charge externe produite à l'exercice induit une réponse psycho-physiologique chez l'individu. La réponse dite charge interne de l'individu est celle qui va moduler l'adaptation à l'activité. Sachant que la charge interne est plus complexe à mesurer et interpréter, des questionnaires subjectifs et échelles auto-rapportées sont utilisés comme le RPE qui mesure une difficulté ressentie, considérée estimer la charge interne [31]. Cependant, les informations subjectives ne sont pas systématiquement communiquées par les athlètes et elles sont donc affectées par un nombre de valeurs manquantes souvent significatif. Les entraîneurs et préparateurs physiques ont besoin des données relatives aux deux types de charge de travail, interne et externe : en particulier, la charge interne qui caractérise l'état des individus et leur réponse à l'effort doit être suivie de façon continue sans être affectée par le problème des valeurs manquantes.

Cela fait de l'imputation de données manquantes pour la charge de travail un enjeu essentiel pour le suivi longitudinal des individus, en sachant que la mesure de la charge de travail comme le sRPE qui est le produit de la durée de l'activité par le RPE, peut être affecté par un grand nombre de valeurs manquantes [67].

Nous avons déduit de la littérature que la prédiction de la charge interne devrait être un outil puissant pour l'optimisation et la personnalisation des entraînements sportifs. Si on illustre nos propos par une situation donnée, *e.g.* un athlète dont on voudrait estimer la performance pour le match à venir, nous avons besoin de l'historique des charges internes antérieures de l'athlète car ce sont elles qui vont nous renseigner sur les états de formes physique et psychologique de l'athlète, et qui par la suite auront un impact direct sur ses performances et ses risques de blessure. La figure 1.4 schématise la situation que nous venons d'exposer dans un contexte sportif (sport collectif).

Au vu de ces constats, la multiplicité des indicateurs de performance évalués lors



**Figure 1.4. Schéma représentatif de la relation charge de travail → performances dans un contexte de sport collectif**

de matches, en particulier dans les sports collectifs, la question des données manquantes pour un suivi fiable des athlètes et une vision dynamique des relations entre charges et performances constitueront le coeur de nos propositions. Si ces travaux de thèse ont pour cadre applicatif principal celui des sports collectifs, les notions de RPE et de performance se retrouvent dans des applications cliniques en phase de rééducation, comme dans le contexte de la rééducation des patients post-AVC en particulier. Aussi, nous proposerons un transfert des approches et résultats des travaux issus du milieu sportif au monde de la rééducation. Ce transfert s’explique également par les deux grands domaines d’application de notre laboratoire EuroMov Digital Health in Motion qui sont la santé et le sport.

Nous traiterons donc de :

- la prédiction de la charge interne en fonction des plannings d’entraînement,
- l’imputation de données manquantes pour les charges internes,
- la modélisation permettant la mise en relation de charges de travail avec des indicateurs de performances individuelles en sports collectifs,
- la prédiction de la charge interne de patients post-AVC en fonction de leur activité en session de rééducation,
- la prédiction de performances de patients post-AVC.

Le chapitre 2 présentera les outils mathématiques utilisés pour adresser ces problématiques.

La prédiction de charge interne sera la première de nos propositions qui sera présentée dans le chapitre 3. Ce travail propose une nouvelle approche pour l'imputation des données manquantes. La méthode proposée utilise le formalisme d'une des *théories de l'incertain*, le formalisme des *fonctions de croyance*. Nous utilisons le formalisme des fonctions de croyance pour modéliser l'incertitude associée aux valeurs imputées, et nous proposons plusieurs modèles d'incertitude qui seront comparés grâce à l'utilisation de *l'algorithme de prédiction des k plus proches voisins crédibiliste*, qui utilise le même formalisme et qui permet de tenir compte de l'incertitude associée aux données imputées. La charge interne considérée est le RPE et celui-ci sera discrétisé en 3 classes de proportions comparables pour être ramené à un problème de classification. L'originalité de ce travail est aussi dans l'utilisation des plannings d'entraînement en entrée des modèle, cela est un premier pas qui a vocation à proposer un outil de prédiction de la charge interne et de personnalisation des entraînements pour le staff sportif.

Dans un second temps, au chapitre 3, nous présenterons un travail mené sur des données COVID 19 pour évaluer la modélisation incertaine pour les valeurs manquantes sur un jeu de données complet. Nous avons choisi un jeu de données COVID 19 pour valider notre approche de l'imputation en raison de l'impossibilité de collecter des données sportives sur terrain durant la pandémie COVID 19. Lors de ce travail, nous proposons une extension de *l'algorithme des k plus proches voisins crédibiliste* aux problèmes de régression, ainsi qu'une méthode d'imputation basée sur ce dernier et des améliorations pour les modèles d'incertitudes.

Le chapitre 4 sera consacré à la prédiction de performances individuelles en match de Handball féminin en fonction des charges de travail (externes) antérieures. Nous avons fait le constat que la performance est multidimensionnelle, d'autant plus dans les sports collectifs. Nous proposons alors d'utiliser des algorithmes de régression dits *multisorties* qui permettent de tirer avantage des relations qui existent entre les différentes sorties à prédire, *e.g.* les différents indicateurs de performances techniques et athlétiques individuelles. Nous comparons les performances prédictives des algorithmes de régression classiques dits *monosortie* avec celles des algorithmes multisorties.

Dans le dernier chapitre 5, nous proposons deux applications cliniques qui reprennent les méthodologies utilisées lors des précédents chapitres. La première application vise à prédire le RPE de patients post-AVC lors de sessions de jeu en réalité virtuelle au cours de leur programme de rééducation. La prédiction est faite en fonction de performances réalisées dans les jeux auxquels ils ont participé lors de la session. La seconde application consiste à prédire les performances aux jeux en amont de la session en fonction des états ressentis de fatigue et de douleur du patient.

Et pour finir, une conclusion générale sur les travaux menés et les résultats obtenus

est proposée.

## Chapitre

### 2

## **Théories et formalisme**

Dans le chapitre précédent nous avons présenté le contexte applicatif général de la thèse et les problématiques qui lui sont associées. Nous avons introduit le lien entre les performances d'un athlète en match en sports collectifs, la charge de travail antérieure et les caractéristiques propres à l'athlète et l'activité. Nous avons montré que le suivi des athlètes en termes de charge de travail pouvait viser plusieurs objectifs. Nous avons souligné l'importance de l'imputation des valeurs manquantes de charge de travail pour le suivi de l'athlète, l'optimisation des performances et la prévention des risques de blessure.

Le chapitre précédent a montré qu'il est très courant d'utiliser des questionnaires subjectifs tels que le RPE pour l'estimation de la charge interne. Le plus souvent, c'est une mesure mixte, le sRPE [67] correspondant au produit du RPE et de la durée de l'activité (en minutes), qui est utilisé. Tous ces constats renforcent l'importance de la considération du RPE comme mesure de la charge de travail.

Il est relativement fréquent que les athlètes ne renseignent pas les questionnaires RPE (pour cause d'oubli, de forte fatigue, ...), des données manquantes de RPE sont inévitablement en grand nombre dans la plupart des jeux de données de clubs sportifs, ce qui nuit au suivi de la charge d'entraînement qui vise principalement 2 objectifs: garder un niveau de charge adéquat au maintien et à l'optimisation de la performance tout en évitant la sur-charge pour prévenir les blessures. Ces objectifs aux enjeux importants motivent le développement de méthodes d'imputation des données de charges de travail manquantes.

Si le RPE est une note subjective, donc par nature déjà imprécise et par suite incertaine, les valeurs de RPE imputées le sont encore davantage. La problématique abordée couvre alors deux questions distinctes :

- Quelle méthode d'imputation est la plus appropriée au contexte de prédiction en sport collectif ?
- Comment pouvons-nous tenir compte des incertitudes associées aux valeurs de charge imputées lors de leur utilisation à des fins de prédiction ?

Le chapitre 4 de ce manuscrit concerne la performance individuelle en sport collectif. Beaucoup de critères peuvent caractériser la performance individuelle d'un athlète, ces critères dépendent du sport concerné, du poste de l'athlète ainsi que d'autres facteurs. La performance individuelle de l'athlète dépend évidemment aussi fortement de l'état de forme physique et psychologique de l'athlète. L'ambition de ce travail de thèse est d'étudier en profondeur la relation (*charge de travail*)  $\rightarrow$  (*performance individuelle*) en sport collectif.

Nous avons voulu répondre aux 2 problématiques énoncées ci-dessus grâce à l'usage de méthodes qui font appel à plusieurs domaines scientifiques interconnectés:

- l'apprentissage automatique
- la théorie des fonctions de croyances
- l'imputation des données manquantes

L'apprentissage automatique ou *machine learning* en anglais, fait partie des domaines au coeur de l'intelligence artificielle. L'apprentissage automatique supervisé consiste en l'apprentissage de modèles de classification ou de régression à partir de données d'apprentissage labellisées. Une fois appris ou entraînés, ces modèles peuvent être utilisés pour prédire les labels d'exemples non-labellisés.

L'apprentissage est dit automatique car automatisé à l'aide d'une approche algorithmique d'optimisation. L'algorithme d'un modèle représente une suite d'instructions à réaliser pour la phase d'apprentissage. Certains algorithmes existants ont une ou plusieurs fonctions objectives à minimiser ou maximiser.

L'apprentissage automatique est devenu de plus en plus populaire ces 15 dernières années, notamment du fait de l'augmentation des capacités de stockage et de la puissance de calcul en informatique. Durant les 4 dernières décennies, l'apprentissage automatique a largement été utilisé dans des domaines nombreux et variés tels que le diagnostic médical, le traitement automatique du langage naturel, la classification de document, la prédiction de consommation électrique, la classification d'image et bien d'autres.

L'apprentissage automatique englobe 2 grands types d'approche :

- l'apprentissage supervisé qui consiste à entraîner un modèle sur des données labellisées ou étiquetées, ce qui signifie que chaque exemple ou observation de l'ensemble d'apprentissage a une réponse associée.
- l'apprentissage non-supervisé qui consiste à extraire des informations de données non-étiquetées de manière, par exemple, à regrouper les observations en clusters.

Dans ces travaux de thèse, nous nous intéressons dans un premier temps (sous-section 3.2) à un problème de classification supervisée, les observations labellisées

sont des plannings d'entraînement auxquels sont associés des labels correspondant aux niveaux de charge de travail (charge interne : RPE discrétisé). Cette tâche de classification aura pour particularité d'être réalisée en présence de labels *incertains* dans les données d'apprentissage.

L'incertitude associée aux données manquantes imputées dans les applications du chapitre 3 sur la charge de travail puis sur les contaminations au COVID-19 nous a orientés vers l'utilisation d'une théorie moderne de l'incertain : la théorie des fonctions de croyance [100, 101]. Son formalisme permet de modéliser de manière fine les différentes incertitudes associées aux données. Le choix de ce formalisme a été appuyé par l'existence de plusieurs algorithmes de classification supervisée étendus aux fonctions de croyances (apprentissage incertain) dans la littérature scientifique.

Après avoir proposé une première approche d'*imputation crédibiliste* appliquée à la prédiction de charge interne en Section 3.2, nous l'avons validée sur un jeu de données COVID-19 en Section 3.3, ce qui nous a permis de proposer une extension aux problèmes de régression de l'algorithme des  $K$  plus proches voisins basé sur la théorie des fonctions de croyance: l'*EKNN*.

Dans le Chapitre 4, nous avons étudié la relation charge de travail  $\rightarrow$  performance individuelle en sport collectif sous une perspective prédictive et multi-critères: la régression multisorties. Ce type de modèle met en relation plusieurs espaces multidimensionnels en entrée et en sortie et permet des gains de performance prédictive et de temps de calcul en tirant profit des corrélations entre labels (sorties des modèles).

## 2.1 Apprentissage supervisé

Nous allons présenter dans cette section les fondamentaux de l'apprentissage automatique, nous débuterons par le formalisme associé aux données puis nous présenterons la classification, la régression, les mesures d'erreurs associées et quelques algorithmes usuels.

### 2.1.1 Données

Soit un jeu de données  $D = (x, y)$  avec

- $x = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \dots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$  la matrice d'attributs de taille  $n \times p$  où  $\forall i \in \{1, \dots, n\}$ ,  
 $x_i = (x_i^1, \dots, x_i^p) \in \Omega^X = \Omega^{X^1} \times \dots \times \Omega^{X^j}$  représente le vecteur d'attributs de la  $i^{\text{ième}}$  observation avec  $\forall j \in \{1, \dots, p\}, x_i^j \in \Omega^{X^j}$  (l'espace  $\Omega^{X^j}$  peut être numérique ou catégoriel).

- $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$  l'ensemble des labels (classes) des observations où  $\forall i \in \{1, \dots, n\}$ ,  $y_i \in \Omega^Y$  est le label de la  $i^{\text{ième}}$  observation (donc associé à  $x_i$ ).

Les différentes observations  $(x_i, y_i)$  du jeu données  $D$  sont des réalisations de variables aléatoires  $X$  et  $Y$  pouvant modéliser divers phénomènes. Les espaces  $\Omega^Y$  des labels  $y_i$  peuvent être de différents types selon les problématiques abordées. Pour les problèmes de classification à  $Q$  classes on a  $\Omega^Y = \{c_1, \dots, c_Q\}$ , pour les problèmes de régression, les labels  $y$  appartiennent à un espace numérique  $\Omega^Y \subset \mathbb{R}$ , *e.g.* le RPE est renseigné à l'aide d'une échelle allant de 0 à 10, on a donc  $\Omega^Y = \{0, \dots, 10\}$ .

Comme mentionné en introduction de ce chapitre, le but de l'apprentissage supervisé est l'élaboration de modèles de classification ou de régression à partir de données d'apprentissage. Ce type d'apprentissage est dit supervisé car le modèle est appris sur des données labellisées, donc selon une forme de direction. Après la phase d'entraînement, un modèle d'apprentissage automatique peut être utilisé pour prédire les labels d'observations pour lesquels on ne dispose que des vecteurs d'attributs.

Dans cette section nous introduisons les bases de l'apprentissage supervisé, nous abordons les principes de la classification et de la régression avec pour chacune de ces tâches, les métriques d'évaluation considérées dans ce manuscrit. Nous définissons ensuite le protocole d'évaluation chronologique utilisé dans l'ensemble de ces travaux de thèse, et nous finissons par passer en revue les algorithmes d'apprentissage supervisé utilisés dans la suite du manuscrit.

## 2.1.2 Classification

L'apprentissage supervisé d'un modèle de classification dit *classifieur* a pour but d'apprendre à associer  $X$  à  $Y$ , de manière à pouvoir prédire le label  $y_* \in \Omega^Y = \{c_1, \dots, c_Q\}$  d'une observation à partir de son seul vecteur d'attributs  $x_* = (x_*^1, \dots, x_*^p)$ . Par exemple, un algorithme de classification supervisé permet de déterminer si un athlète va se blesser lors d'une activité physique ( $\Omega^Y = \{oui, non\}$ ) en fonction de ses  $p$  charges de travail passées ( $x = \{charge^1, \dots, charge^p\}$ ). La phase d'apprentissage permet d'estimer ou d'*apprendre* les paramètres du modèle.

### 2.1.2.1 Evaluation d'un classifieur

Une fois appris, un classifieur est évalué sur de nouvelles données, les *données de test*, où les labels prédits par le modèle sont comparés aux labels réels. Durant la phase de test, de nouvelles données  $(x_{*i}, y_{*i})_{i \in \{1, \dots, T\}}$  sont utilisées. Un classifieur entraîné permettra de prédire les labels de ces données de test, pour ensuite comparer les prédictions obtenues  $(\hat{y}_{*1}, \dots, \hat{y}_{*T})$  avec les vrais labels  $(y_{*1}, \dots, y_{*T})$ . Cette comparaison se fait

		réalité	
		+	-
prédiction	+	vp	fp
	-	fn	vn

**Tableau 2.1. Matrice de confusion**

généralement à partir d'une *matrice de confusion* de taille  $q \times q$  comptant le nombre de bonnes/mauvaises prédictions par classe.

**Exemple:** nous voulons apprendre un classifieur qui permet de prédire si oui ou non un athlète va se blesser au prochain match. Nous considérons la blessure comme la classe positive (+), et l'absence de blessure comme la classe négative (-), alors la matrice de confusion est définie par la Table 2.1 où vp, fp, fn et vn sont respectivement les nombres de vrais positifs (bien classifiés), faux positifs (mal classifiés), faux négatifs (mal classifiés) et vrais négatifs (bien classifiés). A partir de la matrice de confusion, plusieurs mesures de performances de classification peuvent être calculées.

Le *taux d'erreur* qu'on note  $\epsilon$  mesure la proportion des  $T$  exemples de test mal classifiés:

$$\epsilon = \frac{fp + fn}{T} \quad (2.1)$$

La *justesse* ou accuracy en Anglais, est notée  $acc$  et représente la proportion d'exemples de test bien classifiés:

$$acc = \frac{vp + vn}{T} = 1 - \epsilon \quad (2.2)$$

Outre ces deux mesures, d'autres mesures comme le *rappel* (noté  $rapp$ ) et la *précision* (notée  $prec$ ) permettent une évaluation plus fine de la qualité de classification en tenant compte du sens des erreurs de prédiction commises (les erreurs sont elles majoritairement constituées de blessures prédites à tort ou de blessures non-prédites ?) :

$$rapp = \frac{vp}{vp + fp} \quad (2.3)$$

$$prec = \frac{vp}{vp + fn} \quad (2.4)$$

Dans un contexte multiclassés (avec  $q > 2$ ), on a

$$acc = \frac{\sum_{i=1}^T \mathbb{1}_{\{\hat{y}_{*i} = y_{*i}\}}}{T} = 1 - \epsilon \quad (2.5)$$

### 2.1.3 Régression

La régression consiste à établir le lien entre une variable numérique cible, le label  $Y \in \Omega^Y$  et d'autres variables, les attributs  $X = (X^1, \dots, X^p) \in \Omega^{X^1} \times \dots \times \Omega^{X^j}$ .

L'apprentissage supervisé permet d'entraîner des modèles prédictifs dits *régresseurs*, capables de prédire  $Y$  à partir de  $X$  dans le cas où  $Y$  est numérique, *i.e.*  $\Omega^Y \subseteq \mathbb{R}$ . L'apprentissage de régresseurs notés  $f : \Omega^{X^1} \times \dots \times \Omega^{X^j} \rightarrow \Omega^Y$  est réalisé à partir d'observations  $(x_i, y_i)_{i=1, \dots, n}$  (les données d'apprentissage) permettant d'estimer les paramètres du modèle. On note  $\widehat{y}_* = f(x_*)$  la prédiction du régresseur  $f$  calculée à partir d'un vecteur d'attribut  $x_*$ .

### 2.1.3.1 Evaluation d'un régresseur

Considérons à nouveau un jeu de données de test  $(x_{*i}, y_{*i})_{i \in \{1, \dots, T\}}$ . Alors, les labels numériques de  $Y$  prennent leurs valeurs dans  $\Omega^y$ . Soit  $f : \Omega^{X^1} \times \dots \times \Omega^{X^j} \rightarrow \Omega^Y$  un régresseur appris sur les  $n$  exemples d'apprentissage, nous voulons tester ses performances prédictives sur les  $T$  nouveaux exemples  $(x_{*i}, y_{*i})$ . Plusieurs mesures permettent de réaliser cette tâche, généralement à partir de l'écart entre les labels prédits  $\widehat{y}_{*i}$  et les vrais labels  $y_{*i}$ .

*La racine carré de l'erreur quadratique moyenne*, ou Root Mean Squared Error en anglais notée *RMSE*, évalue la justesse des prédictions  $\widehat{y}_{*i}$  faite avec le modèle  $f$ . Plus cette valeur est proche de 0, plus le modèle est performant.

$$RMSE = \sqrt{\frac{\sum_{i=1}^T (\widehat{y}_{*i} - y_{*i})^2}{T}} \quad (2.6)$$

*La moyenne des erreurs absolues*, ou Mean Absolute Error en anglais notée *MAE*, évalue l'écart absolu moyen entre les valeurs prédites et les labels numériques réels :

$$MAE = \frac{\sum_{i=1}^T |\widehat{y}_{*i} - y_{*i}|}{T} \quad (2.7)$$

Pour le *MAE* tout comme pour le *RMSE*, plus la valeur est proche de 0, meilleures sont les prédictions du modèle. Il est à noter que le *RMSE* a tendance à amplifier le poids des grandes erreurs par rapport aux petites erreurs contrairement au *MAE* qui est linéaire par rapport à l'amplitude des erreurs de prédiction observées au sein du jeu de test.

### 2.1.4 Protocoles d'évaluation

D'un point de vue statistique, pour qu'un modèle d'apprentissage supervisé soit capable de généraliser ce qu'il apprend des données d'apprentissage à l'ensemble de la population étudiée, les données d'apprentissage doivent être le plus représentatives possible du phénomène étudié. De cette manière, les biais d'échantillonnage sont réduits et le modèle peut limiter son *sur-apprentissage* [102]. En effet des modèles entraînés sur

des données trop spécifiques auront tendance à trop "coller à ces données" et perdront ainsi en capacité de généralisation, ce qui peut se traduire par de grandes différences entre les performances prédictives calculées d'une part sur les données d'apprentissage et d'autre part sur les données de test.

Les performances prédictives des modèles d'apprentissage automatique sont donc généralement calculées sur des données de test différentes des données d'apprentissage. Cette séparation d'un jeu de données en apprentissage et en test constitue la base de la méthode de *validation croisée* qui consiste à évaluer un modèle sur d'autres données que celles ayant servi à son apprentissage. Les données de test doivent être issues d'un même modèle génératif que les données d'apprentissage, les 2 sous-ensembles de données doivent ainsi avoir des distributions en labels suffisamment proches. De manière à estimer les performances prédictives le plus justement possible, *la validation croisée à  $k$  couches* (k-fold cross validation) est souvent utilisée. Elle consiste à séparer un jeu de données en  $k$  sous-ensembles ayant des distributions en labels proches, puis en  $k$  apprentissages successifs du modèle considéré sur  $k - 1$  couches et en l'évaluation sur la  $k^{ième}$  couche (non-utilisée pour l'apprentissage), en faisant varier la couche de test de manière à ce que chaque couche soit utilisée 1 fois en données de test. L'estimateur de la performance est alors la moyenne des  $k$  performances calculées itérativement. De cette manière chaque donnée a été utilisée à la fois en apprentissage et en test (pour couvrir tous les cas possibles) et l'estimateur de la performance prédictive est alors statistiquement optimal [103, 104, 105].

**Remarque:** La validation croisée convient parfaitement au contexte de données statiques (*i.e.* lorsque les exemples de  $D$  ne sont pas ordonnés) mais pose problème dans un environnement dynamique. En effet, de manière à ne pas utiliser d'information *future* pour la prédiction de phénomènes *passés*, un pipeline d'évaluation chronologique est généralement défini comme suit : pour chaque date  $t$  d'évaluation considérée, le modèle est entraîné sur l'ensemble des données antérieures à  $t$  et évalué sur des données ultérieures à  $t$ .

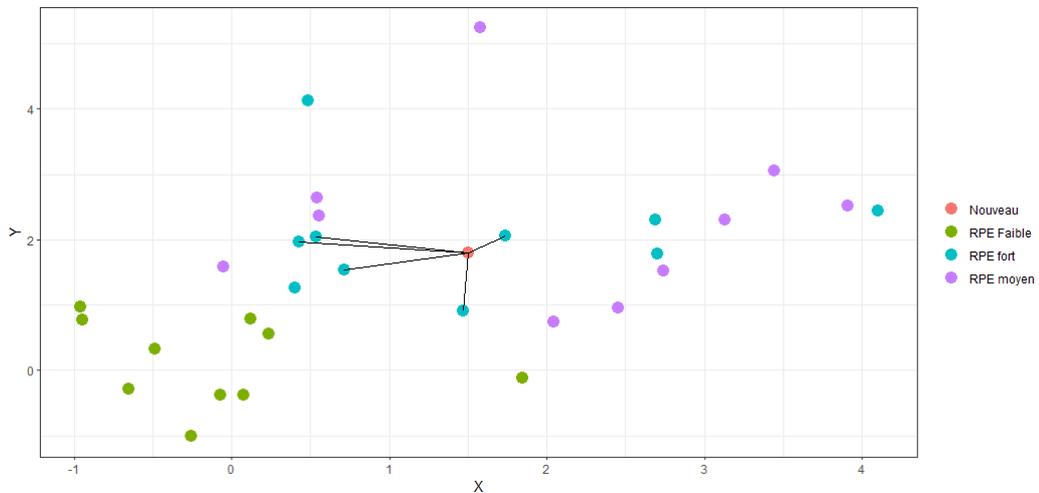
### 2.1.5 Algorithmes usuels

De très nombreux modèles d'apprentissage supervisé ont été proposés dans la littérature depuis les années 1950 [106], nous présentons ici certains de ces modèles bien connus de la littérature et utilisés dans la suite de ce manuscrit. Les algorithmes présentés dans cette section sont utilisables en classification et en régression, bien évidemment avec quelques adaptations méthodologiques entre les deux variantes, notamment au niveau de l'apprentissage des modèles et ou au stade prédictif.

### 2.1.5.1 K plus proches voisins

Les K plus proches voisins ou  $k$  nearest neighbours en anglais, noté  $KNN$  [107, 108], est un algorithme d'apprentissage supervisé non paramétrique basé sur une distance permettant de déterminer les exemples les plus proches d'un exemple à classifier. L'entraînement d'un  $KNN$  consiste uniquement en le calcul d'une matrice de (di-)similarité  $n \times n$  répertoriant toutes les distances entre les vecteurs d'attributs des observations. Pour déterminer le label  $y_*$  d'un nouvel exemple à partir de son vecteur d'attributs  $x_*$ , l'algorithme  $KNN$  cherche dans un premier temps les  $K$  plus proches voisins de  $x_*$  dans  $D$  et prédit le label majoritaire du voisinage de  $x_*$  pour un problème de classification ou le label moyen de ce voisinage pour un problème de régression. L'hyperparamètre  $K$  est généralement fixé au préalable (éventuellement par calibration). Toutes les distances peuvent être utilisées pour calculer le voisinage de  $x_*$ , néanmoins la distance euclidienne reste l'usage le plus fréquent :

$$d(x_*, x_i) = \sqrt{\sum_{j=1}^p (x_*^j - x_i^j)^2} \quad (2.8)$$

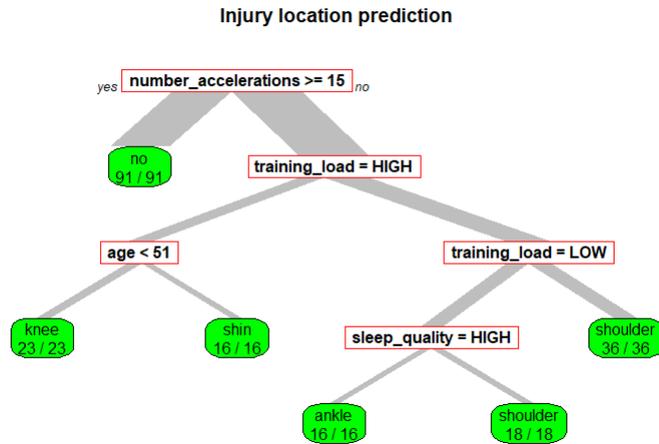


**Figure 2.1. Algorithme des  $k$  plus proches voisins : Nous voulons déterminer la charge de travail d'une activité, la nouvelle observation doit être assignée à un label  $y_i \in \Omega^Y = \{RPE\_faible, RPE\_moyen, RPE\_fort\}$  Le point rouge est la nouvelle observation à classifier, les points verts appartiennent à la classe RPE faible, les points violets à la classe RPE moyen et les points bleus à classe RPE fort. Les  $k=5$  plus proches voisins de la nouvelle observation sont liés à elle par un segment sur la figure. Le vote majoritaire assignera donc la classe RPE fort à la nouvelle observation**

### 2.1.5.2 Arbre de décision

L'arbre de décision ou decision tree en anglais est un algorithme d'apprentissage supervisé permettant de construire des modèles de classification (arbres de classification) et de régression (arbres de régression) pouvant être représentés par des graphes orientés, ce qui les rend particulièrement interprétables. Chaque chemin de l'arbre part du noeud initial, qui contient l'ensemble des données d'apprentissage, et mène à une feuille (noeud terminal de l'arbre de décision) qui est associée à une prédiction sur  $Y$ . La littérature scientifique propose de nombreuses méthodes d'arbres de décision [109, 110, 111], l'algorithme CART (Classification and Regression Tree [112]) est celui qui est utilisé pour tous les arbres de décision de ce manuscrit, cet algorithme a la particularité de construire des arbres binaires, *i.e.* chaque noeud interne de l'arbre donne naissance à 2 noeuds enfants.

L'apprentissage d'un arbre de décision est réalisé selon un critère d'impureté (classification) ou de variabilité (régression) en termes de labels. Pour chaque noeud, l'algorithme effectue une recherche exhaustive des attributs (et seuils) qui permettent d'obtenir les feuilles qui minimisent le critère d'impureté ou de variabilité des labels.



**Figure 2.2. Algorithme Arbre de décision CART :** Nous voulons déterminer si un athlète va se blesser et si oui ou non il va se blesser durant une activité. Les critères sont l'âge, le nombre d'accélération, la charge de travail, si l'athlète a déjà été blessé dans le passé et la qualité du sommeil  $x_i = (age, past\_injuries, sleep\_quality_i, training\_load_i, number\_accelerations_i)$ , les vecteurs d'attributs sont associés à des labels  $y_i \in \Omega^Y = \{no, shoulder, knee, ankle, shin\}$

Dans les problèmes de classification, la mesure d'impureté  $imp$  à minimiser la plus courante est l'indice de Gini appliqué à la distribution de probabilités  $(p_1, \dots, p_q)$  des labels au sein d'un noeud (voir Eq 2.9). Elle permet de calculer, pour chaque potentielle coupure  $(F \rightarrow F_L, F_R)$  d'une feuille  $F$  en 2 noeuds enfants  $F_L$  et  $F_R$ , des gains de

pureté (*i.e.* baisse d'impureté) en comparant les niveaux de pureté du noeud parent, des noeuds enfants en tenant compte du *poids*  $\pi_L$  et  $\pi_R$  de chaque feuille (qui correspondent aux proportions d'exemples terminant dans chaque feuille enfant) (voir Eqs 2.9 et 2.10). Au stade prédictif, pour une nouvelle observation  $x_*$ , son label prédit  $\hat{y}_*$  correspondra à la classe majoritaire de la feuille associée à  $x_*$ .

$$imp(F) = Gini(p_1, \dots, p_q) = \sum_{k=1}^q p_k(1 - p_k) \quad (2.9)$$

$$gain(F \rightarrow F_L, F_R) = imp(F) - \pi_L imp(F_L) - \pi_R imp(F_R) \quad (2.10)$$

En régression, le critère à minimiser est la variabilité des labels au sein des feuilles. Cette variabilité est considérée en termes de variance ou d'erreur quadratique notée  $SSE$  et calculée pour chaque noeud coupure (voir Eq. 2.11), ce qui permet le calcul de *baisse* de variabilité. Pour une nouvelle observation  $x_*$  le label prédit correspondra à la moyenne des labels de la feuille  $F$  dans laquelle elle est rangée, on a  $\hat{y}_*(F) = \bar{y}(F)$ .

$$SSE(F) = \sum_{i \in F_L} (y_i - \bar{y}_L)^2 + \sum_{i \in F_R} (y_i - \bar{y}_R)^2 \quad (2.11)$$

$$baisse(F \rightarrow F_L, F_R) = SSE(F) - \pi_L SSE(F_L) - \pi_R SSE(F_R) \quad (2.12)$$

Les valeurs  $\bar{y}_L$  et  $\bar{y}_R$  sont respectivement les moyennes des labels  $y_i$  dans les noeuds enfants  $F_L$  et  $F_R$ .

### 2.1.5.3 Forêt aléatoire

La forêt aléatoire est un algorithme d'apprentissage automatique basé sur l'utilisation d'un ensemble d'arbres de décision [113]. Chaque arbre de décision de la forêt aléatoire est entraîné sur un ré-échantillonnage aléatoire (avec remise) de taille  $n$  du jeu de données d'apprentissage. Lors de la recherche de coupures "optimales", seul un sous-ensemble aléatoire des attributs est considéré. Le nombre d'attributs considérés pour l'évaluation d'une coupure potentielle est égal à  $\sqrt{j}$  en classification et  $\frac{j}{3}$  en régression où  $j$  est le nombre total d'attributs. Au stade prédictif, les prédictions des arbres d'une forêt sont agrégées par un vote (entre les différentes classes) en classification ou une moyenne (des labels prédits) en régression. La forêt aléatoire permet ainsi de réduire les risque de sur-apprentissage et d'améliorer les performances prédictives des arbres de décision. Cet algorithme peut être utilisé en classification et en régression.

### 2.1.6 Régression multisortie

Afin de réaliser des tâches de prédiction impliquant plusieurs variables de sorties (labels) numériques, nous nous intéressons à une approche spécifique de l'apprentissage

supervisé : la *régression multisortie*. Ce type de tâche vise à prédire plusieurs variables cibles à partir d'un ensemble d'attributs.

La régression multisortie présente deux bénéfices principaux:

- améliorer les performances prédictives (en termes de *RMSE*, *MAE*, etc.)
- réduire les temps de calcul

L'idée sous-jacente à la régression multisortie est de tirer profit des corrélations potentielles entre les variables de sortie. Le fait d'entraîner un unique modèle capable de prédire plusieurs labels à la fois, plutôt que d'entraîner un modèle (monosortie) par label permet naturellement de diminuer les temps de calcul par un facteur correspondant au nombre de labels considérés. Par exemple, nous voudrions pouvoir prédire le nombre de tirs cadrés et le nombre de buts réalisés par un joueur de football au match à venir en fonction de ses charges de travail passées et de ses moyennes de tirs cadrés et de buts aux matchs précédents.

Il existe plusieurs types d'approches de régression multisortie, en voici les principales : la *Transformation de Problèmes* (TP) qui consiste à transformer un problème multisortie impliquant  $d$  labels en  $d$  problèmes de régression classiques, *i.e.* monosortie ou *Single output*) *e.g.* pour l'exemple ci-dessus un modèle pour prédire le nombre de tirs cadrés et un modèle pour prédire le nombre de buts. Il existe des méthodes TP qui ont pour ambition de tirer profit des éventuelles corrélations entre les différentes sorties *e.g.* le nombre de buts est très positivement corrélé au nombre de tirs cadrés. La deuxième famille d'approches appelée *Adaptation d'Algorithme* (AA) consiste à adapter les algorithmes classiques d'apprentissage supervisé aux problèmes multisortie.

#### 2.1.6.1 Transformation de Problème (TP)

Il existe plusieurs approches de TP telles que *Regression chain* (RC) [114] ou *Multi Target Regression Stacking* [115, 116] (MTRS). [117]. La méthode RC consiste à mettre en cascade des modèles monosorties dans un ordre choisi, où à chaque niveau de la "cascade", les prédictions des modèles précédents sont les entrées des modèles suivants *e.g.* faire un premier modèle pour prédire le nombre de tirs cadrés puis un second modèle auquel on ajoute le nombre de tirs cadrés en entrée pour prédire le nombre de buts. La méthode MTRS consiste à apprendre  $d$  modèles monosorties pour chaque sortie, et  $d$  méta-modèles monosorties qui corrigent les premiers en tenant compte des prédictions des autres modèles *e.g.* faire deux premiers modèles (tirs cadrés, buts) puis deux autres modèles (tirs cadrés, buts) qui prennent en entrée les prédictions des deux premiers modèles (tirs cadrés prédits, buts prédits).

### 2.1.6.2 Adaptation d'Algorithme (AA)

Les algorithmes d'apprentissage multisortie de la catégorie AA sont généralement des extensions d'algorithmes connus comme les arbres de décision où la fonction de coupure est adaptée pour le cas multisortie, ou des algorithmes nativement multisortie comme les  $k$  plus proches voisins où l'algorithme calcule une seule fois les  $k$  voisins au lieu de  $d$  fois. Les extensions multisortie sont généralement réalisées au niveau du choix de la fonction objectif à minimiser dans les approches d'apprentissage supervisé.

En régression, avec  $d$  variables de sortie, la fonction objectif *gain* (voir Eq. 2.10), impliquant le critère de variabilité *SSE* calculé au niveau des noeuds et feuilles  $F$  (voir Eq. 2.11) est étendue au contexte multisortie en considérant la somme des variabilités de chaque variable de sortie dans  $F$  :

$$SSE(F) = \sum_{i \in F_L} (y_i - \bar{y}_L)^T (y_i - \bar{y}_L) + \sum_{i \in F_R} (y_i - \bar{y}_R)^T (y_i - \bar{y}_R) \quad (2.13)$$

où  $\cdot^T$  représente l'opérateur *transposé* et avec  $\forall i \in \{1, \dots, n\} : y_i \in \Omega^Y \subseteq \mathbb{R}^d$ .

