Apprentissage profond sous contraintes biomédicales pour la prédiction de la glycémie future de patients diabétiques

Maxime de Bois

université
PARIS-SACLAY

Apprentissage profond sous contraintes biomédicales pour la prédiction de la glycémie future de patients diabétiques

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : informatique
Unité de recherche : université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 18 décembre 2020, par

Maxime De Bois

Composition du jury

Anne Vilnat Professeur, CNRS-LIMSI	Présidente
Mohamed Chetouani Professeur, ISIR	Rapporteur & Examinateur
Nicole Vincent Professeure, LIPADE	Rapporteur & Examinatrice
Stéphane Gazut Ingénieur de recherche, CEA	Examinateur
Jean-Daniel Zucker Professeur, IRD, UMMISCO	Examinateur
Mehdi Ammi Professeur, université Paris 8	Directeur de thèse
Mounîm A. El Yacoubi Professeur, CNRS-SAMOVAR	Co-directeur de thèse

Thèse de doctorat

NNT : 2020UPASG065

A notre société tout entière, pour un monde meilleur.

Remerciements

Les travaux effectués tout au long de ce doctorat n'auraient pas été rendus possible sans l'aide et le support de mon entourage personnel et professionnel. Ainsi, avant toute chose, je tiens à remercier ces personnes.

En premier lieu, mes remerciements vont à mes tuteurs, Mehdi et Mounîm. Merci de m'avoir initié au monde de la recherche ainsi que de m'avoir accompagné intellectuellement et humainement pendant ces trois années. Travailler avec vous a été très agréable et j'espère que nous aurons l'opportunité de continuer notre collaboration dans le futur !

Je remercie l'ensemble des membres du jury, Anne Vilnat en sa qualité de présidente, les rapporteurs Mohamed Chetouani et Nicole Vincent, ainsi que les examinateurs Jean-Daniel Zucker et Stéphane Gazut. Merci d'avoir lu mes travaux avec l'intérêt que vous avez témoigné pendant la soutenance ainsi que pour les échanges fort intéressants que nous avons pu avoir.

Cette thèse est le fruit de la collaboration avec le réseau de santé pour personnes diabétiques, Revesdiab. Je remercie l'ensemble du bureau, et en particulier Sylvie Joannidis, ainsi que les infirmières coordinatrices Carole Huberson et Corinne Nkondjock pour m'avoir accompagné dans la construction et le déroulement de la campagne de collecte de données. Enfin, merci à tous les membres de l'association qui y ont participé. Vous rencontrer a été essentiel dans la prise en compte des besoins des patients dans les différents projets menés pendant le doctorat.

Enfin, je remercie toutes les personnes que j'ai pu croiser de près ou de loin lors de mes séjours au LIMSI et à Télécom SudParis. Vous êtes beaucoup à avoir croisé mon chemin et vous côtoyer au quotidien a été une source de plaisir constant.

Sur une note plus personnelle, je tiens à remercier tout d'abord mes parents qui m'ont insufflé l'amour des sciences, et en particulier celui de l'informatique. Merci à vous deux pour m'avoir écouté, soutenu et conseillé pendant toutes ces années. Merci aussi à toi Aurélien pour être le meilleur des frères ! J'ai de la chance de pouvoir partager autant de choses avec toi.

Je tiens aussi à remercier tous mes amis pour tous ces moments à jouer aux jeux vidéos ou Donjons & Dragons, à faire de la musique, mais aussi à faire la fête. Vous êtes nombreux et c'est un plaisir de vous avoir dans ma vie.

Je te remercie, Audrey ma princesse. Merci pour ta gentillesse et ton amour sans égal, pour tous ces moments passés ensemble, ainsi que pour m'avoir accompagné dans cette aventure. Si celle-ci s'arrête aujourd'hui, la nôtre

ne fait que commencer !

Enfin, je remercie mon chat Ronron, mon compagnon de thèse au quotidien et particulièrement pendant cette dernière année de pandémie avec la rédaction du manuscrit dans la Creuse. Ses bêtises, ses siestes devant mon clavier ainsi que son besoin constant d'attention ont été déterminants dans ma réussite.

Publications

Journaux internationaux

1. **M. De Bois**, M. A. El Yacoubi, M. Ammi. Enhancing the Interpretability of Deep Models in Heathcare Through Attention : Application to Glucose Forecasting for Diabetic People. *arXiv preprint arXiv:2009.03732, accepté à International Journal of Pattern Recognition and Artificial Intelligence*, 2020 [37].
2. **M. De Bois**, M. A. El Yacoubi, M. Ammi, . Integration of Clinical Criteria into the Training of Deep Models : Application to Glucose Prediction for Diabetic People. *arXiv preprint arXiv:2009.10514*, 2020 [30].
3. **M. De Bois**, M. A. El Yacoubi, M. Ammi. M. Ammi, M. Adversarial multi-source transfer learning in healthcare : Application to glucose prediction for diabetic people. *Computer Methods and Programs in Biomedicine*, 199 : 105874–105874, 2020 [35].
4. **M. De Bois**, M. A. El Yacoubi, M. Ammi. GLYFE : Review and Benchmark of Personalized Glucose Predictive Models in Type-1 Diabetes. *arXiv preprint arXiv:2006.15946*, 2020 [34].

Conférences internationales

5. **M. De Bois**, M. A. El Yacoubi, M. Ammi. Interpreting deep glucose predictive models for diabetic people using RETAIN. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 685–694.Springer, 2020 [36].
6. **M. De Bois**, M. A. El Yacoubi, M. Ammi, M. Study of short-term personalized glucose predictive models on type-1 diabetic children. In *2019 International Joint Conference on Neural Networks* (IJCNN), pages 1–8. IEEE, 2019 [33].
7. **M. De Bois**, M. A. El Yacoubi, M. Ammi, M. Prediction-coherent lstm-based recurrent neural network for safer glucose predictions in diabetic people. In *International Conference on Neural Information Processing*, pages 510–521. Springer, 2019 [32].

8. **M. De Bois**, M. A. El Yacoubi, M. Ammi. Model fusion to enhance the clinical acceptability of long-termglucose predictions. In *BIBE 2019 : 19th International Conference on Bioinformatics and Bioengineering*, pages 258–264. IEEE Computer Society, 2019 [31].

Posters

9. **M. De Bois**, M. A. El Yacoubi, M. Ammi. Study of Short-Term Personalized Glucose Predictive Models in Diabetes. Journée IMT IA & Santé, 2018.

Logiciels Open Source

10. **M. De Bois.** Multi-source adversarial transfer learning in glucose prediction for type-2 diabetic patients, 2020. URL https://github.com/dotXem/GlucosePredictionATL. DOI : 10.5281/zenodo.3699846 [29].
11. **M. De Bois.** Interpreting deep glucose predictive models through the retain architecture, 2020. URL https://github.com/dot DOI : 10.5281/zenodo.3951702 [28].
12. **M. De Bois.** Integration of clinical criteria into the training of deep models : Application to glucose prediction-for diabetic people, 2020. URL https://github.com/dotXem/DeepClinicalGlucosePrediction. DOI : 10.5281/zenodo.3904234 [27].
13. **M. De Bois.** CG-EGA Python Implementation, 2019. URL https://github.com/dotXem/CG-EGA. DOI : 10.5281/zenodo.3459590 [26].
14. **M. De Bois.** GLYFE, 2019. URL https://github.com/dotXem/GLYFE. DOI : 10.5281/zenodo.3234605 [25].

Table des matières

Table des figures

Liste des tableaux

Acronymes

AP *Accurate Prediction*. 50, 88, 90–92, 106, 107, 116, 117, 122–124, 126, 128, 129, 165, 174, 175

APAC Amélioration Progressive de l'Acceptabilité Clinique. 107, 117–119, 121, 122, 124, 126, 128, 129, 175

AR processus autorégressif. 18, 25, 34, 35, 50, 84, 85, 88, 90, 92

ARIMA processus autorégressif intégré avec moyenne mobile. 25, 35, 96

ARIMAX processus autorégressif intégré avec moyenne mobile et données exogènes. 25, 35, 85

ARMA processus autorégressif avec moyenne mobile. 25, 34, 35

ARMAX processus autorégressif avec moyenne mobile et données exogènes. 25, 34, 35

ARX processus autorégressif avec données exogènes. 25, 34, 35, 42, 84, 85, 88, 90, 92, 95, 115

ATL apprentissage par transfert adverse, *Adversarial Transfer Learning*. 139

BE *Benign Error*. 50, 88, 107, 109, 116, 165, 174

CG-EGA *Continuous Glucose-Error Grid Analysis*. 48–50, 87–90, 92, 93, 98, 105–107, 112, 117, 122, 128, 129, 139, 163, 165, 170, 174

CGM capteur de glycémie en continu, *Continuous Glucose Monitoring device*. 16–18

cMSE erreur quadratique moyenne cohérente. 106, 107, 113, 114, 121–123, 125, 126, 129, 174

CNN réseau de neurones convolutifs, *Convolutional Neural Networks*. 25, 41, 43, 55

DT arbre de décision, *Decision Tree*. 25, 36, 160–163, 165, 170

ELM *Extreme Learning Machine*. 25, 42, 85, 86, 88, 90, 92

EP *Erroneous Prediction*. 50, 88, 90–92, 106, 107, 109, 111, 116, 117, 123–126, 128, 129, 165, 174

ES lissage exponentiel, *Exponential Smoothing*. 25, 43

ESN *Echo State Network*. 25, 43

ESOD *Energy of the Second Order Differences*. 48

FCN réseau de neurones entièrement convolutifs, *Fully Convolutional Neural Network*. 137, 138, 141–144, 160, 161, 163, 165, 168, 170, 175–177

FFNN réseau de neurones de type passe-en-avant, *Feedforward Neural Network*. 25, 40, 51–53, 85, 86, 88, 90, 92, 93, 95, 96

FSL capteur de glycémie FreeStyle Libre de Abbott. 59, 61, 64, 66, 67

GBM machine à boosting de gradient, *Gradient Boosting Machine*. 25, 37, 51, 52, 96, 160–163, 165, 170

gcMSE erreur glycémique quadratique moyenne cohérente. 107, 115, 116, 119, 121, 123, 125, 126, 128, 129, 174, 175

GE *Grammatical Evolution*. 25, 43, 52, 53

GP processus gaussien, *Gaussian Process*. 25, 39, 85, 88, 90, 92, 93, 168

gRMSE *Glucose Root Mean Squared Error*. 48

KAF filtre adaptatif à noyau, *Kernel Adaptive Filter*. 25, 39

KF filtre de Kalman, *Kalman Filter*. 25, 43

KRR *Kernel Ridge Regression*. 25, 40

LDP perplexité de domaine locale, *Local Domain Perplexity*. 145–147, 149

LR régression linéaire, *Linear Regression*. 25, 43

LSTM réseau de neurones récurrents ou unités *Long Short-Term Memory*. 25, 41–43, 51, 52, 55, 85, 86, 88, 90, 92, 96, 106, 107, 114, 119, 121–123, 126, 128, 129, 132, 149, 153, 160, 161, 163, 165, 168, 170, 174–177

LVX régression par Variables Latentes à données eXogènes. 25, 43

MA moyenne mobile, *Moving Average*. 25, 34, 43, 85

MAPE erreur absolue moyenne en pourcentages, *Mean Absolute Percentage Error*. 45–48, 87, 89–94, 122

MARD deviation absolue relative moyenne, *Mean Absolute Relative Deviation*. 45–48, 87

MASE *Mean Absolute Scaled Error*. 118–122, 124, 126, 128

MSE erreur quadratique moyenne, *Mean Squared Error*. 40, 86, 106, 113–115, 118, 121, 129, 138

P-EGA *Point-Error Grid Analysis*. 48, 49, 88, 106, 107, 109, 111, 115, 116, 124, 129, 139, 140, 174

PH horizon de prédiction, *Prediction Horizon*. 33, 34, 40, 44

Phys modèle physiologique. 25, 43

Poly régression temporelle polynomiale. 84, 85, 88, 90, 92, 93, 98

R-EGA *Rate-Error Grid Analysis*. 49, 106, 107, 109, 111, 115, 116, 129, 139, 174

Ref modèle de référence naïf. 25, 43, 84, 90–93, 96

RF forêt aléatoire, *Random Forest*. 25, 37, 160–163, 165, 170

RGPD règlement général sur la protection des données. 61, 65, 66

RMSE racine carrée de l'erreur quadratique moyenne, *Root Mean Squared Error*. 45–48, 52, 87, 89–93, 95, 96, 105, 122–124, 139, 140, 163, 165, 170, 174

RNN réseau de neurones récurrents, *Recurrent Neural Network*. 25, 41

SARIMA processus autorégressif intégré saisonnier avec moyenne mobile. 25, 35

SARIMAX processus autorégressif intégré saisonnier avec moyenne mobile et données exogènes. 25, 35

SOM carte auto adaptative, *Self-Organizing Map*. 25, 43

SVM machine à vecteurs de support, *Support Vector Machine*. 38, 53

SVR SVM pour les tâches de régression. 25, 38, 39, 42, 52, 53, 55, 85, 88, 90, 92, 93, 95, 97, 121, 122, 126, 137, 141–144, 149, 168, 174

TG gain temporel, *Time Gain*. 45, 47, 48, 87, 89, 90, 92, 93, 98, 119, 120, 139

TL apprentissage par transfert, *Transfer Learning*. 138, 139, 141

WFNN *neuro-fuzzy Neural Network with Wavelets activation Function*. 25

1 | Introduction

Sommaire

1.1 Contexte

1.1.1 Enjeux du diabète

Le *diabete mellitus*, ou diabète sucré, aussi appelé simplement diabète est l'une des maladies majeures de notre monde moderne. En 2019, la Fédération Internationale pour le Diabète a estimé que 463 millions d'adultes dans le monde sont diabétiques et que la maladie est directement à l'origine de 4.2 millions de morts [50]. La prévalence de la maladie dans le monde est prédite d'évoluer de 9.3% à 10.9% (700 millions) d'ici 2045. En France, sa prévalence est estimée à 7.6%, soit près de 3.5 millions de personnes.

Le diabète se caractérise par un trouble de la régulation naturelle du taux de sucre dans le sang (glycémie). Cette dernière, dont la Figure 1.1 schématise le procédé chez une personne saine, met en jeu deux organes importants du corps humain : le pancréas et le foie. Lorsque le taux de sucre est trop élevé, suite à un repas par exemple, le pancréas sécrète une hormone, l'insuline. Celle-ci permet le stockage du sucre (ou glucose) sous la forme de glycogène par le foie et les cellules de l'organisme. Cela a pour conséquence de faire baisser la glycémie. À l'inverse, si la glycémie est trop basse, le pancréas sécrète une seconde hormone, le glucagon, qui permet au foie de relâcher dans le sang le sucre qui a été stocké sous la forme de glycogène. L'objectif de la régulation de la

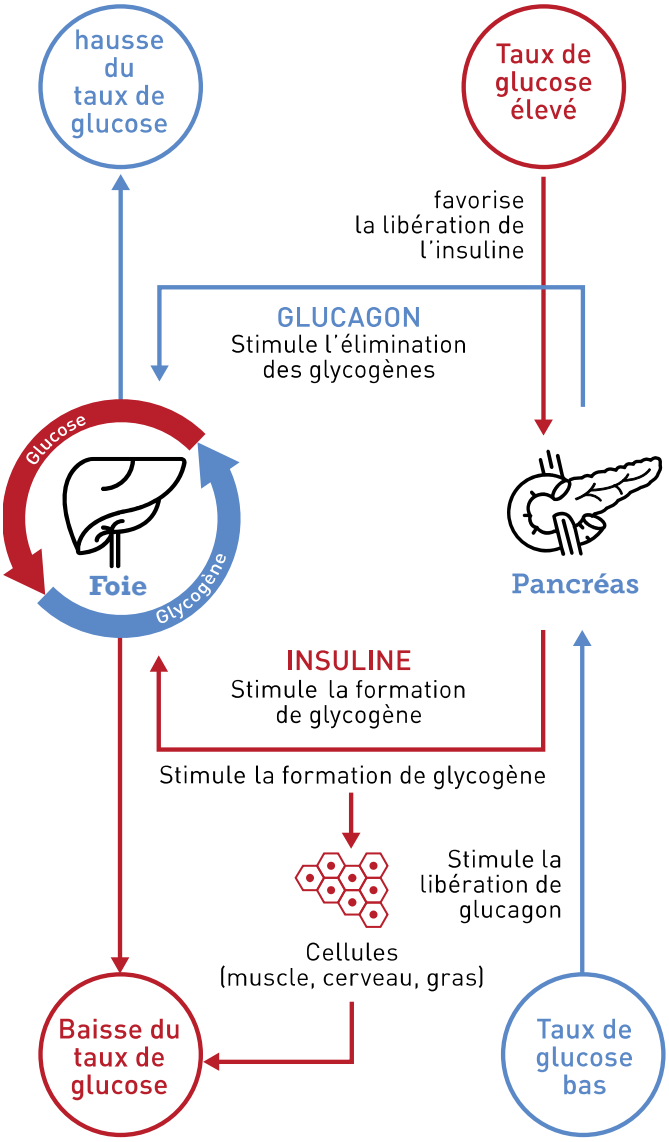


FIGURE 1.1: Régulation de la glycémie dans le corps humain [49].

glycémie est d'atteindre l'homéostasie qui se trouve autour de 90 mg/dL de glucose dans le sang.

Il existe trois types principaux de diabète : le type 1, le type 2, et le diabète gestationnel. Le diabète de type 1, aussi appelé « diabète insulinodépendant » ou « diabète juvénile », est provoqué par une réaction auto-immune détruisant les cellules bêta du pancréas, responsables de la sécrétion d'insuline. Survenant généralement pendant l'enfance, le diabète de type 1 représente entre 7 et 12% de la population diabétique française. Quant à lui, le diabète de type 2 est le plus fréquent et représente entre 87 et 91% de la population diabétique française et environ 90% dans le monde. Il se caractérise par une résistance accrue et progressive des cellules de l'organisme à l'insuline. Pour pallier cette réduction de l'action de l'insuline, le pancréas en sécrète davantage. Cependant, à terme, le pancréas n'est plus capable de produire suffisamment d'insuline pour combattre la résistance des cellules. Enfin, le diabète gestationnel, représentant entre 1 et 3% de la population diabétique française, se caractérise simplement

par un taux de sucre élevé pendant la grossesse. Bien que celui-ci disparaisse généralement après la grossesse, il peut aussi se transformer en diabète de type 2.

Que ce soit à cause de la non-production d'insuline ou la résistance accrue à son action, tous les types de diabète sont sujets à une perturbation de la régulation de la glycémie. Cette dérégulation peut avoir à la fois des conséquences à court terme et des conséquences à long terme. Lorsque la glycémie est trop basse (en dessous de 70 mg/dL), la personne diabétique en état d'hypoglycémie est sujette à des étourdissements, risque de tomber dans le coma, voire même de mourir (conséquences courts-termes). À l'inverse, si la glycémie est trop élevée (supérieure à 180 mg/dL), la personne diabétique en état d'hyperglycémie est sujette au développement d'une rétinopathie ou d'un ulcère des membres inférieurs (conséquences longs-termes).

1.1.2 Solutions pour la gestion du diabète

Afin d'éviter ces conséquences à la fois à court terme et à long terme induites par l'hypo ou l'hyperglycémie, les personnes diabétiques doivent elles-mêmes réussir à réguler leur glycémie. Nous pouvons identifier trois axes d'action permettant de faciliter sa réalisation.

Le premier est la modification du style de vie des personnes diabétiques. Cela passe par avoir une alimentation plus saine et moins calorique pour les personnes en surpoids (très courant chez les personnes diabétiques de type 2), à remplacer les graisses saturées (e.g., beurre) par des graisses insaturées (e.g., huile d'olive), à éviter le tabac ainsi que la consommation excessive d'alcool. La pratique d'une activité physique régulière, combinant exercice d'endurance et de résistance, est conseillée. Afin d'aider les personnes dans l'adoption d'un meilleur rythme de vie, de nombreux programmes d'éducation thérapeutique sont proposés par les hôpitaux ainsi que les associations.

Le second axe d'action visant à aider les personnes diabétiques au quotidien est l'utilisation de médicaments. Les personnes diabétiques de type 1, ne produisant pas d'insuline, ont besoin d'un apport en insuline au quotidien. Celui-ci peut se faire au moyen d'injection par seringue, stylo ou pompe (appareil portable distribuant automatiquement l'insuline). Nous pouvons distinguer plusieurs types d'insuline dont les plus courants sont l'insuline à action rapide et l'insuline à action lente. L'insuline à action lente correspond à l'insuline dont le corps a besoin pour fonctionner équilibrant la glycémie tout au long de la journée. Elle est aussi appelée insuline *basale* lorsqu'elle est administrée par une pompe à insuline. Quant à elle, l'insuline à action rapide, aussi appelée insuline en *bolus*, correspond à l'insuline dont le corps a besoin pour couvrir les apports en glucides liés aux repas. Pour les personnes diabétiques de type 2, certains médicaments comme la metformine ou le gliclazide peuvent être utilisés pour améliorer la réponse naturelle du corps aux aliments ingérés, réduisant le taux de sucre après un repas.

Enfin, des outils technologiques peuvent être utilisés. De ce point de vue, l'avancée la plus marquante ces dernières années est la commercialisation de capteurs de glycémie en continu, *Continuous Glucose Monitoring devices* (CGM). La Figure 1.2 représente le CGM FreeStyle Libre de Abbott. Ces capteurs prennent la forme d'un patch à coller sur la peau et permettent de mesurer la glycémie de la personne à intervalle de 5 à 15 minutes. Cette

FIGURE 1.2: Capteur de glycémie en continu FreeStyle Libre de Abbott.

mesure se fait dans le liquide interstitiel, liquide présent entre les cellules de l'organisme, et non directement dans le sang. Le taux de sucre dans le liquide interstitiel se caractérise par un retard d'environ 12.5 minutes sur le taux de sucre dans le sang [87]. Bien que donnant une mesure moins représentative de l'état actuel de la glycémie du patient, les CGM sont bien plus pratiques que l'action de se piquer le bout du doigt à travers l'utilisation de lancettes. De plus, pour un patient, connaître sa glycémie en continu permet à la fois d'opérer un contrôle plus fin de celle-ci, mais aussi de lui apporter des connaissances sur le fonctionnement de sa maladie. En effet, l'ensemble de la population diabétique est caractérisé par sa très grande variabilité. Cela implique que la régulation de la glycémie de chaque personne diabétique possède ses propres spécificités, et que chaque traitement doit lui être personnalisé. Par ailleurs, nous pouvons noter l'émergence d'applications smartphone visant à aider les personnes diabétiques au quotidien. Par exemple, c'est le cas de l'application Gluci-Check permettant de compter les glucides présents dans son repas [54] ou de l'application de coaching pour personnes diabétiques mySugr [117].

Du point de vue de la recherche, de nombreux efforts ont été faits ces dernières années afin de construire des modèles prédisant la glycémie future de personnes diabétiques [162]. Avoir la connaissance de la glycémie dans 30 minutes ou une heure permet au patient d'anticiper les hypoglycémies ou hyperglycémies à venir. Un tel outil pourrait aussi participer à l'éducation thérapeutique du patient en mettant en évidence les éléments importants dans la régulation de sa glycémie. Enfin, ces prédictions pourraient être incluses dans le développement d'un pancréas artificiel [19]. Ce dernier, prenant la forme d'une boucle prédictive de contrôle, permettrait de réguler automatiquement la glycémie de la personne diabétique à l'instar d'un pancréas naturel.

1.2 Problématique et objectifs

En 2007, Sparacino *et al.* montrent, à travers un processus autorégressif (AR) simple, qu'il est possible de prédire la glycémie des patients diabétiques 30 minutes dans le futur, leur permettant ainsi d'anticiper d'éventuelles hypoglycémies. Depuis lors, grâce notamment à la démocratisation de l'utilisation de CGM, de nombreux chercheurs se sont intéressés à la tâche de la prédiction de la glycémie [119]. Ces recherches ont mis en évidence la grande complexité de la tâche. Tout d'abord, de nombreux facteurs, comme les prises d'insuline, les ingestions de glucides lors des repas, mais aussi l'activité physique [73], le sommeil [82] ou même l'humeur [138], influencent la régulation de la glycémie. Aussi, les modèles prédictifs doivent tenir compte de la grande variabilité à la fois individuelle, mais aussi de celle de la population diabétique en général. Au fil des années, de nombreux modèles différents ont été utilisés. En particulier, de nombreuses recherches ont exploré ces dernières années l'utilisation prometteuse de l'apprentissage profond pour la prédiction de la glycémie [114, 115, 168, 169, 95, 96].

