Université de Montréal

Prédiction d'états mentaux futurs à partir de données de phénotypage numérique

Par

Thierry Jean

Université de Montréal

Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de Maîtrise

en Sciences biomédicales

option Médecine Computationnelle

Décembre 2023

Université de Montréal

Université de Montréal, Faculté de Médecine

Ce mémoire intitulé

Prédiction d'états mentaux futurs à partir de données de phénotypage numérique

Présenté par

Thierry Jean

A été évalué par un jury composé des personnes suivantes

Paul Lespérance
Président-rapporteur

Pierre Orban
Directeur de recherche

Marc Lanovaz
Membre du jury

Résumé

Le phénotypage numérique mobilise les nombreux capteurs du téléphone intelligent (p. ex. : accéléromètre, GPS, Bluetooth, métadonnées d'appels) pour mesurer le comportement humain au quotidien, sans interférence, et les relier à des symptômes psychiatriques ou des indicateurs de santé mentale. L'apprentissage automatique est une composante intégrale au processus de transformation de signaux bruts en information intelligible pour un clinicien. Cette approche émerge d'une volonté de caractériser le profil de symptômes et ses variations dans le temps au niveau individuel.

Ce projet consistait à prédire des variables de santé mentale (p. ex. : stress, humeur, sociabilité, hallucination) jusqu'à sept jours dans le futur à partir des données du téléphone intelligent pour des patients avec un diagnostic de schizophrénie. Le jeu de données CrossCheck, composé d'un échantillon de 62 participants, a été utilisé. Celui-ci inclut 23,551 jours de signaux du téléphone avec 29 attributs et 6364 autoévaluations de l'état mental à l'aide d'échelles ordinales à 4 ancrages.

Des modèles prédictifs ordinaux ont été employés pour générer des prédictions discrètes interprétables sur l'échelle de collecte de données. Au total, 240 modèles d'apprentissage automatique ont été entrainés, soit les combinaisons de 10 variables de santé mentale, 3 horizons temporels (même jour, prochain jour, prochaine semaine), 2 algorithmes (XGBoost, LSTM) et 4 tâches d'apprentissage (classification binaire, régression continue, classification multiclasse, régression ordinale). Les modèles ordinaux et binaires ont performé significativement au-dessus du niveau de base et des deux autres tâches avec une erreur moyenne absolue macro entre 1,436 et 0,767 et une exactitude balancée de 58% à 73%. Les résultats montrent l'effet prépondérant du débalancement des données sur la performance prédictive et soulignent que les mesures n'en tenant pas compte surestiment systématiquement la performance.

Cette analyse ancre une série de considérations plus générales quant à l'utilisation de l'intelligence artificielle en santé. En particulier, l'évaluation de la valeur clinique de solutions d'apprentissage automatique présente des défis distinctifs en comparaison aux traitements conventionnels. Le rôle grandissant des technologies numériques en santé mentale a des conséquences sur l'autonomie, l'interprétation et l'agentivité d'une personne sur son expérience.

Mots-clés : santé numérique, soi quantifié, intelligence artificielle, régression ordinale, prévision, explicabilité, échelle clinique, apprentissage automatique, débalancement de classes

Abstract

Digital phenotyping leverages the numerous sensors of smartphones (e.g., accelerometer, GPS, Bluetooth, call metadata) to measure daily human behavior without interference and link it to psychiatric symptoms and mental health indicators. Machine learning is an integral component of processing raw signals into intelligible information for clinicians. This approach emerges from a will to characterize symptom profiles and their temporal variations at an individual level.

This project consisted in predicting mental health variables (e.g., stress, mood, sociability, hallucination) up to seven days in the future from smartphone data for patients with a diagnosis of schizophrenia. The CrossCheck dataset, which has a sample of 62 participants, was used. It includes 23,551 days of phone sensor data with 29 features, and 6364 mental state self-reports on 4-point ordinal scales.

Ordinal predictive models were used to generate discrete predictions that can be interpreted using the guidelines from the clinical data collection scale. In total, 240 machine learning models were trained, i.e., combinations of 10 mental health variables, 3 forecast horizons (same day, next day, next week), 2 algorithms (XGBoost, LSTM), and 4 learning tasks (binary classification, continuous regression, multiclass classification, ordinal regression). The ordinal and binary models performed significantly better than the baseline and the two other tasks with a macro-averaged mean absolute error between 1.436 and 0.767 and a balanced accuracy between 58% and 73%. Results showed a dominant effect of class imbalance on predictive performance and highlighted that metrics not accounting for it lead to systematic overestimation of performance.

This analysis anchors a series of broader considerations about the use of artificial intelligence in healthcare. In particular, assessing the clinical value of machine learning solutions present distinctive challenges when compared to conventional treatments. The growing role of digital technologies in mental health has implication for autonomy, sense-making, and agentivity over one's experience.

Keywords: digital health, quantified self, artificial intelligence, ordinal regression, forecast, explainability, clinical scale, machine learning, class imbalance

Table des matières

Liste des tableaux

Corps du texte

Article

Liste des figures

Corps du texte

Article

Liste des sigles et des abréviations

ACC : exactitude (*accuracy*)

BACC : exactitude balancée (*balanced accuracy*)

DSM : *Diagnostic and Statistical Manual*

GBDT : boosting de gradient avec arbres décisionnels (*gradient boosted decision trees*)

GPS : *global positioning system*

LSTM : *long short-term memory*

MAE : erreur absolue moyenne (*mean absolute error*)

MAMAE : moyenne macro de l'erreur absolue moyenne (*macro-averaged mean absolute error*)

RMSE : erreur quadratique moyenne (*root mean squared error)*

RNN : réseau récurrent de neurones (*recurrent neural network*)

ROC AUC : aire sous la courbe d'efficacité du récepteur (*receiver operating characteristic area under the curve)*

SHAP : *SHapley Additive exPlanations*

XGBoost : *eXtreme Gradient Boosting*

Remerciements

Pour commencer, je remercie Pierre pour l'ensemble du support offert au cours de ma maitrise et de mon cheminement Honor au baccalauréat. Plonger dans l'apprentissage automatique pour la première fois au sein d'un laboratoire dynamique et stimulant aura marqué mon parcours. Tu m'as guidé et accompagné à travers le monde scientifique.

Je veux remercier Guillaume Dumas et Michaël Chassé, responsables du programme de médecine computationnelle, pour leur enseignement riche et terre à terre ainsi que les membres de ma cohorte: Johann, Ghazaleh, Vincent, pour les cours dynamiques.

Merci à mes amis et collègues pour les échanges qui ont nourri mes réflexions et m'ont motivé à avancer : Thibault, Rose, Shivam, Alexia, Alexei, Marie-Odette, Julien, Carlos, Elijah, Stefan.

Merci à mon ami Clément de m'avoir donné la poussée dont j'avais besoin pour me lancer en programmation. Merci à mon oncle Gilles d'avoir partagé son amour des probabilités et d'avoir rendu les mathématiques cool.

Je remercie chaudement mes parents Chantal et Clément, toujours disponibles pour un appel, pour leurs bons conseils et leur patience indéfectible.

Finalement, un grand merci à Ioana, que j'aime, d'être présente tous les jours, les bons comme les moins bons, et d'illuminer mon quotidien.

Chapitre 1 – Contexte théorique

1.1 Psychiatrie et santé mentale

1.1.1 Définition du trouble mental

Le concept de maladie mentale réfère à une source de dysfonction cognitive, émotive, comportementale ou sociale entrainant un mal être ou de la détresse (*Diagnostic and Statistical Manual of Mental Disorders*, 2013). La maladie se manifeste à travers des symptômes, soit des pensées, des sentiments, des comportements ou des expériences subjectives dysfonctionnels. D'autre part, la santé mentale est associée à un bon fonctionnement et une capacité à surmonter les tensions normales de la vie (Organisation mondiale de la Santé, 2021). Elle implique une capacité à s'adapter en dépit de symptômes plutôt que simplement leur absence. Ces conceptions de la maladie et de la santé mentale sont ancrées dans les travaux du psychiatre Emil Krapelin réalisés au 19^{e} siècle. Il crée un système de classification de la maladie mentale selon le syndrome, c'est-à-dire un schéma de symptômes concomitants (Diefendorf & Kraepelin, 1907). Dans la même lignée, la première édition du Diagnostic and Statistical Manual (DSM; 1952), soit l'ouvrage de référence en psychiatrie, est publiée en 1952. Il contient des définitions de *troubles mentaux* composées de listes de symptômes servant de critères diagnostiques. Un trouble mental peut à la fois décrire une condition passagère ou durable. L'appellation *troubles mentaux graves* (*serious mental illness*) regroupe des conditions qui persistent dans le temps et qui sont associées avec des conséquences importantes, soit la dépression majeure, la schizophrénie, le trouble bipolaire, le trouble obsessif compulsif, trouble de panique, stress post-traumatique et la personnalité borderline (National Institute of Mental Health, 2023). Le DSM est d'abord nosologique; il établit un langage standardisé facilitant la pratique clinique et l'étude des troubles mentaux (Kendler, 2009).

Malgré les révisions, l'utilisation du DSM regroupe sous un même diagnostic des gens avec des troubles aux manifestations hétérogènes (Hyman, 2010). Simultanément, les taux de comorbidité élevés parmi les patients suggèrent un manque de spécificité du système de classification. Depuis la publication initiale du DSM, les différentes révisions ont tenté de répondre à ce problème en passant de considérations binaire « présence ou absence » à des dimensions continues allant de

normal à pathologique et en incluant des spécifications ou des sous-types aux diagnostics (Maser & Patterson, 2002; Regier et al., 2013).

En plus de la variabilité au sein d'une catégorie diagnostique, la gravité des symptômes d'un individu et leurs conséquences sur son fonctionnement varient dans le temps. En particulier, les troubles mentaux graves peuvent être conçus comme une alternance de périodes de rémission partielle et de dysfonction symptomatique à différentes échelles temporelles (Burcusa & Iacono, 2007; Emsley et al., 2013). La notion de *dynamiques temporelles* décrit les patrons de symptômes propres à une personne (Nelson et al., 2017) et la *trajectoire* réfère à une tendance vers l'amélioration ou la dégradation du fonctionnement (Fojo et al., 2017). Au niveau pratique, il existe une tension entre les systèmes de classification basés sur des différences interindividuelles et les traitements cherchant à entrainer des changements au niveau intra-individuel (Wright & Woods, 2020).

Les critiques envers les systèmes de classification ne sont pas seulement théoriques; plusieurs chercheurs déplorent que ces faiblesses aient freiné le progrès dans la compréhension des troubles mentaux et de l'amélioration des traitements (Insel, 2015). Puisque l'étalon-or en recherche psychiatrique et médicale est l'essai clinique randomisé, il est difficile d'observer des effets entre des groupes basés sur le diagnostic et constitués de gens aux défis dissimilaires (Molenaar & Campbell, 2009; Richters, 2021). Les systèmes de classification avaient initialement des visées descriptives, mais leur valeur prédictive quant à l'efficacité d'un traitement s'avère limitée.

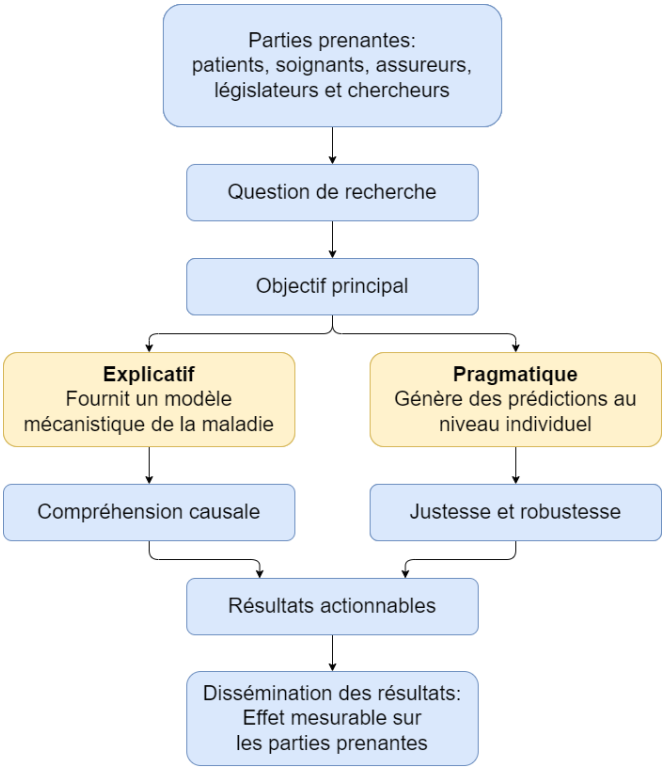
En réponse aux limitations des approches existantes, la psychiatrie pragmatique propose que la recherche et la pratique clinique soient guidées par leur finalité. Selon cette perspective, un modèle descriptif ou mécanistique d'un trouble est utile seulement s'il permet de formuler des prédictions justes, d'ajuster les soins et d'améliorer l'état de santé du patient (Paulus, 2017). Cela est cohérent avec l'école de philosophie pragmatique selon laquelle plusieurs explications d'un phénomène peuvent être valides et qu'elles devraient être jugées selon leur utilité et non de leur capacité à décrire une réalité de manière exacte.

Ailleurs en recherche biomédicale, les termes *pragmatique*, *pratique* et *écologique* peuvent servir à décrire un type d'étude qui s'intéresse à des phénomènes ou à l'évaluation de traitement dans le milieu de vie en opposition au contexte contrôlé de recherche randomisée, lui qualifié *d'explicatif* (Patsopoulos, 2011; Schwartz & Lellouch, 1967). En psychiatrie, le pragmatisme prend un sens un peu différent et suggère principalement de diminuer le rôle du diagnostic dans la recherche, compte tenu de sa faiblesse prédictive, et de recentrer le domaine sur l'évaluation de la santé des participants (Steele & Paulus, 2019). En 2017, Paulus appelle à une psychiatrie pragmatique fondée sur les preuves (*evidence-based pragmatic psychiatry*) (Paulus, 2017). Comme point de départ, il note que les différentes parties prenantes (patients, soignants, assureurs, législateurs et chercheurs) sont généralement peu alignées dans leurs objectifs qui incluent entre autres d'innover les soins, de minimiser les coûts, d'offrir des traitements efficaces et de prioriser les problématiques (Figure 1). Néanmoins, l'auteur identifie qu'elles convergent vers le besoin d'explications et de prédictions au niveau individuel permettant une prise d'action clinique. On note une réconciliation des études explicatives et prédictives, qui peuvent toutes deux avoir une valeur pragmatique dans la mesure où elles entrainent une amélioration des soins.

1.1.2 Soins en santé mentale

Aujourd'hui, les soins psychiatriques s'organisent autour de rencontres périodiques en contexte médical. Selon une étude de Chiu et al. (2018) réalisée en Ontario, parmi les 14,4 % de la population ayant nécessité des soins de santé mentale, une personne a vu son omnipraticien 3 à 6 fois durant 2014. Une autre étude rapporte que les rendez-vous de suivi pour soins de santé mentale incluent seulement 15 minutes pour l'évaluation du risque en moyenne (Cohen et al., 2019). Le psychiatre décide alors du traitement ou de ses ajustements. Une étude rapporte que dans seulement 15 sur 70 cas des psychiatres ont adéquatement détecté une dégradation de l'état du patient (Hatfield et al., 2010). Après une première hospitalisation aux soins psychiatriques, plus de 12 % des patients ont été réadmis deux fois ou plus au Québec en 2018 selon l'Institut canadien d'information sur la santé (Institut canadien d'information sur la santé, 2023).

Figure 1. *Cadre pour étude pragmatique en psychiatrie*



Note. Inspiré de *Evidence-Based Pragmatic Psychiatry—A Call to Action.* Paulus (2017)

Le paradigme de soins actuel peine à prendre en charge les individus et à intervenir de manière préventive. D'une part, le système de santé opère avec des ressources restreintes et tente de desservir un grand nombre de demandeurs de soins. Le nombre limité de professionnels rend le rôle central de l'évaluation et de l'intervention un à un insoutenable. D'autre part, organiser les soins en rencontres périodiques ne permet pas de rendre compte des caractéristiques particulières, des dynamiques temporelles et de la trajectoire du trouble d'une personne (Chiauzzi & Wicks, 2021). Cela pose un problème sachant que plusieurs diagnostics nécessitent la persistance d'un trouble pour une période, nécessitant plus d'une rencontre avant qu'un diagnostic soit établi et laissant la personne vulnérable. Ensuite, la variation au sein d'une catégorie diagnostique rend la sélection de traitement pour une personne donnée est ardue et souvent requiert plusieurs tentatives. De longues périodes d'essais sont nécessaires pour évaluer les effets souhaités, les effets secondaires, puis calibrer le traitement au fil des variations intra-individuelles (Insel, 2015). Autrement, l'information obtenue en cabinet médical à une généralisabilité limitée au vécu quotidien de la personne. En psychiatrie, la majorité des

observations du clinicien est obtenue de manière rétrospective et peut être limitée par les capacités mnésiques du patient (pouvant être affectées par la condition) ainsi qu'un ensemble de biais cognitifs tel que la désirabilité sociale ou l'effet de récence (Rogler et al., 2001).

Les approches centrées sur le patient (*patient-centered care*), ou patient-partenaire, considèrent la personne experte de sa condition idiosyncrasique et l'encouragent à participer activement aux soins (Greene et al., 2012). En facilitant les échanges proactifs à propos de la santé mentale, les professionnels de la santé pourraient être mieux informés de la trajectoire du patient. Cette pratique recentre la santé mentale vers une notion de qualité de vie subjective plutôt qu'une grille statique de symptômes. Ces objectifs sont cohérents avec la conception pragmatique. Néanmoins, la portée de ces efforts demeure limitée par les rencontres peu fréquentes et le manque de données objectives, soulignant le besoin de nouvelles méthodes de collecte de données rendant compte de l'évolution de l'état du patient.

1.1.3 Évaluation du patient

Parmi les professionnels offrant des soins de santé mentale, la méthode d'évaluation du risque principale était l'examen formel, l'examen informel, ou l'entrevue structurée pour 80% des répondants (Cohen et al., 2019). Deux tiers rapportent l'entrevue structurée comme étant la méthode la plus fiable, contre 5 % pour les échelles cliniques. L'entrevue clinique implique des questions structurées ou non structurées, puis une appréciation subjective et qualitative de la santé mentale du patient selon ses réponses, son comportement et son attitude au cours de l'interaction. Les notes des rencontres sont colligées afin de rendre compte de l'évolution du patient. Pour pallier la subjectivité de l'approche, l'Association Psychiatrique Américaine suggère l'utilisation d'échelles standardisées (American Psychiatric Association, 2015). Elles servent à évaluer quantitativement des construits définis avec précision et permettent d'observer plus clairement les variations d'une rencontre à l'autre. Les propriétés psychométriques et l'usage prévu sont établis avec des études de validation, permettant une utilisation standardisée et facilitant la mise en relation du contexte clinique et des résultats de recherche (p.ex. : évaluation de l'efficacité d'un traitement). Cependant, un sondage de plus de 300 psychiatres a révélé que plus de 80 % d'entre eux n'utilisaient pas d'échelles cliniques de manière courante pour traiter la

dépression (Zimmerman & McGlinchey, 2008). Les raisons données principales incluent un doute envers leur efficacité clinique, le temps requis et un manque de formation.

Le concept de *soins fondés sur la mesure* (*measurement-based care*) fait la promotion de la collecte routinière (plutôt que ponctuelle) d'information structurée sur l'état de patients (Lewis et al., 2019). Cette pratique a pour but d'impliquer le patient, de suivre et d'évaluer le traitement et de guider le clinicien dans l'ajustement idiosyncrasique du traitement. Les soins fondés sur la mesure peuvent être considérés comme une composante des soins centrés sur le patient et un élément nécessaire à une pratique pragmatique. D'ailleurs, différentes études rapportent l'amélioration de l'état de santé grâce à ce type d'approche (Trivedi, 2009).

Néanmoins, les soins organisés en rencontres périodiques demeurent une barrière majeure au suivi de l'état d'un patient. Les *soins fondés sur la mesure à distance* (*remote measurement-based care*) étendent le concept de soins fondés sur la mesure à la collecte de données à l'extérieur du laboratoire ou du contexte médical (Chiauzzi & Wicks, 2021; Goldberg et al., 2018). Par exemple, l'équipe clinique peut administrer des échelles cliniques par téléphone ou en ligne entre les rencontres régulières. Cette pratique augmente la résolution temporelle des observations et permet un suivi plus précis de la trajectoire de la personne. Une revue de ces approches conclut qu'en moyenne cette pratique contribue à une diminution cliniquement significative des symptômes (Goldberg et al., 2018).

Similairement, le domaine de la psychologie possède une riche littérature portant sur les mesures écologiques. La méthode d'échantillonnage d'expériences (*experience sampling method*) (Larson & Csikszentmihalyi, 1983) consiste à autoévaluer l'expérience d'une situation au moment où elle a lieu (p. ex. : noter une situation ayant entrainé des pleurs et l'expérience subjective associée). Plus récemment, les évaluations écologiques momentanées (*ecological momentary assessment*) (Shiffman et al., 2008) servent à mesurer de manière spontanée les comportements ou l'état d'une personne selon un intervalle de temps fixe ou aléatoire. La différence principale étant la contingence de l'expérience et de son évaluation, ces méthodes sont à toute fin pragmatique jugées équivalentes (Birk & Samuel, 2020; Insel, 2017). Aujourd'hui, les moyens technologiques (sites web, des applications pour téléphone, montre intelligente) facilitent ces méthodes de

collecte de données. La méthode d'échantillonnage d'expériences et l'évaluation écologique momentanée ont toutes deux montré une utilité clinique (Kwasnicka et al., 2021; Morgenstern et al., 2014). Cependant, elles demeurent principalement limitées par l'attrition de patients fatigués par les mesures répétées fréquentes (Goldberg et al., 2018; Torous et al., 2019).

La *psychiatrie computationnelle* est proposée pour développer des modèles mathématiques de la santé mentale intégrant une variété de sources de données (p. ex. : imagerie cérébrale, médias sociaux, données épidémiologiques) afin de représenter la complexité du phénomène (Insel, 2017; Paulus et al., 2016). Plus précisément, il est question de *modèles explicatifs* qui permettraient une meilleure compréhension mécanistique des troubles et de *modèles prédictifs* supportant la prise de décision clinique (p. ex. : diagnostic, pronostic, sélection de traitement). Dans les deux cas, les autoévaluations écologiques remplissent le rôle critique de ponts entre des mesures plus objectives et l'expérience subjective de bien-être ou de mal-être de l'individu. À travers l'augmentation de la quantité de données et de la complexité des modèles, la psychiatrie computationnelle tente de mieux rendre compte de variations intra-individuelles et interindividuelles des symptômes et des états mentaux (Hitchcock et al., 2022).

1.2 Le phénotypage numérique

1.2.1 Le téléphone comme instrument de mesure

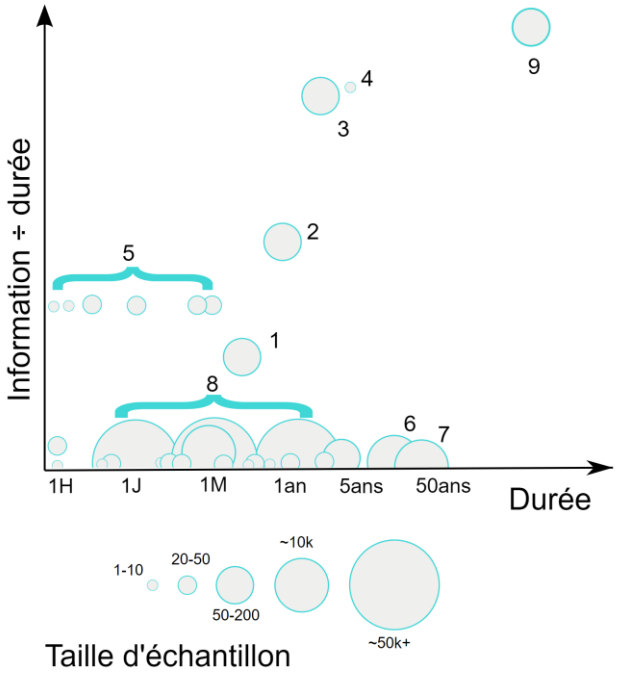
L'invention et la marchandisation du téléphone portable ont radicalement changé les liens sociaux en gardant interconnectés ses usagers. Depuis le début des années 2000, le nombre d'abonnements pour téléphone mobile par 100 personnes est passé de 52 à 108 (International Telecommunication Union, 2022). Progressivement, les téléphones mobiles ont été équipés de capteurs supplémentaires et ont gagné accès aux réseaux Wi-Fi, Bluetooth et de Global Positioning System (GPS) (Wikipedia contributors, n.d.). Depuis, les appareils mobiles n'ont cessé de gagner en fonctionnalités (p. ex. : caméra, réseaux sociaux, assistance routière, mode de paiement), devenant un outil central à la vie moderne. On parle alors de téléphone intelligent.

Selon Eagle & Pentland (2006) du MIT Human Dynamics Lab, la littérature scientifique sur les relations sociales et les dynamiques de groupe a toujours été limitée par sa méthode de collecte

de données principale, le questionnaire structuré. Ils souhaitaient obtenir des données riches et continues de vraies interactions sociales, à l'extérieur du laboratoire, notamment afin d'entrainer des modèles d'apprentissage automatique à prédire les dynamiques de groupe. En 2003, ils lancent l'étude *Reality Mining*, la première de son genre, au cours de laquelle les données de téléphone mobile (réseau téléphonique, appels, SMS, Bluetooth, GPS, etc.) de 100 participants sont recueillies pendant 9 mois (Eagle & Pentland, 2006). En 2008, une seconde étude, *Social Evolution*, inclut 80 étudiants universitaires partageant un même dortoir (Dong et al., 2011). Les analyses déterminent des relations entre les données du téléphone et des phénomènes contagieux incluant l'attitude par rapport à l'exercice physique, les symptômes de toux, les opinions politiques, les goûts musicaux et la performance académique. En 2010, leur étude de plus grande envergure, *Friends and Family*, suit 130 adultes pendant 12 mois (Aharony et al., 2011). À l'aide de données de téléphone, d'autoévaluations quotidiennes et de questionnaire mensuels, un grand nombre de phénomènes est étudié: interactions sociales, activité physique, habitude d'achat, statut financier, personnalité.

Les études du laboratoire de Pentland sont les premières à étudier finement des comportements naturels grâce à des données à haute résolution temporelle (Figure 2). Leurs travaux tracent le schéma pour les études se servant du téléphone comme instrument de mesure qui suivront. Du côté expérimental, on montre l'utilité de mesures externes (p. ex. : questionnaire, résultat académique, poids) pouvant être mises en relation avec les données du téléphone. Les analyses emploient généralement des modèles mathématiques ou des algorithmes d'apprentissage automatique permettant d'extraire du signal pertinent parmi un grand volume de données. Au niveau technique, ils déterminent l'architecture de système permettant de connecter un serveur de recherche aux téléphones de participants et de récolter les mesures provenant des capteurs et de l'utilisation de l'appareil.

Figure 2 *Quantités de données générées par les études en sciences sociales*



Note. Revues qualitatives d'études et de jeux de données en sciences sociales en fonction de la durée, de la quantité d'information générée et du nombre de participants. 1) Reality Mining, 2) Social Evolution, 3) Friends and Family, 4) Pionniers des "données riches", 5) Études avec badges sociométriques, 6) Midwest field station, 7) Framingham Heart Study, 8) Données d'appels téléphoniques, 9) Omniscience.
Inspiré de *Social fMRI: Investigating and shaping social mechanisms in the real world*. Aharony et al. (2011).

1.2.2 Définition du phénotypage numérique

L'objectif de quantification du comportement n'est pas unique aux sciences sociales et ses défis techniques conséquents en font une méthode de recherche multidisciplinaire. Le domaine en informatique de *pervasive, unobtrusive, and ubiquitous computing* traite des objets du quotidien interconnectés et l'intersection avec des applications en santé est intitulée *ubiquitous health* ou *uHealth* (traduis *santé intégrée*). Au sens large, la captation personnelle et la santé intégrée incluent l'analyse de comportements sur les réseaux sociaux et de signaux de montre intelligente en plus des données de téléphone intelligent. L'expression *mobile health* ou *mHealth* (traduis *santé mobile*) s'inscrit dans la santé intégrée et se concentre sur l'utilisation de données de téléphone mobile. Finalement, le terme *phénotypage numérique* est prévalent en sciences psychiatriques pour décrire des méthodes variées de santé intégrée et de santé mobile (Mohr et al., 2017).

Dans une série d'articles en 2015 et 2016, Onnela, Torous et collègues plaident pour l'utilisation du téléphone comme instrument de mesure pour la recherche et la pratique clinique en psychiatrie (Onnela & Rauch, 2016; Torous et al., 2015; Torous & Baker, 2016). Le phénotypage numérique (*digital phenotyping*) est défini comme « le processus de quantification du phénotype humain, au niveau individuel, moment par moment, dans l'environnement de vie quotidienne par l'entremise du téléphone intelligent » (Onnela & Rauch, 2016). Ces publications font suite à des études expérimentales de 2013 et 2014 utilisant le téléphone afin d'étudier et de développer des interventions pour les personnes avec un trouble bipolaire, de schizophrénie, de dépression majeure, d'alcoolisme, ou d'idéation suicidaire (Ben-Zeev et al., 2014; BinDhim et al., 2015; Gustafson et al., 2014; Pramana et al., 2014). Venant lui-même du domaine de l'étude des dynamiques de groupe, Onnela cite le travail de Pentland comme étant d'importance majeure (Torous et al., 2015). Par ailleurs, le phénotypage numérique s'inscrit explicitement comme l'extension des méthodes d'échantillonnage d'expérience et d'évaluation écologique momentanée (Onnela & Rauch, 2016). En plus des questionnaires d'autoévaluation nécessitant une participation *active*, le phénotypage numérique capture les comportements de manière *passive* grâce aux capteurs de l'appareil.

Peu avant, Jain et al. proposent en 2015 le terme *phénotype numérique* (*digital phenotype*) qui se veut la signature numérique de la manifestation d'une maladie biologique (Jain et al., 2015). Ce concept suit la notion de *phénotype étendu* proposée par Richard Dawkings en biologie évolutionniste (Dawkins, 1982). Les technologies numériques (p. ex. : téléphone intelligent, réseaux sociaux, montre intelligente) seraient des médiums où le phénotype de la maladie (p. ex. : rhume, insomnie, idéation suicidaire, bipolarité) s'exprimerait. Le concept de phénotype numérique est donc davantage biologisant. Il demeure peu utilisé.

Dans leur article de 2016, Onnela et Rauch mentionnent le phénotypage numérique comme étant distinct du phénotype numérique sans fournir de justification (Onnela & Rauch, 2016). Dans un article de perspective de 2017, Insel appelle à l'utilisation du phénotypage numérique sans le distinguer du phénotype numérique (Insel, 2017). Les deux approches convergent vers l'objectif de développer des mesures valides de l'activité humaine à l'extérieur du laboratoire (Birk & Samuel, 2020). Insel (2017) déplore que les patients soient étiquetés avec leur diagnostic et voit

le phénotypage numérique comme l'instrument permettant de rendre compte des variations inter- et intra-individuelles. Il souhaite réaliser les objectifs des soins fondés sur la mesure et fournir de l'information actionnable permettant de livrer de meilleurs soins. Sa conception est fondamentalement pragmatique.

1.2.3 Mesurer le comportement

Le téléphone intelligent est l'instrument de collecte de données principal du phénotypage numérique. Bien qu'il existe différents profils d'utilisations, les gens sont généralement motivés à transporter leur téléphone et à interagir avec leur appareil quotidiennement. À l'inverse, les appareils de collecte de données spécialisés risquent de diminuer l'engagement des participants (p. ex. : oublier l'appareil, décider de ne pas le porter) ou de dénaturer le phénomène étudié (p. ex. : saillant, encombrant). De plus, déployer des appareils spécialisés à l'extérieur du contexte de recherche serait éventuellement un défi. Vaizman (2018) propose 4 principes pour acquérir des mesures écologiques et authentiques: 1) utilisation naturelle de l'instrument, 2) placement non contraignant de l'instrument, 3) environnement naturel, et 4) comportement naturel. L'ubiquité et l'aspect discret du téléphone intelligent en font une plateforme idéale pour le phénotypage numérique.

Les données collectées grâce au téléphone intelligent peuvent être divisées en deux types: passive et active. Les données passives proviennent des capteurs (p. ex. : accéléromètre, GPS) et des métadonnées d'utilisation (p. ex. : appels, déverrouillage de l'écran) de l'appareil. Le Tableau 1 présente une synthèse des données passives pouvant être collectées basée sur des articles de revues (Lee et al., 2023; Rohani et al., 2018). Les données actives résultent d'une participation délibérée de l'utilisateur. Cela inclut principalement des questionnaires et des échelles cliniques autorapportés, mais aussi des mesures physiologiques (p. ex. : poids), tenir un journal personnel ou des tests cognitifs. Finalement, certaines études incluent des mesures externes obtenues sans l'aide du téléphone provenant d'évaluations cliniques ou de dossier médical.

Tableau 1 *Données passives*

Capteur	Attributs de bas niveau	Attributs de haut niveau
Accéléromètre	Mouvement de l'appareil, gravité	Activité physique, moyen de transport, sédentarité, tremblements
GPS	Emplacement	Lieux visités, temps passé au domicile, sédentarité, absentéisme
Gyroscope	Orientation de l'appareil	Activité physique, tremblements
Bluetooth	Proximité d'appareil Bluetooth	Contacts sociaux, lieux visités
Baromètre	Pression atmosphérique	Altitude
Capteur lumineux	Luminosité	Qualité du sommeil
Microphone	Audio	Conversations ambiantes, interactions sociales
Métadonnées téléphoniques	Appels, SMS, contacts	Interactions sociales, réciprocité
Gestionnaire d'applications	Applications utilisées, installées	Qualité du sommeil, utilisation de réseaux sociaux
Signal cellulaire	Position par rapport aux tours téléphoniques	Mobilité
Batterie	Niveau de charge	Sédentarité
État de l'écran	Ouverture, fermeture, luminosité	Qualité du sommeil, temps passé sur l'appareil
Verrouillage de l'appareil	Ouverture, fermeture	Qualité du sommeil, concentration
Écran tactile	Interaction avec l'écran	Tremblements
Caméra	Mouvement des yeux	Concentration, exploration visuelle
Capteur électrodermal	Conductance de la peau	Stress physiologique
photoplethysmogramme	Rythme cardiaque	Stress physiologique

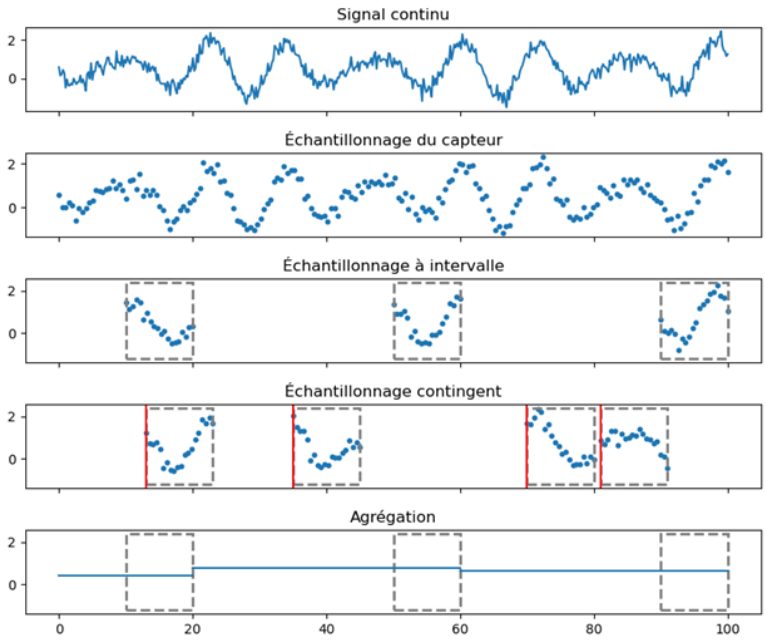
La collecte de données génère un large volume de données passives sans utilité immédiate ainsi qu'un nombre plus restreint de mesures actives avec une certaine validité préétablie. Des modèles mathématiques et des algorithmes d'apprentissage automatique servent à extraire des relations complexes entre les données passives et actives afin de générer de l'information pertinente.

1.2.4 Acquisition de mesures

Le phénotypage numérique offre une opportunité unique de mesurer les comportements en milieu naturel à haute fréquence et en continu. Cependant, la collecte de données passives et actives a ses limites pratiques. Une application mobile faisant fréquemment des lectures des capteurs du téléphone taxe la batterie, diminue la performance, écourte la durée de vie de l'appareil et les données générées peuvent rapidement excéder la mémoire du téléphone. Des effets négatifs marqués sur l'expérience d'utilisation du téléphone risquent de causer de l'attrition (p. ex. : fermer ou désinstaller l'application) ou de dénaturer le phénomène étudier en enfreignant les critères 1, 2 et 4 de Vaizman (2018). Similairement, des mesures actives trop fréquentes peuvent être lourdes pour l'utilisateur, soit la limitation principale des mesures autorapportées et potentiellement interagir avec le phénomène d'intérêt (Torous et al., 2019). Par exemple, remplir des questionnaires sur le stress à répétition pourrait influencer le niveau de stress réel ou perçu (Fortney et al., 2017).

Afin de diminuer la pression sur le téléphone et les risques de vie privée, la collecte de données utilise un processus d'échantillonnage et d'agrégation sur l'appareil avant d'effectuer un transfert vers un serveur centralisé (Figure 3). La fréquence d'échantillonnage du capteur (p. ex. : 100Hz) constitue la résolution maximale, soit celle que l'on obtiendrait avec des mesures continues. L'application de phénotypage numérique fait un échantillonnage d'observation en fonction de deux paramètres: la fréquence et la durée de la période d'échantillonnage (p. ex. : 20 secondes à 100Hz toutes les 5 minutes). Les échantillons peuvent agrégés afin de compresser l'information et réduire le volume de données à transférer (p. ex. : min, max, mean, std). L'échantillonnage peut être à intervalle régulier ou contingent à un signal.

Figure 3 *Processus d'échantillonnage d'un capteur*



D'autres types de transformations sont utilisées afin d'éviter de transférer des données sensibles vers le serveur de recherche. Le plus simple est d'offusquer les données en remplaçant les numéros de téléphone par des identifiants uniques ou en centrant les coordonnées GPS sur un 0 arbitraire par exemple. À l'autre extrême, l'application peut extraire de l'information provenant de données trop sensibles pour être transférées. Par exemple, on peut classifier le signal audio provenant du microphone comme étant une conversation ou non. Ces transformations définissent la granularité de données qui seront disponibles pour les analyses et il est généralement impossible de reconstruire les données brutes dont elles proviennent.