On peut remarquer que le critère de coupure (et donc d'apprentissage) inclut ainsi la variance des labels numériques mais aussi la covariance entre eux. La forêt aléatoire multisortie reprend simplement l'algorithme forêt aléatoire en appliquant le critère de coupure multidimensionnel défini par l'Eq. 2.13 durant l'apprentissage de chacun des arbres de la forêt. Il existe plusieurs méthodes d'évaluation des modèles de régression multisortie dont la moyenne des racines des erreurs quadratiques moyennes (2.15) qui est une extension multisortie de la racine de l'erreur quadratique moyenne (2.14) :

$$RMSE(l) = \frac{1}{n} \sum_{i=1}^n (y_i^l - \hat{y}_i^l)^2 \quad (2.14)$$

$$aRMSE = \frac{1}{L} \sum_{l=1, \dots, L} RMSE(l) \quad (2.15)$$

où les données d'apprentissage contiennent  $N$  exemples labellisés chacun avec  $d$  labels,  $y_i^l$  est la valeur réelle du  $l^{\text{ème}}$  label du  $i^{\text{ème}}$  exemple et  $\hat{y}_i^l$  sa valeur prédite. Les algorithmes d'apprentissage automatique qui ont été présentés sont des algorithmes qui se basent sur le cadre des probabilité classique, cela impliquant que ces algorithmes ne proposent pas de gestion de l'incertitude au sein des modèles. Ceci peut être problématique notamment quand il y a présence de valeurs manquantes dans les données d'apprentissage. Ceci nous a amenés à choisir un cadre nous permettant la gestion de l'incertitude relative aux valeurs imputés. Nous avons choisi le cadre de la théorie des fonctions des croyances [100, 101] car il propose des algorithmes d'apprentissage supervisé reconnus et qui permettent une gestion plus fine de l'incertitude relative aux

données. Nous avons été d'autant plus motivés par notre contexte applicatif, le sport, les valeurs manquantes sont très présentes.

## 2.2 Théorie des fonctions de croyances

Dans cette section nous justifions l'utilisation de théories modernes de l'incertain et présentons le formalisme des fonctions de croyances.

### 2.2.1 Motivations

En général les différentes incertitudes des données sont soit ignorées, soit prises en compte par des modèles ensemblistes ou probabilistes. Il existe néanmoins d'autres formalismes plus souples permettant une modélisation fine et spécifique à la nature de l'incertitude en question :

- théorie des sous ensembles flous [118]
- théorie des possibilités [119]
- théorie des fonctions de croyance [100] [101]
- théorie des probabilités imprécises [120]

Dans notre contexte, nous choisissons le formalisme des fonctions de croyances pour sa souplesse et sa flexibilité. De plus, il existe des algorithmes d'apprentissage automatique qui reprennent ce formalisme et qui ont montré leur efficacité dans la littérature [121, 122, 123].

### 2.2.2 Formalisme

Dans cette section, nous exposons les bases de la théorie des fonctions de croyance, également connue sous le nom de théorie de Dempster-Shafer ou théorie de l'évidence [101, 100].

#### 2.2.2.1 Définitions

Soit  $\Omega = \{\omega_1, \omega_2, \dots, \omega_H\}$  le cadre de discernement ou *powerset*, *i.e.* l'espace des possibles. Une fonction de masse  $m : 2^\Omega \rightarrow [0, 1]$  représente le degré de connaissance relatif aux sous-ensembles de  $\Omega$ . Les éléments  $A \subseteq \Omega$  tels que  $m(A) > 0$  sont appelés éléments focaux et leur somme vaut 1 :

$$\sum_{A \subseteq \Omega} m(A) = 1 \quad (2.16)$$

La quantité  $m(\Omega)$  représente le degré d'ignorance. A partir de la fonction de masse  $m$ , différentes mesures d'incertitude peuvent être calculées telles que les fonctions de *croyance* et de *plausibilité* définies dans les équations (2.17) et (2.18) qui peuvent être respectivement interprétées comme des bornes inférieure et supérieure de probabilités :

$$Bel(A) = \sum_{B \subseteq A} m(B) \quad (2.17)$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad (2.18)$$

Différentes fonctions de masse peuvent représenter différentes sources d'information. Au niveau décisionnel, il peut être nécessaire de les combiner en une seule fonction de masse exprimant la connaissance que nous pouvons extraire de l'ensemble des sources.

### 2.2.2.2 Combinaison de masses

Il existe plusieurs méthodes de fusion d'information via des règles de combinaison de masses de croyance [124, 125]. L'une des plus connues est la règle de combinaison conjonctive de Dempster  $\oplus$  [100]:

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B)m_2(C) \quad (2.19)$$

où  $\kappa = \sum_{B \cap C = \emptyset} m_1(B)m_2(C)$  représente le degré de *conflict* entre les sources  $m_1$  et  $m_2$ .

L'idée principale de cette règle est de considérer toutes les sources comme fiables. Après la combinaison, nous obtenons une nouvelle fonction de masse qui peut être utilisée au niveau de la prise de décision.

Dans le cas où une source peut être peu fiable, une règle de combinaison dite règle disjonctive a été proposée :

$$(m_1 \cup m_2)(A) = \sum_{B \cup C = A} m_1(B)m_2(C) \quad (2.20)$$

Cette règle est moins utilisée car elle élargit les éléments focaux et peut rendre la prise de décision difficile [126, 127], d'autres règles de combinaison existent et s'adaptent à différents contextes [128, 129].

### 2.2.2.3 Prise de décision

Dans les cas où le degré d'ignorance  $m(\Omega)$  est trop important, certains auteurs recommandent de rejeter la prise de décision [121]. Sinon, le choix de la mesure d'incertitude à considérer pour la prise de décision reste un dilemme [130].

Par exemple, selon l'objectif de l'application et la stratégie choisie en termes de conservatisme [131, 132, 133], toute mesure d'incertitude comprise entre la *croyance* (voir équation (2.17)) et la *plausibilité* (voir équation (2.18)) peut être utilisée. Cependant, ces mesures ne sont pas additives, *i.e.*  $\forall A, B \in \Omega$  telles que  $A \cap B = \emptyset$ ,  $Bel(A \cup B) \neq Bel(A) + Bel(B)$  (idem pour la fonction  $Pl$ ). Pour cette raison, de nombreux outils en science des données sont incompatibles avec ces mesures d'incertitude *soft* puisque la plupart d'entre eux ont été développés dans le cadre classique de la théorie des probabilités.

Afin de se ramener à un formalisme probabiliste certains auteurs [123] ont proposé des alternatives permettant d'estimer les distributions de probabilité en maximisant leur vraisemblance à partir de données incertaines. Pour y parvenir, la fonction de vraisemblance étendue aux fonctions de croyances [121] peut être maximisée, éventuellement à l'aide de l'algorithme *Evidential Expectation Maximization* ( $E^2M$ ) correspondant à une extension crédibiliste de l'algorithme EM [131] pour les cas où la vraisemblance crédibiliste ne serait pas analytiquement maximisable. Il est à noter que la nature itérative de l'algorithme  $E^2M$  peut introduire des niveaux de complexité importants.

Pour des raisons pragmatiques, de nombreux chercheurs choisissent ainsi de projeter le contenu informationnel des fonctions de masse dans le cadre des probabilités [121, 123]. Le *modèle des croyances transférables* (*transferable belief model*) a été proposé par Smets [134, 130] où la transformation *pignistique* permet de convertir les fonctions de masse en distributions de probabilité pour lesquelles les degrés d'ignorance sont projetés dans des distributions uniformes. Cette uniformité est une des faiblesses les plus critiquées de la transformation pignistique, car celle-ci ne fait aucune distinction envers les différents éléments de l'espace  $\Omega$ . Malgré les critiques de la communauté crédibiliste envers le choix de probabilité uniforme pour modéliser l'ignorance [135], la transformation pignistique définie dans l'équation (2.21) reste une solution simple et naturelle pour calculer des distributions de probabilité à partir de fonctions de masse :

$$BetP(\omega) = \frac{1}{1 - m(\emptyset)} \sum_{A \ni \omega} \frac{m(A)}{|A|} : \forall \omega \in \Omega \quad (2.21)$$

Dans le contexte de l'apprentissage automatique, de nombreux classifieurs permettent le calcul de prédictions incertaines exprimées dans des espaces plus complexes que les probabilités classiques [121, 136]. La littérature propose ainsi plusieurs algorithmes qui utilisent le formalisme des fonctions de croyances. Dans la section suivante, nous en présentons un parmi les plus connus les *K plus proches voisin crédibiliste* [121, 122, 123].

### 2.2.3 K plus proches voisins crédibiliste

Dans cette section nous introduisons un algorithme d'apprentissage automatique incertain largement utilisé dans les travaux présentés dans ce manuscrit (voir chapitres 3 et 5): l'algorithme des  $K$  plus proches voisins crédibiliste qui est une extension aux fonctions de croyances de l'algorithme bien connu en apprentissage automatique l'algorithme des  $K$  plus proches voisins. Cet algorithme noté EKNN (Evidential  $K$  Nearest Neighbours) étend l'algorithme des  $K$  plus proches voisins au formalisme des fonctions de croyance [121] et est basé sur la règle conjonctive de combinaison de Dempster. Dans cette méthodologie, les labels  $(y_i)_{i=1,\dots,n}$  peuvent être incertains et représentés par des masses de croyance.

Soit  $x_s$  une nouvelle observation à classer. La première étape consiste à calculer les distances entre  $x_s$  et tous les vecteurs d'attributs  $x_i$  des exemples d'apprentissage afin de déterminer l'ensemble des  $K$  voisins "les plus proches" de  $x_s$ . Dans l'algorithme EKNN, chaque voisin est considéré comme une source d'information. Pour chaque voisin  $x_k$  tel que  $y_k = C_q$ , une fonction de masse  $m_{s,k}$  est calculée:

$$\begin{cases} m_{s,k}(\{C_q\}) &= \alpha_0 \times \exp(-\gamma_q^2 \times d_{s,k}^\lambda) \\ m_{s,k}(\Omega^Y) &= 1 - m_{s,k}(\{C_q\}) \end{cases} \quad (2.22)$$

La quantité  $m_{s,k}(\{C_q\})$  représente la masse attribuée au label  $C_q$  par le voisin  $x_k$  de  $x_s$ . Les hyperparamètres  $\gamma_q$  permettent de pénaliser certaines classes de façon spécifique. Les hyperparamètres  $\alpha_0, \gamma_q$  peuvent être calibrés à l'aide d'une procédure d'optimisation classique telle que la descente de gradient. Le paramètre  $\gamma_q > 0$  est lié à la classe/élément focal  $C_q$ , dans [121] l'auteur recommande de fixer le paramètre  $\alpha_0$  (qui empêche les fonctions de masse dogmatiques i.e. fonctions de masse avec masse sur  $m(\Omega) = 0$ ) à 0.95,  $d_{s,k}$  représente la distance euclidienne entre  $x_s$  et son voisin  $x_k$  et  $\lambda \in \{1, 2, 3, \dots\}$  est un paramètre qui pénalise les voisins les plus éloignés.

Une fois toutes les masses  $m_{s,k}$  calculées, elles sont combinées à l'aide de la règle de combinaison de Dempster en une fonction de masse finale  $m_s = m_{s,1} \oplus \dots \oplus m_{s,K}$  où  $m_{s,1}, \dots, m_{s,K}$  représentent les connaissances associées aux  $K$  plus proches voisins de  $x_s$ . L'idée sous-jacente à cette combinaison est de considérer tous les voisins  $x_k$  de  $x_s$  comme fiables. Enfin, une décision peut être prise en fonction de  $m_s$ , l'approche choisie dans [121] est de prédire la classe correspondant au maximum de croyance (2.17).

Dans le cas où les labels sont incertains ou imparfaits et modélisés par des fonctions de masse, les données d'apprentissage sont sous la forme  $(x_i, m_{y_i})_{i=1,\dots,n}$ , pour chaque voisin  $x_k$  nous avons une fonction de masse  $m_{y_k}$  relative au label  $y_k$ . Dans [121], l'auteur propose d'actualiser les fonctions de masse de tous les voisins avec le niveau d'incertitude de leurs labels. Dans l'équation suivante, un facteur correspondant au

niveau d'incertitude des labels  $m_{y_k}$  est ainsi ajouté:

$$\forall A \subseteq \Omega^Y, \begin{cases} m_{s,k}(A) & = \alpha_0 \times \exp(-\gamma_A^2 \times d_{s,k}^\lambda) \times m_{y_k}(A) \\ m_{s,k}(\Omega^Y) & = 1 - \sum m_{s,k}(A) \end{cases} \quad (2.23)$$

Les éléments focaux  $A$  peuvent contenir plusieurs classes  $C_q$ . Plus il y a d'éléments focaux, plus il y a de paramètres  $\gamma_A$ : un trop grand nombre d'éléments focaux peut être très coûteux. La possibilité de tenir compte de labels incertains a motivé nos travaux sur l'imputation des valeurs manquantes. Les jeux de données qui sont utilisés dans les applications réelles souffrent très souvent de la présence de valeurs manquantes, alors l'imputation de celles-ci représente un enjeu important surtout dans le contexte de l'apprentissage automatique où les algorithmes classiques ne peuvent pas apprendre sur des données incomplètes. Dans ce qui suit nous introduisons les bases de l'imputation des valeurs manquantes ainsi que des travaux qui traitent de l'incertitude à la procédure d'imputation.

### 2.3 Imputation de valeurs manquantes

Malgré un contexte d'augmentation du nombre d'applications en apprentissage automatique et d'une disponibilité grandissante des données, les jeux de données collectés sont souvent incomplets. Cela peut s'expliquer notamment par différentes causes incluant les défauts de mesure, des capteurs défectueux ou partiellement fiables, des questionnaires partiellement remplis, etc. Le traitement des valeurs manquantes des applications d'apprentissage automatique est devenu une question importante qui a donné lieu à un large éventail de recherches [137, 138, 139]. De nombreuses méthodes de gestion des valeurs manquantes ont été proposées : moyenne, omission, apprentissage, etc. [137, 139, 140], L'une des méthodes les plus populaires consiste simplement à exclure les exemples incomplets du jeu de données en raison de l'incapacité de la plupart des modèles prédictifs à traiter les valeurs manquantes [141, 142, 140]. Ce type de traitement n'est cependant pas souhaitable lorsque les quantités de données disponibles sont limitées ou que la structure des données est chronologique.

La méthode d'imputation choisie dépend également du modèle génératif des valeurs manquantes, qui sont souvent classées en *Missing Completely At Random* (MCAR) pour les valeurs manquantes qui sont indépendantes des valeurs observées ou non observées *e.g.* un athlète n'a pas renseigné son RPE ni à cause de la valeur d'une variable observée ni à cause d'un ressenti de la difficulté trop élevé (RPE non observé), *Missing At Random* (MAR) pour les valeurs manquantes qui ne dépendent que des valeurs observées *e.g.* le RPE d'un athlète n'est pas renseigné car il a été blessé durant sa séance, ou *Not Missing At Random* (NMAR) pour les valeurs manquantes qui dépendent des valeurs non observées *e.g.* un athlète n'a pas renseigné son RPE car la difficulté ressentie était

trop forte [137, 139]. Ces catégories indiquent la raison pour laquelle les données sont manquantes, cette information doit, dans l'idéal, être prise en compte dans le choix de la méthode d'imputation [143]. Dans ce manuscrit, nous aborderons la problématique des labels manquants dans les problèmes de classification et de régression.

Considérons un vecteur de labels  $y = (y_1, \dots, y_n)$ ,  $N = \{1, \dots, n\}$  avec certains de ses éléments  $(y_i)_{i \in I}$  qui sont manquants. Soit  $I$  l'ensemble des valeurs manquantes. L'objectif des méthodes d'imputation est d'estimer les valeurs  $(\hat{y}_i)_{i \in I}$  les plus proches possible des valeurs réelles  $(y_i)_{i \in I}$ .

Il existe deux familles de méthodes d'imputation : les méthodes d'imputation simple [144] visant à imputer les valeurs manquantes par des valeurs uniques, et les méthodes d'imputation multiple [138] où, pour chaque donnée manquante, un ensemble de valeurs peut être considéré. Parmi les méthodes d'imputation simples les plus connues, l'*imputation par moyenne*:

$$\forall i \in I, \hat{y}_i = \bar{y} = \frac{1}{|N \setminus I|} \sum_{i \in N \setminus I} y_i \quad (2.24)$$

La valeur manquante est imputée par la moyenne des valeurs connues de  $y$ . Cette méthode est définie pour les variables numériques, pour les variables catégorielles, l'*imputation par maximum de fréquence* consiste à imputer les valeurs manquantes par les valeurs les plus observées. L'imputation par moyenne et par maximum de fréquence représentent de loin les méthodes d'imputation les plus utilisées en pratique dans le cas où les valeurs manquantes sont effectivement imputées (hormis le cas de l'imputation par omission donc) [145].

Des méthodes d'*imputation par apprentissage* ont été proposées en tirant profit de modèles prédictifs d'apprentissage automatique [146]. Ces modèles d'apprentissage automatique utilisent les exemples  $(x_i, y_i)$  pour lesquels le label  $y_i$  n'est pas manquant. L'imputation par apprentissage est particulièrement adaptée pour les valeurs manquantes de type MAR car ce type d'absence suppose que la valeur est manquante en raison d'autres variables observées.

L'*imputation multiple* a pour but de traiter le problème des valeurs manquantes tout en minimisant les biais dus aux méthodes d'imputation simple [138]. Par exemple, un biais possible de l'imputation par moyenne peut être de fausser les corrélations de la variable imputée avec les autres variables dans le cas d'une analyse multivariée. L'imputation multiple permet de restreindre ce biais en générant plusieurs jeux de données et combinant les résultats des analyses comme l'analyse de corrélation. Il existe plusieurs méthodes pour faire de l'imputation multiple, mais toutes suivent les mêmes étapes :

1. avec une méthode d'imputation simple, on impute plusieurs fois les valeurs manquantes en considérant différentes valeurs associées à différentes probabilités, ce qui crée  $m$  jeux de données complets associés chacun d'eux à une probabilité.

2. les  $m$  jeux de données complets sont analysés individuellement
3. les résultats des analyses sont agrégés, par exemple par une moyenne pondérée

A la différence de l'imputation multiple, l'imputation simple ne prend pas en compte l'incertitude associée à l'imputation. Une fois imputées, les données sont traitées comme si elles étaient certaines. Cela peut mener à des résultats d'analyse très biaisés et ainsi grandement affecter les conclusions qui peuvent en être déduites [147].

L'imputation multiple a donc pour avantage de tenir compte de l'incertitude associées aux imputations par le biais des distributions de probabilité des variables partiellement observée. Cependant, les analyses individuelles des jeux de données générés sont réalisées indépendamment les unes des autres, sans tenir compte du niveau (potentiellement faible) de la vraisemblance des valeurs imputées. D'autres approches utilisent des théories de l'incertain plus générales que les probabilités classiques comme la théorie des fonctions de croyances qui permettent, entre autre, de tenir compte de la nature des incertitudes considérées.

### 2.3.1 Imputation incertaine

Au-delà des méthodes classiques d'imputation des valeurs manquantes [142, 141], certaines méthodes d'imputation par soft computing ont été proposées [148, 149, 150]. Dans [151], une méthode est proposée pour catégoriser les données manquantes et supprimer le bruit qu'elles génèrent à l'aide d'une approche basée sur les noyaux afin d'améliorer l'imputation des données manquantes qui sont assignées à des classes (classification non supervisée) relatives à leur degrés de fiabilité et représentées par des fonctions de croyances basées sur les distances entre les observations et les centres des classes. L'objectif de la méthode est de concevoir une stratégie d'imputation offrant une résistance à l'incertitude. Toutefois, la méthode n'est pas destinée aux données manquantes dans le cadre d'algorithmes d'apprentissage automatique.

Dans [152], les auteurs proposent une méthode pour minimiser les erreurs de classification dues à l'incertitude associée aux valeurs manquantes. Plusieurs estimations précises pour les valeurs manquantes sont effectuées et les prédictions correspondantes sont finalement combinées conjonctivement dans des fonctions de croyance prédictives, l'approche proposée peut être assimilée à une méthode d'imputation multiple crédibiliste.

Dans le contexte de la recherche d'information, Jousselme *et al.* ont proposé une représentation de l'incertitude associée aux valeurs manquantes [153]. Les données manquantes sont modélisées par des fonctions de croyance définies sur l'espace de chaque variable. La méthode montre de bonnes performances pour les tâches de recherche d'information. Cette méthode reste cependant plus adaptée aux données catégorielles qui ont des espaces ou cadres de discernement finis et généralement de petites tailles.

Aucune de ces méthodes ne permet de prendre en compte l'incertitude associée à l'imputation au niveau prédictif dans le cadre de l'apprentissage automatique. L'approche pour la modélisation des valeurs manquantes présentée dans le chapitre 3 s'inscrit dans le formalisme des fonctions de croyances, nous proposons plusieurs modèles d'incertitude associés à différentes méthodes d'imputation. Dans la seconde partie du chapitre 3 nous proposons une extension de l'E $K$ NN aux problèmes de régression ainsi qu'une méthode d'imputation crédibiliste, *i.e.* dans le formalisme des fonctions de croyance, associée à l'algorithme E $K$ NN régression qui se base sur les distances temporelles entre les valeurs manquantes et leurs plus proches voisins connus. Deux applications en apprentissage automatique sont présentées dans ce chapitre, une en classification dans un contexte sportif, la seconde en régression dans le contexte de la pandémie COVID-19 .

## Chapitre

### 3

## Modélisation incertaine des valeurs manquantes

### 3.1 Introduction

Dans ce chapitre nous présentons nos travaux sur l'imputation de données manquantes dans deux applications. La première concerne la prédiction de charge interne (RPE) à partir de plannings d'entraînement et des données d'historique de charge interne. Au cours de ce travail, nous avons proposé une approche pour la gestion des données manquantes et des pas de temps irréguliers (des entraînements et donc des remplissages de RPE).

En raison de la crise sanitaire mondiale démarrée à l'automne 2019, les travaux suivants ont été menés sur des données relatives à la pandémie COVID-19. L'objectif de ces travaux était de tester l'efficacité de l'approche proposée pour la gestion des valeurs manquantes et l'incertitude associée à ces dernières via la modélisation incertaine sur un jeu de données complet et précis que nous avons bruité artificiellement.

Dans la suite de ce chapitre, nous présenterons les détails des travaux menés pour la prédiction de charge interne sur un jeu de données de football où la charge interne est mesurée par le RPE et ceux des travaux pour la prédiction des décès quotidiens à partir des décès et contaminations passés sur un jeu de données COVID-19. Nous y exposerons les détails et résultats des deux applications de la modélisation incertaine pour la gestion des valeurs manquantes ainsi qu'une adaptation de l'E $K$ NN [121] aux problèmes de régression [154].

### 3.2 Prédiction de charge interne de joueurs de football

Dans le milieu sportif, les préparateurs physiques et entraîneurs doivent planifier les entraînements (rythme, contenu, etc.) en fonction de la charge de travail de leurs effectifs, cette dernière est grandement influencée par la charge interne [40, 11]. Comme nous l'avons constaté dans le chapitre 1, bien quantifier et anticiper la charge de travail représente des enjeux cruciaux pour l'optimisation des performances et la minimisation des risques de blessure sans contact.

Le volume d'une activité physique est souvent mesuré par des indicateurs issus de capteurs GPS et accéléromètres, et son intensité par des indicateurs subjectifs comme le RPE. Après chaque session d'entraînement ou match, les athlètes doivent renseigner un formulaire à propos de la charge interne/difficulté ressentie durant leur activité. En pratique, ils remplissent occasionnellement les RPE, ce qui cause un grand nombre de valeurs manquantes au sein du jeu de données ainsi collecté. Modéliser la charge interne comme un processus temporel régulier (*i.e.* à pas de temps constant) pose ainsi un certain nombre de problèmes évidents (incomplétude des données), l'imputation des charges internes manquantes représente alors une solution importante pour le suivi du RPE et sa prédiction basée sur des données antérieures.

Dans les travaux que nous présentons ici, l'imputation de données manquantes au sein de séries temporelles est revisitée sous l'angle de l'incertitude de manière à pouvoir gérer les pas de temps irréguliers pour la prédiction de charge interne (RPE). Le RPE est discrétisé en 3 classes de proportions comparables, les 3 classes représentent 3 niveaux de difficulté: faible, normale et élevée. La nature subjective du RPE et l'irrégularité dans son remplissage induisent une forte incertitude quant à la vraie valeur de la charge interne ressentie par les joueurs.

Le formalisme des fonctions de croyances [100, 101] permet une modélisation souple de l'incertitude relative au RPE et l'algorithme EKNN [121] basé sur ce formalisme est capable de prendre en compte cette incertitude, inhérente aux données d'apprentissage (labels incertains), pour des tâches de prédiction. L'algorithme EKNN, permet ainsi de prendre en compte l'incertitude des labels modélisée sous forme de fonctions de croyance lors de prédictions.

Dans ce qui suit, la problématique de prédiction de la charge interne et les données utilisées sont présentées en section 3.2.1, ensuite nous détaillons l'approche crédibiliste proposée en termes de modélisation de l'incertitude des labels imputés en section 3.2.2 et de prédiction, ensuite nous exposons les expériences menées lors de ces travaux en section 3.2.3, puis pour finir des conclusions et perspectives sont proposées pour ce travail en section 3.2.4.

### 3.2.1 Problématique et données

Après avoir formalisé le problème de prédiction de charge interne et les différentes variables considérées, les données disponibles sont présentées.

#### 3.2.1.1 Modélisation du problème

La charge interne (RPE) est une conséquence directe de l'activité physique. En outre, on fait l'hypothèse qu'elle constitue un processus dynamique dont les valeurs aux différentes dates ne sont pas indépendantes. La manière de réagir à l'effort dépend aussi des car-

actéristiques physiques d'un athlète. On pose alors:

$$Y_{t_i}^p = f(Y_{t_{i-1}}^p, \dots, Y_{t_{i-q}}^p, X_{t_i}^p, X_{t_{i-1}}^p, \dots, X_{t_{i-q}}^p, Z^p) \quad (3.1)$$

avec

- $Y_{t_i}^p \in \Omega^Y = \{C_1, C_2, C_3\}$  le RPE discrétisé du joueur  $p$  au temps  $t_i$
- $X_{t_i}^p = (X_{t_i}^{p,1}, \dots, X_{t_i}^{p,14})$  synthétise l'entraînement suivi par l'athlète  $p$  à la date  $t_i$  avec  $X_{t_i}^{p,j} \in \Omega^{X^j} = \{0, 1\}$  une variable binaire spécifiant si l'exercice  $j$  a été réalisé par le joueur  $p$  au temps  $t_i$
- $Z^p = (Z_1^p, Z_2^p, Z_3^p) \in R_+^{*3}$  le vecteur d'attributs personnels de l'athlète  $p$  contenant son âge (années), son poids (kg) et sa taille (cm)
- $q$  la taille de l'historique des données considérées.

L'enjeu est alors d'estimer la "meilleure" fonction  $f$  capable, pour chaque athlète  $p$ , de prédire  $Y_{t_i}^p$  à partir de son historique de charge interne  $(Y_{t_{i-1}}^p, \dots, Y_{t_{i-q}}^p)$ , de son historique d'entraînement  $(X_{t_i}^p, X_{t_{i-1}}^p, \dots, X_{t_{i-q}}^p)$  et de ses attributs personnels  $Z^p$ . L'échelle de RPE utilisée dans ces travaux est l'échelle modifiée proposée par [155]. L'échelle utilisée sert à faire une correspondance avec la fréquence cardiaque *e.g.* un RPE de 8 équivaut à 80% de la fréquence cardiaque.

### 3.2.1.2 Données disponibles

Dans cette section, les données utilisées pour les expérimentations sont décrites. Elles concernent 27 joueurs professionnels de football d'une équipe française de deuxième division lors de la saison 2017. Les données collectées couvrent la période allant du 26 Juillet 2017 au 15 Novembre 2017. Sur cette période nous avons 51 dates de session (entraînements ou matchs) pour lesquelles au moins un joueur a renseigné un RPE.

Soit  $D = (X, Y, Z)$  avec

$$\bullet X = \begin{pmatrix} x_{t_1}^1 & \dots & x_{t_1}^P \\ \vdots & & \vdots \\ x_{t_T}^1 & \dots & x_{t_T}^P \end{pmatrix}$$

la matrice de variables d'entraînements ordonnées chronologiquement de taille  $T \times P$  avec  $P = 27$  et  $T = 51$  où chaque élément contient 14 variables binaires associées aux différents exercices possibles lors d'un entraînement (voir Tableau 3.1).

$$\bullet Y = \begin{pmatrix} y_{t_1}^1 & \dots & y_{t_1}^P \\ \vdots & & \vdots \\ y_{t_T}^1 & \dots & y_{t_T}^P \end{pmatrix}$$

la matrice de labels numériques  $\Omega^Y \in \{0, \dots, 10\}$  ordonnés chronologiquement de taille  $T \times P$  avec  $P = 27$  et  $T = 51$  où chaque colonne représente les labels d'un athlète (joueur) évoluant dans le temps. Certains labels numériques de  $Y$  sont manquants, ce qui motive une procédure d'imputation. Après imputation, les labels numériques sont transformés en labels catégoriels à travers une discrétisation en 3 classes  $\Omega^{\tilde{Y}} \in \{C_1, C_2, C_3\}$ .

$$\bullet Z = \begin{pmatrix} z_1^1 & \dots & z_1^P \\ z_2^1 & \dots & z_2^P \\ z_3^1 & \dots & z_3^P \end{pmatrix}$$

une matrice contenant les valeurs des âges ( $z_1^p$ ), tailles ( $z_2^p$ ) et poids ( $z_3^p$ ) des  $P$  joueurs.

Le planning d'entraînement ( $X_{t_i}^{p,1}, \dots, X_{t_i}^{p,14}$ ) d'un joueur  $p$  contient 14 variables binaires représentant des exercices physiques ou évènements (soins) ayant éventuellement eu lieu. Le tableau 3.1 présente les distributions des variables d'entraînement.

**Tableau 3.1. Distributions des variables d'entraînement**

Variable	0	1
Versus	0.45	0.55
Toro	0.84	0.16
Étirements	0.96	0.04
Réath	0.93	0.07
Abdos	0.96	0.04
Gainage	0.95	0.05
Aerobie	0.77	0.23
Brz	0.98	0.024
Mob	0.98	0.02
Mult	0.72	0.28
Non	0.56	0.45
PPG	0.93	0.066
Tech	0.84	0.16
Soins	0.86	0.14

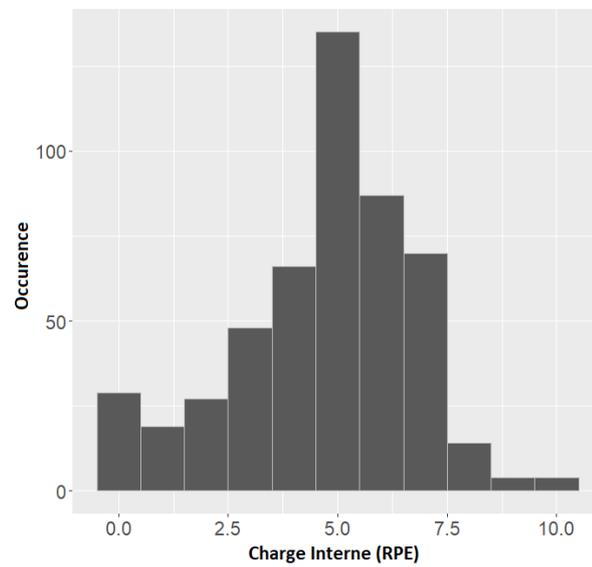
Le tableau 3.2 expose la moyenne et l'écart-type de chaque attribut personnel.

La matrice  $Y$  contient 503 RPE remplis, la figure 3.1 expose la distribution du RPE.

Dans la figure 3.1 on constate une concentration des RPE autour de la valeur 5. Pour simplifier le travail préliminaire, nous avons souhaité nous ramener à un problème de classification, alors nous avons choisi d'effectuer une discrétisation des labels (RPE) numériques (après imputation des valeurs manquantes) basée sur la distribution du RPE dans les données initiales. Au vu de cette distribution, la discrétisation suivante est donc adoptée :

Attribut	Moyenne	écart-type
Poids	76.2	8.3
Taille	180.3	7.7
Âge	29	6

**Tableau 3.2. Moyenne et écart-type des attributs personnels**



**Figure 3.1. Distribution du RPE**

- Classe 0 : RPE < 5
- Classe 1 : RPE = 5
- Classe 2 : RPE > 5

On obtient ainsi 3 classes d'effectifs relativement proches (33%). Nous discrétisons le RPE après l'imputation des valeurs manquantes, à partir du seuil "RPE=5" observé sur la distribution initiale.

### 3.2.2 Approche crédibiliste

Dans cette section nous justifions l'utilisation de théories modernes de l'incertain et présentons deux modèles d'imputation de labels numériques manquants avec modélisation de l'incertitude associée.

#### 3.2.2.1 Motivations

En général, les différentes incertitudes des données issues d'applications "réelles" sont soit ignorées, soit prises en compte avec des approches comme l'imputation multiple (voir Section 2.24). Il existe néanmoins d'autres formalismes plus souples comme celui de la théorie des fonctions de croyances que nous utilisons, ce dernier permet une modélisation fine et spécifique à la nature de l'incertitude en question. Dans notre contexte, les données contiennent plusieurs niveaux d'incertitude :

- pas de temps irrégulier : de l'irrégularité "naturelle" des entraînements (week-ends, jours fériés, vacances) et du remplissage occasionnel des RPE par les joueurs résulte une matrice  $Y$  très incomplète.
- nature subjective du PRÉ : la perception de la charge interne en est une mesure subjective et donc naturellement imprécise.

L'imputation des valeurs manquantes ajoute de l'incertitude supplémentaire aux valeurs imputées, il est alors souhaitable de modéliser l'incertitude relative aux RPE imputés à un niveau plus fin *i.e.* modéliser directement les incertitudes relatives aux valeurs imputées.

#### 3.2.2.2 Méthodes d'imputation incertaines

Nous proposons ici deux méthodes d'imputation impliquant une modélisation de l'incertitude par fonction de croyance. La première correspond à une interpolation linéaire associée à un degré d'incertitude déterminé par la taille de la fenêtre temporelle durant laquelle les joueurs n'ont pas rempli de RPE. La deuxième peut être assimilée à de l'imputation par apprentissage basée sur l'EKNN (voir Section 2.2.3).

Les charges internes (RPE) collectées couvrent une période de 114 jours, sur cette période seulement 51 jours ont au moins 1 RPE rempli.

### ***Interpolation incertaine (evid\_imp)***

La charge interne étant considérée comme un processus temporel il est envisageable d'imputer les RPE manquants à partir des précédents et des suivants. Pour chaque RPE manquant  $y_{t_i}^p \in \Omega^Y$ :

- on calcule la moyenne arithmétique entre les RPE suivants et précédents  $y_{t_{i-j}}^p$  et  $y_{t_{i+k}}^p$ :

$$\widehat{y}_{t_i}^p = \frac{1}{2}y_{t_{i-j}}^p + \frac{1}{2}y_{t_{i+k}}^p \quad (3.2)$$

- on définit ensuite la fonction de masse suivante:

$$\begin{cases} m(\widehat{y}_{t_i}^p) &= \alpha_i^p \\ m(\Omega^Y) &= 1 - \alpha_i^p \end{cases} \quad (3.3)$$

où  $\alpha_i^p$  représente le degré de certitude associé à  $\widehat{y}_{t_i}^p$  défini par

$$\alpha_i^p = \frac{\beta}{\Delta t_i^p} \quad (3.4)$$

avec  $\beta \in [0, 2]$  une constante à fixer (empiriquement ou de façon experte) et  $\Delta t_i^p = t_{i+k-1} - t_{i-j+1}$  correspond à la période durant laquelle le joueur  $p$  n'a pas rempli de RPE.

Le modèle (3.3) permet d'attribuer une masse à  $\Omega^Y$ , *i.e.* à l'ignorance, d'autant plus grande que la période sans RPE renseigné par le joueur  $p$  est longue. Pour les RPE remplis (charges internes disponibles) et donc certains, on a  $m(y_{t_i}^p) = \alpha_i^p = 1$ . L'hyperparamètre  $\beta$  permet de gérer la vitesse de décroissance du poids de l'ignorance lorsque  $(k, j) \rightarrow (0, 0)$  (*i.e.* lorsque la période sans RPE renseignée devient de plus en plus courte), plus  $\beta$  est proche de 2, plus la masse décroît lentement.

### ***Imputation basée sur l'EKNN (EKNN\_imp)***

Dans cette approche l'idée est d'utiliser les joueurs qui ont suivi le même entraînement au temps  $t$  comme voisinage du joueur qui n'a pas rempli son RPE (pas de charge interne attribuée) en  $t$ , le nombre de voisins varie selon le nombre de joueurs présents à l'entraînement  $t$ . Les attributs personnels des voisins permettent le calcul des distances entre le joueur pour lequel la charge interne est à prédire/imputer et ses voisins (de manière à pondérer leur votes).

On utilise ensuite l'E $K$ NN pour imputer le RPE manquant. Pour cette méthode d'imputation, on considère le RPE comme une variable catégorielle à 11 classes et nous appliquons la règle du maximum de croyance (2.17) pour l'imputation de  $\widehat{y}_{t_i}^p$ . Sachant que les fonctions de masse des joueurs qui ont participé à l'entraînement sont des masses dogmatiques *i.e.*, qu'on n'attribue de masse qu'au singleton (le RPE rempli par le joueur), la fusion conjonctive dans l'E $K$ NN donne une masse  $m_s$  où les éléments focaux ne sont que des singletons en plus de la masse sur l'ensemble  $\Omega^Y$ , cela fait que  $\max(\text{masse}) = \max(\text{croyance}) = \max(\text{plausibilité}) = \max(\text{pignistique})$ <sup>1</sup>. On peut remarquer que vu que le modèle d'incertitude (2.22) associé aux voisins dans le modèle de l'E $K$ NN ne considère comme éléments focaux que les singletons  $\{C_q\}$  observés et l'espace  $\Omega^Y$  entier, toute fusion conjonctive des masses des voisins résultera en une masse de croyance  $m$  ne possédant comme élément focaux que des singletons et l'espace  $\Omega^Y$  entier. Pour une telle masse, au niveau prédictif, on peut remarquer que  $\max(m) = \max(\text{Bel}) = \max(\text{Pl}) = \max(\text{BetP})$ .