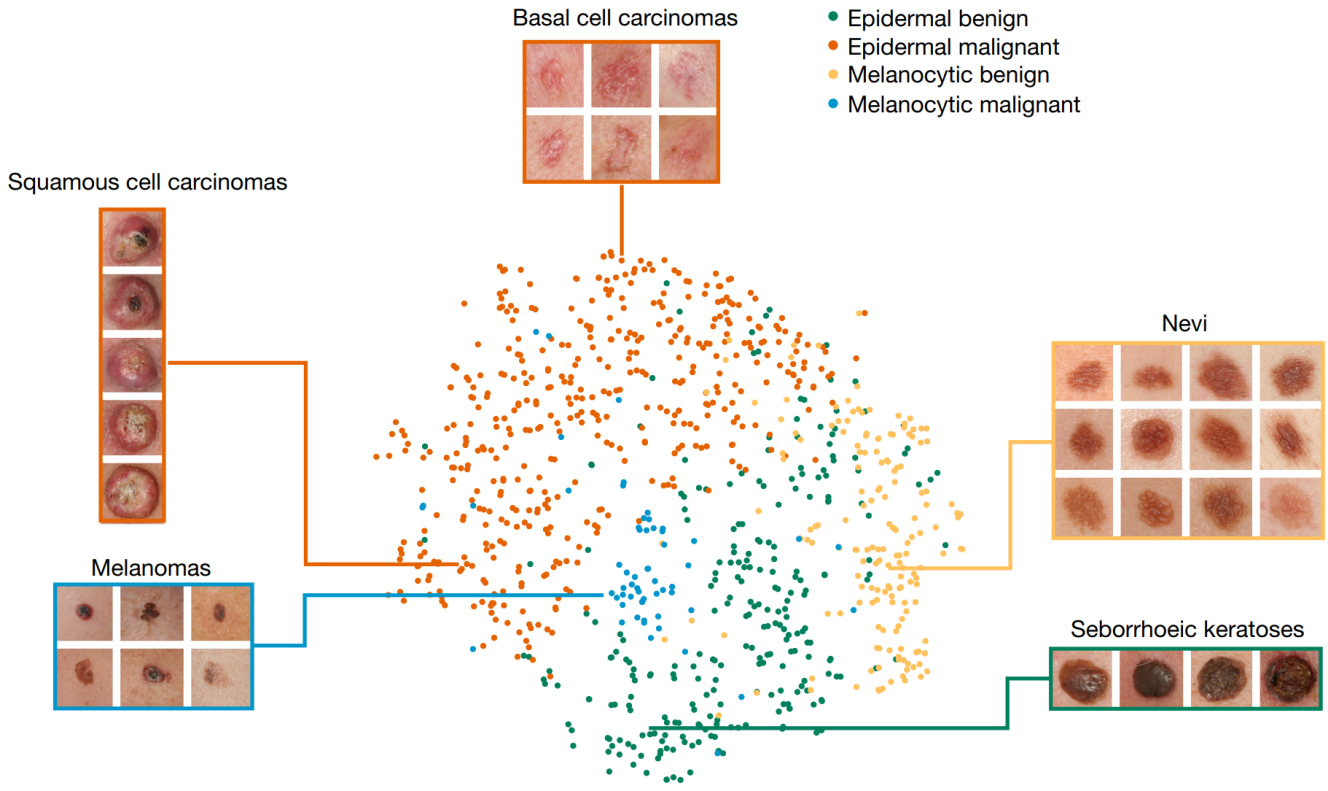


FIGURE 1.3: Représentation en 2 dimensions des connaissances apprises par un réseau de neurones convolutifs sur des images de lésions de la peau pour la classification du cancer de la peau [47]. Nous pouvons voir que les différents types de maladie de la peau sont regroupés dans des régions spécifiques, indiquant que le modèle profond réussit à les distinguer.

L'apprentissage profond est basé sur l'usage de réseaux de neurones artificiels cherchant à imiter le fonctionnement du cerveau humain. L'accès grandissant à des quantités de données de plus en plus importantes, ainsi que l'amélioration des puissances de calcul lui a permis de réaliser de grand progrès dans le domaine de la recon-

naissance d'images [136] ou le traitement du langage naturel [41]. Ces dernières années ont vu l'émergence de plusieurs succès permis par l'apprentissage profond dans le domaine de la santé. Par exemple, en 2016, Gulshan *et al.* ont montré que l'utilisation de l'apprentissage profond permet de détecter automatiquement à plus de 97.5% une rétinopathie diabétique à partir de rétinophotographies [68]. Toujours en 2016, Wang *et al.* [152] ont proposé un modèle basé sur l'apprentissage profond capable de détecter efficacement un cancer du sein métastatique [1] [152]. De plus, son utilisation comme outil a permis aux médecins de réduire de 85% le taux d'erreurs de détection. Enfin, en 2017, Estava *et al.* ont montré que l'apprentissage profond est capable d'égaler, voire de surpasser, des dermatologues dans la classification de différents cancers de la peau [47]. La Figure 1.3 donne une représentation graphique de leurs résultats.

Cependant l'usage pratique de l'apprentissage profond dans la santé, et en particulier pour la prédiction de la glycémie, fait face à de nombreux challenges [16, 153] :

— **Quantité des données** : Pour être performants dans leur tâche, les modèles profonds ont besoin de très grandes quantités de données d'apprentissage. Aujourd'hui la plupart des applications qui ont du succès doivent leur réussite aux grandes quantités de données obtenues au fil des années (e.g., images de cancer). Cependant, de manière générale, obtenir des données de santé en quantité suffisante reste difficile. Cela est dû à leur coût d'obtention nécessitant souvent des connaissances expertes (e.g., labellisation d'images de cancer), à la rareté de la donnée d'intérêt (e.g., détection de maladies rares), ainsi qu'à leur nature sensible compliquant leur partage entre différentes structures de santé. Dans le prédiction de la glycémie, le besoin d'avoir des modèles personnalisés au patient réduit considérablement les données disponibles pour l'entraînement des modèles. Ces données doivent avoir été récoltées en quantité suffisante pour chaque patient.

— **Qualité des données** : Les données récoltées doivent non seulement être en grande quantité, mais aussi être de qualité. Une donnée de mauvaise qualité peut être caractérisée par un bruit important au sein du signal (e.g., données capteur), ou être inconstante dans son format (e.g., notes non structurées écrites à la main dans un dossier de santé électronique). Comme pour l'humain, ces données sont difficilement exploitables par les modèles d'apprentissage profond. Ces données peuvent également être très hétérogènes, décrivant des phénotypes similaires, mais pas identiques, ou provenant de matériels ou protocoles expérimentaux différents. Pour la prédiction de la glycémie, la qualité des données est variable. Par exemple, les capteurs de glycémie peuvent être sujets à des erreurs de mesures. Aussi, les patients peuvent se tromper lors de l'estimation des glucides ingérés pendant les repas.

— **Interopérabilité des modèles** : Les modèles ont le risque de ne fonctionner que dans le cadre de l'étude dans lequel ils ont été conçus. Par exemple, un modèle détectant les cancers de la peau, entraîné sur une population occidentale majoritairement de couleur de peau blanche, ne fonctionnera pas s'il est utilisé sur des populations différentes. Ce problème peut aussi survenir si les modèles sont appliqués sur des données

1. Un cancer est dit métastatique lorsqu'il s'est propagé à un ou plusieurs autres endroits du corps.

récoltées avec un matériel expérimental différent. L'interopérabilité des modèles prédictifs de glycémie est presque impossible à obtenir à cause de la très grande variabilité de la population diabétique. À cause de cela, un modèle entraîné sur un patient ne sera pas précis, et donc dangereux, si utilisé sur un autre patient. Cette non-interopérabilité est accrue si les patients ont un diabète très différent (différent type ou stade de la maladie) ou utilisent des capteurs de glycémie différents.

— **Interprétabilité des modèles** : Bien que potentiellement très performants, les modèles issus de l'apprentissage profond sont très complexes. Cette complexité réduit considérablement leur interprétabilité, ce qui leur doit le surnom de « boîtes noires ». L'interprétabilité des modèles dans la santé est particulièrement importante, car elle permet d'avoir confiance dans les prédictions qui sont faites. Si le modèle fait une prédiction surprenante pour le médecin, savoir pourquoi elle a été faite lui permet de prendre une décision finale plus éclairée. L'interprétabilité des modèles est aussi importante en ce sens qu'elle peut permettre de faire des découvertes en identifiant, par exemple, des liens non connus entre les données. Dans la prédiction de la glycémie, la confiance dans la prédiction est importante, car le patient s'en sert pour prendre des décisions visant à mieux réguler sa glycémie. Aussi, pouvoir expliquer les prédictions du modèle permettrait au patient de comprendre mieux le fonctionnement très personnel de sa maladie.

Dans cette thèse, nous étudions, à la lueur de la tâche de la prédiction de la glycémie future de personnes diabétiques, comment combattre ces différentes limitations de l'utilisation de l'apprentissage profond dans la santé. En particulier, l'objectif est de proposer de nouveaux modèles qui soient à la fois précis, sans danger pour le patient et interprétables.

1.3 Contributions et démarche

1.3.1 Contributions

Les contributions de cette thèse sont les suivantes :

— La plupart des études s'intéressant à la prédiction de la glycémie se focalisent uniquement sur les personnes diabétiques de type 1. Les personnes diabétiques de type 2 représentant près de 90% de la population diabétique, les études traitant de la prédiction de la glycémie gagneraient à leur être étendues. Dans cette optique, nous avons conduit une campagne de collecte de données sur des personnes diabétiques de type 2 en collaboration avec l'association Revesdiab. L'ensemble des travaux présentés dans cette thèse utilise cet ensemble de données combiné à deux autres jeux similaires provenant de personnes, virtuelles et réelles, diabétiques de type 1. Cela nous permet de montrer, en particulier, la validité de nos études sur les deux types de diabète principaux.

— Nous procédons à une analyse approfondie de l'état de l'art de la prédiction de la glycémie de personnes

diabétiques. En particulier, nous construisons une base open source de résultats de référence de modèles prédictifs de glycémie. Celle-ci répond au besoin de littérature de posséder des procédés de traitement standardisés afin de comparer les performances des modèles prédictifs. Nous montrons tout d'abord que la tâche de la prédiction de la glycémie favorise davantage des modèles complexes relevant de l'apprentissage automatique ou profond. Toutefois, nous mettons en évidence la difficulté générale qu'ont les modèles à effectuer les prédictions à cause de manque général de données d'entraînement. Cette difficulté est en particulier exacerbée pour les modèles profonds, basés sur des réseaux de neurones. Par ailleurs, nous mettons en évidence le challenge de faire des prédictions à la fois précises, utiles pour le patient, et sans danger pour celui-ci. En effet, une bonne précision statistique ne garantit pas l'acceptabilité clinique des modèles, ceux-ci étant en pratique particulièrement dangereux en région d'hypoglycémie.

— Dans l'optique d'améliorer l'acceptabilité clinique des modèles, nous proposons d'inclure des contraintes liées à celle-ci dans lors de l'entraînement des modèles profonds. En particulier, nous proposons de nouvelles fonctions de coût pour l'apprentissage des modèles. Celles-ci permettent d'améliorer la cohérence des prédictions successives ainsi que de se focaliser davantage sur les régions glycémiques dangereuses (e.g., hypoglycémie). Enfin, nous explorons leur utilisation pratique en proposant une méthodologie permettant d'obtenir le compromis optimal entre précision et acceptabilité clinique des prédictions, selon des critères cliniques a priori.

— Pour améliorer les performances des modèles profonds, nous proposons d'explorer l'apprentissage par transfert afin d'utiliser plus efficacement les données que nous avons à disposition. L'apprentissage par transfert vise à réutiliser les connaissances apprises sur plusieurs patients diabétiques lors de l'entraînement d'un modèle sur un nouveau patient. Afin de faciliter ce transfert de connaissance, nous proposons le cadre de l'apprentissage par transfert multi-sources adverse permettant d'extraire des connaissances plus générales. Nous montrons que ces connaissances sont plus facilement transférables à un nouveau patient, donnant lieu à une amélioration significative des performances. De plus, nous montrons qu'il est possible de transférer des connaissances entre patients diabétiques de différents types, ainsi qu'entre patients virtuels et réels.

— La complexité accrue des modèles prédictifs de glycémie, et en particulier des modèles profonds, entraîne une baisse en interprétabilité des prédictions. Nous nous intéressons à l'amélioration de l'interprétabilité des modèles profonds à travers le principe d'attention. En particulier, nous proposons un modèle profond et interprétable pour la prédiction de la glycémie. Celui-ci implémente un double mécanisme d'attention lui permettant d'estimer la contribution de chaque variable en entrée à la prédiction finale. Nous montrons empiriquement l'intérêt d'un tel modèle pour la prédiction de glycémie en analysant le comportement du modèle dans le calcul de ses prédictions.

— Enfin, le code lié à l'ensemble de la chaîne de traitement des données (prétraitement des données, entraîne-

ment des modèles, évaluations des résultats) a été mis à disposition en open source sur GitHub. Cela permet en particulier de faciliter l'utilisation de nos travaux par la communauté scientifique.

1.3.2 Organisation de la thèse

La thèse est organisée comme suit :

— **Chapitre 2** : Dans un premier temps, nous présentons l'état de l'art de la prédiction de la glycémie. Celui-ci se focalise sur les modèles étudiés et leur évaluation, mais aussi sur les données utilisées ainsi que leurs étapes de prétraitement. Son analyse nous permet de soulever plusieurs problématiques qui sont traitées dans cette thèse.

— **Chapitre 3** : Nous introduisons les trois jeux de données expérimentales utilisés tout au long des travaux présentés dans cette thèse. En particulier, nous présentons la base de données IDIAB que nous avons créée en collaboration avec l'association Revesdiab. Puis nous détaillons les étapes de prétraitement des données avant leur utilisation pour l'entraînement et l'évaluation des modèles prédictifs.

— **Chapitre 4** : Nous approfondissons la revue de littérature du Chapitre 2 par la création d'une base de résultats de référence pour la prédiction de la glycémie de personnes diabétiques. Nous alimentons cette base de référence par les résultats de 9 modèles prédictifs utilisés dans la littérature.

— **Chapitre 5** : Dans ce chapitre, nous présentons de nouvelles fonctions de coût pour l'apprentissage des modèles profond dans le but améliorer l'acceptabilité clinique des modèles. Nous explorons ensuite leur utilisation pratique.

— **Chapitre 6** : Ce chapitre s'intéresse à l'utilisation de l'apprentissage par transfert afin d'améliorer les performances des modèles profonds prédictifs de glycémie en utilisant plus efficacement les données à disposition. En particulier, nous étudions la transférabilité des modèles en faisant varier les patients utilisés comme sources du transfert.

— **Chapitre 7** : Ce chapitre a pour objectif d'améliorer l'interprétabilité des modèles prédictifs de glycémie. Puis, nous explorons dans ce chapitre l'utilisation pratique d'un modèle profond et interprétable.

— **Chapitre 8** : Pour conclure, nous passons en revue l'ensemble des études présentées dans cette thèse en synthétisant les résultats et les principaux enseignements dégagés. Nous discutons également les limites de ces études et nous explorons quelques perspectives pour de futurs travaux de recherche.

2 | État de l'art

Sommaire

2.1 Introduction

Les premiers travaux traitant de la prédiction de la glycémie future de patients diabétiques datent de la fin des années 1990 [134]. Depuis, l'intérêt du domaine n'a cessé de grandir auprès de la communauté scientifique. L'objectif de ce chapitre est de dresser un état de l'art de la prédiction de la glycémie, dans toute sa diversité,

ses avancées et ses limitations. Cet état de l'art a été construit à partir d'une base de données que nous avons constituée en début de doctorat, puis actualisée en continu. En nous appuyant sur les revues de littérature faites par Oviedo *et al.* en 2017 [119], puis Woldaregay *et al.* en 2019 [156], nous recensons 50 articles publiés entre 2007 et 2020. Afin d'apporter une cohésion entre les articles, nous avons utilisé les restrictions suivantes dans la construction de l'état de l'art :

— Nous nous intéressons à la prédiction de la glycémie future uniquement. Ainsi, nous ne répertorions pas les recherches traitant de la prédiction de la glycémie courante telles que [146, 2]. Ces recherches se caractérisent par la non-utilisation de données de glycémie provenant de acrshortcgm. Leur objectif n'est ainsi pas le même. Tandis que ces recherches ont pour objectif de se passer de acrshortcgm pour l'estimation de la glycémie courante pour améliorer le confort des patients, nous cherchons plutôt à profiter davantage de l'utilisation de ces appareils.

— Un des axes de recherche numérique sur le diabète est le développement d'un pancréas artificiel. Visant à remplacer les fonctions d'un pancréas naturel, celui-ci est modélisé comme une commande prédictive (*model predictive control*). En anticipant les variations futures de glycémie de la personne, l'insuline (bolus ou basale) du patient est modifiée dynamiquement pour réguler la glycémie. Ainsi, cette tâche peut se scinder en deux distinctes : la prédiction de la glycémie et la prise de décision pour la régulation. Nous avons décidé de ne pas inclure les recherches traitant de commandes prédictives pour la régulation de la glycémie, car ces travaux ne se focalisent pas sur la tâche de la prédiction de la glycémie (qui devient alors secondaire) et ne concernent qu'une petite quantité de la population diabétique (type 1 et type 2 très avancé).

No.	Ref.	Auteurs	Année	Modèles	Données											Évaluation									
									Nature							Pré-traitement				Métriques					
					Nom[*]	*Type*[†]	*N. patients*	*N. jours*[‡]	Glucose	Insuline	Glucides	Activité Physique	Temps	Sommeil	Évènements	Humeur	Nettoyage	Augmentation	Descripteurs	*Globale ou perso.*	*PH*	RMSE	MAPE+MARD	r+R^2+Fit	TL+TG	ESOD	Precision/recall	P-EGA+CG-EGA	Autres
1	[5]	Aliberti	2019	LSTM, FFNN, AR, RNN	-	T1	451	3	✓								✓			globale	30,45,60,90	✓		✓	✓				✓
2	[4]	Ben Ali	2018	FFNN, SVR, AR, ELM	-	T1	12	14	✓											personnalisée	15,30,45,60	✓	✓	✓				✓	✓
3	[10]	Bertachi	2018	FFNN	OhioT1DM	T1	6	55	✓	✓	✓	✓							✓	personnalisée	30,60	✓					✓		
4	[14]	Bunescu	2013	SVR, ARIMA, Ref	-	T1	10	28	✓	✓	✓	✓							✓	personnalisée	30,60	✓							
5	[21]	Contreras	2018	GE	OhioT1DM	T1	6	55	✓	✓	✓	✓					✓		✓	personnalisée	30,60,90	✓						✓	✓
6	[20]	Contreras	2017	GE	T1DMS	V1	100	14	✓	✓	✓								✓	personnalisée	120,240,360	✓	✓					✓	✓
7	[24]	Daskalaki	2013	ARX, RNN	-	T1	23	10	✓	✓										personnalisée	15,30,45	✓		✓	✓				
8	[45]	Eren-Oruklu	2012	ARMAX, ARMA	-	T2	5	24	✓			✓								personnalisée	30		✓					✓	
9	[44]	Eren-Oruklu	2009	ARMA, AR	- / -	T2 / T2	11 / 8	2 / 2	✓								✓			personnalisée	5,10,15,20,25,30		✓					✓	✓
10	[53]	Fiorini	2017	KRR, LSTM, ARIMA, KF	- / -	T1 / T2	72 / 34	7 / 7	✓											personnalisée	30,60,90	✓							✓
11	[56]	Gani	2009	AR	-	T1	9	5	✓								✓			personnalisée	30,60,90	✓			✓				
12	[64]	Georga	2019	KAF, SVR	-	T1	15	12	✓	✓	✓	✓							✓	personnalisée	5,15,30,45,60	✓	✓		✓	✓	✓		✓
13	[63]	Georga	2017	KAF, SVR, FFNN	-	T1	7	14	✓											personnalisée	15,30,60	✓		✓	✓		✓		
14	[62]	Georga	2016	KAF	-	T1	15	12	✓											personnalisée	5,15,30,45,60	✓						✓	
15	[61]	Georga	2015	ELM	-	T1	15	12	✓	✓	✓	✓							✓	personnalisée	30	✓			✓	✓			
16	[59]	Georga	2013	SVR	-	T1	27	13	✓	✓	✓	✓	✓				✓		✓	personnalisée	15,30,60,120	✓		✓				✓	
17	[60]	Georga	2013	SVR, FFNN, GP	-	T1	15	12	✓	✓	✓	✓	✓	✓					✓	personnalisée	30,60,				✓		✓		
18	[58]	Georga	2012	RF	-	T1	27	13	✓	✓	✓	✓	✓						✓	personnalisée	15,30,60,120	✓		✓				✓	
19	[69]	Hamdi	2018	SVR	-	T1	12	14	✓											personnalisée	15,30,45,60	✓	✓	✓					
20	[77]	Jankovic	2016	ARX+ELM+FFNN	-	T1	6	4	✓	✓	✓	✓								personnalisée	15,30,45	✓		✓	✓				
21	[78]	Jeon	2020	GBM, RF	OhioT1DM	T1	6	55	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	personnalisée	30	✓		✓					✓
22	[97]	Li'	2018	ESN, ELM	-	T1	8	3	✓										.	personnalisée	15,30,45	✓			✓			✓	
23	[96]	Li"	2019	CNN, SVR, ARX, FFNN, LVX	T1DMS / - / OhioT1DM	V1 / T1 / T1	10 / 10 / 6	180 / 42 / 55	✓	✓	✓		✓				✓		✓	personnalisée	30,60	✓	✓		✓				
24	[95]	Li"	2019	CNN+LSTM, SVR, ARX, FFNN, LVX	T1DMS / -	V1 / T1	10 / 10	360 / 42	✓	✓	✓						✓			personnalisée	30,60	✓	✓						✓
25	[99]	Liu	2018	Phys, LVX	T1DMS / -	V1 / T1	10 / 10	14 / 14	✓		✓	✓							✓	personnalisée	30,60,90,120	✓					✓	✓	✓
26	[105]	Macas	2017	ARMAX, ARX	-	T1	3	30	✓	✓	✓		✓						✓	personnalisée	60	✓						✓	
27	[110]	Martinsson	2020	LSTM, Ref	OhioT1DM	T1	6	55	✓											globale	30,60	✓							✓
28	[111]	Mayo	2019	SVR, DT, LR, FFNN, RF, GBM	OhioT1DM	T1	6	55	✓								✓	✓		personnalisée	30		✓					✓	
29	[113]	Midroni	2018	GBM, RF, LSTM	OhioT1DM	T1	6	55	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	personnalisée	30	✓							
30	[115]	Mirshekarian	2019	LSTM, ARIMA, Ref	AIDA / T1DMS / OhioT1DM	V1 / V1 / T1	40 / 10 / 6	600 / 1350 / 55	✓	✓	✓	✓	✓				✓	✓		personnalisée	30,60	✓						✓	
31	[114]	Mirshekarian	2017	LSTM, SVR, ARIMA, Ref	-	T1	10	28	✓	✓	✓						✓	✓		personnalisée	30,60	✓							
32	[116]	Montaser	2017	SARIMA, ARIMA, SARIMAX, ARIMAX	-	T1	10	0,33	✓	✓										globale	30,60,120,180,240,300	✓	✓						✓
33	[122]	Pappada	2011	FFNN	-	T1	27	3,6	✓	✓	✓				✓	✓			✓	globale	75	✓						✓	✓
34	[125]	Perez-Gandia	2010	FFNN, AR	- / -	T1 / T1	9 / 6	12 / 3	✓				✓				✓			globale	15,30,45	✓							✓
35	[126]	Phadke	2020	MA, ES, LR, ARIMA, Ref	- / OhioT1DM	T1 / T1	10 / 6	10 / 55	✓											personnalisée	15,30,45	✓	✓					✓	✓
36	[130]	Reymann	2016	SVR	AIDA / -	V1 / T1	5 / 1	25 / 35	✓											globale	30,60		✓						
37	[140]	Sparacino	2007	AR	-	T1	28	2	✓								✓			personnalisée	30,45					✓			✓
38	[142]	Stahl	2009	Phys., ARMA, ARMAX	-	T1	1	180	✓	✓	✓						✓		✓	personnalisée	15,60,120								✓
39	[143]	Sun	2018	LSTM, ARIMA, SVR	- / -	T1 / V1	20 / 11	23 / 38	✓								✓	✓		globale	15,30,45,60	✓		✓	✓				
40	[148]	Valetta	2009	GP	-	T1	1	9	✓	✓	✓	✓							✓	personnalisée	25,60,240								✓
41	[150]	Vehi	2020	GE	- / OhioT1DM / T1DMS	T1 / T1 / V1	10 / 6 / 100	7 / 55 / 14	✓	✓	✓						✓		✓	personnalisée	60						✓	✓	✓
42	[154]	Wang	2014	KF, ARX	T1DMS / -	V1 / T1	30 / 5	1 / 2	✓	✓	✓								✓	personnalisée	30		✓	✓	✓	✓	✓		✓
43	[159]	Yu	2018	KAF	-	T1	3	5	✓											globale	30	✓	✓				✓		
44	[160]	Yu	2018	KAF	T1DMS / -	V1 / T1	30 / 3	3 / 5	✓											globale	30	✓	✓				✓		
45	[163]	Zarkogianni	2015	FFNN, SOM, WFNN, LR	-	T1	10	6	✓			✓							✓	personnalisée	30,60,120	✓	✓	✓			✓		
46	[166]	Zecchin	2014	FFNN	-	T1	20	3	✓		✓								✓	globale	30	✓			✓	✓			
47	[164]	Zecchin	2012	FFNN+AR	T1DMS / -	V1 / T1	20 / 15	11 / 7	✓		✓								✓	globale	30	✓			✓	✓			✓
48	[167]	Zhao	2012	LVX, AR, ARX	T1DMS / -	V1 / T1	10 / 7	7 / 2,	✓	✓	✓						✓			personnalisée	30,60	✓					✓		
49	[169]	Zhu	2020	RNN, SVR, ARX, FFNN	T1DMS / OhioT1DM	V1 / T1	10 / 6	360 / 55	✓	✓	✓		✓				✓	✓		personnalisée	30	✓	✓		✓				
50	[168]	Zhu	2018	CNN	OhioT1DM	T1	55	6	✓	✓	✓		✓				✓	✓		personnalisée	30	✓	✓						

[*] Les différents jeux de données utilisés sont séparés d'un / ;
[†] T1 : type 1 réel ; T2 : type 2 réel ; V1 : type 1 virtuel ;
[‡] nombre de jours moyen par patient au sein des jeux respectifs ;

Tableau 2.1: Description détaillée des études traitant de la prédiction de la glycémie entre 2007 et 2020.