1.3 L'apprentissage automatique

1.3.1 Brève histoire l'intelligence artificielle

Depuis les débuts de l'informatique, les mathématiciens ont réfléchi aux conséquences profondes des machines et des approches algorithmiques. Alan Turing, souvent nommé le père de l'informatique, a publié une série de travaux séminaux incluant un article intitulé *Can machines think* en 1950 (Turing, 1950). Cela a mis la table pour les notions *d'intelligence artificielle* proposée en 1955 par John McCarthy (McCarthy et al., 2006) et *d'apprentissage automatique*

(*machine learning*) en 1959 par Arthur Samuel (Samuel, 1959). L'intelligence artificielle est un champ d'étude dont l'objectif est de créer des machines capables de répliquer l'intelligence humaine. L'apprentissage automatique est l'approche la plus populaire en IA et se base sur des algorithmes permettant de s'améliorer sur une tâche en apprenant à partir d'exemples. L'*apprentissage profond* (*deep learning*) s'inscrit dans l'apprentissage automatique et se concentre sur les modèles utilisant plusieurs couches de neurones. Le premier réseau de neurones a été créé en 1951 par Marvin Minsky (Minsky, 1961), mais le potentiel de l'approche est réalisé à partir de 2010 grâce à l'amélioration des performances de calcul et de la quantité de données disponibles permettant de créer des modèles « profonds ».

1.3.2 Termes clés et formalisations mathématiques

L'apprentissage automatique est un sujet d'étude multidisciplinaire trouvant ses racines dans l'informatique et les statistiques. Un ensemble de disciplines, incluant les neurosciences, l'économétrie, la physique et les sciences biomédicales, contribuent à la recherche théorique et aux applications pratiques du domaine. Cependant, cette diversité de contributions amène à une littérature chargée de jargons et de formalismes mathématiques hétérogènes. Pour ce mémoire, la terminologie et la notation sont adaptées et traduites du livre *Deep Learning* de Goodfellow, Bengio et Courville (Goodfellow et al., 2016) et du *Machine Learning Glossary* par Google (Google, 2023).

Une *entrée* (*input*) est un vecteur de *d* dimensions où chacune décrit un attribut d'un objet. Un *attribut* (*feature*) est une caractéristique d'un objet représentée numériquement. Une *cible* ou *sortie* (*output*) est une valeur numérique que le modèle doit produire. On qualifie de « vraie » la valeur cible jugée correcte. Une *prédiction* est une valeur numérique produite par le modèle. C'est une estimation de la valeur de la cible. Un *exemple* est une paire contenant une entrée et une cible qui décrivent un objet. Un *ensemble d'entrainement* (*training set*) contient *n* exemples qui sont présentés au modèle lors de son entrainement. L'*apprentissage supervisé* est un paradigme selon lequel le modèle est entrainé à partir d'*exemples* où la sortie est connue et le succès est défini par la capacité à bien prédire la *sortie*.

Un *modèle* (*model*) est une fonction recevant une entrée et produisant une prédiction. Un *modèle entrainé* (*trained model*) se distingue par la valeur de ses *paramètres* (p. ex. : les bêtas d'une régression linéaire). Deux modèles ayant appris des choses différentes possèdent des valeurs de paramètres distinctes. Le *processus d'apprentissage* ou *d'entrainement* consiste à minimiser la *fonction de coût* (*loss function*). On cherche les paramètres produisant un modèle entrainé avec le coût minimum.

Une *tâche d'apprentissage* ou *tâche* définit au sens large le type de prédictions que le modèle devrait générer (p. ex. : classification, régression). Le choix de la tâche influence plusieurs décisions méthodologiques concernant les exemples, l'algorithme, l'entrainement, etc. Une *mesure de performance* ou *mesure d'erreur* est une fonction servant à évaluer la qualité des prédictions par rapport aux sorties dans le contexte d'une tâche. Contrairement à la fonction de coût, la mesure de performance n'influence pas directement le processus d'entrainement.

Un *algorithme d'apprentissage* ou *algorithme* est un processus par lequel les paramètres d'un modèle sont mis à jour lorsque présenté avec un exemple. Différents algorithmes auront différentes structures et paramètres. Un algorithme à *capacité* élevée est capable de représenter une fonction complexe. Par exemple, un réseau de neurones une capacité supérieure à une régression logistique. Une *famille d'algorithmes* réfère à des algorithmes partageant des caractéristiques communes.

Les *données tabulaires* sont dans un format en deux dimensions où chaque rangée est un exemple et chaque colonne un attribut. Les *séries temporelles* ou *séquences* sont des données avec un format en trois dimensions où chaque rangée et colonne sont des exemples et des attributs et le troisième axe représente la dimension temporelle. Une séquence est composée de données tabulaires répétées. Une *prévision* (traduction de *forecast*) est une prédiction d'une valeur future. La *tâche de prévision* (*forecasting*) est une tâche d'apprentissage supervisé consistant à obtenir des prévisions justes. On note que la *tâche de prévision* est distincte et indépendante de la *tâche d'apprentissage*.

1.3.3 Développer un modèle d'apprentissage automatique

La démarche de l'apprentissage automatique diffère des statistiques fréquentistes qui ont longtemps dominé la littérature scientifique (p. ex. : sciences biomédicales, neurosciences, psychologie, santé publique). De manière plus large, l'intelligence artificielle gagne en adoption et a été employée dans au moins 4 % des études en santé et en télémédecine publiée dans Scopus en 2023 selon une analyse de Nature (Van Noorden & Perkel, 2023). Bien comprendre les différentes mentalités sous-tendant l'apprentissage automatique et les statistiques fréquentistes est nécessaire pour reconnaitre les conclusions scientifiques qu'elles permettent et porter un regard critique les conclusions d'études (Breiman, 2001).

Selon l'approche fréquentiste, la recherche consiste à développer une hypothèse, soit un modèle théorique, et le valider avec des observations empiriques. Le modèle théorique est formalisé par un modèle statistique que l'on applique (*fit*) aux données collectées et l'on obtient une mesure d'erreur. En parallèle, on évalue l'hypothèse nulle (l'absence d'effet) ou un modèle théorique établi sur les mêmes données. Un test statistique permet de déterminer lequel du modèle proposé ou de l'hypothèse nulle correspond le mieux aux données. Si le modèle statistique proposé est significativement « meilleur », on conclut que le modèle théorique correspondant explique mieux le phénomène. À l'inverse, si les modèles sont équivalents, on doit préférer le modèle le plus simple ou celui le mieux supporté par la science existante. On accorde une valeur descriptive / explicative du phénomène aux paramètres statistiques dérivés des données (p. ex. : coefficients de régression); ils expriment quantitativement des relations entre les variables du modèle. Le processus de recherche aboutit en la validation d'un modèle théorique permettant de comprendre et d'interpréter des phénomènes. Crucialement, les chercheurs proposent des modèles théoriques et statistiques basés sur leurs connaissances scientifiques, intuition, biais, etc. L'hypothèse doit être indépendante et précéder les données sinon la validité des analyses statistiques peut être remise en question (*hypothesizing after the results are known*; HARKing)(Breiman, 2001; Kerr, 1998; Molnar, 2022b).

À l'inverse, l'apprentissage automatique ne spécifie pas de modèle théorique préalablement et l'échantillon empirique d'entrées et sorties (X, Y) est fouillé de manière exhaustive afin de trouver le « meilleur » modèle ($\hat{h}$). Le défi central est de démontrer à partir des données disponibles que la performance du modèle est généralisable à d'autres données (dans le futur, d'autres participants, d'autres contextes, etc.). Le principe de *minimisation du risque empirique* (*empirical risk minimization*) indique que le modèle minimisant la fonction de coût a le plus de chance de bien performer sur de nouvelles données. La différence clé avec l'approche fréquentiste est que l'on partitionne les données disponibles afin d'appliquer les modèles à un ensemble d'entrainement, puis l'on calcule une mesure d'erreur (ou de performance) sur un ensemble d'évaluation indépendant. Le but est d'obtenir une estimation non biaisée de la performance du modèle sur de nouvelles données. Durant une étude, les chercheurs doivent rester aveugles à l'ensemble d'évaluation pour maintenir l'intégrité des résultats. La recherche se conclut par l'obtention d'un modèle générant des prédictions auxquelles on peut faire confiance (ou non). En apprentissage automatique, les paramètres du modèle sont jugés indicatifs de ce que le modèle a appris des données plutôt qu'une description « correcte » des propriétés du phénomène étudié. Bien que des modèles théoriques soient rarement spécifiés, l'ensemble des décisions méthodologiques des chercheurs reflètent leurs connaissances, intuition, biais, etc. (Varoquaux & Cheplygina, 2022). Les hypothèses dans les études d'apprentissage automatique portent généralement sur l'effet de décisions méthodologiques (p. ex. : tel algorithme améliore la performance) et non le phénomène prédit.

La conception pragmatique selon laquelle plusieurs explications peuvent être justes et que leur utilité les distingue concorde bien avec l'apprentissage automatique. Au lieu de valider un modèle statistique, l'étude a pour but d'évaluer l'efficacité de la méthode de fouille pour obtenir le meilleur modèle prédictif. La méthode de fouille contient plusieurs étapes dépendantes qui influencent la performance (p. ex. : transformation de données, création d'ensembles d'entrainement, algorithme d'apprentissage). Idéalement, pour arriver à des conclusions robustes, toutes les permutations d'étapes sont évaluées pour contrôler les interactions et cela à plusieurs reprises pour quantifier la part d'aléatoire (Varoquaux & Colliot, 2023). Cependant, les chercheurs doivent se concentrer sur une ou quelques parties du système à la fois pour contrôler

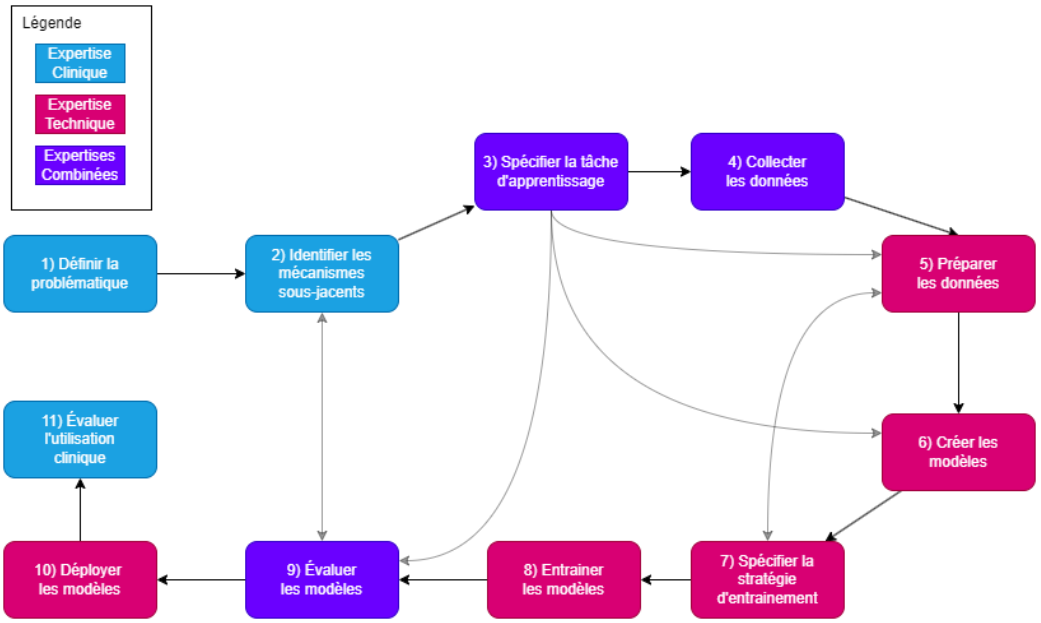
les coûts de calcul. De plus, la fouille peut être guidée par des contraintes du contexte d'application (taille, latence, complexité, interprétabilité, etc. du modèle) et le type de généralisabilité que l'on souhaite valider.

1.3.4 Développer un système d'aide à la décision

La majorité de la recherche académique porte sur l'amélioration de la performance prédictive des modèles. Cela inclut le développement d'algorithme, de transformations de données et de méthodes d'optimisation. C'est un effort louable, car une faible amélioration de performance peut avoir des effets bénéfiques majeurs à l'échelle d'une population. En santé, les modèles prédictifs pourraient supporter les cliniciens dans le processus diagnostique et l'évaluation du risque. Ultimement, l'objectif est d'améliorer le bien-être du patient, réduire le nombre d'événements cliniques et améliorer l'accès aux soins de santé.

En pratique, un modèle d'apprentissage automatique n'est qu'un seul morceau d'un système d'aide à la décision. D'abord, le modèle est développé hors ligne (*offline*), dans un contexte isolé du système. Des données historiques servent à entrainer le modèle et à évaluer sa performance prédictive. Une fois entrainé et jugé suffisamment performant, il est déployé dans le système d'aide à la décision (contexte *online*) et l'on surveille la qualité de ces prédictions (Huyen, 2022). Lorsqu'on évalue l'efficacité d'une intervention médicale, on valide qu'utiliser la règle de décision $h: X \to Y$ améliore les indicateurs de santé. Évaluer un système incorporant l'apprentissage automatique implique d'évaluer à la fois la fouille pour le prédicteur $\hat{h}$ et l'intervention réalisée selon ce prédicteur (c.-à-d., une forme de règle de décision). Le processus de développement doit considérer les interactions et les contraintes que le modèle prédictif entretient avec le système (Klement & El Emam, 2023). La Figure 4 présente un cycle de développement et les sections suivantes détaillent le développement d'une solution de phénotypage numérique à titre d'exemple.

Figure 4 *Cycle de développement de l'apprentissage automatique*



La recherche scientifique sur l'apprentissage automatique en santé se rend rarement aux étapes de *10) Déployer les modèles* et *11) Évaluaer l'utilisation clinique*. Ces étapes correspondent aux *Actionable outcomes* et *Measurable impact on stakeholders* de Paulus (2021) (section 1.1) qui sont essentielles à une approche pragmatique. Par ailleurs, ces étapes sont loin d'être linéaires et ensemble existent dans un cadre légal, un contexte social et une institution. Manquer à rendre compte de ces contraintes mène à la surestimation de la performance prédictive ou au développement de modèles non utilisables en pratique.

1.3.4.1 Définir la problématique

Un problème est posé en termes communs. Des indicateurs permettant d'évaluer le succès et de faire le suivi du problème sont identifiés.

Exemple. Réduire le nombre d'hospitalisations à l'urgence pour épisode psychotique. Le nombre de cas devrait être moindre dans la population suivie avec le système d'aide à la décision.

1.3.4.2 Identifier les mécanismes sous-jacents

Des experts énumèrent les actions ayant le potentiel de résoudre le problème.

Exemple. Accès aux soins, médication, services sociaux, thérapie.

1.3.4.3 Spécifier la tâche d'apprentissage

Chaque mécanisme est évalué afin de déterminer les bénéfices potentiels de l'apprentissage automatique. Face à plusieurs opportunités, il faut considérer la faisabilité et les bénéfices potentiels de chacune.

Exemple. Prédire le risque futur de symptômes, prédire l'efficacité d'un médicament.

1.3.4.4 Collecter les données

Un système est mis en place pour obtenir les données brutes qui serviront d'entrées et de cibles pour entrainer le modèle et d'indicateur de succès.

Exemple. Une application mobile et une plateforme de phénotypage numérique accompagné de dossiers médicaux électroniques.

1.3.4.5 Préparer les données

Les données brutes sont nettoyées, transformées, déidentifiées, etc. pour entrainer des modèles prédictifs. La création d'attributs (*feature engineering*) mobilise les connaissances humaines pour extraire l'information pertinente des données brutes et faciliter l'apprentissage du modèle. Un jeu de données est assemblé pour le développement du modèle.

Exemple. À partir des coordonnées GPS collectées, la distance parcourue durant la journée est calculée.

1.3.4.6 Créer les modèles

Un ou plusieurs algorithmes sont sélectionnés et la tâche est formalisée en tant que problème d'optimisation mathématique.

Exemple. On va évaluer les réseaux de neurones pour prédire un score de risque de la présence d'un symptôme la semaine prochaine.

1.3.4.7 Spécifier la stratégie d'entrainement

Les données sont divisées en ensembles d'entrainement, de validation et d'évaluation selon la généralisation souhaitée. Les ensembles d'entrainement et de validation servent à estimer la

performance qui serait atteignable sur l'ensemble d'évaluation. Le format du jeu de données est ajusté en fonction de l'algorithme évalué.

Exemple. Les données sont partitionnées selon l'axe temporel pour évaluer la prédiction d'état futur.

1.3.4.8 Entrainer les modèles

Plusieurs algorithmes sont entrainés itérativement en suivant la stratégie. Une méthodologie rigoureuse est nécessaire pour évaluer les différentes configurations systématiquement et trouver le meilleur modèle. Jusqu'à cette phase, le processus est encore très itératif; on s'assure que tout fonctionne au niveau méthodologique et technique.

Exemple. Des dizaines de réseaux de neurones sont entrainées pour trouver l'architecture la plus performante sur les données d'entrainement.

1.3.4.7 Évaluer les modèles

Les meilleures configurations sont utilisées pour entrainer les modèles finaux ayant du potentiel. Ensuite, les prédictions sur l'ensemble d'évaluation servent d'indicateurs de la performance. Retourner itérer sur les étapes précédentes romprait le postulat d'indépendance de l'entrainement et de l'évaluation. Il est possible de conduire des tests post hoc pour comparer les modèles et analyser les erreurs.

Exemple. Les 3 architectures identifiées sont entrainées puis font des prédictions sur l'ensemble d'évaluation. Des tests statistiques déterminent s'ils performent au-dessus du niveau de la chance.

1.3.4.8 Déployer les modèles

Le modèle d'apprentissage automatique entrainé le plus performant est rendu disponible aux usagés. C'est probablement l'étape la plus complexe au niveau technique et humain. Le modèle requiert de l'infrastructure technologique spécialisée et une interface doit être développée pour les utilisateurs. Il y a un besoin d'éducation et de gestion du changement pour le personnel de soins et le patient.

Exemple. L'application mobile est rendue disponible et les données collectées permettent de prédire le risque de symptômes. L'équipe médicale utilise ces signaux pour faire des appels de suivi et sélectionner un traitement approprié.

1.3.4.9 Évaluer l'utilisation clinique

Une fois le modèle prédictif adopté, on peut conduire des études de groupe pour évaluer l'effet de son utilisation sur les indicateurs de succès. Surveiller l'indicateur permet aussi d'évaluer l'effet de changement au modèle prédictif ou de dégradation de la performance (Huyen, 2022).

Exemple. Un essai clinique randomisé évalue l'effet de l'application mobile sur le taux d'hospitalisation à l'urgence pour épisode psychotique.

1.4 Revue du phénotypage numérique

Les études de phénotypage numérique couvrent une grande diversité de populations, d'appareils de collecte de données, de phénomènes de santé, d'algorithmes d'apprentissage, de tâches d'apprentissage, de traitement de signal, de modèles statistiques et de méthodes d'évaluation. La revue des défis méthodologiques du phénotypage numérique par Garcia-Ceja et al. (2018) donne une vue d'ensemble du domaine, couvrant les thèmes clés avec clarté. Bien que l'article ait paru en 2018, plusieurs défis demeurent non résolus ou inexplorés, incluant: la variance intra- et inter-individuelle intrinsèque à la maladie mentale, la différence entre évaluer un traitement et un algorithme, le débalancement des classes, et l'utilisation de méthodes ordinales. La variance intra- et inter-individuelle et l'évaluation de traitement sont des défis centraux à la psychiatrie computationnelle (Paulus & Thompson, 2021).

Cette section brosse un portrait de la littérature portant sur l'utilisation de données du téléphone intelligent pour étudier la santé mentale en se basant sur la taxonomie des études de Garcia-Ceja et al. (2018). On dénote trois types d'objectifs de recherche: *association*, *détection* et *prévision*. Les études d'association sont descriptives alors que celles de détection et de prévision sont prédictives. Une étude peut inclure plus d'un objectif et débute généralement avec l'association.

1.4.1 Études descriptives

1.4.1.1 Association

Ce type d'analyse inspecte les relations entre données du téléphone et cibles de santé mentale afin de construire un modèle explicatif ou pour sélectionner les attributs à inclure dans un modèle prédictif (de détection ou de prévision). Un exemple représentatif de l'objectif d'association serait d'étudier les liens que l'humeur, le niveau de stress, les symptômes psychotiques et la qualité du sommeil autoévalués entretiennent avec les mesures extraites du téléphone (p. ex. : temps quotidien au domicile, réciprocité des appels) (Henson et al., 2020). Ces relations sont mesurées à l'aide de corrélations pour des échantillons de la population générale et d'une population clinique, puis les deux groupes sont contrastés. Une revue systématique des associations entre données passives et l'humeur dépressive révèle des résultats contradictoires au niveau de la force et de la direction des corrélations (Rohani et al., 2018). Des niveaux d'activité physique et de mobilité élevées sont fortement associés à une humeur positive et la relation demeure cohérente à travers les études. Au contraire, la relation entre la sociabilité indiquée par les appels et les SMS et l'humeur varie en intensité et en direction. La nature plus complexe de la sociabilité (à travers le téléphone) et le rôle de variables médiatrices ou modératrices comme le genre et l'âge pourraient expliquer cette variabilité (Rohani et al., 2018).

Toutefois, les données de phénotypage numérique tendent à enfreindre plusieurs postulats des méthodes corrélationnelles (Abdullah et al., 2016; Currey & Torous, 2022; Pratap et al., 2019; Strauss et al., 2022; Wang et al., 2015, 2018). D'abord, les données actives et passives ne sont pas indépendantes, car elles proviennent de mesures répétées. Aussi, la majorité des variables de santé mentale sont débalancées avec plus d'instances de valeur « en santé ». L'asymétrie de la distribution rompt le postulat d'homoscédasticité des corrélations paramétriques. Ensuite, les méthodes corrélationnelles (paramétrique ou non) présument une relation monotone entre les variables. Les valeurs extrêmement faibles et élevées (p. ex. : anormalement peu ou beaucoup d'appels) risquent d'être indicatives d'un état symptomatique. Les relations de deuxième degré (parabolique) ou plus sont potentiellement fréquentes. En plus de ces faiblesses, la validation

d'un modèle explicatif nécessite des hypothèses précises *a priori* (section 1.3.3), ce qui est rarement le cas en phénotypage numérique.

L'objectif de déterminer les données passives avec la meilleure valeur prédictive demeure pertinent, mais les coefficients de corrélation ne permettent pas de répondre à ce type de question. Une corrélation décrit une relation linéaire entre deux variables X et Y. L'approche adéquate est d'entrainer un modèle de régression linéaire et d'examiner ses coefficients pour évaluer la valeur prédictive $X \rightarrow Y$ des variables. Dans tous les cas, un modèle de régression linéaire univarié demeure un mauvais proxy de la valeur prédictive d'une variable incluse dans un modèle prédictif complexe et non linéaire. Une corrélation élevée peut suggérer une bonne valeur prédictive, mais une corrélation faible ne permet pas de conclusion (p. ex. : une relation parabolique aurait une corrélation faible).

La valeur prédictive d'une variable n'existe pas dans l'absolu; elle peut seulement être déterminée une fois le modèle prédictif d'intérêt entrainé. Toutefois, il demeure complexe d'évaluer l'apport d'une variable isolée, car les modèles d'apprentissage automatique extraient des interactions parmi l'ensemble des variables. Plusieurs approches d'interprétation de modèle et de sélection d'attributs permettent d'évaluer le rôle de variables de manière systématique. Les analyses d'association telles qu'elles existent devraient être fortement remises en question.

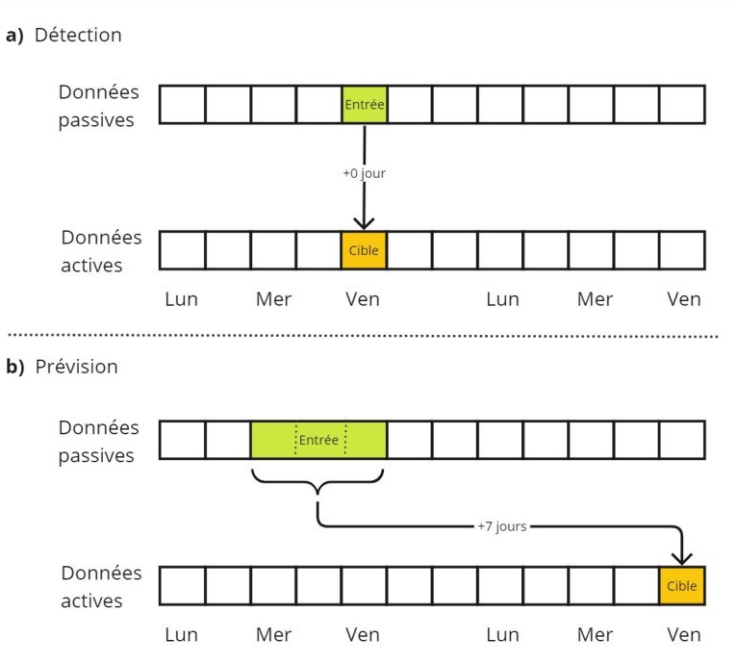
1.4.2 Études prédictives

1.4.2.1 Détection

Les études de détection sont les plus nombreuses et consistent à prédire une variable de santé mentale concurrente, tel un diagnostic, à partir de données passives (Figure 5a). Ces études utilisent typiquement des données provenant d'un grand échantillon sur une courte durée afin de maximiser la variation inter-individuelle. Dans le domaine de la santé mentale, on retrouve à la fois des études sur la santé mentale générale (p. ex. : humeur, stress, bien-être; Aharony et al., 2011; Madan et al., 2010; Rhim et al., 2020; Umematsu, Sano, & Picard, 2019; Wang et al., 2014), les troubles mentaux graves (p. ex. : bipolarité (Abdullah et al., 2016; Zulueta et al., 2018), schizophrénie (Tseng et al., 2020), dépression majeure (Canzian & Musolesi, 2015; Opoku Asare

et al., 2021), anxiété généralisée (Jacobson & Feng, 2022)) et des troubles neurodégénératifs (p. ex. : Parkinson). Lee et al. (2023) ont fait une bonne revue du phénotypage numérique en santé mentale. Les caractéristiques prévalentes incluent avoir comme cible des scores d'échelles cliniques standardisées (p. ex. : PHQ, GAD), utiliser un algorithme prédictif de boosting de gradient avec arbres décisionnels (*gradient-boosted decision trees*, GBDT) et résoudre une tâche de classification binaire. Néanmoins, il est difficile de comparer deux études et les performances obtenues compte tenu des jeux de données variés et des nombreuses décisions méthodologiques.

Figure 5 *Types d'études prédictives de phénotypage numérique*



Note. **a.** Schéma d'étude prédictive de détection. **b.** Schéma d'étude prédiction de prévision

Une limite courante des études de détection est l'utilisation de l'échantillonnage aléatoire pour créer les ensembles d'entrainement, de validation et d'évaluation. Cette approche présume que l'ordre des données n'est pas essentiel et que le passé ressemble au futur (Hewamalage et al., 2023). Plus rarement, les ensembles contiennent des participants différents afin d'évaluer la performance diagnostique pour une nouvelle personne (Vaizman, 2018; Wang et al., 2016; Zhou et al., 2022). Ces deux approches ne permettent pas d'évaluer la performance d'un modèle face aux changements de dynamique ou de transition de phase des symptômes (p. ex. : changement

de phase bipolaire, épisode psychotique). Le concept de *dérive distributionnelle* (*distribution drift*) décrit le fait que la relation $P(X, Y)$ entre les données entrées (X) et les cibles (Y) changent au fil du temps (Huyen, 2022). La relation peut être décomposée de deux manières:

$$P(X, Y) \quad = P(X|Y)P(Y) = P(Y|X)P(X)$$

Il existe trois types de dérives distributionnelles, soit la *dérive de covariable* (*covariate shift*), *dérive de cible* (*label shift*) et *dérive de concept* (*concept shift*). La dérive de concept se produit lorsque les entrées $P(X)$ sont constantes, mais la cible $P(Y|X)$ change. Cela s'applique notamment aux effets saisonniers ou encore à l'effet de nouvelles politiques de santé (Davis et al., 2017; Thai & Ebell, 2019). La dérive de covariable se produit lorsque la distribution des données en entrée $P(X)$ change, mais non la condition $P(Y|X)$. À l'inverse, la dérive de cible est lorsque la distribution des cibles $P(Y)$ change, mais non la condition $P(X|Y)$. En pratique, les deux phénomènes ont tendance à se manifester ensemble lorsqu'un modèle est entrainé sur des données non représentatives du cas d'utilisation (p. ex. : prévalence dans un hôpital régional vs urbain) (Riley et al., 2016; Testa et al., 2014). Le passage du temps et les événements de la vie entrainent nécessairement des changements des distributions de signaux provenant du téléphone $P(X)$ (p. ex. : déménagement, nouvel emploi) et les symptômes $P(Y)$ (p. ex. : médication, thérapie).

Contrairement aux études d'association, les études de détection permettent de valider la valeur prédictive du phénotypage numérique pour des variables de santé mentale. Cependant, les données collectées et les stratégies d'entrainement sont souvent insuffisantes pour évaluer la performance prédictive face à des variations de dynamiques temporelles intra-individuelles. De bons résultats de détection ne devraient pas être extrapolés à de bonnes prévisions.

1.4.2.2 Prévision

Les modèles de prévision tentent de prédire les risques futurs à partir des trajectoires des signaux passifs en entrée (Figure 5b). Pour bien estimer la performance, il faut employer une stratégie d'entrainement représentant les dépendances temporelles et un jeu de données acquis sur une période assez longue pour observer de la variabilité intra-individuelle. C'est le type d'étude le plus

complexe et celle décrite en moins grand nombre malgré la pertinence de la tâche pour la psychiatrie clinique. En 2017, Suhara et al. (2017) ont utilisé des données actives d'humeur, de comportement et de sommeil pour prédire l'état binaire « sévèrement déprimé » ou non un, trois, ou sept jours dans le futur grâce à des réseaux de neurones. En 2019, Spathis et al. (2019a) et Umematsu, Sano, & Picard (2019) utilisent les signaux passifs du téléphone pour prévoir l'humeur future, dans sept et un jours respectivement, pour des échantillons de la population générale.

Néanmoins, les études de prévision demeurent hétérogènes, peu nombreuses et espacées dans le temps alors que l'apprentissage automatique a connu des progrès fulgurants. L'ère de l'apprentissage profond débute autour de 2010 (Goodfellow et al., 2016) et différents papiers suggèrent l'utilisation des réseaux de neurones récurrents (*recurrent neural network*; RNN; Rumelhart et al., 1986) pour modéliser les dépendances temporelles et formuler des prévisions en santé et en phénotypage numérique (Durstewitz et al., 2019; Koppe et al., 2019). La majorité des études de prévision emploient les RNNs de type LSTM (Hochreiter & Schmidhuber, 1997) ou parfois GRU (Cho et al., 2014). Ces modèles sont souvent comparés de manière peu systématique à des algorithmes très simples (et encore plus vieux) comme la régression logistique, les machines à vecteur de support (Vapnik & Chervonenkis, 2015), etc. Au contraire, peu ont examiné les algorithmes GBDTs incluant XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), Catboost (Prokhorenkova et al., 2018) à des fins de prévision malgré le succès de ce type de modèle pour la tâche de détection (Lee et al., 2023). Les algorithmes N-Beats (Oreshkin et al., 2020) et N-Hits (Challu et al., 2023) ont été spécifiquement conçus pour fournir des prévisions explicables et n'ont jamais été évalués pour le phénotypage numérique. Les RNNs eux-mêmes ont été complètement remplacés par les transformeurs (Vaswani et al., 2017) pour le traitement du langage naturel.

Le peu d'études de phénotypage numérique de prévision laisse plusieurs modèles prédictifs non évalués. Une évaluation systématique de ces algorithmes permettrait de situer leur performance pour la tâche. Dans l'éventualité où les performances sont comparables, cela ouvre une discussion sur les autres propriétés pertinentes d'un modèle pour une utilisation clinique (interprétabilité, quantité de données nécessaires, prédiction probabiliste, etc.).

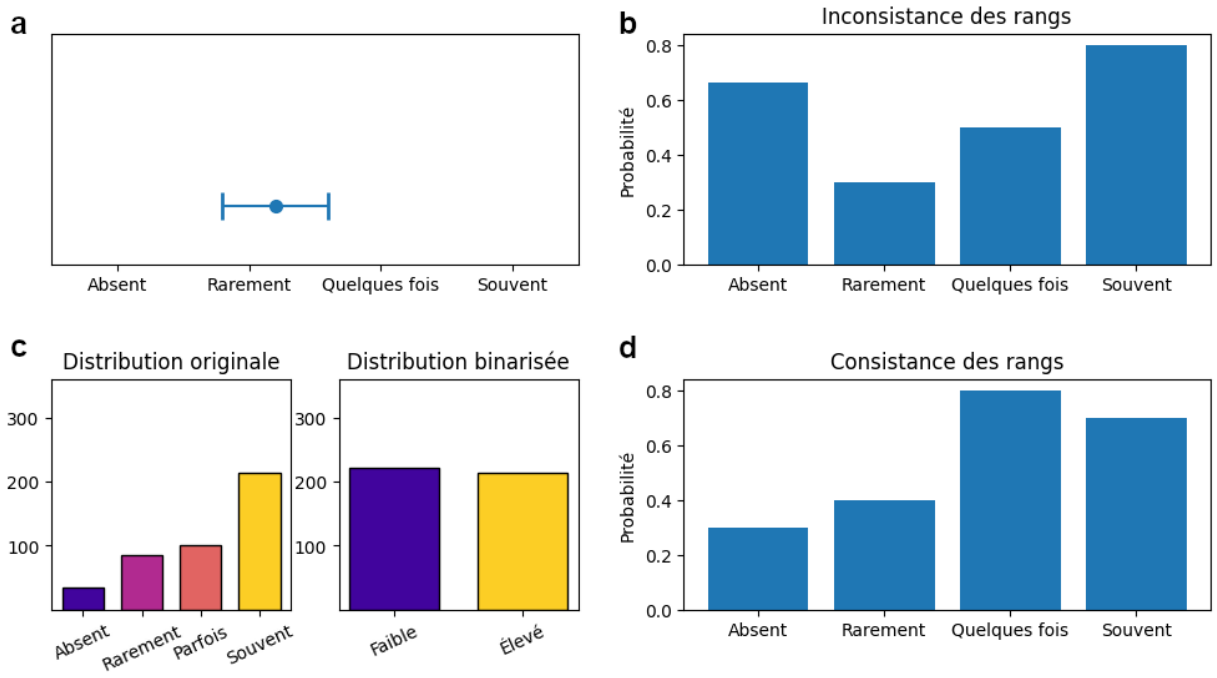
1.5 Objectifs de recherche

Ce travail vise à prédire des états mentaux futurs à partir de données provenant du téléphone intelligent grâce à l'apprentissage automatique. Le premier objectif est de développer des modèles prédictifs ordinaux capables de rendre compte des scores obtenus sur des échelles standardisées ordinales. La relation entre le débalancement des classes et la tâche d'apprentissage est explorée. Le second objectif est d'évaluer la qualité des prévisions pour des cibles plus éloignées dans le futur. L'algorithme XGBoost, peu utilisé pour les tâches de prévision, est comparé de manière systématique au LSTM, soit l'un des algorithmes les plus populaires, à l'aide d'une méthodologie standardisée.

1.5.1 Objectif 1: Prédire des autoévaluations ordinales

La majorité des études du domaine reposent sur des données ordinales provenant de questionnaires, d'échelles cliniques, ou d'évaluation écologique momentanée (Rohani et al., 2018). Une échelle ordinale inclut des valeurs discrètes suivant un ordre strict (<), mais ne définit point de distance entre les valeurs. Pourtant, la majorité des études prédictives (détection et prévision) utilisent la classification binaire et la régression ordinale (Lee et al., 2023; Rohani et al., 2018) et rarement la classification multiclasse. Garcia-Ceja et al. (2018) identifient la régression ordinale (parfois appelée classification ordinale) comme idéale pour prédire les cibles ordinales fréquentes en santé mentale. Cependant, leur revue de la littérature n'en répertorie aucun exemple. Par conséquent, les modèles modèles ne respectant pas la nature des données produisent des prédictions qui ne correspondent à aucun construit clinique bien défini. Pour illustrer ce problème, la prochaine section compare les différentes tâches d'apprentissage pour prédire la fréquence d'un symptôme sur l'échelle {Absent : 0, Rarement : 1, Parfois : 2, Souvent : 3}.

Figure 6 *Propriétés de différentes tâches d'apprentissage*



Note. **a**. Prédiction continue sur une échelle ordinale. **b**. Prédiction multiclasse montrant l'inconsistance des rangs. **c**. Distribution d'une variable binarisée (0, 1, 2 vs 3) **d**. Prédiction ordinale montrant la consistance des rangs.

La régression continue préserve l'ordonnancement des données, mais perd les classes discrètes en supposant une distance fixe entre les observations (c.-à-d., une échelle à intervalle). Une continue du modèle (p. ex. : 1.2) ne correspond à aucun concept défini sur l'échelle. On pourrait naïvement conclure que 1.2 suggère « un peu plus que Rarement », mais la certitude du modèle est inconnue. Maintenant, admettons une prédiction probabiliste entre 0,8 à 1,6 (Figure 6a). L'interprétation demeure ambiguë étant donné qu'une minorité de l'intervalle se situe « moins que Rarement », une majorité « plus que Rarement », mais seule la valeur Rarement est incluse et non Parfois.

Au contraire, la classification multiclasse conserve les classes discrètes, mais perd leur ordonnancement. Le modèle génère une probabilité pour chaque classe et prédit celle avec la probabilité la plus élevée. Comme l'ordre n'est pas représenté, un modèle pourrait prédire à 80% que le symptôme est Souvent présent et à 70 % qu'il est Absent (Figure 6b). Il y a inconsistance des rangs (*rank inconsistency*; Shi et al., 2023). Une telle prédiction est difficile à interpréter et probablement peu utile.

La classification binaire peut paraitre comme une version simplifiée d'un problème ordinal, mais en fait elle brouille à la fois la nature discrète et l'ordre des classes. D'abord, il faut choisir comment binariser. On peut créer l'échelle {Absent : 0, Présent : 1} en mettant (Absent vs Rarement, Parfois, Souvent) par exemple. En rassemblant plusieurs classes, on perd leur nature discrète et l'ordre qu'elles entretenaient (Figure 6c). Les classes binaires ne font plus directement référence à l'échelle ordinale originale et il devient impossible de reconstruire l'échelle ordinale à partir d'une valeur binaire; la variable binaire contient donc au plus la même quantité d'information que la variable originale. C'est une démonstration de l'inégalité du traitement de données (*data processing inequality*) (Beaudry & Renner, 2012; Cover & Thomas, 2006).

$$\text{Inégalité du traitement de données:} \quad I(X;Y) \geq I(X;Z)$$
$$o\text{ù } Z\ est\ la\ transformation\ de\ Y$$

Finalement, la régression ordinale préserve à la fois les classes discrètes et leur ordonnancement (Figure 6d). Comme la tâche partage des propriétés avec la régression continue et la classification multiclasse, il en existe différentes formulations mathématiques. La plus populaire est de convertir la tâche ordinale de prédire *n* ancrages en une tâche multiclasse de *n-1* catégories (>0, >1, …, >n-1) avec une contrainte pour respecter la consistance des rangs. Jusqu'à présent, seuls les réseaux de neurones ont la flexibilité suffisante pour incorporer la contrainte des rangs dans le modèle lui-même. Pour les autres types de modèles, on commence avec une tâche de régression continue puis l'on applique des étapes de post-traitement pour rendre les prédictions discrètes. Les seuils pour créer les classes discrètes sont appris à partir des données.