### 3.2.2.3 Omission des valeurs manquantes

Conjointement aux méthodes d'imputation incertaines proposées en Section 3.2.2.2, une approche par omission des valeurs manquantes est aussi considérée dans cette étude. L'approche par omission (notée *precise*) consiste à ne pas imputer les valeurs manquantes, la subtilité de cette approche est d'associer une ancienneté aux valeurs historiques de RPE. Par exemple, si nous souhaitons prédire le RPE d'un joueur  $p$  lors d'une session  $t$  en fonction de son planning d'entraînement et de son RPE lors de l'entraînement précédente, si le RPE précédent du joueur  $p$  à l'entraînement  $t - 1$  est manquant, nous considérons l'entraînement le plus récent  $t - j$  (par rapport à  $t$ ) où le joueur  $p$  a rempli son RPE et nous lui associons une valeur d'ancienneté qui correspond au nombre de jours  $j - 1$  durant lesquels le joueur  $p$  n'a pas rempli de RPE.

## 3.2.3 Expériences

Pour évaluer les 3 méthodes d'imputation proposées en section 3.2.2.2 dans un cadre de classification, nous définissons ici une méthode alternative de référence (notée *baseline*), nous décrivons le protocole d'évaluation défini de façon chronologique et présentons les résultats obtenus.

### 3.2.3.1 Protocole d'évaluation chronologique

Nous choisissons de nous placer dans un contexte de prédiction chronologique du futur à partir du passé car il correspond à la réalité terrain du préparateur physique.

<sup>1</sup>Ces travaux sont antérieures à l'élaboration de l'espérance pignistique : ces travaux pourront être refait avec l'usage de la régression pour le RPE

L'intervalle temporel sur lequel les RPE (charges internes) ont été renseignés s'étale sur une période de 114 jours contenant 51 dates d'entraînement. De manière à disposer de suffisamment de données d'apprentissage et de test, les 25 premières et 11 dernières dates sont respectivement toujours utilisées en apprentissage et en test, les dates intermédiaires sont initialement contenues dans les données de test puis progressivement ajoutées aux données d'apprentissage. La prédiction naïve (baseline) à chaque itération d'apprentissage est constituée de la classe majoritaire dans les données d'apprentissage (*i.e.* la valeur de RPE passée la plus fréquente). Pour la méthode d'imputation par  $EKNN$  ( $EKNN\_imp$ ), pour chaque date  $t$ , le voisinage de chaque joueur correspond aux joueurs ayant renseigné un RPE en  $t$ .

Une fois les RPE manquants (dans les données d'apprentissage) imputés, au niveau prédictif (du RPE suivant  $t$ ), les prédictions par  $KNN$  ou  $EKNN$  sont calculées relativement aux 5 plus proches voisins au sein des données d'apprentissage (une rapide étude de sensibilité a montré un faible impact de ce paramètre sur les performances prédictives).

### 3.2.3.2 Design expérimental

Dans cette section, les performances prédictives de différents modèles d'apprentissage supervisé (classification) associés à différents modèles d'imputation sont comparées relativement à différentes tailles d'historique considérées. Les modèles prédictifs considérés utilisent les valeurs passées de RPE pour prédire les valeurs futures, la sensibilité des performances prédictives obtenues dépend donc naturellement de l'horizon temporel passé considéré et dans le contexte de l'entraînement, la question du choix de l'amplitude de cet horizon reste une question ouverte de par la multiplicité des facteurs intervenant dans le repos et la régénération musculaire. En première approche, nous nous sommes restreints à de petites valeurs de  $q$  (horizon temporel passé) du fait de l'absence d'information allant dans le sens d'une grande autocorrélation temporelle de la charge de travail. La valeur  $q = 0$  (voir Equation prédictive (3.1)) indique qu'aucun historique n'est pris en compte, seul le planning d'entraînement présent ainsi que les attributs personnels sont utilisés comme facteurs prédictifs. 3 niveaux de comparaison sont ainsi proposés pour quantifier les apports des :

- 3 modèles d'imputations proposées dans cette étude: *evind\_imp*, *EKNN\_imp* et *precise*.
- 2 modèles d'incertitude associés aux labels basés pour l'un, sur la taille des périodes sans données de RPE (Equations (3.3) et (3.4)) et pour l'autre sur les similarités entre exemples comme initialement proposé dans l' $EKNN$  (Equation (2.22) et [121]).

<b>q</b>	<b>baseline</b>	<b>precise</b>	<b>evid_imp</b>	<b>EKNN_imp</b>
0	0.34 ±0.02	<b>0.50</b> ±0.02	0.43 ±0.02	0.44 ±0.02
1	0.34 ±0.01	<b>0.42</b> ±0.05	0.40 ±0.04	0.39 ±0.03
2	0.34 ±0.02	<b>0.41</b> ±0.06	0.38 ±0.03	0.35 ±0.03
3	0.36 ±0.02	<b>0.42</b> ±0.04	0.40 ±0.03	0.38 ±0.02
4	0.35 ±0.02	<b>0.42</b> ±0.05	0.38 ±0.03	0.35 ±0.02
5	0.36 ±0.02	0.40 ±0.04	<b>0.40</b> ±0.03	0.37 ±0.02

**Tableau 3.3. Comparaison des performance prédictives de la baseline et du KNN associées à différents modèles d'imputation avec labels précis pour différentes tailles d'historique  $q$**

- 2 modèles prédictifs sont considérés dans ce travail: le KNN et l'EKNN relativement à la baseline (moyenne des RPE passés).

Une première expérimentation est effectuée sur les 3 jeux de données issus des 3 modèles d'imputation (*precise*, *EKNN\_imp* et *evid\_imp*) sans considérer les incertitudes associés aux imputations. Cette comparaison permet de mesurer l'intérêt des différents modèles d'imputation dans le contexte classique de données précises (*i.e.* sans modèle d'incertitude). Par souci de cohérence avec les autres expérimentations de cette étude, le seul modèle prédictif utilisé ici est le KNN. Différentes tailles d'historique ( $q = 0, \dots, 5$ ) sont considérées, l'évaluation des performances prédictives (du KNN) est réalisée en termes de justesse (*accuracy*) moyenne de prédiction ( $\pm$  écarts-types) (Tableau 3.3).

Une seconde expérimentation est réalisée en remplaçant le KNN par l'EKNN, de manière à confirmer mesurer l'intérêt de la fusion conjonctive entre voisins par rapport à leur vote.

Enfin, pour la troisième expérimentation, nous considérons les 2 jeux de données incertains car imputés par les méthodes d'imputation crédibiliste (*EKNN\_imp*, *evid\_imp*). Pour ces 2 jeux de données, les labels (RPE) sont représentés par des masses de croyance (voir Section 3.2.2.2). Nous comparons donc les performances prédictives de l'EKNN sur ces 2 jeux de données avec prise en compte des incertitudes d'imputation crédibilistes (Tableau 3.5).

### 3.2.3.3 Résultats

Sur le Tableau 3.3 on constate que les données possèdent un pouvoir prédictif, les différentes approches d'imputation associées au KNN obtenant presque toujours de meilleures performances que la baseline. Globalement, les meilleures justesses de prédiction, à savoir 0.50 et 0.51 (Tableaux 3.3 et 3.4) sont obtenues sur le jeu de données imputées par omission (modèle *precise*) pour  $q = 0$ . De manière générale il apparaît que, pour tous les jeux de données imputés, la prise en compte de données historiques

q	baseline	precise	evid_imp	EKNN_imp
0	0.34 ±0.02	<b>0.51</b> ±0.01	0.44 ±0.01	0.46 ±0.02
1	0.34 ±0.02	<b>0.46</b> ±0.03	0.42 ±0.04	0.42 ±0.02
2	0.34 ±0.02	<b>0.44</b> ±0.03	0.42 ±0.02	0.38 ±0.02
3	0.36 ±0.02	<b>0.44</b> ±0.02	0.38 ±0.02	0.37 ±0.02
4	0.35 ±0.02	<b>0.43</b> ±0.04	0.40 ±0.04	0.36 ±0.02
5	0.36 ±0.02	0.40 ±0.03	<b>0.41</b> ±0.03	0.39 ±0.02

**Tableau 3.4.** Comparaison des performance prédictives de la baseline et de l’EKNN associées à différents modèles d’imputation avec labels précis pour différentes tailles d’historique  $q$

q	evid_imp	EKNN_imp
0	<b>0.48</b> ±0.01	0.47 ±0.02
1	<b>0.48</b> ±0.01	0.45 ±0.02
2	<b>0.44</b> ±0.02	0.36 ±0.03
3	<b>0.41</b> ±0.01	0.38 ±0.03
4	<b>0.39</b> ±0.02	<b>0.39</b> ±0.01
5	<b>0.42</b> ±0.02	0.41 ±0.03

**Tableau 3.5.** Comparaison des performance prédictives de l’EKNN en fonction du modèle d’imputation incertaine considéré pour différentes tailles d’historique  $q$

nuit aux performances prédictives. Ceci implique l’absence de mémoire (*i.e.* d’auto-corrélation) du processus de charge interne entre les entraînements. Ce résultat est donc surprenant du point de vue de informationnel, mais peut constituer un résultat important du point de vue cognitif et sportif. La charge interne ne semble ainsi pas impactée par ses niveaux antérieurs, même si on pouvait s’attendre à ce que la difficulté ressentie par un joueur dépende de ses niveaux de difficulté ressentis précédents, au moins dans un passé proche. Une potentielle explication pourrait être liée au sommeil qui permettrait une réelle régénération physico-cognitive qui rendrait la difficulté ressentie par les joueurs surtout dépendante des conditions présentes d’entraînement.

En termes d’imputation, le modèle *precise* permet presque toujours d’obtenir les meilleurs performances prédictives par rapport aux modèles *evid\_imp* et *EKNN\_imp*. Cette observation est un peu contre-intuitive par rapport à la littérature scientifique sur l’imputation de données manquantes qui a largement démontré la faiblesse de l’imputation par omission dans un contexte prédictif de données relativement limitées. L’explication réside probablement dans l’utilisation de l’information d’ancienneté des RPE renseignés par le modèle d’imputation *precise*. Ce résultat va donc dans le sens d’un plus grand rôle du calendrier d’entraînement (et donc de charges attribuées aux joueurs) par rapport à la dynamique des entraînements (et des charges associées).

On remarque aussi sur les Tableaux 3.3 et 3.4 que les prédictions faites par *KNN*

deviennent moins justes au fur et à mesure que la taille de l'historique pris en compte grandit même si l'*EKNN* permet d'obtenir de légèrement meilleures prédictions que le *KNN*, la différence étant plus marquée lorsque des données historiques sont prises en compte ( $q > 0$ ). La prise en compte de l'incertitude semble donc ne présenter ici qu'un léger intérêt et uniquement au niveau de la fusion entre voisins lors de l'apprentissage de l'*EKNN*.

Les performances prédictives obtenues sur les jeux de données issus des modèles d'imputation *evind\_imp* et *EKNN\_imp* semblent montrer que l'*EKNN* est plus adapté à ces derniers que le *KNN* et la prise en compte des modèles d'incertitude des labels imputés permet à l'*EKNN* presque systématiquement d'augmenter les performances prédictives réalisées sur les jeux de données imputés. Le jeu de données résultant de la méthode d'imputation *evind\_imp* obtient des prédictions plus justes en moyenne que les prédictions faites sur le jeu de données imputé avec la méthode *EKNN\_imp*. On peut noter que les justesses moyennes obtenues par l'*EKNN* sur le jeu de données avec imputation *evind\_imp* associée au modèle d'incertitude pour de petits historiques sont légèrement inférieures à celles obtenues sur le jeu de données *precise* par *EKNN* mais presque toujours supérieures à celles obtenues par *KNN* pour la même configuration. Le modèle d'incertitude (3.3)-(3.4) semble donc adapté à la méthode d'imputation *evind\_imp* même si son utilisation n'apporte pas de gain prédictif significatif sur ce jeu de données.

### 3.2.4 Perspectives

Ces travaux ont proposé deux méthodes d'imputation crédibilistes de données manquantes pour des séries temporelles à pas de temps variable. L'utilisation d'un classifieur crédibiliste (*EKNN*) pour la prédiction de la charge interne à partir de données historiques de RPE, d'entraînement et d'attributs personnels montre certaines limites qui ouvrent la voie à différentes perspectives:

- l'utilisation d'autres classifieurs capables de prendre en compte des labels incertains [156, 157, 158, 123]
- le développement d'une approche mixte combinant imputation incertaine et l'utilisation des données d'ancienneté
- l'extension du protocole d'évaluation au cas de la régression (le RPE étant naturellement numérique).

L'implémentation de l'extension du package R 'evclass' aux labels incertains pourra servir de benchmark pour différentes méthodes d'apprentissage à partir de données incertaines. Comme il a été évoqué ci-dessus, le RPE et la plupart des indicateurs de

charge/performance étant numériques, l'extension de l'EKNN à la régression a été abordée dans les travaux qui ont suivi.

Du fait de la pandémie COVID-19, nous n'avons pas pu mettre en place une collecte de données sportives pour améliorer l'imputation crédibiliste et la tester sur un jeu de données complet éventuellement bruité artificiellement. Pour les travaux qui vont suivre nous nous sommes donc servis d'un jeu de données COVID-19, ce qui nous a permis de contribuer, à notre niveau, aux recherches visant à endiguer ou gérer la pandémie.

### 3.3 Imputation crédibiliste pour la prédiction de mortalité quotidienne par COVID 19

Dans un contexte d'apprentissage supervisé, le traitement des valeurs manquantes est une question incontournable dans tous les domaines y compris dans le sport de haut niveau ou l'épidémiologie. Du fait de la crise sanitaire COVID-19, qui a empêché la mise en place d'un protocole de collecte de données sportives, nous avons choisi un jeu de données COVID-19 relatif aux nouveaux cas et décès quotidiens de manière à rester dans un contexte de données chronologiques *i.e.* où les observations sont ordonnées dans le temps comme pour les indicateurs de charges de travail et de performances en sport et à contribuer à notre humble niveau à l'effort général de lutte contre la pandémie. Pour une tâche de prédiction de séries temporelles, quelle que soit leur nature, les valeurs manquantes introduisent des pas de temps irréguliers qui contredisent l'hypothèse fondamentale de la grande majorité des méthodes de traitement de séries temporelles standard, *i.e.* le pas de temps constant entre observations (période d'échantillonnage constante). La solution consiste généralement en l'imputation des données manquantes de manière à ramener le pas de temps à une valeur constante.

Le choix de la méthode d'imputation des valeurs manquantes dépend fortement de la nature des valeurs manquantes, - MAR: les valeurs sont manquantes à cause d'autres variables observées, - NMAR les valeurs sont manquantes à cause d'elles mêmes et - MCAR: les valeurs sont manquantes par pur hasard (voir Section 2.3). [137, 139]. Ces catégories indiquent la raison pour laquelle les données sont manquantes, l'origine de l'incomplétude des données qui représente une information importante pour le choix de la méthode d'imputation [143]. Dans l'application (football) de la Section 3.2, les valeurs manquantes n'étaient probablement pas de type MCAR, *e.g.* un athlète qui ne renseigne pas son RPE après un entraînement très/trop difficile. Cette réalité pourrait expliquer les faibles performances prédictives obtenues avec ou sans modèle d'incertitude (Equations 3.3 et 3.4). Dans cette Section, notre objectif est de prédire le nombre de décès quotidiens en France dus au COVID-19 à partir d'un jeu de données artificiellement bruitées. Certains labels numériques (nombre de nouveaux décès) sont supprimés de manière aléatoire avec une fréquence plus ou moins grande, ce qui donne

lieu à des valeurs manquantes MCAR car l'absence de valeurs est purement aléatoire. La structure du jeu de données est simple et adaptée à la prédiction de séries temporelles. Nous proposons une méthodologie pour traiter l'incertitude inhérente aux méthodes d'imputation des valeurs manquantes similaire à celle présentée en Section 3.2.2 avec des modèles d'incertitude légèrement plus sophistiqués. Les indicateurs de charge de travail (externe et interne) et de performances dans le sport sont, pour la plupart, numériques. Du fait de l'incapacité de la plupart des modèles d'apprentissage automatique à traiter les valeurs manquantes [141, 142], nous avons voulu étendre l'approche d'imputation incertaine (Section 3.2) ainsi que l'EKNN (Section 2.2.3) à la régression.

Certaines méthodes de *soft computing* sont conçues pour modéliser l'incertitude des données à partir d'information de type (aléatoire, épistémique etc.) et de niveaux généralement quantitatifs [159, 160, 161]. Dans ce contexte, l'ignorance relative aux données manquantes correspond à des niveaux d'incertitude potentiellement différents qui pourraient être pris en compte dans les modèles d'imputation choisis. Les avantages de l'association de modèles d'incertitude aux méthodes d'imputation sont évalués dans cette étude. Nous précisons que l'incertitude traitée dans ce travail est considérée épistémique pour se placer dans un contexte de données manquantes sans connaissance de leur modèle génératif même si, en l'occurrence, le modèle de bruitage est défini par nous mêmes, mais n'est pas utilisé (dans les modèles d'incertitude). Comme évoqué ci-dessus, ce travail a donné lieu à l'extension de l'EKNN aux problèmes de régression, ce qui a permis la proposition d'une méthode d'imputation basée sur de l'information chronologique et l'EKNN régression (notée 'Time-EKNN'). Nous évaluons la performance prédictive d'une extension de l'algorithme Evidential- $K$  Nearest Neighbors (EKNN) que nous avons proposée pour les problèmes de régression sur la base de [121, 154, 162].

Pour cette application COVID, nous utilisons un jeu de données complet artificiellement et aléatoirement bruité de manière à se placer dans le contexte de données manquantes de type MCAR, où l'indépendance des *absences* de données simplifie grandement les modélisations. En terme d'incertitude, l'incertitude relative aux valeurs manquantes peut être interprétée comme une ignorance totale ou une imprécision complète des valeurs réelles mais non-observables. Les performances prédictives sont évaluées une fois les données manquantes imputées avec et sans modèles d'incertitude (dans ce dernier cas, les labels numériques imputés sont considérés comme certains). Dans la suite, nous présentons le problème prédictif en Section 3.3.1, l'extension de l'EKNN à la régression en Section 3.3.2 et nous proposons trois méthodes d'imputation de valeurs manquantes en Section 3.3.4. Dans la Section 3.3.5, nous présentons les modèles d'incertitude associés aux méthodes d'imputation précédemment introduites, et en Section 3.3.6, nous commentons les résultats expérimentaux.

### 3.3.1 Formalisme de la tâche prédictive considérée

Dans cette section, nous présentons le formalisme du problème de régression et de la tâche d'imputation de valeurs manquantes.

Soit  $D = (X, Y)$  avec  $X = (x_1, \dots, x_T)$  et  $Y = (y_1, \dots, y_T)$  les séries ordonnées chronologiquement représentant respectivement les attributs et les labels des  $T$  observations où  $\forall t = 1, \dots, T, (x_t, y_t)$  représente respectivement les nombres de nouveaux cas et de nouveaux décès COVID-19 quotidiens.

L'objectif de la régression est d'apprendre une fonction  $f$  liant les valeurs futures des labels numériques aux valeurs présentes et passées des attributs et aux valeurs passées des labels:

$$y_t = f(y_{t-h}, \dots, y_{t-(h+q)}, x_{t-h}, \dots, x_{t-(h+q)}) \quad (3.5)$$

où  $h$  est l'horizon de prédiction,  $q$  une taille de fenêtre passée pour les attributs et les labels numériques à prendre en compte.

Nous nous plaçons dans le contexte où le vecteur de labels numériques  $Y$  est incomplet et contient donc des données manquantes. Dans l'équation (3.5), les labels numériques passés  $y_t$  étant des entrées de la fonction  $f$ , le processus d'imputation doit avoir lieu en amont du processus d'apprentissage du régresseur  $f$ .

Du fait de la nature numérique du label  $Y$  considéré dans cette étude, *i.e.* le nombre de décès quotidiens dus au COVID-19, nous proposons une extension de l'algorithme  $EKNN$  aux problèmes de régression basée sur [121] et proche de [154, 162].

### 3.3.2 Extension de l' $EKNN$ à la régression

L'approche  $EKNN$  présentée en Section 2.2.3 et initialement proposée par Denoeux [121] concerne les tâches de classification. De manière à se ramener aux problèmes de régression discrète, nous avons supprimé tous les paramètres de pénalisation  $\gamma_l$  (définis par rapport aux labels des classes catégorielles) des équations (2.22) et (2.23) modélisant l'incertitude associée aux différents voisins (à partir de leur distance à l'exemple à classifier), ce qui donne les équations (3.6) et (3.7).

Nous obtenons le modèle d'incertitude suivant:

$$\text{labels certains} : \begin{cases} m_{s,i}(\{\omega_l\}) &= \alpha_0 \times \exp(-d_{s,i}^\lambda) \\ m_{s,i}(\Omega) &= 1 - m_{s,i}(\{\omega_l\}) \end{cases} \quad (3.6)$$

$$\text{labels incertains} : \begin{cases} m_{s,i}(\{\omega_l\}) &= (\alpha_0 \times \exp(-d_{s,i}^\lambda)) \times m_i(\{\omega_l\}) \\ m_{s,i}(\Omega) &= 1 - m_{s,i}(\{\omega_l\}) \end{cases} \quad (3.7)$$

La première étape de l'algorithme  $EKNN$ , *i.e.* le calcul des distances de chaque observation à l'exemple à classifier  $x_s$ , reste inchangée. A partir du modèle d'incertitude

exposé ci-dessus, les masses de croyance  $m_{s,i}$  associées aux voisins sont combinées conjonctivement et au niveau prédictif, la masse de croyance  $m_s$  issue de cette fusion est transformée en probabilité selon la transformation pignistique  $BetP_s$  proposée par Smets dans le modèle des croyances transférables ([135] et Equation (3.8)) et enfin l'espérance de cette probabilité pignistique est utilisée comme prédiction:

$$E_{BetP_s}[Y] = \sum_{\omega_l \in \Omega} BetP_s(\{\omega_l\}) \times \omega_l \quad (3.8)$$

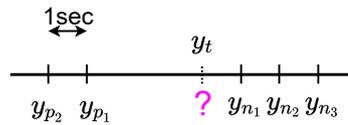
L'implémentation de cette extension de l'algorithme EKNN à la régression est disponible sur notre github<sup>2</sup>.

**Remarque:** du fait de la nature discrète des masses de croyance considérées par le modèle d'incertitude (3.6)-(3.7), l'ordre entre les valeurs de  $Y$  n'est pas considéré dans ce modèle d'incertitude, mais est utilisé dans le calcul de l'espérance pignistique.

### 3.3.3 Problématique de l'imputation temporelle

Soit  $y_t \in Y$  une valeur manquante dans une série chronologique. Considérons  $y_{p_1} \dots y_{p_U}$  les  $U$  valeurs de labels numériques précédentes connues avec  $p_U < \dots < p_1$  avant  $y_t$  et  $y_{n_1} \dots y_{n_R}$ , avec  $n_1 < \dots < n_R$ , les  $N$  valeurs connues suivantes,  $U$  et  $R$  sont des hyperparamètres à ajuster.

Dans l'exemple présenté à la figure 3.2, nous avons  $P = 2$  et  $N = 3$ .



**Figure 3.2. Imputation chronologique des données**

L'objectif du processus d'imputation est de calculer ou imputer des valeurs  $\hat{y}_t$  pour tous les  $y_t$  manquants.

### 3.3.4 Méthodes d'imputation

Dans cette section, nous présentons trois méthodes d'imputation. Les trois méthodes d'imputation étudiées dans ces travaux sont basées sur l'utilisation des valeurs de labels passées et/ou futures.

La première est appelée la méthode *Dernière Observation Rapportée* en anglais Last Observation Carried Forward notée LOCF qui remplace les valeurs manquantes par la dernière valeur connue. En reprenant la Figure 3.2 on obtient  $\hat{y}_t = y_{p_1}$ .

<sup>2</sup>[https://github.com/lgi2p/evidential\\_imputation](https://github.com/lgi2p/evidential_imputation)

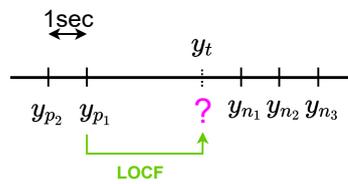
La deuxième méthode d'imputation proposée est appelée *Moyenne Mobile Centrée* en anglais *Centered Moving Average* notée *CMA* qui prend en compte la nature dynamique des données et impute les valeurs manquantes à partir des valeurs futures et passées incluses dans une fenêtre temporelle *win* centrée en  $t$ . On obtient

$$\hat{y}_t = \frac{1}{|i:t_i \in win|} \sum_{i:t_i \in win} y_i.$$

La dernière méthode d'imputation appelée *Time-EKNN* et notée *TEKNN* applique l'algorithme *EKNN* adapté à la régression selon la distance temporelle  $d^*$  définie par  $d^*(x_{t_i}, x_{t_j}) = |t_i - t_j| \quad \forall (x_{t_i}, x_{t_j}) \in \Omega^X \times \Omega^X$  pour prédire les valeurs manquantes. Comme la méthode *CMA*, cette méthode prend en compte la nature dynamique des données grâce à l'utilisation de la distance temporelle  $d^*$ .

### 3.3.4.1 Dernière Observation Rapportée (LOCF)

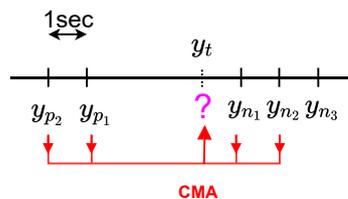
Dans cette méthode les valeurs manquantes sont remplacées par les dernières valeurs connues ou observables. En reprenant la Figure 3.2 on obtient  $\hat{y}_t = y_{p_1}$ . Soit  $y_t$  une valeur de label numérique manquant et  $y_{p_1}$  la dernière valeur connue. L'imputation *LOCF* est simplement  $\hat{y}_t = y_{p_1}$ . Pour l'exemple de la figure 3.3, nous obtenons  $y_{p_1}$ .



**Figure 3.3. Imputation par méthode LOCF**

### 3.3.4.2 Moyenne Mobile Centrée (CMA)

Nous présentons ici une méthode qui prend en compte la nature dynamique des données et impute les valeurs manquantes à partir des valeurs futures et passées incluses dans une fenêtre temporelle *win* centrée en  $t$ .



**Figure 3.4. Imputation par Mméthode CMA**

Cette méthode repose sur l'hypothèse que des labels numériques proches dans le temps sont susceptibles d'avoir des valeurs proches. Plus simplement, l'idée de la

méthode d'imputation CMA est d'imputer les  $y_t$  manquants à partir des labels numériques passés et futurs connus les plus proches. Les valeurs des labels numériques passés et futurs sont moyennées en fonction de la durée (en jours) qui les sépare du label numérique manquant  $y_t$  puis arrondis aux entiers les plus proches.

On pose  $\delta_t^p = |t - p|$  et  $\delta_t^n = |t - n|$  les décalages temporels entre l'instant  $t$  d'un label manquant  $y_t$  et respectivement les labels numériques connus précédents  $y_{p_u}$  et suivants  $y_{n_r}$ , on a alors:

$$\hat{y}_t = \sum_{u=1}^U Sim(t, p_u) \times y_{p_u} + \sum_{r=1}^R Sim(t, n_r) \times y_{n_r} \quad (3.9)$$

avec pour tout  $(t, U, R) \in \{0, \dots, T\} \times \mathbb{N}^{*2}$  :

•

$$Sim(t, p_u) = \frac{sim(t, p_u)}{\sum_{u=1}^U sim(t, p_u) + \sum_{r=1}^R sim(t, n_r)} \quad (3.10)$$

•

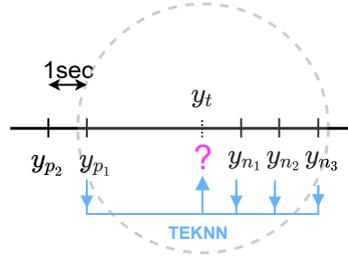
$$sim(t, p_u) = 1 - \frac{\delta_t^{p_u}}{\sum_{u=1}^U \delta_t^{p_u} + \sum_{r=1}^R \delta_t^{n_r}} \quad (3.11)$$

Les Équations (3.10) et (3.11) définissent une similarité temporelle normalisée  $Sim(t, p_u)$  entre un temps de mesure  $t$  et l'un des temps de mesure précédents  $p_u$ . Ces similarités sont utilisées pour pondérer les valeurs des labels numériques passés dans l'équation (3.9). On notera que ces équations peuvent être directement transposées pour mesurer la similarité  $Sim(t, n_r)$  entre  $t$  et n'importe quel temps de mesure suivant  $n_r$ . Les estimations sont arrondies à l'entier le plus proche.

Pour l'exemple de la figure 3.4, l'imputation CMA avec  $U = R = 2$  conduit à  $\hat{y}_t = \frac{7}{30} \times y_{p_1} + \frac{6}{30} \times y_{p_2} + \frac{9}{30} \times y_{n_1} + \frac{8}{30} \times y_{n_2}$

### 3.3.4.3 Time-EKNN (TEKNN)

L'idée sous-jacente à l'approche d'imputation TEKNN est d'utiliser le modèle de régression EKNN pour prédire les valeurs de labels numériques manquants sur la base des exemples complets, *i.e.* où les valeurs de labels numériques sont connues, qui sont proches dans le temps. Cette méthode peut être considérée comme une extension décentrée de l'approche CMA, dans laquelle les points utilisés pour l'imputation appartiennent à un voisinage temporel de  $y_t$ , quelle que soit leur localisation temporelle autour de  $y_t$  (avant et/ou après), comme le montre la figure 3.5. Pour ce modèle également, les voisins les plus proches dans le temps ont plus de poids sur les valeurs imputées  $\hat{y}_t$  (voir l'équation (3.6)), les estimations sont aussi arrondies à l'entier le plus proche comme pour la méthode CMA.



**Figure 3.5.** imputation par méthode TEKNN

Quelle que soit la méthode envisagée, l'imputation des données manquantes implique par nature une incertitude associée aux valeurs imputées. La section suivante propose des modèles d'incertitude pour les données imputées dans le cadre des fonctions de croyance.

### 3.3.5 Modélisation de l'incertitude pour les méthodes d'imputation

Dans cette section, nous présentons les modèles d'incertitude associés aux trois méthodes d'imputation décrites en Section 3.3.4. Par une modélisation incertaine, des fonctions de masse  $m_t^y$  sont attribuées à chaque label numérique  $y_t \in Y$ , que sa valeur soit manquante ou pas. Pour les valeurs de labels numériques connues, des fonctions de masse dogmatiques, *i.e.*  $m_t^y(\{y_t\}) = 1$  sont attribuées. Les méthodes d'imputation associées à un modèle d'incertitude permettent de convertir un ensemble de données précises mais incomplètes (en termes de label)  $(x, y)$  en un ensemble de données incertaines (mais) complet  $(x_t, m_t^y)_{t=1, \dots, T}$ .

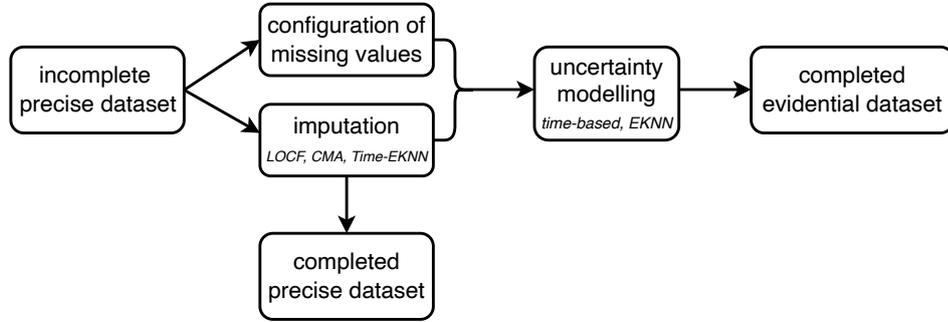
En raison de leur nature dynamique, les méthodes d'imputation LOCF et CMA sont associées au même modèle d'incertitude basé sur le temps. L'idée sous-jacente est que plus les valeurs utilisées pour l'imputation des labels manquants sont proches des labels manquants dans le temps, moins les labels numériques imputés résultants seront incertains.

#### **Modèle d'incertitude associé aux méthodes d'imputation LOCF ET CMA**

Une fois les labels manquants  $y_t$  imputés à l'aide de la méthode LOCF ou CMA, les valeurs imputées précises  $\hat{y}_t$  sont transformées en masse de croyance où l'incertitude relative aux valeurs imputées est fonction de la durée de l'absence de données disponibles avant et après ces valeurs. Ce modèle est donc basé sur les décalages temporels  $\delta_t^p = |t - p|$  et  $\delta_t^n = |t - n|$  entre les valeurs manquantes et les valeurs connues les plus proches. Plus ces décalages sont importants, plus le label imputé correspondant  $\hat{y}_t$  est incertain. Soit  $\beta \in [0, 1]$  un hyperparamètre contrôlant le niveau d'incertitude, *i.e.* la vitesse de décroissance des masses  $m_t(\{\hat{y}_t\})$  en fonction du temps écoulé entre les valeurs manquantes et les valeurs connues les plus proches. La

fonction de masse associée aux méthodes d'imputation CMA et LOCF est la suivante :

$$\begin{cases} m_t(\{\hat{y}_t\}) &= \exp\left(-\beta \times \min(\delta_t^{p1}, \delta_t^{n1})\right) \\ m_t(\Omega^Y) &= 1 - m_t(\{\hat{y}_t\}) \end{cases} \quad (3.12)$$



**Figure 3.6. Schéma d'imputation crédibiliste**

### Modèle d'incertitude associé à la méthode d'imputation TEKNN

Le modèle d'incertitude de l'approche d'imputation TEKNN est constitué de sa sortie crédibiliste, *i.e.* de la masse de croyance issue de la combinaison conjonctive des masses des voisins.

L'articulation globale entre l'imputation et les modèles d'incertitude associés est décrite dans la Figure 3.6.

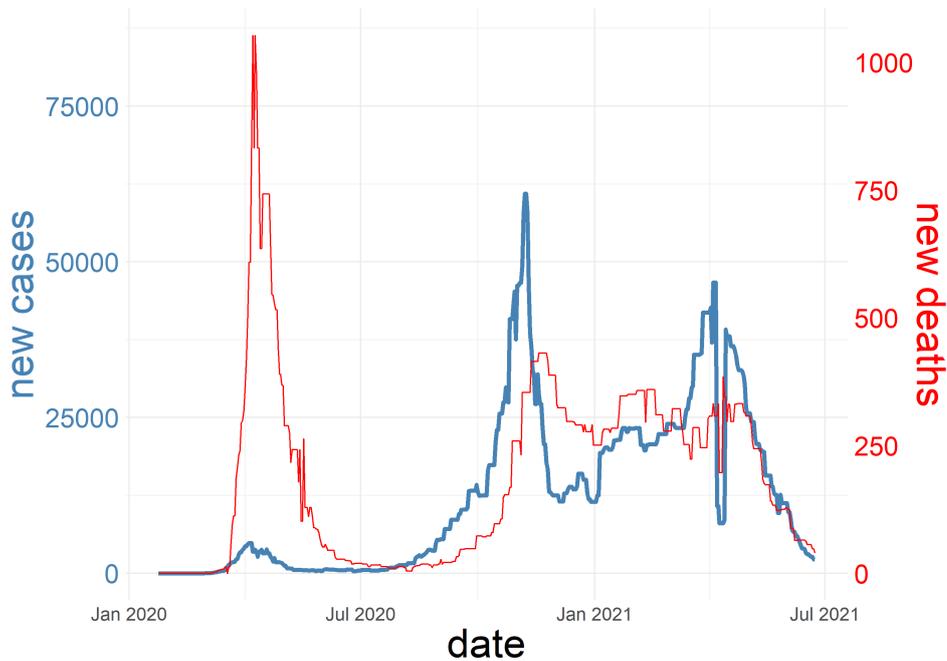
#### 3.3.6 Expérimentation

Dans cette section, nous exposons l'expérimentation menée sur un jeu de données publiques COVID-19 dans lequel certains labels (nombre quotidien de décès) sont bruitées, *i.e.* supprimés de manière aléatoire et ensuite imputés avant les phases d'apprentissage et de test. Après avoir décrit le jeu de données, nous donnons les détails de notre procédure de bruitage et de la configuration expérimentale. Enfin, nous analysons les résultats.

##### 3.3.6.1 Jeu de données

A partir du site web [ourworldindata](https://ourworldindata.org/)<sup>3</sup>, nous avons utilisé le jeu de données contenant les nombres quotidiens de nouveaux cas confirmés  $(x_t)_{t=1,\dots,T}$  et de nouveaux décès  $(y_t)_{t=1,\dots,T}$  en France métropolitaine. La figure 3.7 montre l'évolution de ces indicateurs pendant la pandémie.

<sup>3</sup><https://ourworldindata.org/>



**Figure 3.7. Évolution des nouveaux décès et des nouveaux cas**

La politique de détection du virus ayant évolué entre les deux premières vagues de la pandémie (avant et après juillet 2020), le lien entre le nombre de cas détectés et le nombre de décès semble radicalement différent au cours de ces deux périodes. Le nombre de nouveaux cas quotidiens ayant été clairement sous-estimé au cours de la première vague, nous avons limité l'expérimentation à la deuxième vague. Nous avons finalement gardé 367 observations quotidiennes complètes pour ce jeu de données.

Comme il n'y avait pas de valeurs manquantes dans le jeu de données, nous avons supprimé/bruité au hasard certaines valeurs de labels (*i.e.* nouveaux décès). Dans la sous-section suivante, nous détaillons notre procédure de bruitage.