Ce chapitre est structuré comme suit. Tout d'abord, nous nous intéresserons aux données utilisées dans l'apprentissage de ces modèles. Nous nous focaliserons notamment sur l'hétérogénéité des données, la nature et taille des cohortes de personnes diabétiques, ainsi que sur les étapes de prétraitement appliquées aux données avant leur utilisant dans l'apprentissage des modèles prédictifs. Puis, nous procédons à la revue des algorithmes utilisés dans la prédiction de glycémie, depuis ses débuts autour des modèles autorégressifs, jusqu'aux modèles plus complexes d'apprentissage automatique et profond utilisés ces dernières années. Suite à cela, nous examinerons les différentes métriques statistiques et cliniques utilisées pour l'évaluation des modèles prédictifs de glycémie. Enfin, après une synthèse des résultats de la littérature, nous procéderons à son analyse critique, dégageant ainsi plusieurs pistes de recherches.

2.2 Données expérimentales

Dans cette section nous détaillons les données expérimentales utilisées dans la construction et évaluation des modèles prédictifs de glycémie. Nous analysons ces données expérimentales selon plusieurs perspectives. Tout d'abord, nous nous intéressons à la composition des cohortes de patients. Puis, nous décrivons la nature des données utilisées dans l'apprentissage des modèles. Enfin, nous détaillons les étapes de prétraitement de ces données avant leur utilisation dans les modèles prédictifs.

2.2.1 Composition des cohortes expérimentales

Tout d'abord, dans l'ensemble des études nous pouvons distinguer trois types de patients étudiés : les patients diabétiques de type 1, les patients diabétiques de type 2, et les patients diabétiques *virtuels* de type 1. Seulement 3 études utilisent des patients diabétiques de type 2 [44, 45, 53]. Excepté [20], toutes les études utilisent des données provenant de patients réels. 14 études utilisent des patients provenant de jeux de données différents, 12 d'entre elles utilisent des données virtuelles.

L'extrême sous-représentation des personnes diabétiques de type 2 au sein des études, alors qu'ils représentent près de 90% de population diabétique [49], peut s'expliquer par la raison suivante. La gravité du dérèglement de la régulation de la glycémie du type 2 évoluant avec le temps, la population diabétique du type 2 est bien plus hétérogène, et donc difficile à étudier, que celle du type 1.

Quant à l'utilisation de plusieurs jeux de données différents, et notamment de celle de données virtuelles, cela peut s'expliquer par le faible nombre de patients composants généralement les cohortes. En effet, au sein des jeux de données utilisés, le nombre de patients est très variable. En effet, le nombre de patients varie entre 1 [142, 148] et 451 [5] pour les patients réels (médiane de 10), et entre 5 [130] et 100 [20, 150] (médiane de 10.5) pour les patients virtuels. Les quantités de données par patient sont aussi très variables avec en moyenne 10.10 jours de

données pour les patients réels et 127.20 pour les données simulées [1].

Parmi les études analysées, plusieurs utilisent des données provenant de jeux de données publiques ou de simulateurs : AIDA (2), OhioT1DM (12), T1DMS (11). Créé au début des années 90 [92], AIDA est un logiciel qui simule les effets de l'insuline et des changements d'alimentation sur la glycémie des patients de type 1 [93]. Toutefois, les auteurs du logiciel déconseillent l'utilisation du logiciel hors d'un cadre éducationnel. OhioT1DM est un jeu de données public mis à disposition par Marling *et al.* en 2018 [109]. Il est constitué de 6 patients diabétiques de type 1 pour lesquelles des données variées (glycémie, insuline, glucides, activité physique, etc.) ont été récoltées pendant 8 semaines. Enfin, T1DMS est un environnement de simulation de 30 patients *in silico*, virtuels, de type 1 [106]. En 2018, il a été accepté par la Food & Drug Administration aux États-Unis, dans le cadre du développement de nouvelles stratégies de traitement du diabète de type 1, en remplacement des tests précliniques sur les animaux.

2.2.2 Nature des données expérimentales

Les données utilisées pour la prédiction de la glycémie sont très variées. Bien que chaque étude n'utilise pas exactement les mêmes données, nous pouvons noter plusieurs similarités. Nous proposons de séparer ces données en plusieurs groupes distincts :

— **Glycémie (100%) :** Toutes les études utilisent au minimum les valeurs passées de glycémie pour prédire la glycémie future. La glycémie des patients, généralement mesurée en mg/dL ou g/L, est récoltée à travers l'utilisation de acrshortcgm tels que le FreeStyle Libre [4, 126, 130] ou Medtronic Enlite [109]. Ces capteurs ont généralement une fréquence d'échantillonnage d'un échantillon toutes les 5 à 15 minutes. La glycémie peut aussi s'obtenir à travers l'utilisation de lancettes comme pour le jeu de données OhioT1DM [113, 78]. En comparaison avec les données obtenues via les capteurs de glycémie, la glycémie obtenue par lancettes est plus éparse (une valeur par utilisation de lancette, soit environ trois valeurs par jour) mais permet d'accéder aux taux réels de glucose dans le sang. En effet, les capteurs de glycémie mesurent le taux de sucre dans le liquide interstitiel, taux présentant un retard d'environ 12.5 minutes avec le taux de sucre dans le sang [87].

— **Insuline (56 %) :** 28 études reportent utiliser des données liées aux injections d'insuline (en unités ou pmol). Nous pouvons distinguer deux types d'insuline : l'insuline basale ou l'insuline en bolus. L'insuline en bolus représente l'injection à un instant précis d'une grande quantité d'insuline, souvent en prévision de l'augmentation postprandiale de la glycémie (suite à un repas). L'insuline basale, quant à elle, représente de faibles quantités d'insuline dispensées en continu par les pompes à insuline. Bien que souvent constantes, elles peuvent être modifiées au besoin par la personne diabétique à travers le réglage de la pompe. Toutes les études utilisant des données d'insuline utilisent les données liées aux bolus d'insuline. Certaines études mentionnent toutefois l'inclusion de l'insuline basale comme donnée supplémentaire [14, 105, 114, 115, 143].

1. Certaines études comme [167] ne mentionnent pas exactement le nombre de jours par patient. Pour ces études, nous avons approximé cette quantité en utilisant les informations données par les auteurs.

— **Glucides (56%) :** 28 études utilisent les valeurs des apports en glucides (en g ou niveaux [105]) pour la prédiction de la glycémie. Ces valeurs sont généralement enregistrées au sein de carnet de bord physique ou électronique par les patients diabétiques.

— **Activité physique (32%) :** Avec 16 études, les données liées à l'activité physique sont les données, après celles de glycémie, d'insuline et de glucides, les plus utilisées pour la prédiction de la glycémie. Les données sont très variées et dépendent grandement du bracelet de fitness utilisé. Nous retrouvons des estimations de dépenses énergétiques [45, 58, 60, 59, 64, 163], de température [45, 148, 78, 113, 115], de sueur [115, 45, 78, 113], de fréquence cardiaque [115, 78, 113], de nombre de pas [10, 21, 150, 78, 113], ou simplement d'accélération [45, 148, 78, 113].

— **Temps (20%) :** 10 études ajoutent aux modèles une notion de temporalité sous la forme de l'heure de la journée [58, 60, 59, 113, 115, 168, 169, 96] ou du jour de la semaine [113]. L'idée est de donner la possibilité au modèle de modéliser le rythme circadien, regroupant tous les processus biologiques cycliques d'une durée d'environ 24h. Aussi, une telle information permet de potentiellement tenir compte du *dawn phenomenon*, caractérisé par une augmentation de la glycémie de la personne diabétique entre 2 et 8h du matin.

— **Sommeil (6%) :** 3 études explorent la prédiction de glycémie nocturne afin de détecter les éventuelles hypoglycémies pendant la nuit. Celles-ci peuvent être particulièrement dangereuses, car le patient peut ne pas se réveiller pour la faire remonter (en mangeant par exemple). Dans leur étude, Georga *et al.* utilisent l'information de détection de sommeil du bracelet de fitness SenseWear [58]. Quant à eux, Midroni *et al.* ainsi que Jeon *et al.* ont utilisé les informations d'heure d'endormissement ainsi que de qualité de sommeil, toutes deux reportées par le patient manuellement et par le bracelet de fitness Basis Peak.

— **Évènements (6%) :** 3 études ont utilisé des données que nous pouvons caractériser comme événementielles. Pappada *et al.* mentionne l'utilisation de données liées aux activités ou style de vie des patients (pas d'informations supplémentaires ne sont données par les auteurs) [122]. À travers le jeu de données OhioT1DM, Midroni *et al.* et *et al.* utilisent des données de début et fin de temps de travail, si le patient se sent malade ou non [113, 78].

— **Humeur (6 %) :** 3 études reportent l'utilisation de données que nous pouvons relier à l'humeur de la personne. Parmi ces données, nous retrouvons le niveau de stress estimé par le patient [113, 78] ou des facteurs émotionnels [122].

L'état de l'art nous montre la grande variété des données utilisées pour la prédiction de la glycémie, avec les valeurs passées de glycémie, d'insuline et de glucides comme données principales. L'ancienneté maximale de ces données utilisées par les modèles est aussi très variable. Bien que souvent non détaillée, l'ancienneté maximale des données, que nous pouvons aussi appeler *longueur de l'historique d'entrée*, varie de 5 minutes pour les modèles autorégressifs simples [140] à 1 heure [168, 169, 110, 78], 2 heures [113, 111, 150] et 3 heures [5, 4].

Quelques études récentes nous permettent de quantifier l'intérêt de chacune de ces données pour la prédiction de la glycémie. Jankovic *et al.* ont montré que l'utilisation des données d'insuline et de glucide permet d'améliorer sensiblement les prédictions [77]. Ces résultats sont soutenus par ceux de Mirshekarian *et al.* [115] et quelque peu contrastés par ceux de Midroni *et al.* [113], suggérant que les données d'insuline n'auraient que peu d'importance. Jeon *et al.* soutiennent que les données d'insuline n'auraient de l'importance que pour certains patients [78]. Les analyses concernant les données d'activité physiques sont elles aussi contrastées. D'un côté, Mirshekarian *et al.* reportent une amélioration avec l'utilisation de données de température de la peau, de conductivité de la peau, et de fréquence cardiaque. D'un autre côté, les meilleurs modèles de Midroni *et al.* sont ceux ne possédant pas ces variables, variables qui sont classées comme étant les moins efficaces par l'analyse d'importance des variables de l'étude de Jeon *et al.*. Quant au reste des données, telles que celles du temps, de sommeil, d'humeur ou évènementielles, elles sont aujourd'hui trop peu étudiées pour pouvoir conclure sur l'intérêt de leur utilisation.

2.2.3 Étapes de prétraitement des données

Avant d'utiliser les données récoltées pour l'entraînement des modèles, de nombreuses études les prétraitent afin de faciliter l'entraînement et ainsi d'améliorer les performances. Parmi ces étapes, nous retrouvons des étapes de nettoyage de données, d'augmentation des quantités de données utilisées et d'extraction manuelle de descripteurs.

Nettoyage des données (40%)

Nous avons vu que les données de glycémie sont les plus importantes dans le milieu. Collectées avec des capteurs de glycémie en continu, celles-ci présentent néanmoins quelques problèmes pouvant heurter l'apprentissage des modèles.

— **Nettoyage manuel des anomalies (4%) :** Il n'est pas rare que ces capteurs renvoient des valeurs erronées, dues à un mauvais positionnement du capteur par exemple. Ces erreurs sont généralement en quantités assez faibles et peuvent être retirées à la main [143, 150]. Elles peuvent être identifiées comme des valeurs aberrantes en comparaison avec les valeurs avoisinantes [143].

— **Lissage des signaux de glycémie (14%) :** Par ailleurs, le signal de glycémie comporte du bruit. Ce bruit, caractérisé par des variations hautes-fréquences, peut être atténué par l'utilisation de filtres. Plusieurs filtres différents ont été utilisés ces dernières années pour la prédiction de la glycémie. Nous recensons l'utilisation de filtres passe-bas [44, 139], de filtres de Kalman [125, 14], de filtres gaussiens [95] ou médians [168, 169]. Ces filtres sont généralement utilisés sur le signal de glycémie donné en entrée au modèle, et non sur le signal de glycémie à prédire, forçant une évaluation des modèles sur les observations réelles et bruitées de glycémie. L'importance de filtrer le signal de glycémie pour améliorer les prédictions de glycémie est incertaine. En effet,

pour être utilisés en pratique, ces filtres doivent être causals (i.e., ne pas utiliser les valeurs futures pour opérer le filtrage), impliquant un fort déphasage du signal de glycémie. Ce déphasage induit une perte d'information non négligeable pour la prédiction de la glycémie future.

— **Imputation des données manquantes (36%) :** Les signaux de glycémie comportent généralement un nombre variable de valeurs manquantes. Celles-ci peuvent provenir à la fois de défauts capteurs, d'erreurs techniques de la part du patient (e.g., remplacement tardif du capteur), du nettoyage des valeurs aberrantes ou bien du suréchantillonnage du signal de glycémie pour le synchroniser avec les autres signaux. Bien que certains chercheurs aient décidé de ne pas utiliser les séquences comportant des valeurs manquantes [5, 111, 59], il est possible de les générer artificiellement. La méthode la plus simple est sans doute d'utiliser la dernière valeur connue à la place de la valeur manquante [126, 113, 78]. Il est aussi possible d'interpoler les valeurs manquantes à partir des valeurs avoisinantes. Plusieurs niveaux de complexité d'interpolation peuvent être utilisés, comme l'interpolation linéaire [78, 113, 114, 115, 143, 168, 169, 142] ou l'interpolation spline [95, 96, 78, 125, 130, 142]. Jeon *et al.* ont conduit une étude visant à définir les méthodes d'imputation les plus efficaces pour la prédiction de la glycémie [78]. Parmi les 11 techniques évaluées, les plus efficaces sont celles réutilisant la valeur précédente, l'interpolation linéaire, l'interpolation spline, l'interpolation Stineman et les filtres de Kalman.

Augmentation des quantités de données (12%)

La quantité des données utilisées pour l'apprentissage des modèles est tout aussi importante que leur qualité. Un des challenges principaux de la prédiction de la glycémie est le besoin d'avoir des modèles personnalisés au patient pour tenir compte de la grande inter/intra variabilité de la population diabétique. En pratique, les modèles n'utilisent souvent que les données d'un même patient pour l'entraînement du modèle lui étant personnalisé, souvent en quantités assez faibles. En conséquence, les modèles sont généralement assez peu performants, performances variant grandement en fonction des régions glycémiques (e.g., la région d'hypoglycémie est difficilement traitée en raison du peu de données d'apprentissage dans cette région). Pour pallier ces limitations, quelques récentes approches de suréchantillonnage et d'apprentissage par transfert ont été utilisées :

— **Suréchantillonnage des données d'apprentissage (4%) :** Mayo *et al.* ont étudié l'utilisation de trois techniques différentes de suréchantillonnage pour permettre à leurs modèles d'être plus performants dans les régions sous-représentées d'hypo et d'hyperglycémie [111]. Parmi ces méthodes, la méthode *adaptive synthetic* (ADASYN) génère des échantillons artificiels basés sur leur difficulté de prédiction relative. Cette difficulté est évaluée à l'aide de la méthode des K plus proches voisins, un échantillon isolé étant supposé être plus difficile à prédire qu'un autre. De leur côté, Zhu *et al.* ont combiné les données de chaque patient à 10% des données des autres patients (représentant une augmentation totale de 50% des données initiales par patient) [168].

— **Apprentissage par transfert (6%) :** L'apprentissage par transfert, consistant à entraîner un premier modèle global sur tous les patients, puis à affiner le modèle sur le patient d'intérêt, a été utilisé par Sun *et al.* [143] ainsi que par Mirshekarian *et al.* [115] et Zhu *et al.* [169]. Globalement, tous les auteurs reportent l'intérêt d'augmenter intelligemment la quantité des données utilisées pour l'apprentissage des modèles, et en particulier pour ceux basés sur l'apprentissage profond.

Extraction de descripteurs (46%)

Afin d'aider les modèles dans l'apprentissage de la prédiction de la glycémie future, il est possible d'extraire manuellement certaines informations expertes. Celles-ci peuvent être à la fois statistiques ou décrivant des phénomènes physiologiques.

— **Descripteurs statistiques (18%) :** Parmi les descripteurs statistiques utilisés communément, nous retrouvons première dérivée du signal de glycémie [10, 78, 122, 163, 113] pour connaître sa vitesse, l'accumulation de dépenses énergétiques depuis un point dans le temps [58, 60, 59, 64, 163, 60], ou bien le temps passé depuis un évènement comme une prise d'insuline, un apport en glucides [113], ou depuis l'endormissement du patient [60]. Enfin, nous avons vu précédemment que les signaux peuvent posséder des données manquantes et qu'elles peuvent être générées artificiellement. Jeon *et al.* et Midroni *et al.* proposent d'inclure un descripteur indiquant si les valeurs ont été générées artificiellement [78, 113].

— **Descripteurs physiologiques (30%) :** Par le passé, de nombreux chercheurs ont tenté de modéliser le système d'autorégulation de la glycémie en utilisant des équations mathématiques. Ces approches reposent généralement sur l'interconnexion de compartiments, chaque compartiment décrivant les dynamiques d'une portion du système pour lequel il est impossible de faire des mesures directes. Ils permettent notamment d'estimer l'absorption et les dynamiques de l'insuline, l'absorption et la digestion des glucides, ou l'effet de l'activité physique du patient :

  — *Modélisation de l'insuline :* Les deux approches principales pour modéliser les dynamiques de l'insuline sont l'estimation de l'insuline toujours active dans le corps après injection, $IoB(t)$ [10, 20, 21, 150], ainsi que l'estimation de la concentration d'insuline dans le plasma $I_p(t)$ [58, 60, 59, 64, 99, 96]. L'Équation 2.1 décrit le calcul de $IoB(t)$ à travers 2 compartiments C_1 et C_2, l'injection d'insuline $u(t)$ ainsi que la constante de durée de l'action de l'insuline K_{DIA}. L'Équation 2.1) décrit l'évolution de $I_p(t)$ en fonction des constantes V_1 et k_e décrivant respectivement le volume de distribution de l'insuline dans le plasma et le ratio d'élimination de l'insuline par le plasma. D'autres modélisations de l'insuline sont utilisées comme le modèle de Berger prenant en compte l'insuline lente et rapide [142] et l'étalement dans le temps des

prises d'insuline [105, 154].

$$\dot{C}_1(t) = u(t) - K_{DIA}C_1(t) \tag{2.1a}$$

$$\dot{C}_2(t) = K_{DIA}(C_1(t) - C_2(t)) \tag{2.1b}$$

$$IoB(t) = C_1(t) + C_2(t) \tag{2.1c}$$

$$\dot{I}_p(t) = \frac{u(t)}{V_1} - k_e I(t) \tag{2.2}$$

— *Modélisation des glucides :* L'approche principale pour modéliser l'absorption des glucides se fait à travers l'estimation de la vitesse d'apparition dans le système de sucre provenant des repas $Ra(t)$. Celle-ci peut s'estimer directement à travers l'Équation 2.3, où C_{in} est la quantité de glucides ingérés, et où C_{bio} et t_{max} sont des constantes représentant respectivement la biodisponibilité et le temps d'apparition maximal des glucides dans le compartiment [10, 20, 21, 99, 150, 150]. La vitesse d'apparition du sucre dans le système peut aussi s'exprimer en fonction du glucose dans les intestins q_{gut}, de sa constante d'absorption k_{abs}, et d'évacuation depuis l'estomac vers les intestins en fonction des quantités en glucides $G_{empt}(t, D)$, voir Équation 2.4 [58, 60, 59, 64, 164, 166]. D'autres modélisations de l'absorption des glucides sont utilisées, comme l'utilisation de deux compartiments symbolisant la consommation et la digestion [14, 114], la modélisation des glucides rapides et lents [142], l'étalement dans le temps de la prise en compte des glucides [105, 154].

$$Ra(t) = \frac{C_{in}C_{bio}te^{-t/t_{max}}}{t_{max}^2} \tag{2.3}$$

$$\dot{q}_{gut} = -k_{abs}q_{gut}(t) + G_{empt}(t, C_{in}) \tag{2.4a}$$

$$Ra(t) = k_{abs}q_{gut}(t) \tag{2.4b}$$

— *Modélisation de l'activité physique :* L'effet de l'activité physique peut se modéliser par l'estimation des équivalents métaboliques [148] ou à travers l'Équation 2.5 [10, 21]. Celle-ci mesure l'activité à bord $AoB(t)$ comme le nombre de pas total à l'instant t depuis le début de la journée, $steps(t)$, multiplié par une décroissance exponentielle en fonction du temps et de la constante k_s liée à la durée de l'activité

physique.

$$AoB(t) = steps(t) \cdot e^{-k_s t} \tag{2.5}$$

Bien que de nombreux descripteurs différents sont utilisés à travers l'ensemble de ces études, leur impact sur la prédiction de la glycémie n'est que très peu étudié. Seule Georga *et al.* ont analysé l'utilisation de différentes combinaisons de leurs descripteurs [64]. Ils montrent, notamment, que l'estimation de la quantité d'insuline dans le plasma, I_p, ainsi que l'accumulation des dépenses énergétiques, sont des informations utiles à la prédiction de la glycémie. L'estimation de la vitesse d'apparition du sucre dans le système, Ra, quant à elle, ne semble pas améliorer les prédictions.

2.3 Modèles prédictifs de glycémie future de patients diabétiques

Dans cette section, nous décrivons les différents modèles et algorithmes utilisés au cours des dernières années pour la prédiction de la glycémie chez des patients diabétiques.

Bien que de natures très variées, ces modèles appartiennent au domaine de l'apprentissage supervisé. Ils peuvent être décrits comme une fonction $f_\theta : \mathcal{X} \to \mathcal{Y}$, paramétrisée par θ et faisant une approximation de l'espace de données de sortie $\mathcal{Y}$ à partir de l'espace de données d'entrée $\mathcal{X}$. L'objectif de l'apprentissage est de trouver la paramétrisation θ qui minimise l'erreur de prédiction de l'espace $\mathcal{Y}$ à partir de l'espace $\mathcal{X}$. Cet apprentissage se fait à partir de N échantillons d'apprentissage (x_n, y_n), $n \in [1, N]$, $x_n \in \mathcal{X}$, et $y_n \in \mathcal{Y}$. Nous dénotons par $\hat{y}_n$ la prédiction faite par le modèle pour l'échantillon n. Nous pouvons séparer les tâches relevant de l'apprentissage supervisé en deux catégories : les tâches de régression et les tâches de classification. Pour une tâche de classification, l'espace de sortie $\mathcal{Y}$ est fini et discret. Il contient les différentes classes à prédire pour la tâche en question. Quant aux tâches de régression, elles sont caractérisées par un espace $\mathcal{Y}$ continu appartenant à $\mathcal{R}$.