Tableau 2 *Propriétés des tâches d'apprentissage*

Propriété	Régression continue	Classification multiclasse	Classification binaire	Régression ordinale
Préserve l'ordre	Oui	Non	Non	Oui
Préserve les classes	Non	Oui	Non	Oui

Bien que la majorité des études de phénotypage numérique utilise des cibles ordinales, la classification binaire ou la régression continue sont les tâches les plus employées. Une première raison potentielle est le grand nombre d'implémentations disponibles en ligne contrairement à la régression ordinale. Ensuite, on pourrait croire qu'il est plus facile de résoudre ces tâches et d'obtenir de bonnes performances prédictives. Cependant, ces affirmations devraient être validées empiriquement et le choix de tâche devrait plutôt être guidé par la nature du problème. Finalement, la classification multiclasse et la régression ordinale semblent implicitement mises de côté pour éviter le problème de débalancement des classes.

Le débalancement des classes (ou débalancement des données; *class imbalance*) réfère au nombre inégal d'exemples collectés pour chaque classe. La notion est équivalente à l'asymétrie (*skewness*) pour des distributions continues. Lors de l'entrainement du modèle, la classe avec moins d'exemples a un poids plus faible dans la fonction de coût qui guide l'apprentissage. Différentes méthodes adaptent les données, l'algorithme ou la fonction de coût afin de réduire ce problème et d'améliorer la performance (Krawczyk, 2016). La binarisation peut délibérément réduire le débalancement et la régression continue permet de l'ignorer bien que la distribution des cibles devrait tout de même être rapportée (Klement & El Emam, 2023). La tâche d'apprentissage détermine les propriétés des données que le modèle va respecter et influence le type d'erreur de prédiction. Pour des données ordinales, la régression ordinale rend le mieux compte des données (Tableau 2). Les prédictions du modèle sont directement sur l'échelle de collecte de données et peuvent être directement interprétées par le clinicien. À l'inverse, les prédictions continues, multiclasses et binaires risquent d'être ambiguës et difficiles à intégrer dans le processus de décision clinique, et donc moins utiles.

1.5.2 Objectif 2: Prévoir les états futurs

Prédire les états mentaux futurs et évaluer la trajectoire des symptômes sont des tâches essentielles de la psychiatrie clinique afin de prévenir du mal être. Cependant, il y a relativement peu d'études de phénotypage numérique de prévision. La majorité d'entre elles utilisent les RNNs, plus particulièrement les LSTMs, et rapportent de bonnes performances. Aucune étude n'a évalué rigoureusement la performance de prévision de modèles en arbre malgré leur succès dans les études de phénotypage numérique de détection (Lee et al., 2023).

Pour des données tabulaires (équivalent aux études de détection), la littérature en apprentissage automatique a conclu à plusieurs reprises que les GBDTs étaient supérieurs aux réseaux de neurones (*deep tabular learning*). Les GBDTs montrent un avantage marqué lorsqu'un jeu de données contient des attributs déviants de la normalité (asymétrie, etc.). Exceptionnellement, les réseaux de neurones ont un avantage pour les petits jeux de données avec des données « propres » (aucune valeur manquante, attributs distribués normalement, etc.), car le biais inductif de l'architecture compense le peu de données (McElfresh et al., 2023). En sommes, les GBDTs démontrent une supériorité pour les données tabulaires en contexte appliqué où les données sont volumineuses, avec des valeurs manquantes et des distributions irrégulières. D'ailleurs, sept sur dix compétitions en ligne d'apprentissage automatique sur données tabulaires ont été remportées par des GBDTs en 2022 (Carlens, 2023). Toutefois, contrairement aux modèles statistiques typiques en santé publique et aux RNNs, les GBDTs n'acceptent pas en entrée des séries temporelles. Bien que les GBDTs n'aient pas de représentation explicite de la dimension temporelle, il est possible de leur fournir de l'information à propos du passé en entrée afin de formuler des prévisions.

La littérature existante a conduit une fouille de modèles biaisée vers les réseaux de neurones, en particulier les LSTMs. Ces modèles ont des propriétés intéressantes pour traiter les séquences de données, mais ils sont complexes à entrainer et à interpréter. Comme détaillé à la section 1.4.2.2, plusieurs algorithmes de prévision ont été inventés depuis 1997 dont la famille des GBDTs qui ont montré de bonnes performances dans les études de détection et d'autres domaines. Une évaluation systématique des algorithmes de prévision permettrait d'améliorer la performance, mais aussi d'évaluer leurs autres propriétés pertinentes à une utilisation clinique (interprétabilité, quantité de données nécessaires, prédiction probabiliste, etc.). L'utilisation de tests statistiques est nécessaire pour déterminer si des niveaux de performance sont significativement différents, ce qui est parfois ignoré en apprentissage automatique (Hewamalage et al., 2023). Typiquement, les modèles sont comparés, mais seulement une minorité d'études les compare au hasard, à des heuristiques (p. ex. : humeur selon le jour de la semaine), ou aux prévisions d'experts. Cette seconde étape est importante, car deux modèles peuvent être significativement différents, mais performer sous le niveau de base pertinent.

Chapitre 2 – Méthodologie

Ce chapitre détaille la méthodologie selon le cycle de développement de système d'aide à la décision présenté à la section 1.3.4. L'apprentissage automatique consiste à conduire la meilleure fouille pour obtenir un modèle prédictif optimal. Sachant que les décisions méthodologiques sont nombreuses, l'accent est mis sur les aspects particuliers qui sont particuliers à ce projet ou qui contrastent avec la littérature. Klement et El Emam (2023) ont conduit une revue de 17 guides pour rapporter les résultats d'apprentissage automatique en santé jugés de haute qualité. Leur analyse a permis de déterminer 37 éléments à rapporter divisés en 5 catégories (Tableau 3). Cette section inclut des indicateurs (p. ex. : i-2.3) pour chaque élément couvert. Certains éléments sont seulement mentionnés brièvement, car ils sont détaillés ailleurs dans cet ouvrage.

2.1 Définir la problématique

Actuellement, les soins psychiatriques présentent de la difficulté à prévenir des symptômes ou événements cliniques majeurs comme indiqué par le nombre élevé de réadmissions aux urgences. Il serait d'intérêt d'aider les cliniciens dans leur processus de pronostic afin d'anticiper la dégradation de l'état de patients avec un trouble de schizophrénie (i-1.1, i-1.7). On tente de vérifier si les modèles prédictifs basés sur le phénotypage numérique permettent d'améliorer le pronostic (i-1.2). Présentement, la majorité des cliniciens utilisent l'entrevue clinique comme méthode de collecte de données principale (i-1.3).

2.2 Identifier les mécanismes sous-jacents

La schizophrénie est associée à des symptômes positifs qui interfèrent avec le bon fonctionnement (p. ex. : hallucination, pensée invasive) et des symptômes négatifs, soit une réduction des fonctions normales (p. ex. : isolation sociale, affect plat). Les études antérieures de phénotypage numérique ont identifié les signaux de l'accéléromètre et du GPS comme étant de bons prédicteurs de symptômes variés (i-1.4; Faurholt-Jepsen et al., 2022; Rohani et al., 2018). On peut supposer que différents patrons de signaux soient associés aux symptômes positifs et négatifs.

Tableau 3 *Comment rapporter une étude en santé utilisant l'apprentissage automatique*

	1. Détails de l'étude	2. Description des données	3. Méthodologie	4. Évaluation du modèle	5. Explicabilité du modèle
1	La tâche médicale / clinique d'intérêt	Inclusion ou exclusion de patients dans l'étude	Données manquantes	Mesure de performance	Attributs importants
2	La question de recherche	Méthode de collecte de données	Débalancement des classes	Conséquence d'une erreur	Plausibilité des prédictions
3	La pratique clinique actuelle	Biais introduits par la méthode de données	Réduction de la dimensionnalité	Validation interne	Interprétation des prédictions par un utilisateur
4	Les prédicteurs ou variables confondantes de la cible connues	Caractéristiques des données	Gestion des données extrêmes	Hyperparamètres	
5	Le protocole de l'étude	Transformations des données	Augmentation de données	Validation externe	
6	Le contexte de l'institution médicale	Qualité des données	Préentrainement du modèle	Dérive distributionnelle	
7	La population cible	Taille d'échantillon	Sélection d'algorithmes		
8	L'utilisation attendue du modèle d'apprentissage automatique	Disponibilité des données	Évaluation du modèle durant entrainement		
9	Performance référence de l'apprentissage automatique	Optimisation d'hyperparamètres	Ajustement des prédictions du modèle		
10	Approbation éthique				

2.3 Spécifier la tâche d'apprentissage

L'apprentissage automatique supervisé servira à entrainer des modèles à prédire les valeurs futures de dix états mentaux (section 3.3.1). L'horizon temporel sera accru pour évaluer les prédictions pour le même jour, le prochain jour et la semaine prochaine. La régression ordinale sera utilisée et comparée à la régression continue, la classification multiclasse et la classification binaire. Tseng et al. (2020) ont entrainé des modèles de régression continue sur le même jeu de données avec les mêmes cibles. Chouduhary et al. (2022) ont évalué des modèles de régression continue, de classification binaire et de classification multiclasse sur un autre jeu de données. Ces deux études serviront de références primaires (i-1.9).

2.4 Collecter les données

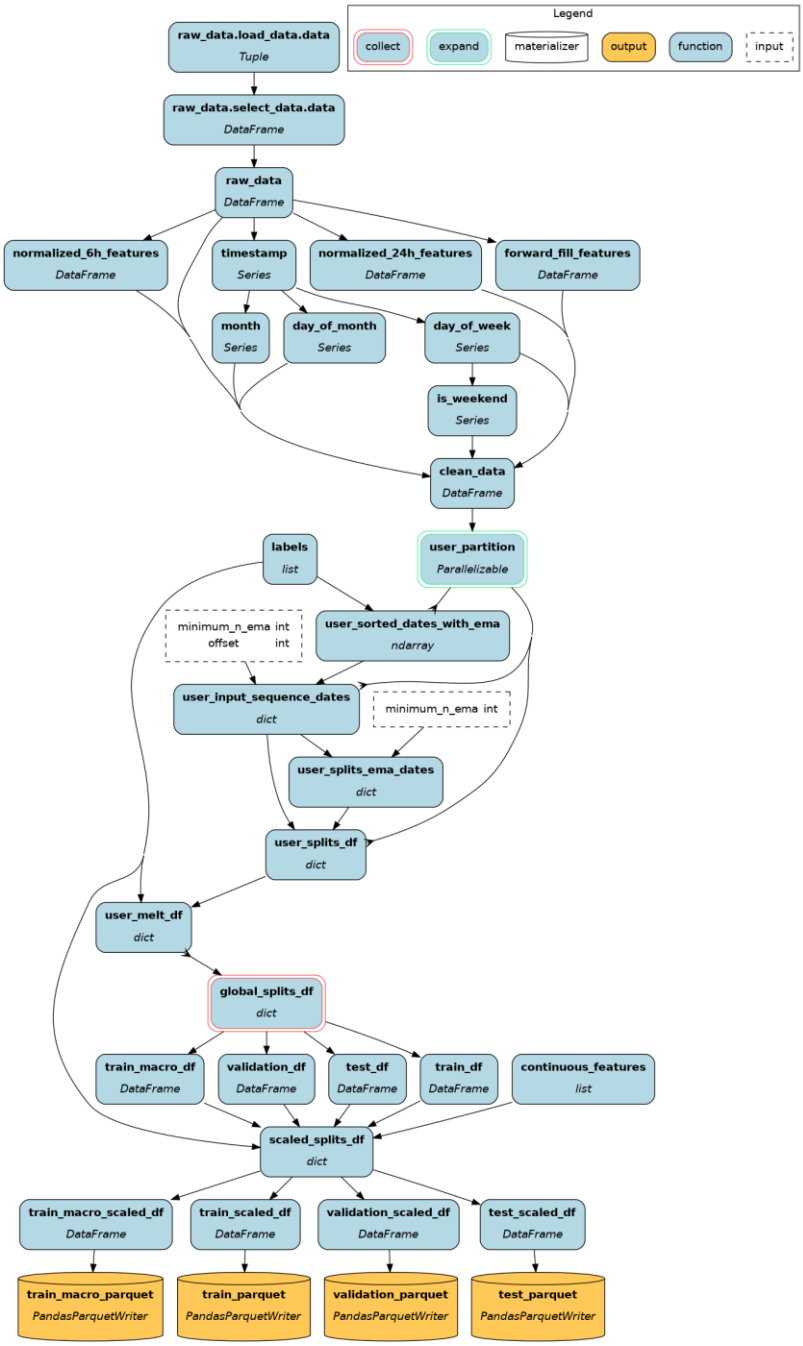
Les données proviennent de l'essai clinique randomisé CrossCheck, dont certaines données sont publiquement accessibles. Plus de détails sur la collecte de données sont fournis à la section 3.3.1 (i-1.6, i-1.10, i-2.1, i-2.2, i-2.7, i-2.8). Le protocole de l'étude CrossCheck introduit un biais dans les données en collectant les autoévaluations d'états mentaux presque exclusivement les lundis, mercredis et vendredis (i-2.3; détails à la section 3.3.1).

2.5 Préparer les données

L'équipe de CrossCheck a réalisé un travail considérable pour traiter les signaux bruts et agréger les attributs en périodes de 6 heures. Seuls les attributs sans données manquantes ont été inclus, ce qui explique l'exclusion de ceux liés au capteur de luminosité (i-3.1). Quatre attributs liés au temps ont été ajoutés, soit le jour de la semaine, le jour du mois, le mois de l'année et si c'est la fin de semaine pour un total de 29 attributs. Le nombre d'attributs ajoutés a été limité pour établir une performance de base sur ces données et favoriser la reproductibilité des résultats. Les exemples sont composés d'une séquence de 3 jours consécutifs ($\times$ 4 périodes de 6h) et d'un état mental cible du même jour (+0 jour), du prochain jour (+1 jour) ou de la semaine prochaine (+7 jours). Un exemple est composé de 348 mesures (29 attributs $\times$ 12 périodes) et 3 jeux de

données (un par horizon temporel) sont créés avec 10 états mentaux chacun. La Figure 7 est générée à partir du code pour traiter les données et donne un aperçu du processus[1].

Figure 7 *Processus de prétraitement des données*



1 La figure a été générée avec la librairie Python *Hamilton*. J'ai implémenté la fonction pour créer la visualisation et ma contribution a été intégrée au projet https://github.com/DAGWorks-Inc/hamilton/pull/512.

Les fuites de données entre les ensembles d'entrainement, de validation et d'évaluation. Doivent être prévenues pour préserver leur indépendance et la fiabilité des estimations de performance. Une erreur de prétraitement commune est de normaliser simultanément les attributs des différents ensembles. Ce risque de fuite est aggravé pour les séquences, car une observation peut se trouver dans plus d'un ensemble (Hewamalage et al., 2023). On veille à éviter ce problème en prétraitant les données avant de créer les séquences (Figure 7).

Le nombre de transformations a été gardé au minimum, mais certaines étaient nécessaires pour permettre l'entrainement des différents algorithmes (i-2.5). Les attributs ont été transformés vers des distributions normales avec la méthode Yeo-Johnson (Raymaekers & Rousseeuw, 2021), car les LSTMs ont besoin d'une échelle commune, idéalement autour de zéro, pour un apprentissage stable via descente de gradient (Goodfellow et al., 2016). Les modèles en arbres comme le XGBoost sont invariants à l'échelle des attributs (Müller & Guido, 2017). Comme les attributs temporels (p. ex. : jour de la semaine) sont cycliques, un encodage catégoriel {lundi : 0, mardi : 1, …, dimanche : 6} qui présente lundi et dimanche comme étant les valeurs les plus éloignées rend l'apprentissage de relations temporelles difficile. Pour obtenir des valeurs autour de zéro pour les LSTMs, les attributs liés au temps ont été projetés sur des splines (Scikit-learn, 2023b). Pour les XGBoosts, les attributs ont été encodés « one-hot », ce qui facilite l'apprentissage d'interactions pour les modèles en arbre (Scikit-learn, 2023a).

De plus, le débalancement des cibles (i-3.2) et les données extrêmes (i-3.4) n'ont pas été ajustés et il n'y a pas eu non plus d'augmentation de données (i-3.5). On détaille l'importance d'utiliser les données telle qu'acquises pour une estimation robuste de la performance en contexte appliqué à la section 3.5. Autrement, la réduction de la dimensionnalité (i-3.3) et le préentrainement de modèles (i-3.6) ne concernent pas notre type de problème.

2.6 Créer les modèles

Les algorithmes LSTM et XGBoost ont été sélectionnés, car ils avaient le plus de succès dans les études de prévision et de détection respectivement (i-3.7). Un problème potentiel pour les tâches de régression est l'incapacité des arbres décisionnels (Müller & Guido, 2017) et la capacité limitée des LSTMs à extrapoler les prédictions à l'extérieur des cibles d'entrainement (Hewamalage et

al., 2023). Cependant, l'étendue des prédictions est contrainte à deux ou quatre valeurs pour l'ensemble des tâches d'apprentissage, évitant cette limitation. Si n'était pas le cas, une solution efficace est de prétraiter la cible avec les valeurs qui la précède afin que le modèle prédise la variation entre les deux observations plutôt que la valeur absolue (p. ex. : les cibles [1, 0, 2, 3, 5] deviennent [0, -1, +2, +1, +2]; Hewamalage et al., 2023).

2.6.1 Algorithme LSTM

Durstewitz et al. (2019) et Koppe et al. (2019) recommandent les RNNs pour traiter les signaux de capteurs médicaux ou de phénotypage numérique. Le graphe du modèle repose sur une fonction récurrente, ce qui permet d'apprendre des fonctions à partir de longues séquences et de longueurs variables.

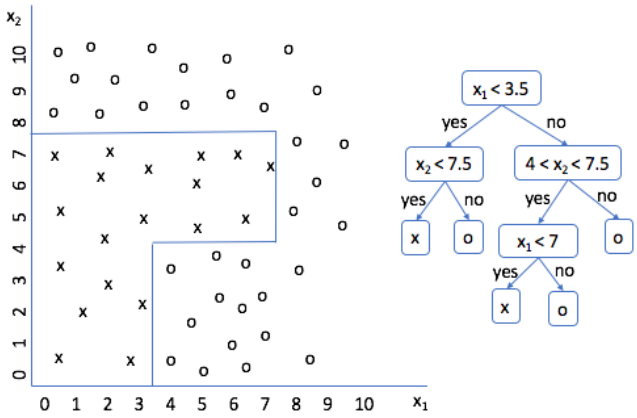
Cependant, la chaine d'opérations rend la représentation de dépendances à long terme difficile. Formellement, cela entraine les problèmes d'explosion et de disparition du gradient (*exploding gradient* et *vanishing gradient*), soit la croissance ou la décroissance exponentielle du gradient. L'explosion du gradient rend la mise à jour de paramètres et l'apprentissage instable ou impossible alors que la disparition de gradient rend la transmission d'informations entre neurones éloignés difficile. L'algorithme *long short-term memory* (LSTM; Hochreiter & Schmidhuber, 1997) maintient un état interne lui permettant de représenter les séquences et sous-séquences contenues dans les donnes passives et les autoévaluations d'états mentaux. Pour mitiger l'explosion et la disparition de gradient, on utilisera la fonction d'activation *rectified linear unit* (ReLU) avec des seuils au gradient (*gradient clipping*; Goodfellow et al., 2016).

2.6.2 Algorithme XGBoost

Un algorithme de la famille GBDT est en fait un *ensemble* de modèles *prédicteurs simples* (*weak learners*), soit des arbres de décision peu profonds (*shallow*), performant juste au-dessus du niveau de la chance. Le *boosting* est un méta-algorithme qui assemble plusieurs prédicteurs simples en un *prédicteur fort* (*strong learner*) (Chen & Guestrin, 2016). Durant l'entrainement, chaque prédicteur simple vient combler les erreurs des précédents, permettant de dédier plus de ressources aux exemples compliqués. Parmi les GBDTs, c'est le processus de boosting et d'entrainement des prédicteurs simples qui varient. Les algorithmes GBDT ont énormément de

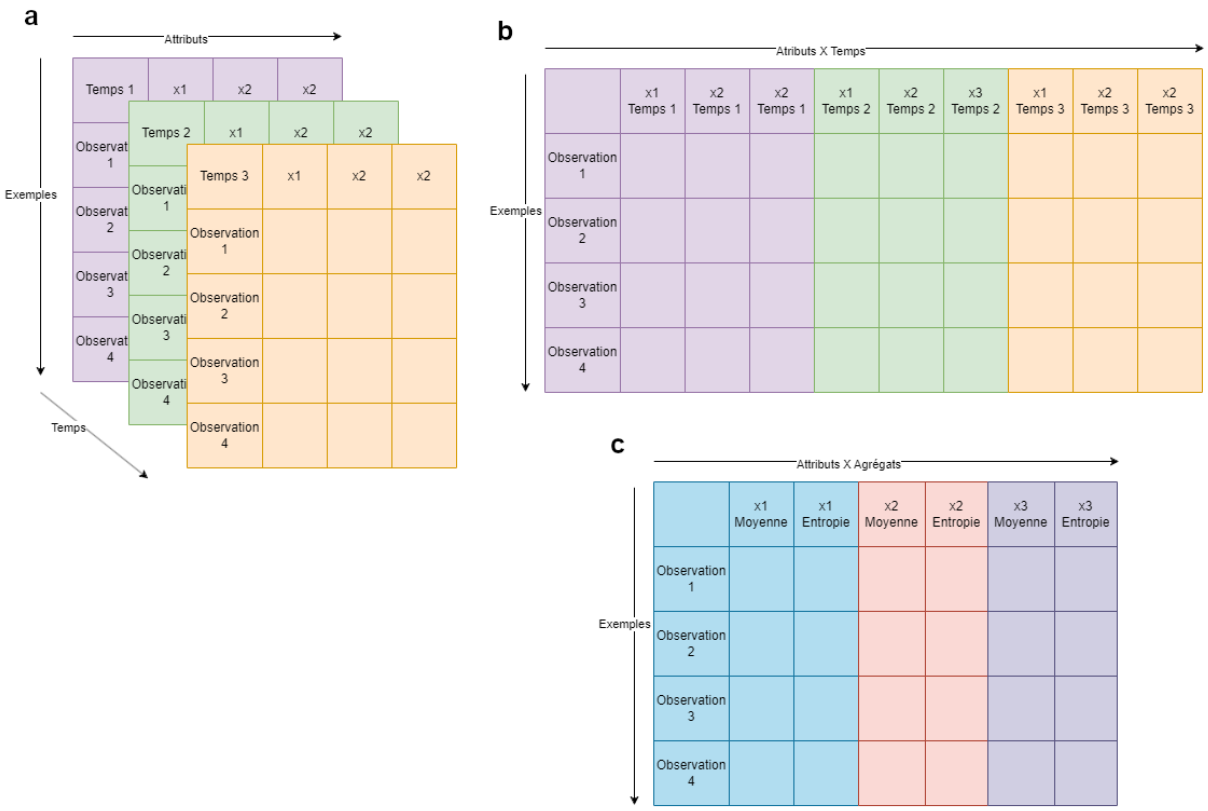
succès sur les données tabulaires, car ils peuvent efficacement gérer des variables catégorielles, des données manquantes, des relations non linéaires et des interactions complexes (Grinsztajn et al., 2022). La Figure 8 présente une frontière de décision difficile à définir pour un modèle linéaire, mais trivial pour un arbre décisionnel.

Figure 8 *Frontière de décision en 2-dimensions d'un arbre décisionnel*



Alors que les RNNs acceptent en entrée des tenseurs tridimensionnels (n exemples, 29 attributs, 12 temps; Figure 9a), les GBDTs utilisent des données tabulaires, c'est-à-dire des matrices à deux dimensions. Pour cette conversion, la dimension temporelle a été « déroulée » pour obtenir (n exemples, 29 attributs × 12 temps = 348 attributs-temps). Ainsi, les deux algorithmes reçoivent exactement la même information (n × 29 × 12 valeurs numériques) en entrée pour permettre de les comparer (Figure 9b). Cependant, le GBDT ne peut pas discerner l'ordre des attributs (p. ex. : [attribut 1, temps 1] précède [attribut 1, temps 2]).

Figure 9 *Restructurer des séquences pour passer de 3 à 2 dimensions*



Note. **a.** Séquences en 3 dimensions (exemples, attributs, temps). **b.** Séquences en 2 dimensions (exemples, attributs × temps). **c.** Agrégats de séquences en 2 dimensions (exemples, attributs × agrégats)

Cette approche devient problématique avec les séries à haute résolution temporelle, elle créerait un nombre trop élevé de colonnes. Comme un arbre décisionnel apprend en divisant un attribut à la fois récursivement, un faible ratio exemple / colonne l'empêche d'apprendre des relations généralisables. Dans ce cas, une alternative est d'écraser la dimension temporelle en la résumant par une ou plusieurs statistiques (moyenne, déviation, entropie, maximum, etc.; Figure 9c). Toutefois, le nombre de colonnes peut facilement exploser à nouveau.

2.7 Spécifier la stratégie d'entrainement

La stratégie d'entrainement est détaillée dans l'article (sections 3.3.2 et 3.3.5), ce qui inclut les items (i-2.9, i-3.8, i-3.9, i-4.1, i-4.3).

2.8 Entrainer les modèles

On entrainera 240 modèles prédictifs (10 états mentaux × 3 horizons temporels × 2 algorithmes × 4 tâches). Ce projet est composé de plus de 3000 lignes de code en langage Python dont un script paramétrable pour l'entrainement des 120 XGBoosts (Figure 10) et un autre pour les 120 LSTMs (Figure 11).

Voici quelques librairies importantes à la réalisation de ce projet:

- *Hamilton* : librairie centrale pour structurer l'ensemble des analyses
- *XGBoost* : implémentation de l'algorithme du même nom
- *PyTorch Lightning* : implémente les bonnes pratiques d'entrainement de réseaux de neurones; les algorithmes LSTM pour les quatre tâches ont été implémentés manuellement
- *PyTorch CORAL* : implémentation de l'algorithme CORN pour la régression ordinale avec réseau de neurones
- *Optuna* : Optimisation d'hyperparamètres
- *Hydra* : gestion de configurations et de résultats d'expériences
- *Scikit-learn* : mesures de performance de base; la MAMAE a été implémentée manuellement
- *Scipy* : tests statistiques de rangs; le calcul de rangs appariés a dû être implémenté
- *Scikit-posthoc* : test statistique post hoc Nemenyi et visualisations de diagramme critique
- *Pandas* et *Numpy* : manipulation et transformations de données
- *Matplotlib* : création de figures

Figure 10 *Processus d'entrainement pour un XGBoost*

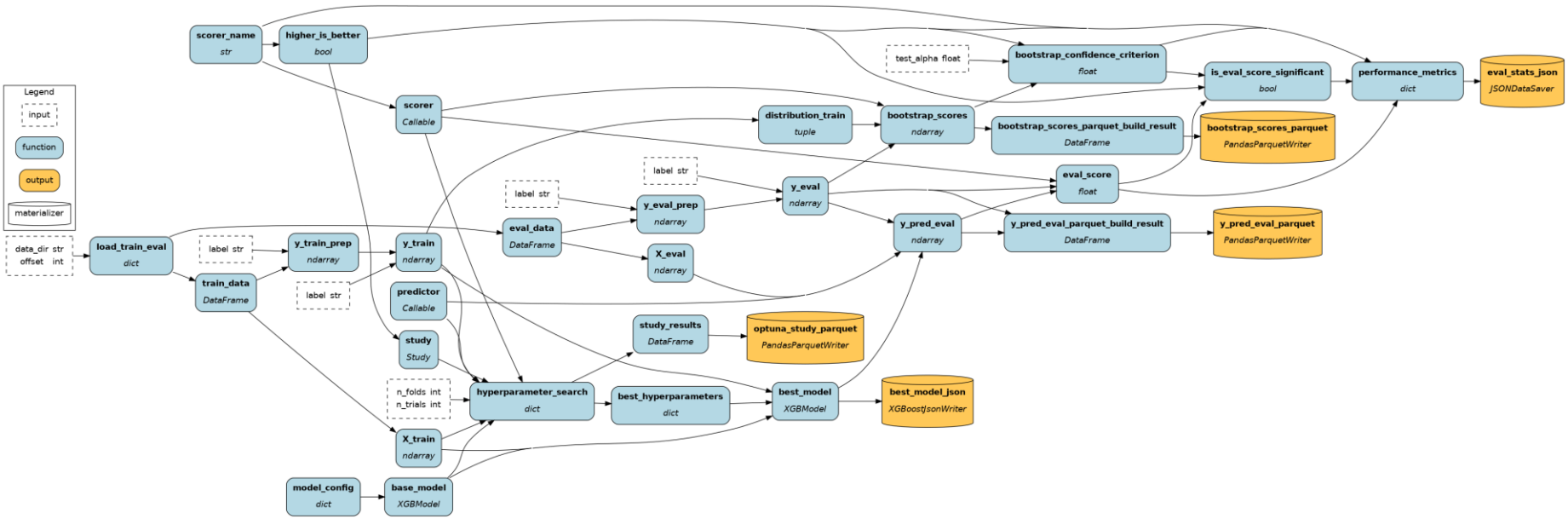
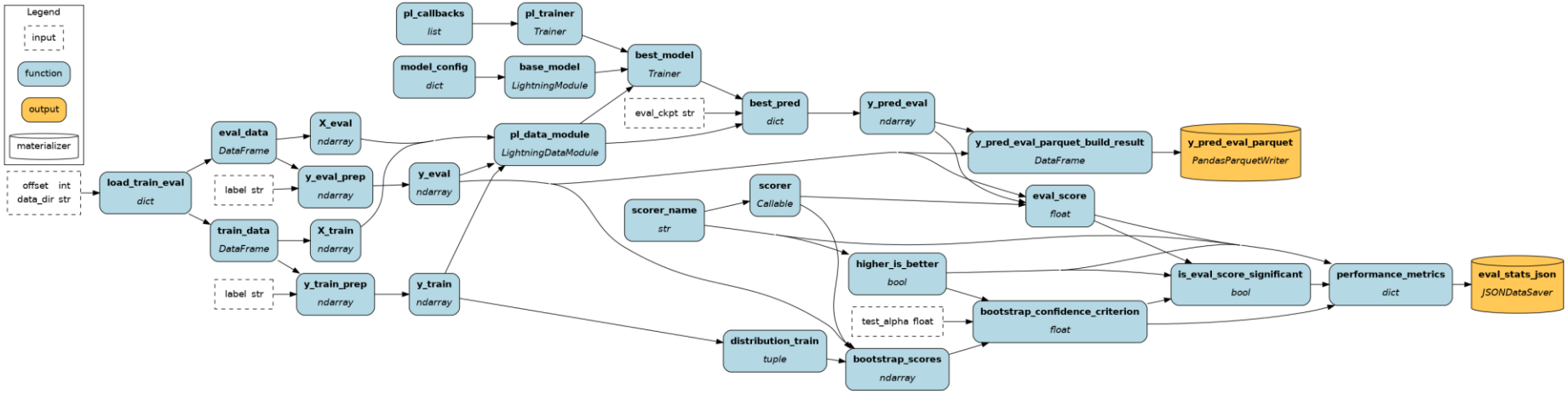


Figure 8 *Processus d'entrainement pour un LSTM*



2.9 Évaluer les modèles

Les modèles ont été évalués sur l'ensemble d'évaluation avec les mesures balancées (p. ex. : MAMAE), ainsi que non balancées (p. ex. : MAE) pour être comparées à la littérature (i-4.1, i-4.3, section 3.5). Des tests statistiques ont permis de tirer des conclusions généralisables quant à l'influence sur la performance de l'état mental, de l'horizon temporel, de l'algorithme et de la tâche prédictive.

Ensuite, l'importance des attributs a été calculée avec les valeurs *SHapley Additive exPlanation* (SHAP; i-5.1; Lundberg & Lee, 2017). Après avoir entrainé et évalué tous les modèles, les distributions ordinales et binaires des états mentaux ont été inspectées pour vérifier l'absence de disparités entre les ensembles d'entrainement, de validation et d'évaluation (i-4.6). Cette validation doit être faite à la fin de l'étude pour éviter une fuite de données guidant les choix méthodologiques des chercheurs (Varoquaux & Cheplygina, 2022).

Il n'y a pas eu de validation externe sur d'autres données, mais il serait facile d'adapter le code d'analyse à cet effet (i-4.5). De plus, une collaboration avec des cliniciens permettrait de spécifier un cas d'utilisation clinique (i-1.8) et de pondérer les types d'erreurs selon les conséquences potentielles (i-4.2). Par la même occasion, on pourrait juger de la facilité d'interprétation et de la plausibilité des explications (i-5.2, i-5.3).

2.10 Déployer les modèles et Évaluer l'utilisation clinique

Ce projet ne couvre pas les étapes de déploiement et d'évaluation clinique des modèles entrainés.

Chapitre 3 - Article : Forecasting mental states in schizophrenia using digital phenotyping data

Auteurs : Thierry Jean[1,2], BSc; Rose Guay Hottin[1], BSc; Pierre Orban[1,2], PhD

[1] Centre de recherche de l'Institut universitaire en santé mentale de Montréal

[2] Département de psychiatrie et d'addictologie, Université de Montréal

Cette section contient l'article en révision *Forecasting mental states in schizophrenia using digital phenotyping data* tel que soumis au journal scientifique PLOS Digital Health le 30 novembre 2023 (#PDIG-D-23-00450). J'ai contribué à la conceptualisation de l'étude et du plan d'analyse. Ensuite, j'ai réalisé l'exploration et la préparation de données. J'ai écrit le code et mené les expériences pour entrainer et évaluer les modèles d'apprentissage automatique. Finalement, j'ai effectué les tests statistiques et participé à la rédaction du manuscrit.

3.1 Abstract

The promise of machine learning successfully exploiting digital phenotyping data to forecast mental states in psychiatric populations could greatly improve clinical practice. Previous research focused on binary classification and continuous regression, disregarding the often ordinal nature of prediction targets derived from clinical rating scales. In addition, mental health ratings typically show important class imbalance or skewness that need to be accounted for when evaluating predictive performance. Besides it remains unclear which machine learning algorithm is best suited for forecast tasks, the eXtreme Gradient Boosting (XGBoost) and long short-term memory (LSTM) algorithms being 2 ţwo popular choices in digital phenotyping studies. The CrossCheck dataset includes 6364 mental state surveys using 4-point ordinal rating scales and 23,551 days of smartphone sensor data contributed by patients with schizophrenia. We trained 120 machine learning models to forecast 10 mental states (e.g., Calm, Depressed, Seeing things) from passive sensor data on 2 predictive tasks (ordinal regression, binary classification) with 2 learning algorithms (XGBoost, LSTM) over 3 forecast horizons (same day, next day, next week). A majority of ordinal regression and binary classification models performed significantly above baseline, with macro-averaged mean absolute error values between 1.19 and 0.77, and balanced accuracy

between 58% and 73%, which corresponds to similar levels of performance when these metrics are scaled. Results also showed that metrics that do not account for imbalance (mean absolute error, accuracy) systematically overestimated performance, XGBoost models performed on par with or better than LSTM models, and a significant yet very small decrease in performance was observed as the forecast horizon expanded. In conclusion, when using performance metrics that properly account for class imbalance, ordinal forecast models demonstrated comparable performance to the prevalent binary classification approach without losing valuable clinical information from self-reports, thus providing richer and easier to interpret predictions.

3.2 Author Summary

Symptoms associated with mental health disorders vary greatly over time. Periods of partial remission unfortunately alternate with relapses defined by a marked worsening of symptoms. Hence, assessing future risk and adopting preventive measures is a key challenge for clinical psychiatry. With their many sensors, smartphones can provide novel insights into human behavior outside the medical office. By using machine learning, a branch of artificial intelligence, it is possible to use such smartphone sensor data to predict future mental states and symptoms in psychiatric patients. The present work highlights the importance of predicting fine-grained levels of symptom severity, as commonly reported by patients using so-called ordinal rating scales. Such ordinal predictions were not less accurate than the simplified binary predictions (on/off, high/low) often reported in previous efforts. Besides, we underscore that severe mental states are rare compared to healthy ones, and that this imbalance brings methodological challenges that need to be taken into account to develop valid predictive models.

3.3 Introduction

Although severe psychiatric disorders such as schizophrenia are often chronic, they are also notoriously temporally dynamic, with the severity of symptoms varying over time (1,2). In a uniquely individual way, periods of partial remission alternate with recurrent relapses defined by a marked worsening of symptoms (3–5). Monitoring a patient's symptom trajectory and predicting future risks are therefore key clinical tasks in order to implement required preventive measures (6). Unfortunately, routine medical appointments provide too few and distant

observations to adequately monitor complex individual temporal dynamics (7,8). Additionally, clinical information is primarily collected through interviews in the medical office, which has limited generalizability to the patient's day-to-day life (9) and heavily depends on the partly flawed patient's memory (10). Digital phenotyping holds promise in this regard, as it allows continuously characterizing human behavior and mental health outside the medical environment using smartphones (11,12). First, patients can use their device to periodically rate their symptoms on clinical scales as their daily life unfolds, the ability of remotely tracking symptoms over time leading to improved clinical outcomes (13). Second, digital phenotyping leverages passive data from the device's sensors (e.g., Wi-Fi, Bluetooth, GPS, accelerometer) to render rich facets of behavior. For instance, sedentariness may be associated with low GPS activity, and sleep disruption may be represented by nightly phone unlocks. Machine learning plays a key role in transforming this unprecedented volume and granularity of data into insights into mental health (14). Critically, machine learning may be further exploited to develop models that accurately predict future fluctuating symptoms (e.g., frequency of hallucinations) and acute events (e.g. hospitalisations), which could improve clinical practice in the future (15–17).

Garcia-Ceja et al. (18) distinguish three types of studies that demonstrate the relevance of digital phenotyping to the characterization of mental illness: association studies merely explore the statistical relationships between inputs (e.g. sensor data) and a target (symptom level); detection studies use inputs to predict with machine learning the target at the current time, akin to diagnosis; and forecasting studies use inputs to predict with machine learning a target in the future, similar to prognosis. First, research validated the presence of statistical associations between passive sensing data and mental states, both in healthy (19) and clinical populations (20,21). For instance, relevant digital markers can be established to differentiate healthy from clinical populations for accelerometry (22) and mobility (23,24) features, as highlighted in a review of 46 studies on this topic (25). Second, several studies have demonstrated that mental health-related outcomes can be successfully predicted from smartphone sensor data using machine learning. In major depression, participants with an established diagnosis can be classified from non-depressed participants (26) and the absence or presence of specific symptoms can be predicted in this clinical population (27). Similar works led to symptom-level

detection in bipolar disorder (28) and schizophrenia samples (29). The binary accuracy of such predictive models ranged from 65% to 98% across 40 studies (30). The large majority of these studies used supervised machine learning, either classification or regression, with gradient boosted decision trees, support vector machines, linear models, and neural networks being the most commonly used algorithms, in that order. Third, forecast studies providing predictions about future health outcomes have also been published, although they are scarcer than association and detection studies. The feasibility of predicting future mood and stress in a healthy population has been replicated a few times (31,32). Similar approaches were successful in predicting clinical scale scores and specific psychiatric symptoms for depression (33), bipolar disorder (34), and anxiety (35). The predictive task was either binary classification (i.e., low/high categories) or continuous regression (i.e., an outcome score), with the forecast horizon extending up to a week in the future. Most of the aforementioned forecast studies investigated the predictive performance of recurrent neural networks amongst other machine learning algorithms, in line with the idea that these types of algorithms are best suited for a forecasting task given their ability to model long-term dependencies and latent variables (36,37). Despite their success in detection studies, gradient boosted decision trees models were not thoroughly investigated for forecasting.