### 3.3.6.2 Procédure de bruitage

La proportion  $v \in [0, 1]$  des valeurs de labels  $y_t$  à supprimer est l'entrée de la procédure. Afin de simuler des défauts de mesure plausibles, nous avons supprimé les labels  $y_t$  par période de temps. Après avoir choisi au hasard les centres de fenêtre de bruitage, nous avons généré aléatoirement des tailles de fenêtre raisonnables. Après des premiers labels supprimés, la procédure est itérative jusqu'à ce que la proportion de labels supprimés atteigne  $v$ .

**Data:** jeu de données original,  $v$  (niveau de bruit)  
**Result:** jeu de données bruitées contenant  $v\%$  de labels manquants  
 $\#supprimés \leftarrow 0$  ;  $s = \lfloor v \cdot T \rfloor$  nombre de labels à supprimer ;  
**while**  $\#supprimés < s$  **do**  
    centre de la fenêtre  $c \in [1 : T] \leftarrow$  génération aléatoire uniforme;  
    taille de la fenêtre  $s \in \{1, 2, 3\} \leftarrow$  génération aléatoire uniforme;  
    calcul des indices de labels à supprimer  $\{c - s, \dots, c + s\}$ ;  
    suppression des labels :  $\{y_{t-s}, \dots, y_{t+s}\}$ ;  
     $\#supprimés \leftarrow \#supprimés + 1 + 2s$ ;  
**end**

**Algorithm 1: Procédure de bruitage pour la suppression aléatoire de labels**

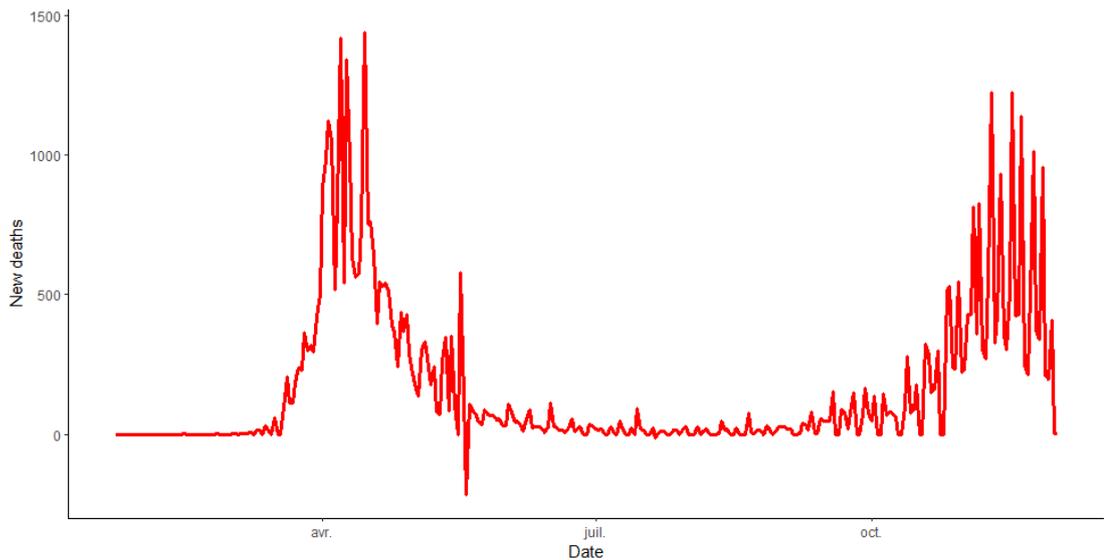
### 3.3.6.3 Lissage

En raison des contraintes hebdomadaires de la politique de collecte de données de santé, les données brutes de COVID-19 apparaissent généralement bruitées, donnant des courbes d'évolution en "dents de scie" Figure 3.8. Une méthode de lissage peut permettre d'obtenir des valeurs correspondant davantage à la réalité que les valeurs brutes. Pour y parvenir nous avons choisi d'utiliser une médiane mobile de 7 jours (labels et attributs). Ce faisant, nous avons évité les biais sans créer de valeurs déraisonnables et sans introduire de déphasage dans les données. Étant donné que les données lissées sont généralement très régulières, l'imputation des valeurs manquantes sur des données lissées n'est pas un véritable problème. Tous les modèles prédictifs mis en œuvre dans notre travail ont été entraînés et testés sur des données lissées en raison de leur niveau de fiabilité plus élevé que celui des modèles entraînés sur des données très bruitées. L'ensemble du processus expérimental est représenté dans la figure 3.9.

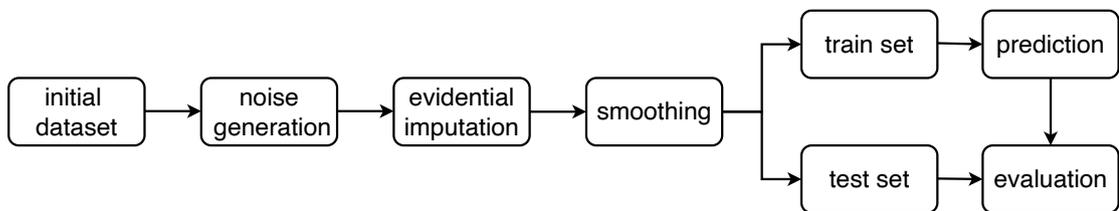
### 3.3.6.4 Mise en place de l'expérimentation

Dans cette section, nous présentons le déroulement chronologique de notre expérience et les espaces d'hyperparamètres considérés. Certains hyperparamètres ont été fixés a priori :  $R$  et  $U$  représentant la largeur de la fenêtre temporelle dans l'approche CMA 3.3.4.2 ont tous deux été fixés à 5 jours. Du fait de la réorganisation des hôpitaux ayant eu lieu durant la pandémie COVID-19 et impliquant de fortes contraintes administratives incompatibles avec un horizon trop court ou trop large, l'horizon de prédiction considéré  $h$  était de 7 jours. Enfin, l'hyperparamètre d'incertitude du modèle d'incertitude basé sur le temps  $\beta$  (voir Equation (3.12)) a été fixé à 0.05.

Pour les autres hyperparamètres, plusieurs configurations ont été envisagées :



**Figure 3.8. Données brutes : Évolution du nombre de nouveaux morts quotidien**



**Figure 3.9. Processus expérimental**

- niveau de bruit (noise level) :  $v \in \{0, 0.1, \dots, 0.7\}$
- longueur de l'historique des données :  $q \in \{1, 2, \dots, 7\}$
- nombre de voisins pour l'EKNN régression :  $K \in \{1, 10, 20\}$

La longueur de l'historique des données  $q$  représente le nombre de jours passés (décès et cas) utilisés pour décrire chaque exemple d'apprentissage. Les 21 premières dates (sur 367 dates au total) sont réservées à l'entraînement des modèles, les prédictions  $y_t$  sont ensuite calculées itérativement à chaque date  $t$  à partir de toutes les données passées des couples  $(x_{t'}, y_{t'})_{t'=t-h, \dots, t-(h+q)}$ . Lors de la première itération de l'évaluation chronologique, les 21 premiers jours sont utilisés comme données d'apprentissage afin de prédire la valeur du label du  $21+7 = 28^{\text{ème}}$  jour (avec un horizon de prédiction de  $h = 7$  jours). Ensuite, les données d'apprentissage sont augmentées d'une date à chaque itération ; *e.g.*, à la deuxième itération, nous utilisons les 22 premiers jours pour prédire le nombre de décès au  $29^{\text{ème}}$  jour. Chaque évaluation chronologique complète, *i.e.* pour les  $367 - 21 - h$  dates d'évaluation, est répétée 50 fois en raison du caractère aléatoire de la procédure de bruitages et les prédictions sont moyennées avant d'être évaluées.

Il est important de remarquer que toute approche par  $K$ -NN ne peut produire de prédiction en dehors de l'espace prédéfini des labels  $\Omega^Y$  du fait de l'agrégation (souvent par vote ou moyenne) des labels des voisins qui sont, en l'occurrence, issus des données d'apprentissage. Pour limiter ce défaut nous redéfinissons  $\Omega^Y$  à chaque itération conformément à l'Eq (3.13) en mettant à jour la valeur maximale du label dans les données d'apprentissage.

$$(\Omega^Y)_t = \{0, 1, \dots, \max_{t' < t} (y_{t'}) \times 1.15\} \quad (3.13)$$

Nous avons choisi une marge de sécurité de 15% de façon à légèrement étendre  $\Omega^Y$  et ainsi limiter les erreurs de prédiction en cas de nouvelle vague de contaminations d'une ampleur supérieure aux vagues précédentes. Comme approche de référence naïve (baseline), nous avons considéré la moyenne mobile qui prédit le nombre de décès pour la semaine suivante comme la moyenne des deux semaines précédentes.

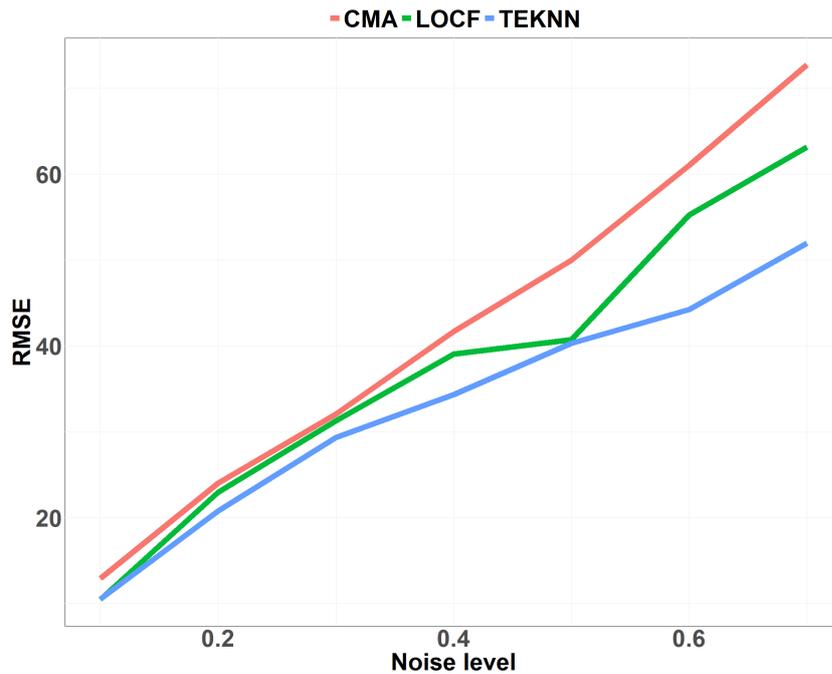
Deux types de résultats sont présentés. Les évaluations chronologiques (voir Figures 3.11 et 3.12) qui nous permettent d'évaluer visuellement les prédictions pour différents niveaux de bruit pour la méthode d'imputation TEKNN en comparant les labels prédits aux labels réels. Les modèles EKNN régression sont évalués sans et avec modèles d'incertitude sur les valeurs imputées ("EKNN" et "EKNN uncertain labels"). Dans le premier cas, les labels d'apprentissage imputés sont considérés comme certains ou précis, tandis que dans le second cas, les modèles incertains (voir les Equations (3.12) et (2.22)) permettent à l'EKNN régression de prendre en compte l'incertitude liée à l'imputation des valeurs manquantes.

La sensibilité au niveau de bruit des évaluations complètes est représentée dans les Figures 3.14, 3.15 et 3.13, où la performance prédictive est mesurée en fonction du niveau de bruit (taux de valeurs manquantes) avec et sans modèle d'incertitude. La mesure d'évaluation considérée est la racine de l'erreur quadratique médiane ou root median squared error notée RMedSE. Nous l'avons choisie plutôt que la racine de l'erreur quadratique moyenne (RMSE) standard en raison de la grande sensibilité de l'opérateur moyenne aux valeurs extrêmes, qui sont courantes dans les données COVID-19.

A contrario, nous évaluons les méthodes d'imputation en comparant le jeu de données initial et les données imputées en termes de RMSE afin d'évaluer les erreurs d'imputation qui sont, elles, peu susceptibles de contenir des valeurs extrêmes (grâce au lissage préalable des données) contrairement aux erreurs de prédiction (voir Figure 3.10).

### 3.3.7 Résultats

Tout d'abord, nous observons dans la Figure 3.10 que les trois méthodes d'imputation sont comparables en termes de performances d'imputation pour des faibles niveaux de

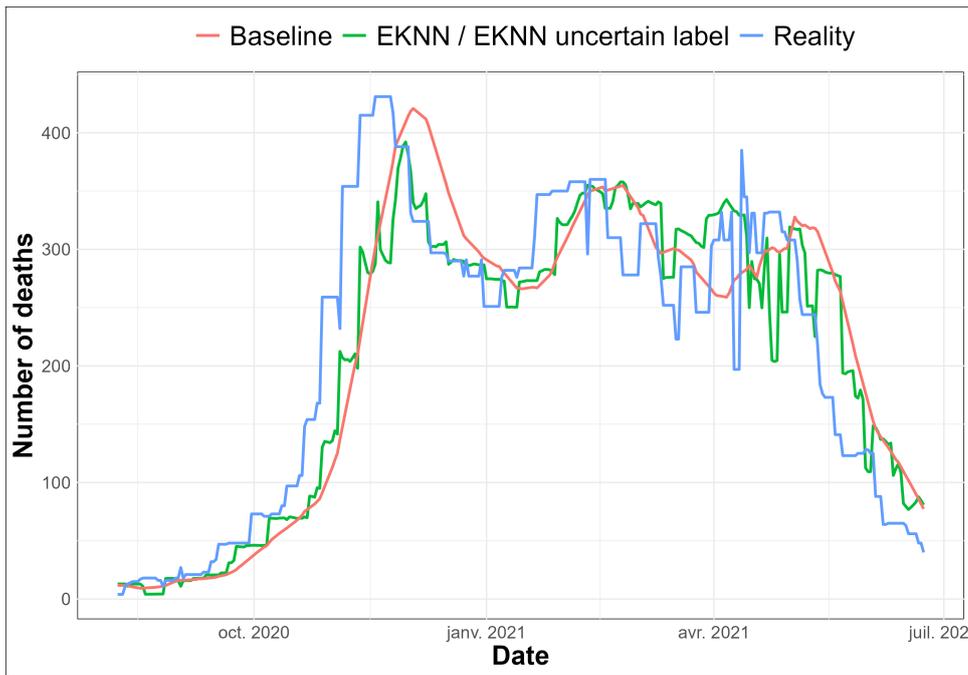


**Figure 3.10. Erreurs d'imputation**

bruit (jusqu'à 0.3 environ). Pour ces faibles niveaux de bruit la méthode TEKNN semble tout de même légèrement plus performante (RMSE plus faible) que les méthodes LOCF et CMA et cette supériorité augmente avec le niveau de bruit. Globalement, le modèle TEKNN semble être la meilleure méthode d'imputation en termes de RMSE.

Sur l'évaluation chronologique avec un niveau de bruit de 0 (Figure 3.11), nous observons que les prédictions avec et sans modèle d'incertitude associé à l'imputation des labels sont confondues. Cette observation était attendue car les données ne sont ici pas bruitées, *i.e.* il n'y a pas d'incertitude associée aux labels d'apprentissage. Pendant les phases d'augmentation et de diminution du nombre de décès, l'E $K$ NN régression semble être plus performant que l'approche de référence baseline (la courbe verte est plus proche de la courbe bleue que la courbe rouge pendant ces périodes). Pendant les périodes de stabilité relative, lorsque l'évolution de la pandémie ralentit, il ne semble pas y avoir de différences significatives entre les approches considérées.

Nous constatons un léger décalage temporel entre le nombre réel de décès quotidiens et les modèles prédictifs (baseline et E $K$ NN), en particulier au début de la vague. Cela est dû au fait que ces modèles ont besoin d'un nombre suffisant d'observations de décès dans le passé pour pouvoir prédire des valeurs correspondant au début de futures vagues. Nous observons, ainsi un déphasage au début de la vague, dû à un manque d'observations en nombre suffisant avec un nombre élevé de décès, les valeurs prédites

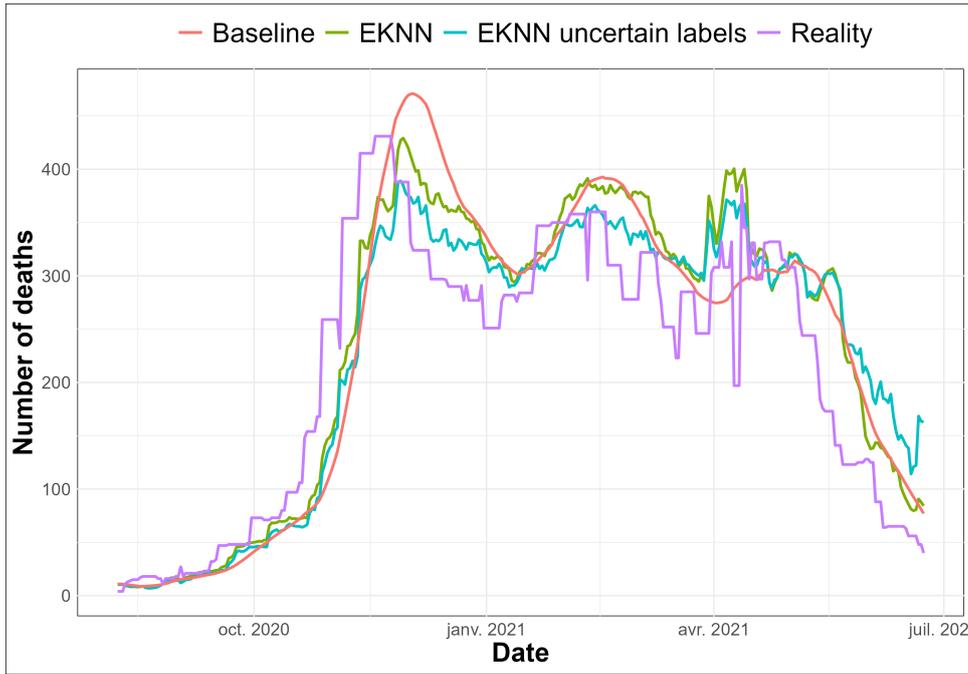


**Figure 3.11. Résultats prédictifs sur des données imputées avec la méthode d'imputation TEKNN, hyperparamètres:  $K = 1$ ,  $q = 4$ ,  $v = 0$**

sont alors sous-estimées jusqu'à ce que les données d'apprentissage contiennent suffisamment de nombres élevés de décès. Nous constatons également que le déphasage se réduit par la suite. Les meilleurs résultats avec un niveau de bruit nul ont été obtenus pour la configuration  $K = 1$ , et  $q = 4$ .

Sur la Figure 3.12, nous observons qu'avec les imputations TEKNN, l'EKNN associé au modèle d'incertitude TEKNN ('EKNN uncertain labels') et l'EKNN sans modélisation incertaine ('EKNN') font des prédictions obtenant des performances assez similaires avec une légère supériorité du 'EKNN uncertain labels' au milieu de la vague et une inversion des performances (meilleures performances de l'EKNN') pendant la fin de vague pandémique.

La figure 3.15 montre que grâce au modèle d'incertitude de la méthode d'imputation TEKNN, nous obtenons une meilleure performance prédictive que sans prise en compte de l'incertain jusqu'à un niveau de bruit élevé (environ 0.45). La supériorité de la méthode EKNN standard (*i.e.* sans modèle d'incertitude) après un niveau de bruit d'environ 0.45 est due au fait qu'un taux élevé de valeurs manquantes induit des voisinages très incertains et donc des prédictions très incertaines, une grande partie de la masse de croyance étant attribuée à l'ignorance. En effet, la transformation pignistique appliquée à la sortie de la fonction de masse de l'EKNN régression distribue la masse sur l'espace  $\Omega^Y$  de manière uniforme sur tous les singletons ; si cette masse est trop



**Figure 3.12. Résultats prédictifs sur des données imputées avec la méthode d'imputation TEKNN, hyperparamètres:  $K = 1$ ,  $q = 3$ ,  $v = 0.3$**

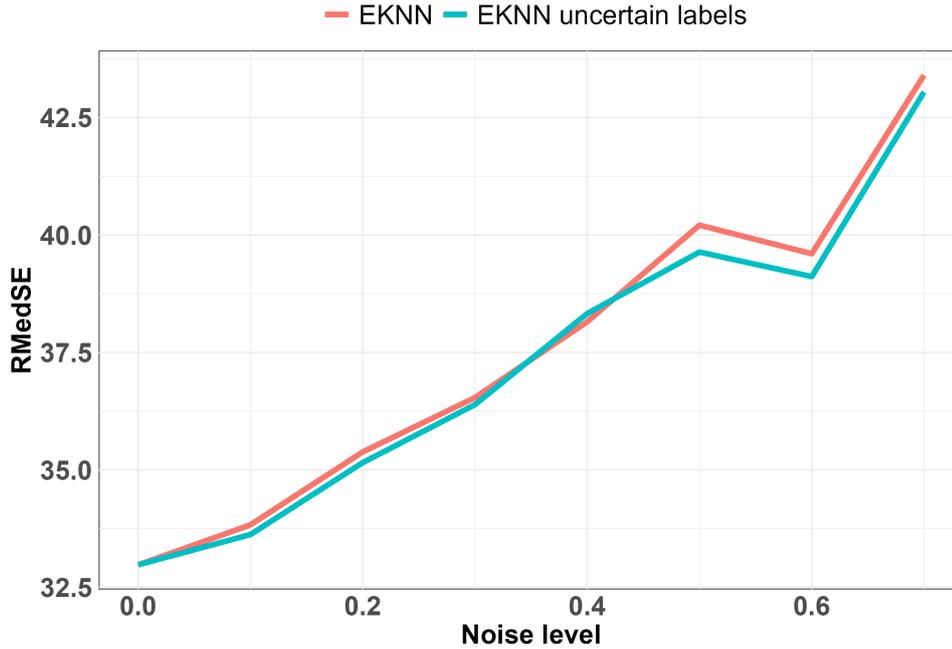
élevée, les prédictions tendent vers le centre de l'espace. A l'exception de la configuration expérimentale où  $K = 1$  avec l'imputation par TEKNN, pour toutes les autres méthodes d'imputation (CMA et LOCF) et toutes les configurations d'hyperparamètres :  $K = \{10, 20\}$  et  $q = \{1, 2, \dots, 7\}$ , l'utilisation de modèles d'incertitude pour les labels imputés nous permet d'obtenir de meilleures performances prédictives (Figures 3.13 et 3.14).

### 3.4 Conclusion

Le but de ces travaux était de proposer des modèles d'incertitude associés à des méthodes d'imputation adaptées aux données chronologiques et d'évaluer leur effet sur des tâches de prédiction. L'objectif était de prédire le nombre de décès quotidiens dus au COVID-19 à un horizon de prédiction de 7 jours avec un jeu de données artificiellement bruitées. Nous avons proposé trois méthodes d'imputation (LOCF, CMA, TEKNN) qui ont montré de bonnes performances d'imputation.

Pour notre expérimentation, nous avons étendu la méthodologie EKNN proposée dans [121] aux problèmes de régression (EKNN régression). Nous avons comparé les performances prédictives de ce modèle associé à la méthode d'imputation TEKNN avec et sans modèle d'incertitude selon différents niveaux de bruit. L'expérimentation a montré l'avantage de la modélisation de l'incertitude associée aux valeurs imputées dans plusieurs configurations d'hyperparamètres.

L'utilisation de valeurs passées imputées  $(x_t, y_t)_{t=t-q, \dots, t}$  en entrée du modèle de



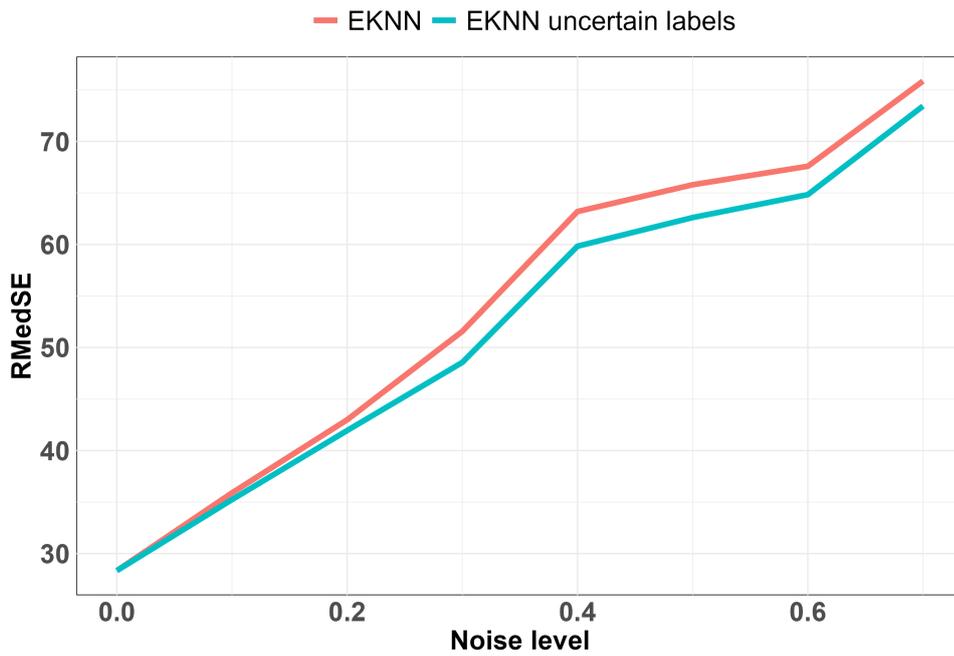
**Figure 3.13. Performances prédictives par rapport aux niveaux de bruit pour les données imputées avec la méthode d'imputation CMA :  $K = 20$  :  $q = 4$**

prédiction conduit à des valeurs d'attributs incertaines. Du fait de l'incapacité de l'E $K$ NN à gérer les attributs incertains, nous n'avons pas utilisé de modèle incertain relatif à ces valeurs de labels passées imputées pourtant incertaines. Une suite logique de ce travail pourrait être d'utiliser d'autres modèles prédictifs issus du soft computing, en particulier ceux capables de gérer les attributs incertains pendant l'apprentissage [163, 123, 164].

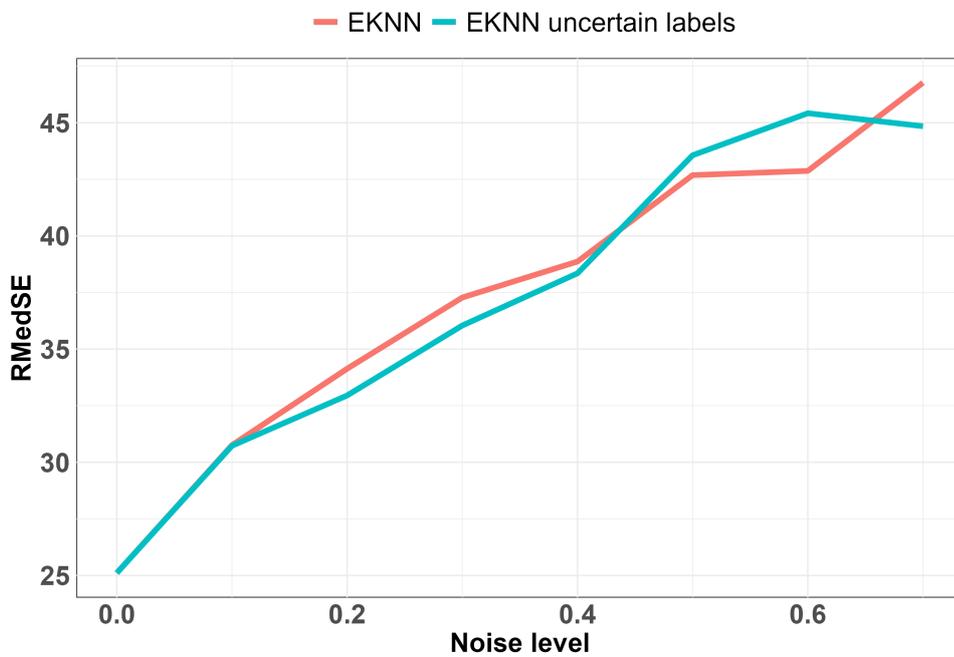
La tâche de prédiction des décès quotidiens de COVID-19 nous a conduits à un problème de régression numérique, le modèle d'incertitude basé sur le temps pourrait néanmoins être adapté à la classification (voir Section 3.12). Il sera alors nécessaire d'adapter la méthode CMA (voir 3.9) aux séries temporelles catégorielles.

En outre, certains algorithmes comme l'E $K$ NN utilisent ce formalisme. La théorie des fonctions de croyance nous permettrait d'améliorer la modélisation de l'incertitude, *e.g.* en utilisant l'imputation avec des intervalles au lieu de valeurs précises.

Les travaux présentés dans le chapitre 3 avaient pour objectif de traiter la question des valeurs manquantes, problématique très importante dans le suivi des activités physiques. La seconde problématique qui va être traitée dans ce manuscrit est celle relative à la multidimensionalité des performances. En sport les préparateurs physiques observent plusieurs indicateurs pour juger la performance d'un sportif. Dans le chapitre 4 nous présentons une étude qui traite de la prédiction de performance multidimensionnelle en Handball féminin.



**Figure 3.14.** Performances prédictives par rapport aux niveaux de bruit pour les données imputées avec la méthode d'imputation LOCF :  $K = 10$  :  $q = 4$



**Figure 3.15.** Performances prédictives par rapport aux niveaux de bruit pour les données imputées avec la méthode d'imputation TEKNN :  $K = 1$  :  $q = 4$

## Chapitre

### 4

# Régression multisortie pour la prédiction de performance en handball féminin

## 4.1 Introduction

Dans le précédent chapitre nous avons abordé la problématique de la gestion des valeurs manquantes qui représente un aspect capital du suivi des athlètes durant leurs activités physiques. Durant ces activités, les athlètes subissent des charges (externes et internes) qui impactent directement leurs performances au cours d'un match. Ce chapitre aborde la performance individuelle réalisée en match par des joueuses professionnelles de handball (sport collectif).

La gestion de la charge de travail pour l'optimisation des performances et la minimisation du risque de blessure constitue une activité importante des préparateurs physiques et entraîneurs. Le lien entre charge de travail et performances sportives a été largement étudié en sports individuels [12, 22, 165] et dans une moindre mesure en sports collectifs [23, 14, 22, 20]. Le choix des indicateurs de performance à considérer reste à ce jour un problème ouvert, propre aux sports considérés et aux choix stratégiques et techniques des entraîneurs. Les différentes dimensions de la performance sportive, individuelle ou collective rendent sensible le choix des indicateurs de performance à optimiser surtout lorsque les critères d'évaluation peuvent paraître antagonistes.

Dans le cadre de ces travaux de thèse, nous avons étudié un jeu de données provenant d'une équipe d'élite de handball féminin. Nous avons proposé l'utilisation de 7 indicateurs de performances comprenant 4 indicateurs techniques et 3 indicateurs athlétiques basés sur des données accélérométriques. Nous avons choisi d'inclure des indicateurs de performances techniques, malgré le nombre très limité d'études basées sur ceux-ci [21, 20, 166] car il est reconnu par les staffs sportifs que la performance technique est un élément crucial de la performance individuelle en match, même en sport collectif. Pour répondre à ces problématiques, nous avons choisi d'utiliser les modèles de régression multisorties présentés dans le chapitre 2. Ce type de modèle nous permet principalement de bénéficier de la relation entre les différentes variables de sortie, ici les différents indicateurs de performance tout en divisant les temps de calcul par le nombre de sorties

(*i.e.*, de variables de performance) considérées. La suite de ce chapitre va exposer les détails de ces travaux et les résultats obtenus.

## 4.2 Prédiction de performances en Handball féminin

### 4.2.1 Données et problématique

Dix-huit joueuses d'une équipe d'élite de handball ont été incluses dans ces travaux. En amont de la collecte des données, les joueuses ont donné leur consentement éclairé par écrit et ce travail a été réalisé conformément à la déclaration d'Helsinki <sup>1</sup> sur les sujets humains. L'étude a été approuvée par le département de santé de l'institution (*i.e.*, de l'équipe) où la recherche a été menée. Les données utilisées pour ce travail font partie de la routine de surveillance quotidienne effectuée dans le club. L'équipe a été autorisée à participer à ce travail de recherche et l'approbation du comité d'éthique n'était donc pas requise.

Parmi les joueuses de l'équipe, il y avait 3 défenseuses centrales, 3 défenseuses gauches, 3 ailières gauches, 3 pivots, 2 ailières droit et 2 latérales droit (les gardiennes de but ont été exclues du jeu de données en raison de la nature très spécifique de leurs performances athlétiques et techniques). Ces joueuses ont été équipées de centrales inertielles (IMU) Catapult pendant les entraînements et les matchs afin de mesurer les accélérations et orientations nécessaires au calcul de différents indicateurs de charge de travail (nombre d'accélérations à haute intensité, nombre de sauts à haute intensité,...).

Les vidéos des matchs ont été analysées par des experts de l'équipe afin de comptabiliser des actions spécifiques, offensives (buts, ...) et défensives (duels gagnés, ...) de toutes les joueuses. Le jeu de données comprend 132 sessions d'entraînement et 42 matchs.

#### 4.2.1.1 Indicateurs

Vingt-quatre variables de charge de travail ont été mesurées pendant les entraînements et les matchs. Dix-huit d'entre elles sont liées aux accélérations (*accel*), aux changements de direction à gauche et à droite (*cod\_left*, *cod\_right*), aux décélérations (*decel*), aux sauts (*jump*) et aux impacts (*impact*). Tous ces indicateurs ont été mesurés à trois niveaux d'intensité différents (faible : *low*, modéré : *med*, élevé : *high*), ce qui correspond approximativement à une discrétisation uniforme des espaces observés des indicateurs.

Les cinq autres indicateurs de charge de travail (charge externe) comprennent des indicateurs propriétaires Catapult et le temps de jeu (*playtime*) passé en entraînements

---

<sup>1</sup><https://www.wma.net/fr/policies-post/declaration-dhelsinki-de-lamm-principes-ethiques-applicables-a-la-recherche-medicale-impliquant-des-etres-humains/>

et en matchs. Les indicateurs propriétaires sont dérivés du PlayerLoad (PlayerLoad Totale : TPL, PlayerLoad par minute : PL\_min), qui correspond à la norme de la dérivée de l'accélération et est souvent utilisé comme mesure de la charge de travail totale [62, 63, 64], et du Repeat High Intensity Efforts (RHIE\_tot\_bouts, RHIE\_eff\_recov, RHIE\_eff\_bouts), qui est un indicateur de l'adéquation entre le repos (durée écoulée entre chaque effort de haute intensité) et le niveau de l'effort produit [65, 66].

Cinq facteurs contextuels concernant les matchs et les joueuses ont été considérés, la localisation du match (localisation\_match), la compétition relative au match, la position de la joueuse (player\_role) et l'identité de la joueuse (player\_id).

Vingt-six actions de jeu prédéfinies (buts, tirs manqués, passes bloquées, passes décisives, ...) ont été annotées par les experts de l'équipe pour chaque joueuse. Les experts ont classé chaque action en offensive/défensive et en positive/négative. La liste complète des actions considérées et leurs caractéristiques sont présentées dans le tableau 4.3.

Pour chaque match et chaque joueuse, la performance a été considérée au moyen de 7 labels : le nombre d'accélération à haute intensité (accel\_high), les nombres de changements de direction gauches et droits (cod\_left\_high et cod\_right\_high), les nombres d'actions offensives/défensives positives/négatives (attack\_pos, attack\_neg, defence\_pos et defence\_neg). Les trois premiers labels représentent l'aspect athlétique de la performance, tandis que les quatre derniers correspondent à la dimension technique de la performance. L'ambition de ce travail est de considérer la performance en match comme multidimensionnelle où chaque dimension de la performance est au même niveau afin d'éviter l'isolement ou l'écrasement de labels en raison d'une hiérarchie prédéfinie. Cette multidimensionnalité de la performance repose donc sur l'hypothèse que chacune de ces dimensions a une importance non-négligeable dans la performance globale des joueuses.

#### 4.2.1.2 Protocole expérimental

Les expériences se sont concentrées sur l'exploration des relations entre les charges de travail, les performances en match (athlétiques et techniques) et les facteurs contextuels dans une perspective prédictive.

Les variables de charge de travail, facteurs contextuels et indicateurs de performance pris en compte sont les suivants:

- **charges de travail:** playtime, TPL (Total PlayerLoad), PL\_min (PlayerLoad per minute), accel\_high, accel\_med, accel\_low, decel\_high, decel\_med, decel\_low, cod\_right\_high, cod\_right\_med, cod\_right\_low, cod\_left\_high, cod\_left\_med, cod\_left\_low, jump\_high, jump\_med, jump\_low, impact\_high, impact\_med, impact\_low, RHIE\_tot\_bouts, RHIE\_eff\_recov, RHIE\_eff\_bouts.

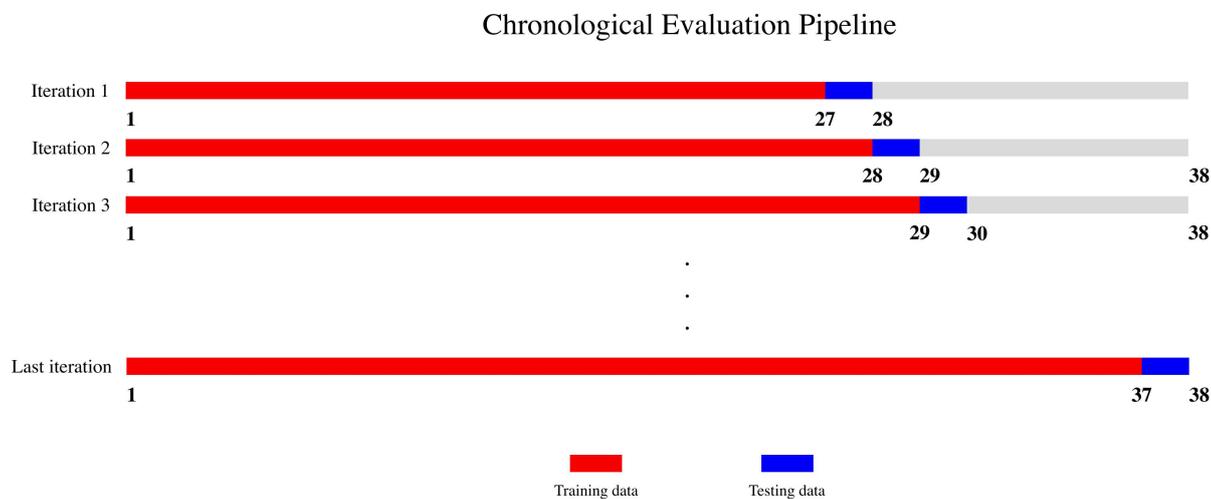
- **facteurs contextuels:** `player_id`, `player_role`, compétition (saison régulière, champion's league, coupe de france, playoffs), `localisation_match` (à domicile, extérieure) et équipe adverse.
- **performances labels:** `accel_high`, `cod_right_high`, `cod_left_high`, `attack_pos`, `attack_neg`, `defence_pos` `defence_neg`.