La prédiction de la glycémie future de patients diabétiques est une tâche de régression. L'espace de sortie $\mathcal{Y}$ est représenté par la glycémie du patient dans le futur à un horizon de prédiction, *Prediction Horizon* (PH), $PH \in \mathcal{R}$ (minutes), y_{t+PH}. Tandis que t représente l'instant présent, $t + PH$ représente l'instant futur pour lequel nous cherchons à prédire la glycémie. Les horizons de prédiction considérés vont généralement de très court terme (15 minutes) à long terme (120 minutes) [119]. Quant à l'espace d'entrée $\mathcal{X}$, il varie d'une étude à une autre, en fonction des données considérées pour effectuer la prédiction de la glycémie. De manière générale, $\mathcal{X}$ est décrit par les valeurs de glycémies passées du patient diabétique jusqu'à H minutes dans le passé. Dans nos travaux, nous nous référons à H comme la longueur de l'historique des valeurs passées utilisées pour faire les prédictions. Comme nous l'avons vu précédemment, d'autres informations telles que les valeurs passées de prise d'insuline, d'ingestion de glucides, d'activité physique ou de sommeil peuvent être utilisées. Ainsi, nous pouvons représenter

l'espace de données d'entrées $\mathcal{X}$ par $\mathcal{R}^{H \times r}$, où r est le nombre de signaux différents utilisés pour la prédictions (e.g., glycémie, insuline, glucides, etc.).

2.3.1 Modèles autorégressifs

Cadre théorique

Un modèle autorégressif standard, aussi appelé processus autorégressif (AR), est un modèle de régression pour séries temporelles dans lequel les valeurs futures de la série considérée sont décrites par ses valeurs passées. Nous pouvons définir un processus autorégressif AR d'ordre p par l'Équation 2.6. L'ordre p d'un processus autorégressif décrit l'ancienneté maximale du signal étant utile à la prédiction. Il est, en ce sens, similaire à longueur de l'historique d'entrée H. Tandis que α_i, $i \in [1, p]$, représente les paramètres du modèle, b est une constante qui lui est ajoutée et ϵ_t est du bruit blanc. L'Équation 2.6 nous donne la prédiction à l'instant d'après, correspondant à un horizon de prédiction PH de 1. Pour obtenir une prédiction à l'horizon de prédiction PH, nous devons calculer un à un les prédictions des instants $t+2, ..., t+PH$, en utilisant la prédiction de l'instant précédent comme entrée au modèle.

$$AR(p) : \hat{y}_{t+1} = \alpha_1 y_t + ... + \alpha_p y_{t-p-1} + b + \epsilon_t \tag{2.6a}$$

$$\hat{y}_{t+1} = \sum_{i=1}^{p} \alpha_i y_{t-i-1} + b + \epsilon_{t+1} \tag{2.6b}$$

Un des problèmes avec les modèles AR simples est que ces derniers ne peuvent pas prendre en compte des bruits blancs ϵ_i dépendants les uns des autres au sein de la série temporelle. Pour pallier ce problème, un modèles autorégressif standard peut être étendu à processus autorégressif avec moyenne mobile (ARMA). Décrit par l'Équation 2.8, un modèle ARMA(p,q) est la combinaison d'un modèle AR(p) et d'un modèle MA(q). Le modèle MA(q) est définit par l'Équation 2.7, où q représente l'ordre du modèle et γ_i, $i \in [1, q]$, les paramètres du modèle.

$$MA(q) : \hat{y}_{t+1} = \epsilon_{t+1} + \sum_{i=1}^{q} \gamma_i \epsilon_{t-i-1} \tag{2.7}$$

$$ARMA(p, q) : \hat{y}_{t+1} = AR(p) + MA(q) \tag{2.8}$$

Une série temporelle ne peut pas toujours être expliquée uniquement par ses valeurs précédentes. Ainsi, il est possible d'ajouter des données extérieures à la série temporelle d'intérêt au modèle AR ou ARMA afin d'affiner les prédictions. Ces données exogènes donnent lieu aux modèles ARX(p,e) et ARMAX(p,q,e) définis respectivement

par les Équations 2.9 et 2.10. Ces données exogènes, d_i, $i \in [1, e]$ sont pondérées par le vecteur de coefficients $\boldsymbol{\eta}$, paramètres du modèle.

$$ARX(p, e) : \hat{y}_{t+1} = AR(p) + \sum_{i=1}^{e} \eta_i d_{t-i-1} \tag{2.9}$$

$$ARMAX(p, q, e) : \hat{y}_{t+1} = ARMA(p, q) + \sum_{i=1}^{e} \eta_i d_{t-i-1} \tag{2.10}$$

Les modèles AR, ARX, ARMA et ARMAX supposent que les séries temporelles sont stationnaires. Un processus est dit stationnaire lorsque se moyenne et sa variance ne varient pas au cours du temps. Une série temporelle non stationnaire peut être rendue stationnaire en prenant sa dérivée. Cette méthodologie peut être incluse aux processus autorégressifs à travers une composante d'Intégration, donnant ainsi lieu aux modèles ARIMA et ARIMAX. Par ailleurs, il est possible de modéliser une composante saisonnière, récurrente dans le temps, au sein d'un processus autorégressif à travers les modèles SARIMA et SARIMAX.

Application à la prédiction de la glycémie

Beaucoup des premiers travaux dans le domaine de la prédiction de la glycémie utilisent les modèles autorégressifs. Parmi eux nous pouvons citer les travaux de Sparacino *et al.* [140]. Ces derniers ont notamment montré que prédire la glycémie de patients diabétiques 30 minutes dans le futur était possible avec un modèle AR et que cela permettait de prévenir des hypoglycémies et hyperglycémies futures. Suite à ces travaux, Gani *et al.* ont investigué les conditions pour rendre le modèle AR stable et précis [56]. Quant à eux, Eren-Oruklu *et al.* ont étudié l'utilisation des modèles AR et ARMA à travers une identification récursive et dynamique de leurs paramètres afin de réduire l'impact de l'inter/intra variabilité de la population diabétique [44]. Parallèlement, Stahl *et al.* ont exploré l'utilisation des processus ARMA et ARMAX pour la prédiction de la glycémie 2 heures dans le futur [142]. Bien que les modèles n'aient pas été jugés suffisamment précis, ces travaux ont permis de mettre en évidence l'importance des données exogènes comme les apports en glucides ou les prises d'insuline. Suite à ces travaux préliminaires, de nombreux chercheurs ont étudié l'utilisation des processus autorégressifs ARX [24, 105] et ARMAX [45, 105]. Quant à elles, les composantes d'intégration et de saisonnalité n'ont vu que peu d'utilisations. Nous pouvons noter les récents travaux de Montaser *et al.* montrant que lorsque les repas sont à heure fixe et constants dans leurs apports en glucides, la composante de saisonnalité d'un modèle SARIMAX permet de significativement améliorer la précision des modèles [116]. Bien que moins recherchés ces dernières années, les processus autorégressifs sont très souvent utilisés comme modèles de référence [154, 5, 4, 95, 96, 125, 126, 143, 164, 167, 169, 77, 114, 115]

2.3.2 Modèles basés sur les arbres de décision

Cadre théorique

Les arbres de décision, *Decision Trees* (DT), sont des méthodes d'apprentissage automatiques très populaires en raison de leur relative simplicité d'interprétation. En partitionnant récursivement les échantillons d'entrée (x_n, y_n) en créant une fonction prédictive simple pour chaque partition, le processus de prédiction des arbres décisions peut être représenté graphiquement [101]. La Figure 2.1 donne un exemple simple d'un arbre de décision pour la prédiction de la glycémie.

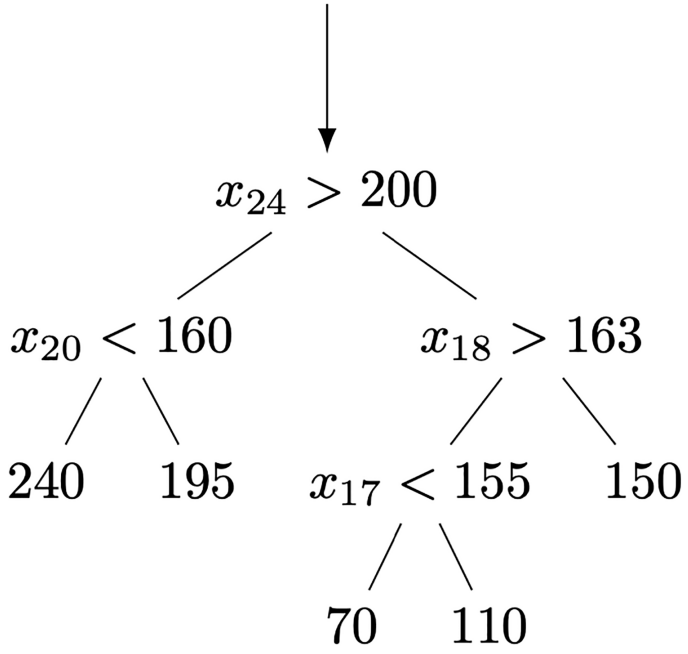


FIGURE 2.1: Exemple d'arbre de décision pour la prédiction de la glycémie future de personnes diabétiques [111].

De manière plus formelle [52], un arbre de décision peut être caractérisé comme un graphe orienté acyclique possédant une *racine*, des *nœuds* et des *feuilles*. À l'exception de la racine, le degré d'entrée de chaque nœud est de 1 et celui de sortie de 2. Lors du parcours de l'arbre, chaque embranchement se fait au moyen d'un critère de séparation $(x_{n,j}$ OP $\alpha)$, où $x_{n,j}$ représente la j-ème variable du n-ème échantillon d'apprentissage, OP représente un opérateur de comparaison quelconque (e.g., $>$, $\leq$, $=$), et α le seuil de décision. L'apprentissage des critères de décision se fait au moyen de la mesure d'impureté, quantifiant la similarité des échantillons d'apprentissage arrivant dans un nœud donné [101]. Différentes mesures d'impureté peuvent être utilisées. Pour des tâches de classification, l'algorithme C4.5 utilise l'entropie comme mesure d'impureté et l'algorithme CART utilise une généralisation de la variance appelée l'index Gini. Pour des tâches de régression, la mesure d'impureté est simplement la somme quadratique de la déviation à la moyenne. À chaque nouveau nœud, le critère de décision est choisi comme celui réduisant la somme totale d'impureté des deux nœuds résultants de la séparation.

Bien qu'intéressants en raison de leur représentation graphique et leur interprétabilité, les arbres de décision sont généralement assez peu performants à cause de leur structure simple. Pour pallier ces limitations, de nombreuses méthodes ensemblistes basées sur les arbres de décision sont utilisées. Parmi elles, les forêts aléatoires, *Random Forests* (RF), ainsi que les machines à boosting de gradient, *Gradient Boosting Machines* (GBM), sont sans doute les plus populaires. Ces techniques utilisent un grand nombre d'arbres de décision dont les prédictions individuelles sont moyennées pour calculer la décision finale. Ces deux méthodes diffèrent par la méthode de construction individuelle des arbres, ainsi que par la construction de la forêt dans son ensemble. Pour les forêts aléatoires, tous les arbres sont construits simultanément. La construction individuelle des arbres comporte une part d'aléatoire pour encourager la diversité des décisions au sein de la forêt [132]. Chaque arbre est créé sur un sous-ensemble aléatoire des échantillons d'apprentissage. Lors du choix du critère de séparation de nœud, seulement un sous-ensemble de variables d'entrée est considéré. Quant aux forêts créées par boosting de gradient, chaque arbre est créé itérativement en ayant pour objectif de réduire les erreurs résiduelles de la forêt. Bien que généralement plus performants, les méthodes ensemblistes basées sur les arbres de décision s'accompagnent d'une importance baisse en interprétabilité. Cela est dû à la grande taille des forêts rendant l'analyse humaine laborieuse et non aisée. Toutefois, ces modèles le sont tout de même plus que la plupart des modèles issus de l'apprentissage automatique grâce à l'importance Gini. Pour un arbre seul, l'importance Gini d'une variable d'entrée est calculée comme la baisse en impureté par le nœud opérant la décision sur cette variable, pondérée par la probabilité d'atteindre le nœud. Pour un modèle RF ou GBM, l'importance des variables est moyennée sur l'ensemble de la forêt.

Application à la prédiction de la glycémie

Dans le domaine de la prédiction de la glycémie, les arbres de décision simples ne sont généralement pas souvent suffisamment précis et sont donc surtout utilisés comme modèles de référence [111]. Quant aux forêts aléatoires, Georga *et al.* ont été les premiers à explorer leur utilité pour la prédiction de la glycémie [58]. Les forêts aléatoires ont été utilisées par Jean *et al.* pour déterminer l'importance des variables d'entrées sur la prédiction de la glycémie [78]. Leurs résultats ont montré que le signal de glycémie passé est le plus important pour la prédiction de la glycémie. En revanche, les mesures d'activités physiques n'ont pas révélé d'intérêt probant. Toutefois, d'autres études ont montré que les forêts aléatoires étaient inférieures aux forêts construites autour du boosting de gradient [113, 111, 78].

2.3.3 Modèles basés sur l'utilisation de noyaux

Cadre théorique

Transformation de l'espace d'entrée par l'utilisation de noyaux

Souvent il n'est pas possible d'approximer correctement l'espace objectif $\mathcal{Y}$ à partir de l'espace d'entrée $\mathcal{X}$ de

manière linéaire. Certains modèles linéaires (e.g., SVM que nous décrivons plus bas) utilisant une mesure d'erreur basée sur le produit scalaire peuvent implémenter une technique du nom d' « astuce du noyau » pour prendre des décisions non linéaires [135]. Nous pouvons définir un noyau par une fonction $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{R}$ calculant le produit scalaire de l'espace d'entrée $\mathcal{X}$ dans un espace intermédiaire $\mathcal{F}$. L'espace intermédiaire $\mathcal{F}$ est un espace de plus grande dimension que $\mathcal{X}$ et peut être obtenu à partir de $\mathcal{X}$ à travers la fonction $\phi : \mathcal{X} \rightarrow \mathcal{F}$. L'Équation 2.11 donne la transformation mathématique opérée par le noyau k sur les échantillons de données x_n et $x_m \in \mathcal{X}$.

$$k(x_n, x_m) = \langle(\phi(x_n), \phi(x_m)\rangle \tag{2.11}$$

En prenant une décision linéaire basée sur le produit scalaire des variables d'entrée dans l'espace $\mathcal{F}$ de plus grande dimension, le modèle peut prendre une décision non linéaire dans l'espace initial $\mathcal{X}$. L'intérêt principal de l'astuce de noyau est que les modèles n'ont pas besoin de passer par l'espace intermédiaire $\mathcal{F}$ pour calculer son produit scalaire, accélérant grandement les calculs et permettant l'utilisation d'espaces $\mathcal{F}$ de dimension infinie. Il existe de nombreux noyaux différents dont les plus utilisés sont le noyau linéaire (Équation 2.12), le noyau polynomial (Équation 2.12), ainsi que le noyau gaussien (Équation 2.13), aussi connu sous le nom de *radial basis function* (Équation 2.14). Dans ces équations, la constante c contrôlant l'homogénéité du noyau, la pente α, le degré d et σ sont des paramètres à optimiser lors de l'apprentissage du modèle.

$$k(x_n, x_m) = \langle x_n, x_m \rangle + c \tag{2.12}$$

$$k(x_n, x_m) = (\alpha\langle x_n, x_m \rangle + c)^d \tag{2.13}$$

$$k(x_n, x_m) = \exp(-\frac{||x_n - x_m||^2}{2\sigma^2}) \tag{2.14}$$

Machines à vecteurs de support

Les machines à vecteurs de support, *Support Vector Machines* (SVM), aussi connues sous le nom de séparateurs à vastes marges, sont sans doute les modèles utilisant l'astuce du noyau les plus populaires. Ils peuvent être appliqués à la fois aux tâches de classification et aux tâches de régression (SVR). Au lieu de minimiser l'erreur quadratique moyenne comme la plupart des modèles de régression, un modèle SVR optimise l'Équation 2.15 [43]. Le SVR minimise la norme L2 de ses paramètres w avec comme contrainte que l'erreur absolue de prédiction $\left|y_n - w^T x_n\right|$ soit inférieure à la marge ϵ. Tous les échantillons ne pouvant respecter cette contrainte sont ajoutés à l'objectif de minimisation sous la forme d'une pénalité définie par leur déviation ζ_n à la marge ϵ. Le coefficient C sert à pondérer cette pénalité. L'Équation 2.16 représente l'optimisation duale de l'Équation 2.15 (où $\alpha_n, \alpha_n^* \geq 0$ sont les variables duales). En utilisant l'équation duale, nous faisons apparaître le produit scalaire entre deux échantillons

de l'espace $\mathcal{X}$. Nous pouvons utiliser l'astuce du noyau pour remplacer ce produit scalaire, correspondant au noyau linéaire décrit par l'Équation 2.12, par n'importe quel autre noyau. Cela permet ainsi au modèle SVR de faire des prédictions non linéaires.

$$\min \frac{1}{2} ||w|| + C \sum_{n=1}^{N} |\zeta_n| \text{ , avec } \left|y_n - w^T x_n\right| \leq \epsilon \tag{2.15}$$

$$\max_{\alpha,\alpha^*} -\frac{1}{2} \sum_{n,m=1}^{N} (\alpha_n^* - \alpha_n)(\alpha_m^* - \alpha_m)\langle x_n, x_m \rangle - \epsilon \sum_{n=1}^{N} (\alpha_n^* + \alpha_n) + \sum_{n=1}^{N} y_n(\alpha_n^* - \alpha_n) \tag{2.16}$$

Processus gaussiens

Les processus gaussiens, *Gaussian Processes* (GP), sont une classe de modèle utilisant l'astuce du noyau. Un processus gaussien peut être défini comme un ensemble fini de variables aléatoires ayant une distribution conjointe gaussienne [91, 128]. Suivant l'Équation 2.17, un processus gaussien peut être défini par son espérance $E[f(x)]$ et sa matrice de covariance représentée par sa fonction noyau $k(x_n, x_m)$. Les prédictions d'un processus gaussien entraîné sur des échantillons $\{\boldsymbol{x}, \boldsymbol{y}\} = (x_n, y_n), n \in [1, N]$ pour l'échantillon de test non observé $\boldsymbol{x_*}$ sont faites à travers la distribution a posteriori définie par l'Équation 2.18. Dans cette équation, $\boldsymbol{\mu}$ sont le vecteur de moyennes calculées sur le jeu d'entraînement, μ_* la moyenne sur l'échantillon non observé de test, et K la matrice de covariance calculée à partir de la fonction noyau.

$$f(x) \sim \mathcal{GP}(E[f(x)], k(x_n, x_m)) \tag{2.17}$$

$$\begin{bmatrix} \boldsymbol{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(\boldsymbol{x}, \boldsymbol{x}) & K(\boldsymbol{x}, \boldsymbol{x_*}) \\ K(\boldsymbol{x_*}, \boldsymbol{x}) & K(\boldsymbol{x_*}, \boldsymbol{x_*}) \end{bmatrix} \right) \tag{2.18}$$

Filtres adaptatifs

Les noyaux, et en particulier les noyaux gaussiens, peuvent aussi être utilisés au sein de filtres adaptatifs à noyau, *Kernel Adaptive Filters* (KAF). Ces filtres possèdent la particularité d'avoir une fonction de transfert évoluant au cours du temps en fonction de l'évolution du signal. Ces filtres peuvent être utilisés pour des tâches de régression au sein d'un espace de Hilbert à noyau reproduisant [100]. Leur principal intérêt est qu'ils peuvent fonctionner en ligne, en s'adaptant dynamiquement aux changements au sein du signal de glycémie.

Application à la prédiction de la glycémie

Les SVR ont suscité beaucoup d'intérêt chez la communauté de la prédiction de la glycémie en raison de leur facilité de compréhension, d'implémentation ainsi que de leurs résultats. Bunescu *et al.* ont montré que pour des horizons courts et moyens termes (30 et 60 minutes), un modèle SVR est capable d'être plus précis dans

la prédiction de la glycémie que des praticiens experts en diabétologie [14]. Parallèlement, ils ont été étudiés par Georga *et al.* [59, 60, 63, 64] révélant à la fois l'importance de l'ajout de données exogènes au signal de glycémie passé, mais aussi de ses bonnes capacités prédictives. Aujourd'hui, son utilisation continue d'être explorée avec les travaux d'Hamdi *et al.* [69], de Mayo *et al.* [111], et notamment dans le cadre de l'utilisation pratique du modèle sur smartphone [130]. Sa facilité d'implémentation ainsi que ses bons résultats lui permettent d'être souvent utilisé comme modèle de référence [4, 95, 96, 114, 143, 169]. Bien que moins populaires, les processus gaussiens ont tout de même été utilisés pour la prédiction de la glycémie future [148, 60]. Quant aux filtres adaptatifs utilisant les noyaux, ils ont été étudiés dans la prédiction de la glycémie en ligne par Georga *et al.* [62, 63, 64] ainsi que par Yu *et al.* [160, 159]. Enfin, Fiorini *et al.* ont expérimenté une simple régression Ridge en ayant les variables d'entrées transformées à l'aide de noyaux (KRR) [53]. Dans l'ensemble, à l'exception de rares utilisations de noyaux linéaires [14, 114, 111], le noyau gaussien est majoritairement utilisé [59, 60, 69, 130, 62, 63, 64, 159, 160, 148, 96, 169, 53].

2.3.4 Modèles basés sur les réseaux de neurones

Cadre théorique

Les réseaux de neurones artificiels sont inspirés du réseau neuronal biologique constituant notre cerveau. Un neurone, aussi appelé perceptron, calcule la somme pondérée de ses entrées appliquée à une fonction d'activation comme le montre l'Équation 2.19. Dans cette équation, $y \in \mathcal{R}$ est la sortie du neurone, $\boldsymbol{x} \in \mathcal{R}^r$ le vecteur d'entrées au neurone, $\boldsymbol{w} \in \mathcal{R}^r$ les poids de la combinaison linéaire, $b \in \mathcal{R}$ une constante prenant le nom de biais, et la fonction d'activation f. Il existe un grand nombre de fonctions d'activation dont les plus connues sont les fonctions d'activation identité, tangente hyperbolique et ReLU [118]. Un réseau de neurones de type passe-en-avant, *Feedforward Neural Network* (FFNN), est constitué de plusieurs couches de neurones, chaque couche étant composée d'un nombre variable de neurones dont leur sortie est donnée en entrée à la couche suivante. La Figure 2.2 donne un exemple de réseau de neurones à une couche cachée prédisant la glycémie à un horizon PH, $\hat{y}_{t+PH}$, à partir du vecteur des entrées $x_{i,j}$, $j \in [0, r-1]$ représentant le j-ème variable à l'instant $i \in [t-H, t]$ passé.

$$y = f(\boldsymbol{w}^T \boldsymbol{x} + b) \tag{2.19}$$

L'entraînement d'un réseau de neurones artificiels se fait par rétropropagation du gradient d'erreur sur les poids des neurones couche après couche [131]. Le gradient d'erreur est calculé à partir d'une fonction de coût mesurant l'erreur moyenne de prédiction. La fonction de coût utilisée dépend de la tâche en question, les plus connues étant l'entropie croisée pour des tâches de classification, et l'erreur quadratique moyenne, *Mean Squared Error* (MSE), pour des tâches de régression (voir Équation 2.20). Pour faire face aux grandes quantités d'échantillons d'entraînement ralentissant le calcul du gradient d'erreur, l'entraînement se fait généralement de manière itérative sur des sous-ensembles aléatoires du jeu d'entraînement, aussi appelé mini-batchs. Pour être entraîné efficacement

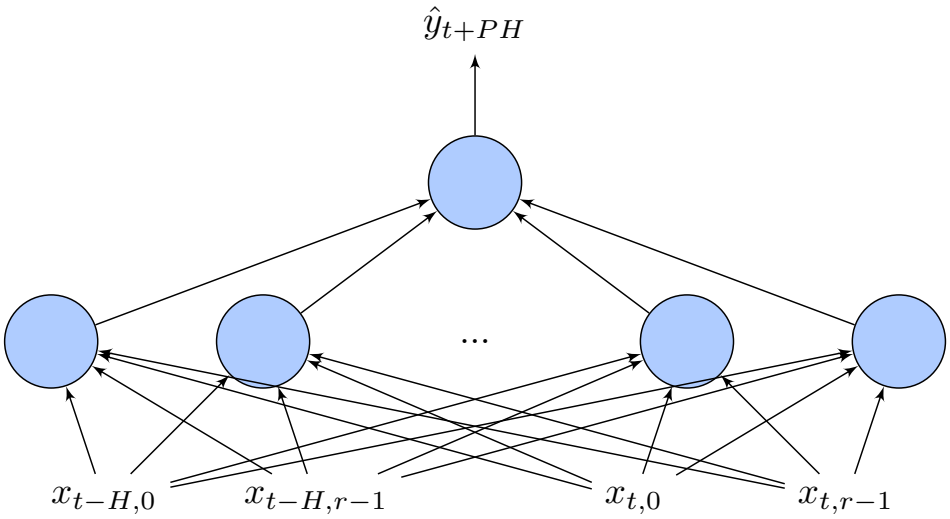


FIGURE 2.2: Exemple de réseau de neurones de type passe en-avant à une couche.

en évitant le surapprentissage aux données d'entraînement, un réseau de neurones a généralement besoin d'un grand nombre de données. Cette contrainte n'étant que rarement satisfaite, il est commun de faire appel à des méthodes de régularisation lors de l'entraînement du modèle. Parmi ces techniques, nous pouvons citer l'utilisation de la norme L2 pénalisant l'utilisation de poids trop importants au sein du réseau, la méthode du *dropout* désactivant une partie du réseau aléatoirement pendant l'entraînement [141], ou bien la méthode d'arrêt anticipé permettant d'arrêter l'entraînement avant que le surapprentissage ne débute.

$$MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 \tag{2.20}$$

Les réseaux de neurones récurrents, *Recurrent Neural Networks* (RNN), sont un type de réseau de neurones artificiels spécifiquement conçu pour les données temporelles et, en particulier, les séries temporelles. Dans un réseau de neurones récurrents, les variables d'entrées sont traitées de manière séquentielle, la sortie à chaque instant étant donnée en entrée à l'instant suivant. La Figure 2.3 représente un réseau de neurones récurrents déroulé sur H instants temporels. L'un des freins principaux à l'utilisation des réseaux de neurones récurrents est le problème de la disparation du gradient empêchant le réseau de retenir les informations les plus anciennes [65]. Ce problème a été résolu grâce à l'utilisation d'unités *Long Short-Term Memory* (LSTM) à la place de neurones traditionnels [71]. Un réseau de neurones récurrents, utilisant des unités LSTM ou non, s'entraîne comme un réseau de neurones artificiels classique par rétropropagation du gradient d'erreur.