To date, detection and forecasting studies have focused on solving binary classification or continuous regression tasks even though the target of the prediction often comes from ordinal rating scales (18,25). Consequently, the resulting binary or real-numbered predictions do not match the ordinal scale interpretation guidelines nor refer to well-defined constructs, leaving key clinical information behind. Previous work evaluated XGBoost models on the same dataset for the tasks of binary classification, continuous regression, and multiclass classification (38). While multiclass classification preserves the original response items, it loses their ordering and faces the rank inconsistency problem (39). Ordinal regression (or ordinal classification) models preserve classes and ordering resulting in rank-consistent discrete predictions easy to interpret with existing validated guidelines. Implicitly, binary classification and continuous regression are often used to mitigate the effect of the small number of examples per class (i.e., class imbalance) on performance. Alas, the data processing inequality from information theory states that variable transformations such as binarization cannot increase the variable's information content (40,41).

Possible gains in predictive performance come at the cost of solving a problem that ignores nuances of the collected data. Still, transforming ordinal scale ratings into a binary target may be a well-motivated modelling decision if done based on a scale's interpretation guidelines and not merely to simplify the predictive task or reduce class imbalance (26,42). For all learning tasks, dedicated evaluation metrics are required to properly evaluate model performance when dealing with class imbalance (43).

Our first objective was to assess the potential performance cost of using ordinal regression compared to binary classification to forecast future mental states, using passive sensing data exclusively. We investigated the potential mediating effect of binarization on the relationship between class imbalance and performance. Our second objective was to provide a comprehensive benchmark of recurrent neural networks and gradient boosted decision trees models for digital phenotyping forecasting to question the implicitly assumed superiority of the former, while systematically exploring the effect of the forecast horizon on predictive performance.

3.3 Methods

3.3.1 Dataset

We obtained the publicly available de-identified data from the CrossCheck study (44,45). This digital phenotyping dataset was collected as part of a randomized controlled trial (clinical trial registration: ClinicalTrials.gov, #NCT01952041) conducted at the Zucker Hillside Hospital in New York City, New York. Ethics approval was obtained from the institutional review boards of Dartmouth College (#24356) and North Shore-Long Island Jewish Health System (#14-100B), and all psychiatric outpatients provided informed consent to participate. Inclusion criteria were a diagnosis of schizophrenia, schizoaffective disorder, or psychosis not otherwise specified; 18 years of age; a significant psychiatric event such as inpatient psychiatric hospitalization or psychiatric hospital emergency room visit within the last 12 months. Data used in the present project comes from 62 patients assigned to the smartphone arm of the clinical trial. They were provided with a Samsung Galaxy S5 Android smartphone on which a mobile app continuously

collected passively sensed data for up to one year. Previous works have reported analyses based on the CrossCheck smartphone data (29,44,45,46–49).

A series of high-level passive sensing features were made available in the dataset (Table 1). Features were computed daily and separately for 6-hour periods: morning (6am-12pm), afternoon (12pm-6pm), evening (6pm-12pm) and night (12am-6am). In total, 23,551 days of passive sensing data were available for analysis. Participants were prompted to provide self-reports about their mental states (Table 2) every Monday, Wednesday, and Friday, with only a minority (3%) of self-reports being obtained on other days of the week. In each self-report survey, 10 distinct items asked the participant about a particular mental state over the recent past, as rated on a 4-point ordinal scale ("Not at all", "A little", "Moderately", "Extremely"). There were 5 positive items for which a high score describes a positive outcome, and 5 negative items for which a high score describes a negative outcome. A total of 6364 surveys were completed, corresponding to 63,640 mental state items being rated.

Tableau 1 *Passive sensing data*

Features	Source
Duration of detected physical activity, informing on sedentary behavior and modes of transportation (in vehicle, on bike, on foot, walking, running, still, tilting, unknown).	Accelerometer
Sleep patterns: start time / end time and sleep duration.	Accelerometer, light sensor, microphone
Location: the distance travelled and number of distinct locations visited were computed.	GPS
Phone usage: lock/unlock frequency and duration as well as the number and duration of incoming or outgoing calls, missed calls or SMS, which may be indicative of social interactions but not their content.	Phone metadata as well as call and SMS logs
Number and duration of ambient conversations, which inform on the presence of people around the phone owner but does not necessarily involve his/her active participation	Microphone
Time: day of the week, day of the month, weekend (yes/no) which reflect important cycles that structure our lives.	Clock

Note. High-level features available in the CrossCheck dataset come from various sources. Features extracted from sensors were indicative of physical activity, sleep patterns, mobility and social interactions, among other things. Features were computed separately for multiple consecutive 6-hour periods.

Tableau 2 *Surveys*

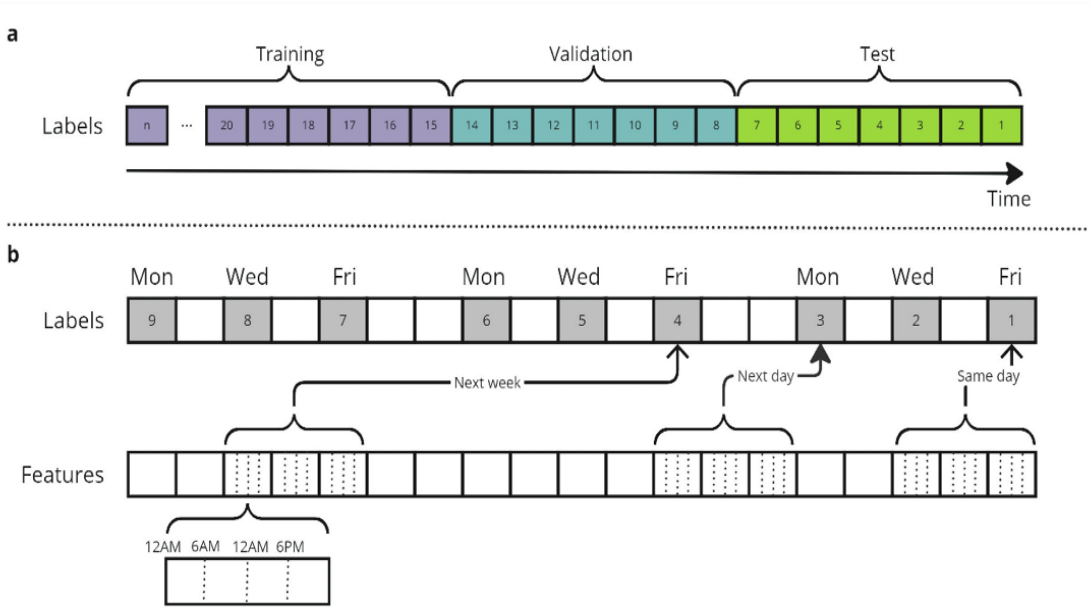
Item name	Item question: “Have you been…”	Valence	Lower	Higher
Calm	feeling calm	Positive	0, 1, 2	3
Hopeful	hopeful about the future	Positive	0, 1, 2	3
Sleep	sleeping well	Positive	0, 1, 2	3
Social	social	Positive	0, 1	2, 3
Think	able to think clearly	Positive	0, 1, 2	3
Depressed	depressed	Negative	0	1, 2, 3
Harm	worried about people trying to harm you	Negative	0	1, 2, 3
Seeing things	seeing things other people can't see	Negative	0	1, 2, 3
Stressed	feeling stressed	Negative	0	1, 2, 3
Voices	bothered by voices	Negative	0	1, 2, 3

Note. 10 types of mental states were self-reported on a 4-point ordinal rating scale (Classes/labels: “Not at all” = 0, “A little” = 1, “Moderately” = 2, “Extremely” = 3). High scores indicated a positive outcome for 5 items (e.g., Calm) and a negative outcome for the 5 others (e.g., Depressed). Recoding the original ordinal labels into binary classes consisted in contrasting 1 class against the 3 others, except for one mental state (Social).

3.3.2 Data Processing

The dataset was partitioned into time-based training, validation and test sets that accounted for temporal dependencies in the data (43,50). The test set contained the latest 7 surveys from each participant, the validation set contained the previous 7 surveys, and the training set included all (>7) earlier surveys (Figure 1). Participants with less than 21 surveys in total were excluded. Consequently, the number of participants decreased from 62 to 61 for models predicting the next week horizon. Depending on the forecast horizon, the training set included from 5163 to 5307 surveys while the validation and test sets each included 427 to 434 surveys. By representing each participant equally in the test set, model evaluation was not biased by participants contributing more data. As a trade-off, the number of training examples per participant varied importantly from 9 to 181 (median = 90.5, interquartile range = 74.5). The training and validation sets were used for model development (preliminary experiments, hyperparameter tuning, etc.) while the test set served for final model performance evaluation. Since our splitting strategy does not control for distribution drift (51,52), we assessed distribution variation across time splits, especially for rarer classes.

Figure 1 *Time-based splitting strategy and forecast horizons*



Note. **a.** The test set contained the latest 7 surveys from each participant, the validation set contained the previous 7 surveys, and the training set included all (>7) earlier surveys **b.** Each label to predict was paired with 3 days of input data (12 6-hour periods), separately for 3 forecast horizons (same day, next day, next week).

Since the CrossCheck dataset includes high-level features extracted from raw sensor data, our preprocessing pipeline primarily served to ensure the XGBoost (53) and LSTM (54) models received equivalent input information while meeting their respective requirements. After dataset splitting, features were standardized at the group level to Gaussian-like distributions using the Yeo-Johnson method (55). While tree-based methods such as the XGBoost algorithm are insensitive to scaling transformations (56), this preprocessing step helps LSTM models converge (57). For LSTM models, each self-report was paired with a sequence of 3 consecutive days divided in 6-hour periods of passive sensing data. Given surveys were mostly completed on some specific days, there was an over-representation of some days in the passive sensing input sequences, as a function of the forecast horizons and targets. For XGBoost models, input sequences were reshaped into tabular format.

3.3.3 Forecasting

We aimed to predict future self-reported mental states using passive smartphone data exclusively. Past self-reports were not included in the models' input contrary to predictive models described in some previous work (31–33). In total, 120 distinct machine learning models were trained. We forecasted 10 mental states (Table 2) over 3 forecast horizons (same day, next day, next week) with 2 machine learning algorithms (XGBoost, LSTM) on 2 predictive tasks (ordinal regression, binary classification). The forecast horizon was the time gap between the input data and the predicted label, which increased from 0 day to 1 day and 7 days (Figure 1). XGBoost and other gradient boosted decision trees algorithms were successful for regression of current day and future self-report aggregates (45,48), and future clinical scale ratings (58) on Crosscheck data. LSTM models and other recurrent neural networks have also provided accurate forecasts of mood and stress in healthy subjects (31,32) and of depressive states in self-identified depressed individuals (33).

3.3.4 Learning Task

3.3.4.1 Ordinal Regression

Like multiclass classification, ordinal regression involves multiple discrete classes, and like continuous regression, it considers an ordering of values. To predict values from [0, 1, 2, 3] corresponding to "Not at all", "A little", "Moderately" and "Extremely", our XGBoost implementation simultaneously learned a continuous regression task and tuned the default thresholds [0.5, 1.5, 2.5] to discretize the continuous predicted values into [0, 1, 2, 3]. For LSTM models, we used the Conditional Ordinal Regression for Neural networks approach and their open source implementation in coral-pytorch (39). The neural network architecture allows a single model to decompose the ordinal regression task of predicting values [0, 1, 2, 3] into 3 independent binary tasks of predicting >0, >1, and >2. XGBoost and LSTM models were respectively trained using the regression squared loss and the conditional ordinal regression loss for neural networks. The performance of both model types was optimized for the macro-averaged mean absolute error (MAMAE), which is robust to class imbalance (59) observed in the CrossCheck dataset. This metric computes the mean absolute error (MAE) per class then averages

results, giving equal weight to each class (S1 Appendix). For the sake of comparison with previous works, models were also evaluated using regular MAE which does not appropriately handle class imbalance (18).

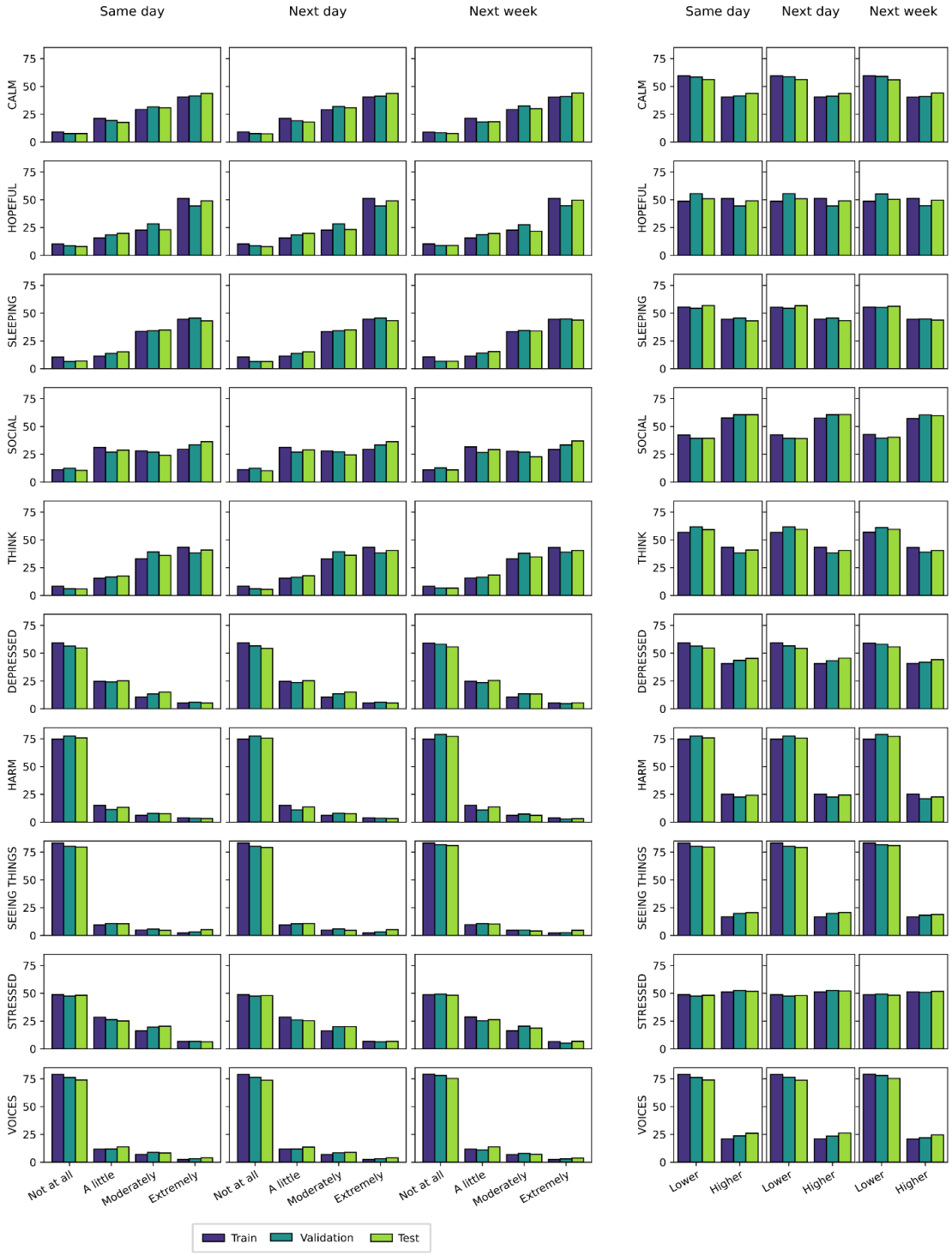
3.3.4.2 Binary Classification

The original 4 classes ("Not at all", "A little", "Moderately" and "Extremely") were binarized using the cutoff resulting in the two best-balanced classes. Due to the skewed nature of the original label distributions, this consisted in contrasting one class against the other 3, except for one variable. Specifically, "Extremely" was converted to "Higher" and other labels to "Lower" for positive items, while "Not at all" was converted to "Lower" and all other labels to "Higher" for negative items. Due to a flatter distribution, the variable Social had "Extremely" and "Moderately" converted to "Higher" and the other two values to "Lower". Binary classes remained imbalanced to some degree, very much so in some instances (Figure 2). Both XGBoost and LSTM models were trained using the binary cross entropy loss and evaluated with balanced accuracy (BAcc) to deal with class imbalance (60–62). BAcc is the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate). For the sake of comparison with a large part of the relevant literature (30), models were also evaluated using regular accuracy (Acc).

3.3.5 Model Training and Validation

For each XGBoost model, the development phase included an automated 10 rounds of hyperparameter optimization using the train and validation sets. Once the best hyperparameters were determined, the final model was trained on both the train and validation sets and evaluated on the test set. For each LSTM model, the PyTorch Lightning Tune functionality was used at the beginning of training to find the optimal learning rate and batch size (63). Manual hyperparameter tuning was done to reach a set of hyperparameters used in all LSTM models: 150 epochs and 1 LSTM layer with 128 nodes and 10% dropout. While this architecture may be considered to have lower capacity, it is well in line with previous relevant works (31–33).

Figure 2 *Label distributions across mental states and dataset splits*



Note. The left and right panels show the distributions of the original ordinal labels and recoded binary labels, respectively. Each row is associated with a mental state and each inner column with a forecast horizon. Bar plots display the proportion (%) of examples for the 4 or 2 classes (labels) in each dataset split.

3.3.6 Statistical Testing

Statistical analyses were conducted to assess if models performed significantly above baseline and compare them across algorithms, forecast horizons, mental states, and learning tasks. For the 60 conditions (2 tasks, 3 forecast horizons, 10 mental states), a baseline distribution of performance scores (MAMAE or BAcc) was generated using a Monte Carlo method drawing 1000 samples with replacement and computer performance against the test set for each (S2 Appendix). A model significantly outperformed the baseline if its test performance was better than the baseline quantile corresponding to a p value < .05 with a Bonferroni correction for 120 models ($p < 4.2 \times 10\text{-}4$). Considering ordinal regression and binary classification tasks separately, model performances on the test set were compared between the XGBoost and LSTM algorithms using the non-parametric Wilcoxon signed-rank test and across forecast horizons (3 horizons) and mental states (10 variables) with non-parametric Friedman tests (51,64). To compare performance between the ordinal regression and binary classification tasks, the MAMAE and BAcc values were transformed to a scale normalized balanced error which ranges from 0 to 1 (S1 Appendix). The relationship between the scale normalized balanced error of ordinal and binary models was compared to a perfect correlation and the residuals were inspected for discrepancies between the two learning tasks. Finally, the effect of class imbalance was assessed using the Spearman rank correlation between the predictive performance and the class imbalance of each of the 10 mental states. Class imbalance was quantified by the difference between the number of examples of the majority and the minority class, normalized by the total number of examples (0 to 1 range). The same definition was applied for ordinal regression and binary classification.

3.4 Results

3.4.1 Descriptive Statistics

Participants tended to rate more frequently high (“Extremely”) on positive items (Calm, Hopeful, Sleeping, Social, Think) and low (“Not at all”) on negative items (Depressed, Harm, Seeing Things, Stressed, Voices) (Figure 2). With 76% to 81% of labels belonging to “Not at all” and only 3% to “Extremely”, the class imbalance for the negative items Harm, Voices and Seeing Things was major. A lesser but still considerable imbalance between majority (33% to 48%) and minority (7

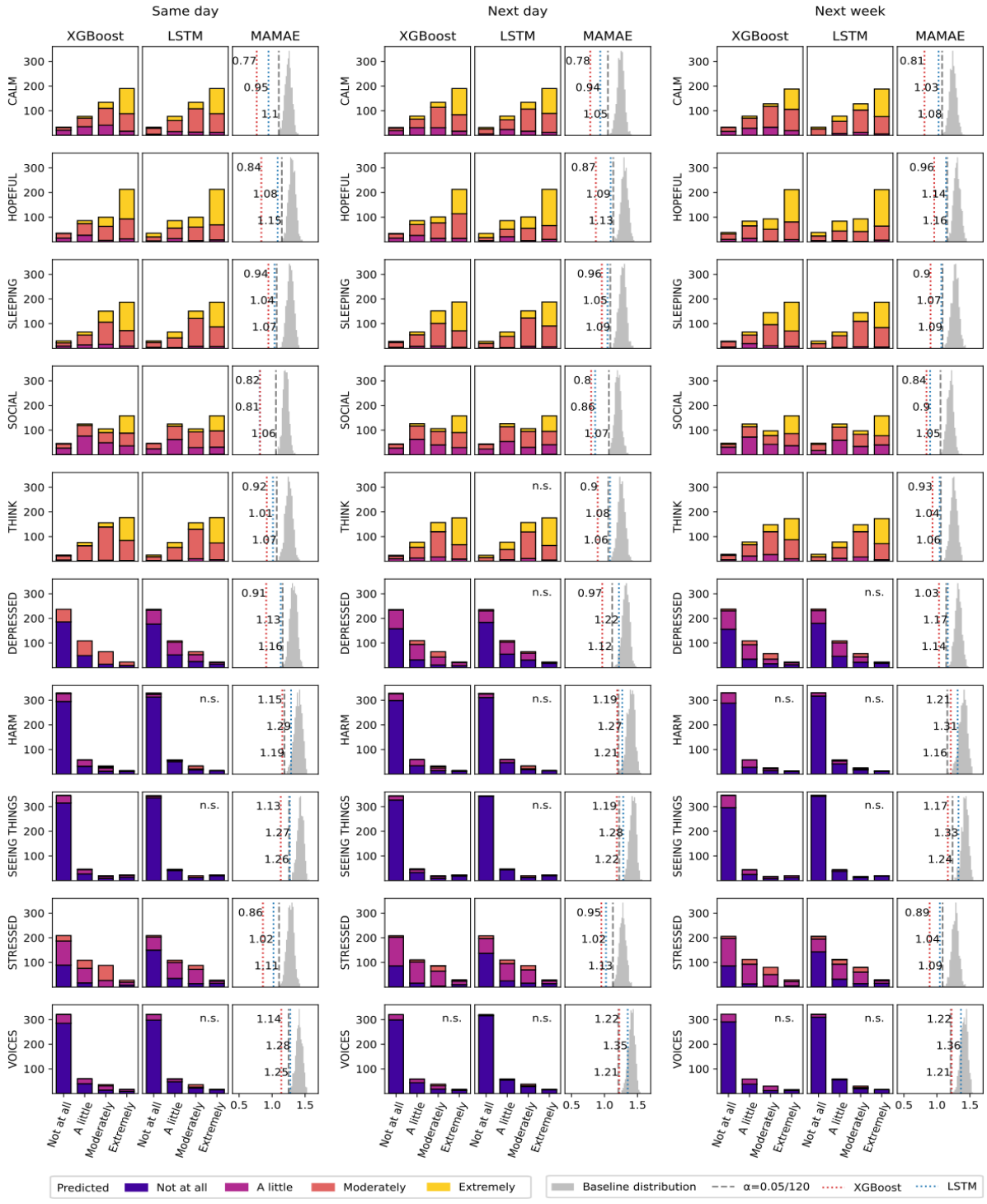
to 11%) classes was observed for positive items. Recoding ordinal classes into binary classes reduced imbalance, yet it remained very large in some cases (Harm, Voices and Seeing Things) with 76-81% for the majority class ("Lower") and only 19-24% for the minority class ("Higher"). Upon visual inspection, no notable distribution shifts were observed between the training, validation and test sets created via time-based splitting (Figure 2). The negligible variation between splits suggests they are representative of the full dataset and the validation split will properly estimate test performance.

3.4.2 Forecasting performance

For ordinal regression, 45 out of 60 models across algorithms, forecast horizons and mental states performed significantly above baseline (Bonferroni corrected P<.05) (Figure 3). Non-significant models were associated with negative mental states (Depressed, Harm, Seeing things, Stressed, Voices) and the LSTM algorithm. MAMAE values ranged from 1.36 to 0.77 (median = 1.04) across all models, with the 45 significant models not exceeding MAMAE = 1.19 (median = 0.96). There was a significant difference in performance between the 2 algorithms (Z = 1, P<.001) with XGBoost models (median = 0.94) performing better than LSTM models (median = 1.08). Similarly, a significant effect of forecast horizon was observed (Q = 19.9, P<.001) although the decrease in performance as the horizon increased from same day (median = 1.02) to next day (median = 1.03) then next week (median = 1.04) was very small.

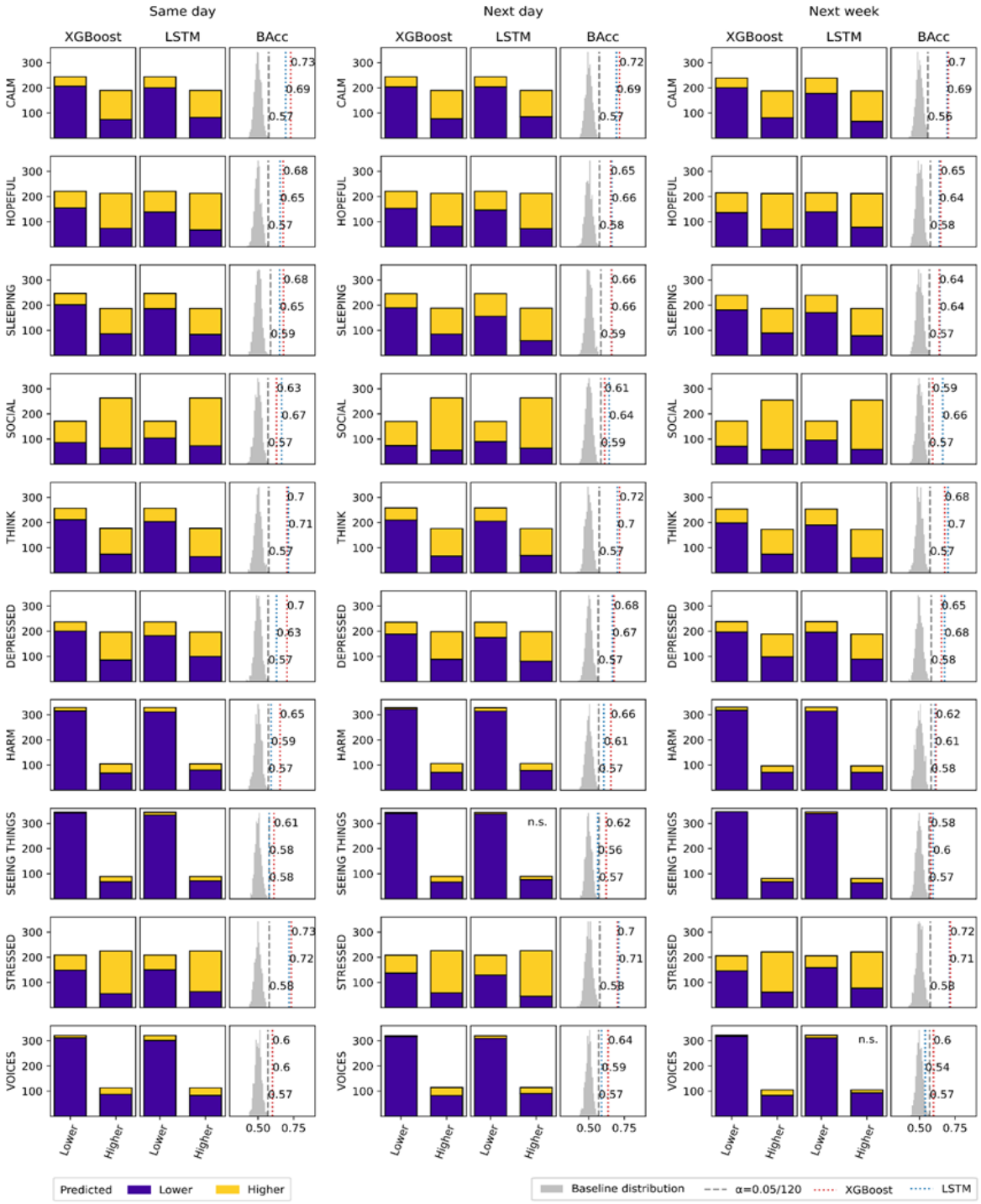
On the binary classification task, 58 out of 60 models were significantly superior to baseline (Bonferroni corrected P<.05) (Figure 4). BAcc values ranged from 54% to 73% (median = 66%) with significant models all performing above 58%. Contrary to ordinal regression, no significant difference in performance was found (Z = 150, P=.1) between the XGBoost (median = 66%) and LSTM (median = 66%) algorithms. A significant effect of the forecast horizon was detected (Q = 7.9, P=.02). The decrease in performance over same day (median = 66%), next day (median = 66%), and next week (median = 65%) was consistent with the effect observed for ordinal regression. When comparing the scale normalized balanced error of ordinal regression and binary classification to a perfect correlation, the very low average of residuals (= 0.003) suggests equivalent performance on the two tasks on average, with no predictive task clearly outperforming the other (Figure 5).

Figure 3 *Ordinal regression task performance*



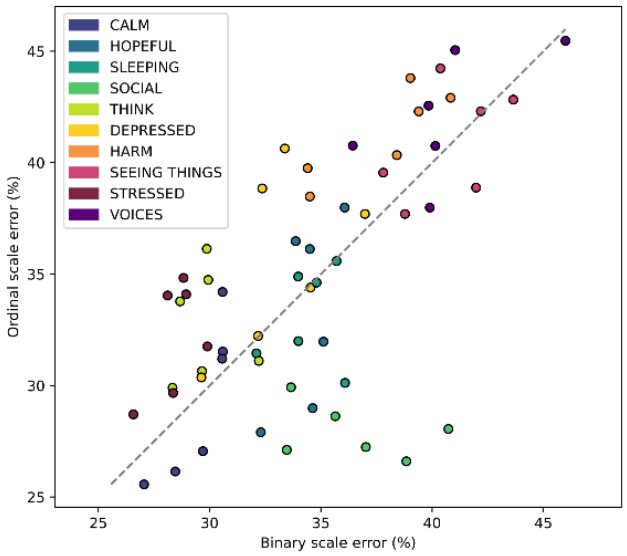
Note. Each row is associated with a mental state and each outer column with a forecast horizon. For each of the 30 conditions, bar plots show XGBoost and LSTM test predictions. The bar height corresponds to the number of examples with each bar indicating the true class and colors reflecting the predicted classes. The MAMAE for these XGBoost and LSTM model predictions are shown against the MAMAE significance threshold (Bonferroni corrected *P*<.05) and the corresponding Monte Carlo baseline distribution. n.s., nonsignificant.

Figure 4 *Binary classification task performance*



Note. Each row is associated with a mental state and each outer column with a forecast horizon. For each of the 30 conditions, bar plots show XGBoost and LSTM test predictions. The bar height corresponds to the number of examples with each bar indicating the true class and colors reflecting the predicted classes. The BAcc for these XGBoost and LSTM model predictions are shown against the BAcc significance threshold (Bonferroni corrected P<.05) and the corresponding Monte Carlo baseline distribution. n.s., nonsignificant.

Figure 5 *Relationship between ordinal regression and binary classification performance*

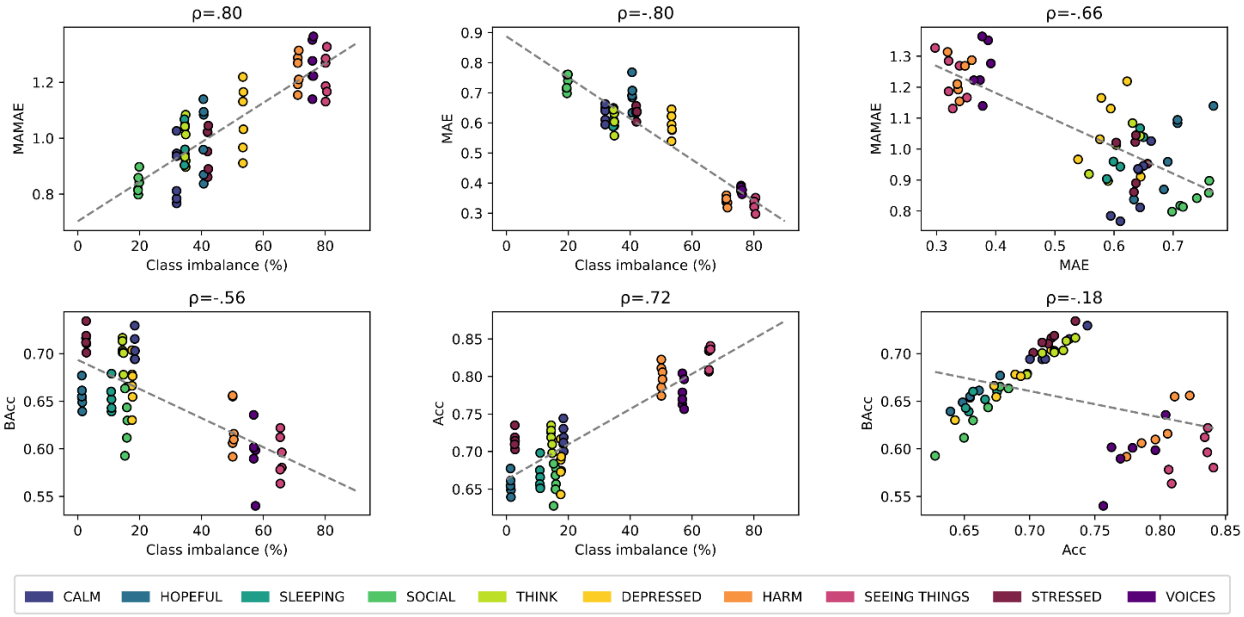


Note. For each of the 60 conditions (10 mental states x 2 algorithms x 3 forecast horizons), the performance of the ordinal regression model (y-axis) is displayed against the binary classification model (x-axis) under the same condition. Balanced performance metrics (MAMAE, BAcc) were normalized to a common scale ranging from 0 to 1 (S1 Appendix).

3.4.3 Class imbalance

Very strong effects on performance were unravelled for mental states, both for ordinal regression (Q = 49.4, P<.001) and binary classification (Q = 47, P<.001) (Figure 3 and Figure 4). Further inspection revealed that the effect of mental states could be explained by their class imbalance, as it correlated with MAMAE values of ordinal regression (Spearman r = .8) and BAcc for binary classification (r = -.56) (Figure 6). Large class imbalance observed for the variables Harm, Voices and Seeing Things, was associated with poor performance (high MAMAE, low BAcc). In stark contrast, opposite effects were observed when using metrics that do not account for class imbalance (MAE, Acc). Indeed, large class imbalance was associated with higher performance, both for the ordinal regression MAE (r=-.8) and the binary classification Acc (r=.72). As a corollary, performance metrics that do and do not account for class imbalance should be negatively correlated, which held for MAMAE and MAE (r = -.66) as well as BAcc and ACC (r=-.18). The latter association was weaker given 7 out of 10 mental states were fairly balanced and their BAcc and Acc strongly positively correlated (r = .95).

Figure 6 *Effect of class imbalance on predictive performance*



Note. Each scatter plot displays the test performance of 60 models under different conditions (10 mental states x 2 algorithms x 3 forecast horizons)-. Rows show the effect of class imbalance on performance for ordinal regression (top) and binary classification (bottom). The left and middle columns respectively highlight the relationship between class imbalance and predictive performance for balanced and unbalanced metrics, and the right column reveals the relationship between the two metrics. Each correlation is quantified by Spearman's ρ (rho).

3.5 Discussion

We show that ordinal regression, which best preserves key clinical information, can forecast self-reported mental states from passive smartphone data with predictive performance levels comparable to those of binary classification. Class imbalance, which may be particularly pronounced for ordinal data, strongly affects model training and performance is inadequately rendered when unfit evaluation metrics are used. The XGBoost performs as well or even better than the LSTM algorithm for forecasting. Increasing the forecast horizon incurred a negligible decrease in performance.

While mental states are often collected on ordinal rating scales in digital phenotyping research, the majority of past studies formulated predictions using binary classification (24,26,32) or continuous regression (29,48,58). Ordinal models adequately preserve the order of discrete classes without assuming continuity between them. Importantly, ordinal predictions exist on the data collection scale and can be interpreted by clinicians using the scale's interpretation

guidelines. Critically, our findings demonstrate that using an ordinal regression modelling that best meets this clinical motivation does not lead to any systematic cost in predictive performance compared to using binary classification.

The rarity of clinically relevant mental-health events (e.g., psychotic episode) leads to datasets composed mainly of healthy examples, both in terms of inputs and labels. In this study, mental states with larger class imbalance (e.g., Harm, Voices, Seeing things) were associated with lower performance, a typical challenge in machine learning (65). Since imbalance is a core property of mental health data, it is best to leave it unadjusted and use adapted methods (62). The majority of past digital phenotyping studies only report performance metrics that do not account for class imbalance (e.g., Acc, root mean squared error, MAE) (30), thereby allowing models that predict only the majority class to appear deceptively good. To overcome these issues, we used metrics that account for class imbalance (BAcc and MAMAE) to train and evaluate models. We show that models appear to perform best on variables describing rare events (e.g., Harm, Voices, Seeing things) when using unbalanced metrics, but that relationship is reversed when using the adequate balanced metrics. Consequently, inappropriate metrics can lead to systematically incorrect conclusions when evaluating model performance or the impact of procedures such as feature engineering, feature selection, or model selection. Besides, resampling techniques have been used in past digital phenotyping studies in an attempt to improve performance by mitigating class imbalance (66). However, these methods provide little to no benefits for modern algorithms like XGBoost (67). Instead, tuning the model's decision threshold is a more sensible approach. Resampling provides no performance improvement for binary models of medical diagnosis and deteriorates model calibration, which should be a key performance criterion when a probabilistic interpretation is necessary (e.g., risk scores) (68). Cost-sensitive learning, which incorporates the outcome of a prediction to the learning process, has also been shown to be an effective solution to overcome class imbalance (69).

Recurrent neural networks, in particular the LSTM approach, are specialized for sequence data such as sensor data (36,37) and are popular forecasting algorithms in digital phenotyping (32,33,35,70). On the other hand, gradient boosted decision trees have been consistently successful in diagnostic studies (45,58,71,72), but few investigated it for forecasting (48,73). Our

results show that XGBoost models are equally capable to LSTM models for forecasting, and even superior under certain conditions. Gradient boosted decision trees were previously found to be superior to neural networks on tabular datasets with skewed features, uninformative features, or rare classes (74,75), all typical characteristics of digital phenotyping. Given similar levels of performance, algorithms should be selected according to other practical implications such as explainability, the amount of required training data, or computational costs.

Increasing the forecast horizon from “same day” to “next week” was associated with a negligible performance degradation, in line with previous findings (31,34). Given the strong weekly seasonality in behavior and mental health self-reports (31,33), one should proceed with care when extrapolating performance to different days of the week. For instance, the CrossCheck study collected self-reports primarily on Monday, Wednesday, and Friday, limiting our ability to train more robust models. Besides, increasing the input history length up to 3 weeks could improve model performance (31–33,48,66,76), given early warning signs have been observed up to 30 days prior to symptom worsening in mental illness (46,77,78).