La majorité des variables de charge de travail correspondent à des comptages (nombre d'accélération, de décélérations, de changements de direction, etc.). Les valeurs passées correspondent aux sommes cumulées calculées sur la dernière semaine ainsi que les 2, 3 et 4 dernières semaines. Pour les indicateurs `playtime`, `TPL`, `RHIE_eff.bouts`, `RHIE_tot.bouts` et `RHIE_eff_recov`, les valeurs sont des moyennes car ce ne sont pas des indicateurs de comptage. Le jeu de données mis en forme contient  $24 \times 4 + 5 = 101$  attributs (variables d'entrée) et 7 labels (variables de sortie) : 3 athlétiques (`accel_high`, `cod_right_high`, `cod_left_high`) et 4 techniques (`attack_pos`, `attack_neg`, `defence_pos` et `defence_neg`).

L'objectif de la tâche prédictive considérée dans ce travail est de prédire les 7 labels (athlétique et technique) en fonction de toutes les charges de travail passées (entraînement et match), des performances techniques passées et des facteurs contextuels (voir Eq (4.1)). Nous avons mis en place un pipeline chronologique où, pour chaque date, les valeurs des 7 labels sont prédites pour chaque joueuse à partir de modèles appris sur les données correspondant aux dates précédentes. Les modèles de régression considérés sont les suivants :  $K$  plus proches voisins ( $KNN$ ), arbres de régression (AR) et forêts aléatoires (FA), ces modèles sont nativement capables de gérer des tâches multisorties avec la bibliothèque python *scikit-learn*. Pour chaque modèle de régression, des régresseurs à sortie unique (...\_MONO) (pour chaque label) et à sorties multiples (...\_MULTI) ont été entraînés et comparés à une baseline naïve qui prédit systématiquement la valeur moyenne de chaque label calculée pour chaque joueuse sur la base de ses valeurs passées. L'évaluation des prédictions a été réalisée en termes de  $aRMSE$  (voir Eq (2.14) et (2.15)).

$$(\text{charges passées, performances passées, contexte}) \longrightarrow \text{performances (7 labels)} \quad (4.1)$$

Les deux premières dates (exemples qui concernent les deux premiers matchs) sont utilisées pour entraîner les modèles à prédire la valeur des 7 labels pour chaque joueuse du 3<sup>ème</sup> match (3<sup>ème</sup> date). Après cette phase initiale, les données des 3 premières dates sont utilisées pour ré-entraîner les modèles à prédire les 7 labels de la 4<sup>ème</sup> date et le processus continue jusqu'à la prédiction de la 42<sup>ème</sup> et dernière date de la saison (voir Figure 4.1).



**Figure 4.1. Évaluation chronologique : À chaque itération, les matchs précédents (en rouge) sont utilisés comme données d'apprentissage pour la prédiction des performances au match suivant (en bleu). La couleur grise représente la partie des matchs suivants qui n'est pas utilisée pour l'apprentissage. Les nombres situés sous les barres chronologiques correspondent aux indices des matchs**

#### 4.2.2 Résultats

Dans cette section nous présentons les résultats obtenus. Nous débutons avec les statistiques descriptives relatives aux différents indicateurs de charges de travail et performances 4.2.2.1. Nous passons après aux résultats de l'analyse de corrélation 4.2.2.2 et en dernier nous abordons les résultats de la tâche prédictive 4.2.2.3.

##### 4.2.2.1 Statistiques descriptives

Le tableau 4.1 décrit les différents indicateurs de charge (moyenne  $\pm$  écart-type) en fonction de leur niveau d'intensité. Le tableau 4.2 décrit les indicateurs propriétaires dans leur unité arbitraire (*au*) et le temps de jeu en minutes. Les tableaux 4.1 et 4.2 concernent toutes les sessions (matchs et entraînements confondus).

Le tableau 4.3 décrit les différentes actions en jeu par leur moyenne  $\pm$  écart-type, leur catégorie (offensive, défensive) et leur polarité (+/-). Le tableau 4.4 présente

Indicateur	intensité faible	intensité modérée	intensité élevée
Acceleration	124.6 ± 49.5	47.4 ± 17.3	19.4 ± 10.2
Deceleration	119.3 ± 44.7	38.1 ± 16.4	8.2 ± 4.7
Jump	2.5 ± 2.3	6.1 ± 4.9	4.3 ± 4.4
Impact	26.6 ± 19.1	2.8 ± 2.8	0.6 ± 1.0
cod_left	661.3 ± 223.6	119.54 ± 44.2	14.5 ± 9.4
cod_right	663.9 ± 226.4	127.5 ± 54.2	15.7 ± 11.5

**Tableau 4.1. Moyenne ( $\pm$  écart-type) comptage de mesures dérivées d'accéléromètre à différents niveaux d'intensité**

Indicateur	Moyenne $\pm$ écart-type
RHIE Total bout (RHIE_tot_bouts)	5.6 ± 4.0
RHIE Effort Recovery - Mean (RHIE_eff_recov)	1.2 ± 1.2
RHIE Efforts Per Bout - Mean (RHIE_eff_bouts)	3.6 ± 1.4
Total PlayerLoad (TPL)	384.6 ± 107.6
PlayerLoad per minute (PL_min)	7.8 ± 1.6
Playtime (min)	44.0 ± 25.9

**Tableau 4.2. Valeurs ( $\pm$  écart-type) des indicateurs propriétaire et du temps de jeu**

le niveau de rareté, *i.e.* le taux de 0 sur les labels de performance. On constate un niveau de rareté élevé pour tous les indicateurs de performances, ceci s'explique pour les performances techniques par le fait que les vidéos de certains matchs n'ont pas été analysées. En ce qui concerne les performances athlétiques, certaines compétitions n'autorisaient pas le port des équipements (centrales inertielles) lors des matchs *e.g.* Coupe de France. Pour adresser cette problématique au niveau de la tâche prédictive, les labels à 0 ne sont pas pris en compte dans l'évaluation des performances prédictives.

La figure 4.2 représente les distributions des labels de performances. Les labels de performances athlétiques (accel\_high, cod\_left\_high et cod\_right\_high) ont des variances et des niveaux moyens plus élevés. Cela s'explique par la nature creuse (*i.e.* beaucoup de 0) des labels de performances techniques, qui ne représentent que les principales actions remarquées par le personnel de l'équipe et contiennent donc principalement des zéros, ce qui peut expliquer le bas niveau des moyennes observées des labels de performances techniques.

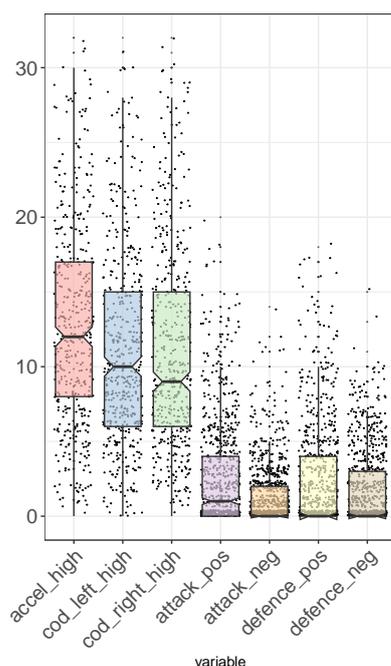
Les figures 4.3 et 4.4 représentent la dynamique des labels de performance durant les premières et deuxièmes parties de saison. La taille des points en arrière-plan est proportionnelle au nombre de joueuses obtenant la même valeur de performance (les couleurs dépendent des indicateurs de performances). Des courbes lissées ont été calculées pour chaque indicateur de performance par le biais d'une régression linéaire locale, ce qui équivaut approximativement à une interpolation temporelle par *KNN*.

Sur les 2 parties de la saison, les indicateurs athlétiques augmentent légèrement au cours de la première moitié de la saison (3 mois), puis diminuent progressivement sur 6

Action	Moyenne $\pm$ écart-type	Catégorie	Polarité
Buts	2.2 $\pm$ 2.0	attack	+
7m	0.3 $\pm$ 0.8	attack	+
7m raté	0.1 $\pm$ 0.2	attack	-
Arrêt gardien	0.8 $\pm$ 1.0	attack	-
Tir raté	0.4 $\pm$ 0.7	attack	-
Tir bloqué	0.1 $\pm$ 0.4	attack	-
2min pr	0.2 $\pm$ 0.5	attack	+
Penalty pr	0.3 $\pm$ 0.6	attack	+
Passe bloquée	0.9 $\pm$ 1.1	attack	-
Passe décisive	0.6 $\pm$ 1.1	attack	+
Avantage gardé	0.8 $\pm$ 1.4	attack	+
Avantage perdu	0.1 $\pm$ 0.5	attack	-
2min concédé	0.2 $\pm$ 0.5	defence	-
7m concédé	0.3 $\pm$ 0.6	defence	-
Duel gagné	1.1 $\pm$ 1.6	defence	+
Duel perdu	1.4 $\pm$ 1.8	defence	-
Neutralisations	0.5 $\pm$ 1.0	defence	+
Glissement	0.1 $\pm$ 0.2	defence	-
Contre +	0.7 $\pm$ 1.3	defence	+
Contre -	0.3 $\pm$ 0.7	defence	-
Couv réussie	0.4 $\pm$ 0.8	defence	+
Couv ratée	0.3 $\pm$ 0.6	defence	-
Aide réussie	0.7 $\pm$ 1.0	defence	+
Aide ratée	0.3 $\pm$ 0.7	defence	-
Montée inv/int réussie	0.6 $\pm$ 1.0	defence	+
Montée inv/int ratée	0.2 $\pm$ 0.4	defence	-

**Tableau 4.3. Moyenne ( $\pm$  écart-type) comptage d'actions en jeu catégorisées en attack / defence et positive (+) / negative (-)**

mois. Cette dynamique met probablement en évidence une augmentation de la performance athlétique obtenue grâce à des sessions d'entraînement dont l'intensité augmente progressivement et montre ses limites à travers les baisses de performances athlétiques observées potentiellement dues à la fatigue accumulée pendant la première moitié de la période (les 2 parties de la saison sont séparées par une période de vacances permettant aux joueuses de se reposer). Les indicateurs de performance technique présentent une tendance à peu près contraire à celle des indicateurs de performance athlétique pendant la première partie de la saison, caractérisée par une diminution suivie d'une augmentation. Cette observation met en évidence un éventuel compromis global entre les performances athlétiques et techniques en début de chaque partie de saison.



**Figure 4.2. Boxplots caractérisant les distributions des labels de performances : Le premier, second (*i.e.* médiane) et troisième quartiles sont respectivement représentés par les boîtes basse, milieu et haute**

Indicateur	Taux de rareté
accel_high	0.40
cod_left_high	0.41
cod_right_high	0.41
attack_pos	0.40
attack_neg	0.52
defence_pos	0.51
defence_neg	0.54

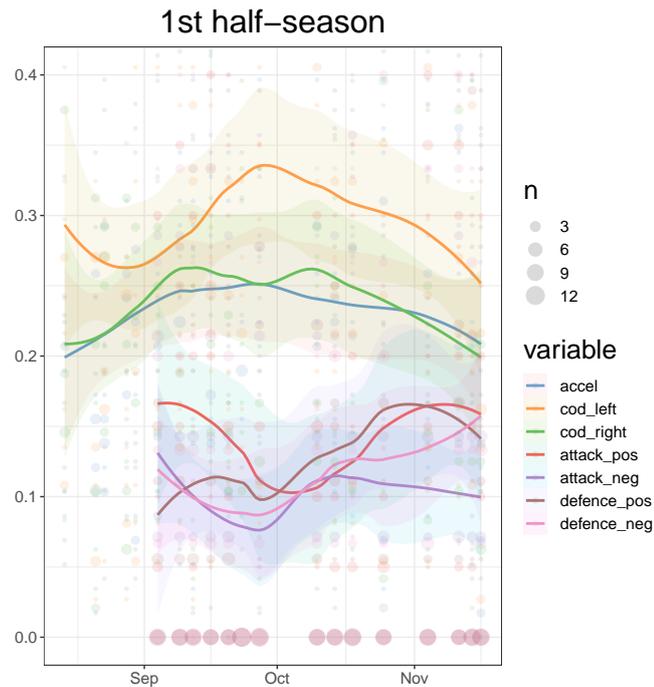
**Tableau 4.4. Taux de rareté des labels de performances (3 athlétiques et 4 techniques)**

#### 4.2.2.2 Analyse de corrélation

Dans cette sous-section, les corrélations ont été analysées entre toutes les variables numériques (charges et performances), les facteurs contextuels ont été exclus étant tous catégoriels.

La figure 4.5 représente les corrélations de Spearman entre les indicateurs de charge de travail et les indicateurs de performance en match sous la forme d'une carte de chaleur où les corrélations négatives sont représentées en bleu et les corrélations positives en rouge. Les corrélations non-significatives (test de Pearson [167]) sont barrées.

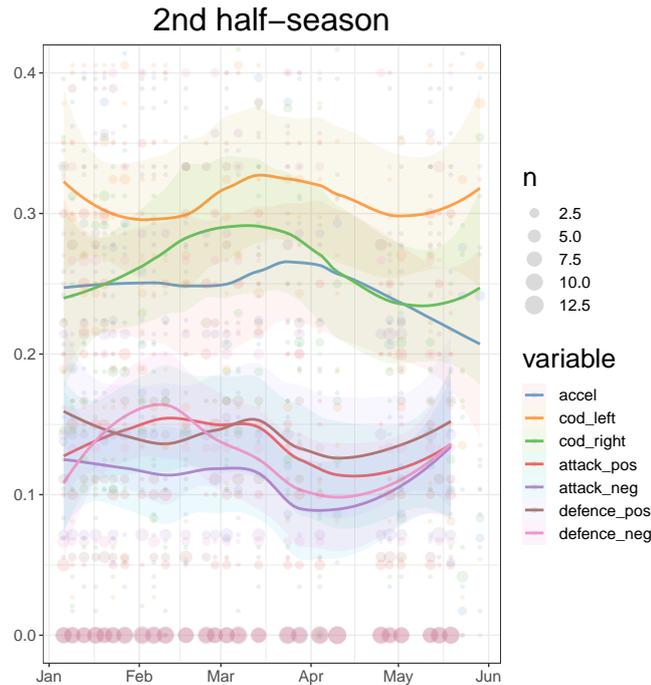
Nous observons que la plupart des corrélations sont positives, ce qui n'est pas surprenant puisque lorsqu'une joueuse joue de manière intensive, elle est à la fois suscep-



**Figure 4.3.** Dynamiques des 7 labels de performances durant la 1<sup>ère</sup> partie de la saison : Chaque point représente la valeur moyenne d'un indicateur de performance (code couleur) pour une joueuse donnée à un match donné. Quand plusieurs joueuses ont les mêmes valeurs de performance (Pour un indicateur donné à un match donné) Les points correspondants sont fusionnés et la taille des points représente le nombre de points fusionnés. Des courbes lissées sont proposées (régression locale) avec un intervalle de confiance à 95%

tible d'avoir une activité athlétique élevée et d'effectuer de nombreuses actions techniques.

L'indicateur "PL\_min" (PlayerLoad par minute) semble être assez fortement et négativement corrélé avec le "field\_time" (Temps de jeu) (-0,33). Cela s'explique par le fait que les joueuses qui disputent de petites périodes de jeu jouent généralement de manière très intensive. Cette corrélation négative significative semble avoir un impact indirect sur de nombreuses autres corrélations : PL\_min est négativement corrélé (de façon modérée) avec la plupart des indicateurs de décélération, de saut, d'impact et de défense. En effet, le "temps de jeu" est positivement corrélé avec la plupart de ces indicateurs, plus une joueuse joue longtemps en match, plus elle a de chance de réaliser d'actions et de travail physique. Ce transfert de corrélation peut également expliquer la non-significativité de la corrélation entre "PL\_min" et près de la moitié des variables, pour lesquelles on s'attendrait naturellement à des corrélations positives : plus une joueuse accélère, plus son PlayerLoad par min devrait être élevé, cependant la corrélation négative issue de la variable "temps de jeu" annule partiellement la corrélation positive *naturelle*.



**Figure 4.4.** Dynamiques des 7 labels de performances durant la 2<sup>ème</sup> partie de la saison : Chaque point représente la valeur moyenne d'un indicateur de performance (code couleur) pour une joueuse donnée à un match donné. Quand plusieurs joueuses ont les mêmes valeurs de performances (pour un indicateur donné à un match donné), les points correspondants sont fusionnés et la taille des points représente le nombre de points fusionnés. Des courbes lissées sont proposées (régression locale) avec un intervalle de confiance à 95%

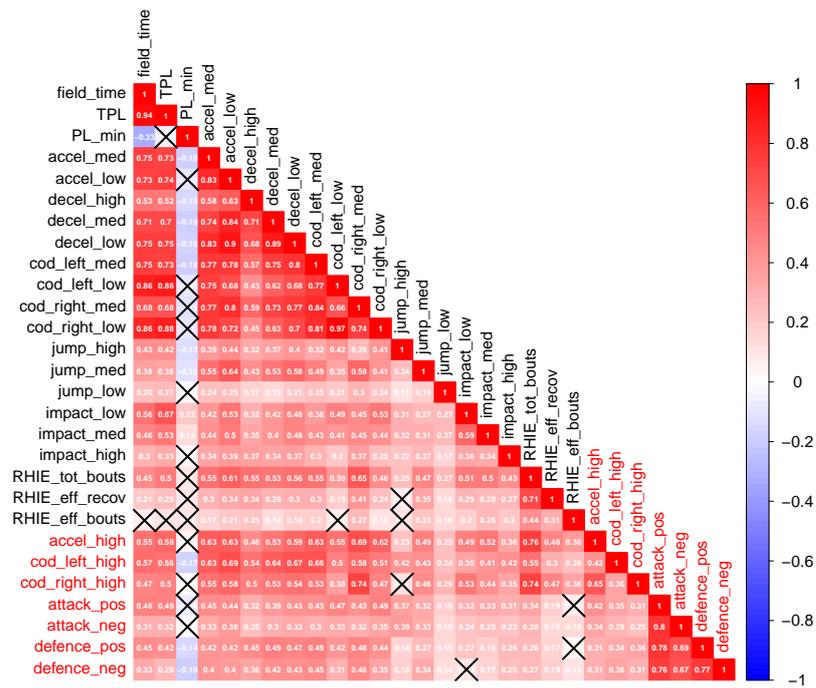
De la même façon, de fortes corrélations ( $\approx [0.5, 1]$ ) sont observées entre 'TPL', les accélérations, les décélérations et les changements de direction, ce qui est conforme à notre interprétation précédente : plus une joueuse est active, plus ses performances athlétiques sont élevées.

Des corrélations modérées mais significatives ( $\approx [0.36, 0.65]$ ) sont observées entre les labels athlétiques en contraste avec les corrélations élevées observées entre les labels techniques ( $\approx [0.67, 0.8]$ ). Si on part du principe qu'une action entraîne une autre, ceci explique les corrélations observées élevées entre les labels techniques ; d'autre part, dans les périodes sans actions, les labels athlétiques continuent eux de varier, ce qui explique également de fortes corrélations.

#### 4.2.2.3 Résultats prédictifs

Cette section présente les performances prédictives en termes d'aRMSE et de temps de calcul de tous les modèles évalués :

- *baseline* naïve qui prédit la moyenne de la joueuse pour chaque label (performance) calculé sur le dernier mois pour le prochain match



**Figure 4.5.** Carte de chaleur des corrélations entre les indicateurs de charge de travail (en noir) et les indicateurs de performance en jeu (en rouge) : Les corrélations négatives sont représentées en bleu et les corrélations positives en rouge, l'intensité des couleurs représentant les valeurs absolues des corrélations. Les corrélations non significatives sont barrées

- régresseurs *KNN* monosortie et multisorties (*KNN\_MONO*, *KNN\_MULTI*)
- arbres de régression monosortie et multisorties (*AR\_MONO*, *AR\_MULTI*)
- régresseurs forêt aléatoire monosortie et multisorties (*FA\_MONO*, *FA\_MULTI*)

Modèle	aRMSE $\pm$ écart-type
baseline naïve	4.803 $\pm$ 2.2
<i>KNN_MONO</i>	5.470 $\pm$ 2.5
<i>KNN_MULTI</i>	5.446 $\pm$ 2.5
<i>AR_MONO</i>	6.204 $\pm$ 2.7
<i>AR_MULTI</i>	6.023 $\pm$ 2.9
<i>FA_MONO</i>	4.785 $\pm$ 2.2
<i>FA_MULTI</i>	4.744 $\pm$ 2.1

**Tableau 4.5.** Performances prédictives issues de l'évaluation chronologique

Nous pouvons voir dans le tableau 4.5 que presque tous les aRMSE des modèles sont plus élevés que l'aRMSE de la baseline naïve, à l'exception de ceux des 2 forêts

aléatoires (FA\_MONO et FA\_MULTI) (4,79 et 4,74), qui sont légèrement inférieurs à l'aRMSE de la baseline naïve (4,80). Cela signifie que les modèles considérés ont un pouvoir prédictif très faible en ce qui concerne ce jeu de données. Cette observation souligne la difficulté de la tâche consistant à prédire les performances en match à partir des données d'entraînement. Cela peut paraître surprenant car l'objectif principal de l'entraînement est justement de maximiser les performances pendant les matchs, mais certaines études ont déjà souligné la difficulté de prédire les performances en match à partir de données de charge de travail [14, 21]. En effet, même si l'entraînement permet généralement d'augmenter les performances sportives, les performances athlétiques et les actions techniques réalisées pendant les matchs ne sont pas nécessairement influencées directement par les charges supportées par les athlètes au cours des sessions d'entraînement et de match précédentes.

Néanmoins, l'approche dite "naïve" (*i.e.* moyenne des performances de chaque joueuse au cours du dernier mois) semble fournir des prédictions assez précises, ce qui implique que les fluctuations des indicateurs de performance considérés autour de leurs moyennes ne sont pas très élevées. Cette observation met en évidence un effet plateau connu des performances au cours d'une saison pour les sportifs de très haut niveau: lorsque des joueuses d'un niveau d'élite sont en bonne forme, leurs performances en match tendent à se stabiliser.

Nous remarquons que pour les 3 modèles prédictifs considérés, la version multi-sorties ("\_MULTI") de chaque modèle est systématiquement plus performante que sa version monosortie ("\_MONO"). Ceci confirme l'intérêt d'utiliser des modèles multi-sorties pour la prédiction d'indicateurs de performance dès lors que les indicateurs considérés sont suffisamment corrélés. En effet, l'information fournie par les corrélations entre les sorties du modèle impose indirectement des contraintes aux sorties et permet d'obtenir au final de meilleurs résultats prédictifs.

Modèle	Temps de calcul (seconds)
KNN_MONO	0.026
KNN_MULTI	0.008
AR_MONO	0.396
AR_MULTI	0.056
FA_MONO	4.750
FA_MULTI	0.818

**Tableau 4.6. Temps de calcul pour tous les modèles**

En outre, cet avantage est renforcé par la réduction considérable des temps de calcul qu'offrent les modèles à sorties multiples. En effet, le tableau 4.6 montre que les temps de calcul moyens des modèles monosortie sont approximativement divisés par 7 pour les modèles à multisorties par rapport à leurs versions monosortie.

### 4.2.3 Discussion et conclusion

L'objectif de ce travail était tout d'abord d'analyser la relation entre différents indicateurs (charges de travail passées, performances passées, facteurs contextuels) et différentes performances (athlétiques et techniques) de joueuses de Handball d'élite au cours d'une saison sportive complète. Le deuxième objectif était d'évaluer les modèles de régression multivariés dans le contexte de labels (indicateurs de performances) corrélés. Pour ce faire, nous avons évalué plusieurs modèles de régression multivariés dans un pipeline chronologique où 7 indicateurs de performances ont été prédits simultanément pour chaque match et chaque joueuse.

Les résultats ont montré que, pour la prédiction des performances en match, la modélisation multivariée a permis une légère augmentation des performances prédictives par rapport à leurs homologues univariés et une réduction substantielle des temps de calcul (voir les tableaux 4.5 et 4.6).

Nous avons également remarqué que la baseline naïve (moyennes de chaque joueuse calculées sur le dernier mois) est plus performante que la plupart des modèles multivariés et univariés. Cette constatation sous-tend l'idée que les charges de travail antérieures ne sont pas des prédicteurs efficaces de ces types conjoints de performances en match (athlétiques et techniques).

En raison de la quantité limitée des données, aucune individualisation des modèles prédictifs n'a été envisagée dans ce travail. Une approche intermédiaire pour les travaux futurs pourrait consister à individualiser les modèles en fonction des facteurs anthropométriques, comme dans [45] et [18]. En perspective, nous recommandons fortement aux chercheurs de futures études sur la prédiction de la performance dans les sports d'équipe d'envisager des approches multivariées, car elles peuvent fournir des informations précieuses aux entraîneurs et aux préparateurs physiques sur les capacités de leurs athlètes et sur la manière dont ils réagissent à l'entraînement à travers l'utilisation conjointe de différents indicateurs de performance.

En outre, il est important de choisir des indicateurs de charge de travail qui sont faciles à interpréter par les entraîneurs/préparateurs physiques et qui résument bien l'activité des athlètes, comme le propose [20]. Dans cette expérimentation, 21 indicateurs de charge de travail ont été utilisés, couvrant uniquement la charge externe. Les indicateurs de charge interne (*e.g.*, fréquence cardiaque, questionnaires subjectifs, etc.) pourraient fournir des informations complémentaires sur la charge de travail qui contribueraient à accroître la capacité prédictive et donc explicative des prochaines performances en match. En outre, il serait utile d'envisager des indicateurs subjectifs tels que le *Rating of Perceived Exertion* ou le *Total Quality of Recovery* pour tenir compte au mieux de l'état de fatigue physique et mental des athlètes (voir [10], [32], [168] et [29]).

La plupart des études qui ont mené une analyse de corrélation entre la charge de travail et les actions en match ont examiné un nombre limité d'indicateurs, comme [21], où la charge de travail n'était représentée que par le temps passé en entraînement et en match. Dans d'autres études récentes, comme [25], des corrélations positives ont été observées entre les charges de travail antérieures et certaines actions en match, et les auteurs de [22] ont observé des corrélations linéaires négatives entre les charges de travail en entraînement hebdomadaire et les performances athlétiques en match. L'utilisation de modèles de régression multivariés pour la prédiction des performances en match à partir de données de charge passée, de performance passée, et de contexte est l'aspect le plus original de ce travail. A notre connaissance, aucune autre étude de ce type n'a été proposée dans la littérature des sciences du sport à ce jour. La prédiction d'indicateurs de performance antagonistes ou négativement corrélés par des modèles multivariés pourrait permettre d'identifier des niveaux de compromis optimaux entre indicateurs.

Les résultats de ce travail ouvrent des perspectives pour l'étude de différents types de performances en match, notamment par l'enrichissement d'autres facteurs que les charges de travail antérieures pour la prédiction des performances en match. Enfin, il est important de garder à l'esprit que nos expériences ont été menées sur un jeu de données collectées dans une équipe de très haut niveau, qui fait partie de l'élite de la discipline. Nos conclusions pourraient ne pas être aussi pertinentes pour une équipe de niveau inférieur.

## Chapitre

### 5

## Applications cliniques en rééducation de patients post-AVC

### 5.1 Introduction

L'accident vasculaire cérébral (AVC) est la deuxième cause de décès et la première cause d'apparition de handicap chez l'adulte dans les pays développés. Les accidents vasculaires cérébraux entraînent fréquemment des limitations physiques sensori-motrices et fonctionnelles responsables de 2 à 4% des dépenses totales en santé [169], ce qui exacerbe le besoin de développer de nouveaux traitements de rééducation [169]. L'hémi-parésie est l'un des handicaps les plus courants causés par l'AVC. La mauvaise récupération de la fonction motrice des membres supérieurs parétiques affecte plus de 80% des individus qui ont survécu à un AVC [170]. Les limitations fonctionnelles du bras persistent le plus souvent 3 mois après l'AVC, même après une rééducation intensive [171]. Bien que l'entraînement moteur répétitif semble efficace pour améliorer la fonction motrice des membres supérieurs parétiques [172], il n'existe pas de protocole spécifique aux patients ayant subi un AVC [173]. L'absence de rééducation entraîne fréquemment la diminution voir l'arrêt de l'utilisation du membre supérieur touché [174], ce qui provoque une réorganisation des réseaux fonctionnels cérébraux [175] et limite la récupération fonctionnelle dépendante de l'utilisation du membre touché [176]. Suite à un AVC, la plasticité neuronale induit des mécanismes de réorganisation/récupération spontanés du système nerveux central [177]. Au fil du temps, la plasticité neuronale est de moins en moins efficace et il est plus difficile de récupérer la fonction motrice du membre supérieur parétique [177]. La récupération réalisée grâce à la plasticité neuronale dépend principalement de l'expérience de l'individu (activité, rééducation,...), qui impacte fortement la récupération fonctionnelle du membre supérieur parétique [178]. Certains aspects du programme de rééducation (*e.g.*, l'intensité, le volume, la variété, la spécificité, la motivation et le biofeedback) peuvent moduler l'effet de la plasticité neuronale sur la récupération [179, 180].

Dans ce contexte, la Thérapie par Réalité Virtuelle (TRV), qui associe des exercices de rééducation à des éléments de jeu, semble être bien adaptée à la récupération fonctionnelle [181]. La TRV peut être définie comme l'utilisation d'une méthode

d'interaction homme-machine ou cerveau-ordinateur [182] pour fournir aux patients des informations sur leur état de santé de manière à guider leur rééducation. Ces interactions se font à travers des stimuli multimodaux sensoriels et moteurs mis en oeuvre dans un espace de travail immersif. Les auteurs de [183] ont proposé un programme de rééducation (ReArm) qui implique la participation des patients post-AVC à des sessions de rééducation en réalité virtuelle. Lors de ces sessions, les patients post-AVC jouent à plusieurs jeux nécessitant l'usage de leur membre supérieur parétique. Les patients post-AVC participent au total à 13 sessions de rééducation en réalité virtuelle. Pour chaque session, les patients post-AVC renseignent des questionnaires en amont et en aval de la session. Ces questionnaires sont relatifs à des ressentis subjectifs relatifs à la fatigue générale, la fatigue du bras, et l'évaluation de la douleur. Les patients post-AVC renseignent aussi le RPE après chaque session. L'échelle de RPE utilisée dans le programme de rééducation ReArm correspond à l'échelle proposée à l'origine par [49] *i.e.* le RPE est sur une échelle de 6 à 20. Une session de rééducation correspond à une succession de jeux auxquels jouent les patients post-AVC durant une vingtaine de minutes. Leurs performances à ces jeux sont généralement évaluées principalement à l'aide de 2 indicateurs, un indicateur de score en pourcentage (plus le score est élevé, meilleure est la performance), et un indicateur de durée (plus la durée passée sur un jeu est faible, plus le patient post-AVC a bien performé).

Un des jeux mis en place dans cette approche intéresse plus particulièrement les thérapeutes pour l'expressivité du niveau de réhabilitation qu'il offre aux patients : *LadyBug*. Ce jeu consiste à attraper successivement des cibles au sein d'un quadrillage en 2D. Pour chaque cible, le patient dispose d'une minute pour l'attraper. S'il l'attrape durant le temps imparti ou si la minute s'écoule sans qu'il l'ait attrapée, la cible disparaît et une autre cible apparaît dans un nouvel endroit du quadrillage. Les patients post-AVC jouent à *LadyBug* au début et à la fin de chaque session de rééducation. Une particularité de ce jeu est qu'il est associé à un indicateur de performance supplémentaire relatif à la qualité des mouvements: le *PathRatio*  $\in [1, +\infty]$  représentant le rapport moyen (calculé sur les 48 cibles) entre le trajet effectué par le patient du curseur à la cible et le trajet optimal (*i.e.* le plus court possible entre les deux), plus ce ratio est proche de 1, plus la trajectoire du mouvement est considérée de "bonne qualité". Ce travail s'inscrit dans le cadre de l'essai clinique ReArm évaluant de nouveaux programmes de rééducation [183]. Nous avons tout d'abord défini les objectifs, d'un point de vue clinique puis d'un point de vue statistique. Les objectifs cliniques qui ont été retenus sont (i) anticiper les réponses des patients post-AVC au programme de rééducation en réalité virtuelle de manière à détecter précocement les patients répondant mal au programme de rééducation et (ii) contourner la nécessité de faire remplir le RPE par les patients post-AVC à la fin de chaque session (gain de temps). Ces objectifs nous amènent à poser deux problématiques statistiques distinctes :

- La prédiction des performances aux prochaines sessions en fonction des performances et niveaux de RPE passés des patients post-AVC.
- La prédiction de RPE en fonction des performances lors des sessions de rééducation.

La première tâche prédictive permettrait aux thérapeutes de détecter les patients pour lesquels l'approche de rééducation n'est pas adaptée et ainsi moduler les exercices de rééducation de ce type de patient, voire de les exclure du programme de rééducation.

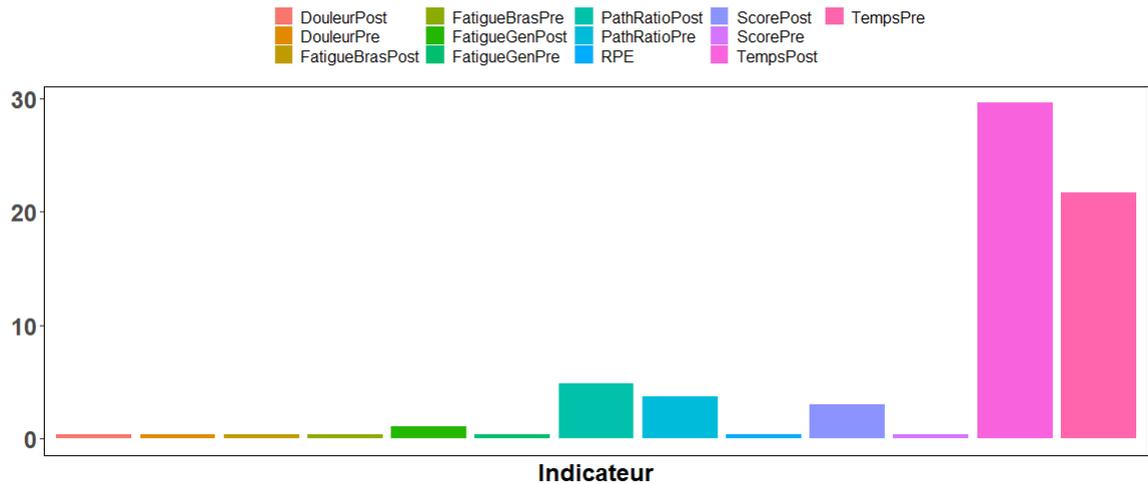
## 5.2 Analyses préliminaires

Dans cette section, nous présentons le jeu de données utilisé pour cette étude et quelques résultats (statistiques descriptives, visualisation) obtenus lors d'une première analyse préliminaire. Le jeu de données, issu du projet ReArm, contient 21 patients post-AVC participant au programme de rééducation ReArm. Parmi ces 21 patients, 17 ont participé à 13 sessions de rééducation, 2 à 12 sessions et 2 à 11 sessions.