Il existe de nombreux autres types de réseaux de neurones pouvant être utilisés pour les tâches de régression. Nous pouvons citer en particulier les réseaux de neurones convolutifs, *Convolutional Neural Networks* (CNN), célèbres pour leur utilisation pour des tâches de reconnaissance et segmentation d'images. Les réseaux convolutifs peuvent être adaptés au traitement de séries temporelles en utilisant des convolutions à une seule dimension, et

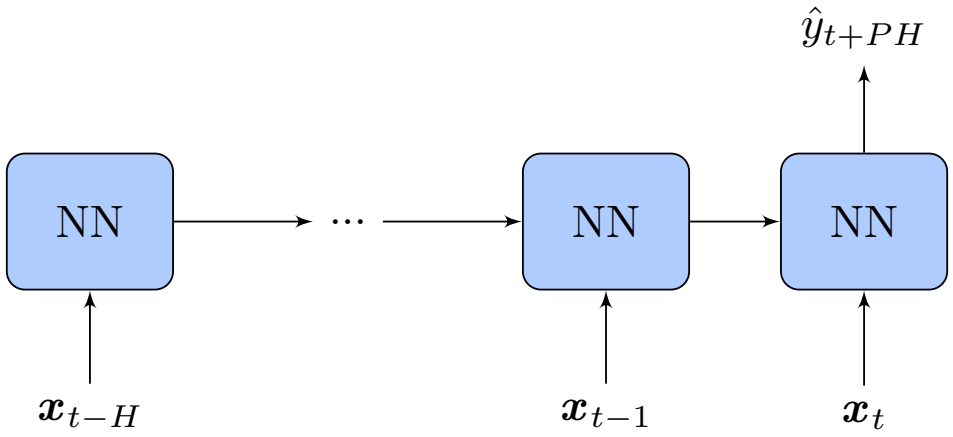


FIGURE 2.3: Exemple de réseau de neurones récurrents déroulé sur H instants temporels.

non deux pour des images. Un autre type de réseau de neurones populaire ces dernières années sont les *Extreme Learning Machines* (ELM) [75], dérivés des *random vector functional link* [121]. Ces réseaux sont particulièrement intéressants, car ils sont très facilement et rapidement entraînables, grâce à leur unique couche de neurones dont les poids sont fixes et aléatoires.

Application à la prédiction de la glycémie

Dans le domaine de la prédiction de la glycémie future de personnes diabétiques, les réseaux de neurones artificiels sont sans doute la classe de modèle la plus utilisée ces dernières années. Les premières approches étudiées ont été les réseaux standards de type passe en avant. Ceux-ci ont montré des résultats prometteurs, notamment en comparaison avec les processus autorégressifs grâce à leur prise de décision non linéaire [125, 122, 164]. Ces résultats sont contrastés par les travaux de Georga *et al.* [60], de Zarkogianni *et al.* [163], ainsi que de Mayo *et al.* [111] montrant des performances plus faibles que d'autres modèles comme le modèle SVR notamment. Toutefois, des variantes des réseaux de neurones standards continuent d'être étudiées pour la prédiction de la glycémie [166, 4, 5].

Les réseaux de neurones récurrents dans leur forme classique n'ont pas été grandement utilisés dans le domaine. Nous pouvons retenir les travaux de Daskalaki *et al.* [24] ainsi que Jankovic *et al.* [77] plaidant en faveur d'une approche ensembliste, combinant les réseaux récurrents à d'autres modèles (e.g., ARX, ELM). Quant à eux, les réseaux récurrents de type LSTM ont eu plus de succès. Diverses études ont montré que les modèles LSTM sont plus performants que des modèles autorégressifs [5], ou SVR utilisant des descripteurs physiologiques experts [114, 143], et qu'ils profitent de l'usage de signaux bruts divers comme la fréquence cardiaque ou la conductivité de la peau [110, 115]. Toutefois, ces résultats prometteurs ne se retrouvent pas dans toutes les études [53, 113].

De leur côté, grâce à leur grande rapidité d'entraînement et simplicité, l'utilisation des modèles ELM a été explorée dans le cadre de modèles prédictifs ensemblistes [77] ou bien de modèles en ligne [61]. Récemment,

différentes architectures utilisant des CNN sont expérimentées, notamment en tant qu'extracteur de caractéristiques pour un réseau LSTM [95], ou comme architecture bout-en-bout implémentant des convolutions causales et dilatées [168, 96]. Enfin, il existe un grand nombre de variantes différentes de réseaux de neurones dont certaines comme les modèles alliant réseaux de neurones et logique floue [163], les cartes auto adaptatives, *Self-Organizing Maps* (SOM) [163], ou les *Echo State Networks* (ESN) [97].

2.3.5 Autres modèles

Mis à part les processus autorégressifs, les arbres de décision, les modèles utilisant l'astuce du noyau, ou les réseaux de neurones, d'autres algorithmes et modèles ont été étudiés pour la prédiction de la glycémie :

— **Filtres récursifs :** Certains filtres récursifs peuvent être utilisés pour faire de la prédiction dans le temps. Ces filtres, comme celui par moyenne mobile (équivalent à un processus MA) [126], par lissage exponentiel, *Exponential Smoothing* (ES) [126], ou par filtre de Kalman, *Kalman Filter* (KF) [154, 53], ont vu quelques utilisations pour la tâche de la prédiction de la glycémie.

— **Régression linéaire, *Linear Regression* (LR) :** Des modèles de régression linéaire simple (avec ou sans régularisation), comparables aux modèles autorégressifs sont parfois utilisés comme modèle de référence [126, 163, 111].

— **Régression par Variables Latentes à données eXogènes (LVX) :** Proposés par Zhao *et al.* pour la prédiction de la glycémie, les modèles LVX ont été utilisées plusieurs fois pour la prédiction de la glycémie [99, 53, 95, 96]. Ces modèles sont des régressions linéaires faites non pas sur les variables d'entrées, mais sur des représentations cachées de celles-ci.

— **Modèles physiologiques (Phys) :** Certains chercheurs ont construit des modèles purement mathématiques décrivant la régulation de la glycémie [142, 99]. Bien que très rarement utilisés à eux seuls pour la prédiction de la glycémie, car assez peu performants [119, 14], ces efforts de modélisation physiologique se retrouvent dans de nombreux descripteurs donnés en entrée aux modèles prédictifs de glycémie [14, 59, 164].

— ***Grammatical Evolution* (GE) :** Contreras *et al.* ont grandement étudié l'utilisation d'algorithmes génétiques à base de grammaire pour la prédiction de la glycémie [21, 20, 150].

— **Modèles ensemblistes :** Plusieurs travaux se tournent vers une approche ensembliste, combinant plusieurs modèles différents et tentant de tirer parti des forces de chacun [77, 160, 10].

— **Modèle de référence naïf (Ref) :** Certaines études utilisent un modèle naïf, prédisant une valeur de glycémie égale à la dernière valeur de glycémie connue, comme modèle de référence [14, 110, 114, 115, 126]. Un modèle possédant une précision inférieure à celui-ci est jugé inutile.

2.4 Évaluation des modèles prédictifs de glycémies

Dans cette section, nous passons en revue la méthodologie générale d'évaluation des modèles prédictifs, ainsi que les variabilités présentes au sein des différentes recherches. Puis, nous détaillons les métriques statistiques et cliniques d'évaluation des performances les plus utilisées dans la littérature.

2.4.1 Méthodologie générale

L'objectif des modèles étudiés est de prédire la valeur de la glycémie des personnes diabétiques PH minutes dans le futur. PH représente l'horizon de prédiction et varie de 15 minutes à plusieurs heures. La Figure 2.4 donne un aperçu de la distribution des horizons de prédiction sur les études. Nous pouvons y voir que les horizons de prédictions courts et moyens termes de 30 et 60 minutes sont les plus représentés avec respectivement 86% et 58% des études. Au sein des horizons plus courts, seul l'horizon de 15 minutes semble présenter un intérêt pour la communauté. En effet, en comparaison avec les horizons proches (e.g., 10 ou 20 minutes), celui de 15 minutes fonctionne avec la plupart des capteurs de glycémie en continu, ayant une fréquence d'échantillonnage de 5 ou 15 minutes. Quant aux horizons plus longs-termes, ils sont globalement assez peu représentés (celui de 120 minutes l'étant le plus). Cela s'explique par la grande difficulté de prédiction à des horizons lointains. En effet, plus l'horizon est grand, plus il y a des chances qu'un évènement important pour la régulation de la glycémie (repas, prise d'insuline, sport) survienne sous qu'il puisse être anticipé par le modèle.

Pour tenir compte de la grande inter/intra variabilité de la population diabétique, la plupart des études (76%) évaluent des modèles prédictifs qui ont été personnalisés aux patients diabétiques. À l'exception des travaux de Zhu *et al.* [168, 169], tous les modèles évalués sur un patient ont été entraînés exclusivement avec les données de ce même patient. Les résultats reportés dans les études sont moyennés sur l'ensemble de la population diabétique étudiée. Bien que moins utilisés (24%), les modèles globaux peuvent être intéressants lorsqu'il n'y a pas assez de données individuelles à disposition [5, 116, 122] ou dans le cadre de l'apprentissage par transfert [169, 115, 143].

Pour permettre une analyse des performances des modèles qui soit non biaisée, les modèles doivent être évalués sur des données qui n'ont pas été utilisées pendant l'entraînement. Ces données portent le nom de *jeu de test*, en opposition au *jeu d'entraînement* utilisé pour l'entraînement des modèles. Pour les modèles globaux, les études reportent utiliser des *patients de test*. Pour les modèles personnalisés, les jeux de tests sont constitués des derniers jours (nombre variable) du patient en question. Contrairement à beaucoup de tâches d'apprentissage automatique, l'évaluation de séries temporelles interdit l'usage de la validation croisée. Consistant en la permutation des jours de test et d'entraînement, elle implique l'évaluation des modèles sur des données du passé pour lesquelles les données d'entraînement donnent des informations. Toutefois, quelques études mentionnent l'utilisation d'un *jeu de validation* [110, 114, 115, 130, 168, 169]. Celui-ci est une fraction du jeu d'entraînement non utilisé pendant l'entraînement et sert d'évaluation intermédiaire pour optimiser les paramètres des modèles. Ces études reportent

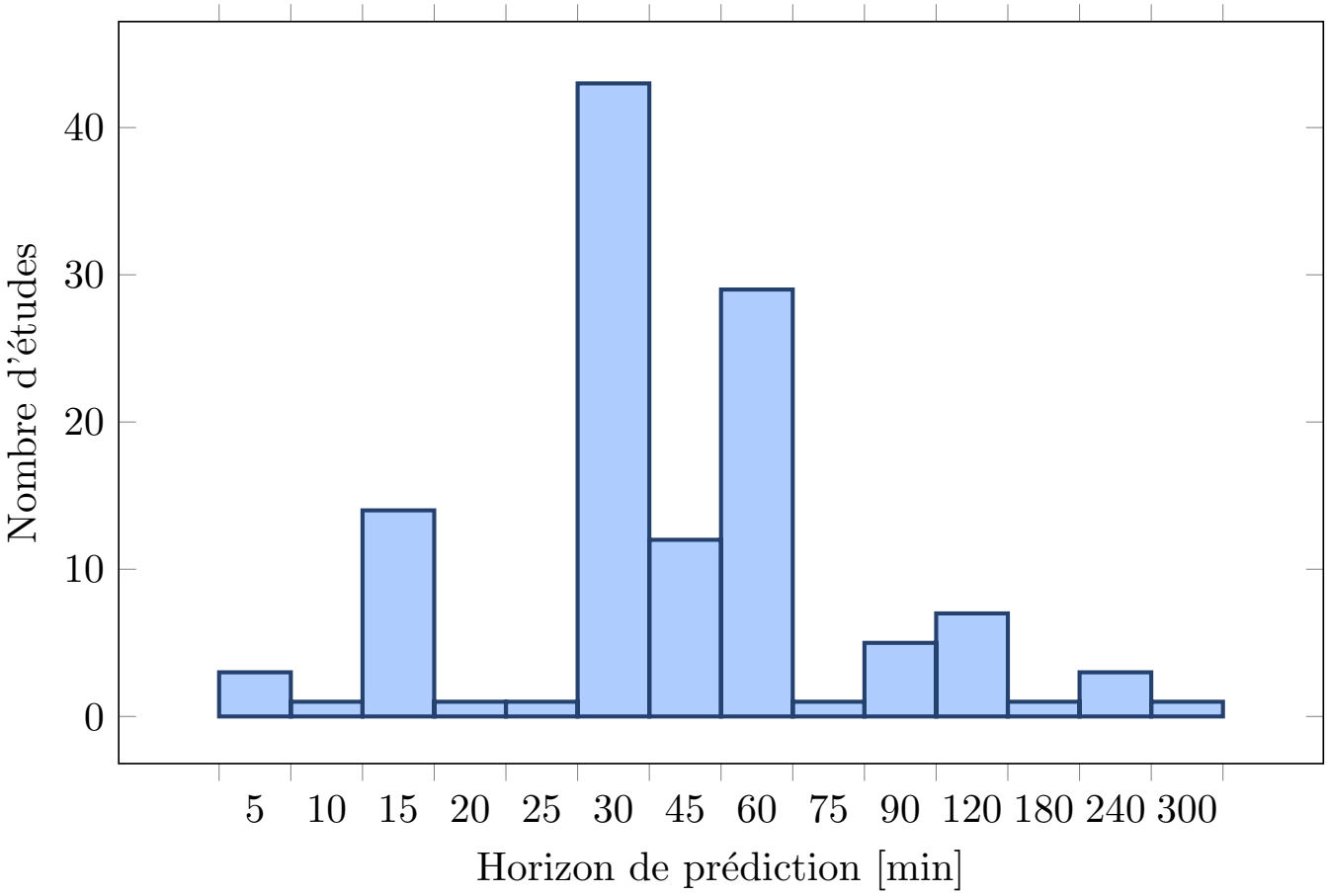


FIGURE 2.4: Histogramme des horizons de prédictions utilisés dans l'état de l'art.

l'utilisation de la validation croisée dans le cadre de la permutation des jours d'entraînement et de validation. Enfin, l'étude de Phadke *et al.* se démarque des autres par l'utilisation de fenêtres temporelles d'apprentissage permettant à la quasi-totalité des données d'être utilisés pour l'évaluation des modèles [126]. Cette méthode a toutefois la limitation de réduire considérablement les données utilisées pour l'entraînement des modèles.

2.4.2 Métriques d'évaluation

Un grand nombre de métriques ont été utilisées pour l'évaluation des modèles prédictifs. La Figure 2.5 représente la distribution des plus importantes d'entre elles. Nous pouvons différencier deux grandes catégories : les métriques statistiques et les métriques cliniques.

Métriques statistiques

Les métriques statistiques sont des mesures générales, non spécialisées au domaine, des performances des modèles. Les plus utilisées sont la RMSE, la MAPE (ou MARD), les coefficients r et R^2, ou le TG/TL :

— **Racine carrée de l'erreur quadratique moyenne, *Root Mean Squared Error* (RMSE) :** utilisée par 80%

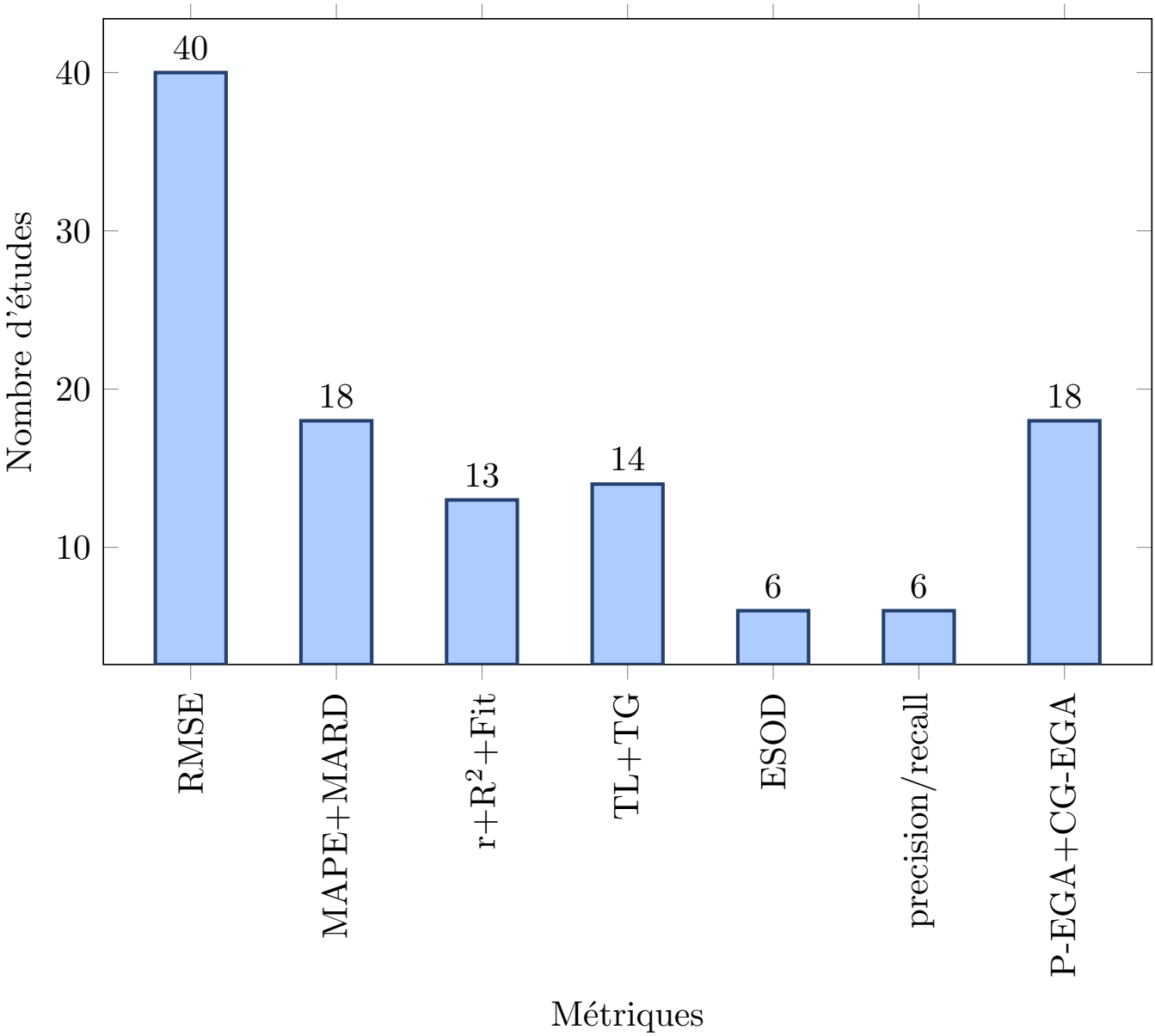


FIGURE 2.5: Histogramme des métriques utilisées dans l'état de l'art.

des études, calculée comme l'erreur quadratique moyenne (voir Équation 2.21), la RMSE mesure la précision moyenne des prédictions. En comparaison avec d'autres mesures comme la l'erreur absolue moyenne (MAE), elle pénalise plus fortement les erreurs de prédiction à forte magnitude.

$$RMSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} (y_n - \hat{y}_n)^2} \tag{2.21}$$

— **Erreur absolue moyenne en pourcentages, *Mean Absolute Percentage Error* (MAPE) :** Aussi utilisée sous le nom de deviation absolue relative moyenne, *Mean Absolute Relative Deviation* (MARD), la MAPE donne, comme la RMSE, une mesure moyenne de la précision des prédictions. En comparaison avec la

RMSE, la mesure de la MAPE est relative, ne dépendant pas de l'échelle des prédictions. Cette particularité est intéressante, considérant que les patients diabétiques peuvent avoir des glycémies très différentes en moyenne. De plus, elle s'exprime en pourcentage, facilitant son interprétation. Enfin, pour des mesures de glycémie (e.g., par acrshortcgm), une MARD (MAPE) $< 10\%$ est généralement regardée comme acceptable [23].

$$MAPE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{100}{N} \cdot \sum_{n=1}^{N} \left| \frac{y_n - \hat{y}_n}{y_n} \right| \tag{2.22}$$

— **Coefficient de corrélation (r) ou de détermination (R^2) ou Fit :** Le coefficient de corrélation r mesure, entre -1 et 1, la corrélation entre les prédictions et les observations de glycémie. Tandis qu'une corrélation de 1 ou -1 indique une parfaite corrélation positive ou négative, une corrélation de 0 indique que les deux variables ne sont pas corrélées. Le coefficient de détermination R^2, aussi appelé Fit, mesure entre 0 et 1 la qualité de la régression. Une valeur de 1 indique que les valeurs prédites par le modèle égalent parfaitement les valeurs observées.

$$r(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\sigma_{\boldsymbol{y}\hat{\boldsymbol{y}}}}{\sigma_{\boldsymbol{y}}\sigma_{\hat{\boldsymbol{y}}}} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \overline{y}_n)(\hat{y}_n - \overline{\hat{y}}_n)}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \overline{y}_n)^2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - \overline{\hat{y}}_n)^2}} \tag{2.23}$$

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = r(\boldsymbol{y}, \hat{\boldsymbol{y}})^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \overline{y}_n)^2} \tag{2.24}$$

— **Retard temporel ou gain temporel, *Time Gain* (TG) :** Le retard temporel mesure en minutes le retard temporel entre le signal de glycémie prédit et celui qui est observé. Il peut être calculé soit comme le déphasage du signal prédit maximisant la corrélation entre le signal prédit et le signal observé (Équation 2.25a), ou comme celui minimisant l'erreur quadratique moyenne de prédiction (Équation 2.25b). Le gain temporel TG, quant à lui, quantifie le nombre de minutes gagnées par le patient par anticipation grâce aux prédictions. Il se calcule simplement comme la différence entre l'horizon de prédiction PH et le retard temporel TL (voir Équation 2.26).

$$TL(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \underset{i \in [0, PH]}{\operatorname{argmax}}\, r(\boldsymbol{y}_{1\ldots N-i}, \hat{\boldsymbol{y}}_{1+i\ldots N}) \tag{2.25a}$$

$$TL(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \underset{i \in [0, PH]}{\operatorname{argmin}}\, \frac{1}{N-i}\sum_{n=1}^{N-i}(y_n - \hat{y}_{n+i})^2 \tag{2.25b}$$

$$TG(\boldsymbol{y}, \hat{\boldsymbol{y}}, PH) = PH - TL(\boldsymbol{y}, \hat{\boldsymbol{y}}) \tag{2.26}$$

— ***Energy of the Second Order Differences*** **(ESOD) :** L'ESOD caractérise les oscillations présentes dans le signal prédit. Plus l'ESOD est élevée, plus il présente des oscillations importantes. Ces oscillations peuvent être dangereuses pour la personne diabétique, pour laquelle la stabilité de la glycémie prédite est importante. En effet, si la glycémie prédite présente trop d'oscillations, la personne diabétique aura du mal à lire ces prédictions, identifier les tendances générales, et prendre des actions suite à celles-ci. Selon l'Équation 2.27, l'ESOD se calcule comme la somme normalisée de l'accélération quadratique du signal de glycémie prédit.

$$ESOD(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\sum_{n=3}^{N} \Delta^2 \hat{y}_n^2}{\sum_{n=3}^{N} \Delta^2 y_n^2} \tag{2.27}$$

Métriques cliniques

Les métriques cliniques sont des mesures qui sont spécifiques au domaine de la prédiction de la glycémie. Elles ont pour objectif de répondre certains problèmes spécifiques du domaine. Dans cette section, nous détaillons la gRMSE, l'indice J, la P-EGA et la CG-EGA :

— ***Glucose Root Mean Squared Error*** **(gRMSE) :** La gRMSE est une modification de la RMSE par Del Favero *et al.* [40] visant à amplifier les erreurs de prédictions dans les zones d'hypo et d'hyperglycémie. En effet, bien que ces erreurs soient particulièrement graves pour les personnes diabétiques, elles ne sont pas forcément bien prises en compte par la RMSE à cause du faible nombre d'échantillons dans ces régions. De plus, les erreurs en région d'hypoglycémie sont plus faibles en magnitude que les erreurs en région d'hyperglycémie, réduisant davantage leur impact sur les performances finales. Pour résoudre ces problèmes, la gRMSE donne un poids plus important aux erreurs commises dans ces régions glycémiques. Lorsque l'observation de glycémie est inférieure à 70 mg/dL (hypoglycémie), l'erreur de prédiction est multipliée par 2.5. À l'inverse, lorsque l'observation de glycémie est supérieure à 180 mg/dL (hyperglycémie), l'erreur de prédiction est multipliée par 1.5. L'inclusion de cette pénalité supplémentaire peut aussi être faite sur les autres métriques statistiques du domaine comme la MARD (MAPE), devenant ainsi la gMARD. Bien qu'intéressantes, nous notons toutefois que ces métriques n'ont pas été beaucoup utilisées dans le domaine. Dans notre état de l'art, nous n'enregistrons que 2 études provenant des mêmes auteurs [20, 21].