Although the levels of forecast performance we achieved are encouraging, predictions are not sufficiently accurate to consider their implementation in the clinical realm. Future works are thus required to significantly improve the performance of forecast models using digital phenotyping data, whether by exploiting much larger datasets, improving feature engineering, or redefining more optimal clinical targets, among other things. Furthermore, it will be key to explain why a given model makes a specific prediction, a critical subject the present study did not explore. Explainable machine learning systems will indeed be necessary for clinicians to be able to decide whether to trust or not a prediction and to comply with regulations in some jurisdictions (79). A popular technique is to attribute SHapley Additive exPlanations (SHAP) values (80) to determine the contribution of each feature towards a prediction. However, Shapley additive explanations can be misleading since they are detached from prediction certainty. Poorer model calibration associated with imbalanced data decreases their reliability. The calibrated explanations method (81) produces probabilistic prediction intervals and scores each feature based on its contribution towards the prediction and its uncertainty. Empirical studies showed that predictions and explanations need to be grounded in the users’ task and fit their mental model

to be useful (82). For digital phenotyping, this means explanations would benefit from higher-level features that relate to mental health constructs instead of features closer to raw smartphone data, which may be at odds with improving predictive performance.

Our uniform methodology was applied to train and evaluate 120 models and provide a fair and comprehensive benchmark. As a trade-off, there might be better achievable performance on each individual task. For instance, the public version of the CrossCheck dataset was used without further feature engineering or selection. Adding well-crafted features could have benefited XGBoost and LSTM models unevenly since neural networks are more sensitive to uninformative features. On another note, we only investigated group models. Each participant was represented equally in the test set to prevent the number of self-reports biasing the evaluation. However, participants with more training examples may indirectly bias results since per-person models typically perform better (18,45,58).

3.6 Conclusion

The unprecedented volume and richness of data about the individual generated by smartphones opens a unique window onto mental health. Beyond the precise monitoring of psychiatric conditions in everyday life, digital phenotyping data paired with machine learning models allows to forecast future mental states. Accurate prediction of the likely progression of the illness would be key in implementing personalized prevention measures. While research in this field is burgeoning, issues remain to be addressed before such forecast models can be implemented as clinical decision support tools. In this study, we explored modelling approaches that preserve the maximum of clinical information from the collected ordinal data by training ordinal regression instead of binary classification models to forecast mental states. Importantly, we show that the clinically motivated ordinal approaches do not incur a trade-off in predictive performance. Given class imbalance is challenging for learning algorithms and is a core property of mental health data, we argue that using binary classification is an unsatisfactory mitigation method and using evaluation metrics that account for its impact is essential. Finally, we question recurrent neural networks as the de facto superior forecast algorithm and thus encourage a more systematic benchmarking of machine learning algorithms, especially gradient boosted decision trees, to predict future mental states using digital phenotyping data.

3.7 Funding

TJ received student fellowships from the Fonds de recherche du Québec - Santé (#303584) and the Canadian Institute for Health Research. PO was supported by a salary award “chercheur boursier junior 1” of the Fonds de recherche du Québec - Santé (#266630, #280391) and the Courtois foundation through the Courtois NeuroMod project (https://www.cneuromod.ca).

3.8 Acknowledgements

The authors acknowledge the work of the CrossCheck research team in completing the original data collection study and creating the public dataset. We are grateful to Hien Nguyen for his comments on a previous draft of our manuscript.

3.9 Author’s Contributions

TJ and PO contributed to the conceptualization of the study and the analysis plan. TJ contributed to data preparation and exploration. TJ and RGH contributed to the machine learning model training and evaluation. TJ and PO contributed to the statistical analysis and prepared the original draft. PO obtained funding. All authors reviewed the draft and approved the final manuscript.

3.10 Data Availability Statement

De-identified digital phenotyping data from the CrossCheck study are openly available from : https://pbh.tech.cornell.edu/data.html. The Python code used to generate all the results reported in this paper can be obtained from a dedicated GitHub repository : https://github.com/zilto/ordinal-forecasting-digital-phenotyping.

3.11 Competing interests

The authors have declared that no competing interests exist.

3.12 Supporting Information

3.12.1 S1 Appendix: Metric definitions

- TP: True positive
- TN: True negative
- FP: False positive
- FN: False negative
- i: Example index
- j: Class index
- k: Number of classes
- x_i: Input
- y_i: True label
- f: Predictor
- $f(x_i)$: Predicted label
- S: Set of all examples
- S_j: Set of examples of class j

Binary Classification metrics

$$\text{Accuracy (Acc)} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\begin{aligned}\text{Balanced Accuracy (BAcc)} &= \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \\ &= \frac{1}{2}\left(\text{sensitivity} + \text{specificity}\right) \\ &= \frac{1}{2}\left(\text{precision} + \text{recall}\right)\end{aligned}$$

Ordinal regression metrics

$$\text{Mean absolute error (MAE)} = \frac{1}{|S|}\sum_{x_i, y_i \in S} |f(x_i) - y_i|$$

$$\text{Macro-averaged mean absolute error (MAMAE)} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|S_j|}\sum_{x_i, y_i \in S_j} |f(x_i) - y_i|$$

Binary BAcc to MAMAE conversion

$$
\begin{aligned}
\text{BAcc} &= \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \\
&= \frac{1}{2}\left(\frac{\text{TP}}{|S_1|} + \frac{\text{TN}}{|S_2|}\right) \\
&= \frac{1}{2}\left(\frac{\sum_{x_i \in S_1} 1_{\{f(x_i)=1\}}}{|S_1|} + \frac{\sum_{x_i \in S_2} 1_{\{f(x_i)=2\}}}{|S_2|}\right) \\
&= \frac{1}{2}\sum_{j=1}^{2}\left(\frac{\sum_{x_i \in S_j} 1_{\{f(x_i)=j\}}}{|S_j|}\right) \\
&= \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} 1_{\{f(x_i)=j\}} \\
&= \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} (1 - 1_{\{f(x_i)\neq j\}}) \\
&= \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} 1 - \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} 1_{\{f(x_i)\neq j\}} \\
&= \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}|S_j| - \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} 1_{\{f(x_i)\neq j\}} \\
&= 1 - \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} 1_{\{f(x_i)\neq j\}} \\
&= 1 - \frac{1}{2}\sum_{j=1}^{2}\frac{1}{|S_j|}\sum_{x_i \in S_j} |f(x_i) - y_i| && y \in \{0,1\} \\
&= 1 - \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|S_j|}\sum_{x_i \in S_j} |f(x_i) - y_i| && k = 2 \\
&= 1 - \text{MAMAE} \\
1 - \text{BAcc} &= \text{MAMAE}
\end{aligned}
$$

$$
\text{Scaled normalized balanced error} = \frac{\text{MAMAE}}{k-1}
$$

Class imbalance

$$
\begin{aligned}
S_{maj}, S_{min} &= \max\left(\{|S_1|, |S_2|, \ldots, |S_k|\}\right), \min\left(\{|S_1|, |S_2|, \ldots, |S_k|\}\right) \\
\text{class imbalance} &= \frac{|S_{maj}| - |S_{min}|}{k-1}
\end{aligned}
$$

3.12.2 S2 Appendix: Pseudo code for the Monte Carlo baseline distribution

Inputs

Train: Array of training labels

Test: Array of testing labels

f(true, prediction): Metric to compute on true labels and predictions (MAMAE or BAcc)

k: Number of samples

n: Sample size

Output

Array of k values of metric f()

Procedure

```
k = 1000

n = length(Test)

distribution = []  # empty array

for i in k samples:

    sample := draw a sample of n items with replacement from array Train

    metric := f(Test, sample)

    distribution.append(metric)

return distribution
```

3.13 References

1. Wright AGC, Woods WC. Personalized Models of Psychopathology. 2020;29.

2. Nelson B, McGorry PD, Wichers M, Wigman JTW, Hartmann JA. Moving From Static to Dynamic Models of the Onset of Mental Disorder: A Review. JAMA Psychiatry. 2017 May 1;74(5):528.

3. Burcusa SL, Iacono WG. Risk for recurrence in depression. Clin Psychol Rev. 2007 Dec;27(8):959–85.

4. Koopmans PC, Bültmann U, Roelen CAM, Hoedeman R, Van Der Klink JJL, Groothoff JW. Recurrence of sickness absence due to common mental disorders. Int Arch Occup Environ Health. 2011 Feb;84(2):193–201.

5. Emsley R, Chiliza B, Asmal L, Harvey BH. The nature of relapse in schizophrenia. BMC Psychiatry. 2013 Dec;13(1):50.

6. Cohen AS, Fedechko T, Schwartz EK, Le TP, Foltz PW, Bernstein J, et al. Psychiatric Risk Assessment from the Clinician's Perspective: Lessons for the Future. Community Ment Health J. 2019 Oct;55(7):1165–72.

7. Insel TR. Digital Phenotyping: Technology for a New Science of Behavior. JAMA. 2017 Oct 3;318(13):1215.

8. Chiauzzi E, Wicks P. Beyond the Therapist's Office: Merging Measurement-Based Care and Digital Medicine in the Real World. Digit Biomark. 2021 Jul 29;5(2):176–82.

9. Mouchabac S, Conejero I, Lakhlifi C, Msellek I, Malandain L, Adrien V, et al. Improving clinical decision-making in psychiatry: implementation of digital phenotyping could mitigate the influence of patient's and practitioner's individual cognitive biases. Dialogues Clin Neurosci. 2021 Jan 1;23(1):52–61.

10. Rogler LH, Mroczek DK, Fellows M, Loftus ST. The Neglect of Response Bias in Mental Health Research. J Nerv Ment Dis. 2001 Mar;189(3):182–7.

11. Onnela JP, Rauch SL. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Neuropsychopharmacology. 2016 Jun;41(7):1691–6.

12. Torous J, Gershon A, Hays R, Onnela JP, Baker JT. Digital Phenotyping for the Busy Psychiatrist: Clinical Implications and Relevance. Psychiatr Ann. 2019 May 1;49(5):196–201.

13. Goldberg SB, Buck B, Raphaely S, Fortney JC. Measuring Psychiatric Symptoms Remotely: a Systematic Review of Remote Measurement-Based Care. Curr Psychiatry Rep. 2018 Oct;20(10):81.

14. Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annu Rev Clin Psychol. 2017 May 8;13(1):23–47.

15. Chen ZS, Kulkarni P (Param), Galatzer-Levy IR, Bigio B, Nasca C, Zhang Y. Modern views of machine learning for precision psychiatry. Patterns. 2022 Nov;3(11):100602.

16. Hauser TU, Skvortsova V, De Choudhury M, Koutsouleris N. The promise of a model-based psychiatry: building computational models of mental ill health. Lancet Digit Health. 2022 Nov;4(11):e816–28.

17. Soyiri IN, Reidpath DD. An overview of health forecasting. Environ Health Prev Med. 2013 Jan;18(1):1–9.

18. Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive Mob Comput. 2018 Dec;51:1–26.

19. DaSilva AW, Huckins JF, Wang R, Wang W, Wagner DD, Campbell AT. Correlates of Stress in the College Environment Uncovered by the Application of Penalized Generalized Estimating Equations to Mobile Sensing Data. JMIR MHealth UHealth. 2019 Mar 19;7(3):e12084.

20. Henson P, D'Mello R, Vaidyam A, Keshavan M, Torous J. Anomaly detection to predict relapse risk in schizophrenia. Transl Psychiatry. 2021 Jun;11(1):28.

21. Ranjan T, Melcher J, Keshavan M, Smith M, Torous J. Longitudinal symptom changes and association with home time in people with schizophrenia: An observational digital phenotyping study. Schizophr Res. 2022 May;243:64–9.

22. Strauss GP, Raugh IM, Zhang L, Luther L, Chapman HC, Allen DN, et al. Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia. Schizophrenia. 2022 Apr 15;8(1):37.

23. Depp CA, Bashem J, Moore RC, Holden JL, Mikhael T, Swendsen J, et al. GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. Npj Digit Med. 2019 Nov 8;2(1):108.

24. Faurholt-Jepsen M, Busk J, Rohani DA, Frost M, Tønning ML, Bardram JE, et al. Differences in mobility patterns according to machine learning models in patients with bipolar disorder and patients with unipolar disorder. J Affect Disord. 2022 Jun;306:246–53.

25. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. JMIR MHealth UHealth. 2018 Aug 13;6(8):e9691.

26. Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study. JMIR MHealth UHealth. 2021 Jul 12;9(7):e26540.

27. Ware S, Yue C, Morillo R, Lu J, Shang C, Bi J, et al. Predicting depressive symptoms using smartphone data. Smart Health. 2020 Mar;15:100093.

28. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker SA, et al. Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata: A BiAffect Digital Phenotyping Study. J Med Internet Res. 2018 Jul 20;20(7):e241.

29. Tseng VWS, Sano A, Ben-Zeev D, Brian R, Campbell AT, Hauser M, et al. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. Sci Rep. 2020 Sep 15;10(1):15100.

30. Lee K, Lee TC, Yefimova M, Kumar S, Puga F, Azuero A, et al. Using digital phenotyping to understand health-related outcomes: A scoping review. Int J Med Inf. 2023 Jun;174:105061.

31. Spathis D, Servia-Rodriguez S, Farrahi K, Mascolo C, Rentfrow J. Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining [Internet]. Anchorage AK USA: ACM; 2019 [cited 2021 May 2]. p. 2886–94. Available from: https://dl.acm.org/doi/10.1145/3292500.3330730

32. Umematsu T, Sano A, Picard RW. Daytime Data and LSTM can Forecast Tomorrow's Stress, Health, and Happiness. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. Berlin, Germany: IEEE; 2019 [cited 2021 Jan 10]. p. 2186–90. Available from: https://ieeexplore.ieee.org/document/8856862/

33. Suhara Y, Xu Y, Pentland A "Sandy." DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In: Proceedings of the 26th International Conference on World Wide Web [Internet]. Perth Australia: International World Wide Web Conferences Steering Committee; 2017 [cited 2021 Sep 12]. p. 715–24. Available from: https://dl.acm.org/doi/10.1145/3038912.3052676

34. Busk J, Faurholt-Jepsen M, Frost M, Bardram JE, Vedel Kessing L, Winther O. Forecasting Mood in Bipolar Disorder From Smartphone Self-assessments: Hierarchical Bayesian Approach. JMIR MHealth UHealth. 2020 Apr 1;8(4):e15028.

35. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther. 2022 Feb;149:104013.

36. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. Mol Psychiatry. 2019 Nov;24(11):1583–98.

37. Koppe G, Guloksuz S, Reininghaus U, Durstewitz D. Recurrent Neural Networks in Mobile Sampling and Intervention. Schizophr Bull. 2019 Mar 7;45(2):272–6.

38. Choudhary S, Thomas N, Alshamrani S, Srinivasan G, Ellenberger J, Nawaz U, et al. A Machine Learning Approach for Continuous Mining of Nonidentifiable Smartphone Data to Create a Novel Digital Biomarker Detecting Generalized Anxiety Disorder: Prospective Cohort Study. JMIR Med Inform. 2022 Aug 30;10(8):e38943.

39. Shi X, Cao W, Raschka S. Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities [Internet]. arXiv; 2022 [cited 2023 Apr 27]. Available from: http://arxiv.org/abs/2111.08851

40. Beaudry NJ, Renner R. An intuitive proof of the data processing inequality. Quantum Inf Comput. 2012 May;12(5 & 6):432–41.

41. Cover TM, Thomas JA. ELEMENTS OF INFORMATION THEORY. Hoboken, New Jersey.: Wiley Interscience; 2006.

42. Palmius N, Saunders KEA, Carr O, Geddes JR, Goodwin GM, De Vos M. Group-Personalized Regression Models for Predicting Mental Health Scores From Objective Mobile Phone Data Streams: Observational Study. J Med Internet Res. 2018 Oct 22;20(10):e10194.

43. Varoquaux G, Colliot O. Evaluating Machine Learning Models and Their Diagnostic Value. In: Colliot O, editor. Machine Learning for Brain Disorders [Internet]. New York, NY: Springer US; 2023 [cited 2023 Nov 17]. p. 601–30. (Neuromethods; vol. 197). Available from: https://link.springer.com/10.1007/978-1-0716-3195-9_20

44. Ben-Zeev D, Brian R, Wang R, Wang W, Campbell AT, Aung MSH, et al. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. Psychiatr Rehabil J. 2017 Sep;40(3):266–75.

45. Wang R, Aung MSH, Abdullah S, Brian R, Campbell AT, Choudhury T, et al. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing [Internet]. Heidelberg Germany: ACM; 2016 [cited 2021 Jan 10]. p. 886–97. Available from: https://dl.acm.org/doi/10.1145/2971648.2971740

46. Ben-Zeev D, Brian R, Campbell A, Scherer E, Hauser M, Kane J. CrossCheck [Internet]. 2020. Available from: https://pbh.tech.cornell.edu/data.html

47. Adler DA, Ben-Zeev D, Tseng VWS, Kane JM, Brian R, Campbell AT, et al. Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks. JMIR MHealth UHealth. 2020 Aug 31;8(8):e19962.

48. Buck B, Scherer E, Brian R, Wang R, Wang W, Campbell A, et al. Relationships between smartphone social behavior and relapse in schizophrenia: A preliminary report. Schizophr Res. 2019 Jun;208:167–72.

49. He-Yueya J, Buck B, Campbell A, Choudhury T, Kane JM, Ben-Zeev D, et al. Assessing the relationship between routine and schizophrenia symptoms with passively sensed measures of behavioral stability. Npj Schizophr. 2020 Nov 23;6(1):35.

50. Zhou J, Lamichhane B, Ben-Zeev D, Campbell A, Sano A. Predicting Psychotic Relapse in Schizophrenia With Mobile Sensor Data: Routine Cluster Analysis. JMIR MHealth UHealth. 2022 Apr 11;10(4):e31006.

51. Hewamalage H, Ackermann K, Bergmeir C. Forecast evaluation for data scientists: common pitfalls and best practices. Data Min Knowl Discov. 2023 Mar;37(2):788–832.

52. Thiagarajan JJ, Rajan D, Sattigeri P. Understanding Behavior of Clinical Models under Domain Shifts [Internet]. arXiv; 2019 [cited 2023 Nov 8]. Available from: http://arxiv.org/abs/1809.07806

53. Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F. Characterizing Concept Drift. Data Min Knowl Discov. 2016 Jul;30(4):964–94.

54. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. 2016 [cited 2023 Nov 3]. p. 785–94. Available from: http://arxiv.org/abs/1603.02754

55. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735–80.

56. Raymaekers J, Rousseeuw PJ. Transforming variables to central normality. Mach Learn [Internet]. 2021 Mar 21 [cited 2023 Jul 26]; Available from: http://link.springer.com/10.1007/s10994-021-05960-5

57. Muller A, Guido S. Introduction to machine learning with python. O'Reilly Media; 2016.

58. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

59. Wang R, Wang W, Aung MSH, Ben-Zeev D, Brian R, Campbell AT, et al. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2017 Sep 11;1(3):1–24.

60. Baccianella S, Esuli A, Sebastiani F. Evaluation Measures for Ordinal Regression. In: 2009 Ninth International Conference on Intelligent Systems Design and Applications [Internet]. Pisa, Italy: IEEE; 2009 [cited 2023 Jul 24]. p. 283–7. Available from: http://ieeexplore.ieee.org/document/5364825/

61. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. In: 2010 20th International Conference on Pattern Recognition [Internet]. Istanbul, Turkey: IEEE; 2010 [cited 2023 Nov 3]. p. 3121–4. Available from: http://ieeexplore.ieee.org/document/5597285/

62. Rashidi HH, Albahra S, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised Machine Learning arena. Front Oncol. 2023 Feb 14;13:1130229.

63. Thölke P, Mantilla-Ramos YJ, Abdelhedi H, Maschke C, Dehgan A, Harel Y, et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage. 2023 Aug;277:120253.

64. Falcon W, dummy, dummy2. PyTorch Lightning [Internet]. 2019. Available from: https://www.pytorchlightning.ai

65. Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res. 2006;7(1–30).

66. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016 Nov;5(4):221–32.

67. Wang W, Mirjafari S, Harari G, Ben-Zeev D, Brian R, Choudhury T, et al. Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems [Internet]. Honolulu HI USA: ACM; 2020 [cited 2021 Jan 10]. p. 1–15. Available from: https://dl.acm.org/doi/10.1145/3313831.3376855

68. Elor Y, Averbuch-Elor H. To SMOTE, or not to SMOTE? [Internet]. arXiv; 2022 [cited 2023 Sep 26]. Available from: http://arxiv.org/abs/2201.08528

69. Van Den Goorbergh R, Van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022 Aug 16;29(9):1525–34.

70. Mienye ID, Sun Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Inform Med Unlocked. 2021;25:100690.

71. Sükei E, Norbury A, Perez-Rodriguez MM, Olmos PM, Artés A. Predicting Emotional States Using Behavioral Markers Derived From Passively Sensed Data: Data-Driven Machine Learning Approach. JMIR MHealth UHealth. 2021 Mar 22;9(3):e24465.

72. Saeb S, Lattie EG, Kording KP, Mohr DC. Mobile Phone Detection of Semantic Location and Its Relationship to Depression and Anxiety. JMIR MHealth UHealth. 2017 Aug 10;5(8):e112.

73. Sarda A, Munuswamy S, Sarda S, Subramanian V. Using Passive Smartphone Sensing for Improved Risk Stratification of Patients With Depression and Diabetes: Cross-Sectional Observational Study. JMIR MHealth UHealth. 2019 Jan 29;7(1):e11041.

74. Jacobson NC, Feng B. Digital phenotyping of generalized anxiety disorder: using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life. Transl Psychiatry. 2022 Aug 17;12(1):336.

75. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? [Internet]. arXiv; 2022 [cited 2023 Oct 17]. Available from: http://arxiv.org/abs/2207.08815

76. McElfresh D, Khandagale S, Valverde J, C VP, Ramakrishnan G, Goldblum M, et al. When Do Neural Nets Outperform Boosted Trees on Tabular Data? [Internet]. arXiv; 2023 [cited 2023 Oct 17]. Available from: http://arxiv.org/abs/2305.02997

77. Pratap A, Atkins DC, Renn BN, Tanana MJ, Mooney SD, Anguera JA, et al. The accuracy of passive phone sensors in predicting daily mood. Depress Anxiety. 2019 Jan;36(1):72–81.

78. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology. 2018 Jul;43(8):1660–6.

79. Cohen A, Naslund JA, Chang S, Nagendra S, Bhan A, Rozatkar A, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. Schizophrenia. 2023 Jan 27;9(1):6.

80. Schneeberger D, Stöger K, Holzinger A. The European Legal Framework for Medical AI. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E, editors. Machine Learning and Knowledge Extraction [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Jan 20]. p. 209–26. (Lecture Notes in Computer Science; vol. 12279). Available from: http://link.springer.com/10.1007/978-3-030-57321-8_12

81. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. ArXiv170507874 Cs Stat [Internet]. 2017 Nov 24 [cited 2021 Dec 8]; Available from: http://arxiv.org/abs/1705.07874

82. Lofstrom H, Lofstrom T, Johansson U, Sonstrod C. Calibrated Explanations: with Uncertainty Information and Counterfactuals [Internet]. arXiv; 2023 [cited 2023 Oct 24]. Available from: http://arxiv.org/abs/2305.02305

83. Amarasinghe K, Rodolfa KT, Jesus S, Chen V, Balayan V, Saleiro P, et al. On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods [Internet]. arXiv; 2023 [cited 2023 Apr 27]. Available from: http://arxiv.org/abs/2206.13503

Chapitre 4 – Résultats supplémentaires

Cette section contient des résultats d'analyses supplémentaires. D'abord, la portée de l'article est étendue en comparant la régression ordinale aux tâches de régression continue et de classification multiclasse. Ensuite, une nouvelle dimension est considérée, soit l'interprétabilité des modèles prédictifs.

4.1 Régression continue et classification multiclasse

L'article justifie l'importance d'utiliser des méthodes ordinales pour prédire les scores d'échelles cliniques standardisées. En moyenne, des performances équivalentes ont été obtenues pour la régression ordinale et la classification binaire, soit la tâche d'apprentissage la plus populaire en phénotypage numérique. Cependant, la tâche binaire est la plus éloignée de l'ordinale, car la binarisation transforme irréversiblement les états mentaux autoévalués, perdant l'ordre et les classes discrètes (section 1.5.2.1). Puisque l'on devait arbitrairement choisir une méthode de binarisation (comme le font les autres études), la comparaison entre les deux tâches reste imparfaite. Afin d'étendre les conclusions de l'article, les tâches de régression continue et de classification multiclasse ont été évaluées, chacune préservant respectivement l'ordonnancement ou la nature discrète des classes. Les performances ont pu être comparées directement à partir des autoévaluations ordinales telles que collectées.

4.1.1 Méthode

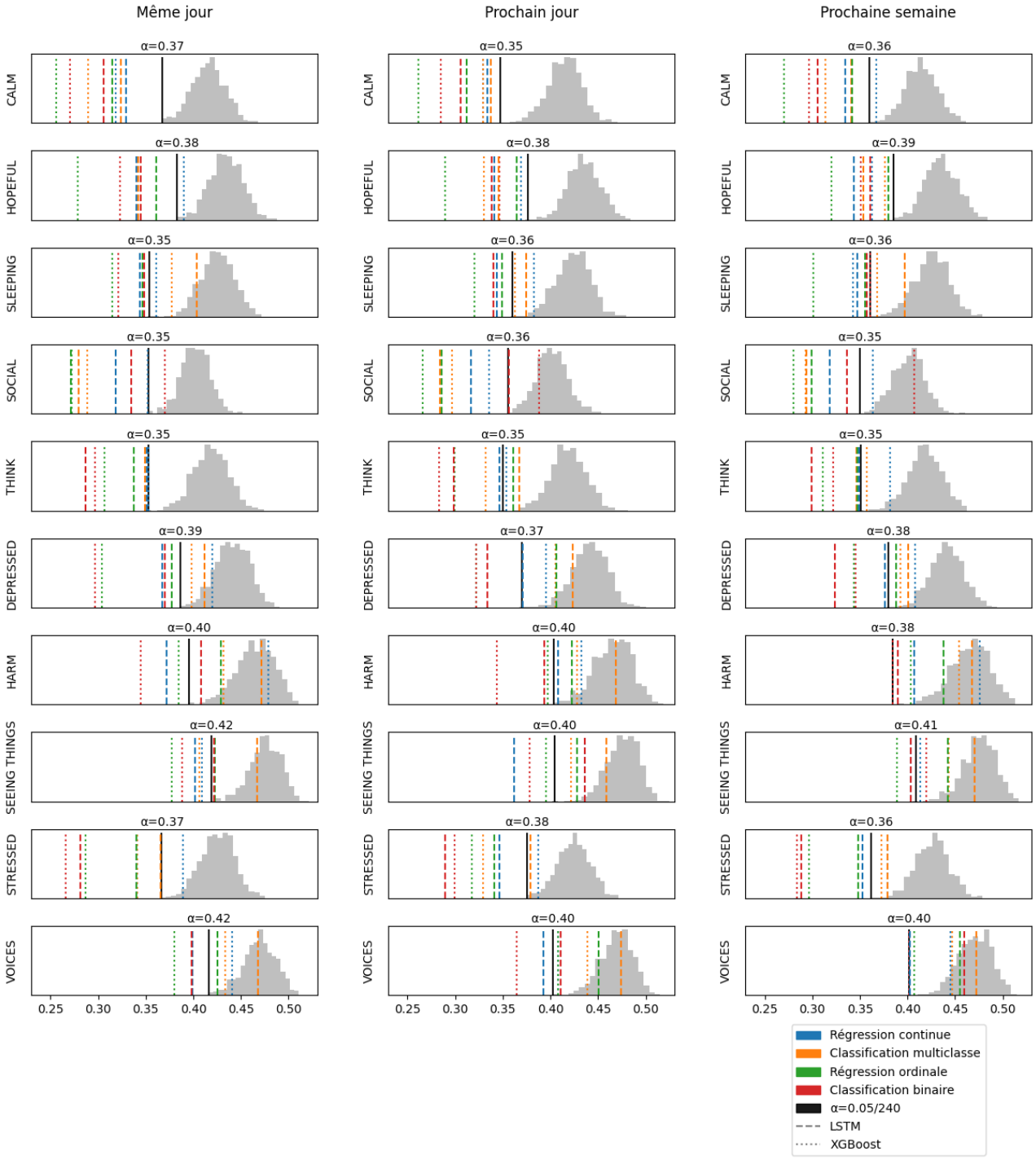
De nouveaux modèles XGBoost et LSTM ont été définis puis entrainés pour ces tâches, aboutissant à un total de 240 modèles (4 tâches $\times$ 2 algorithmes $\times$ 3 horizons temporels $\times$ 10 états mentaux). Les modèles sont comparés avec les mesures balancées (MAMAE et BACC) portées sur une échelle d'erreur commune (*scaled error*). Pour chacun des 60 jeux de données (3 horizons temporels $\times$ 10 états mentaux), une distribution Monte-Carlo a été créée pour tester la significativité des modèles (4 tâches $\times$ 2 algorithmes). Les analyses statistiques (Friedman avec Nemenyi post hoc, Wilcoxon) pour comparer la performance à travers les tâches, algorithmes, horizons temporels et états mentaux ont été réutilisées. On ajoutera une analyse de variance des

rangs (*rank-adjusted analysis of variance*) avec les quatre facteurs afin d'évaluer l'effet relatif de chacun sur la performance prédictive.

4.1.2 Performance des modèles

Au total, 154/240 modèles ont été significativement supérieur au niveau de base avec un seuil alpha de ,05 avec correction Bonferroni (α=,0002). La Figure 12 et le Tableau 4 montrent les performances sur les quatre tâches par rapport au seuil critique et à la distribution du niveau de base pour toutes les combinaisons d'état mental et d'horizon temporel. La différence de performance entre les tâches d'apprentissage est significative (Q = 45,6, p=6,89e-10) avec la classification binaire (médiane = 34,45%; 48/60 sig; *p/q sig* signifie p des q modèles entrainés ont performé significativement au-dessus du niveau de base) et la régression ordinale (médiane = 34,51%; 45/60 sig) ressortant supérieure à la régression continue (médiane = 36,74%; 35/60 sig) et la classification multiclasse (médiane = 37,53%; 26/60 sig) (Figure 13a). La différence de performance entre les algorithmes est significative (Z = 2332, p=6,76e-4) avec le XGBoost (médiane = 35,91%; 75/120 sig) étant supérieur au LSTM (médiane = 35,62%; 79/120 sig) (Figure 13b). La différence de performance entre les horizons temporels est significative (Q = 30,68, p=2,18e-7) avec une dégradation alors que l'horizon augmente du jour même (médiane = 35,02%; 59/80 sig), au jour suivant (médiane = 35,15%; 49/80 sig) et à la semaine suivante (médiane = 36,29%; 46/80 sig) (Figure 13c). La différence de performance pour les états mentaux est significative (Q = 160,32, p=6,38e-30). Le test post hoc indique cinq groupes aux performances significativement différentes. Le premier est composé de CALM (médiane=31,46%; 23/24 sig), SOCIAL (médiane = 30,80%; 19/24 sig), STRESSED (médiane=34,07%; 18/24 sig), THINK (médiane = 34,64%; 19/24 sig) et HOPEFUL (médiane = 34,53%; 23/24 sig). Le second contient SOCIAL, STRESSED, THINK, HOPEFUL et SLEEPING (médiane=34,85%; 16/24 sig). Le troisième inclue STRESSED, THINK, HOPEFUL, SLEEPING et DEPRESSED (médiane = 37,67%; 12/24 sig). Finalement, le quatrième regroupe DEPRESSED, HARM (médiane = 41,57%; 7/24 sig) et SEEING THINGS (médiane = 41,67%; 11/24 sig) et le cinquième HARM, SEEING THINGS et VOICES (médiane = 41,80%; 6/24 sig) (Figure 13d).

Figure 9 *Performance d'évaluation des 240 modèles prédictifs*



Note. Chaque rangée est associée avec un état mental et chaque colonne avec un horizon temporel. Pour chacune des 30 conditions, 8 barres verticales indiquent l'erreur en % de l'échelle (0 indique une performance parfaite) des XGBoosts et des LSTMs sur les 4 tâches d'apprentissage. Une neuvième barre indique le seuil critique de la distribution Monte Carlo pour $P<.05$ avec correction Bonferroni pour 240 modèles.

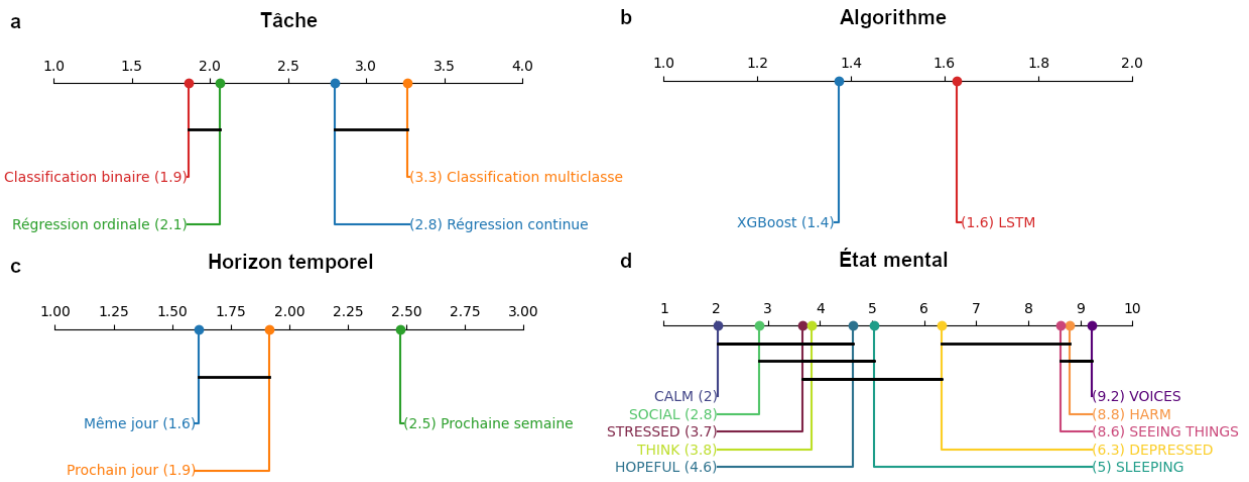
Tableau 4 Performance d'évaluation des 240 modèles prédictifs

		Tâche	Régression continue		Classification multiclasse		Régression ordinale		Classification binaire	
		Algorithme	XGBoost	LSTM	XGBoost	LSTM	XGBoost	LSTM	XGBoost	LSTM
État mental	Horizon temporel	Seuil critique (%)	Erreur sur l'échelle (%)							
CALM	Même jour	36.71	31.82	32.90	28.95	32.36	**25.57**	31.52	27.06	30.60
	Prochain jour	34.78	32.89	33.42	32.95	33.73	**26.14**	31.20	28.46	30.57
	Prochaine semaine	35.98	n.s. 36.71	33.51	31.39	34.13	**27.06**	34.20	29.70	30.58
DEPRESSED	Même jour	38.67	n.s. 41.96	36.76	n.s. 39.79	n.s. 41.18	30.36	37.70	**29.63**	36.98
	Prochain jour	36.97	n.s. 39.58	n.s. 37.10	n.s. 40.56	n.s. 42.34	32.22	n.s. 40.64	32.18	33.38
	Prochaine semaine	37.95	n.s. 40.78	37.63	n.s. 39.28	n.s. 40.08	**34.40**	n.s. 38.84	34.54	32.37
HARM	Même jour	39.53	n.s. 47.89	37.16	n.s. 43.13	n.s. 47.16	38.48	n.s. 42.91	**34.51**	n.s. 40.83
	Prochain jour	40.34	n.s. 43.26	n.s. 40.84	n.s. 42.78	n.s. 46.87	39.75	n.s. 42.30	**34.41**	39.40
	Prochaine semaine	38.44	n.s. 47.57	n.s. 40.73	n.s. 45.46	n.s. 46.79	n.s. 40.34	n.s. 43.80	**38.42**	n.s. 39.02
HOPEFUL	Même jour	38.24	n.s. 38.98	34.05	34.43	34.22	**27.90**	36.13	32.30	34.51
	Prochain jour	37.60	36.93	34.12	33.03	34.56	**28.99**	36.48	34.63	33.87
	Prochaine semaine	38.55	36.23	34.36	37.58	35.34	**31.97**	37.99	35.12	36.07
SEEING THINGS	Même jour	41.87	40.88	40.18	40.66	n.s. 46.67	**37.70**	n.s. 42.31	38.78	n.s. 42.19
	Prochain jour	40.48	n.s. 43.65	**36.21**	n.s. 42.19	n.s. 45.83	39.55	n.s. 42.83	37.80	n.s. 43.65
	Prochaine semaine	40.89	n.s. 41.36	40.37	n.s. 44.29	n.s. 47.06	**38.88**	n.s. 44.23	n.s. 41.98	40.37
SLEEPING	Même jour	35.41	n.s. 36.09	34.41	n.s. 37.78	n.s. 40.33	**31.45**	34.62	32.10	34.81
	Prochain jour	36.03	n.s. 38.25	34.38	n.s. 36.25	n.s. 37.47	**31.99**	34.89	33.99	33.98
	Prochaine semaine	36.12	34.27	34.76	n.s. 36.84	n.s. 39.71	**30.12**	35.58	36.09	35.71
SOCIAL	Même jour	35.25	35.18	31.84	28.83	27.94	27.24	**27.11**	n.s. 37.02	33.47
	Prochain jour	35.51	33.58	31.67	29.71	28.38	**26.60**	28.62	n.s. 38.84	n.s. 35.65
	Prochaine semaine	35.03	n.s. 36.35	31.84	29.31	29.37	**28.05**	29.92	n.s. 40.73	33.66
STRESSED	Même jour	36.69	n.s. 38.89	34.04	34.12	36.60	28.71	34.04	**26.58**	28.11
	Prochain jour	37.56	n.s. 38.72	34.67	32.91	n.s. 37.93	31.75	34.09	29.90	**28.95**
	Prochaine semaine	36.17	n.s. 37.90	35.30	n.s. 37.28	n.s. 37.88	29.67	34.83	**28.36**	28.83
THINK	Même jour	35.31	35.31	35.07	34.90	34.96	30.65	33.78	29.66	**28.67**
	Prochain jour	34.98	n.s. 35.40	34.65	33.24	n.s. 36.70	29.90	n.s. 36.13	**28.34**	29.87
	Prochaine semaine	35.05	n.s. 38.16	34.94	n.s. 35.74	34.62	31.11	34.74	32.21	**29.94**
VOICES	Même jour	41.67	n.s. 44.06	39.89	n.s. 43.33	n.s. 46.76	**37.99**	n.s. 42.56	39.90	39.84

	Prochain jour	40.29	n.s. 40.83	39.25	n.s. 43.86	n.s. 47.42	n.s. 40.77	n.s. 45.05	**36.43**	n.s. 41.04
	Prochaine semaine	40.13	n.s. 44.54	n.s. 40.23	n.s. 44.71	n.s. 47.22	n.s. 40.75	n.s. 45.47	n.s. **40.14**	n.s. 45.99

Note. Les résultats en gras indiquent la meilleure combinaison (algorithme, tâche) par rangée. n.s. indique les résultats non significatifs.

Figure 10 *Différences de rangs de la performance des modèles prédictifs*



Note. La position indique le rang moyen de la variable à travers les comparaisons multiples. Les barres horizontales regroupent les valeurs qui ne sont pas statistiquement différentes. **a.** rangs selon la tâche d'apprentissage **b.** rangs selon l'algorithme d'apprentissage **c.** rangs selon l'horizon temporel **d.** rangs selon l'état mental.