Lors de chaque session, les patients post-AVC commencent d'abord par répondre à des questionnaires relatifs à leur état de fatigue générale (FatigueGenPre), à la fatigue du bras (FatigueBrasPre) et à l'évaluation de la douleur (DouleurPre). Après avoir renseigné leurs réponses, les patients jouent au jeu LadyBug à chaque début de session (Pre). Pendant la session, les patients jouent à une dizaine de jeux puis rejouent au jeu LadyBug (Post). Les jeux LadyBug Pre et Post sont évalués par un score en pourcentage (ScorePre, ScorePost), une durée (TempsPre, TempsPost) et les PathRatios (PathRatioPre, PathRatioPost). Après la session de jeu, les patients doivent répondre aux questionnaires sur la fatigue générale (FatigueGenPost), la fatigue du bras (FatigueBrasPost), l'évaluation de la douleur (DouleurPost) et renseignent aussi la difficulté ressentie durant la session (RPE). Pour ces premiers travaux, nous portons notre attention sur les jeux LadyBug (performances) et sur les questionnaires subjectifs (charges ressenties). Les questionnaires (Douleur, FatigueGen, FatigueBras) sont sur des échelles de 0 à 10. Le patient post-AVC donne une note en positionnant un curseur sur une échelle, c'est pour cela que les notes peuvent être des réels

### 5.2.1 Description des données

La figure 5.1 présente les taux de données manquantes par variable. Nous pouvons observer que la plupart des valeurs manquantes sont associées aux variables de performance en durées (TempsPre, TempsPost) qui présentent respectivement 20% et 30% de données manquantes. Ceci s'explique par le fait que le thérapeute collecte manuellement cette information et dispose d'un intervalle de temps très court pour y parvenir (le temps mis pour attraper une cible ne reste affiché qu'1 ou 2 secondes). Les autres



**Figure 5.1. Taux de valeurs manquantes par jeu/questionnaire en %**

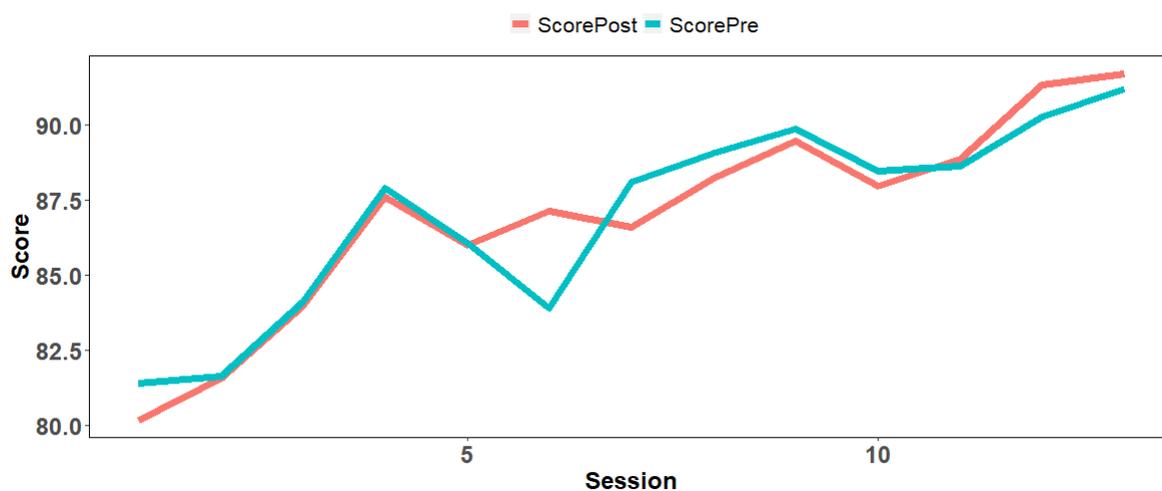
variables présentent toutes moins de 5% de données manquantes, la plupart étant dues au processus de collecte des données qui n'est que partiellement automatisé.

Le tableau 5.1 présente quelques statistiques descriptives relatives aux indicateurs de performance des jeux LadyBug Pre (ScorePre, TempsPre, PathRatioPre) et Post (ScorePost, TempsPost, PathRatioPost), ainsi qu'aux questionnaires Pre (DouleurPre, FatigueGenPre, FatigueBrasPre) et Post (DouleurPost, FatigueGenPost, FatigueBrasPost, RPE). On peut remarquer que les ScorePre et ScorePost ont des médianes (95%) relativement éloignées de leurs moyennes ( $\approx 87\%$ ). Nous observons également que pour beaucoup de variables (toutes sauf TempsPre, TempsPost et le RPE), les moyennes et médianes sont bien plus proches d'un extremum que de l'autre (e.g.  $\min(\text{ScorePre}) = 10\%$ ,  $\max(\text{ScorePre}) = 100\%$ ,  $\text{médiane}(\text{ScorePre}) \approx 95\%$ ,  $\text{mean}(\text{ScorePre}) = 87\%$ ). Ceci indique que pour certains patients (vraisemblablement peu nombreux) ces variables prennent des valeurs très différentes des autres patients. Les profils cliniques et les trajectoires de réhabilitation seront donc analysés de manière à détecter d'éventuels patients présentant des particularités propres (voir Section 5.2.3). En comparant les performances en début ("Pre") et en fin de session ("Post"), on n'observe qu'une très légère amélioration des indicateurs de temps ( $\text{mediane}(\text{TempsPre}) > \text{mediane}(\text{TempsPost})$ ) et de qualité de mouvement ( $\text{mediane}(\text{PathRatioPre}) > \text{mediane}(\text{PathRatioPost})$ ), les scores restant au même niveau (en moyenne) avant et après les sessions de rééducation. Les questionnaires subjectifs (FatigueBrasPre, FatigueBrasPost, FatigueGenPre, FatigueGenPost, DouleurPre et DouleurPost) montrent une légère détérioration générale entre Pre et Post, probablement due à la fatigue (et donc à la charge de rééducation) des sessions.

La figure 5.2 représente l'évolution des scores moyens (ScorePre, ScorePost) calculées sur l'ensemble des 21 patients pour chacune des 13 sessions impliquant le jeu LadyBug. Les patients qui ont une valeur manquante à une session  $i$ , ne sont pas inclus

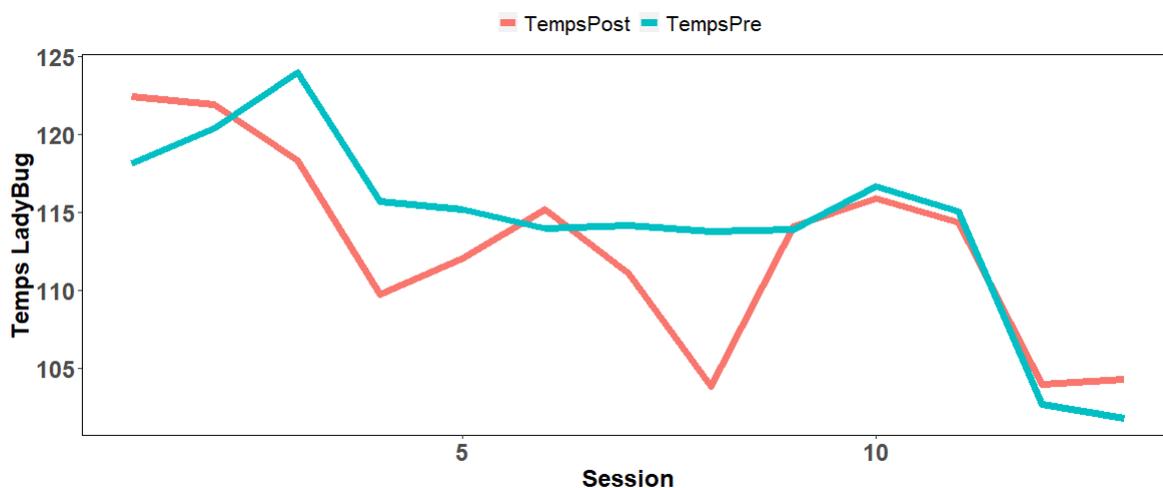
Indicateur	Min	Max	Médiane	Moyenne	Ecart-type
ScorePre	10.00	100.00	95.00	86.88	20.49
TempsPre	63.00	189.00	108.00	114.39	27.51
PathRatioPre	1.15	4.41	1.38	1.47	0.42
ScorePost	12.00	100.00	95.00	86.90	19.33
TempsPost	75.60	187.00	107.50	112.86	27.16
PathRatioPost	1.02	4.38	1.36	1.47	0.45
DouleurPre	0.00	7.00	0.00	1.33	1.88
FatigueGenPre	0.00	10.00	0.50	1.70	2.27
FatigueBrasPre	0.00	9.00	1.15	1.97	2.20
DouleurPost	0.00	8.00	1.00	1.62	1.96
FatigueGenPost	0.00	10.00	2.00	2.36	2.52
FatigueBrasPost	0.00	10.00	2.00	2.63	2.57
RPE	9.00	20.00	13.00	12.81	1.79

**Tableau 5.1. Statistiques descriptives : des indicateurs des jeux LadyBug (Score, Temps, PathRatio) et des questionnaires Pre et Post (RPE, FatigueGen, FatigueBras, Douleur)**

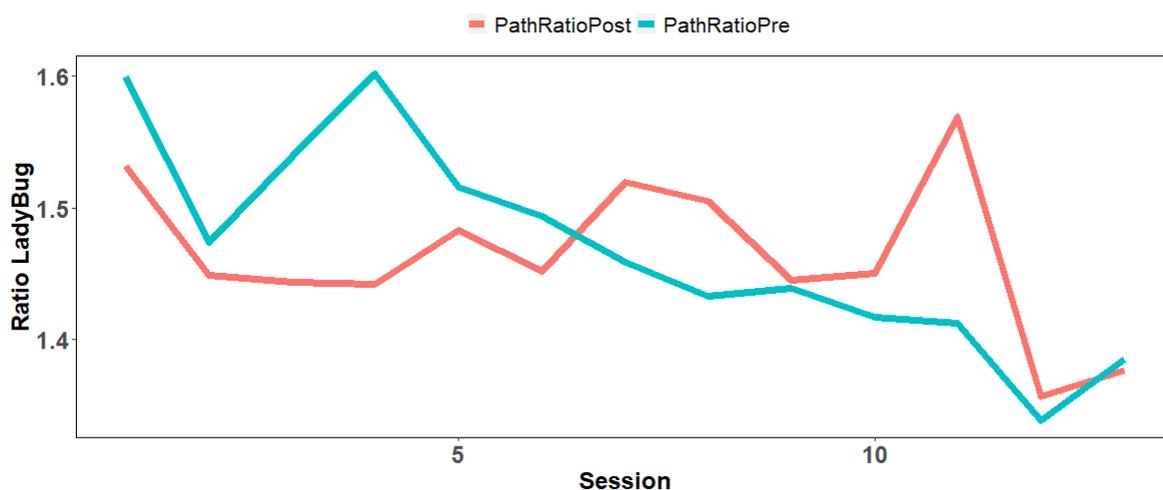


**Figure 5.2. Évolution des scores moyens à LadyBug**

dans le calcul de la moyenne à la session  $i$ . Nous observons clairement une tendance à la hausse, cela peut renforcer l'efficacité du programme de rééducation.

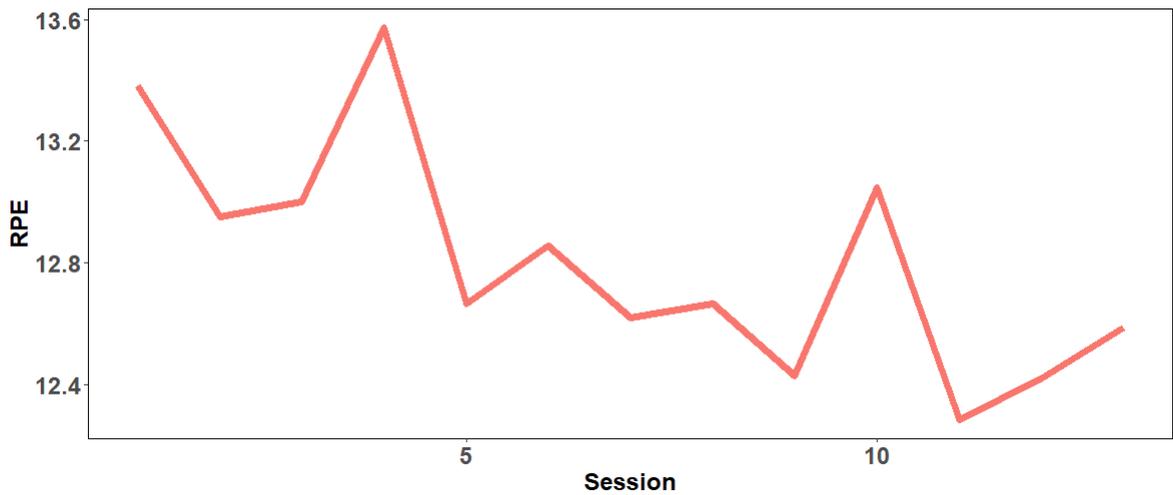


**Figure 5.3.** Évolution des temps de jeu à LadyBug

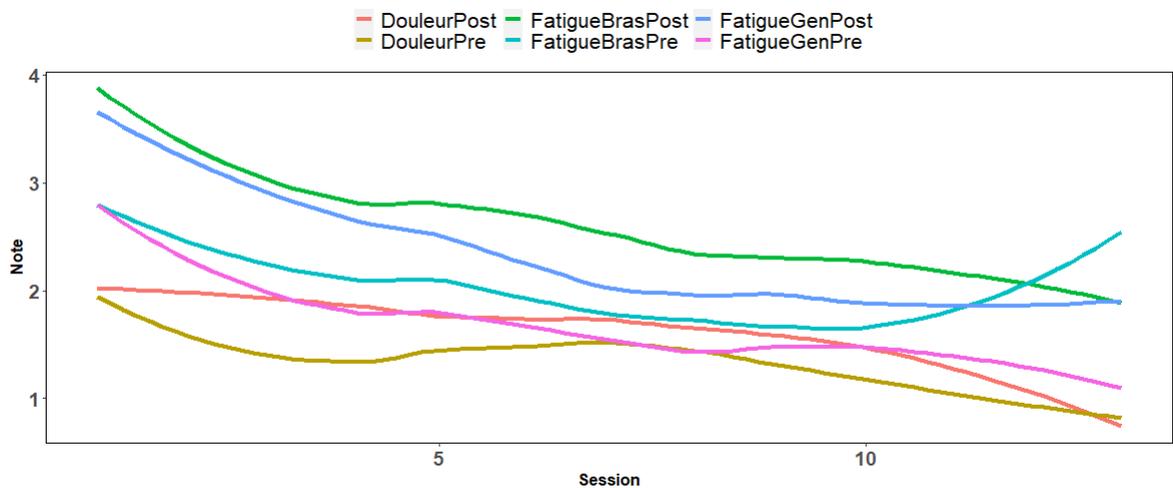


**Figure 5.4.** Évolution des path-ratios à LadyBug

Les figures 5.3 et 5.4 présentent respectivement les évolutions des temps de jeu (TempsPre et TempsPost) et des path-ratios (PathRatioPre, PathRatioPost) (pre et post à chaque fois) du jeu LadyBug. Les temps de jeu présentent une décroissance moins affirmée que la croissance des scores mais avec cependant des temps de jeu ponctuellement très bas (voir 4<sup>ème</sup> et 8<sup>ème</sup> sessions), ce qui illustre là aussi le succès de la thérapie à long terme (décroissance des scores), à moyen terme (remarquables succès à certaines sessions) et à court terme en remarquant que les temps de jeu 'post' sont presque tous inférieurs aux temps de jeu 'pre' (les patients sont donc plus rapides en fin qu'en début de session). Les path-ratios montrent eux-aussi une décroissance (+ faible est le path-ratio, meilleure est la qualité de mouvement) modérée, ce qui souligne la encore le succès de la rééducation. Nous observons aussi que pendant les 6 premières

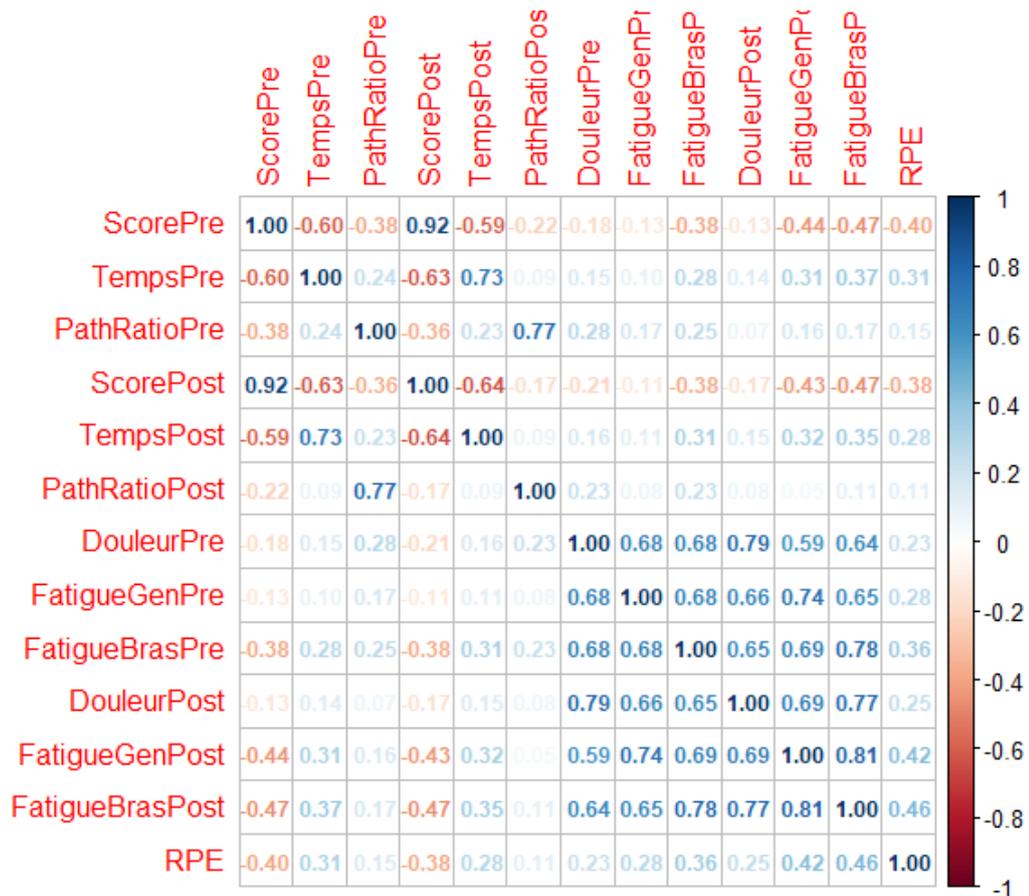


**Figure 5.5.** Évolution RPE



**Figure 5.6.** Évolution notes

sessions, les mouvements sont de meilleure qualité en 'post', donc après la séance de rééducation, qu'en 'pre' ( $\text{PathRatioPre} > \text{PathRatioPost}$ ) mais que ceci s'inverse sur la deuxième moitié de la thérapie. Le fait que durant toute la thérapie les temps de jeu diminuent semble indiquer qu'en milieu de thérapie les patients changent de comportement, durant les premières sessions ils s'améliorent de session en session et au cours de la session alors qu'après la 7<sup>ème</sup> session environ, même si la qualité de leur mouvements s'améliore en 'pre' (le path-ratio continue de baisser) elle ne s'améliore plus en 'post' malgré que leur rapidité augmente (temps de jeu décroissant en 'post' aussi). Ce phénomène pourrait s'expliquer par une fatigue accumulée au fur et à mesure des sessions et/ou par une lassitude, les patients semblent fournir des efforts plus poussés en début qu'en fin de thérapie. Il pourrait être donc pertinent de changer l'ordre voir la nature des jeux au milieu de la thérapie de manière à mieux maintenir le niveau d'implication des patients à leur thérapie.

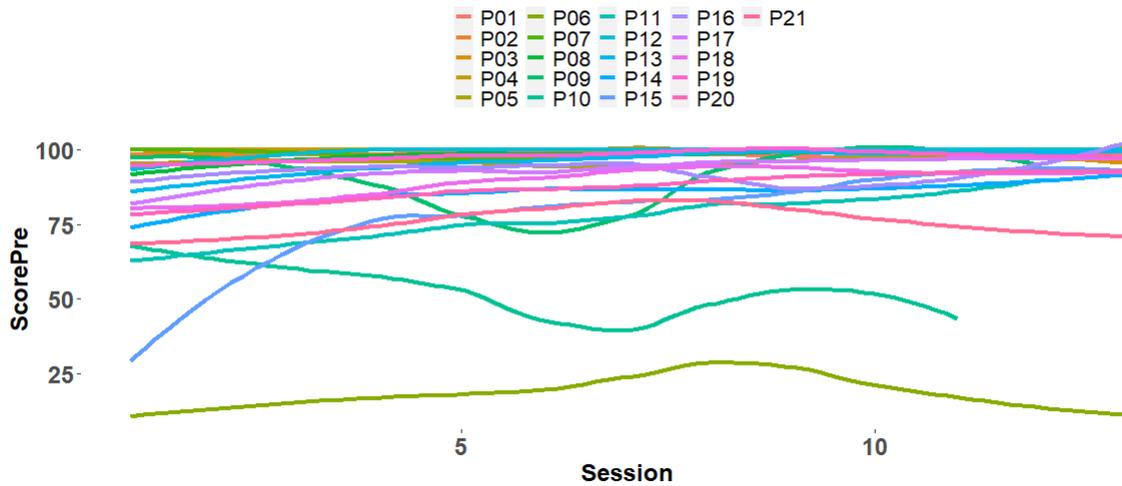


**Figure 5.7. Corrélations de Pearson entre indicateurs de performance et de charge de rééducation au jeu LadyBug**

La figure 5.5 représente l'évolution du RPE moyen (calculé sur l'ensemble des patients). On observe une décroissance relativement constante, au cours des sessions, de la difficulté ressentie par les patients durant leurs sessions de rééducation. La figure 5.6 montre l'évolution des niveaux de fatigue et de douleur moyens renseignés avant (FatigueBrasPre, FatigueGenPre, DouleurPre) et après (FatigueBrasPost, FatigueGenPost, DouleurPost) chaque session. Nous observons une décroissance générale des indicateurs de douleur et de fatigue sauf la FatigueBrasPre qui semble ré-augmenter à partir de la 10<sup>ème</sup> session, probablement du fait de la charge de travail qui semble fatiguer le bras entre les dernières sessions.

### 5.2.2 Analyse de corrélation

Après avoir analysé les évolutions des différentes variables au cours des sessions, les corrélations entre ces variables sont ici calculées et analysées. La figure 5.7 représente les corrélations de Pearson entre les différents indicateurs (performance et questionnaires). Nous constatons tout d'abord des corrélations attendues, notamment les corrélations négatives entre les scores en pourcentage, et les durées



**Figure 5.8. Évolution score LadyBugPre par patient**

(ScorePre/TempsPre, ScorePost/ TempsPost). Aussi nous observons sans surprise que les performances à LadyBugPre et LadyBugPost sont positivement corrélées, *i.e.* plus le score ScorePre est élevé, plus ScorePost le sera, et plus la durée TempsPre sera basse, plus TempsPost le sera. En ce qui concerne les ratios (PathRatioPre, PathRatioPost), il y a des corrélations négatives plus modérées avec les scores en pourcentage (ScorePre, ScorePost), les deux indicateurs de ratio PathRatioPre et PathRatioPost sont positivement liés entre eux, *i.e.* plus le mouvement est qualitatif au jeu Pre, plus il le sera au jeu Post.

Le RPE possède des corrélations négatives avec les scores aux jeux ScorePre et ScorePost ainsi que des corrélations positives avec les performances en durées TempsPre et TempsPost, ce qui semble normal : plus un patient fournit d'efforts, meilleures sont ses performances et plus grand est son état de fatigue en fin de session. Nous constatons aussi que le RPE est positivement corrélé avec les notes aux questionnaires pré- et post-session, ce qui indique que le RPE est très influencé par l'état général du patient au moment de la session, les corrélations avec les notes post-session étant logiques partant du principe que le RPE est une note générale qui englobe la fatigue, la douleur etc.

De manière à évaluer les spécificités des patients nous analysons dans la sous-section suivante les trajectoires de performance et de charge de chaque patient au jeu LadyBug 'pre' (moins de valeurs manquantes qu'en 'post' et moins de dépendance à d'autres variables associées au contenu des sessions qu'en 'post').

### 5.2.3 Trajectoires de performance et de charge de travail du jeu LadyBugPre

Les figures 5.8, 5.9, 5.10 et 5.11 représentent respectivement les trajectoires des performances (score, temps et path-ratio) et de la charge de travail (appréciée par RPE) de

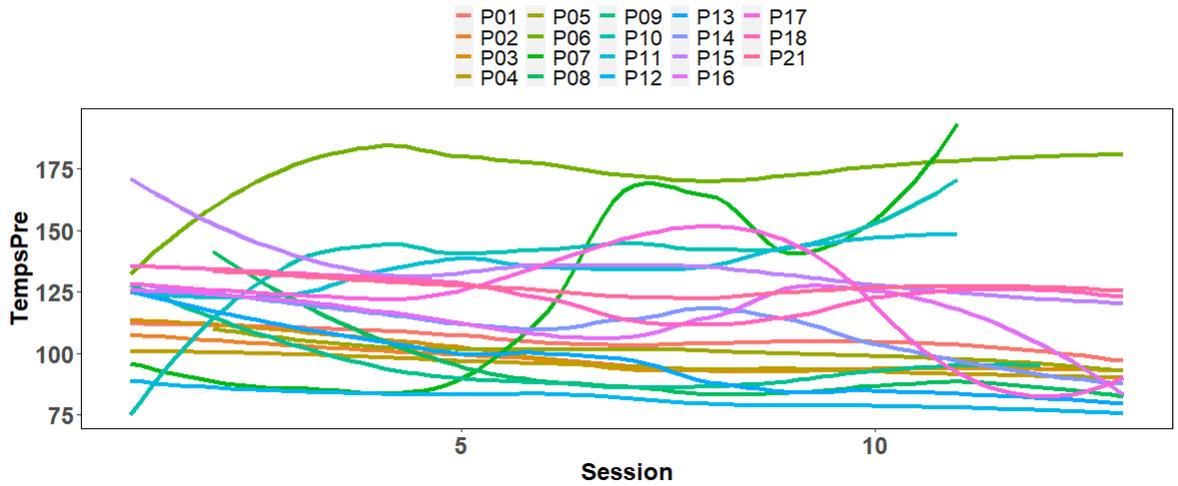


Figure 5.9. Évolution durées LadyBugPre par patient

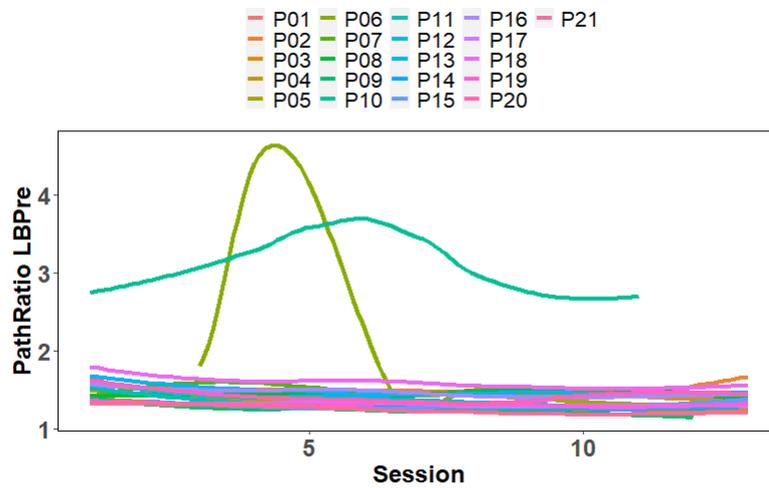


Figure 5.10. Évolution PathRatio LadyBugPre par patient

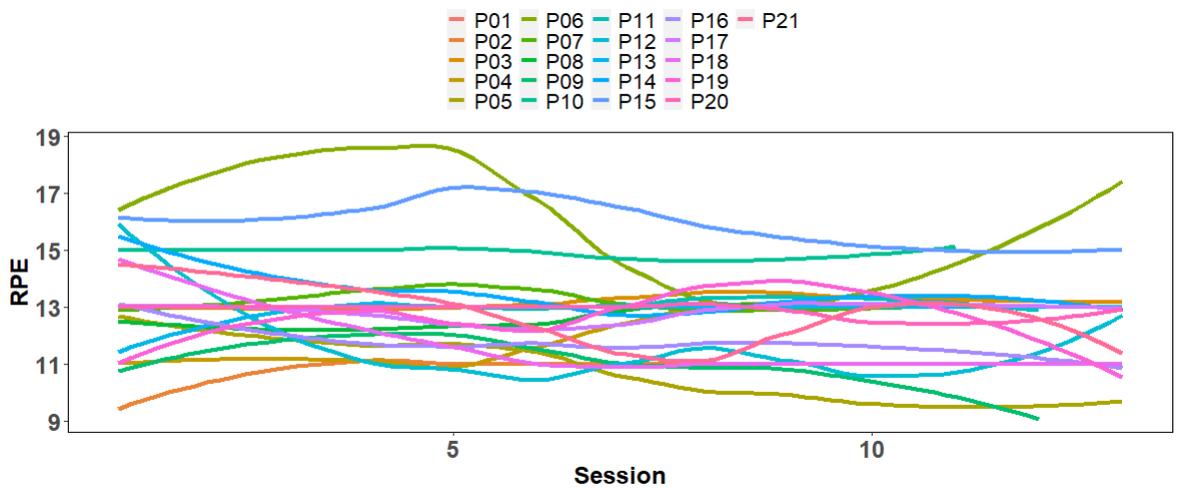


Figure 5.11. Évolution RPE par les patients

chaque patient. Par soucis de lisibilité toutes les courbes ont été lissées par régression polynomiale locale de degré 2. Sur la figure 5.8 on observe pour la plupart des patients une évolution positive au cours des sessions. Cependant, nous observons quelques exceptions tel que le patient P10 qui semble subir une dégradation plutôt qu’une amélioration. Nous remarquons aussi une certaine variabilité dans les scores de départ des patients, ce qui caractérise l’hétérogénéité des niveaux de handicap initiaux des patients.

La figure 5.9 confirme le constat global d’amélioration des performances des patients malgré quelques exceptions, *e.g.* le patient P10 qui ici aussi semble avoir des performances en durée qui se dégradent. La figure 5.10 montre aussi que la plupart des patients post-AVC ont des PathRatioPre relativement constant au cours des sessions, à l’exception des patients ”P10” et ”P06” qui présentent des PathRatios beaucoup plus instables que les autres patients. On remarque également que le patient ”P06” semble avoir la majorité de ses PathRatios manquants.

De manière générale nous constatons donc une amélioration des performances de rééducation pour la plupart des patients. En miroir de ce constat, nous remarquons sur la Figure 5.11 que la fatigue ressentie diminue au cours des sessions pour la plupart des patients (à part le patient P10).

### 5.3 Prédiction du RPE

Dans cette section nous présentons la problématique de prédiction du RPE, l’expérimentation réalisée ainsi que les résultats.

#### 5.3.1 Problématique

Le remplissage du RPE à la fin de chaque session représente du temps et des efforts que les thérapeutes aimeraient pouvoir économiser. L’objectif est donc ici d’évaluer les capacités de prédiction du RPE à partir des données de performance au jeu lors des sessions afin de ne plus avoir à faire renseigner les RPE par les patients post-AVC après leurs sessions de rééducation. Nous avons choisi le RPE comme variable cible (*i.e.* à prédire) car il est positivement corrélé aux autres notes de questionnaire de ressenti post sessions 5.7 ce qui fait de lui un indicateur suffisamment général (la difficulté ressentie est un concept assez large qui peut englober fatigue et douleur). Pour atteindre cet objectif, nous mettons en place un protocole expérimental où les RPE des patients sont prédits en fonction des performances réalisées au jeu LadyBug lors de la session de rééducation considérée, il s’agit donc d’un problème de régression:

$$RPE_{p,t} = f(p, ScorePre_{p,t}, TempsPre_{p,t}, PathRatioPre_{p,t}) + \epsilon \quad (5.1)$$

où  $p$  et  $t$  représentent respectivement l'identité du patient et la date ou session considérée et où  $\epsilon$  représente l'erreur de prédiction ou le bruit.

L'identité du patient est utilisée pour procéder à une individualisation intermédiaire du modèle de prédiction, une individualisation complète impliquant un apprentissage du modèle de régression sur des données restreintes au patient considéré et donc en très faible quantité, serait illusoire dans ce contexte applicatif. Dans un second temps, nous avons voulu évaluer la robustesse de nos modèles de prédiction aux données manquantes, car dans beaucoup d'applications les données de RPE sont souvent incomplètes (remplissage partiellement systématique).

### 5.3.2 Expérimentation

Pour notre expérimentation nous mettons en place une évaluation chronologique, *i.e.* nous fixons une session de départ, à partir de laquelle nous ajouterons à chaque itération 1 session aux données d'apprentissage jusqu'à la dernière. A chaque itération, nous prédisons le RPE des patients post-AVC de la session en cours à partir des données de performance (scores, temps et path-ratios) au jeu LadyBug grâce à des modèles entraînés sur les sessions antérieures. Nous comparons une méthode de référence naïve ("baseline") qui prédit systématiquement la moyenne des RPE du patient considéré calculée sur toutes les sessions antérieures, à deux modèles de prédiction: le  $KNN$  pondéré et l' $EKNN$  régression (voir Eq. (3.7)).

Pour évaluer la robustesse des modèles prédictifs aux données manquantes nous procédons à un bruitage artificiel et aléatoire correspondant à différents niveaux de bruit comme cela a été fait dans le chapitre 2 sur le jeu de données COVID-19. Les différents niveaux de bruits considérés sont  $\{0, 0.1, \dots, 0.6\}$ . Pour chaque niveau de bruit  $v$ , nous retirons aléatoirement une proportion  $v$  de données de performance des données d'apprentissage puis les imputons avec une méthode d'imputation crédibiliste (*i.e.* qui attribue un degré de certitude à la valeur imputée) à chaque itération de l'évaluation chronologique avant l'apprentissage des modèles, l' $EKNN$  peut ainsi utiliser la connaissance de l'incertitude relative à l'imputation à travers le modèle d'incertitude (3.7). Il est à noter que seules les données de RPE (*i.e.* de label ici) ont été bruités. En effet l' $EKNN$  est adapté aux données de labels incertaines mais, dans leur versions initiales, ni le  $KNN$ , ni l' $EKNN$  ne peuvent gérer les données d'attributs incertaines ou manquantes. Pour cette raison, sur les 260 exemples du jeu de données, les 60 exemples initialement incomplets ont été retirés en amont de l'expérimentation qui est ainsi restreinte aux 200 exemples restants. Cette expérimentation est reproduite deux fois, une fois avec la méthode d'imputation crédibiliste par moyenne et une fois la méthode d'imputation crédibiliste LOCF (voir Section 3.3.4).

### 5.3.3 Imputation crédibiliste

#### - Imputation par moyenne

La méthode d'imputation par moyenne consiste à imputer toutes les données manquantes par la moyenne de la variable considérée et calculée sur les données connues donc passées ici (voir Eq. (2.24)). Le modèle d'incertitude crédibiliste associé aux données imputées par moyenne est basé sur l'erreur d'imputation qui est calculée en termes de RMSE (voir Eq. (2.14)) sur les données de RPE connues.

$$RMSE^{norm}(RPE) = \frac{RMSE(RPE)}{\max_{i \in I} \left( \widehat{RPE}_i - RPE_i \right)^2} \in [0, 1] \quad (5.2)$$

où  $I$  = l'ensemble des exemples complets.

Ce RMSE est ensuite normalisé de manière à prendre une valeur dans  $[0, 1]$  (voir Eq. 5.2). Le modèle d'incertitude associé aux données imputées par moyenne est le suivant:

$$\forall i \in J, \begin{cases} m_i(\{\widehat{RPE}_i\}) & = 1 - RMSE^{norm}(RPE) \\ m_i(\Omega^{RPE}) & = 1 - m_i(\{\widehat{RPE}_i\}) \end{cases} \quad (5.3)$$

où  $J$  = l'ensemble des exemples incomplets.

Pour le RPE utilisé dans le programme ReArm, le cadre de discernement  $\Omega^{RPE} = \{0, \dots, 20\}$  Plus l'incertitude d'imputation est grande, *i.e.* plus le  $RMSE^{norm}$  est proche de 1, plus la masse  $m_i(\{\widehat{RPE}_i\})$  attribuée à la valeur imputée est proche de 0, plus donc la valeur imputée sera considérée incertaine par le modèle prédictif.

#### - Imputation par LOCF

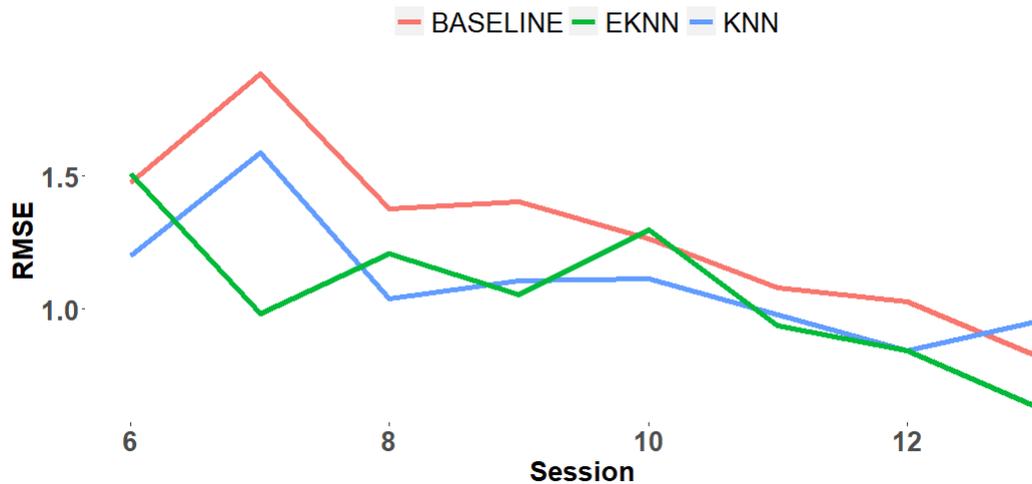
La méthode d'imputation LOCF utilisée ici se distingue de celle utilisée lors de l'application COVID-19 sur deux points. Globalement, nous imputons les valeurs manquantes d'un patient uniquement à partir de ses données passées et le modèle d'incertitude Eq. (3.12). Cependant, quand nous n'avons pas de données passées disponibles, nous considérons la valeur manquante comme une ignorance totale. La fonction de masse attribuée au label manquant est alors:

$$\forall i \in K, m_i(\Omega^{RPE}) = 1 \quad (5.4)$$

avec  $K$  = l'ensemble des exemples incomplets pour lesquels aucune valeur passée de RPE n'est connue ( $K \subset J$ ).