— **Indice J :** Proposé par Facchinetti *et al.*, l'indice J combine les mesures d'ESOD et de gain temporel TG [48]. Calculé selon l'Équation 2.28, l'indice J mesure la régularité du profile de prédiction ainsi que le temps gagné à travers le mécanisme de prédiction. D'après les auteurs, l'indice J peut tout aussi bien s'utiliser comme métrique d'évaluation ou comme métrique à optimiser pendant l'entraînement des modèles. Toutefois, comme pour la gRMSE, à ce jour, seulement 3 des études répertoriées l'ont utilisée [164, 64, 154].

$$J(\boldsymbol{y}, \hat{\boldsymbol{y}}, PH) = PH \cdot \frac{ESOD(\boldsymbol{y}, \hat{\boldsymbol{y}})}{TG(\boldsymbol{y}, \hat{\boldsymbol{y}}, PH)} \tag{2.28}$$

— ***Point-Error Grid Analysis* (P-EGA) :** Connue aussi sous le nom de *Clarke Error-Grid Analysis*, la P-EGA est la métrique clinique la plus répandue dans le domaine de la prédiction de la glycémie. Celle-ci mesure la précision clinique des prédictions de glycémie en les catégorisant selon 5 zones d'acceptabilité : A, B, C, D, E. Tandis que la zone A indique que la prédiction est suffisamment cliniquement précise, la zone E indique que la prédiction est très dangereuse pour la personne diabétique. Les scores sont attribués à travers la segmentation de l'espace en deux dimensions représentées par les observations et les prédictions de glycémie. La Figure 2.6 donne l'attribution des scores de précision clinique par la P-EGA. Pour évaluation des modèles, il est courant d'utiliser le pourcentage de prédictions se trouvant dans les zones A ou B (A+B).

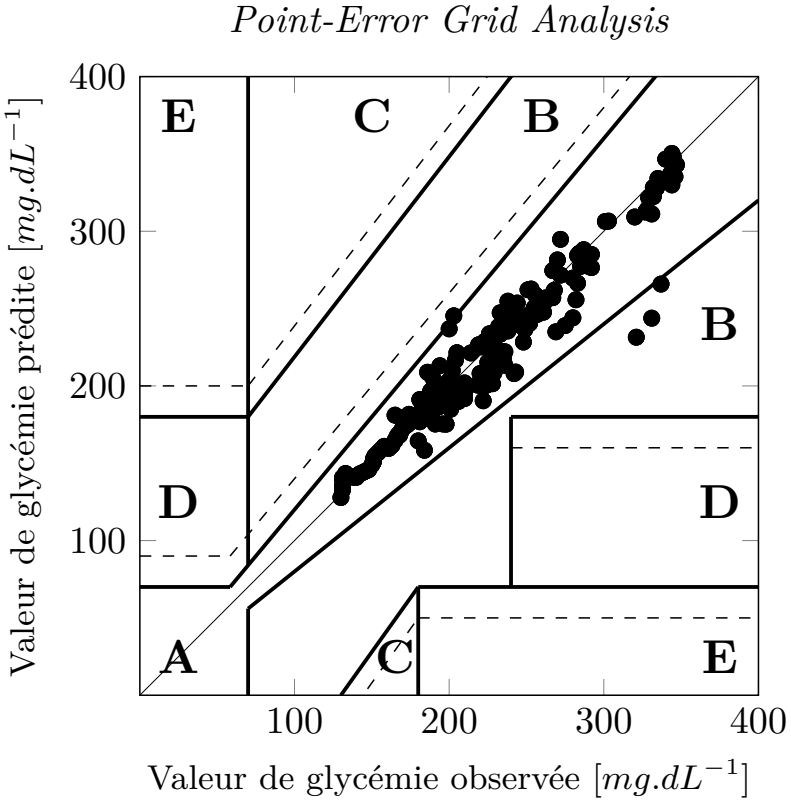


FIGURE 2.6: Exemple d'utilisation de la P-EGA pour estimer la précision clinique des prédictions de glycémie.

— ***Continuous Glucose-Error Grid Analysis* (CG-EGA) :** Originellement proposée par Kovatchev *et al.* pour l'évaluation de l'acceptabilité clinique des capteurs de glycémie [85], la CG-EGA est une extension de la P-EGA. Dans la CG-EGA, cette dernière est combinée avec une seconde grille, la *Rate-Error Grid Analysis* (R-EGA), analysant la précision clinique des variations prédites de glycémie. En effet, une prédiction peut être dangereuse non seulement si elle n'est pas précise, mais aussi si la variation prédite depuis la dernière prédiction n'est pas représentative de la vraie variation. La Figure 2.7 donne une représentation graphique de l'utilisation de la R-EGA. Celle-ci fonctionne comme la P-EGA, en attribuant un score aux prédictions en fonction de l'erreur de la variation prédite. Celle-ci est calculée comme la différence entre une prédiction et

la prédiction précédente, divisée par l'intervalle de temps en minutes entre chaque prédiction. La CG-EGA combine les résultats des deux grilles en attribuant à chaque prédiction le label d'*Accurate Prediction* (AP), de *Benign Error* (BE) ou de *Erroneous Prediction* (EP). Alors que les prédictions AP sont globalement des prédictions sans danger pour le patient, les prédictions BE sont des erreurs bénignes et les prédictions EP sont des erreurs mettant en danger le patient. Pour qu'un modèle soit cliniquement acceptable, il doit avoir un taux d'AP élevé ainsi qu'un taux d'EP faible.

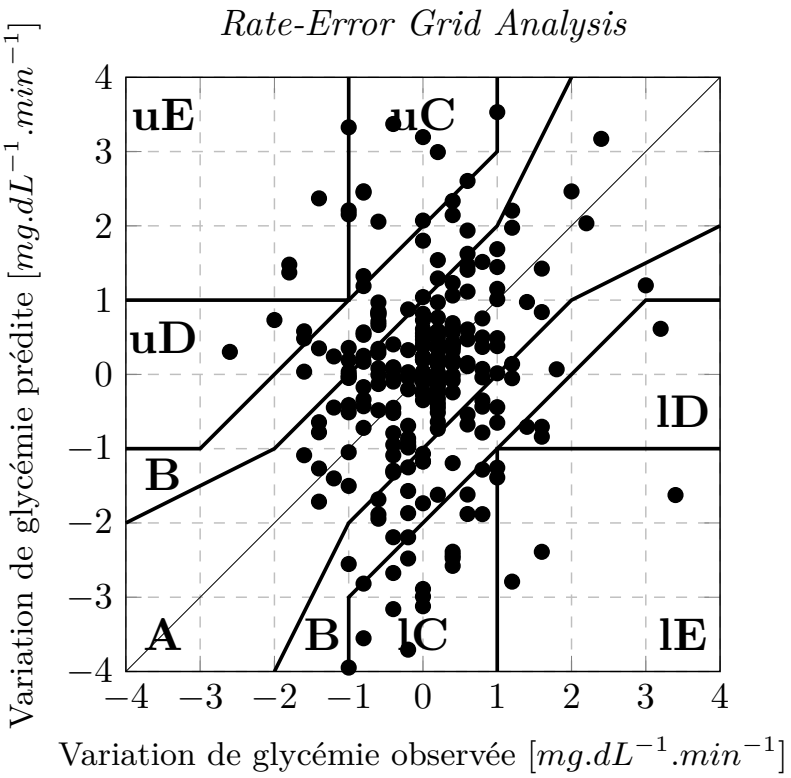


FIGURE 2.7: Exemple d'utilisation de la R-EGA pour estimer la précision clinique des prédictions de glycémie.

2.5 Synthèse des résultats

Les premiers travaux visant à construire des modèles prédisant la glycémie future de personnes diabétiques se focalisent sur l'utilisation de processus autorégressifs. En 2007, Sparacino *et al.* montrent qu'un modèle AR de faible ordre est capable de faire des prédictions permettant d'anticiper les hypoglycémies des patients 20 à 25 minutes en avance [140]. Ces résultats sont confirmés sur d'autres patients par Gani *et al.* [56]. En particulier, ils attribuent cette réussite au lissage du signal de glycémie qui comporte un bruit trop important dans sa forme brute. Toutefois, comme ces derniers le soulèvent, le lissage des prédictions, tel qu'il a été fait dans ces deux études, est incompatible avec une utilisation réelle, en ligne. En effet, ces filtres ne sont pas causals, car ils utilisent les données du futur. Toujours avec des processus autorégressifs, en 2012, les travaux de Zhao *et al.* et de Eren-

		P-EGA										
		Hypoglycémie			*Euglycémie*			*Hyperglycémie*				
		A	D	E	A	B	C	A	B	C	D	E
R-EGA	A	AP	EP	EP	AP	AP	EP	AP	AP	EP	EP	EP
	B	AP	EP	EP	AP	AP	EP	AP	AP	EP	EP	EP
	uC	BE	EP	EP	BE	BE	EP	BE	BE	EP	EP	EP
	lC	BE	EP	EP	BE	BE	EP	BE	BE	EP	EP	EP
	uD	EP	EP	EP	BE	BE	EP	BE	BE	EP	EP	EP
	lD	BE	EP	EP	BE	BE	EP	EP	EP	EP	EP	EP
	uE	EP	EP	EP	EP	EP	EP	EP	EP	EP	EP	EP
	lE	BE	EP	EP	EP	EP	EP	EP	EP	EP	EP	EP

AP : *Accurate Prediction* ; BE : *Benign Error* ; EP : *Erroneous Prediction*

Tableau 2.2: Classification des prédictions de glycémie opérée par la CG-EGA. En fonction des scores obtenus sur la P-EGA et la R-EGA, une prédiction est classifiée comme étant une prédiction cliniquement précise (AP), une erreur benigne (BE) ou une erreur dangereuse (EP).

Oruklu *et al.* suggèrent qu'ajouter des informations supplémentaires comme les prises d'insuline et de glucides [167] ou concernant l'activité physique [45] permet d'améliorer les prédictions.

Suite à ces travaux, de nombreux efforts ont été faits pour identifier les données d'entrées les plus adaptées pour la prédiction de la glycémie :

— L'intérêt de l'ajout d'informations concernant les repas a été confirmé par les travaux de Zecchin *et al.* utilisant des réseaux de neurones FFNN [164, 166] et de Midroni *et al.* utilisant des modèles GBM [113]. Dans leurs travaux, Montaser *et al.* ont étudié le caractère saisonnier des excursions de glycémie postprandiales (i.e., suite à un repas) [116]. Ils montrent en particulier que lorsque les repas sont à heure et quantité de glucides fixes, il est possible de prédire la glycémie du patient à un horizon bien plus lointain (jusqu'à 5h). Bien que les contraintes imposées soient incompatibles avec une utilisation quotidienne d'un tel dispositif, leurs travaux montrent néanmoins l'importance des repas dans la modélisation de la glycémie future des patients.

— Quant à l'ajout d'informations liées à l'activité physique du patient, les résultats sont plus contrastés. Jankovic *et al.* reportent un gain en performance lié à leur utilisation [77], utilisation qui serait particulièrement bénéfique en région d'hypoglycémie selon les travaux de Zarkogianni *et al.* [163]. Cependant, en utilisant un modèle GBM permettant d'évaluer l'importance des variables d'entrées pour la prédiction, les travaux de Midroni *et al.* montrent à la fois que les modèles donnent beaucoup d'importance aux variables décrivant l'activité physique du patient, mais que ces modèles n'arrivent pas à avoir une meilleure précision que les modèles ne les utilisant pas [113]. Enfin, bien que Mirshkarian *et al.* reportent un gain en performances en utilisant des informations comme la fréquence cardiaque ou la conductivité de la peau sur un modèle LSTM, nous notons que ces améliorations sont faibles [115]. Ainsi, l'intérêt de l'utilisation d'informations décrivant l'activité physique du patient est incertain.

— Comme pour l'activité physique, l'intérêt de l'utilisation de descripteurs physiologiques est contrasté dans les

différentes études. Bunescu *et al.* et Georga *et al.* montrent qu'un modèle SVR utilisant des descripteurs physiologiques modélisant les dynamiques d'insuline et de glucides, ainsi que les dépenses énergétiques des patients est capable d'effectuer de très bonnes prédictions [14, 59]. Toutefois, Mirshekarian *et al.* montrent qu'un modèle LSTM n'utilisant pas ces connaissances expertes est capable d'obtenir de meilleures performances [114]. En effet, à condition d'utiliser suffisamment de données d'entraînement, les réseaux de neurones ont la capacité d'apprendre à extraire les informations pertinentes à partir des données brutes sans intervention humaine.

Néanmoins, les principaux efforts ont été portés sur la recherche de modèles prédictifs plus performants. Dans cet état de l'art, nous reportons un très grand nombre de modèles prédictifs différents, allant des modèles autorégressifs historiques, à des modèles plus complexes relevant de l'apprentissage profond. Dans l'ensemble, il est difficile de déterminer lequel des modèles est le plus adapté à la tâche de prédiction de glycémie, chaque étude reportant souvent des conclusions différentes. De plus, il est impossible de comparer les résultats de chaque étude entre eux, car ces dernières utilisent des ensembles de données différents. Toutefois, nous pouvons faire les analyses suivantes :

— Les modèles autorégressifs utilisés historiquement ont montré être surpassés par de nombreux autres modèles caractérisés par leur complexité accrue, en particulier par des SVR [14], réseaux de neurones FFNN [125, 164, 4], réseaux de neurones récurrents [5, 114, 115, 143, 169] ou convolutifs [95, 96].

— La publication du jeu de données OhioT1DM nous permet de comparer les publications l'utilisant pour l'évaluation de leurs modèles. Sur un total de 12 études, seules 7 peuvent être rigoureusement comparées, car celles-ci utilisent strictement le même ensemble de test ainsi qu'une métrique d'évaluation commune (RMSE). Au sein des modèles présentés, les réseaux de neurones récurrents de Mirshekarian *et al.* [115] et de [169] montrent avoir la meilleur précision. Ils sont suivis par les modèles GBM de Jeon *et al.* [78], les réseaux de neurones FFNN de Bertachi *et al.* [10], et les réseaux de neurones convolutifs de Li *et al.* [96]. De leur côté, les modèles SVR [96, 169], autorégressifs [96, 169, 115], ou utilisant les modèles d'évolution grammaticale GE [21] ne semblent pas donner de résultats satisfaisants. Globalement ces résultats plaident en faveur de l'apprentissage profond pour la prédiction de la glycémie. Nous notons que ce qui démarque les réseaux de neurones performants des autres moins performants est, outre la nature du modèle, l'utilisation de l'apprentissage par transfert. Dans le cadre de l'entraînement d'un modèle personnalisé au patient, l'apprentissage par transfert permet de réutiliser les connaissances apprises sur plusieurs patients pour faciliter l'apprentissage du modèle sur un nouveau patient.

— Dans leurs travaux Mayo *et al.* montrent que les performances relatives des modèles sont très variables en fonction des différentes régions glycémiques (hypoglycémie, euglycémie, hyperglycémie) [111]. Ces régions n'étant pas représentées équitablement au sein des sous-ensembles de test, le meilleur modèle général n'est

pas nécessairement le meilleur modèle dans toutes les régions individuelles. En particulier, dans leur étude, tandis que le modèle SVR est le meilleur dans les régions d'euglycémie ou d'hyperglycémie, c'est le modèle FFNN qui est le plus adapté aux prédictions en hypoglycémie. Ce constat a amené plusieurs chercheurs à explorer des méthodes ensemblistes pour la prédiction de la glycémie. Dans leurs travaux, Vehí *et al.* proposent de combiner l'utilisation d'un modèle basé sur l'évolution grammaticale GE pour la prédiction à moyen terme de la glycémie, un modèle SVM pour la détection d'hypoglycémie postprandiale, et un réseau de neurones FFNN pour la détection d'hypoglycémie nocturne [150]. Quant à eux, Bertachi *et al.* proposent d'utiliser plusieurs réseaux FFNN, chacun spécialisé dans une zone région glycémique spécifique [10].

— Certains travaux permettent de soulever la disparité entre les mesures statistiques utilisées dans l'entraînement des modèles prédictifs de glycémie et leurs objectifs cliniques. Dans leurs travaux, plutôt que de prédire la glycémie future des patients, Martinsson *et al.* proposent d'entraîner les modèles à paramétriser une distribution gaussienne centrée sur la prédiction afin d'estimer l'incertitude de la prédiction [110]. Quant à eux, Vehi *et al.* proposent d'utiliser une fonction de coût donnant plus d'importance aux régions d'hypoglycémie et d'hyperglycémie [150].

2.6 Analyse critique et perspectives

2.6.1 Analyse des performances relatives des modèles prédictifs

Comme nous venons de le voir dans ce chapitre, l'état de l'art de la prédiction de la glycémie est très fourni. La Figure 2.8, représentant la distribution des études analysées sur les années, nous montre que le domaine de la prédiction de la glycémie suscite de plus en plus d'intérêt. Cela s'explique par la démocratisation du logiciel de simulation T1DMS ces dernières années, avec plus de 70% des publications datant d'après 2017, ainsi que par la mise à disposition du jeu de données OhioT1DM en 2018. Ceux-ci permettent notamment à des chercheurs extérieurs au domaine de participer aux efforts bien plus facilement que s'ils devaient récolter eux-mêmes les données. Du point de vue de cette thèse, nous notons que la plupart des études analysées dans l'état de l'art, tout comme le jeu de données OhioT1DM, n'existaient pas lors de son commencement en octobre 2017.

Cependant, il est aujourd'hui difficile d'évaluer l'état de la littérature, d'identifier les meilleurs modèles, données ou étapes de prétraitement. En effet, aujourd'hui, les études traitant de la prédiction de la glycémie sont grandement hétérogènes :

— Tout d'abord, bien que cela est en train de changer avec le jeu de données OhioT1DM, très peu d'études utilisent les mêmes données, rendant les comparaisons entre études délicates, voire impossibles. De plus, en étant utilisées en petites quantités (faible nombre de patients, petites quantités de données par patient), les résultats ne peuvent refléter l'intégralité de la population diabétique caractérisée par sa grande variabilité.

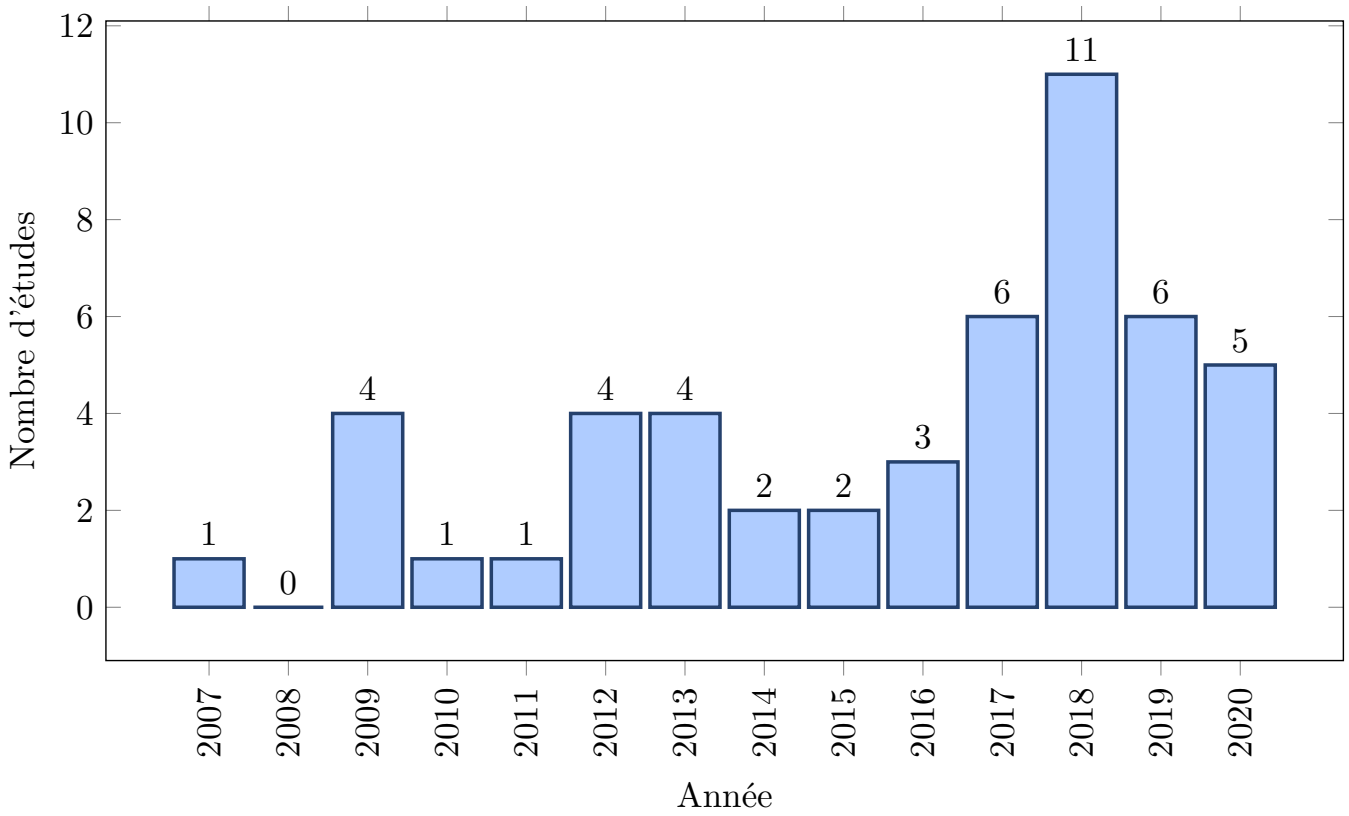


FIGURE 2.8: Histogramme représentant le nombre d'études traitant de la prédiction de la glycémie par année. L'année 2020 étant en cours, le nombre affiché n'est pas représentatif du nombre futur d'études publiées sur cette année.