En considérant les quatre facteurs simultanément, sans interaction, l'analyse de variance des rangs conclut en un effet principal significatif pour la tâche d'apprentissage (F(3, 236) = 25,21, p=4,3e-14), l'état mental (F(9, 230) = 57,86, p=1,4e-53), l'algorithme (F(1, 238) = 10,04, p=,0017), et l'horizon temporel (F(2, 237) = 3,99, p=,02). Les tailles d'effets indiquent un rôle majeur de l'état mental (η^2=,621), un effet important pour la tâche d'apprentissage (η^2=,090) et des effets mineurs pour l'algorithme (η^2=,012) et l'horizon temporel (η^2=,009).

4.2 Interprétabilité des modèles

Les modèles d'apprentissage automatique sont optimisés pour fournir la meilleure performance prédictive. Cependant, il est essentiel de faire sens d'un modèle pour avoir confiance en ses prédictions (Doshi-Velez & Kim, 2017). L'interprétabilité (ou explicabilité) est « le degré selon lequel un humain peut comprendre la cause d'une décision » (Miller, 2018). En pratique, les méthodes d'interprétabilité décrivent les relations qu'un modèle a apprises à partir des données. Cela permet entre autres de s'assurer que les prédictions sont plausibles et d'éviter qu'elles soient biaisées par des attributs protégés (p. ex. : âge, sexe). Dans certaines juridictions, rendre son modèle interprétable est une obligation légale lorsque son utilisation influence une décision légale, médicale ou autres activités critiques (Schneeberger et al., 2020).

4.2.1 Méthode

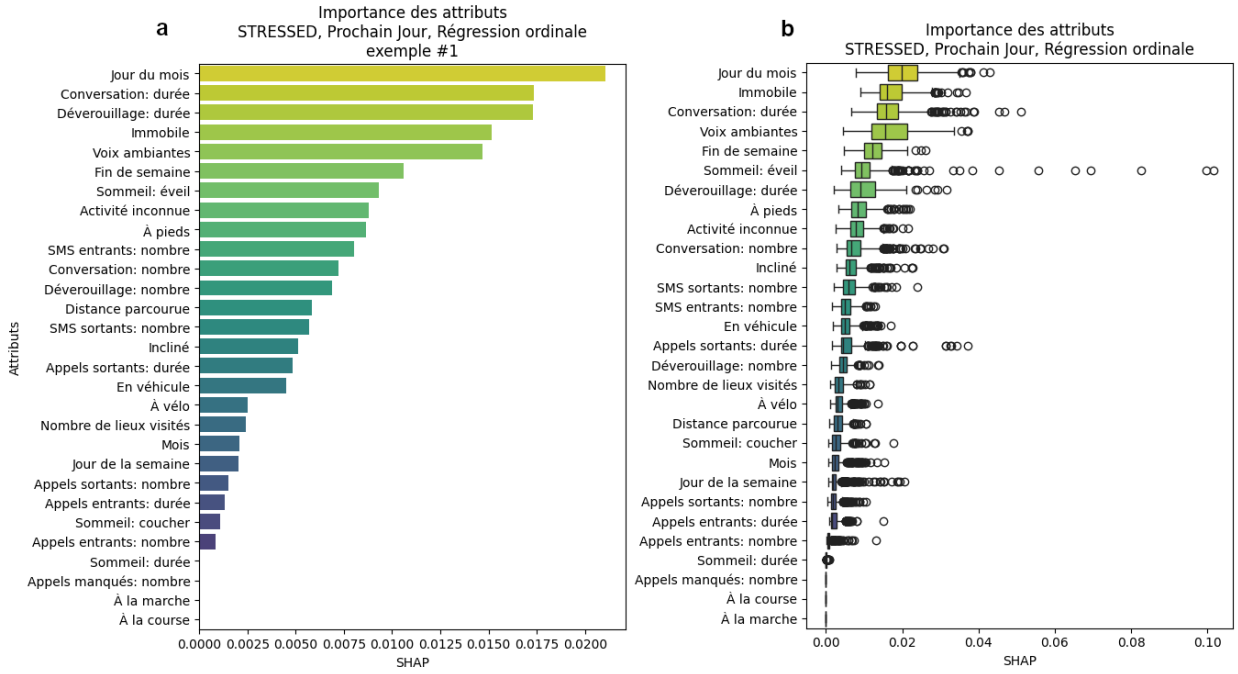
La variété de modèles entrainés sera interprétée avec la méthode SHAP (Lundberg & Lee, 2017). L'approche SHAP est l'une des méthodes d'interprétabilité les plus populaires en apprentissage automatique, incluant le phénotypage numérique (Jacobson & Feng, 2022; Sükei et al., 2021). Elle fournit une explication pour chaque prédiction du modèle en attribuant une importance (valeur de Shapley) à chaque attribut en entrée selon sa contribution marginale à la prédiction. Pour chaque prédiction, il est donc possible de décortiquer quel attribut augmente ou diminue la valeur de la prédiction par rapport à la prédiction moyenne (explication locale). En accumulant les explications, on obtient l'importance moyenne de l'ensemble des attributs ce qui permet une interprétation du fonctionnement global du modèle (explication globale). La méthode SHAP a été choisie parce qu'elle est agnostique à l'algorithme et donc compatible tant avec les XGBoost que les LSTMs.

Cependant, le format des valeurs SHAP varie selon la tâche de prédiction. Comme la classification multiclasse produit une explication binaire par classe (quatre par prédiction), une explication synthèse a été calculée à partir de la moyenne des valeurs absolues. Pour rendre les comparaisons possibles, les explications des autres tâches ont aussi été converties en valeurs absolues, puis les valeurs d'importance ont été normalisées par modèle. Les valeurs SHAP ont été calculées sur l'ensemble d'évaluation pour les 240 modèles. Chaque modèle avait 348 attributs (3 jours $\times$ 4 périodes $\times$ 29 attributs). Le résultat est une matrice à haute dimension qui peut être agrégée et tranchée pour répondre à des questions variées. En autres, des analyses ont été faites sur l'axe temporel et l'axe des attributs et en groupant les attributs par capteurs.

4.2.2 Explications

Les résultats suivants illustrent le type de question et de réponses que peuvent fournir les explications SHAP, mais ne sont pas exhaustifs. Ils peuvent servir à formuler des hypothèses, mais statistiques inférentielles sont nécessaires pour tirer des conclusions robustes. L'importance par attribut a été calculée en écrasant la dimension temporelle et faisant la moyenne à travers les 12 intervalles (3 jours $\times$ 4 périodes). La variable STRESSED est présentée comme exemple puisque c'est l'un des états mentaux les plus étudiés.

Figure 11 *Explications de l'état STRESSED du prochain jour avec un XGBoost ordinal*

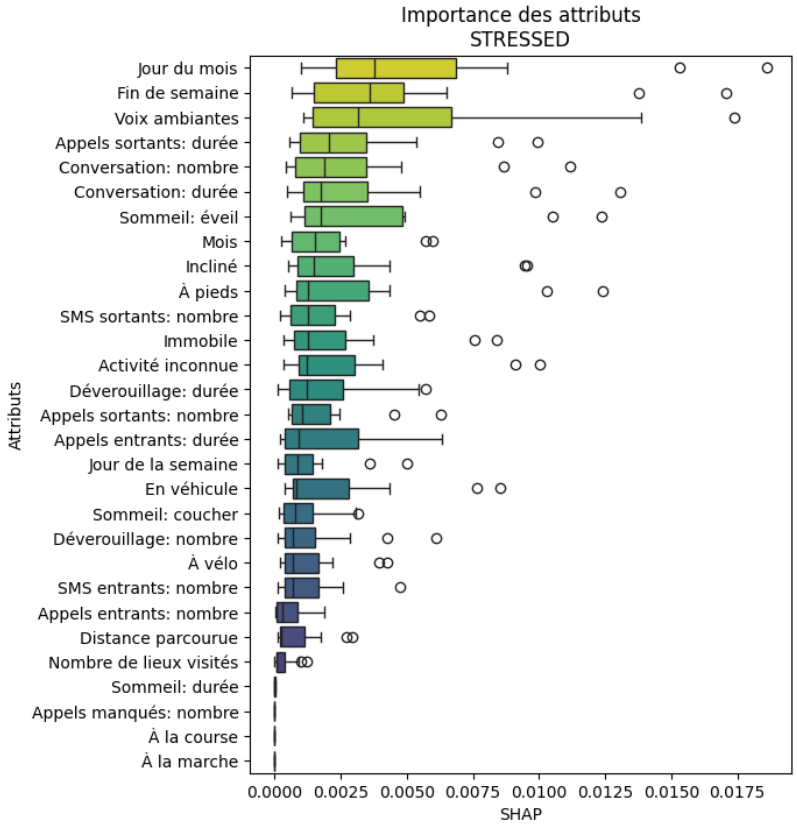


Note. **a.** Explication locale pour une seule prédiction selon les valeurs SHAP des 29 attributs. **b**. Explications globales à partir des 434 prédictions de l'ensemble d'évaluation selon les valeurs SHAP des 29 attributs. Les ronds indiquent des valeurs extrêmes pour l'attribut.

L'explication locale assigne une importance à chaque attribut (Figure 14a). On note que les attributs audios du CrossCheck ne peuvent pas différencier la voix de l'utilisateur et valider sa participation dans les conversations détectées (Wang et al., 2017). Pris ensemble, le jour du mois, les conversations, le déverrouillage, le fait d'être immobile et les voix ambiantes peuvent suggérer que la personne a une journée occupée au travail avec plusieurs interactions humaines ou, à l'inverse, qu'elle est seule dans un endroit public.

En générant les explications pour les 434 exemples de l'ensemble d'évaluation, on obtient un portrait du fonctionnement du modèle (Figure 14b). On note qu'il y a plusieurs données extrêmes relativement à leur distribution entre autres pour *Sommeil: éveil*, *Appels sortants: durée* et *Jour de la semaine*. Notamment, cela indique que le modèle considère certains moments d'éveil comme un excellent prédicteur du niveau de stress du lendemain (p. ex., après avoir travaillé tard pour une date de tombée). Aussi, les explications globales contextualisent l'explication locale. L'importance des attributs semble bien alignée avec le motif global. Par exemple, on n'observe pas de valeur surprenante pour *Jour de la semaine*.

Figure 12 *Explications globales pour l'état mental STRESSED*

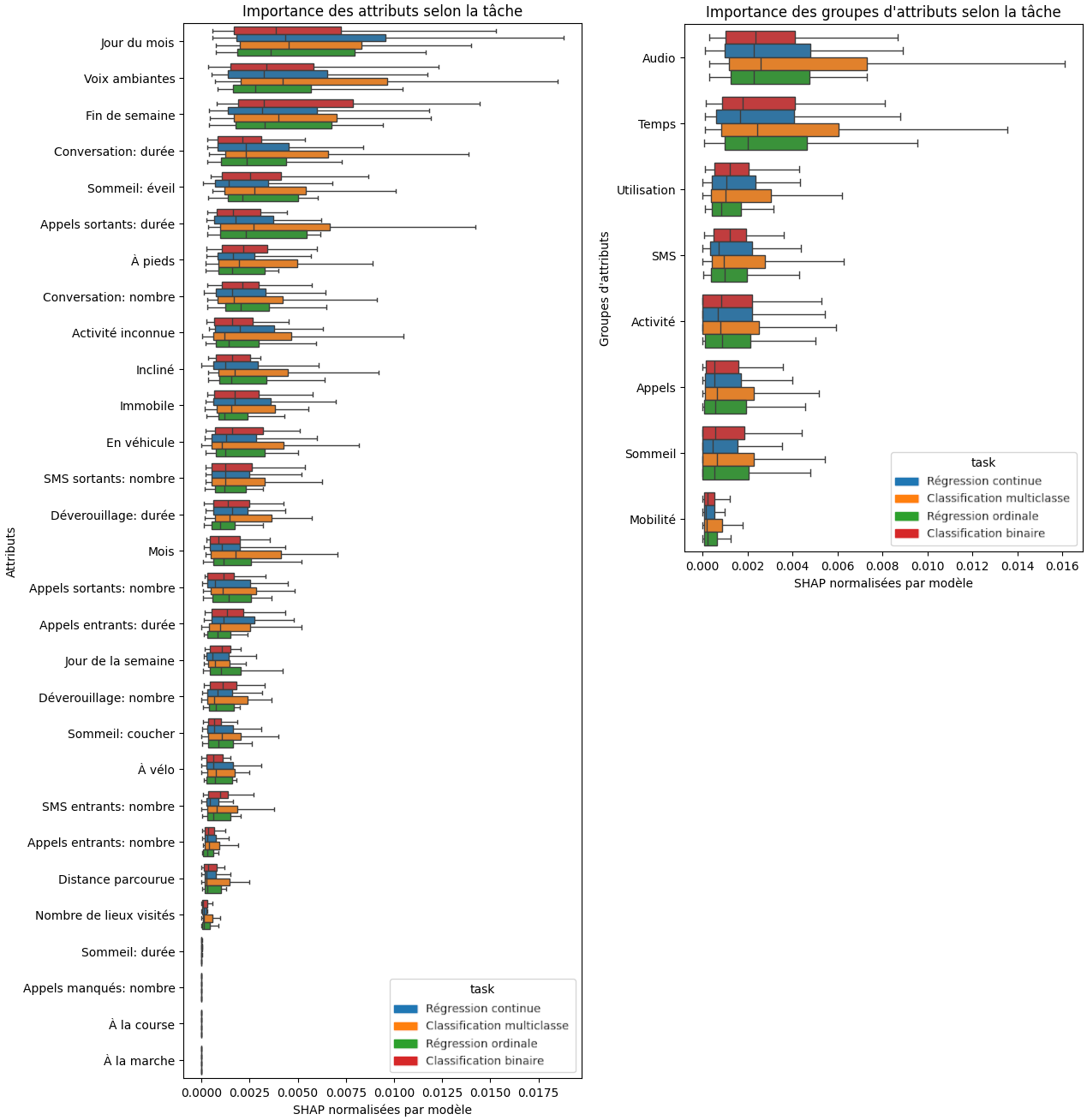


Note. Valeurs SHAP des attributs selon 24 modèles (4 tâches, 2 algorithmes, 3 horizons temporels).

Ensuite, agréger les explications à un niveau supérieur permet de tirer des conclusions plus généralisables (Figure 15). Par exemple, les variables les moins importantes pour le modèle ordinal, soit *Sommeil: durée*, *Appels manqués: nombre*, *À la course*, *À la marche* sont aussi les moins importantes pour les modèles binaire, continu et multiclasse. La grande variabilité de *Voix ambiantes* suggère que certains modèles ont su exploiter cet attribut davantage.

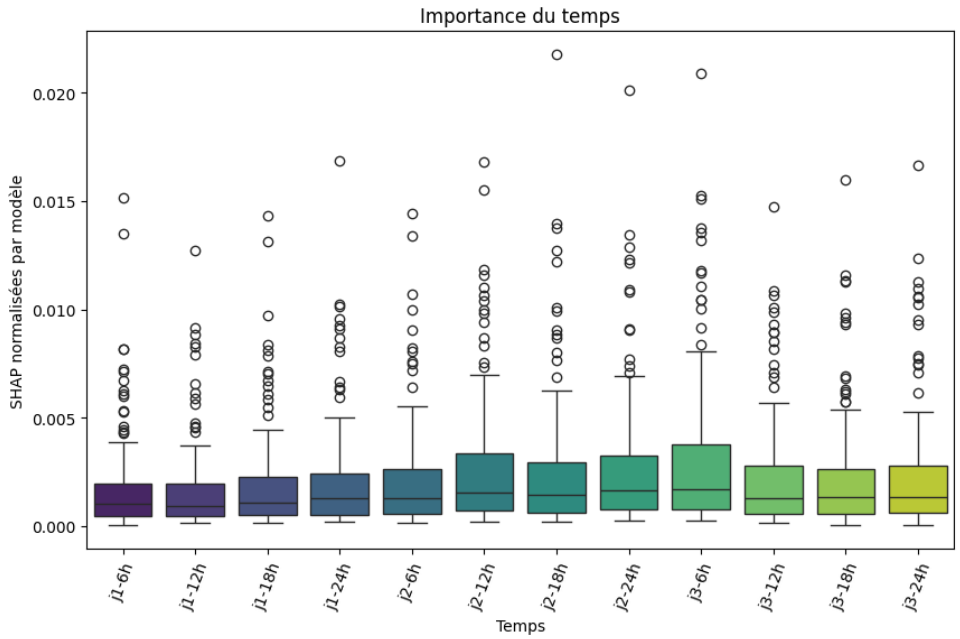
On note les distributions très asymétriques des valeurs SHAP, en particulier pour la tâche de régression multiclasse (Figure 16a). Il est surprenant de voir *Jour du mois* comme étant la variable la plus importante. Il est moins surprenant de trouver *Fin de semaine* en troisième, car l'effet saisonnier des semaines a été maintes fois rapporté (Saeb et al., 2016; Spathis et al., 2019a, 2019b; Suhara et al., 2017). En regroupant les attributs (Figure 16b), les groupes *Audio* et *Temps* ressortent comme plus importants, suivis d'*Utilisation*. Le groupe *Mobilité* est le moins important et les autres semblent équivalents.

Figure 13 *Explications globales par tâche d'apprentissage*



Note. Valeurs SHAP des attributs selon 240 modèles groupés par tâche (4 tâches, 10 états mentaux, 2 algorithmes, 3 horizons temporels). **a.** Explications par attributs. **b.** Explications par groupes d'attributs.

Figure 14 *Explications selon la position dans la séquence d'entrée*



Des explications pour la dimension temporelle peuvent être obtenues en faisant la moyenne des 29 attributs pour chaque intervalle de temps (Figure 17). On note une importante asymétrie des valeurs SHAP avec une distribution assez uniforme en nombre et en intensité des données extrêmes. En moyenne, l'importance semble augmenter pour les jours 2 et 3, qui sont plus près de la cible de prédiction. La distribution assez uniforme suggère que le modèle base ses prédictions selon les signaux forts qu'il perçoit, peu importe la position temporelle.

Chapitre 5 – Discussion

5.1 Conclusions principales

Ce projet consistait à développer des modèles d'apprentissage automatique pour prédire des états mentaux futurs dans une population psychiatrique à partir de données passives du téléphone intelligent. Le premier objectif a été d'évaluer la performance de méthodes prédictives ordinales afin de mieux rendre compte des autoévaluations de participants ayant un diagnostic de schizophrénie et de fournir des prédictions plus riches. Le deuxième objectif était d'évaluer le potentiel des algorithmes GBDT à des fins de prévision et de les comparer aux populaires RNNs. Au cours de cette étude, le débalancement des classes s'est avéré un aspect critique. Finalement, l'interprétation des modèles entrainés ajoute de la perspective aux performances observées.

5.1.1 Tâche prédictive

L'article et les analyses supplémentaires démontrent la bonne performance des modèles de régression ordinale allant de MAMAE = 1,436 (47,89 % de l'échelle) à MAMAE = 0,767 (25,57 % de l'échelle). Les tests statistiques suggèrent la supériorité de la régression ordinale par rapport à la régression continue et à la classification multiclasse. D'ailleurs, moins de modèles ont dépassé le niveau de base pour la régression continue et la classification multiclasse. Adopter la régression ordinale qui respecte à la fois l'ordonnancement et la nature discrète de la cible semble donc être bénéfique pour la performance. Cependant, il n'y avait pas de différence significative entre la régression ordinale et la classification binaire sur l'échelle commune. À performance équivalente, le modèle ordinal fournit des prédictions plus riches sur toute l'étendue de l'échelle contrairement au modèle binaire. De plus, comparer la performance en normalisant par l'étendue de l'échelle constitue un biais optimiste pour les modèles binaires puisqu'ils ont une plus petite étendue pour se tromper. Autrement dit, les modèles ordinaux accomplissent une tâche plus difficile, et ce de manière comparable aux modèles binaires en termes de performance. Finalement, l'analyse de variance multifacteurs attribue un effet significatif de la tâche prédictive sur la performance. La tâche prédictive est le deuxième facteur le plus important parmi ceux étudiés avec une taille d'effet moyenne à élevée (η^2 = ,09).

À notre connaissance, ce projet est parmi les seuls à employer l'apprentissage automatique pour une tâche de régression ordinale consistant à prédire autoévaluations ordinales à partir de signaux du téléphone intelligent. Une autre étude sur le jeu de données CrossCheck a utilisé des modèles de régression continue pour une étude de détection des états mentaux (Tseng et al., 2020). Comparer les résultats aux MAMAE obtenus est difficile, car ils sont rapportés dans une figure, sans valeur numérique et la mesure non balancée *root mean squared error* (RMSE). Nos modèles débalancées (HARM, SEEING THINGS, VOICES) obtiennent des MAE comparables aux RMSE représentées visuellement, autour de 0,3, mais nos autres modèles sont inférieurs (MAE = 0,75 vs RMSE = 0,4). On devine que les modèles de cette étude optimisent la mesure de performance en prédisant naïvement la moyenne.

La classification multiclasse a eu le moins de succès (MAMAE médiane = 1.13 [37.53%]; 26/60 sig). Une étude a évalué des modèles XGBoost à prédire les scores d'anxiété généralisée basés sur le GAD-7 avec la régression continue, la classification binaire et la classification multiclasse (Choudhary et al., 2022). L'échelle contient 7 items avec des valeurs de 0 à 3. La somme des réponses crée une cible de régression continue de 0 à 21 et les classes [Aucune, Faible, Modérée, Sévère] et [Non sévère, Sévère] sont établies selon les seuils d'interprétation cliniques. Puisqu'aucune mesure de performance balancée n'a été utilisée (p. ex. : MAMAE, MAMSE), il n'y a pas eu de comparaison directe entre tâches d'apprentissage. À partir des scores d'exactitude et de la proportion des classes ordinales, la BACC par classe pour le modèle de classification multiclasse peut être calculée (Aucune = 51,5%, Faible = 61,5%, Modérée = 39,5%, Sévère = 46.0%). Donc, le XGBoost multiclasse a performé sous le niveau de la chance, soit un niveau de base beaucoup plus facile à surpasser que la distribution Monte Carlo employée dans ce projet, pour certaines classes et en agrégat.

En plus de conserver les données originales, faire la promotion des méthodes ordinales prévient la binarisation arbitraire (c.-à-d., non motivée par des seuils cliniques). Notamment, certaines études déterminent les classes « faibles » et « élevées » selon la distribution des autoévaluations collectées (Canzian & Musolesi, 2015; Umematsu, Sano, Taylor, et al., 2019). C'est particulièrement problématique, car la définition des classes et des prédictions du modèle dépend alors des données. Autrement dit, un exemple considéré « dépression faible » pourrait

être reconsidéré « dépression élevée » si l'on collectait des données additionnelles. Dans d'autres cas, les directives d'interprétation de l'échelle sont redéfinies arbitrairement « (1) none to minimal depression (range 0-4); (2) mild depression (range 5-9) […]. The score can also be interpreted as no depression (range 0-9) » (Wang et al., 2018). L'utilisation de la binarisation perd le sens des phénomènes cliniques mesurés et la variété d'approches pour définir les classes crée de la confusion dans la littérature.

Les études ayant des cibles ordinales devraient a priori utiliser la régression ordinale et justifier leur choix dans le cas contraire. D'une part, la classification binaire demeure pertinente lorsqu'elle distingue des catégories cliniques bien définies. D'autre part, la régression continue peut-être appropriée pour une échelle ordinale avec une grande étendue et une résolution fine. Un avantage de la régression ordinale est que les prédictions sont contraintes à une échelle prédéterminée. Un effet similaire peut être obtenu en post-traitant les prédictions de régression continue, mais cette option ne semble pas avoir été explorée. De plus, la classification multiclasse devrait être évitée pour prédire des cibles ordinales compte tenu des problèmes d'inconsistance des rangs et de débalancement des classes. Autrement, une mesure de performance comme le MAMAE indique l'erreur en nombre de points sur l'échelle, ce qui demeure simple à interpréter.

5.1.2 Algorithme de prévision

L'algorithme XGBoost a des performances significatives avec une erreur représentant entre 47,88 % et 25,57 % de l'échelle, contre 47,42 % à 27,11 % d'erreur pour le LSTM. Les tests statistiques ont conclu que le XGBoost était statiquement supérieur au LSTM. Aussi, l'algorithme d'apprentissage était le troisième facteur le plus important pour la performance avec une taille d'effet mineure (η^2=,012). Une étude conclut que la majorité des différences de performance entre réseaux de neurones et GBDTs, et entre les différents algorithmes de GBDTs sont négligeables (Grinsztajn et al., 2022; McElfresh et al., 2023). Autrement, les prédictions du même et du prochain jour étaient statistiquement supérieures à celles pour la semaine prochaine. L'horizon temporel avait un effet mineur (η^2=,009), comparable à celui de l'algorithme. De faibles dégradations similaires ont été rapportées dans d'autres études (Spathis et al., 2019b). On note que les tests de significativité sont basés sur les rangs et ne concluent pas quant à la taille des

écarts de performance. Similairement, les effets rapportés pour la tâche d'apprentissage, l'algorithme et l'horizon temporel sont relatifs au poids prépondérant de l'état mental.

L'ampleur et la rigueur de la méthode de fouille employée déterminent en grande partie la valeur et la généralisabilité d'une étude d'apprentissage automatique. Malheureusement, la majorité des études de phénotypage numérique inclut un seul modèle prédictif et parfois des modèles qui sont *a priori* dépassés ou non compétitifs (Sarda et al., 2019; Umematsu, Sano, & Picard, 2019). Une évaluation systématique d'un grand nombre de modèles performants ajoute du poids aux conclusions. Autrement, les comparaisons d'approches prédictives à travers plusieurs études sont toujours imparfaites à cause du grand nombre de décisions méthodologiques à reproduire.

Par ailleurs, les généralisations quant à la supériorité d'un modèle doivent être particulièrement conservatrices pour les séries et la tâche de prévision. Une méthode performe bien relativement aux données (entrées et cibles) et le phénotypage numérique gère une multitude de sources de variances (comportements, états mentaux, capteurs, application de collecte, etc.). Montero-Manso et al. (2020) ont créé un ensemble de neuf algorithmes de prévision et ont développé un méta-algorithme pour assigner un poids à chacun selon les attributs de la série à prédire. Ce méta-algorithme a regroupé les 97 000 séries du jeu de données M4 (Makridakis et al., 2020) en cinq prototypes pour lesquels différentes combinaisons de modèles performaient le mieux. Ces résultats démontrent qu'aucun prédicteur n'est le meilleur dans l'absolu et qu'une évaluation systématique des options disponibles est toujours de mise.

Les forces du LSTM et du XGBoost peuvent être combinées en les incluant dans un ensemble de modèles, mais aussi en les assemblant en un seul. En effet, le LSTM peut encoder de longues séquences de données à haute résolution (m attributs, n temps) en un vecteur aux dimensions fixes (k attributs) qui servira d'entrée à un XGBoost. Essentiellement, le LSTM apprend à représenter les variations de signaux pertinentes à la tâche prédictive (c.-à-d., crée un *embedding*). Ainsi, le XGBoost bénéficie d'une représentation riche des dépendances temporelles tout en évitant une inflation du nombre d'attributs (section 2.6.2). Une telle approche a eu beaucoup de succès pour prédire les phases du sommeil à partir de l'accéléromètre d'une montre intelligente (K_MAT, 2023). Les modèles prédictifs de phénotypage numérique (détection et

prévision) utilisent principalement des attributs dérivés de données passives qui varient dans le temps. Cependant, ils pourraient bénéficier de l'inclusion des attributs constants ou durables, similaires à un trait, afin de caractériser les participants. Par exemple, les variables *groupe d'âge* et *détient un emploi* ont le potentiel d'enrichir les signaux passifs pour prédire des états mentaux. Pour une tâche de prévision, il suffit de répéter la valeur à tous les intervalles de temps.

5.1.3 Interprétabilité

Les analyses supplémentaires incluent les explications SHAP des modèles prédictifs. D'une part, un clinicien pourrait inspecter l'explication d'une prédiction pour évaluer sa crédibilité. D'autre part, les explications globales permettent aux experts et aux développeurs d'effectuer un contrôle qualité et de fournir des pistes pour résoudre des problèmes de données.

Notamment, l'importance du cycle hebdomadaire et de la fin de semaine a été rapportée maintes fois, mais l'importance prépondérante de la variable *Jour du mois* a été un résultat surprenant. L'explication la plus probable est la présence d'une corrélation hasardeuse. Les ensembles de validation et d'évaluation ont été créés avec sept exemples consécutifs par participant, couvrant une période d'environ deux semaines. Cependant, la corrélation entre le jour du mois et le jour de la semaine est parfaite pour une période inférieure à un mois. Alors, l'optimisation d'hyperparamètres basée sur l'ensemble de validation a renforcé l'utilisation du jour de la semaine comme prédicteur et cela a aussi été récompensé par l'ensemble d'évaluation. Toutefois, cette relation est fortuite, car le jour du mois (1 à 31) ne corrèle pas avec le jour de la semaine (1 à 7) pour une année entière et la performance prédictive aurait été dégradée sur de nouvelles données. Sans SHAP, ce défaut serait resté invisible. Par ailleurs, le jour de la semaine, un attribut plus robuste a obtenu une importance faible, car SHAP peine à estimer l'importance d'attributs corrélés (Lundberg & Lee, 2017).

Un autre aspect surprenant a été le poids faible de la catégorie d'attributs *Mobilité* puisqu'elle a été maintes fois rapportées comme prédictive de l'humeur et du stress (Faurholt-Jepsen et al., 2022; Rohani et al., 2018). Contrairement aux autres études, nos modèles prédictifs incluaient seulement deux attributs simples de mobilité. Il est fort probable que les attributs fournis n'exploitent pas le plein potentiel des signaux bruts.

Seulement 5 des 17 articles de directives pour les résultats d'apprentissage automatique en santé mentionnent de rapporter l'importance des variables d'un modèle (Klement & El Emam, 2023). Les relations entre signaux du téléphone et état mental sont typiquement explorées à travers des études d'association, mais les analyses corrélationnelles employées ont de nombreuses faiblesses (section 1.4.1.1). Les méthodes d'interprétabilité sont des outils mieux adaptés pour comprendre le rôle de variables en entrées. En particulier, les explications globales peuvent guider la sélection d'attributs à inclure et à exclure afin de retirer l'information superflue. Ce processus diminue la complexité du modèle et simplifie son interprétation en plus de diminuer son coût computationnel. À plus long terme, on peut cesser de collecter les données à faible valeur prédictive afin de mieux préserver la vie privée des participants.

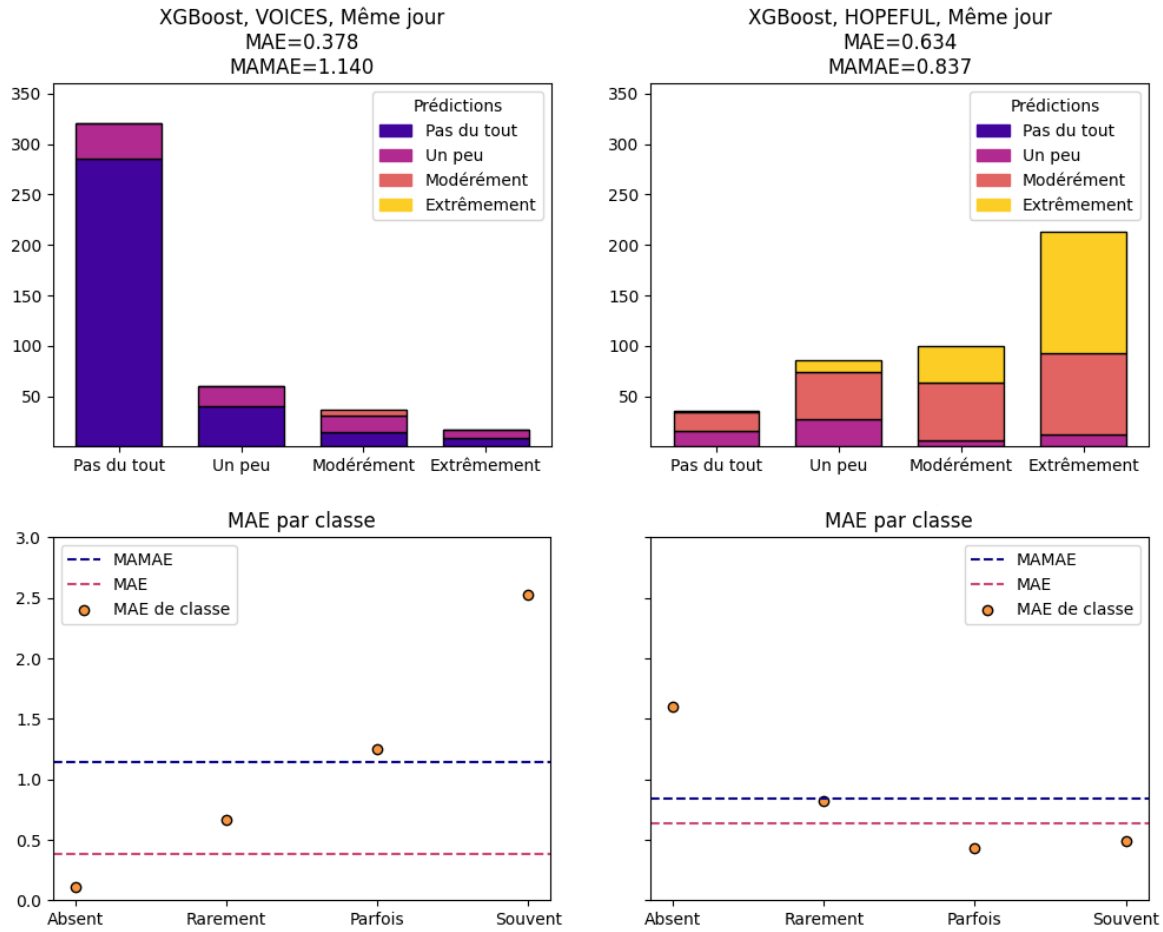
L'interprétabilité des modèles est une composante importante de l'apprentissage automatique en santé. Néanmoins, les méthodes actuelles constituent essentiellement un deuxième modèle prédictif à gérer, introduisant bon nombre de décisions méthodologiques sans bonnes pratiques établies. D'abord, il faut choisir la définition « d'importance » et l'approche pour la calculer. Ensuite, un ensemble représentatif d'exemples doit être sélectionné pour calculer les explications globales. Il faut prendre en compte l'effet des attributs corrélés sur la qualité des explications. Puis, comme le calcul d'importance peut être stochastique, plusieurs itérations sont alors nécessaires pour estimer la stabilité et la marge d'erreur des explications. Les *modèles interprétables* constituent une solution potentielle, car ils annulent le besoin de méthodes d'explications ad hoc, mais ils limitent les méthodes prédictives possibles. L'adoption de modèles interprétables en santé pourrait améliorer la fiabilité des explications. Revenant au pragmatisme, peu importe la méthode, il faut s'assurer que les explications sont utiles et contribuent à l'amélioration des soins.

5.1.4 Débalancement

Durant la phase exploratoire du projet, l'inspection des distributions des états mentaux a permis d'identifier un débalancement des classes important. Toutefois, la MAMAE n'a pas été choisie pour contrôler le débalancement, mais bien parce qu'elle était désignée comme idéale pour évaluer la régression ordinale (Baccianella et al., 2009). Les problèmes liés au débalancement sont

devenus apparents après avoir comparé les MAMAEs obtenus aux RMSE rapportées (une mesure non balancée) dans une autre étude sur le CrossCheck avec les mêmes cibles de prédiction (Tseng et al., 2020). Les deux travaux arrivaient à des conclusions inverses quant à quel modèle performe le mieux. Alors, les MAEs ont été calculés pour nos modèles et les distributions de prédictions ont été inspectées (Figure 18). Le modèle VOICES montre une très faible diversité de prédiction contrairement à HOPEFUL. La MAE favorise VOICES à HOPEFUL et la MAMAE l'inverse. Les MAEs par classe révèlent une grande disparité de la performance selon la classe prédite pour VOICES, se traduisant par un écart important entre MAE et MAMAE. La meilleure calibration de HOPEFUL réduit cet écart.

Figure 15 *Performance de modèles ordinaux selon leur débalancement*



Note. Le haut présente la MAE, la MAMAE et la distribution des prédictions. Le bas détaille la MAE par classe.

Dans notre article, les corrélations de Spearman opposées du débalancement avec la MAMAE et avec la MAE formalisent ces conclusions. Par conséquent, la relation entre MAMAE et MAE d'un même modèle est négative lorsqu'il y a débalancement. Cette conclusion peut être déduite des formules mathématique ou montrée par simulation (Thölke et al., 2023); nos résultats le montrent de manière empirique. D'ailleurs, l'analyse de facteurs attribue à l'état mental un effet majeur sur la performance (η^2=,6). Le débalancement des classes étant une propriété d'un état mental, on conclut que le débalancement a un effet sur la performance supérieure à la tâche prédictive (η^2=,09) ou à l'algorithme d'apprentissage.

En inspectant la Figure 12 et le Tableau 4 (section 5.1.2), on remarque une interaction entre tâche d'apprentissage et débalancement. La classification multiclasse performe très bien pour la cible SOCIAL qui suit une distribution plus uniforme et la régression continue ne performe pas si mal pour les cibles plus débalancées (VOICE, HARM). Ces tâches présentent une double dissociation quant aux classes discrètes et ordonnées, et entretiennent chacune une simple dissociation avec la régression ordinale. Cela suggère que représenter l'ordonnancement est important pour prédire la tendance centrale et les classes discrètes aident à couvrir l'étendue de l'échelle. Lorsque le débalancement est élevé, exploiter la tendance centrale en respectant l'ordre des classes contribue davantage à de bonnes prédictions.

Parmi 17 articles de directives pour l'apprentissage automatique en santé jugés de haute qualité, seulement 76% indiquaient de justifier le choix de mesure de performance et 36% mentionnait de décrire le débalancement des données (Klement & El Emam, 2023). À notre connaissance, aucune étude n'utilise les mesures de performance balancée BACC ou MAMAE pour évaluer leur modèle de phénotypage numérique. Pour la classification débalancée, les mesures de rappel/sensibilité, de précision/valeur prédictive positive et de score F1 sont parfois rapportées pour une appréciation plus nuancée de la performance (Lee et al., 2023). Toutefois, ces mesures sont asymétriques; les résultats dépendent de la « classe positive » et seraient différents si elles étaient inversées (Thölke et al., 2023). Elles sont utiles lorsque l'on peut juger de l'importance de ne pas se tromper sur une classe ou l'autre (p. ex. : dépistage vs preuve d'expert), mais l'objectif de prédire un état mental n'est pas assez spécifique pour conclure à cet effet. Certaines études rapportent aussi l'aire sous la courbe (*area under the curve*; AUC) de l'efficacité du récepteur

(*receiver operating characteristic*; ROC) lorsque les données sont débalancées. Un défaut de la ROC AUC est qu'un changement de quelques prédictions peut faire varier de beaucoup le score obtenu lorsque les données sont très débalancées (Van Calster et al., 2019; Varoquaux & Colliot, 2023). Cette mesure est définie pour la classification binaire, mais demeure difficilement adaptable à d'autres tâches d'apprentissage.

Comme discuté à la section 1.4.1.1, le débalancement des données pose problème pour les mesures corrélationnelles et les études d'association. Pour les mêmes raisons (hétéroscédasticité, non monotone), utiliser des coefficients linéaires (Pearson r, Spearman ρ, Kendall τ) entre les prédictions et les cibles est une erreur méthodologique importante (Bati & Singh, 2018; Depp et al., 2019; Jacobson & Feng, 2022; Wang et al., 2016, 2017). En effet, les mesures de performance comme la MAE évaluent l'erreur pour des paires prédiction-cible puis font une moyenne. Un coefficient de corrélation évalue la relation entre deux variables sans apparier observations et prédictions, ce qui pose un problème (Poldrack et al., 2020). Pour une tâche de régression continue avec les cibles [0, 1, 2, 3], les prédictions [1, 2, 3, 4] ou [0, 4, 8, 12] obtiendraient des corrélations parfaites (=1) pour la majorité des coefficients employés.