### 5.3.4 Résultats

L'hyperparamètre  $K$  pour le  $KNN$  et l' $EKNN$  a été fixé à 5. Pour obtenir cette valeur, nous avons comparé empiriquement les résultats obtenus en termes de RMSE pour des



**Figure 5.12. Évaluation chronologique (jeu de données non bruité)**

baseline	$KNN$	EKNN régression
$1.29 \pm 0.32$	$1.10 \pm 0.22$	$1.06 \pm 0.27$

**Tableau 5.2. RMSE moyen  $\pm$  écart-type par approche pour la prédiction du RPE (jeu de données non bruité)**

valeurs de  $K \in 1, \dots, 10$  sur le jeu de données non bruité.

La session de départ pour l'évaluation chronologique a été fixée à la 5<sup>ème</sup> session. De manière à analyser les performances prédictives initiales (*i.e.* avant bruitage) des modèles d'un point de vue chronologique et au global, une première expérience est ici réalisée sans bruitage (*i.e.* sur les données complètes). La figure 5.12 représente l'évaluation chronologique qui compare les 3 approches prédictives (baseline,  $KNN$  pondéré, EKNN) pour un niveau de bruit 0 *i.e.* sans valeurs manquantes (donc sans besoin d'imputation). Nous observons que les 2 approches KNN et EKNN ont des performances sensiblement similaires, et de meilleures performances que l'approche baseline naïve. Le tableau 5.2 présente les performances prédictives obtenues en termes de RMSE sur le jeu de données non bruité par chaque approche prédictive. L'approche EKNN fait une erreur moyenne de 1,06, le KNN 1,10, et l'approche naïve 1,29. Ces résultats rejoignent les résultats de [121] qui montrent que l'EKNN donne des performances prédictives légèrement supérieures au  $KNN$  pondéré en classification. Cet avantage viendrait de la fusion conjonctive par la règle de Dempster qui permettrait une agrégation légèrement plus fine de l'information des différents voisins qu'un simple vote (même pondéré).

Dans un second temps, le pipeline d'évaluation chronologique décrit précédemment est mis en oeuvre avec un niveau de bruitage progressif (de 0 à 0,6) pour les 2 méthodes d'imputation considérées (*i.e.* l'imputation par moyenne et par LOCF). La figure 5.13 représente les résultats obtenus pour les évaluations chronologiques où les données

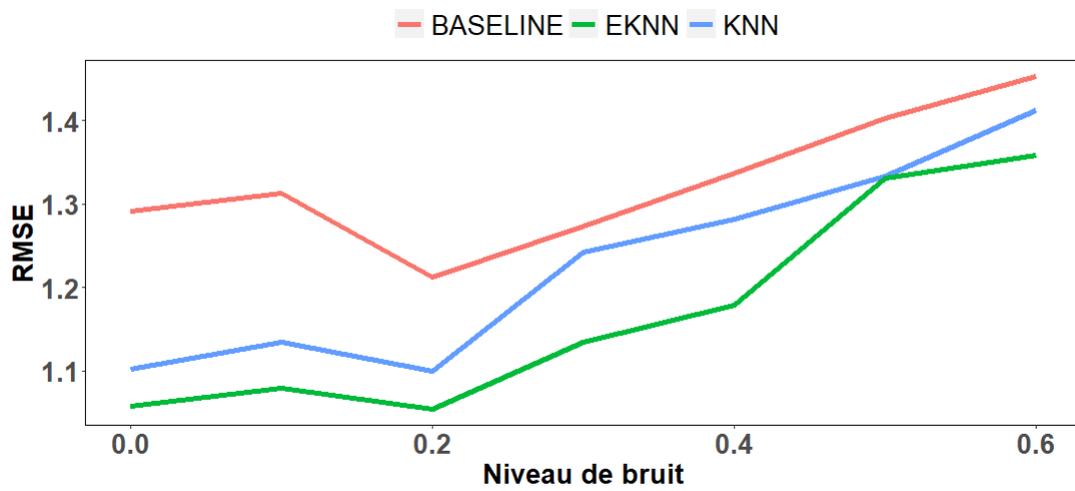


Figure 5.13. Résultats prédictifs pour l'imputation par moyenne

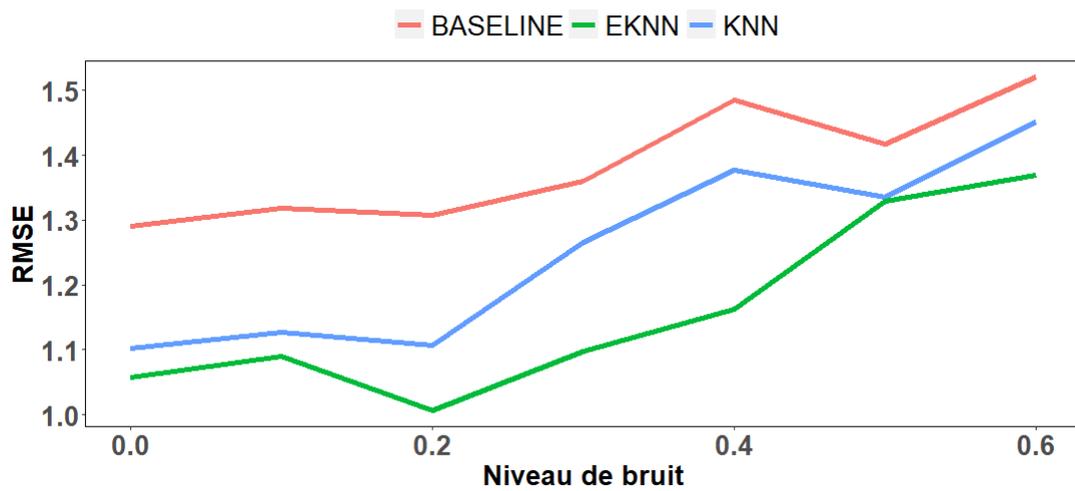


Figure 5.14. Résultats prédictifs pour l'imputation par la méthode LOCF

manquantes ont été imputées par moyenne avec le modèle d'incertitude (5.3) pour chaque niveau de bruit. Les erreurs de prédiction du  $KNN$  et de l' $EKNN$  apparaissent toutes deux inférieures aux erreurs du modèle "baseline" pour tous les niveaux de bruits. Cette observation met en évidence le pouvoir prédictif du  $KNN$  et de l' $EKNN$  sur ce jeu de données. Nous constatons aussi sur cette figure que les erreurs des 3 modèles prédictifs augmentent avec le niveau de bruit. Partant du principe qu'en retirant (même aléatoirement) des données d'apprentissage on diminue l'information disponible pour les modèles prédictifs et ainsi leur capacité de généralisation, ce constat était attendu. On constate aussi que l' $EKNN$  montre de meilleures performances que les 2 autres modèles prédictifs quelque soit le niveau de bruit. Cet avantage prédictif s'explique probablement par la fusion conjonctive (règle de Dempster plus informative que le vote) et par le modèle d'incertitude (5.3) qui permet à l' $EKNN$  d'utiliser la connaissance de l'incertitude associée aux données imputées (via les performances de la méthode d'imputation considéré, *i.e.* l'imputation par moyenne ici).

La figure 5.14 représente les résultats obtenus pour les évaluations chronologiques où les données manquantes ont été imputées par la méthode LOCF crédibiliste (voir Section 3.3.4) avec le modèle d'incertitude (5.3) pour chaque niveau de bruit. On peut observer que la méthode d'imputation crédibiliste LOCF donne des résultats très proches de l'imputation par moyenne. On peut toutefois observer que ponctuellement (pour des niveaux de bruits de 0,3 et 0,4) l'imputation par moyenne donne des résultats légèrement meilleurs que l'imputation par LOCF, ce qui rejoint les conclusions de [184] qui met en évidence l'avantage (voir l'optimalité) de l'imputation par moyenne dans un contexte prédictif précis (*i.e.* sans modèle d'incertitude associé aux données imputées). Ici encore, l' $EKNN$  obtient les meilleures performances prédictives.

#### 5.4 Prédiction de performances aux prochaines sessions

Pour répondre à l'objectif principal relatif aux réponses des patients post-AVC au programme de rééducation ReArm pour la détection précoce de patients non-répondants à la rééducation, nous avons décidé d'estimer des *courbes prévisionnelles* en emboîtant des prédictions de modèles prédisant les réponses (*i.e.* performances) des patients à la prochaine session. Une courbe prévisionnelle consiste en une chaîne de prédictions à partir d'une session donnée, jusqu'à la dernière session de rééducation du patient post-AVC.

Nous comparons 5 méthodes de prédiction :

- modèle de référence naïf ('baseline') : prédiction systématique de la moyenne de chaque indicateur de performance (ScorePre, TempsPre, PathRatioPre) calculées sur les sessions passées du patient considéré.

- arbres de régression monosorties (AR-MONO) et multisorties (AR-MULTI)
- forêts aléatoires monosorties (FA-MONO) et multisorties (FA-MULTI)

Les entrées des modèles AR et FA (MONO et MULTI) sont les performances (ScorePre, TempsPre, PathRatioPre) à la session précédentes et le RPE à la session précédente  $t-1$  en plus de l'identité du patient  $p$  :

$$(ScorePre_{t,p}, TempsPre_{t,p}, PathRatioPre_{t,p}) = f(ScorePre_{t-1,p}, TempsPre_{t-1,p}, PathRatioPre_{t-1,p}, RPE_{t-1,p}, p) + \epsilon \quad (5.5)$$

4 patients ont été exclus de cette expérimentation car ils avaient un trop grand nombre de données de TempsPre manquantes. En ce qui concerne les valeurs manquantes restantes, elles sont imputées par la dernière valeur observée du patient considéré. Si aucune valeur antérieure n'est disponible, la valeur manquante est imputée par la moyenne des autres patients durant la session considérée. Nous obtenons un total de 17 patients. Certains ont fait 11 ( $n = 2$ ) ou 12 ( $n = 2$ ) sessions de rééducation, la majorité a réalisé 13 sessions de rééducation ( $n = 13$ ).

#### 5.4.1 Expérimentation

Pour une meilleure compréhension, nous expliquons l'expérimentation à travers un exemple. Admettons que nous voulons générer une courbe prévisionnelle pour le patient post-AVC "P01". Nous choisissons de générer la courbe prévisionnelle à partir de la 5<sup>ème</sup> session de rééducation patient post-AVC "P01". Les modèles de régression sont entraînés sur toute les sessions des autres patients post-AVC en plus des sessions connues du patient "P01".

Pour la première itération de l'évaluation chronologique, les performances (ScorePre, TempsPre, PathRatioPre) et le RPE à la session précédente du patient post-AVC "P01" sont connus. Les modèles de régression AR et FA monosorties et multisorties utilisent ces données connues en entrée *i.e.* les performances (ScorePre, TempsPre, PathRatioPre) et le RPE du patient "P01" à la 4<sup>ème</sup> sessions sont utilisés en entrées des modèles de régression. À partir de la deuxième itération, les modèles de régression AR et FA vont utiliser en entrée les performances et RPE prédits à l'itération précédente.

Pour l'imputation des données de RPE qui seront forcément manquantes pour les sessions futures (nécessaires à l'emboîtement des prédictions), une approche par *K*NN pondéré indépendante des modèles AR et FA, qui prédisent eux les 3 indicateurs de performance, a été utilisée. Le *K*NN est entraîné sur toutes les sessions des autres patients post-AVC, et les sessions connues du patient post-AVC "P01" *i.e.* ses 4 premières sessions. La prédiction du RPE se fait en fonction des performances au jeu LadyBug Pre ainsi que de l'identité du patient:

$$RPE_{t,p} = f(\text{ScorePre}_{t,p}, \text{TempsPre}_{t,p}, \text{PathRatioPre}_{t,p}, p) + \epsilon \quad (5.6)$$

Dès la première itération de l'évaluation (*i.e.* 5<sup>ème</sup> session), le RPE sera prédit en fonction des performances au jeu LadyBug Pre prédites par les modèles de régression. Nous utilisons ensuite le RPE prédit en entrée des modèles prédisant les 3 indicateurs de performance pour la 6<sup>ème</sup> session. Le protocole d'évaluation se poursuit ainsi jusqu'à la prédiction de la dernière session de rééducation du patient. Pour le calcul de l'erreur de prédiction à chaque itération lors de l'évaluation chronologique, nous calculons le aRMSE (Eq. 2.15) avec les résultats obtenus sur l'ensemble des patients qui ont participé à la session concernée par l'itération. Nous calculons également un aRMSE moyen pour chaque approche relativement à toute l'évaluation chronologique *i.e.* calculé sur toutes les itérations de l'évaluation chronologique. Les indicateurs ScorePre et TempsPre ont été ramenés à des échelles plus similaires à celles de PathRatioPre (*i.e.*  $\approx [1, 3]$ ). L'indicateur TempsPre qui à la base est en secondes a été transformé en minutes (moyenne  $1,907 \pm 0,459$ ). Le score ScorePre qui est un pourcentage de 0 à 100 est exprimé par une échelle de 0 à 1 (moyenne  $0,869 \pm 0,205$ ).

#### 5.4.2 Résultats

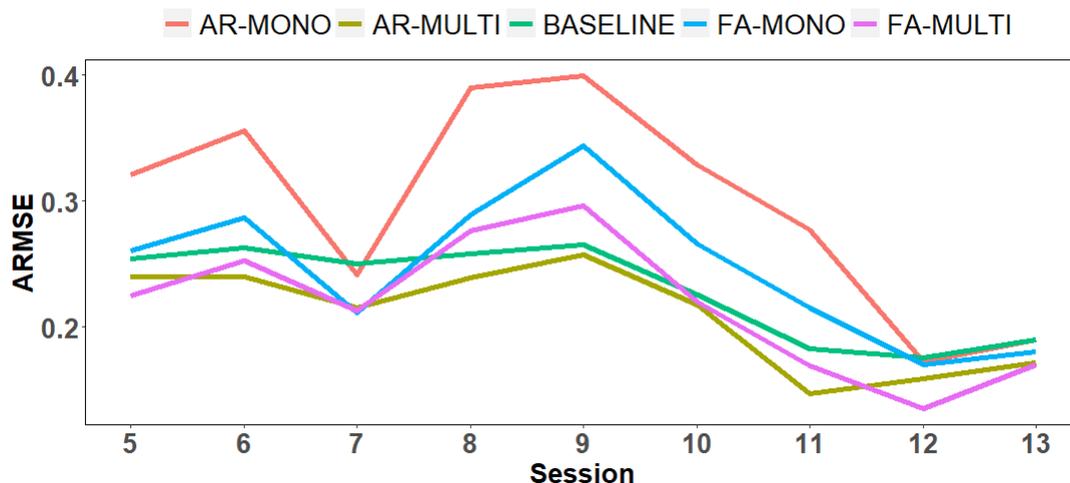
baseline	AR-MONO	FA-MONO	AR-MULTI	FA-MULTI
$0.229 \pm 0.037$	$0.297 \pm 0.083$	$0.247 \pm 0.057$	<b><math>0.206 \pm 0.040</math></b>	$0.217 \pm 0.053$

**Tableau 5.3. aRMSE moyen  $\pm$  écart-type par approche pour les prédictions emboîtées de performances de rééducation**

Les hyperparamètres des modèles AR et FA monosorties et multisorties ont été obtenus grâce à l'algorithme d'optimisation *GridSearch* [185]. L'algorithme *GridSearch* consiste à tester toutes les combinaisons possibles dans une grille de valeurs d'hyperparamètres sélectionnées en amont. L'algorithme a été appliqué aux données qui correspondent à toutes les sessions des patients jusqu'à la session de départ de l'évaluation. Pour le modèle de prédiction du RPE *KNN*, l'hyperparamètre *K* avait la valeur 5 comme pour l'expérimentation précédente.

Le tableau 5.3 présente les résultats obtenus par les différents modèles prédictifs en termes de aRMSE moyen  $\pm$  écart-type. Nous constatons que les 2 approches monosorties (AR-MONO, FA-MONO) obtiennent des performances prédictives inférieures à celle de la méthode de référence 'baseline'. Le meilleur résultat (*i.e.* plus bas aRMSE) est obtenu par l'arbre de décision multisortie (AR-MULTI).

La figure 5.15 expose les résultats de l'évaluation chronologique en termes de aRMSE pour chaque session. Il est à retenir que la première itération est la seule où sont utilisées des entrées réelles, *i.e.* les entrées des modèles de régression sont les



**Figure 5.15. Évaluation chronologique par modèle prédictif**

véritables performances et RPE à la session précédente. Dès la deuxième itération, les prédictions faites à l'itération précédente sont utilisées en entrées des modèles de régression en tant que performances et RPE précédents. Nous observons sur la figure 5.15 que le modèle AR-MULTI obtient les meilleurs eRMSE (*i.e.* les plus bas) pour chaque session à part la 12<sup>ème</sup>. Le modèle FA-MULTI (2<sup>ème</sup> meilleur modèle d'après le tableau 5.3) obtient des résultats proches mais légèrement moins bons pour chaque session. Nous observons que les modèles multisorties (FA-MULTI, AR-MULTI) obtiennent tous deux de meilleures performances prédictives que leurs homologues monosorties (FA-MONO, AR-MONO). La méthode de référence 'baseline' suit en termes de ARMSE une trajectoire proche des modèles AR-MULTI et FA-MULTI tout en ayant toujours des performances prédictives inférieures à celles du modèle AR-MULTI. Nous constatons aussi que l'erreur prédictive de tous les modèles diminue globalement au fur et à mesure de l'évaluation chronologique à part un pic aux 8<sup>ème</sup> et 9<sup>ème</sup> sessions surtout pour les modèles monosorties. D'un côté, ceci peut sembler paradoxal car cela voudrait dire que les modèles de régression monosorties et multisorties réalisent de meilleures prédictions en fonction des performances et RPE prédits plutôt qu'en fonction des valeurs réelles à la première itération. D'un autre côté cela est éventuellement rendu possible par l'augmentation de la taille des données d'apprentissage au cours de l'évaluation chronologique. Ce constat tend à vouloir dire que la taille de données d'apprentissage est un critère plus important que l'incertitude qui est associée (et qui grandit naturellement au fur et à mesure des sessions).

Le tableau 5.4 nous montre les performances prédictives moyennes de chaque modèle prédictif pour chaque indicateur de performance de rééducation. Nous constatons que même si l'arbre de régression multisortie (AR-MULTI) a les meilleures performances globales en termes de aRMSE, les résultats sont plus diffus quand on considère les indicateurs de performance individuellement. Le modèle AR-MULTI sur-

Indicateur	baseline	AR-MONO	FA-MONO	AR-MULTI	FA-MULTI
ScorePre	0.111 ± 0.026	0.134 ± 0.042	0.095 ± 0.037	<b>0.088 ± 0.034</b>	0.092 ± 0.036
TempsPre	0.425 ± 0.080	0.418 ± 0.099	<b>0.363 ± 0.100</b>	0.370 ± 0.094	0.370 ± 0.107
PathRatioPre	<b>0.151 ± 0.067</b>	0.340 ± 0.163	0.282 ± 0.112	0.169 ± 0.063	0.190 ± 0.101

**Tableau 5.4. RMSE moyen ± écart-type par modèle prédictif et par indicateur de performance de rééducation**

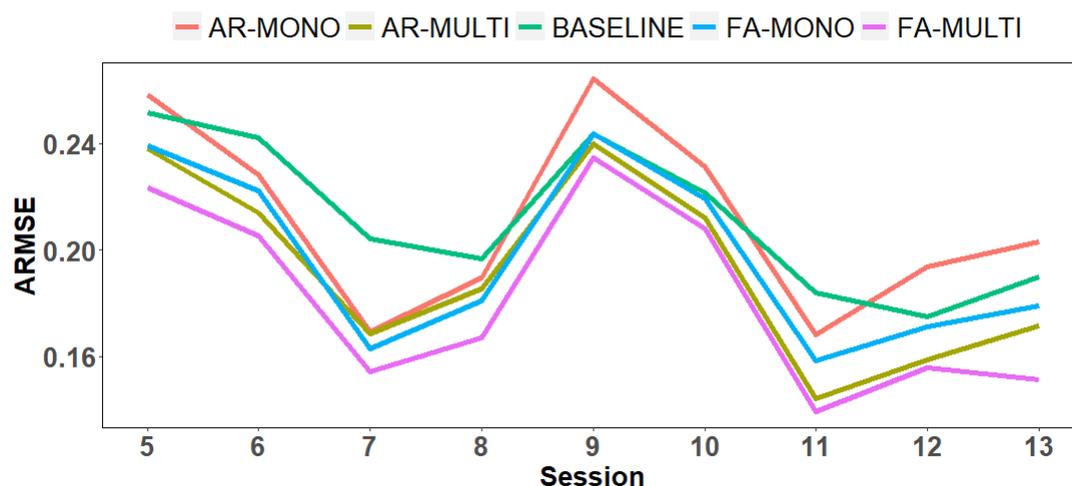
baseline	AR-MONO	FA-MONO	AR-MULTI	FA-MULTI
0.212 ± 0.028	0.211 ± 0.036	0.198 ± 0.034	0.193 ± 0.035	<b>0.182 ± 0.036</b>

**Tableau 5.5. aRMSE moyen ± écart-type par modèle prédictif après suppression des données du patient "P10"**

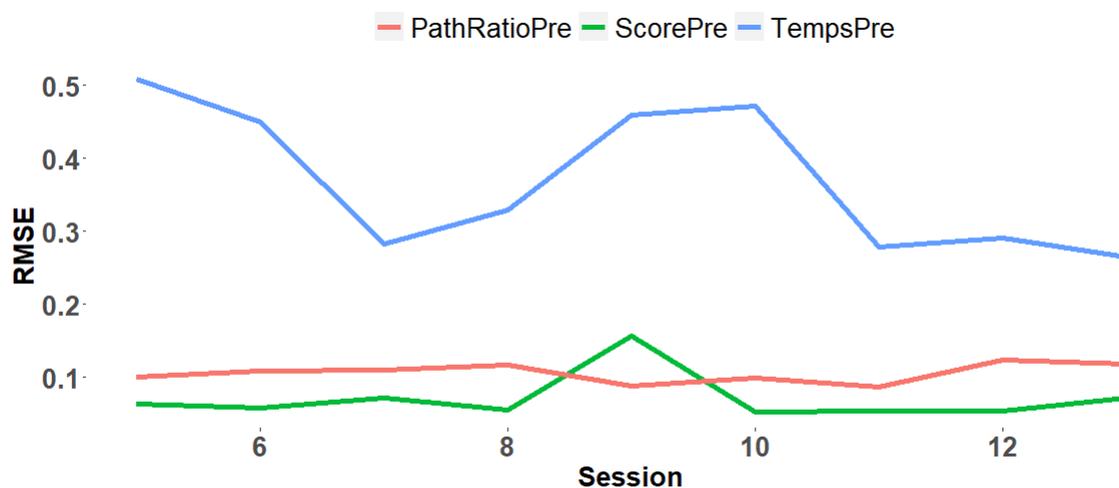
passé la méthode de référence baseline pour les indicateurs ScorePre et TempsPre et lui est légèrement inférieure pour l'indicateur PathRatioPre. Comme nous pouvons le voir sur la figure 5.15 et le tableau 5.4 les modèles FA-MONO et FA-MULTI sont proches du modèle AR-MULTI en termes de performance et le modèle FA-MONO lui est même supérieur sur l'indicateur TempsPre.

Nous avons pu voir lors des premières analyses descriptives du jeu LadyBug (Section 5.2.3) que le patient post-AVC "P10" se distinguait nettement des autres patients du fait d'une tendance à la baisse de ses performances de rééducation contrairement à tous les autres patients du jeu de données. Le tableau 5.5 montre les résultats obtenus par toutes les approches en termes de aRMSE moyen ± écart-type après exclusion du patient "P10" du jeu de données. On observe que les performances prédictives de tous les modèles sont meilleures et plus stables après exclusion du patient P10 en comparaison du Tableau 5.3. En effet l'exclusion du patient P10 permet une réduction de tous les aRMSE et des écarts-types associés. En outre, on peut constater que tous les modèles prédictifs obtiennent des aRMSE inférieurs à celui de l'approche 'baseline' ce qui signifie que le patient P10, par ses spécificités cliniques, inhibait voire annulait le pouvoir prédictif des modèles. Nous notons aussi que la forêt aléatoire multisortie (FA-MULTI) obtient la meilleure performance prédictive.

La figure 5.16 représente l'évaluation chronologique avec les modèles entraînés et testés sur le jeu de données sans le patient "P10". Nous observons en premier lieu que la plupart des modèles monosorties et multisorties obtiennent des performances prédictives supérieures à celles obtenues par la méthode de référence 'baseline' tout au long de l'évaluation chronologique. Le modèle AR-MONO est le seul qui obtient de moins bonnes performances prédictives que la méthode 'baseline' à certains moments de l'évaluation chronologique. Le résultat principal mis en évidence par cette figure est que la forêt aléatoire multisortie (FA-MULTI) obtient de meilleures performances prédictives que tous les autres modèles prédictifs pendant toute l'évaluation chronologique. Nous pouvons constater que toutes les approches présentent un pic de



**Figure 5.16. Évaluation chronologique après suppression des données du patient "P10"**



**Figure 5.17. Evaluation chronologique du modèle FA-MULTI par indicateur après suppression des données du patient "P10"**

aRMSE à la 9<sup>ème</sup> session de rééducation.

Sur la figure 5.17 nous pouvons voir que le meilleur modèle (FA-MULTI) obtient les meilleures performances prédictives pour l'indicateur ScorePre, les moins bonnes performances étant obtenues pour l'indicateur TempsPre. Un second constat concerne le fait que le modèle AR-MULTI obtient les meilleures performances prédictives avec ou sans patient "P10".

Le tableau 5.6 expose les résultats obtenus par chaque approche pour chaque indicateur de performance de rééducation. Nous pouvons observer que la plupart des approches ont eu un gain de performance prédictive sur les différents indicateurs par rapport à l'expérimentation précédente qui incluait le patient P10. Néanmoins, il n'y a pas de modèle qui obtient les meilleurs résultats sur tous les indicateurs. Le modèle FA-MULTI se démarque sur l'indicateur ScorePre et il rivalise avec la méthode

Indicateur	baseline	AR-MONO	FA-MONO	AR-MULTI	FA-MULTI
ScorePre	0.107 ± 0.023	0.092 ± 0.030	0.082 ± 0.031	0.082 ± 0.032	<b>0.070 ± 0.033</b>
TempsPre	0.423 ± 0.076	0.417 ± 0.092	0.387 ± 0.099	<b>0.365 ± 0.091</b>	0.370 ± 0.099
PathRatioPre	<b>0.106 ± 0.016</b>	0.127 ± 0.007	0.124 ± 0.019	0.130 ± 0.006	<b>0.106 ± 0.013</b>

**Tableau 5.6. RMSE moyen ± écart-type par modèle prédictif et par indicateur de performance de rééducation après suppression des données du patient P10**

de référence baseline sur l'indicateur PathRatioPre. En ce qui concerne l'indicateur TempsPre, le modèle AR-MULTI obtient le meilleur RMSE même si le modèle FA-MULTI obtient des performances assez proches. Ces résultats nous indiquent qu'aucun modèle ne domine globalement tous les autres, même si on constate tout de même une supériorité des modèles multisorties (par rapport aux modèles monosorties). Le modèle FA-MULTI dans ce cas de figure (sans patient P10) est un modèle qui représente un bon compromis (en termes de performances par indicateur).

## 5.5 Discussion et Conclusion

Dans ce chapitre nous avons voulu faire une première analyse des données collectées durant le programme de rééducation ReArm. Une première analyse descriptive des données montre en moyenne (sur l'ensemble des patients) des évolutions croissantes des performances de rééducation et une baisse de la difficulté ressentie (RPE) (voir Section 5.2). L'analyse de corrélation (Figure 5.7) a permis de mettre en évidence les corrélations entre indicateurs de performance et entre indicateurs de fatigue/difficulté ressentie. Nous avons pu constater que plus un patient performe (en termes de ScorePre, TempsPre) sur le jeu LadyBug, plus son RPE est élevé. Pour l'indicateur PathRatioPre qui est relatif à la qualité des mouvements effectués durant le jeu, nous avons observé de faibles corrélations avec les autres indicateurs de performance (ScorePre, TempsPre) et plus globalement avec tout les autres indicateurs à l'exception de l'indicateur PathRatioPost. Cela s'explique par le fait que les indicateurs ScorePre et TempsPre sont plus représentatifs de l'efficacité du patient au jeu, *i.e.* un patient peut être bon au jeu avec des mouvements de mauvaise qualité grâce au mécanisme de compensation qui induit une réorganisation au niveau cérébral [177]. En effet, les mécanismes de compensation permettent aux patients d'être plus efficaces dans leurs activités. Cependant, la compensation n'est pas une réelle rééducation car elle n'implique pas toujours que les patients regagnent un mouvement "normal". De plus, cette dernière peut être néfaste *e.g.* un patient peut compenser ses difficultés en utilisant beaucoup plus souvent son bras sain, ce qui va encore plus affaiblir le bras parétique.

Dans la suite de cette étude nous avons analysé les évolutions des différents indicateurs au niveau individuel (Section 5.2.3). Nous avons pu constater que la majorité des patients post-AVC qui ont participé au programme de rééducation ReArm ont eu

une évolution positive de leurs performances de rééducation pour les différents indicateurs (Figures 5.2, 5.3, 5.4). Néanmoins, cette analyse plus granulaire a aussi montré que parmi l'ensemble des patients, certains avaient des performances qui semblaient avoir une évolution négative. Ce constat peut être dû au fait que les patients post-AVC ont des niveaux d'atteinte différents et propres à chacun et que donc le programme de rééducation ReArm n'est pas forcément adapté à tout les niveaux d'atteinte.

La partie prédictive de ce travail a été déclinée en deux applications pour deux objectifs cliniques distincts. L'objectif clinique secondaire de ce travail visait à prédire la difficulté ressentie (RPE) par les patients post-AVC en fonction de leurs performances au jeu LadyBug Pre durant leur session de manière à éviter aux thérapeutes d'avoir à remplir de RPE lors des sessions de rééducation. Nous avons comparé 2 modèle pour cette application, le *KNN* et l'*EKNN* régression ainsi qu'une méthode de référence baseline qui prédit le RPE moyen du patient aux sessions antérieures. Les résultats ont mis en évidence dans un premier temps un pouvoir prédictif pour les prédictions réalisées par *KNN* ou *EKNN*, qui se démarquent significativement de la méthode de référence baseline (voir tableau 5.2). Dans un seconde temps, nous avons voulu évaluer la résistance des modèles aux données manquantes. Nous avons constaté que, grâce modèles d'incertitude combinés aux méthodes d'imputation (voir Section 5.3.3), l'*EKNN* permettait d'obtenir de meilleurs performances prédictives que le *KNN*, ce qui illustre l'intérêt de l'imputation crédibiliste.

L'objectif clinique principal était de pouvoir anticiper la réponse des patients post-AVC au programme de rééducation ReArm. Nous avons répondu à cette problématique en élaborant des modèles de régression qui prédisent les performances aux sessions suivantes qu'on a ensuite emboîtés de manière à obtenir des courbes prévisionnelles sur toute la durée de la thérapie. Nous avons fixé le départ des courbes prévisionnelles à la 5<sup>ème</sup> session. Nous avons comparé 2 modèles de régression multisorties : la forêt aléatoire et l'arbre de régression (FA-MULTI, AR-MULTI) à leurs homologues monosorties (AR-MONO, FA-MONO), en plus d'une méthode de référence baseline qui prédit la moyenne de chaque indicateur de performance calculée sur les sessions connues du patient pour lequel on génère la courbe prévisionnelle. Les résultats ont montré une supériorité du modèle arbre de régression multiserie sur les autres modèles monosorties et multiserie ainsi que sur la méthode de référence baseline (voir tableau 5.3 et Figure 5.15).

Nous nous sommes ensuite intéressés aux performances du meilleur modèle (arbre de régression multiserie) . Après avoir détecté un patient un peu atypique en termes d'évolution des indicateurs de rééducation (le patient P10), une seconde expérimentation a été menée après avoir exclu les données de ce patient du jeu de donnée. Les résultats ont montré une amélioration des performances prédictives de tout les modèles (voir Figure 5.16 et tableau 5.5). Nous avons pu voir que les modèles

multisorties étaient les meilleurs modèles et que la forêt aléatoire multisortie était devenu le meilleur modèle en termes de aRMSE (voir Eq. (2.15), Figure 5.16, tableau 5.5). Nous avons aussi observé ces améliorations au niveau de chaque indicateur. On a aussi pu constater que le modèle forêt aléatoire multisortie performe particulièrement bien sur l'indicateur ScorePre et qu'il est équivalent à la méthode de référence baseline sur l'indicateur PathRatioPre. Ces résultats nous amènent à penser que lors de la construction de modèles entraînés sur des données patients, il faut s'assurer que les patients présents dans les données d'entraînement ont bien des niveaux de handicaps (ou de pathologie) plus ou moins comparables. Nous recommandons donc aux thérapeutes de ne pas hésiter à fixer des seuils d'inclusion plus discriminants ou précis [183].

Une calibration des modèles et un enrichissement en données d'apprentissage permettraient vraisemblablement d'augmenter les performances prédictives et ainsi de mieux comprendre l'impact des différentes variables sur les indicateurs de performance de rééducation du jeu LadyBug Pre. Globalement, ces travaux nous ont permis de confirmer l'intérêt des modèles multisorties par rapport aux modèles monosorties à partir du moment où les sorties sont corrélées.

## Conclusions et Perspectives

### Conclusion

Dans cette thèse, nous avons exploré les potentialités de l'apprentissage automatique pour aider les experts à optimiser la performance sportive. En premier lieu, dans le chapitre 1 nous avons dressé un état des lieux des problématiques liées à la performance sportive, soulignant la prévalence d'indicateurs subjectifs pour quantifier la charge de travail des athlètes et l'épineuse question de l'imputation des valeurs manquantes. L'arsenal mathématique mobilisé est ensuite introduit dans le chapitre 2. Nous y détaillons les modèles de comportement de l'apprentissage automatique, la gestion de l'incertitude avec la théorie des fonctions de croyance, avec un focus sur l'algorithme EKNN(algorithme des k plus proches voisins évidentiel) et l'imputation de valeurs manquantes.

Une contribution majeure de nos travaux réside dans la proposition faite dans le chapitre 3 : il s'agit d'une approche novatrice pour modéliser l'incertitude relative aux valeurs manquantes imputées et son impact sur les prédictions exploitant ces données. S'appuyant sur les fonctions de croyance, cette approche est appliquée à deux jeux de données distincts, l'un relatif au sport et l'autre à la pandémie de COVID-19. Cette contribution s'avère fructueuse, démontrant l'amélioration des performances prédictives lorsqu'un modèle d'incertitude est associé aux méthodes d'imputation. L'application sur la prédiction des décès quotidiens COVID-19 a été une opportunité pour proposer d'autres modèles d'incertitude ainsi qu'une méthode d'imputation basée sur une variante de l'EKNN pour la régression que nous avons élaborée. L'extension de l'EKNN à la régression a été motivée par l'application aux données sportives : le but était de prédire la charge interne (RPE) de joueurs de football en fonction de leur planning d'entraînement et de l'historique des charges de travail. Nous avons voulu expérimenter notre approche dans le cadre de la régression car la plupart des indicateurs de charges de travail et de performances sont des indicateurs numériques et non pas catégoriels.

Dans le chapitre 4, nous explorons l'application de la régression multi-sorties pour prédire les performances de joueuses de handball élite au match suivant en fonction de leurs charges de travail antérieures et de variables contextuelles, les approches multi-sorties ont souligné leur supériorité par rapport aux modèles monosortie. Les modèles multi-sorties bénéficient en effet d'un contenu informationnel supplémentaire qui cor-

respond aux corrélations qui lient les différents indicateurs de performances techniques en match ou les performances athlétiques.

Enfin, au chapitre 5, nous explorons le domaine de la rééducation post-AVC en transposant les concepts développés dans le monde du sport des précédents chapitres. L'objectif y est double : i) prédire le RPE des patients post-AVC en fonction de leurs performances à une session de rééducation donnée, et ii) prédire leurs performances à venir au jeu sérieux LadyBug en fonction de leurs dernières performances et RPE connus. La forêt aléatoire multi-sorties s'affirme comme l'algorithme le plus performant pour cette tâche, ouvrant la voie à des outils d'aide à la décision pour guider les thérapeutes dans la planification de séances de rééducation.

En conclusion, cette thèse apporte des contributions intéressantes et novatrices à l'optimisation des performances pour les experts du sport comme de la santé. Les modèles et outils proposés permettent d'une part, d'anticiper la réponse des individus aux activités physiques (prédiction de RPE, prédiction de performances) et d'autre part, de mieux traiter les problèmes liés à la collecte de données avec la modélisation de l'incertitude des valeurs manquantes. L'utilisation de la régression multi-sorties et l'application de ces méthodes à la rééducation post-AVC constituent des propositions pertinentes pour des problématiques récurrentes dans le domaine de l'activité physique. Elles nous font aussi prendre conscience des avantages qu'apporte l'utilisation de modèles multi-sorties, notamment en ce qui concerne le gain en temps de calcul, corollaire non négligeable dans un contexte où nous devons réduire notre consommation énergétique.

L'exploration de nouvelles approches d'apprentissage automatique pour la performance sportive, l'adaptation de ces méthodes à d'autres disciplines et populations, et l'intégration de ces outils dans des applications concrètes pour les entraîneurs, préparateurs physiques et thérapeutes constituent des pistes prometteuses pour des recherches futures.

## Perspectives

Les travaux de cette thèse ouvrent la voie à de nombreuses perspectives pour l'optimisation de la performance sportive et la rééducation.

## Développement d'indicateurs de charge de travail

L'utilisation de techniques d'agrégation multicritère pour combiner les charges externes et internes offre la possibilité d'obtenir des indicateurs de charge de travail plus complexes. Cela permettrait une meilleure compréhension de l'impact de la charge de travail sur la performance grâce à des indicateurs qui intègrent différents

points de vue experts pour permettre une optimisation individualisée des programmes d'entraînement/rééducation.