— L'hétérogénéité dans les études se retrouve aussi dans le processus de traitement des données. Les études n'utilisent très souvent pas des données de même nature (insuline, glucides, activité physique, etc.), des étapes de prétraitement identiques (nettoyage des données, extractions de descripteurs), ou des procédés d'évaluation similaires (horizons de prédictions, métriques).

— Au sein d'une même étude, pour laquelle les données, prétraitements et évaluations sont identiques, il reste toutefois difficile de tirer des conclusions. En effet, certaines performances particulièrement bonnes ou mauvaises, ainsi qu'un manque de détails sur les procédés généraux du traitement des données, peuvent soulever des doutes quant à la pertinence des résultats. Toutefois, cela n'est pas le cas pour toutes les études, celles de Jeon *et al.* [78], de Georga *et al.* [64], de Mayo *et al.* [111], ainsi que de Zhu *et al.* [169] proposant des analyses détaillées des performances obtenues pour différents modèles, données ou étapes de prétraitement.

De manière générale, la communauté de la prédiction de la glycémie gagnerait à unifier ses procédés d'évaluation à travers l'utilisation de données et de procédés de traitement identiques. L'existence du logiciel de simulation T1DMS et du jeu OhioT1DM présentent l'opportunité de réunir les chercheurs autour d'un même environnement facilitant et accélérant la recherche de meilleures solutions. Pour répondre à ces challenges, le Chapitre 4 introduit le benchmark GLYFE proposant un procédé de traitement et une base de résultats de référence pour l'évaluation

des modèles prédictifs de glycémie.

2.6.2 Limitations cliniques de l'état de l'art

Du point de vue clinique, l'état de l'art montre posséder plusieurs limitations et opportunités :

— Tout d'abord, la population diabétique de type 2 n'est presque pas étudiée. Pourtant les personnes diabétiques de type 2 représentent près de 90% de la population diabétique [49]. Bien que cela peut s'expliquer par une recherche historiquement plus centrée sur le type 1, lequel présente moins de variabilité individuelle, la prédiction de la glycémie serait tout autant utile pour les patients diabétiques de type 2. De plus, la personnalisation des modèles au patient, grâce aux données accessibles en plus grandes quantités, permettrait de gérer efficacement la grande inter variabilité de la population diabétique de type 2.

— Dans l'ensemble, les modèles sont évalués soit sur peu de patients différents, soit sur un nombre faible de jours. Cela s'explique par la difficulté de récolter des données autres que la glycémie. Tandis que les données de glucides ou d'insuline doivent être manuellement notées par le patient, les données propres à l'activité physique doivent être récoltées via l'utilisation de bracelets d'activité physique. Quant aux données événementielles ou d'humeur, leur très faible occurrence au quotidien demande à qu'elles soient collectées sur de très longues durées. À cause de ces difficultés, il est aujourd'hui difficile de juger de l'intérêt d'utiliser de telles données pour la prédiction de la glycémie. Seules celles de la glycémie, d'insuline ou des glucides ont aujourd'hui un réel impact avéré.

— Quant aux modèles prédictifs, la très grande majorité n'intègre des critères cliniques que pendant l'évaluation des modèles, et non pendant l'apprentissage de ceux-ci. Bien que certains groupes de chercheurs aient proposé des méthodes pour l'inclusion de tels critères (e.g., gMSE [40], indice J [48]), ils ne sont presque pas utilisés dans la littérature. Pourtant, la sécurité des patients utilisant un dispositif prédisant la glycémie future est primordiale. En effet, face à une prédiction, le patient est seul pour décider ou non d'en tenir compte dans ses actions futures (e.g., dois-je prendre de l'insuline, car le modèle prédit une forte glycémie dans 30 minutes ? Dois-je faire mon sport sachant qu'une hypoglycémie est probable dans les minutes à venir ?). Les modèles gagneraient à inclure des contraintes d'acceptabilité clinique dans leur apprentissage, de telle sorte à pouvoir ajuster le risque induit par les prédictions.

— Ce risque induit par les prédictions est amplifié par la complexification des modèles prédictifs ces dernières années. En effet, bien que les modèles autorégressifs historiques soient relativement interprétables, ceux-ci sont peu à peu délaissés en faveur de modèles complexes d'apprentissage automatique (e.g., SVR), ou d'apprentissage profond (e.g., LSTM, CNN). L'interprétabilité des modèles est un aspect particulièrement important, car elle permet d'obtenir la compréhension du mécanisme prédictif, et ainsi la confiance dans la prédiction. L'interprétabilité des modèles prédictifs n'est pas un enjeu spécifique à la prédiction de la glycémie

ou à la santé. Ainsi, beaucoup de travaux ont été effectués ces dernières années pour rendre les modèles plus compréhensibles et interprétables. Parmi ces innovations, nous relevons, par exemple, le principe d'attention permettant de comprendre sur quelles variables les modèles profonds se focalisent pour faire leur prédiction.

Dans cette thèse, nous avons comme objectif d'aborder ces différentes limitations. Tout d'abord, nous proposons de faire nos évaluations à la fois sur des personnes diabétiques de type 1 et de type 2. Pour permettre une évaluation sur le type 2, nous avons procédé à une campagne de collecte de données très diverses (glycémie, glucides, insuline, activité physique, sommeil, humeur). Cette campagne de collecte de données est détaillée dans le Chapitre 3. Afin d'améliorer les performances cliniques des modèles prédictifs, et en particulier des modèles profonds, nous proposons dans le Chapitre 5 une méthodologie permettant d'obtenir un compromis optimal entre précision des modèles et acceptabilité clinique. Pour réduire la quantité de données nécessaires pour l'apprentissage des modèles, nous développons dans le Chapitre 6 une approche pour transférer des connaissances générales a priori apprises sur plusieurs patients vers un nouveau patient. Enfin, la thèse utilisant majoritairement des modèles profonds non interprétables, nous étudions dans le Chapitre 7, une architecture interprétable basée sur des réseaux de neurones récurrents et le principe d'attention.

3 | Données expérimentales

Sommaire

3.1 Introduction

La construction de modèles prédictifs de glycémie basés sur l'apprentissage automatique n'est possible que si une quantité suffisante de données est utilisée pour l'entraînement des modèles. Cette quantité de données se doit d'être d'autant plus grande dans le cadre de l'apprentissage profond. En outre, afin d'assurer un entraînement et une évaluation optimale, ces données doivent être prétraitées. Ainsi la première étape du projet a été de construire les corpus de données et de les préparer à être utilisés pour entraîner les modèles prédictifs de glycémie. Ce chapitre

décrit, dans un premier temps, les différents jeux de données ainsi que leur processus d'acquisition. Puis, il traite de leurs étapes de prétraitement, étapes qui sont uniformes sur l'ensemble des études expérimentales conduites pendant le doctorat.

3.2 Projet IDIAB

3.2.1 Contexte et objectifs

L'étude de l'état de l'art dans le chapitre précédent nous a montré qu'il existe plusieurs moyens d'obtenir des données pour la prédiction de glycémie, notamment à travers le logiciel de simulation T1DMS [106] ou le jeu de données OhioT1DM [1]. Nous avons aussi identifié plusieurs limitations à l'état de l'art. Premièrement, presque aucune étude ne se fait sur les personnes diabétiques de type 2 alors qu'elles représentent plus de 90% de la population diabétique. De plus, le nombre de patients sur lesquels les données sont récoltées est assez faible, n'étant ainsi pas très représentatif de la grande variabilité de la population diabétique. Enfin, l'utilisation des données d'activité physique, d'humeur ou d'évènement est encore mal étudiée.

Le projet IDIAB, mené en collaboration avec l'association Revesdiab [129], réseau de santé pour le diabète, a pour objectif général de palier ces limitations. En particulier, il a pour objet de :

— collecter des données sur patients diabétiques de type 2 afin d'évaluer les modèles de prédictions de glycémie sur cette population diabétique qui gagnerait elle aussi à utiliser des dispositifs de tels dispositifs ;

— à termes augmenter le nombre de patients sur lesquelles les modèles sont évalués en mettant les données à disposition des autres chercheurs ;

— collecter des données variées comme la glycémie du patient, l'insuline injectée, les glucides ingérés, son activité physique, son humeur, et évènements quotidiens en conditions réelles ;

Dans la section suivante, nous détaillons les aspects techniques du projet IDIAB (système et protocole expérimentaux, critères d'inclusion et d'exclusion à l'étude). Ces modalités techniques ont été discutées avec le réseau Revesdiab, partenaire de l'étude, et ont été approuvées par le Comité de Protection des Personnes (CPP) du Sud-Ouest et Outre-Mer dans le cadre de la collecte de données envisagée (numéro d'enregistrement RCB 2018-A00312-53).

3.2.2 Aspects techniques du projet IDIAB

Système expérimental

La régulation de la glycémie est influencée par de nombreux facteurs comme l'ingestion de glucides (repas), les prises d'insulines, l'activité physique [73], le sommeil [82], ou l'humeur [138]. La Figure 3.1 donne une représenta-

1. Publié courant 2018, le jeu de données OhioT1DM n'était pas disponible au début du doctorat.

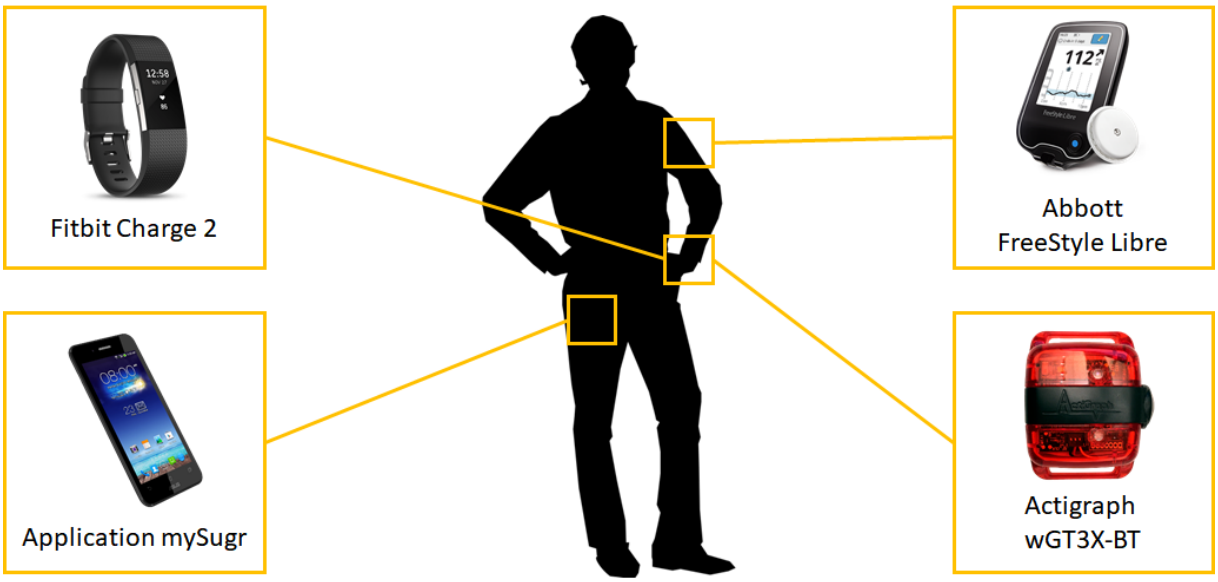


FIGURE 3.1: Système expérimental du projet IDIAB.

tion graphique du système expérimental. Celui-ci a servi à collecter ces données variées en équipant les patients diabétiques du capteur de glycémie en continu FreeStyle Libre (Abbot), des bracelets mesurant l'activité physique Charge 2 (Fitbit) et wGT3X-BT (ActiGraph), ainsi de l'application smartphone mySugr (mySugr) :

— **FreeStyle Libre de Abott** : Le FreeStyle Libre (FSL) est un capteur de glycémie en continu, certifié dispositif médical de classe IIb. Utilisé dans le domaine de la prédiction de glycémie [130, 4], il a l'avantage d'être remboursé par la Sécurité Sociale en France. Le FSL est composé d'un capteur et d'un lecteur. Le capteur prend la forme d'un patch que l'on fixe généralement sur le haut du bras (voir Figure 3.2). Celui-ci produit un courant électrique dont l'intensité varie en fonction du taux de glucose du liquide interstitiel. Ces valeurs sont stockées dans le capteur pendant 8 heures et peuvent être obtenues, toutes les 15 minutes, en scannant le capteur avec le lecteur FSL ou avec son smartphone. De plus, le lecteur FSL permet au patient de consulter différentes statistiques comme la tendance actuelle de sa glycémie. Grâce à son application sur ordinateur, le FSL est aussi utilisé par les diabétologues pour analyser l'évolution du diabète du patient (estimation de l'HbA1c, statistiques sur plusieurs semaines, etc.).

— **Charge 2 de Fitbit** : Le bracelet Charge 2 de Fitbit (que nous pouvons appeler *bracelet Fitbit* pour simplifier) est un bracelet permettant de mesurer en continu l'activité physique de l'utilisateur. Représenté par la Figure 3.3, il permet d'obtenir des informations quant au nombre de pas, au nombre de calories brûlées, à la fréquence cardiaque, ainsi quant au suivi du sommeil. Bien que conçu pour un usage personnel, le bracelet Charge 2 (comme les autres objets Fitbit) fait souvent son apparition dans des études aux buts nombreux et variés [38, 155]. Il est connecté par Bluetooth au smartphone de l'utilisateur avec l'application Fitbit. Les données collectées par les capteurs internes du bracelet sont transférées automatiquement à l'application Fitbit.

FIGURE 3.2: Capteur et lecteur FreeStyle Libre (gauche) et leur utilisation (droite).

FIGURE 3.3: Bracelets Charge 2 (gauche), wGT3X-BT (centre) et leur utilisation (droite).

Ces données sont ensuite stockées au sein du centre de données Fitbit. Il est alors possible de récupérer les données stockées en utilisant une API (*Application Programming Interface* - interface de programmation d'application) que Fitbit met à disposition.

— **wGT3X-BT d'ActiGraph** : Bien que le bracelet Fitbit réponde au besoin de mesurer l'activité physique du patient, celui-ci n'a pas été certifié comme dispositif médical. Ainsi, afin de pouvoir vérifier que son utilisation est sans danger pour le patient dans le cadre de la prédiction de la glycémie, nous avons inclus un second bracelet mesurant l'activité physique : le wGT3X-BT d'ActiGraph (voir Figure 3.3). Ce bracelet (que nous pouvons appeler *bracelet ActiGraph* pour plus de simplicité) est un dispositif médical de classe I. Conçu pour la recherche biomédicale, il est largement utilisé dans ce domaine [89, 3]. Utilisant un accéléromètre de type MEMS (Micro-Electro-Mechanical-System) à 3-axes et des algorithmes de filtrage propriétaires d'ActiGraph, il permet d'enregistrer des descripteurs de l'activité physique (nombre de pas, dépenses énergétiques, sommeil, etc.).

— **mySugr de mySugr GmbH** : Enfin, pour récolter les données discrètes telles que les prises de glucides ou d'insuline, l'application smartphone mySugr de mySugr GmbH est utilisée (voir Figure 3.4). MySugr est une application de coaching pour personnes diabétiques et a été certifiée dispositif médical de classe I. À travers sa fonctionnalité de journal quotidien, la personne diabétique peut y enregistrer ses prises de glucides, ses injections d'insuline, ses changements d'humeur, ou des informations concernant des évènements importants

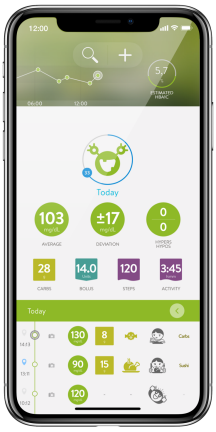


FIGURE 3.4: Application smartphone mySugr.

de la journée. Son format numérique permet aux données récoltées d'être plus précises, notamment grâce à l'horodatage automatique des différentes entrées du journal.

Protocole Expérimental

Nous décrivons ici le protocole expérimental que les participants ont suivi, en particulier, l'utilisation qui doit être faite des différents dispositifs que nous venons de présenter. Ce protocole expérimental est le fruit des discussions que nous avons eues avec le réseau Revesdiab, ainsi que de nos tests effectués en interne (permettant de vérifier sa faisabilité).

Pour chaque participant, la durée de la collecte de donnée a été fixée à 4 semaines. Cette durée a été jugée suffisante en étant supérieure à la durée de nombreuses études portant sur la prédiction de la glycémie, comme a pu le montrer l'analyse de l'état de l'art (10.10 jours en moyenne)[2]. De plus, elle coïncide avec l'utilisation de deux capteurs FSL, chaque capteur ayant une durée de vie de 14 jours exactement.

Pour chaque participant, l'expérience débute par une première rencontre avec une infirmière du réseau Revesdiab et moi-même. Cette rencontre a pour objectif d'équiper le participant, de lui expliquer en détail le fonctionnement des dispositifs et des tâches qui lui incombent. L'objectif a aussi été de le sensibiliser sur ses droits (droit de rétraction, droit de suppression des données, etc.), conformément au règlement général sur la protection des données (RGPD), entré en vigueur en mai 2018. En outre, le participant signe un formulaire de consentement quant à l'utilisation des données récoltées dans le cadre du projet.

Pendant la durée de la collecte, le participant ne doit pas changer ses habitudes et train de vie au quotidien tout en veillant à effectuer les actions suivantes :

— **FreeStyle Libre** : porter le capteur de glycémie sur l'arrière du bras non dominant et le scanner avec le lecteur

2. Bien que le jeu de données OhioT1DM possède environ 8 semaines de données par patient, n'existant pas au début de la thèse, il ne pouvait servir de référence à notre collecte.

au moins 4 fois par jour, dont une fois au coucher et une fois au lever (pour assurer une continuité des données de glycémie collectées tout au long de la journée et de la nuit) ;

— **FreeStyle Libre** : changer le capteur de glycémie aussitôt que celui-ci expire (ce qui arrive toutes les 2 semaines) ;

— **Charge 2** et **wGT3X-BT** : porter au quotidien, jour et nuit, les deux bracelets sur le poignet non dominant sauf pour les activités aquatiques (e.g., douche, vaisselle, piscine) ;

— **Charge 2** et **wGT3X-BT** : recharger les appareils lorsque cela est nécessaire, de préférence de nuit (tous les 5 jours et une fois au bout de 15 jours respectivement) ;

— **Charge 2** : veiller à ce que le Bluetooth et l'accès à Internet soient activés sur le smartphone pour la sauvegarde des données du Charge 2 ;

— **mySugr** : renseigner, à chaque repas, la date et l'heure, sa description, le nombre de glucides ingérés et une photo ;

— **mySugr** : renseigner, à chaque prise d'insuline ou de médicaments (spécifique au diabète, comme la Metformine, ou non, comme le Doliprane), la date et l'heure ainsi que la dose administrée ;

— **mySugr** : renseigner, à chaque changement d'humeur ou évènement important, la date et l'heure ainsi que sa nature.

L'expérimentation se finit par un entretien dont le but est de récupérer le système expérimental et les données collectées, ainsi que de discuter du ressenti du participant tout au long de ces 4 semaines.

Critères d'inclusion et d'exclusion

Tout comme le système et protocole expérimental, les critères d'inclusion et d'exclusion à l'étude ont été établis avec l'aide du réseau Revesdiab. L'objectif de ces critères est d'assurer le bon déroulement de la collecte ainsi la sûreté du participant pendant celle-ci. Les critères d'inclusion, caractérisant les patients potentiels pouvant participer à l'étude, ont été les suivants :

— personnes âgées de 18 à 75 ans ;

— personnes habitant en France ou résidant en France pendant la durée de l'expérience ;

— hommes ou femmes ;

— personnes ayant été diagnostiquées diabétiques de type 2 depuis au moins 1 an ;

— personnes utilisant déjà le dispositif Free Style Libre depuis plus de 3 mois et ayant suivi les formations appropriées ;

— volontaires ;

Le critère d'inclusion concernant les personnes utilisant déjà un FreeStyle Libre a été jugé nécessaire, car son utilisation par une personne non initiée pourrait se révéler dangereuse. Une incompréhension des mesures faites par le capteur peut entraîner un comportement compromettant. Par exemple, si une hyperglycémie est faussement détectée, le patient peut être amené à prendre de l'insuline pouvant l'amener en hypoglycémie et être sujet à des complications graves telles que le coma. Cependant, ce critère d'inclusion implique aussi plusieurs contraintes, notamment celle de restreindre l'étude aux diabétiques de type 2 jugés sévères. En effet, bien que le FreeStyle Libre soit remboursé par la sécurité sociale, il n'est prescrit qu'en cas de nécessité, excluant ainsi les nouveaux patients diabétiques ou ceux possédant un diabète léger.

Les critères d'exclusion ont pour objectif d'exclure certains individus de l'étude, individus non représentatifs de la population diabétique (e.g., diabète déséquilibré) ou pour lesquels l'étude peut se révéler être dangereuse. Les critères d'exclusion ont été les suivants :

— femmes enceintes, allaitantes ou prévoyant une grossesse pendant l'année ;

— personnes sous dialyses ;

— personnes ayant un rythme de vie incompatible avec le rechargement des appareils, et le port au quotidien des dispositifs ;

— personnes incompatibles avec le dispositif FreeStyle Libre (allergies, avis du médecin, etc.) ;

— personnes ayant un diabète très déséquilibré (HbA1c > 9% ou < 5.5%, ou historique d'hypoglycémies sévères) ;

— personnes ayant un smartphone incompatible avec les applications demandées (pour cause d'ancienneté, de système d'exploitation) ;

— patients porteurs d'une maladie aigüe ;

— personnes diabétiques de type 1 (insulinodépendant) ;

3.2.3 Résultats de la campagne de collecte de données

La collecte de données s'est déroulée de juin 2018 jusqu'à décembre 2019. Ici, nous donnons plus de détails sur le suivi de la collecte, la création de la base de données IDIAB ainsi que son utilisation pendant la thèse.

Retours sur la campagne de collecte de données

Les données de 6 patients diabétiques de type 2 ont été récoltées. Le Tableau 3.1 donne un aperçu des caractéristiques physiologiques des participants. Nous pouvons voir que la durée de l'expérimentation a été respectée, avec un minimum de 28 jours (4 semaines) par participants, tout comme le critère d'inclusion en âge. Toutefois, nous pouvons noter une surreprésentation de sexe féminin (seulement un participant était un homme). Cette disparité est

ID	Sexe (F/M)	Âge (années)	Taille (cm)	Poids (kg)	Durée (jours)
1	F	46	163	83	28
2	F	60	163	68	31
3	F	59	167	105	30
4	M	57	163	64	32
5	F	72	160	90	32
6	F	45	167	105	34
Moyenne	5/6 F	56.5 ± 9.14	163.83 ± 2.48	85.83 ± 16.10	31.17 ± 1.86

Tableau 3.1: Description des participants à la collecte de données du Projet IDIAB.

due à la difficulté de recrutement de participant pour une étude relativement contraignante et arrive régulièrement dans le domaine de la prédiction de glycémie (e.g., le jeu de données OhioT1DM possède 2 participants hommes pour 4 participants femmes [109])

Nous avons rencontré des problèmes mineurs durant le déroulement de la collecte de données. Le premier problème a été la présence d'un capteur FSL défectueux chez l'un des participants, entraînant l'absence de données de glycémie pendant deux jours (le temps de trouver un capteur de remplacement). Le deuxième problème a été la présence d'une irritation prononcée au poignet chez l'un des participants. L'irritation serait provoquée par une allergie légère au nickel présent sur le bracelet Fitbit. L'incident a été résolu par l'infirmière coordinatrice en lui recommandant d'appliquer un spray protecteur au niveau de la surface de contact.

Dans l'ensemble, le protocole expérimental a été respecté. La principale difficulté pour les participants a été le calcul du nombre de glucides pour chaque repas. En effet, contrairement aux personnes diabétiques de type 1 qui ont l'habitude de compter les glucides ingérés par repas, ce n'est pas nécessairement le cas pour les personnes diabétiques de type 2. Ainsi, pour les participants n'étant pas à l'aise avec le calcul des glucides, nous leur avons proposé de faire particulièrement attention à la description du repas ainsi qu'aux photos, afin que les glucides soient calculés a posteriori par nous-mêmes.

Quant au ressenti des participants, il a été globalement très positif. Bien que le port simultané des deux bracelets a été vu comme particulièrement contraignant, le système avec un unique bracelet d'activité physique (idéalement, le bracelet Fitbit) a été jugé utilisable dans le cadre de l'éventuelle prédiction des futures valeurs de glycémie. Enthousiasmés par aider à faire « avancer la science », la plupart se sont montrés volontaires pour être sollicités de nouveau dans un cadre expérimental similaire. Les participants ont apprécié la découverte de l'application mySugr ainsi que celle du bracelet Fitbit, les jugeant pertinents pour la gestion de leur diabète au quotidien.