En somme, les mesures balancées permettent d'évaluer des prédictions (discrètes ou continues) pour des cibles discrètes, tout en corrigeant le débalancement des classes. Pour la classification binaire, c'est une bonne pratique de rapporter le nombre de vrais positifs, faux positifs, vrais négatifs, et faux négatifs (Varoquaux & Colliot, 2023). Les lecteurs peuvent alors calculer d'autres mesures qu'ils jugent pertinentes. Similairement, la classification multiclasse et la régression ordinale peuvent fournir des matrices de confusion (ou équivalent), mais il n'est pas simple de résumer l'information pour la régression continue. Une bonne pratique scientifique serait de fournir un fichier à deux colonnes contenant les prédictions et les cibles pour l'ensemble d'évaluation. Cette information permettrait une meilleure appréciation du débalancement et de la qualité des prédictions qui peuvent autrement être résumés et dissimulés derrière la mesure de performance. Les mesures balancées devraient être rapportées par défaut lorsque possible, car elles demeurent simples à interpréter et facilitent les comparaisons entre études. Lorsqu'une étude rapporte seulement des métriques non balancées, sans inclure de matrice de confusion, il est impossible de corriger pour le débalancement spécifique au jeu de données utilisé.

Comme discuté dans l'article, le débalancement des classes est une propriété des données de santé et l'altérer dégrade la performance du modèle en contexte appliqué. En particulier, les données débalancées originales sont nécessaires pour obtenir un modèle *calibré*, soit la capacité à prédire correctement la probabilité qu'un événement ait lieu. La calibration est distincte de la discrimination, soit générer une bonne prédiction, et devrait être évaluée plus systématiquement en santé (Park & Han, 2018; Van Den Goorbergh et al., 2022). Cependant, les mesures de performance les plus communes en apprentissage automatique (ACC, MAE, MAMAE, F1, RMSE, etc.) sont calculées à partir des valeurs prédites et non des probabilités assignées. Elles sont qualifiées de *non-proper scoring rules*, car elles peuvent être optimisées de manière « tactique » sans améliorer la qualité réelle des prédictions. C'est ce qui amène un modèle à prédire principalement la classe majoritaire afin d'optimiser la ACC ou la MAE. Les mesures balancées BACC et MAMAE ne viennent que mitiger ce problème.

Les *proper scoring rules* sont calculées à partir des probabilités et favorisent l'entrainement de modèles calibrés, mitigeant le problème de débalancement des classes (Van Calster et al., 2019). D'ailleurs, l'entropie croisée utilisée pour entrainer les modèles de classification binaire ou multiclasse est « proper », mais peu s'en servent pour calculer la performance prédictive sur l'ensemble d'évaluation. Pour la classification binaire et multiclasse, on suggère d'utiliser le score de Brier (Brier, 1950), pour la régression ordinale le *ranked probability score* (Epstein, 1969) et pour la régression continue le *continuous ranked probability score* (Hersbach, 2000). Les variantes *Brier skill score*, *ranked probability skill score* et *continuous ranked probability skill score* permettent de comparer deux méthodes probabilistes directement afin de quantifier les bénéfices d'utiliser une approche plutôt qu'une autre (Gneiting & Raftery, 2007).

Sous un autre angle, Varoquaux et Colliot (2023) remettent en question l'utilisation de plusieurs mesures de performance binaires comme la précision/sensibilité, le rappel / spécificité, le score F1, la BACC, etc., car ils reposent tous sur la probabilité de prédire la bonne classe pour un exemple ($P(\text{Prédiction} \mid \text{Cible})$). Selon eux, une mesure plus pertinente est la cote post-test (*post-test odds*), soit la probabilité d'être en dépression pour une prédiction de dépression ($P(\text{Cible} \mid \text{Prédiction})$). Le rapport de cotes (*odds ratio*) quantifie l'augmentation des chances que le phénomène prédit soit vrai si le modèle en fait la prédiction. Puisque cette mesure est

indépendante du débalancement des classes et du phénomène, le rapport de cotes obtenu sur un jeu de données peut être appliqué à des populations avec différentes distributions de classes pour obtenir leur cote post-test. Dans une perspective pragmatique, la cote post-test est une estimation de l'utilité d'un prédicteur dans une population.

5.1.5 Avenues futures

Les thèmes abordés relatifs aux tâches d'apprentissage, aux algorithmes de prévision, à l'interprétabilité de modèles et au débalancement de données soulèvent plusieurs questions de recherche. Dans un futur immédiat, le plan d'analyse et le code développé pourraient être réutilisés pour les quelques jeux de données de phénotypage numérique en accès libre (*Reality Mining, Social Evolution, Friends and Family, StudentLife, ExtraSensory*; respectivement Eagle & Pentland, 2006; Madan et al., 2012; Aharony et al., 2011; Wang et al., 2014; Vaizman et al., 2017). De manière générale, cela permettrait d'étendre nos conclusions à d'autres populations (en santé, différents lieux géographiques, différents processus de recrutement, etc.) et à d'autres cibles prédictives liées à la santé mentale recueillies dans ces études. Il serait possible d'évaluer si le niveau de débalancement des classes pour un même construit varie d'un jeu de données à l'autre et si la régression ordinale demeure robuste. Notre plan d'analyse rendrait possible la comparaison entre notre régression ordinale et les publications existantes, peu importe la tâche réalisée, et soulignerait l'importance d'utiliser des mesures de performances balancées. Au niveau plus technique, il serait pertinent d'évaluer un plus grand nombre d'algorithmes d'apprentissage, en particulier ceux plus simples, compte tenu de l'absence d'avantage marqué des RNNs. En particulier, il existe une série d'algorithmes intrinsèquement explicables (XGB1, XGB2, FIGS; respectivement Navas-Palencia, 2020; Lengerich et al., 2020; Tan et al., 2022) qui pourrait avoir une plus grande utilité clinique et éviter le besoin de méthode ad hoc. Finalement, les processus d'entrainement, d'évaluation, et d'analyses post hoc des modèles devraient être probabilistes et favoriser une bonne calibration notamment par l'utilisation de méthodes conformelles. Ainsi, on peut déterminer les circonstances où un modèle prédictif devrait bien performer ou non avant son déploiement et communiquer la certitude des prédictions à l'utilisateur une fois en milieu clinique.

5.2 Perspectives

5.2.1 Cibles de prédiction

Les performances prédictives obtenues semblent alignées avec les études de phénotypage numérique, mais paraissent insuffisantes pour une utilisation clinique. En effet, nos modèles prédictifs les plus performants obtenaient tout de même une erreur d'environ 25 % de l'échelle sur des mesures balancées. Pour le reste de la littérature, la présence de binarisation arbitraire, de mesures non balancées et la faible reproductibilité des études suggèrent un biais optimiste quant à la performance prédictive. Nos résultats montrent que la distribution des cibles est le facteur le plus important pour la bonne performance d'un modèle.

Grâce à une variété d'approches, les travaux de phénotypage numérique tentent progressivement d'améliorer la qualité des prédictions de symptômes et d'états mentaux. Toutefois, peu s'interrogent sur l'existence potentielle d'une limite à la performance atteignable sur cette tâche. Dans le cadre de leur fouille pour le meilleur prédicteur $\hat{h}$ pour les échantillons (X, Y) (section 1.3.3), les études se sont concentrées sur les nombreux signaux du téléphone et leurs transformations (X) ainsi que les algorithmes et tâches d'apprentissages ($\hat{h}$). La nature des données de santé mentale servant de cibles (Y) est plus rarement remise en question.

La majorité des études recueillent des autoévaluations quotidiennes de symptômes ou d'états mentaux grâce à des échelles standardisées. La théorie de la mesure conçoit que toute mesure est une observation avec une part d'erreur irrémédiable et non la vraie valeur du phénomène examiné. Pour plusieurs échelles utilisées en phénotypage numérique, la fiabilité et la validité ont été vérifiées pour détecter des différences interindividuelles et non pour discerner des variations chez un même individu. Par conséquent, la taille de ces erreurs et la capacité des échelles à les capturer sont inconnues; c'est le cas pour la majorité des mesures médicales réappropriées (p. ex. : données administratives, dossier patient) pour d'autres utilisations en apprentissage automatique (Brakenhoff et al., 2018). La performance maximale théorique d'un algorithme pour prédire les états mentaux (cibles) est déterminée par la fiabilité de la mesure de la cible. Les gains additionnels possibles sont le résultat de relations hasardeuses parmi l'erreur aléatoire (*overfit*) qui ne seront pas généralisables (Gell et al., 2023).

Générer un échantillon aléatoire à partir de la population sert à mitiger les sources de variations, mais ne peut réduire la part d'erreur de l'instrument. D'ailleurs, l'amélioration de performance associée à des décisions méthodologiques est souvent inférieure à l'erreur aléatoire de l'ensemble d'évaluation (Varoquaux & Cheplygina, 2022). Plusieurs études de phénotypage numérique mentionnent en discussion qu'en collectant plus de données de meilleures performances seront possibles. Néanmoins, il y a des démonstrations empiriques qu'augmenter la taille d'échantillon diminue la marge d'erreur, mais ne permet pas de dépasser le plafond associé à l'instrument de mesure (Gell et al., 2023; Varoquaux & Cheplygina, 2022). En fait, la littérature rapporte des diminutions de la performance et des tailles d'effet dans de plus grands échantillons, car la marge pour surestimer la performance est diminuée (Varoquaux, 2017).

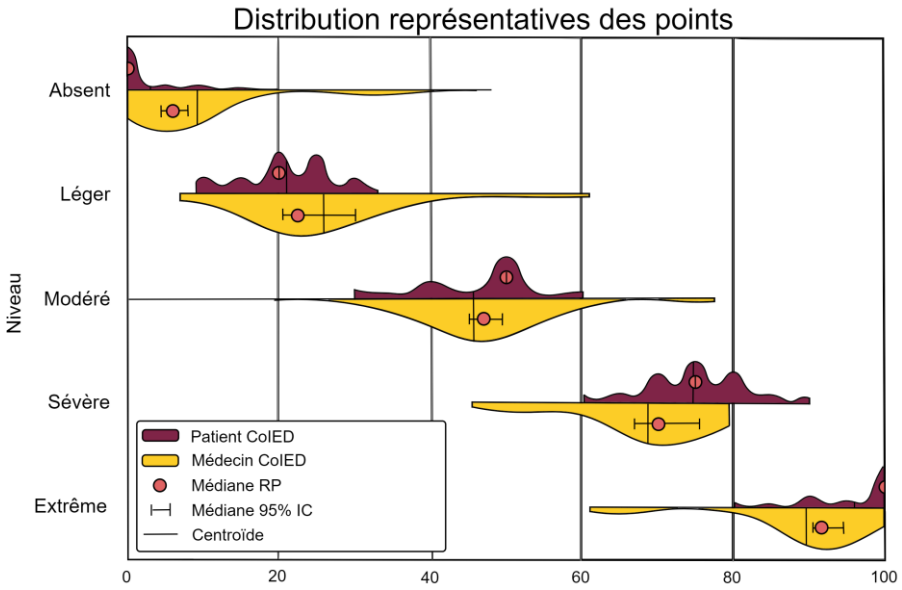
Comme les cibles binaires contiennent moins d'information que des cibles continues ou ordinales, elles sont associées a une plus grande erreur de mesure (Varoquaux & Colliot, 2023). Il serait donc désirable d'utiliser les scores bruts d'échelle clinique, mais le débalancement de classes entraine ses propres défis. Des articles récents soutiennent que les défis associés au débalancement ne proviennent pas des proportions des classes, mais bien du faible nombre d'exemples pour certaines classes (Elor & Averbuch-Elor, 2022). Le ré-échantillonnage pour rééquilibrer les classes étant déconseillé (section 3.5), opter pour des instruments de mesure générant des distributions de symptômes moins débalancées pourrait améliorer les performances prédictives. Même un instrument fiable peut avoir une résolution inadéquate pour mesurer le phénomène d'intérêt (Tiego & Fornito, 2022). Les ancrages d'une échelle de symptôme, par exemple de stress, devraient couvrir un gradient d'expériences afin d'éviter que les mesures répétées soient principalement dans les extrêmes.

Un autre facteur influençant la distribution des cibles est le moment de la mesure. Bien que les études de phénotypage numérique rapportent employer des évaluations écologiques momentanées, ce terme est utilisé très librement. La plupart du temps, ce sont plutôt des questionnaires quotidiens, comme c'est le cas pour CrossCheck (Wang et al., 2016). Dans d'autres cas, la mesure est rétroactive pour la veille (Aharony et al., 2011) ou la semaine précédente (Wang et al., 2018). Pour des symptômes plutôt ponctuels, comme un stress aigu ou des hallucinations, une autoévaluation concurrente représenterait bien le phénomène. L'ajout d'un délai favoriserait

un retour vers le niveau de base ou une évaluation générale de la journée, entrainant une distribution plus concentrée. Il est possible que les capteurs du téléphone capturent correctement l'expression de symptômes ponctuels, mais que l'autoévaluation n'y corresponde pas. Donc, le moment de la mesure semble influencer le construit qui est effectivement mesuré et que le modèle prédira (p. ex. : stress aigu vs stress quotidien).

Les sciences psychiatriques et psychologiques confondent souvent le rôle de l'évaluateur et de l'échelle dans l'évaluation de symptôme (Uher, 2023). De manière critique, la mesure obtenue avec une échelle de stress est le stress « selon le clinicien » ou « selon soi » et non « le stress ». Cette distinction met de l'avant le rôle de la subjectivité et des processus cognitifs de l'évaluateur dans l'obtention de la mesure. Seveso et al. (2020) ont évalué l'interprétation d'une échelle ordinale à 5 ancrages [Absent, Faible, Modérée, Sévère, Très sévère] auprès de 31 médecins de spécialités variées et de 1152 patients (Figure 19). En plus d'une variance importante de l'interprétation dans chaque groupe, les deux groupes avaient des interprétations statistiquement différentes pour les classes Absence, Sévère et Extrême. Une explication probable est que les cliniciens possèdent plusieurs patients comme référent (interindividuel) et qu'un patient utilise principalement son expérience personnelle (intraindividuel). Finalement, les auteurs proposent trois méthodes pour encoder les cibles ordinales selon l'interprétation propre à un évaluateur ou à un groupe d'évaluateurs, entrainant en une amélioration de la performance sur une tâche de régression continue.

Figure 19 *Interprétation continue d'une échelle ordinale*



Note. Diagrammes en violon des IEDs pour chaque niveau de sévérité (n docteur = 62, n patient = [1970, 2155, 1971, 1944, 1670]. Les CoIED sont indiqués par des barres verticales et les cercles désignent la médiane des RP pour chaque niveau de l'échelle et la strate (n docteur =31, n patients = 1152).
Tiré de *Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings*. Seveso et al. (2020).

5.2.2 Protocole d'évaluation d'un système IA

La collecte de données de phénotypage numérique est un défi logistique significatif. Elle implique entre autres de développer la technologie pour le serveur de recherche et une application mobile et de former le personnel clinique et les participants. Ces avancées sont transférables à une éventuelle utilisation clinique. À l'inverse, le développement de modèles prédictifs se fait hors du contexte clinique et la généralisabilité des performances reste à démontrer. Seules les études cliniques permettront d'évaluer justement la performance, l'utilité et les potentiels effets négatifs des modèles prédictifs.

Une étude sur l'efficacité d'un traitement tente de vérifier si une règle de décision $h: X \rightarrow Y$ améliore les indicateurs de santé. Une étude d'apprentissage automatique évalue l'efficacité d'une méthode de fouille pour trouver la meilleure règle de décision $\hat{h}$ à partir d'un échantillon de données (X, Y). L'évaluation d'une solution d'apprentissage automatique doit inclure à la fois la fouille pour la règle de décision par le modèle et l'utilisation de la règle de décision par le personnel médical. Les *prédictions performatives* constituent un défi central pour l'apprentissage

automatique en santé, mais demeurent très peu abordées. La notion de prédiction performative réfère au fait que l'existence de la prédiction influence le phénomène prédit. Par exemple, utiliser une prédiction du stress pour sélectionner un traitement influencera le stress. Au contraire, prédire la météo de demain ne devrait pas changer les risques d'orage. En pratique, les modèles d'apprentissage automatique doivent être réentrainés fréquemment pour bénéficier de nouvelles données et mitiger la dégradation de la performance (Huyen, 2022). Toutefois, si l'utilisation de la règle de décision $\hat{h}_1$ affecte les données (X, Y) qui serviront à la fouille pour la prochaine règle $\hat{h}_2$, il y a des risques de boucle de rétroaction dégénérative.

À titre d'exemple, une étude sur le risque de réhospitalisation pour pneumonie a observé une diminution du risque pour les patients avec une urée sanguine un peu au-dessus de 50 et de 100 (Caruana et al., 2015). Effectivement, le personnel soignant commence à traiter les patients avec une valeur de plus de 50 et commence la dialyse lorsque l'urée sanguine dépasse 100 (c.-à-d., des règles de décision non reliées à l'apprentissage automatique). Le processus de décision clinique a donc un effet performatif et observable sur les données. Un modèle trouvant une règle selon laquelle un patient avec une urée sanguine de 115 a une meilleure survie qu'un à 97 n'aurait pas tort, car la valeur 115 est associée avec recevoir un traitement. Toutefois, en changeant l'application de traitements selon cette règle de décision, les patients avec une valeur de ≈90 à 100 et dé-prioriserait ceux entre ≈100 et 120, ce qui affecterait leur chance de survie et serait observable parmi les données. Cette boucle de rétroaction illustre bien la dérive de concept lorsque l'utilisation de la règle $\hat{h}$ modifie la relation $X \rightarrow Y$. Par contre, une compréhension des mécanismes biologiques (c.-à-d., des modèles explicatifs) permet deux règles bien plus robustes « urée sanguine < 35 → normal » et « urée sanguine > 35 → une augmentation de l'urée sanguine augmente le risque ».

Pour évaluer l'efficacité d'un système d'apprentissage automatique, le système entier (collecte de données, transformation de données, entrainement de modèles, déploiement du modèle, calcul de prédictions) doit être mis en place pour s'assurer qu'il est robuste aux variations inter- et intra-individuelles et aux prédictions performatives. Déterminer l'efficacité d'un traitement (h) est une question causale. Le problème fondamental de la causalité est que l'on ne peut pas observer simultanément l'effet de la présence et de l'absence du traitement sur une personne.

En médecine, l'essai clinique randomisé est la méthode étalon-or pour démêler un effet causal. L'échantillonnage aléatoire est utilisé pour générer des groupes aux propriétés similaires puis le traitement est introduit dans l'un d'eux pour quantifier l'effet différencié sur la santé. La durée de l'essai clinique doit être suffisante pour observer des événements cliniques (c.-à-d., des variations intra-individuelles) et l'effet de prédictions performatives.

L'étude CrossCheck est l'une des rares à avoir complété un protocole d'essai clinique randomisé (Wang et al., 2017). Leur modèle a été entrainé à prédire le score du *7-item Brief Psychiatric Rating Scale* (BPRS) évalué par un clinicien à partir des données passives et actives des 30 derniers jours. Le système en place générait une prédiction BPRS chaque semaine selon laquelle l'équipe médicale appelait le patient si jugé à risque (score ≥ 12 ou augmentation de 10% depuis la semaine précédente) pour discuter de son état. Le système réentraînait le modèle d'un patient chaque fois qu'une nouvelle évaluation du clinicien était disponible (environ chaque semaine ou chaque mois). Cette étude fait plusieurs choses correctement, mais il faut identifier ces limites pour diriger les projets. D'abord, le système entier pour entrainer les modèles et faire des prédictions est correctement mis en place. On note que les modèles utilisent en entrée les autoévaluations quotidiennes en plus des données passives. Cependant, les articles sur l'étude CrossCheck se concentrent sur la performance prédictive et non l'effet sur la santé des patients. Est-ce que le groupe traitement a présenté un meilleur état de santé à la fin de l'étude que le groupe contrôle? L'intervention, soit appeler les patients considérés à risque était simple, mais il faut se demander si les ressources déployées à cet effet dans le cadre de l'étude seraient disponibles en général. Les études futures devraient collecter les données qui servent d'entrée au modèle (dans ce cas, les signaux du téléphone et les autoévaluations) auprès des deux groupes et seulement faire varier l'intervention médicale appliquée selon les prédictions. Les données du téléphone du groupe contrôle permettraient potentiellement de vérifier que les deux groupes sont comparables et recueillir les autoévaluations prévient un effet confondant associé avec remplir les questionnaires (Fortney et al., 2017).

Alors que l'essai clinique randomisé tire des conclusions entre groupes, d'autres méthodes d'inférence causale sont mieux adaptées pour détecter des effets au fil du temps. Les données longitudinales du phénotypage numérique seraient adaptées pour les méthodes de *différence-*

en-différences (*difference-in-difference*; un devis quasi expérimental « avant-après »; Currie & Gruber, 1996), et le *contrôle synthétique* (*synthetic control*; Abadie et al., 2010) notamment. Pour ces analyses, le protocole doit collecter les données du téléphone auprès des groupes traitement et contrôle. Aussi, des observations doivent être recueillies avant et après que les modèles prédictifs influencent les décisions cliniques. Pour estimer des effets causaux, ces méthodes ont besoin de participants qui reçoivent leur traitement régulier et d'autres pour qui le clinicien utilise les prédictions. Elles peuvent s'accommoder de données observationnelles sans randomisation grâce à des techniques d'appariement (*matching*; Imbens, 2014), ce qui pourrait permettre de recruter de plus grands échantillons. Aussi, cela permettrait deux niveaux d'implication, soit contribuer des données et contribuer des données et utiliser les modèles prédictifs. Considérant la nature intrusive et expérimentale du phénotypage numérique, un tel protocole avec un consentement en deux phases semble porter moins d'enjeux éthiques. Il serait même possible de permettre à la population générale de partager ses données. Avec de grands échantillons, il devient possible de calculer l'*effet hétérogène du traitement* (*heterogeneous treatment effect*; Currie & Gruber, 1996), soit une estimation de quels sous-groupes bénéficieraient du traitement (c.-à-d., l'utilisation des modèles prédictifs).

5.2.3 Aide à la décision

Certains supposent qu'obtenir de l'information plus juste ou en plus grande quantité se traduira par une amélioration des soins (Stephens-Davidowitz & Pinker, 2017). Cependant, le clinicien doit pouvoir comprendre pleinement cette information et l'intégrer dans son processus décisionnel. Avec le grand nombre d'algorithmes et de méthodes existantes, c'est une tâche ardue pour le personnel soignant de développer une opinion nuancée de chaque modèle ou prédiction. D'ailleurs, 76 % des étudiants en médecine sondés identifient les connaissances des médecins en matière de santé connectée comme un enjeu important ou très important (un problème similaire aux échelles standardisées; section 1.1.3) et 63 % jugent probable ou très probable de recommander un objet médical à un patient (Paré et al., 2022). Il n'est pas clair comment les prédictions doivent être intégrées dans le raisonnement clinique et comment s'assurer d'une interprétation standardisée des prédictions. Une étude expérimentale rapporte que fournir une prédiction de l'état dépressif à un médecin avant la rencontre clinique n'a pas amélioré la qualité

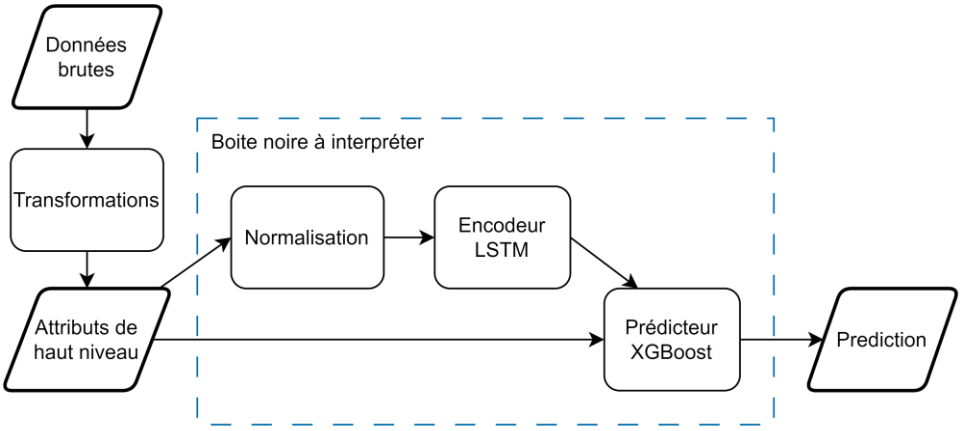
des soins par rapport au groupe contrôle (Rollman et al., 2002). Sans compréhension nuancée des modèles prédictifs, les prédictions risquent d'être acceptées de manière autoritaire ou selon les biais du clinicien (c.-à-d., si elles concordent avec l'avis préétabli). Au sens large, il y a un écart à franchir entre prédiction et décision. Parmi les étapes de développement présentées à la section 1.3.4, l'étape *11) Déploiment et utilisation* met de l'avant le défi humain d'utiliser l'apprentissage automatique.

L'avenue la plus explorée pour aider le processus décisionnel est l'ajout de méthodes d'interprétabilité. Dans l'exemple sur la pneumonie (section 5.2.2), les explications du modèle ont permis une appréciation critique des prédictions. Néanmoins, la proposition d'ajouter des explications ramène au problème initial de la supposition que « plus d'information se traduit par de meilleures décisions » (Torous et al., 2015). À l'inverse, l'hypothèse que plus d'information entraine une surcharge cognitive ou diminue l'efficacité des services mérite d'être examinée. En comparant les pratiques actuelles (contrôle), l'utilisation d'un modèle prédictif et l'utilisation d'un modèle avec explications, une série d'études empiriques auprès d'une compagnie d'assurance a conclu en l'absence d'effets bénéfiques additionnels pour les explications (Amarasinghe et al., 2023). L'utilisation de modèles prédictifs a augmenté l'efficacité (nombre de transactions traitées) des utilisateurs sans améliorer la qualité de leurs décisions. L'ajout d'explications qui provenaient du modèle ou qui avaient été générées aléatoirement a diminué le gain d'efficacité et a augmenté la confiance des utilisateurs envers leurs décisions sans améliorer leur qualité. Autrement, le rôle des explications est souvent mal compris et elles risquent d'être utilisées de manière erronée. La première erreur est d'avoir une interprétation causale d'une explication (Birk & Samuel, 2020). Par exemple, « le rôle important de la variable mobilité suggère que plus de mobilité améliore l'état mental » (traduction libre et paraphrase; Saeb et al., 2015). En plus, certains articles des interprétations causales bidirectionnelles comme à la fois « l'importance forte de la mobilité indique que plus de mobilité améliore l'humeur » et « l'importance forte des appels indique qu'une humeur faible entraine plus d'appels ». La deuxième erreur est d'oublier que les explications sont indicatives du processus de décision du modèle et ne constituent pas un modèle explicatif des états mentaux (section 1.3.3). Une interprétation correcte serait « l'importance forte de la mobilité montre que le modèle considère cette information importante

pour prédire l'état mental autoévalué », rappelant que le modèle prédit la valeur autoévaluée de stress et non « le stress » (section 5.2.1).

Le domaine de l'interprétabilité s'est développé en réponse à la complexité croissante des algorithmes d'apprentissage. Alors qu'il est possible de raisonner sur le processus décisionnel d'un arbre décisionnel, les réseaux de neurones ne sont pas directement interprétables. On parle de « boite noire » (*black box*) parce que seules les entrées et les sorties sont observées et le fonctionnement interne demeure opaque. L'interprétabilité est donc une composante ad hoc pour comprendre les relations extraites à partir des données. Corolairement, la pertinence d'une explication dépend de la nature informative des attributs en entrée. Par exemple, une prédiction de stress élevé pourrait être expliquée par une mesure brute du nombre de réseaux Wi-Fi détectés en matinée ou une mesure de plus haut niveau comme le nombre de lieux visités. La majorité des études de phénotypage transforment les données passives brutes, mais entraine tout de même leurs modèles prédictifs avec des attributs peu interprétables en soi. En utilisant des attributs en entrées qui correspondent à de l'information cliniquement utile, les prédictions et leurs explications seraient plus faciles à intégrer dans le processus décisionnel. Aussi, la littérature existante se concentre à expliquer la boite noire qu'est le modèle, mais laisse de côté toutes les étapes de prétraitement des données qui peuvent elles aussi avoir des effets considérables sur les prédictions. Les méthodes d'interprétabilité agnostiques au modèle peuvent être étendues aux autres étapes de transformation des données pour délimiter une « boite noire à expliquer » dont les entrées et les sorties ont une valeur explicative (Molnar, 2022a). La Figure 20 illustre une approche pour combiner un encodeur LSTM avec un prédicteur XGBoost (décrit à la section 5.2.1) qui préserve son interprétabilité en fournissant des explications « allant des attributs de haut niveau jusqu'aux prédictions ». Les études futures devraient explorer la tension entre fournir des explications interprétables et couvrir pleinement le parcours des données.

Figure 160 *Boite noire interprétable étendue*



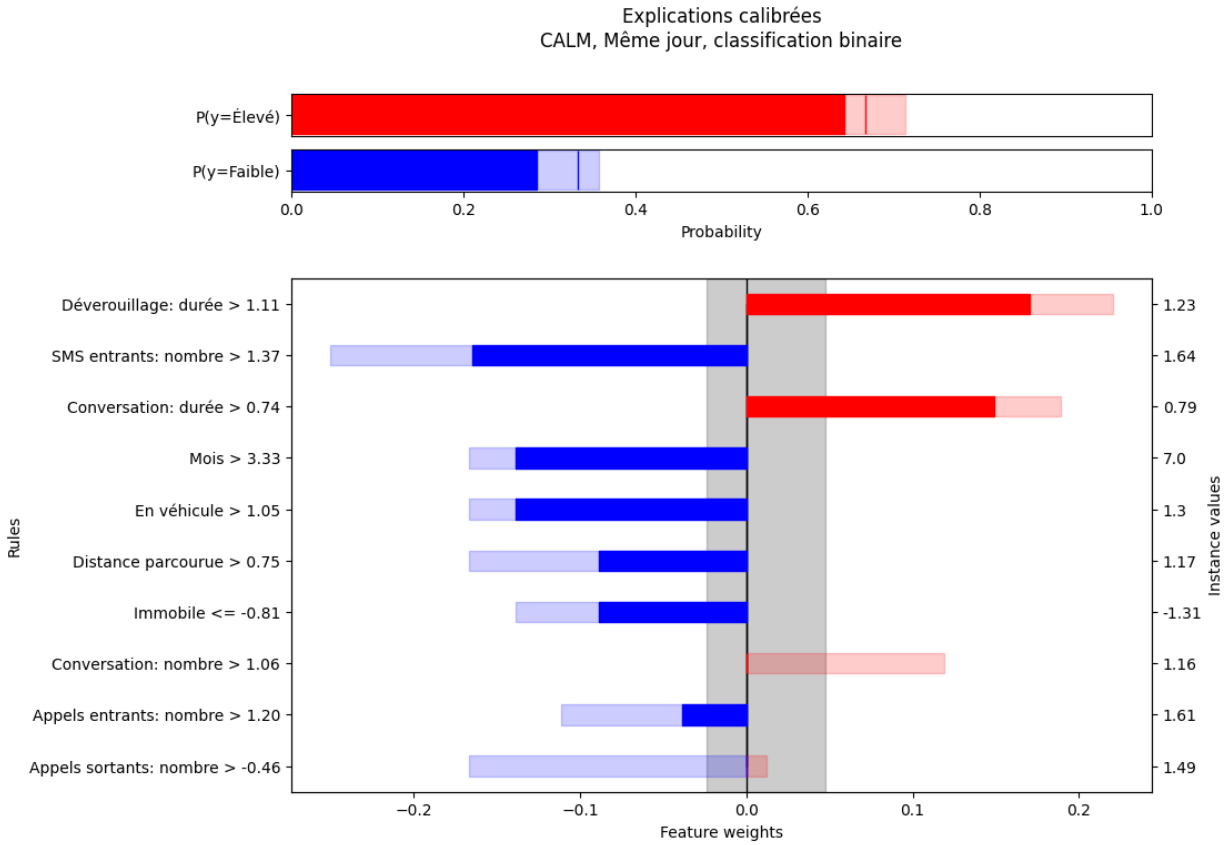
Note. Le schéma montre un modèle prédictif à plusieurs composantes et délimite la partie à expliquer par une méthode d'interprétabilité.

La notion de compromis entre la performance prédictive et l'interprétabilité d'un modèle est prévalente dans la littérature scientifique, mais peu substantifiée (Gunning et al., 2021). Bien qu'un réseau de neurones puisse en principe représenter des fonctions plus complexes qu'un arbre décisionnel (c.-à-d., plus grande capacité), cela ne permet pas de conclure que sur une tâche donnée un modèle non interprétable performera mieux qu'un modèle interprétable. D'ailleurs, le laboratoire de Cyntia Rudin a reproduit plusieurs études avec des performances comparables à partir de modèles plus simples et plus interprétables (Semenova et al., 2022). Leur équipe propose l'ensemble de Rashomon (*Rashomon set*) pour décrire l'ensemble de modèles ayant des performances équivalentes, mais des explications distinctes (Fisher et al., 2019). Ce concept nous invite à sélectionner un modèle ou un algorithme ayant des propriétés explicatives désirables, comme des attributs interprétables au niveau clinique, en plus d'une bonne performance. Par exemple, au lieu de créer des modèles statistiques complexes, l'apprentissage automatique peut servir à fabriquer des règles de décision simples, déjà prévalentes en milieu clinique (Bravo et al., 2023). Leurs travaux critiquent le fait que les méthodes d'interprétabilité sont utilisées pour cautionner des modèles dont la complexité superflue dissimule les biais et diminue la capacité de l'utilisateur de former une opinion critique des prédictions. Le revers de l'ensemble de Rashomon est qu'il est possible de volontairement ou involontairement créer des modèles biaisés et discriminatoires, mais avec des explications anodines et socialement acceptables (Lakkaraju & Bastani, 2020; Slack et al., 2020).

Plusieurs faiblesses des méthodes d'interprétabilité proviennent de leur nature ad hoc au modèle prédictif. Pour une configuration et un jeu de données particuliers, l'entrainement d'un modèle prédictif contient tout de même une part d'aléatoire. Par conséquent, les explications représentant les relations apprises par le modèle varient aussi. En plus, le calcul d'explication peut lui-même contenir une part d'aléatoire. Lorsqu'un modèle déployé en contexte clinique est mis à jour en avec de nouvelles données, il y aura des variations parmi les relations apprises et les explications associées. Peu importe le niveau de performance du modèle, les méthodes d'interprétabilité comme SHAP vont générer une explication. Les études de simulation montrent que des attributs reçoivent une importance élevée même s'ils ne sont aucunement liés à la cible et que la prédiction est erronée (Löfström et al., 2023b). Des méthodes comme SHAP servent à expliquer une prédiction, qu'elle soit bonne ou mauvaise. Outre l'interprétabilité, il serait plus facile d'intégrer les modèles d'apprentissage automatique en contexte clinique s'ils fournissaient des prédictions probabilistes. Les prédictions conformelles constituent une méthode flexible capable de transformer les prédictions ponctuelles en intervalles conformels qui garantissent la présence de la valeur selon un seuil prédéfini.

Contrairement au bootstrapping et aux intervalles de confiance, la prédiction conformelle dépend des données particulières reçues en entrée et génère un intervalle pour chaque prédiction, ce qui indique la difficulté ou la certitude associée à cette prédiction. L'innovation récente de la méthode *calibrated explanations* (Löfström et al., 2023a; T. Löfström et al., 2023) est de générer des prédictions conformelles et d'assigner une importance à chaque attribut en entrée selon sa contribution à la prédiction et l'intervalle de certitude, réglant la dissociation performance-interprétabilité (Figure 21). Cette méthode d'interprétabilité est agnostique au modèle et nous avons pu l'implémenter avec succès pour la classification binaire, la régression continue, la classification multiclasse et la régression ordinale. Il est même possible de convertir directement ces explications en valeurs SHAP.

Figure 17 *Explication calibrée d'une prédiction binaire de l'état CALM du même jour*



Note. En haut, il y a la probabilité de CALM Élevé et Faible. En dessous, on retrouve l'importance du top 10 attributs sur 29 pour la prédiction. Une couleur pâle indique un intervalle conformel. L'axe de droite présente la valeur des attributs et celui de droite les règles utilisées pour calculer les intervalles.

5.2.4 Santé mentale numérique

Dans ce projet, les modèles ont obtenu une erreur balancée moyenne d'environ 1 à 2 points sur des échelles à 4 ancrages. Bien que la performance paraisse basse face à d'autres articles, les résultats de ce travail et la série d'arguments méthodologiques présentés en discussion mènent à conclure en la présence d'un biais optimiste et de lacunes méthodologiques au sein de la littérature. Autrement, il n'y a pas de mesure étalon à laquelle comparer les prédictions, car les cliniciens sont dans l'impossibilité d'évaluer une dizaine de variables de santé mentale quotidienne pour chacun de leurs patients. Là est l'apport unique du phénotypage numérique, soit la production d'une nouvelle source d'information pour les cliniciens. Comme discuté, la performance prédictive nécessaire est relative à l'utilisation des prédictions. Par exemple, les modèles existants pourraient être suffisamment fiables pour informer les questions de l'entrevue

clinique (p. ex. : un niveau de stress anormal a été signalé il y a 20 jours). L'utilité réelle du phénotypage numérique pourra seulement être démontrée à travers des protocoles cliniques.

Maintenant, il est plus intéressant de se questionner quant aux implications de modèles prédictifs arbitrairement performants, atteignant par exemple 90 % de justesse. La psychiatrie fournit un cadre définissant la santé et la maladie mentale et au travers duquel les individus interprètent leurs expériences (Fabián et al., 2023). L'adoption même du phénotypage numérique comme pratique influence notre conception de la santé mentale. Les cadres dominants (p. ex. : psychanalytique, béhavioriste, cognitiviste, neuroscientifique) ont successivement cherché à établir des marqueurs « objectifs » pour se dégager de l'expérience personnelle. Le déplacement vers le numérique est sous-tendu par l'hypothèse que le numérique est transparent et honnête et donc plus fiable que l'expérience subjective (Fabián et al., 2023). Le passage d'un cadre à l'autre crée des tensions, car chaque construit n'a pas une correspondance une à une dans le nouveau cadre (Onnela & Rauch, 2016; Taschereau-Dumouchel et al., 2022). Par exemple, les béhavioristes ont rejeté l'existence d'un inconscient, mais admettaient des sources de renforcement intrinsèque. De la même manière, les catégories du DSM ne concordent pas aux mécanismes neurologiques. En dépit de ces transformations, la détresse subjective demeure un aspect essentiel aux décisions cliniques et beaucoup (Taschereau-Dumouchel et al., 2022).

Pour les courants prénumériques, le travail du clinicien consiste à intégrer de l'information objective et subjective pour guider le patient. Dans l'ère numérique, cette tâche est partiellement déléguée aux modèles prédictifs et le clinicien se trouve à devoir comprendre le modèle en plus du patient. Considérant la complexité des modèles prédictifs, il est improbable que les cliniciens et les patients aient l'occasion d'adopter une position critique pour chaque prédiction. En pratique, ils devront se remettre à faire confiance ou non aux modèles, leur concédant une certaine autorité. Dans ces circonstances, la prédiction du modèle et les justifications relatives « au niveau de Wi-Fi élevé et le nombre de contacts téléphoniques » font office d'explications *en soi*. De la même manière, des résultats d'imagerie cérébrale ou une séance avec un psychanalyste offrent un cadre permettant d'interpréter son expérience.