## Performances et modèles multisorties

La démocratisation des approches multi-sorties et le développement/l'adaptation des algorithmes associés est essentiel pour une évaluation plus réaliste de la performance : la performance qu'elle soit celle d'un patient ou d'un athlète est par essence multifactorielle.

Il est aussi certain que le développement de nouveaux modèles d'incertitude et l'extension d'algorithmes d'apprentissage automatique aux fonctions de croyance constituent des pistes prometteuses pour améliorer la gestion des valeurs manquantes omniprésentes dans les données réelles et gérer l'incertitude inhérente à celles-ci lors de prédictions. Nos travaux portent sur des données manquantes de type purement aléatoire (Missing Completely At Random), néanmoins la flexibilité de la théorie des fonctions de croyance sur laquelle nous basons notre approche ouvre un vaste champ des possibles et permettrait de traiter des valeurs manquantes d'une autre nature. L'exploitation des modèles multi-sorties pour générer des courbes prévisionnelles et anticiper les réponses physiologiques des patients aux programmes de rééducation offre une perspective novatrice pour les thérapeutes. Enrichir les données d'entrée avec des indicateurs précis sur les niveaux d'atteinte et de capacités des patients permettra d'affiner encore la précision des prédictions réalisées. Une perspective possiblement fructueuse serait le développement d'algorithmes multi-sorties qui traitent les données incertaines.

L'ensemble de ces travaux peut contribuer à la conception d'outils d'aide à la décision pour les professionnels du milieu sportif et de la rééducation. Ces outils permettraient d'optimiser les programmes d'entraînement et de rééducation, de mieux gérer la charge de travail ainsi que de suivre la progression des patients/athlètes. Les perspectives ouvertes par cette thèse sont nombreuses. Elles placent l'apprentissage automatique au coeur de l'aide à l'optimisation de la performance sportive et de la rééducation, en offrant des solutions innovantes pour une meilleure prise en compte de la complexité et de l'individualité de chaque athlète et patient.

## Bibliographie

- [1] Gelayol Golkarnarenji, Minoos Naebe, Khashayar Badii, Abbas S Milani, Ali Jamali, Alireza Bab-Hadiashar, Reza N Jazar, and Hamid Khayyam. Multi-objective optimization of manufacturing process in carbon fiber industry using artificial intelligence techniques. *Ieee Access*, 7:67576–67588, 2019.
- [2] Dorina Weichert, Patrick Link, Anke Stoll, Stefan Rüping, Steffen Ihlenfeldt, and Stefan Wrobel. A review of machine learning for the optimization of production processes. *The International Journal of Advanced Manufacturing Technology*, 104(5):1889–1902, 2019.
- [3] Madani Abdu Alomar. Performance optimization of industrial supply chain using artificial intelligence. *Computational Intelligence and Neuroscience*, 2022, 2022.
- [4] Kadir Amasyali and Nora M El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018.
- [5] Paul Delanoë, Dieudonné Tchente, and Guillaume Colin. Method and evaluations of the effective gain of artificial intelligence models for reducing CO<sub>2</sub> emissions. *Journal of Environmental Management*, 331:117261, 2023.
- [6] Zeeshan Ahmed, Khalid Mohamed, Saman Zeeshan, and Xinqi Dong. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, 2020:baaa010, 2020.
- [7] João Gustavo Claudino, Daniel de Oliveira Capanema, Thiago Vieira de Souza, Julio Cerca Serrão, Adriano C Machado Pereira, and George P Nassis. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sports medicine-open*, 5:1–12, 2019.
- [8] Hans Van Eetvelde, Luciana D Mendonça, Christophe Ley, Romain Seil, and Thomas Tischer. Machine learning methods in sport injury prediction and pre-

- vention: a systematic review. *Journal of experimental orthopaedics*, 8:1–15, 2021.
- [9] Jennifer A Bunn, James W Navalta, Charles J Fountaine, and Joel D Reece. Current state of commercial wearable technology in physical activity monitoring 2015–2017. *International journal of exercise science*, 11(7):503, 2018.
- [10] Stephen W West, Jo Clubb, Lorena Torres-Ronda, Daniel Howells, Edward Leng, Jason D Vescovi, Sean Carmody, Michael Posthumus, Torstein Dalen-Lorensen, and Johann Windt. More than a metric: how training load is used in elite sport for athlete management. *International journal of sports medicine*, 42(04):300–306, 2021.
- [11] Franco M Impellizzeri, Samuele M Marcora, Aaron J Coutts, et al. Internal and external training load: 15 years on. *Int J Sports Physiol Perform*, 14(2):270–273, 2019.
- [12] Jill Borresen and Michael Ian Lambert. The quantification of training load, the training response and the effect on performance. *Sports medicine*, 39(9):779–795, 2009.
- [13] Naoki Kawamori and Gregory G Haff. The optimal training load for the development of muscular power. *The Journal of Strength and Conditioning Research*, 18(3):675–684, 2004.
- [14] Jordan L Fox, Robert Stanton, Charli Sargent, Sally-Anne Wintour, and Aaron T Scanlan. The association between training load and performance in team sports: a systematic review. *Sports Medicine*, 48(12):2743–2774, 2018.
- [15] John Orchard. Is there a relationship between ground and climatic conditions and injuries in football? *Sports medicine*, 32(7):419–432, 2002.
- [16] Elisa Bisagno, Sergio Morra, Martina Basciano, Carola Rosina, and Francesca Vitali. Assessing individual performance in team sports: a new method developed in youth volleyball. *Journal of Functional Morphology and Kinesiology*, 4(3):53, 2019.
- [17] Jordi Duch, Joshua S Waitzman, and Luís A Nunes Amaral. Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937, 2010.
- [18] Tindaro Bongiovanni, Athos Trecroci, Luca Cavaggioni, Alessio Rossi, Enrico Perri, Giulio Pasta, F Marcello Iaia, and Giampietro Alberti. Importance of an-

- thropometric features to predict physical performance in elite youth soccer: A machine learning approach. *Research in Sports Medicine*, 29(3):213–224, 2021.
- [19] Talko B Dijkhuis, Matthias Kempe, and Koen APM Lemmink. Early prediction of physical performance in elite soccer matches—a machine learning approach to support substitutions. *Entropy*, 23(8):952, 2021.
- [20] Brendan H Lazarus, Andrew M Stewart, Kevin M White, Amber E Rowell, Alireza Esmaeili, William G Hopkins, and Robert J Aughey. Proposal of a global training load measure predicting match performance in an elite team sport. *Frontiers in physiology*, 8:930, 2017.
- [21] Toni Caparrós, Eduard Alentorn-Geli, Gregory D Myer, Lluís Capdevila, Kristian Samuelsson, Bruce Hamilton, and Gil Rodas. The relationship of practice exposure and injury rate on game performance and season success in professional male basketball. *Journal of sports science and medicine*, 15(3):397, 2016.
- [22] Toni Modric, Sime Versic, and Damir Sekulic. Relations of the weekly external training load indicators and running performances in professional soccer matches. *Sport Mont*, 19(1):31–37, 2021.
- [23] Michel S Brink, Esther Nederhof, Chris Visscher, Sandor L Schmikli, and Koen APM Lemmink. Monitoring load, recovery, and performance in young elite soccer players. *The Journal of Strength and Conditioning Research*, 24(3):597–603, 2010.
- [24] Joseph OC Coyne, Aaron J Coutts, Robert U Newton, and G Gregory Haff. Relationships between different internal and external training load variables and elite international women’s basketball performance. *International Journal of Sports Physiology and Performance*, 16(6):871–880, 2021.
- [25] Romain Dubois, Noëlle Bru, Thierry Paillard, Anne Le Cunuder, Mark Lyons, Olivier Maurelli, Kilian Philippe, and Jacques Prioux. Rugby game performances and weekly workload: Using of data mining process to enter in the complexity. *PloS one*, 15(1):e0228107, 2020.
- [26] Tessitore A Gomez MA. Sansone P, Gasperi L. Training load, recovery and game performance in semiprofessional male basketball: influence of individual characteristics and contextual factors. *Biol Sport*, 38:207–217, 2021.
- [27] Saeid Younesi, Alireza Rabbani, Filipe Manuel Clemente, Rui Silva, Hugo Sarmiento, and António José Figueiredo. Relationships between aerobic performance, hemoglobin levels, and training load during small-sided games: a study in professional soccer players. *Frontiers in Physiology*, 12:649870, 2021.

- [28] María Isabel Piñar, David García, David Mancha-Triguero, and Sergio José Ibáñez. Effect of situational and individual factors on training load and game performance in liga femenina 2 basketball female players. *Applied Sciences*, 12(15):7752, 2022.
- [29] Joseph OC Coyne, Aaron J Coutts, Robert U Newton, and G Gregory Haff. The current state of subjective training load monitoring: follow-up and future directions. *Sports Medicine-Open*, 8(1):53, 2022.
- [30] William G Hopkins. Quantification of training in competitive sports: methods and applications. *Sports medicine*, 12:161–183, 1991.
- [31] Gunnar Borg. *Borg's perceived exertion and pain scales*. Human kinetics, 1998.
- [32] Emmanuel Vallance, Nicolas Sutton-Charani, Abdelhak Imoussaten, Jacky Montmain, and Stéphane Perrey. Combining internal-and external-training-loads to predict non-contact injuries in soccer. *Applied Sciences*, 10(15):5261, 2020.
- [33] Paul B Gastin, Denny Meyer, and Dean Robinson. Perceptions of wellness to monitor adaptive responses to training and competition in elite australian football. *The Journal of Strength and Conditioning Research*, 27(9):2518–2526, 2013.
- [34] Robin T Thorpe, Greg Atkinson, Barry Drust, and Warren Gregson. Monitoring fatigue status in elite team-sport athletes: implications for practice. *International journal of sports physiology and performance*, 12(s2):S2–27, 2017.
- [35] Benny Peiser, Thomas Reilly, Greg Atkinson, Barry Drust, and Jim Waterhouse. Seasonal changes and physiological responses: Their impact on activity, health, exercise and athletic performance. *International SportMed Journal*, 7(1):16–32, 2006.
- [36] Carl Foster, Jose A Rodriguez-Marroyo, and Jos J De Koning. Monitoring training loads: the past, the present, and the future. *International journal of sports physiology and performance*, 12(s2):S2–2, 2017.
- [37] Michael Ian Lambert and Jill Borresen. Measuring training load in sports. *International journal of sports physiology and performance*, 5(3):406–411, 2010.
- [38] Rasmus Pind and Jarek Mäestu. Monitoring training load: necessity, methods and applications. *Acta Kinesiologiae Universitatis Tartuensis*, 23:7–18, 2017.
- [39] VB Issurin. Generalized training effects induced by athletic preparation: a review. *Journal of sports medicine and physical fitness*, 49(4):333, 2009.

- [40] Franco M Impellizzeri, Ermanno Rampinini, Aaron J Coutts, ALDO Sassi, and Samuele M Marcora. Use of rpe-based training load in soccer. *Medicine and Science in sports and exercise*, 36(6):1042–1047, 2004.
- [41] Billy Sperlich, Kamiar Aminian, Peter Dükling, and Hans-Christer Holmberg. Wearable sensor technology for monitoring training load and health in the athletic population, 2020.
- [42] David L Carey, K Ong, Rod Whiteley, Kay M Crossley, Justin Crow, and Meg E Morris. Predictive modelling of training loads and injury in australian football. *International Journal of Computer Science in Sport*, 17(1):49–66, 2018.
- [43] Timothy G Eckard, Darin A Padua, Darren W Hearn, Brett S Pexa, and Barnett S Frank. The relationship between training load and injury in athletes: a systematic review. *Sports medicine*, 48:1929–1961, 2018.
- [44] Emmanuel Vallance, Nicolas Sutton-Charani, Patrice Guyot, and Stéphane Perrey. Predictive modeling of the ratings of perceived exertion during training and competition in professional soccer players. *Journal of Science and Medicine in Sport*, 2023.
- [45] Jonathan D Bartlett, Fergus O’Connor, Nathan Pitchford, Lorena Torres-Ronda, and Samuel J Robertson. Relationships between internal and external training load in team-sport athletes: evidence for an individualized approach. *International journal of sports physiology and performance*, 12(2):230–234, 2017.
- [46] Alessio Rossi, Enrico Perri, Athos Trecroci, Marco Savino, Giampietro Alberti, and F Marcello Iaia. Gps data reflect players’ internal load in soccer. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 890–893. IEEE, 2017.
- [47] Alessio Rossi, Enrico Perri, Luca Pappalardo, Paolo Cintia, and F Marcello Iaia. Relationship between external and internal workloads in elite soccer players: Comparison between rate of perceived exertion and training load. *Applied sciences*, 9(23):5174, 2019.
- [48] David L Carey, K Ong, Meg E Morris, Justin Crow, and Kay M Crossley. Predicting ratings of perceived exertion in australian football players: methods for live estimation. *International Journal of Computer Science in Sport*, 15 (2): 64, 77, 2016.
- [49] Gunnar Borg. Perceived exertion as an indicator of somatic stress. *Scandinavian journal of rehabilitation medicine*, 1970.

- [50] Tannath J Scott, Cameron R Black, John Quinn, and Aaron J Coutts. Validity and reliability of the session-rpe method for quantifying training in Australian football: a comparison of the cr10 and cr100 scales. *The Journal of Strength and Conditioning Research*, 27(1):270–276, 2013.
- [51] Javier Espasa-Labrador, Azahara Fort-Vanmeerhaeghe, Alicia M Montalvo, Marta Carrasco-Marginet, Alfredo Irurtia, and Julio Calleja-González. Monitoring internal load in women’s basketball via subjective and device-based methods: A systematic review. *Sensors*, 23(9):4447, 2023.
- [52] Andrew Watson, Stacey Brickson, Alison Brooks, and Warren Dunn. Subjective well-being and training load predict in-season injury and illness risk in female youth soccer players. *British journal of sports medicine*, 51(3):194–199, 2017.
- [53] Patrick C Maughan, Niall G MacFarlane, and Paul A Swinton. Relationship between subjective and external training load variables in youth soccer players. *International journal of sports physiology and performance*, 16(8):1127–1133, 2021.
- [54] Alice Iannaccone, Daniele Conte, Cristina Cortis, and Andrea Fusco. Usefulness of linear mixed-effects models to assess the relationship between objective and subjective internal load in team sports. *International journal of environmental research and public health*, 18(2):392, 2021.
- [55] Arne Jaspers, Michel S Brink, Steven GM Probst, Wouter GP Frencken, and Werner F Helsen. Relationships between training load indicators and training outcomes in professional soccer. *Sports medicine*, 47(3):533–544, 2017.
- [56] Arne Jaspers, Tim Op De Beéck, Michel S Brink, Wouter GP Frencken, Filip Staes, Jesse J Davis, and Werner F Helsen. Relationships between the external and internal training load in professional soccer: what can we learn from machine learning? *International journal of sports physiology and performance*, 13(5):625–630, 2018.
- [57] Jakub Marynowicz, Mateusz Lango, Damian Horna, Karol Kikut, and Marcin Andrzejewski. Predicting ratings of perceived exertion in youth soccer using decision tree models. *Biology of Sport*, 39(2):245–252, 2022.
- [58] Franco M Impellizzeri, Annie C Jeffries, Asaf Weisman, Aaron J Coutts, Alan McCall, Shaun J McLaren, and Judd Kalkhoven. The ‘training load’ construct: why it is appropriate and scientific. *Journal of Science and Medicine in Sport*, 25(5):445–448, 2022.

- [59] Craig A Staunton, Grant Abt, Dan Weaving, and Daniel WT Wundersitz. Misuse of the term ‘load’ in sport and exercise science. *Journal of science and medicine in sport*, 25(5):439–444, 2022.
- [60] Eric W Banister and Thomas W Calvert. Planning for future performance: implications for long term training. *Canadian journal of applied sport sciences. Journal canadien des sciences appliquees au sport*, 5(3):170–176, 1980.
- [61] John R Fitz-Clarke, R Hugh Morton, and Eric W Banister. Optimizing athletic performance by influence curves. *Journal of Applied Physiology*, 71(3):1151–1158, 1991.
- [62] Daniel P Nicolella, Lorena Torres-Ronda, Kase J Saylor, and Xavi Schelling. Validity and reliability of an accelerometer-based player tracking device. *PLoS one*, 13(2):e0191823, 2018.
- [63] Steve Barrett, Adrian Midgley, and Ric Lovell. Playerload™: reliability, convergent validity, and influence of unit position during treadmill running. *International journal of sports physiology and performance*, 9(6):945–952, 2014.
- [64] Eirik H Wik, Live S Luteberget, and Matt Spencer. Activity profiles in international women’s team handball using playerload. *International journal of sports physiology and performance*, 12(7):934–942, 2017.
- [65] Georgia M Black and Tim J Gabbett. Repeated high-intensity-effort activity in elite and semielite rugby league match play. *International journal of sports physiology and performance*, 10(6):711–717, 2015.
- [66] Rich D Johnston, Tim J Gabbett, David G Jenkins, and Michael J Speranza. Effect of different repeated-high-intensity-effort bouts on subsequent running, skill performance, and neuromuscular function. *International Journal of Sports Physiology and Performance*, 11(3):311–318, 2016.
- [67] Carl Foster, Daniel Boulosa, Michael McGuigan, Andrea Fusco, Cristina Cortis, Blaine E Arney, Bo Orton, Christopher Dodge, Salvador Jaime, Kim Radtke, et al. 25 years of session rating of perceived exertion: historical perspective and development. *International journal of sports physiology and performance*, 16(5):612–621, 2021.
- [68] Christopher Carling, Tom Reilly, and A Mark Williams. *Performance assessment for field sports*. Routledge, 2008.

- [69] Tiziana D’Isanto, Francesca D’Elia, Gaetano Raiola, and Gaetano Altavilla. Assessment of sport performance: Theoretical aspects and practical indications. *Sport Mont*, 17(1):79–82, 2019.
- [70] D’Isanto Tiziana, Gaetano Altavilla, Raiola Gaetano, et al. Teaching method in volleyball service: Intensive and extensive tools in cognitive and ecological approach. *Journal of Physical Education and Sport*, 17:2222–2227, 2017.
- [71] Lars Bojsen Michalsik. On-court physical demands and physiological aspects in elite team handball. *Handball sports medicine: Basic science, injury management and return to sport*, pages 15–33, 2018.
- [72] Jean-Francis Grehaigne, Paul Godbout, and Daniel Bouthier. Performance assessment in team sports. *Journal of teaching in Physical Education*, 16(4):500–516, 1997.
- [73] Carmen Manchado, Juan Tortosa-Martínez, Helena Vila, Carmen Ferragut, and Petra Platen. Performance factors in women’s team handball: Physical and physiological aspects—a review. *The Journal of Strength and Conditioning Research*, 27(6):1708–1719, 2013.
- [74] João Henrique Gomes, Renata Rebello Mendes, Marcos Bezerra de Almeida, Marcelo Callegari Zanetti, Gerson dos Santos Leite, and Aylton José Figueira Júnior. Relationship between physical fitness and game-related statistics in elite professional basketball players: Regular season vs. playoffs. *Motriz: Revista de Educação Física*, 23:e101626, 2017.
- [75] Lucas A Pereira, Sophia Nimphius, Ronaldo Kobal, Katia Kitamura, Luiz AL Turisco, Rita C Orsi, César C Cal Abad, and Irineu Loturco. Relationship between change of direction, speed, and power in male and female national olympic team handball athletes. *The Journal of Strength and Conditioning Research*, 32(10):2987–2994, 2018.
- [76] Toni Modric, Sime Versic, Damir Sekulic, and Silvester Liposek. Analysis of the association between running performance and game performance indicators in professional soccer players. *International journal of environmental research and public health*, 16(20):4032, 2019.
- [77] Toni Modric, Sime Versic, Patrik Drid, Marko Stojanovic, Łukasz Radzimiński, Cyril Bossard, Tomasz Aftański, and Damir Sekulic. Analysis of running performance in the offensive and defensive phases of the game: is it associated with the team achievement in the uefa champions league? *Applied Sciences*, 11(18):8765, 2021.

- [78] Jakub Marynowicz, Karol Kikut, Mateusz Lango, Damian Horna, and Marcin Andrzejewski. Relationship between the session-rpe and external measures of training load in youth soccer training. *The Journal of Strength and Conditioning Research*, 34(10):2800–2804, 2020.
- [79] Vicente de Dios-Álvarez, David Suárez-Iglesias, Sara Bouzas-Rico, Pello Alkain, Adrián González-Conde, and Carlos Ayan-Perez. Relationships between rpe-derived internal training load parameters and gps-based external training load variables in elite young soccer players. *Research in Sports Medicine*, 31(1):58–73, 2023.
- [80] Youri Geurkink, Gilles Vandewiele, Maarten Lievens, Filip De Turck, Femke Ongenaes, Stijn PJ Matthys, Jan Boone, and Jan G Bourgois. Modeling the prediction of the session rating of perceived exertion in soccer: Unraveling the puzzle of predictive indicators. *International journal of sports physiology and performance*, 14(6):841–846, 2019.
- [81] Mohamed Elloumi, Emna Makni, Wassim Moalla, Taieb Bouaziz, Zouhair Tabka, Gérard Lac, and Karim Chamari. Monitoring training load and fatigue in rugby sevens players. *Asian journal of sports medicine*, 3(3):175, 2012.
- [82] Unai Zurutuza, Julen Castellano, Ibon Echeazarra, and David Casamichana. Absolute and relative training load and its relation to fatigue in football. *Frontiers in psychology*, 8:878, 2017.
- [83] Robin T Thorpe, Anthony J Strudwick, Martin Buchheit, Greg Atkinson, Barry Drust, and Warren Gregson. The influence of changes in acute training load on daily sensitivity of morning-measured fatigue variables in elite soccer players. *International journal of sports physiology and performance*, 12(s2):S2–107, 2017.
- [84] Eric W Banister, Thomas W Calvert, Margaret V Savage, and Tim Bach. A systems model of training for athletic performance. *Aust J Sports Med*, 7(3):57–61, 1975.
- [85] Thierry Busso. Variable dose-response relationship between exercise training and performance. *Medicine and Science in Sports and Exercise*, 35(7):1188–1195, 2003.
- [86] Alan Griffin, Ian C Kenny, Thomas M Comyns, Helen Purtill, Caoimhe Tieran, Eoin O’Shaughnessy, and Mark Lyons. Training load monitoring in team sports: a practical approach to addressing missing data. *Journal of sports sciences*, 39(19):2161–2171, 2021.

- [87] Lauren C Benson, Carlyn Stilling, Oluwatoyosi BA Owoeye, and Carolyn A Emery. Evaluating methods for imputing missing data from longitudinal monitoring of athlete workload. *Journal of Sports Science and Medicine*, 20(2):188, 2021.
- [88] Amarah Epp-Stobbe, Ming-Chang Tsai, and Marc Klimstra. Comparison of imputation methods for missing rate of perceived exertion data in rugby. *Machine Learning and Knowledge Extraction*, 4(4):827–838, 2022.
- [89] Thierry Busso, K Häkkinen, A Pakarinen, C Carasso, Jean René Lacour, PV Komi, and H Kauhanen. A systems model of training responses and its relationship to hormonal responses in elite weight-lifters. *European journal of applied physiology and occupational physiology*, 61:48–54, 1990.
- [90] Thierry Busso, Henri Benoit, Regis Bonnefoy, Léonard Feasson, and Jean-René Lacour. Effects of training frequency on the dynamics of performance response to a single training bout. *Journal of Applied Physiology*, 92(2):572–580, 2002.
- [91] Thomas W Calvert, Eric W Banister, Margaret V Savage, and Tim Bach. A systems model of the effects of training on physical performance. *IEEE Transactions on systems, man, and cybernetics*, (2):94–102, 1976.
- [92] Thierry Busso, CLAUDE Carasso, and JEAN-RENE Lacour. Adequacy of a systems structure in the modeling of training effects on performance. *Journal of applied physiology*, 71(5):2044–2049, 1991.
- [93] Thierry Busso, K Häkkinen, A Pakarinen, H Kauhanen, PV Komi, and Jean René Lacour. Hormonal adaptations and modelled responses in elite weightlifters during 6 weeks of training. *European journal of applied physiology and occupational physiology*, 64:381–386, 1992.
- [94] Thierry Busso, Robin Candau, and Jean-René Lacour. Fatigue and fitness modelled from the effects of training on performance. *European journal of applied physiology and occupational physiology*, 69:50–54, 1994.
- [95] David C Clarke and Philip F Skiba. Rationale and resources for teaching the mathematical modeling of athletic training and performance. *Advances in physiology education*, 2013.
- [96] Elisée Gouba, Balira Ousmane Konfe, Ousseynou Nakoulima, Blaise Some, Olivier Hue, et al. Applying a mathematical model to the performance of a female monofin swimmer. *Applied Mathematics*, 4(12):1673, 2013.

- [97] Ben Stephens Hemingway, Leon Greig, Mladen Jovanovic, Ben Ogorek, and Paul Swinton. Traditional and contemporary approaches to mathematical fitness-fatigue models in exercise science: A practical guide with resources. part i. 2021.
- [98] Frank Imbach, Nicolas Sutton-Charani, Jacky Montmain, Robin Candau, and Stéphane Perrey. The use of fitness-fatigue models for sport performance modelling: conceptual issues and contributions from machine-learning. *Sports Medicine-Open*, 8(1):1–6, 2022.
- [99] Michael Svensson and Barry Drust. Testing soccer players. *Journal of sports sciences*, 23(6):601–618, 2005.
- [100] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. pages 57–72, 2008.
- [101] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.
- [102] Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.
- [103] Michael W Browne. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132, 2000.
- [104] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.
- [105] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11(1), 2010.
- [106] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49. Springer, 2008.
- [107] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.
- [108] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.
- [109] Etienne Grossmann. Adatree: boosting a weak classifier into a decision tree. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 105–105. IEEE, 2004.

- [110] Chen Jin, Luo De-Lin, and Mu Fen-Xiang. An improved id3 decision tree algorithm. In *2009 4th international conference on computer science and Education*, pages 127–130. IEEE, 2009.
- [111] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications*, 4(2):13–19, 2014.
- [112] Leo Breiman. *Classification and regression trees*. Routledge, 2017.
- [113] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [114] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.
- [115] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.
- [116] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 22–30. Springer, 2004.
- [117] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581*, pages 1159–1168, 2012.
- [118] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
- [119] D. Dubois and H. Prade. Possibility theory: qualitative and quantitative aspects. pages 169–226, 1998.
- [120] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.
- [121] T. Denœux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.
- [122] T. Denœux. A neural network classifier based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.
- [123] N. Sutton-Charani, S. Destercke, and T. Denœux. Learning decision trees from uncertain data with an evidential em approach. 1:111–116, 2013.
- [124] M. C. Florea, A.-L. Jousselme, É. Bossé, and D. Grenier. Robust combination rules for evidence theory. *Information Fusion*, 10(2):183–197, 2009.

- [125] P. Smets. Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of approximate reasoning*, 9(1):1–35, 1993.
- [126] Didier Dubois and Henri Prade. A set-theoretic view of belief functions logical operations and approximations by fuzzy sets. *International Journal Of General System*, 12(3):193–226, 1986.
- [127] Philippe Smets. Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of approximate reasoning*, 9(1):1–35, 1993.
- [128] Mihai Cristian Florea, Anne-Laure Jousselme, Éloi Bossé, and Dominic Grenier. Robust combination rules for evidence theory. *Information Fusion*, 10(2):183–197, 2009.
- [129] Didier Dubois and Henri Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence*, 4(3):244–264, 1988.
- [130] P. Smets and R. Kennes. The transferable belief model. *Artificial intelligence*, 66(2):191–234, 1994.
- [131] Thierry Denœux. Maximum likelihood estimation from fuzzy data using the em algorithm. *Fuzzy sets and systems*, 183(1):72–91, 2011.
- [132] Thierry Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.
- [133] Thierry Denœux, Orakanya Kanjanatarakul, and Songsak Sriboonchitta. Eknclus: a clustering procedure based on the evidential k-nearest neighbor rule. *Knowledge-Based Systems*, 88:57–69, 2015.
- [134] P. Smets. Constructing the pignistic probability function in a context of uncertainty. 89:29–40, 1989.
- [135] P. Smets. Decision making in the tbm: the necessity of the pignistic transformation. *International journal of approximate reasoning*, 38(2):133–147, 2005.
- [136] B. Yuan, X. Yue, Y. Lv, and T. Denœux. Evidential deep neural networks for uncertain data classification. pages 427–437, 2020.
- [137] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

- [138] Donald B Rubin. Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC, 2018.
- [139] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. 2002.
- [140] Yuzhe Liu and Vanathi Gopalakrishnan. An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data*, 2(1):8, 2017.
- [141] K. Lakshminaryan, S. A. Harp, and T. Samed. Imputation of missing data in industrial databases. *Applied intelligence*, 11(3):259–275, 1999.
- [142] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115, 2010.
- [143] A. Farhangfar, L. Kurgan, and J. Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12):3692–3705, 2008.
- [144] Mark Huisman. Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, 10(1):1–29, 2009.
- [145] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la société française de statistique*, 153(2):79–99, 2012.
- [146] CK Enders. *Applied missing data analysis*. new york, ny. Guilford Press. Everitt, BS (1981). A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, 16:171–180, 2010.
- [147] John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.
- [148] D. Li, J. Deogun, W. Spaulding, and B. Shuart. Towards missing data imputation: a study of fuzzy k-means clustering method. pages 573–579, 2004.
- [149] S. Azim and S. Aggarwal. Hybrid model for data imputation: using fuzzy c means and multi layer perceptron. pages 1281–1285, 2014.
- [150] K. J. Nishanth and V. Ravi. Probabilistic neural network based categorical data imputation. *Neurocomputing*, 218:17–25, 2016.

- [151] J. Hamizadeh and M. Moradi. Enhancing data analysis: uncertainty-resistance method for handling incomplete data. *Applied Intelligence*, 50(1):74–86, 2020.
- [152] Z. Liu, Q. Pan, G. Mercier, and J. Dezert. Pattern classification with missing data using belief functions. In *17th International Conference on Information Fusion (FUSION 2014)*, pages 1–8, 2014.
- [153] A-L. Jusselme and P. Maupin. Comparison of uncertainty representations for missing data in information retrieval. In *Proceedings of the 16th International Conference on Information Fusion*, pages 1902–1909, 2013.
- [154] Simon Petit-Renaud and Thierry Denoeux. Handling different forms of uncertainty in regression analysis: a fuzzy belief structure approach. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 340–351. Springer, 1999.
- [155] Gunnar A Borg. Psychophysical bases of perceived exertion. *Medicine and science in sports and exercise*, 14(5):377–381, 1982.
- [156] T Fenoeux and M Skarstein Bjanger. Induction of decision trees from partially classified data using belief functions. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0, volume 4, pages 2923–2928. IEEE, 2000.*
- [157] Benjamin Quost and Thierry Dencœux. Learning from data with uncertain labels by boosting credal classifiers. In *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery From Uncertain Data*, pages 38–47, 2009.
- [158] Asma Trabelsi, Zied Elouedi, and Eric Lefevre. Decision tree classifiers for evidential attribute values and class labels. *Fuzzy Sets and Systems*, 366:46–62, 2019.
- [159] L. Zadeh. Fuzzy logic, neural networks, and soft computing. pages 775–782, 1996.
- [160] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangooei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare, M. Panahiazar, S. Nahavandi, D. Srinivasan, A. F. Atiya, and U. R. Acharya. Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020). *Annals of Operations Research*, pages 1–42, 2021.

- [161] L. Jacquin, A. Imoussaten, F. Troussel, J. Montmain, and D. Perrin. Evidential classification of incomplete data via imprecise relabelling: Application to plastic sorting. pages 122–135, 2019.
- [162] Simon Petit-Renaud and Thierry Denœux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35(1):1–28, 2004.
- [163] A. Trabelsi, Z. Elouedi, and E. Lefevre. Decision tree classifiers for evidential attribute values and class labels. *Fuzzy Sets and Systems*, 366:46–62, 2019.
- [164] S. Tsang, B. Kao, K. Y Yip, W.-S. Ho, and S. D. Lee. Decision trees for uncertain data. *IEEE transactions on knowledge and data engineering*, 23(1):64–78, 2009.
- [165] Shona L Halson. Monitoring training load to understand fatigue in athletes. *Sports medicine*, 44(2):139–147, 2014.
- [166] Callum J McCaskie, Warren B Young, Brendan B Fahrner, and Marc Sim. Association between preseason training and performance in elite australian football. *International journal of sports physiology and performance*, 14(1):68–75, 2019.
- [167] Myles Hollander and FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS. Tests for dependence. 1979.
- [168] Tim J Gabbett. Debunking the myths about training load, injury and performance: empirical evidence, hot topics and recommendations for practitioners. *British journal of sports medicine*, 54(1):58–66, 2020.
- [169] Friedhelm Hummel, Pablo Celnik, Pascal Giraux, Agnes Floel, Wan-Hsun Wu, Christian Gerloff, and Leonardo G Cohen. Effects of non-invasive cortical stimulation on skilled motor function in chronic stroke. *Brain*, 128(3):490–499, 2005.
- [170] María Antonia Fuentes, Adrián Borrego, Jorge Latorre, Carolina Colomer, Mariano Alcañiz, María José Sánchez-Ledesma, Enrique Noé, and Roberto Llorens. Combined transcranial direct current stimulation and virtual reality-based paradigm for upper limb rehabilitation in individuals with restricted movements. a feasibility study with a chronic stroke survivor with severe hemiparesis. *Journal of medical systems*, 42:1–9, 2018.
- [171] Elaine L Miller, Laura Murray, Lorie Richards, Richard D Zorowitz, Tamilyn Bakas, Patricia Clark, and Sandra A Billinger. Comprehensive overview of nursing and interdisciplinary rehabilitation care of the stroke patient: a scientific statement from the american heart association. *Stroke*, 41(10):2402–2448, 2010.

- [172] Janne Marieke Veerbeek, Erwin van Wegen, Roland van Peppen, Philip Jan van der Wees, Erik Hendriks, Marc Rietberg, and Gert Kwakkel. What is the evidence for physical therapy poststroke? a systematic review and meta-analysis. *PloS one*, 9(2):e87987, 2014.
- [173] Alex Pollock, Sybil E Farmer, Marian C Brady, Peter Langhorne, Gillian E Mead, Jan Mehrholz, and Frederike van Wijck. Interventions for improving upper limb function after stroke. *Cochrane Database of Systematic Reviews*, (11), 2014.
- [174] E Taub, G Uswatte, V Mark, and D Morris. The learned nonuse phenomenon: implications for rehabilitation. *Eura Medicophys*, 42:241–55, 2006.
- [175] Pålvel Lindberg. *Brain plasticity and upper limb function after stroke: some implications for rehabilitation*. PhD thesis, Rehabiliteringsmedicin, 2007.
- [176] Peter S Lum, Sara Mulroy, Richard L Amdur, Philip Requejo, Boris I Prilutsky, and Alexander W Dromerick. Gains in upper extremity function after stroke via recovery or compensation: potential differential effects on amount of real-world limb use. *Topics in stroke rehabilitation*, 16(4):237–253, 2009.
- [177] Theresa A Jones. Motor compensation and its effects on neural reorganization after stroke. *Nature Reviews Neuroscience*, 18(5):267–280, 2017.
- [178] Jeffrey A Kleim. Neural plasticity and neurorehabilitation: teaching the new brain old tricks. *Journal of communication disorders*, 44(5):521–528, 2011.
- [179] Jeffrey A Kleim and Theresa A Jones. Principles of experience-dependent neural plasticity: implications for rehabilitation after brain damage. 2008.
- [180] Sandeep K Subramanian, Christiane B Lourenço, Gevorg Chilingaryan, Heidi Sveistrup, and Mindy F Levin. Arm motor recovery using a virtual reality intervention in chronic stroke: randomized control trial. *Neurorehabilitation and neural repair*, 27(1):13–23, 2013.
- [181] Kate E Laver, Belinda Lange, Stacey George, Judith E Deutsch, Gustavo Saposnik, and Maria Crotty. Virtual reality for stroke rehabilitation. *Cochrane database of systematic reviews*, (11), 2017.
- [182] Wei-Peng Teo, Makii Muthalib, Sami Yamin, Ashlee M Hendy, Kelly Bramstedt, Eleftheria Kotsopoulos, Stephane Perrey, and Hasan Ayaz. Does a combination of virtual reality, neuromodulation and neuroimaging provide a comprehensive platform for neurorehabilitation?—a narrative review of the literature. *Frontiers in human neuroscience*, 10:284, 2016.

- [183] Camille O Muller, Makii Muthalib, Denis Mottet, Stéphane Perrey, Gérard Dray, Marion Delorme, Claire Duflos, Jérôme Froger, Binbin Xu, Germain Faity, et al. Recovering arm function in chronic stroke patients using combined anodal hd-tacs and virtual reality therapy (rearm): a study protocol for a randomized controlled trial. *Trials*, 22:1–18, 2021.
- [184] Marine Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What’s a good imputation to predict with missing values? *NeurIPS 2021 - 35th Conference on Neural Information Processing Systems*, 12 2021.
- [185] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

## LISTE DES PUBLICATIONS

1. Elimam, R., Sutton-Charani, N., Montmain, J., Perrey, S. (2020, October). Imputation crédibiliste pour la prédiction de charge interne de joueurs de football. In LFA'2020- 29èmes Rencontres Francophones sur la Logique Floue et ses Applications.
2. Elimam, R., Sutton-Charani, N., Perrey, S., Montmain, J. (2022). Uncertain imputation for time-series forecasting: Application to COVID-19 daily mortality prediction. PLOS Digital Health, 1(10), e0000115.
3. Elimam, R., Sutton-Charani, N., Perrey, S., Montmain, J. (2023). Multi-output regression for the prediction of world-class performances in women's handball. IEEEAccess, reviewing process.