Le phénotypage numérique promet de mieux capturer les réalités individuelles, de fournir un portrait contextualisé de la santé mentale, mais la valeur explicative d'un modèle est limitée par les attributs en entrée. Les signaux du téléphone échouent à représenter la réalité économique et sociale d'un individu. Par exemple, les infrastructures et les services de transport ne sont pas accessibles de la même manière pour tous (Birk & Samuel, 2020). La concentration sur l'individu dé-collectivise la santé et met la responsabilité sur l'individu. Cependant, les composantes de son environnement non capturées par le téléphone continuent d'imposer des contraintes. Similairement, Gómez-Carrillo et al. (2023) soulignent le cadre limité de la psychiatrie de précision et appelle à intégrer des dimensions développementale, sociale, culturelle et expérientielle.

5.2.5 Enjeux éthiques

Le phénotypage numérique repose sur la collecte massive de données personnelles en continu et vient avec un lot de considérations éthiques et pratiques. L'existence de données brutes à haute résolution sur les comportements d'utilisateur pose un risque pour sa vie privée, permettant notamment de reconstruire ses déplacements. L'anonymisation, l'obfuscation et le chiffrement de données directement sur l'appareil avant d'être transférées sur le serveur de recherche sont idéaux. Comme chaque mesure de sécurité détruit de l'information contenue dans les données, les chercheurs peuvent avoir tendance à adopter des mesures de sécurité des plus sommaires afin de permettre des analyses secondaires non anticipées. Les serveurs de recherche et les téléphones des participants peuvent devenir les cibles d'acteurs malveillants cherchant à intercepter ces données. De plus, la majorité des applications de collecte de données requiert des privilèges élevés non standards pour accéder directement aux capteurs. Cela en fait des cibles de choix pour les pirates informatiques désirant prendre contrôle de l'appareil.

Un enjeu éthique particulier au phénotypage numérique est l'utilisation secondaire des données. Pour conduire une étude scientifique, le consentement des participants doit être obtenu après leur avoir présenté la portée de l'étude ainsi que les risques et les bénéfices associés. Contrairement aux données d'imageries cérébrales ou autres mesures en santé, les données personnelles du téléphone sont d'intérêt pour une multitude d'acteurs externes (p. ex. : agence de publicité, assureur, gouvernement, corps policiers, pirate informatique). Par conséquent, il est

essentiel que l'utilisation des données permises et les politiques de rétention soient explicites pour que le participant puisse apprécier la balance de risques et bénéfices réels et fournir un consentement éclairé. De plus, les partenariats publics-privés comportent un risque significatif pour l'intégrité du consentement étant donné que les lois québécoises et canadiennes sont beaucoup plus flexibles envers les compagnies privées que la recherche scientifique. Notamment, il n'existe pas de recours permettant à un particulier de faire supprimer les données personnelles qu'une entreprise possède.

La littérature en santé numérique doit être considérée avec un regard critique sachant les incitatifs financiers rattachés. En 2015, après 13 ans en tant que directeur du *National Institute for Mental Health* (NIMH) à faire la promotion des neurosciences, Thomas Insel rejoint la jeune entreprise *Verily* en santé numérique affiliée à Google (Wikipedia contributors, n.d.). Il se concentre sur la création d'outils commerciaux de phénotypage numérique. C'est en 2017 qu'il publie l'article de perspective *Digital Phenotyping: Technology for a New Science of Behavior*, cité plus de 600 fois, alors qu'il est en train de lancer sa propre entreprise *Mindstrong*. Une telle prise de position a le potentiel d'orienter le travail scientifique et les fonds investis de manière importante. Cependant, cela peut avoir des effets plus directs sur la production scientifique. Par exemple, l'étude de Pratap et al. (2019) compte deux auteurs cofondateurs d'une entreprise en démarrage en santé mentale, la dernière auteure fait des mandats de consultation pour *Verily* et le protocole de l'étude récompense les participants avec 20$ pour utiliser l'application commerciale *Ginger.io*. D'ailleurs, *Ginger.io* a été acquise par *Headspace* (qui a débuté avec une application de méditation populaire) en 2021. La couverture médiatique rapporte que « Headspace Health now boasts more than 2,700 enterprise and health plan customers with combined bookings by end of 2021 of nearly $300 million, the companies said. The combined company now **has the world's largest mental health data set, which will be leveraged** to deliver highly personalized care, executives said » (mise en gras par l'auteur; Landi, 2021). Avec plus de 160 millions en fonds investi, l'entreprise a fermé et vendu sa technologie à *SonderMind* en 2023 semblablement après avoir reçu la pression d'investisseurs pour commercialiser leur produit (Landi, 2023). Les résultats d'études et la manière dont ils sont rapportés ont donc un effet sur la

recherche, mais aussi sur la technologie déployée en ce moment et la valorisation d'entreprises privées. Considérant la difficulté d'accès aux données de santé, elles valent leur pesant d'or.

5.3 Conclusion

Le phénotypage numérique offre une perspective unique sur le comportement humain grâce à des mesures répétées, fréquentes, en contexte écologique, et sur de grands échantillons. C'est un outil riche pouvant servir la recherche dans un ensemble de domaines, dont les sciences sociales, la santé publique, les sciences biomédicales et la santé mentale. Depuis ses débuts, l'apprentissage automatique joue un rôle clé dans le traitement des larges volumes de données non interprétables générées.

Néanmoins, le phénotypage numérique rassemble plusieurs propriétés qui en font un problème de taille pour l'apprentissage automatique. La complexité de la collecte de données introduit du bruit à plusieurs étapes: les capteurs, l'usure de l'appareil, l'application de collecte, le traitement et l'agrégation des signaux bruts. Ensuite, le comportement humain capté en entrée et les états mentaux servant de cibles présentent une grande hétérogénéité intra- et inter-individuelle, en particulier lorsqu'il est question de santé mentale. Finalement, les données manquantes, le relativement petit nombre d'exemples (pour la difficulté de la tâche), le débalancement des classes, les exigences de niveau de performance, le besoin d'interprétabilité, la tâche de prévision, le risque de dérive distributionnelle et de prédictions performatives rendent l'entrainement de modèles particulièrement compliqué.

Aux termes de ce travail, il m'apparait essentiel d'élargir les objectifs de la recherche de la performance prédictive à d'autres dimensions d'aide à la décision. Plusieurs problèmes déjà identifiés dans la littérature demeurent sous-explorés et doivent être résolus afin de réaliser la vision d'une psychiatrie pragmatique. En spécifiant des cas d'utilisation précis, il sera possible de plus facilement spécifier la création d'attributs interprétables, la tâche d'apprentissage, le type de modèle probabiliste, la méthode d'interprétation et le protocole expérimental permettant d'évaluer concrètement les bénéfices du phénotypage numérique pour les soins. Et par la même occasion, inclure les différentes parties prenantes dans le développement de cette technologie ayant le potentiel de directement transformer leurs soins.

Références bibliographiques

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., & Choudhury, T. (2016). Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association*, *23*(3), 538–543. https://doi.org/10.1093/jamia/ocv200

Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, *7*(6), 643–659. https://doi.org/10.1016/j.pmcj.2011.09.004

Amarasinghe, K., Rodolfa, K. T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., Bizarro, P., Talwalkar, A., & Ghani, R. (2023). *On the importance of application-grounded experimental design for evaluating explainable ML methods.* https://doi.org/10.48550/arXiv.2206.13503

American Psychiatric Association. (1952). *Diagnostic and Statistical Manual of mental disorders (DSM)* (1ère éd.).

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of mental disorders: DSM-5™* (5e éd.). https://doi.org/10.1176/appi.books.9780890425596

American Psychiatric Association. (2015). *The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults* (3e éd.). https://doi.org/10.1176/appi.books.9780890426760

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Evaluation measures for ordinal regression. *2009 Ninth International Conference on Intelligent Systems Design and Applications*, 283–287. https://doi.org/10.1109/ISDA.2009.230

Bati, G. F., & Singh, V. K. (2018). "Trust us": Mobile phone use patterns can predict individual trust propensity. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173904

Beaudry, N. J., & Renner, R. (2012). An intuitive proof of the data processing inequality. *Quantum Information and Computation*, *12*(5), 432–441. https://doi.org/10.26421/QIC12.5-6-4

Ben-Zeev, D., Brenner, C. J., Begale, M., Duffecy, J., Mohr, D. C., & Mueser, K. T. (2014). Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia Bulletin*, *40*(6), 1244–1253. https://doi.org/10.1093/schbul/sbu033

BinDhim, N. F., Shaman, A. M., Trevena, L., Basyouni, M. H., Pont, L. G., & Alhawassi, T. M. (2015). Depression screening via a smartphone app: Cross-country user characteristics and feasibility. *Journal of the American Medical Informatics Association*, *22*(1), 29–34. https://doi.org/10.1136/amiajnl-2014-002840

Birk, R., & Samuel, G. (2020). Can digital data diagnose mental health problems? A sociological exploration of "digital phenotyping." *Sociology of Health & Illness*, *42*(8), 1873–1887. https://doi.org/10.1111/1467-9566.13175

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G. M., Groenwold, R. H. H., & Van Smeden, M. (2018). Measurement error is often neglected in medical literature: A systematic review. *Journal of Clinical Epidemiology*, *98*, 89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023

Bravo, F., Rudin, C., Shaposhnik, Y., & Yuan, Y. (2023). Interpretable prediction rules for congestion risk in intensive care units. *Stochastic Systems*, *0*(0). https://doi.org/10.1287/stsy.2022.0018

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Burcusa, S. L., & Iacono, W. G. (2007). Risk for recurrence in depression. *Clinical Psychology Review*, *27*(8), 959–985. https://doi.org/10.1016/j.cpr.2007.02.005

Canzian, L., & Musolesi, M. (2015). Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, 1293–1304. https://doi.org/10.1145/2750858.2805845

Carlens, H. (2023). State of competitive machine learning in 2022. *ML Contests Research*. Repéré le 20 mai 2023 à https://mlcontests.com/state-of-competitive-data-science-2022

Caruana, R. A., Lou, Y., Gehrke, J. E., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Challu, C., Olivares, K. G., Oreshkin, B. N., Garza Ramirez, F., Mergenthaler Canseco, M., & Dubrawski, A. (2023). NHITS: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(6), 6989–6997. https://doi.org/10.1609/aaai.v37i6.25854

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chiauzzi, E., & Wicks, P. (2021). Beyond the therapist's office: Merging measurement-based care and digital medicine in the real world. *Digital Biomarkers*, *5*(2), 176–182. https://doi.org/10.1159/000517748

Chiu, M., Gatov, E., Zaheer, J., Lebenbaum, M., Fu, L., Newman, A., & Kurdyak, P. (2018). Postdischarge service utilisation and outcomes among chinese and south asian psychiatric inpatients in ontario, canada: A population-based cohort study. *BMJ Open*, *8*(1), e020156. https://doi.org/10.1136/bmjopen-2017-020156

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Choudhary, S., Thomas, N., Alshamrani, S., Srinivasan, G., Ellenberger, J., Nawaz, U., & Cohen, R. (2022). A machine learning approach for continuous mining of nonidentifiable smartphone data to create a novel digital biomarker detecting generalized anxiety disorder: Prospective cohort study. *JMIR Medical Informatics*, *10*(8), e38943. https://doi.org/10.2196/38943

Cohen, A. S., Fedechko, T., Schwartz, E. K., Le, T. P., Foltz, P. W., Bernstein, J., Cheng, J., Rosenfeld, E., & Elvevåg, B. (2019). Psychiatric risk assessment from the clinician's perspective: Lessons for the future. *Community Mental Health Journal*, *55*(7), 1165–1172. https://doi.org/10.1007/s10597-019-00411-x

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2e éd.). Wiley-Interscience.

Currey, D., & Torous, J. (2022). Digital phenotyping correlations in larger mental health samples: Analysis and replication. *BJPsych Open*, *8*(4), 1-7. https://doi.org/10.1192/bjo.2022.507

Currie, J., & Gruber, J. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Business & Economic Statistics*, *14*(2), 146–155. https://doi.org/10.3386/t0147

Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association : JAMIA*, *24*(6), 1052–1061. https://doi.org/10.1093/jamia/ocx030

Dawkins, R. (1982). *The extended phenotype* (1ère éd.). Oxford University Press.

Depp, C. A., Bashem, J., Moore, R. C., Holden, J. L., Mikhael, T., Swendsen, J., Harvey, P. D., & Granholm, E. L. (2019). GPS mobility as a digital biomarker of negative symptoms in schizophrenia: A case control study. *npj Digital Medicine*, *2*, 108. https://doi.org/10.1038/s41746-019-0182-1

Diefendorf, A. R., & Kraepelin, E. (1907). Clinical psychiatry: A textbook for students and physicians, abstracted and adapted from the 7th German edition of Kraepelin's "Lehrbuch der Psychiatrie". MacMillan Co. https://doi.org/10.1037/13656-000

Dong, W., Lepri, B., & Pentland, A. (Sandy). (2011). Modeling the co-evolution of behaviors and social relationships using mobile phone data. *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia - MUM '11*, 134–143. https://doi.org/10.1145/2107596.2107613

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, *24*(11), 1583–1598. https://doi.org/10.1038/s41380-019-0365-9

Eagle, N., & Pentland, A. (Sandy). (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, *10*(4), 255–268. https://doi.org/10.1007/s00779-005-0046-3

Elor, Y., & Averbuch-Elor, H. (2022). To SMOTE, or not to SMOTE? https://doi.org/10.48550/arXiv.2201.08528

Emsley, R., Chiliza, B., Asmal, L., & Harvey, B. H. (2013). The nature of relapse in schizophrenia. *BMC Psychiatry*, *13*(1), 50. https://doi.org/10.1186/1471-244X-13-50

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology*, *8*(6), 985–987. https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2

Fabián, R. D. L., Jiménez-Molina, Á., & Obaid, F. P. (2023). A critical analysis of digital phenotyping and the neuro-digital complex in psychiatry. *Big Data & Society*, *10*(1), 20539517221149097. https://doi.org/10.1177/20539517221149097

Faurholt-Jepsen, M., Busk, J., Rohani, D. A., Frost, M., Tønning, M. L., Bardram, J. E., & Kessing, L. V. (2022). Differences in mobility patterns according to machine learning models in patients with bipolar disorder and patients with unipolar disorder. *Journal of Affective Disorders*, *306*, 246–253. https://doi.org/10.1016/j.jad.2022.03.054

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.

Journal of Machine Learning Research, 177(20), 1-81. https://www.jmlr.org/papers/v20/18-760.html

Fojo, A. T., Musliner, K. L., Zandi, P. P., & Zeger, S. L. (2017). A precision medicine approach for psychiatric disease based on repeated symptom scores. *Journal of Psychiatric Research*, *95*, 147–155. https://doi.org/10.1016/j.jpsychires.2017.08.008

Fortney, J. C., Unützer, J., Wrenn, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., & Harbin, H. T. (2017). A tipping point for measurement-based care. *Psychiatric Services*, *68*(2), 179–188. https://doi.org/10.1176/appi.ps.201500439

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, *51*, 1–26. https://doi.org/10.1016/j.pmcj.2018.09.003

Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller, V. I., & Langner, R. (2023). The burden of reliability: How measurement noise limits brain-behaviour predictions [Document soumis pour publication]. https://doi.org/10.1101/2023.02.09.527898

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Goldberg, S. B., Buck, B., Raphaely, S., & Fortney, J. C. (2018). Measuring psychiatric symptoms remotely: A systematic review of remote measurement-based care. *Current Psychiatry Reports*, *20*(10), 81. https://doi.org/10.1007/s11920-018-0958-z

Gómez-Carrillo, A., Paquin, V., Dumas, G., & Kirmayer, L. J. (2023). Restoring the missing person to personalized medicine and precision psychiatry. *Frontiers in Neuroscience*, *17*, 1041433. https://doi.org/10.3389/fnins.2023.1041433

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Repéré le 20 mai 2023 à http://www.deeplearningbook.org

Google. (2023, November 14). *Machine learning glossary. Machine learning glossary*. Repéré le 20 mai 2023 à https://developers.google.com/machine-learning/glossary

Greene, S. M., Tuzzio, L., & Cherkin, D. (2012). A framework for making patient-centered care front and center. *The Permanente Journal*, *16*(3), 49–53. https://doi.org/10.7812/TPP/12-025

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Dans S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh. *Advances in neural information processing systems* (Vol. 35, pp.

507–520). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA 's explainable AI program: A retrospective. *Applied AI Letters*, *2*(4), e61. https://doi.org/10.1002/ail2.61

Gustafson, D. H., McTavish, F. M., Chih, M.-Y., Atwood, A. K., Johnson, R. A., Boyle, M. G., Levy, M. S., Driscoll, H., Chisholm, S. M., Dillenburg, L., Isham, A., & Shah, D. (2014). A smartphone application to support recovery from alcoholism: A randomized clinical trial. *JAMA Psychiatry*, *71*(5), 566–572. https://doi.org/10.1001/jamapsychiatry.2013.4642

Hatfield, D., McCullough, L., Frantz, S. H. B., & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy*, *17*(1), 25–32. https://doi.org/10.1002/cpp.656

Henson, P., Wisniewski, H., Stromeyer Iv, C., & Torous, J. (2020). Digital health around clinical high risk and first-episode psychosis. *Current Psychiatry Reports*, *22*(11), 58. https://doi.org/10.1007/s11920-020-01184-x

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), 788–832. https://doi.org/10.1007/s10618-022-00894-5

Hitchcock, P., Fried, E., & Frank, M. (2022). Computational psychiatry needs time and context. *Annual Review of Psychology, 73*(1), 243-270. https://doi.org/10.1146/annurev-psych-021621-124910

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huyen, C. (2022). *Designing machine learning systems*. O'Reilly Media.

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology*, *6*(1), 155–179. https://doi.org/10.1146/annurev.clinpsy.3.022806.091532

Imbens, G. (2014). Matching methods in practice: three examples. *National Bureau of Economic Research* (Working Paper 19959). https://doi.org/10.3386/w19959

Insel, T. R. (2015). The NIMH experimental medicine initiative. *World Psychiatry*, *14*(2), 151–153. https://doi.org/10.1002/wps.20227

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. *JAMA*, *318*(13), 1215. https://doi.org/10.1001/jama.2017.11295

Institut canadien d'information sur la santé. (2023). *Health Indicators Interactive Tool*. Repéré le 10 novembre 2022 à https://yourhealthsystem.cihi.ca/epub/

Jacobson, N. C., & Feng, B. (2022). Digital phenotyping of generalized anxiety disorder: Using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life. *Translational Psychiatry*, *12*(1), 336. https://doi.org/10.1038/s41398-022-02038-1

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology*, *33*(5), 462–463. https://doi.org/10.1038/nbt.3223

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Dans I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett. *Advances in neural information processing systems* (Vol. 30, pp. 3146–3154). Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kendler, K. S. (2009). An historical framework for psychiatric nosology. *Psychological Medicine*, *39*(12), 1935–1941. https://doi.org/10.1017/S0033291709005753

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klement, W., & El Emam, K. (2023). Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: Development and validation. *Journal of Medical Internet Research*, *25*, e48763. https://doi.org/10.2196/48763

K_MAT. (2023, December 5). *2nd place solution - child mind institute - detect sleep states. Kaggle*. Repéré le 11 décembre 2023 à https://www.kaggle.com/competitions/child-mind-institute-detect-sleep-states/discussion/459627

Koppe, G., Guloksuz, S., Reininghaus, U., & Durstewitz, D. (2019). Recurrent neural networks in mobile sampling and intervention. *Schizophrenia Bulletin*, *45*(2), 272–276. https://doi.org/10.1093/schbul/sby171

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kwasnicka, D., Kale, D., Schneider, V., Keller, J., Yeboah-Asiamah Asare, B., Powell, D., Naughton, F., Ten Hoor, G. A., Verboon, P., & Perski, O. (2021). Systematic review of ecological

momentary assessment (EMA) studies of five public health-related behaviours: Review protocol. *BMJ Open*, *11*(7), e046435. https://doi.org/10.1136/bmjopen-2020-046435

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85. https://doi.org/10.1145/3375627.3375833

Landi, H. (2021). Ginger and Headspace plan merger to rapidly scale up digital mental health services. In *Fierce Healthcare*. Repéré le 6 décembre 2023 à https://www.fiercehealthcare.com/digital-health/ginger-and-headspace-plan-to-merge-to-scale-up-comprehensive-digital-mental-health

Landi, H. (2023). Mental health provider SonderMind buys Mindstrong's tech assets. In *Fierce Healthcare*. Repéré le 6 décembre 2023 à https://www.fiercehealthcare.com/health-tech/mental-health-provider-sondermind-buys-mindstrongs-tech-assets

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, *15*, 41–56.

Lee, K., Lee, T. C., Yefimova, M., Kumar, S., Puga, F., Azuero, A., Kamal, A., Bakitas, M. A., Wright, A. A., Demiris, G., Ritchie, C. S., Pickering, C. E. Z., & Nicholas Dionne-Odom, J. (2023). Using digital phenotyping to understand health-related outcomes: A scoping review. *International Journal of Medical Informatics*, *174*, 105061. https://doi.org/10.1016/j.ijmedinf.2023.105061

Lengerich, B., Tan, S., Chang, C.-H., Hooker, G., & Caruana, R. (2020). Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the twenty third international conference on artificial intelligence and statistics* (Vol. 108, pp. 2402–2412). PMLR. https://proceedings.mlr.press/v108/lengerich20a.html

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., Hoffman, M., Scott, K., Lyon, A., Douglas, S., Simon, G., & Kroenke, K. (2019). Implementing measurement-based care in behavioral health: A review. *JAMA Psychiatry*, *76*(3), 324. https://doi.org/10.1001/jamapsychiatry.2018.3329

Löfström, H., Löfström, T., Johansson, U., & Sönströd, C. (2023a). Calibrated explanations: With uncertainty information and counterfactuals. https://doi.org/10.48550/arXiv.2305.02305

Löfström, H., Löfström, T., Johansson, U., & Sönströd, C. (2023b). Investigating the impact of calibration on the quality of explanations. *Annals of Mathematics and Artificial Intelligence*, 1–18. https://doi.org/10.1007/s10472-023-09837-2

Löfström, T., Löfström, H., Johansson, U., Sönströd, C., & Matela, R. (2023). Calibrated explanations for regression. https://doi.org/10.48550/arXiv.2308.16245

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Dans I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, *Advances in neural information processing systems* (Vol. 30, pp. 4765-4774). Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Madan, A., Cebrian, M., Lazer, D., & Pentland, A. (2010). Social sensing for epidemiological behavior change. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 291–300. https://doi.org/10.1145/1864349.1864394

Madan, A., Cebrian, M., Moturu, S., Farrahi, K., & Pentland, A. "Sandy." (2012). Sensing the "Health State" of a Community. *IEEE Pervasive Computing*, *11*(4), 36–45. https://doi.org/10.1109/MPRV.2011.79

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Maser, J. D., & Patterson, T. (2002). Spectrum and nosology: Implications for DSM-V. *Psychiatric Clinics of North America*, *25*(4), 855–885. https://doi.org/10.1016/S0193-953X(02)00022-9

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Darthmouth Summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4) 12. https://doi.org/10.1609/aimag.v27i4.1904

McElfresh, D., Khandagale, S., Valverde, J., C, V. P., Ramakrishnan, G., Goldblum, M., & White, C. (2023). When do neural nets outperform boosted trees on tabular data? https://doi.org/10.48550/arXiv.2305.02997

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences https://doi.org/10.48550/arXiv.1706.07269

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, *13*(1), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117. https://doi.org/10.1111/j.1467-8721.2009.01619.x

Molnar, C. (2022a). Interpret Complex Pipelines By Drawing A Box. In *Mindful Modeler*. Repéré le 6 décembre 2023 à https://mindfulmodeler.substack.com/p/interpret-complex-pipelines-by-drawing

Molnar, C. (2022b). *Modeling mindsets: The many cultures of learning from data*.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. https://doi.org/10.1016/j.ijforecast.2019.02.011

Morgenstern, J., Kuerbis, A., & Muench, F. (2014). Ecological momentary assessment and alcohol use disorder treatment. *Alcohol Research*, *36*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4432849/

Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with python: A guide for data scientists*. O'Reilly Media, Inc.

National Institute of Mental Health. (2023). *Mental illness statistics*. Repéré le 6 décembre 2023 à https://www.nimh.nih.gov/health/statistics/mental-illness

Navas-Palencia, G. (2020). Optimal binning: Mathematical programming formulation. https://doi.org/10.48550/arXiv.2001.08025

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T. W., & Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder: A review. *JAMA Psychiatry*, *74*(5), 528. https://doi.org/10.1001/jamapsychiatry.2017.0001

Organisation mondiale de la Santé. *Comprehensive mental health action plan (2013-2030)*. (2021). Repéré le 12 octobre 2023 à https://www.who.int/publications/i/item/9789240031029

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W. et al. (2019). Calibration: The achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study. *JMIR mHealth and uHealth*, *9*(7), e26540. https://doi.org/10.2196/26540

Oreshkin, B. N., Chapados, N., Carpov, D., & Bengio, Y. (2020). N-BEATS: Neural Basis Expansion Analysis for interpretable time series forecasting. https://doi.org/10.48550/arXiv.1905.10437

Paré, G., Raymond, L., Pomey, M.-P., Grégoire, G., Castonguay, A., & Ouimet, A. G. (2022). Medical students' intention to integrate digital health into their medical practice: A pre-peri COVID-19 survey study in Canada. *DIGITAL HEALTH*, *8*, https://doi.org/10.1177/20552076221114195

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*(3), 800–809. https://doi.org/10.1148/radiol.2017171920

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, *13*(2), 217–224. https://doi.org/10.31887/DCNS.2011.13.2/npatsopoulos

Paulus, M. P. (2017). Evidence-based pragmatic psychiatry—a call to action. *JAMA Psychiatry*, *74*(12), 1185-1186. https://doi.org/10.1001/jamapsychiatry.2017.2439

Paulus, M. P., Huys, Q. J. M., & Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 386–392. https://doi.org/10.1016/j.bpsc.2016.05.001

Paulus, M. P., & Thompson, W. K. (2021). Computational approaches and machine learning for individual-level treatment predictions. *Psychopharmacology*, *238*(5), 1231–1239. https://doi.org/10.1007/s00213-019-05282-4

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, *77*(5), 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

Pramana, G., Parmanto, B., Kendall, P. C., & Silk, J. S. (2014). The SmartCAT: An m-health platform for ecological momentary intervention in child anxiety treatment. *Telemedicine and e-Health*, *20*(5), 419–427. https://doi.org/10.1089/tmj.2013.0214

Pratap, A., Atkins, D. C., Renn, B. N., Tanana, M. J., Mooney, S. D., Anguera, J. A., & Areán, P. A. (2019). The accuracy of passive phone sensors in predicting daily mood. *Depression and Anxiety*, *36*(1), 72–81. https://doi.org/10.1002/da.22822

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Dans S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, *Advances in neural information processing systems* (Vol. 31, pp. 6638-6648). Curran Associates, Inc.

Raymaekers, J., & Rousseeuw, P. J. (2021). Transforming variables to central normality. *Machine Learning*. https://doi.org/10.1007/s10994-021-05960-5

Regier, D. A., Kuhl, E. A., & Kupfer, D. J. (2013). The DSM-5: Classification and criteria changes. *World Psychiatry*, *12*(2), 92–98. https://doi.org/10.1002/wps.20050

Rhim, S., Lee, U., & Han, K. (2020). Tracking and modeling subjective well-being using smartphone-based digital phenotype. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 211–220. https://doi.org/10.1145/3340631.3394855

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

Riley, R. D., Ensor, J., Snell, K. I. E., Debray, T. P. A., Altman, D. G., Moons, K. G. M., & Collins, G. S. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ*, *353*, i3140. https://doi.org/10.1136/bmj.i3140

Rogler, L. H., Mroczek, D. K., Fellows, M., & Loftus, S. T. (2001). The neglect of response bias in mental health research. *The Journal of Nervous and Mental Disease*, *189*(3), 182–187. https://doi.org/10.1097/00005053-200103000-00007

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. *JMIR mHealth and uHealth*, *6*(8), e9691. https://doi.org/10.2196/mhealth.9691

Rollman, B. L., Hanusa, B. H., Lowe, H. J., Gilbert, T., Kapoor, W. N., & Schulberg, H. C. (2002). A randomized trial using computerized decision support to improve treatment of major depression in primary care. *J Gen Intern Med*, *17*(7), 493–503. https://doi.org/10.1046/j.1525-1497.2002.10421.x

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, *4*, e2537. https://doi.org/10.7717/peerj.2537

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, *17*(7), e175. https://doi.org/10.2196/jmir.4273

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3(3), 210-229. https://doi.org/10.1147/rd.33.0210

Sarda, A., Munuswamy, S., Sarda, S., & Subramanian, V. (2019). Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: Cross-sectional observational study. *JMIR mHealth and uHealth*, *7*(1), e11041. https://doi.org/10.2196/11041

Schneeberger, D., Stöger, K., & Holzinger, A. (2020). The european legal framework for medical AI. Dans A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), Dans *Machine learning and knowledge extraction* (Vol. 12279, pp. 209–226). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_12

Schwartz, D., & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*, *20*, 637–648. https://doi.org/10.1016/0021-9681(67)90041-0

Scikit-learn. (2023a). *Time-related feature engineering: Gradient boosting*. Dans *Scikit-learn User Guide*. Repéré le 2 décembre 2022 à https://scikit-learn.org/stable/auto_examples/applications/plot_cyclical_feature_engineering.html#gradient-boosting

Scikit-learn. (2023b). Time-related feature engineering: Periodic spline. Dans *Scikit-learn User Guide*. Repéré le 2 décembre 2022 à https://scikit-learn.org/stable/auto_examples/applications/plot_cyclical_feature_engineering.html#periodic-spline-features.

Semenova, L., Rudin, C., & Parr, R. (2022). On the existence of simpler machine learning models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1827–1858. https://doi.org/10.1145/3531146.3533232

Seveso, A., Campagner, A., Ciucci, D., & Cabitza, F. (2020). Ordinal labels in machine learning: A user-centered approach to improve data validity in medical settings. *BMC Medical Informatics and Decision Making*, *20*, 142. https://doi.org/10.1186/s12911-020-01152-8

Shi, X., Cao, W., & Raschka, S. (2023). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, *26*(3), 941–955. https://doi.org/10.1007/s10044-023-01181-9

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. https://doi.org/10.1145/3375627.3375830

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019a). Passive mobile sensing and psychological traits for large scale mood prediction. *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 272–281. https://doi.org/10.1145/3329189.3329213

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019b). Sequence multi-task learning to forecast mental wellbeing from sparse self-reported data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2886–2894. https://doi.org/10.1145/3292500.3330730

Steele, J. D., & Paulus, M. P. (2019). Pragmatic neuroscience for clinical psychiatry. *The British Journal of Psychiatry*, *215*(1), 404–408. https://doi.org/10.1192/bjp.2019.88

Stephens-Davidowitz, S., & Pinker, S. (2017). *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. Dey Street Books.

Strauss, G. P., Raugh, I. M., Zhang, L., Luther, L., Chapman, H. C., Allen, D. N., Kirkpatrick, B., & Cohen, A. S. (2022). Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia. *Schizophrenia*, *8*(1), 37. https://doi.org/10.1038/s41537-022-00241-z

Suhara, Y., Xu, Y., & Pentland, A. 'Sandy'. (2017). DeepMood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. *Proceedings of the 26th International Conference on World Wide Web*, 715–724. https://doi.org/10.1145/3038912.3052676

Sükei, E., Norbury, A., Perez-Rodriguez, M. M., Olmos, P. M., & Artés, A. (2021). Predicting emotional states using behavioral markers derived from passively sensed data: Data-driven machine learning approach. *JMIR mHealth and uHealth*, *9*(3), e24465. https://doi.org/10.2196/24465

Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G., & LeDoux, J. E. (2022). Putting the "mental" back in "mental disorders": A perspective from research on fear and anxiety. *Molecular Psychiatry*, *27*(3), 1322–1330. https://doi.org/10.1038/s41380-021-01395-5

Tan, Y. S., Singh, C., Nasseri, K., Agarwal, A., & Yu, B. (2022). Fast interpretable greedy-tree sums (FIGS). https://doi.org/10.48550/arXiv.2201.11931

Testa, A., Kaijser, J., Wynants, L., Fischerova, D., Van Holsbeke, C., Franchi, D., Savelli, L., Epstein, E., Czekierdowski, A., Guerriero, S., Fruscio, R., Leone, F. P. G., Vergote, I., Bourne, T., Valentin, L., Van Calster, B., & Timmerman, D. (2014). Strategies to diagnose ovarian cancer: New evidence from phase 3 of the multicentre international IOTA study. *British Journal of Cancer*, *111*(4), 680–688. https://doi.org/10.1038/bjc.2014.333

Thai, T. N., & Ebell, M. H. (2019). Prospective validation of the good outcome following attempted resuscitation (GO-FAR) score for in-hospital cardiac arrest prognosis. *Resuscitation*, *140*, 2–8. https://doi.org/10.1016/j.resuscitation.2019.05.002

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Tiego, J., & Fornito, A. (2022). Putting behaviour back into brain–behaviour correlation analyses [Document soumis pour publication]. https://doi.org/10.31219/osf.io/g84j2

Torous, J., & Baker, J. T. (2016). Why psychiatry needs data science and data science needs psychiatry: Connecting with technology. *JAMA Psychiatry*, *73*(1), 3. https://doi.org/10.1001/jamapsychiatry.2015.2622

Torous, J., Gershon, A., Hays, R., Onnela, J.-P., & Baker, J. T. (2019). Digital phenotyping for the busy psychiatrist: Clinical implications and relevance. *Psychiatric Annals*, *49*(5), 196–201. https://doi.org/10.3928/00485713-20190417-01

Torous, J., Staples, P., & Onnela, J.-P. (2015). Realizing the potential of mobile mental health: New methods for new data in psychiatry. *Current Psychiatry Reports*, *17*(8), 61. https://doi.org/10.1007/s11920-015-0602-0

Trivedi, M. H. (2009). Tools and strategies for ongoing assessment of depression: A measurement-based approach to remission. *The Journal of Clinical Psychiatry*, *70*, 26–31. https://doi.org/10.4088/JCP.8133su1c.04

Tseng, V. W.-S., Sano, A., Ben-Zeev, D., Brian, R., Campbell, A. T., Hauser, M., Kane, J. M., Scherer, E. A., Wang, R., Wang, W., Wen, H., & Choudhury, T. (2020). Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Scientific Reports*, *10*(1), 15100. https://doi.org/10.1038/s41598-020-71689-1

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass*, *17*(5), e12740. https://doi.org/10.1111/spc3.12740

Umematsu, T., Sano, A., & Picard, R. W. (2019). Daytime data and LSTM can forecast tomorrow's stress, health, and happiness. *2019 41st Annual International Conference of the IEEE*

Engineering in Medicine and Biology Society (EMBC), 2186–2190. https://doi.org/10.1109/EMBC.2019.8856862

Umematsu, T., Sano, A., Taylor, S., & Picard, R. W. (2019). Improving students' daily life stress forecasting using LSTM neural networks. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 1–4. https://doi.org/10.1109/BHI.2019.8834624

International Telecommunication Union. (2022). *Measuring digital development: Facts and Figures 2022*. Repéré le 6 décembre 2023 à https://www.itu.int/itu-d/reports/statistics/facts-figures-2022/

Vaizman, Y., Ellis, K., & Lanckriet, G. (2017). Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing*, *16*(4), 62–74. https://doi.org/10.1109/MPRV.2017.3971131

Vaizman, Y. (2018). *Behavioral context recognition in the wild* (ProQuest ID: Vaizman_ucsd_0033D_17157; Merritt ID: ark:/13030/m5479744) [thèse de doctorat, University of California]. Repéré le 7 septembre 2022 à https://escholarship.org/uc/item/200910xx

Van Den Goorbergh, R., Van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, *29*(9), 1525–1534. https://doi.org/10.1093/jamia/ocac093

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature*, *621*(7980), 672–675. https://doi.org/10.1038/d41586-023-02980-0

Vapnik, V., & Chervonenkis, A. (2015). *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. Dans V. Vvok, H. Papadopoulos, & A. Gammerman, *Measures of Complexity* (pp. 11-30). Springer, Cham. https://doi.org/10.1007/978-3-319-21852-6_3

Varoquaux, G. (2017). Cross-validation failure: Small sample sizes lead to large error bars https://doi.org/10.48550/arXiv.1706.07581

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, *5*, 48. https://doi.org/10.1038/s41746-022-00592-y

Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. Dans O. Colliot (Ed.), *Machine learning for brain disorders* (Vol. 197, pp. 601–630). Springer US. https://doi.org/10.1007/978-1-0716-3195-9_20

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Dans I. Guyon, U. Von Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett. *Advances in neural information processing systems* (Vol. 30, pp. 3146–3154). Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, R., Aung, M. S. H., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E. A., Tseng, V. W. S., & Ben-Zeev, D. (2016). CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 886–897. https://doi.org/10.1145/2971648.2971740

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14. https://doi.org/10.1145/2632048.2632054

Wang, R., Harari, G., Hao, P., Zhou, X., & Campbell, A. T. (2015). SmartGPA: How smartphones can assess and predict academic performance of college students. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, 295–306. https://doi.org/10.1145/2750858.2804251

Wang, R., Wang, W., Aung, M. S. H., Ben-Zeev, D., Brian, R., Campbell, A. T., Choudhury, T., Hauser, M., Kane, J., Scherer, E. A., & Walsh, M. (2017). Predicting symptom trajectories of schizophrenia using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–24. https://doi.org/10.1145/3130976

Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–26. https://doi.org/10.1145/3191775

Wikipedia contributors. (n.d.). Thomas R. Insel. Dans *Wikipedia, The Free Encyclopedia*. Repéré le 6 décembre 2023 à https://en.wikipedia.org/wiki/Thomas_R._Insel

Wikipedia contributors. (n.d.). Smartphone. Dans *Wikipedia, the Free Encyclopedia*. Repéré le 6 décembre 2023 à https://en.wikipedia.org/wiki/Smartphone

Wright, A. G. C., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 49–74. https://doi.org/10.1146/annurev-clinpsy-102419-125032

Zhou, J., Lamichhane, B., Ben-Zeev, D., Campbell, A., & Sano, A. (2022). Predicting psychotic relapse in schizophrenia with mobile sensor data: Routine cluster analysis. *JMIR mHealth and uHealth*, *10*(4), e31006. https://doi.org/10.2196/31006

Zimmerman, M., & McGlinchey, J. B. (2008). Why don't psychiatrists use scales to measure outcome when treating depressed patients? *Journal of Clinical Psychiatry*, *69*(12), 1916–1919. https://doi.org/10.4088/jcp.v69n1209

Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., Langenecker, S. A., McInnis, M., Ajilore, O., Nelson, P. C., Ryan, K., & Leow, A. (2018). Predicting mood disturbance severity with mobile phone keystroke metadata: A BiAffect digital phenotyping study. *Journal of Medical Internet Research*, *20*(7), e241. https://doi.org/10.2196/jmir.9775