Création de la base de données IDIAB

Suite à leur collecte, la plupart des données sont encore dans un format brut, non uniformisé, et donc difficilement exploitables. À partir de ces données brutes, nous avons créé la base de données IDIAB. Celle-ci a pour

objectif de faciliter l'accès aux données des participants et l'ajout de nouveaux participants. Elle permet aussi de supprimer les données des participants, mettant ainsi en œuvre le droit du participant à la suppression de ses données, en accord avec le RGPD. La base de données, après création, est stockée sur les serveurs sécurisés du CNRS-LIMSI et est seulement accessible par investigateurs du projet IDIAB (i.e., Mounîm A. El Yacoubi, Mehdi Ammi, et moi-même). Nous décrivons ici les étapes qui ont été suivies dans la création de la base de données IDIAB.

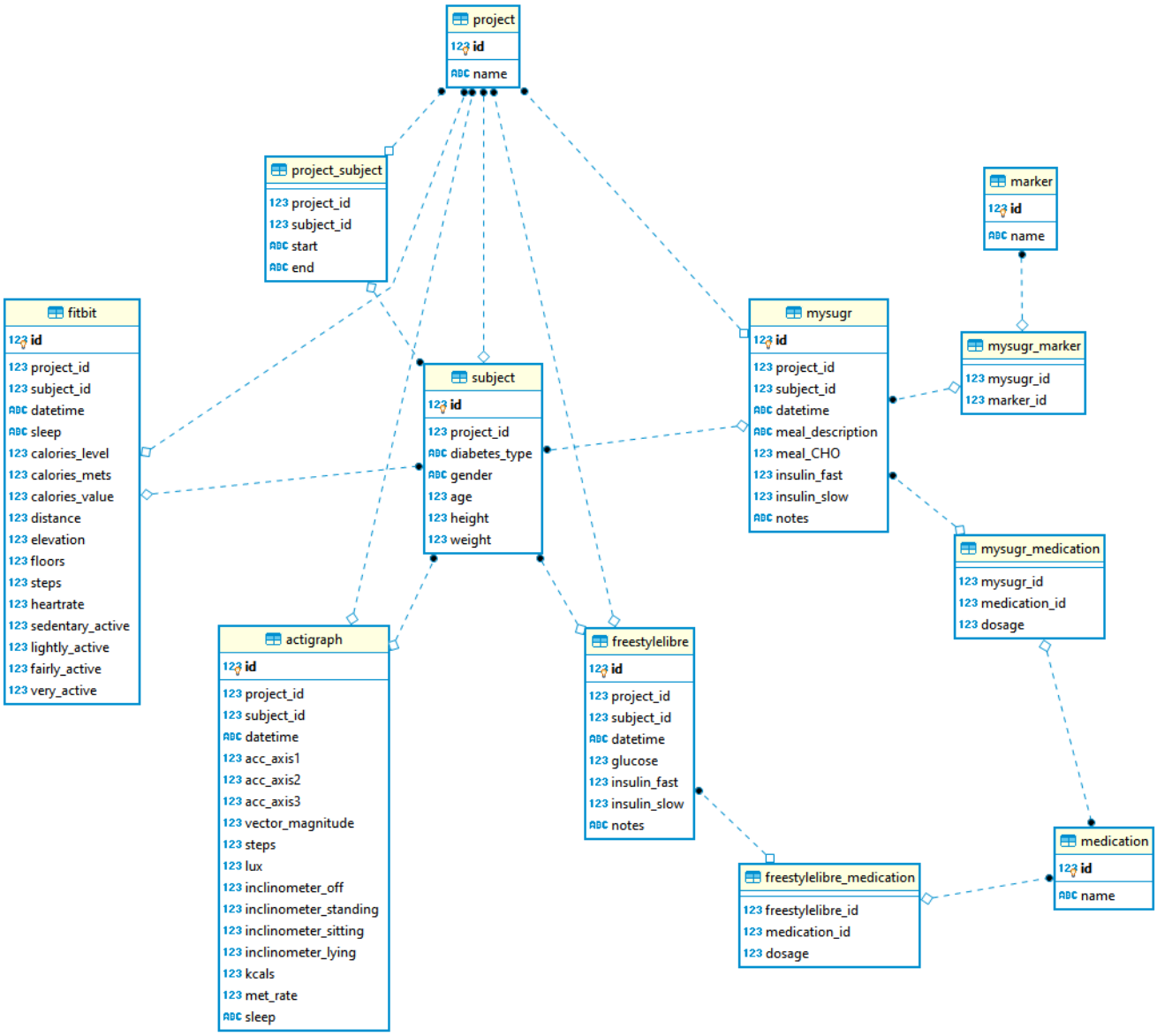


FIGURE 3.5: Schéma de la base de données relationnelle IDIAB - visualisation par DBeaver.

La Figure 3.5 représente le schéma de la base de données relationnelle IDIAB. Chaque participant (table *subject*) est associé à une expérience de collecte de données (table *project*, dont la relation plusieurs à plusieurs avec la table *subject* est représentée par la table intermédiaire *project_subject*). La base de données ne possède aujour-

d'hui qu'une seule entité *project*, le doctorat n'ayant fait l'objet que d'une seule campagne de collecte de données. À l'avenir, la base de données pourra être enrichie de nouvelles données associées à de nouvelles campagnes. Les tables *freestylelibre*, *fitbit*, *actigraph*, et *mysugr* représentent respectivement les données provenant du FreeStyle Libre, du bracelet Fitbit, du bracelet ActiGraph, et de l'application mySugr. Enfin, les tables *medication* et *marker* représentent respectivement les prises de médicaments autres que l'insuline et les évènements importants de la journée reportés par le participant (e.g, changement d'humeur, maladie, etc.). La table *marker* possède une relation plusieurs à plusieurs avec la table *mysugr* à travers la table intermédiaire *mysugr_marker*. De son côté, la table *medication* possède une relation plusieurs à plusieurs avec les tables *freestylelibre* et *mysugr* à travers les tables intermédiaires *freestylelibre_medication* et *mysugr_medication*. En effet, certains participants avaient déjà l'habitude de renseigner leurs prises de médicaments dans le FreeStyle Libre. L'ensemble des champs des différentes tables représentent les données que nous avons pu extraire des dispositifs. Dans les paragraphes suivants, nous donnons les étapes et détails d'extraction des données de chaque dispositif.

Les données brutes du FSL sont regroupées dans un unique fichier CSV. La lecture de ce fichier a permis d'obtenir les données de glycémie (champ *glucose*), d'insuline (champs *insulin_slow* et *insulin_fast*) ainsi que des commentaires laissés par le participant (champ *notes*), le tout étant horodaté (champ *datetime*). Les deux champs *insulin_slow* et *insulin_fast*) représentent le type d'insuline (insuline à action lente ou à action rapide) pris par le participant. Le champ *notes* sert à donner les informations concernant les éventuelles prises de médicaments ou d'évènements liés au FSL (e.g., changement de capteur). Les éventuels médicaments identifiés dans le champ *notes* sont identifiés à travers la table intermédiaire *freestylelibre_medication*.

Les données du bracelet Fitbit, extraites à partir des serveurs Fitbit en utilisant l'API mis à disposition, prennent la forme de descripteurs haut-niveau (e.g., nombre de pas) répartis sur de nombreux fichiers XML (un fichier par jour et par descripteur). L'agrégation de ces nombreux fichiers a permis d'obtenir les données de phase de sommeil (champ *sleep*), de calories dépensées (champs *calories_level*, *calories_mets* et *calories_value*), de déplacement (champs *distance*, *elevation*, *floors*, *steps*), de fréquence cardiaque (champ *heartrate*) et de niveau d'activité physique (champs *sedentary_active*, *lightly_active*, *fairly_active*, et *very_active*), le tout horodaté avec le champ *datetime*. Après extraction des serveurs Fitbit et intégration à la base de données, les données ont été supprimées des serveurs Fitbit (nous rendant uniques détenteurs des données expérimentales, en accord avec la RGPD).

Contrairement aux données Fitbit, les données ActiGraph sont les données brutes d'accéléromètre. Pour obtenir des descripteurs de haut niveau similaires à ceux des données Fitbit, il faut utiliser le logiciel ActiLife, fourni par le constructeur. En plus des données brutes d'accéléromètre moyennées par minute (champs *acc_axis1*, *acc_axis2*, *acc_axis3*, *vector_magnitude*, et *lux*), ActiLife permet d'obtenir des descripteurs d'activité physique (champs *steps*, *kcals*, *met_rate*), de posture du participant (champs *inclinometer_off*, *inclinometer_standing*, *inclinometer_sitting*, et *inclinometer_lying*), ainsi que de sommeil (champ *sleep*).

Enfin, les données mySugr, contenues dans un unique fichier CSV, peuvent être directement incluses à la

base de données. Ces données se caractérisent par des données de nourriture (champs *meal_description* et *meal_CHO*), d'insuline (*insulin_slow* et *insulin_fast*), de commentaires (champ *notes*), ainsi que d'informations sur les différents évènements de la journée (table intermédiaire *mysugr_marker*) et sur les prises de médicaments, autres que l'insuline, pris (table intermédiaire *mysugr_medication*). Cependant, pour les patients pour qui le calcul du nombre de glucides de chaque repas fut une tâche difficile, un grand nombre d'entrées dans le champ *meal_CHO* ont été laissées vides. Pour chaque repas dont les entrées en glucides n'ont pas été renseignées, nous les avons calculés nous-mêmes, a posteriori, à partir de la description des repas (champ *meal_description*). Comme pour les médicaments, certains patients ont préféré renseigner les prises d'insuline dans l'application mySugr plutôt que dans le FSL, expliquant la présence des champs *insulin_slow* et *insulin_fast* dans les deux tables. Le champ *notes* a beaucoup été utilisé par les participants afin de communiquer un éventuel problème dans l'enregistrement des informations (e.g., problème d'horodatage, ou évènement quotidien explicatif). Ces informations nous ont amenés, lorsqu'il s'avérait nécessaire, de modifier certaines entrées du journal (e.g., modification de l'heure de l'enregistrement ou correction de la valeur en glucides du repas). Enfin la table intermédiaire *mysugr_marker* permet d'identifier plusieurs « marqueurs » caractérisant l'enregistrement du journal. Ces marqueurs peuvent prendre le forme d'informations concernant l'humeur du participant (e.g., joyeux, stressé), son état physique (e.g., mal de tête, allergies), de la nature du repas (e.g., déjeuner, dîner), ou d'évènement quotidien (e.g., sport, ménage). Bien que nous ne demandions aux participants de ne renseigner que les informations sur leur humeur, certains participants ont aussi renseigné les autres informations.

Les Figures 3.6, 3.7 et 3.8 représentent les données obtenues sur une journée pour le patient 2. Pour chaque courbe au sein de ces Figures, leur nom dans la légende décrit le champ associé dans la base de données. La Figure 3.6 décrit les données de glycémie, d'insuline, de glucides, ainsi que celles d'humeur et d'évènements extraites par le FreeStyle Libre et mySugr. Quant à elles, les Figures 3.7 et 3.8 décrivent les données d'activité physique et de sommeil récoltées par les bracelets Fitbit et ActiGraph respectivement.

3.3 Jeux de données additionnels

Dans le but de généraliser davantage les résultats obtenus dans les différentes études de cette thèse, nous avons choisi d'utiliser deux autres jeux données en complément du jeu IDIAB : le jeu OhioT1DM et le jeu T1DMS. Contrairement au jeu IDIAB, tous deux concernent le diabète de type 1. Le jeu OhioT1DM est très similaire au jeu IDIAB dans la nature des données récoltées (glycémie, glucides, insuline, activité physique, sommeil). Quant au jeu T1DMS, simulé à partir du logiciel T1DMS, il nous a permis de travailler sur les modèles prédictifs avant d'avoir pu obtenir les données réelles.

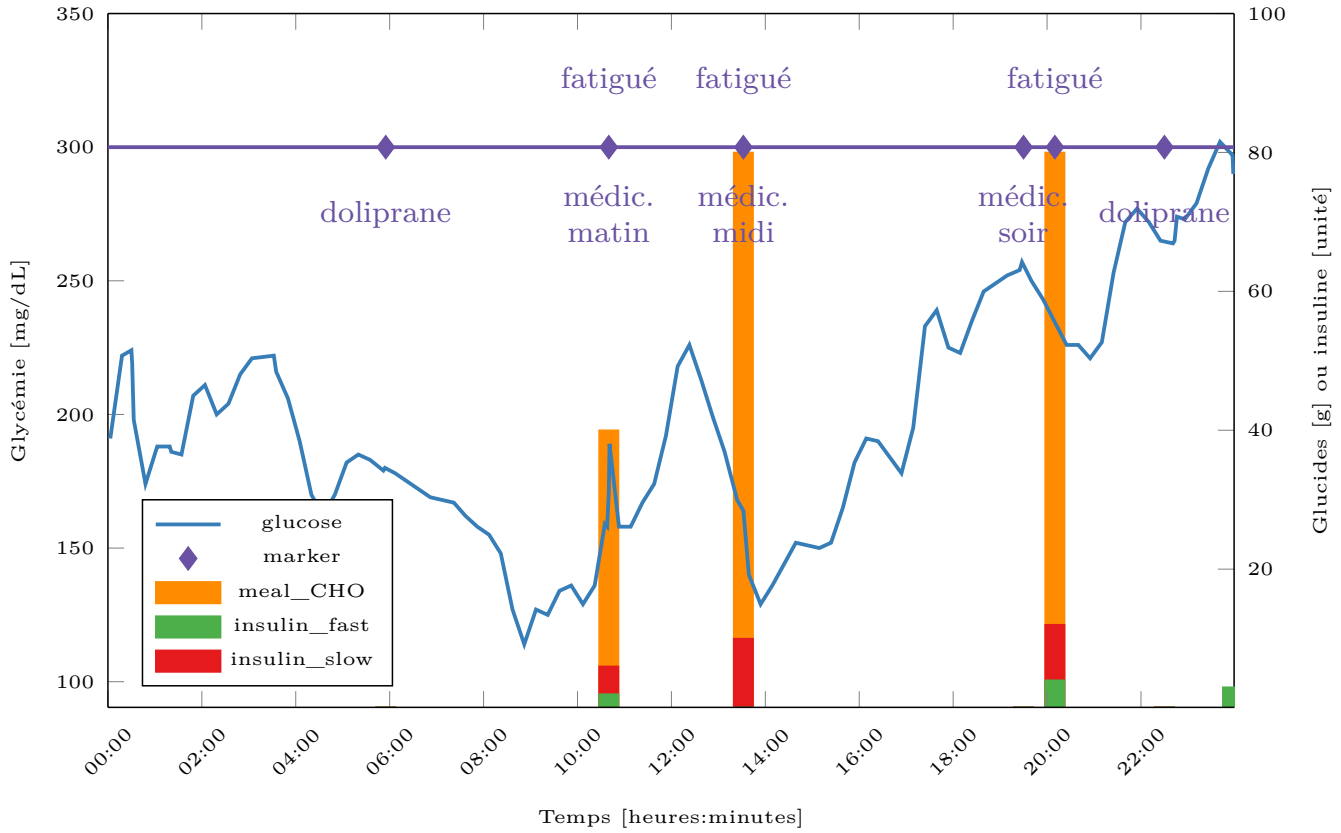


FIGURE 3.6: Données de glycémie, d'insuline, de glucides, d'humeur et évènement extraits du FreeStyle Libre et de mySugr pour le patient 2 sur une journée.

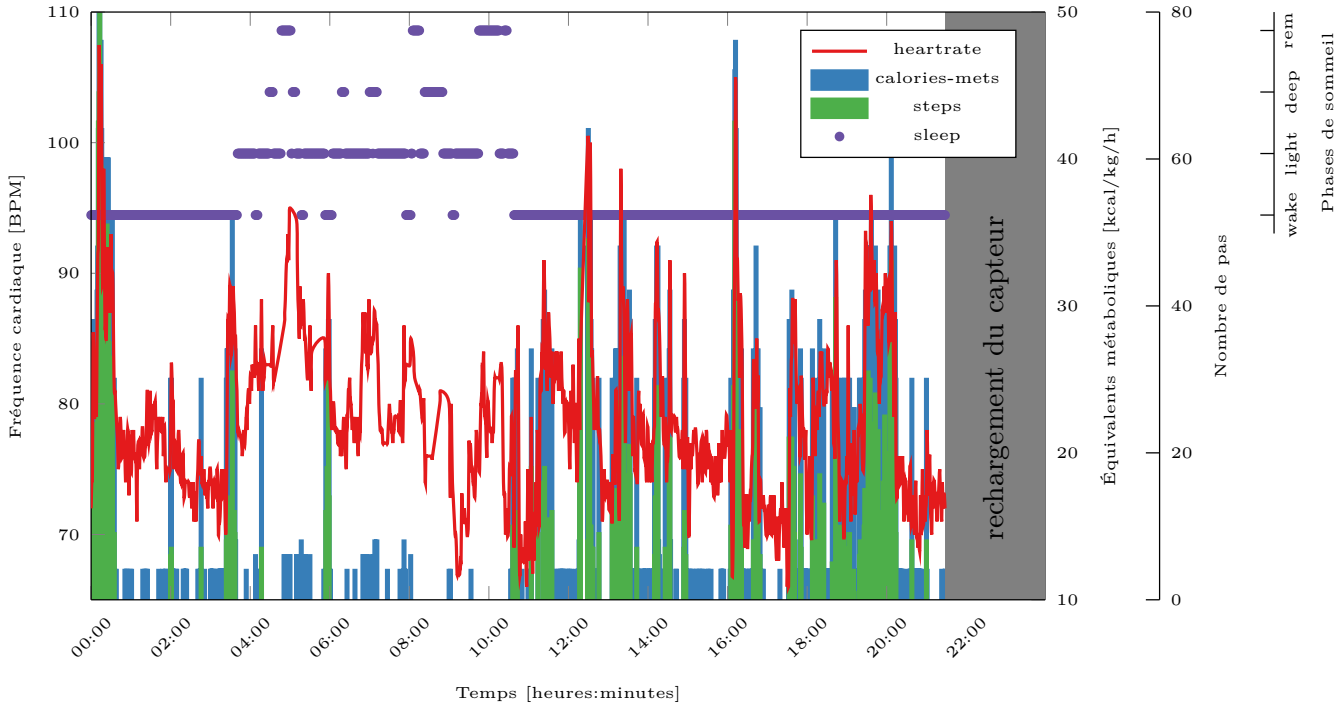


FIGURE 3.7: Données d'activité physique et de sommeil extraites du bracelet Fitbit pour le patient 2 sur une journée.

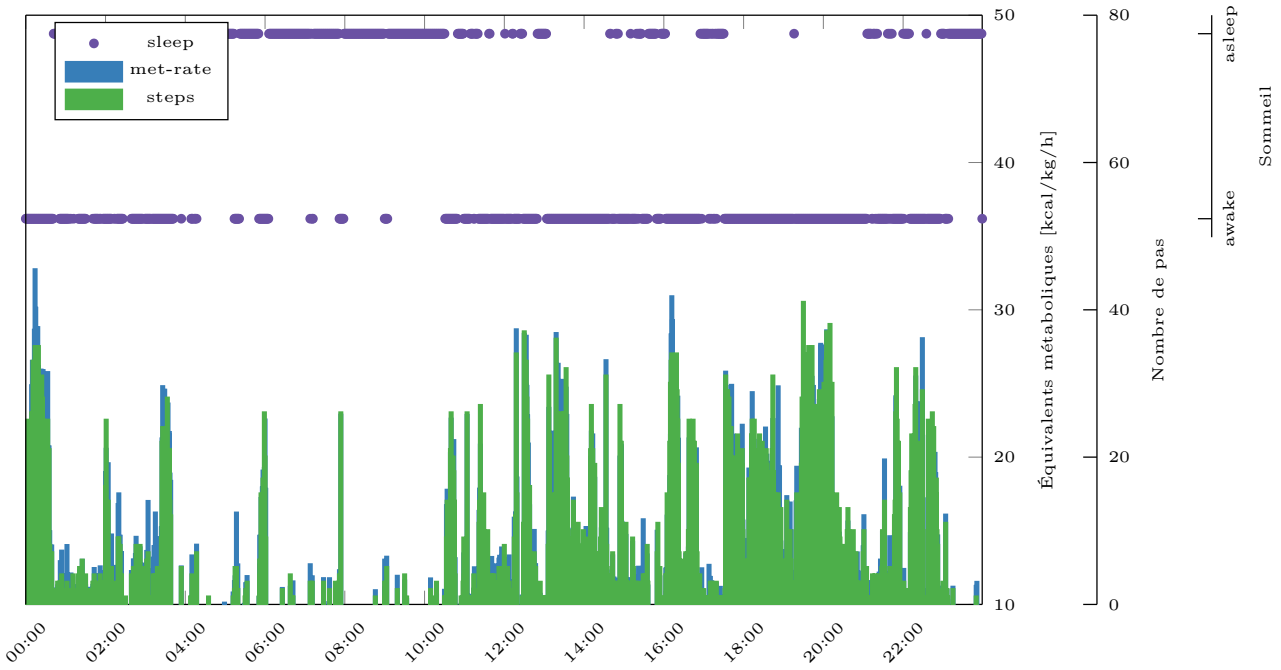


FIGURE 3.8: Données d'activité physique et de sommeil extraites du bracelet ActiGraph pour le patient 2 sur une journée.

3.3.1 Jeu de données OhioT1DM

Présentation du jeu de données

En 2018, Marling et Bunescu ont publié un ensemble de données nommé OhioT1DM pour le *Blood Glucose Level Prediction (BGLP) Challenge* [109][3]. L'ensemble de données comprend 6 patients diabétiques de type 1 (identifiés par leurs identifiants - 559, 563, 570, 575, 588 et 591) qui ont été suivis pendant 8 semaines en environnement non contrôlé. Ils portaient des pompes à insuline Medtronic 530G, des appareils Medtronic Enlite acrshortcgm, des bracelets de fitness Basis Peak et utilisaient une application pour smartphone pour enregistrer les événements quotidiens. Les données suivantes ont été collectées : taux de glucose toutes les 5 minutes à partir du capteur de glycémie, niveau de glucose à partir de lancettes, injections d'insuline (en bolus et basale), repas (heures et quantités de glucides), exercice, sommeil, travail, stress, maladie, fréquence cardiaque (toutes les 5 minutes), réponse galvanique de la peau, température de la peau, température de l'air et nombre de pas. Étant le premier jeu de données public à disposer de cette variété et de cette quantité de données par patient, il suscite un intérêt croissant auprès de la communauté des chercheurs [168, 10, 21, 113, 78, 150, 115, 111, 96, 110, 169, 126].

3. En juin 2020, le jeu de données OhioT1DM a été étendu à 6 nouveaux patients diabétiques de type 1. Compte tenu de la date de cette extension, la version étendue du jeu de données n'a pas pu être utilisée pendant le doctorat.

3.3.2 Jeu de données T1DMS

Présentation du logiciel T1DMS

T1DMS, également connu sous le nom du *UVA/Padova Type 1 Diabetes Metabolic Simulator*, est un environnement de simulation de personnes diabétiques basé sur Matlab. Le simulateur a été annoncé pour la première fois en 2009 [86]. Il a été mis à jour en 2014 [106] et fait en ce moment l'objet d'une nouvelle mise à niveau [151]. En 2018, il a été approuvé par la *Food and Drug Administration* aux États-Unis, dans le cadre du développement de nouvelles stratégies de traitement du diabète de type 1, en tant que substitut possible à des tests précliniques sur les animaux. Dans sa version publique, T1DMS fournit à l'utilisateur une population de 30 patients virtuels de type 1 (10 enfants, 10 adolescents et 10 adultes). Pendant la simulation, chaque patient virtuel est soumis à un scénario quotidien sur lequel l'utilisateur a un contrôle total. La simulation peut se faire en boucle ouverte ou en boucle fermée et est définie par des repas (horaires et quantités), des bolus d'insuline (horaires et quantités). Son fonctionnement en boucle fermé, prenant en compte de potentielles actions prises par le patient modélisés dans la simulation (e.g., prise d'insuline lorsque la glycémie est trop haute), permet aux chercheurs de développer des outils d'automatisation du contrôle de la glycémie [88, 147, 51, 94]. Quant à son fonctionnement en boucle ouverte, sans action différente de celles programmées initialement (e.g., prise d'insuline avant chaque repas), il connaît une utilisation croissante ces dernières années dans le domaine de la prédiction de la glycémie [20, 167, 160, 143, 150, 115, 96, 95, 99, 154, 160, 169, 164]. Le principal inconvénient de l'utilisation de ce simulateur est qu'une licence doit être achetée. Il existe également une version étendue du simulateur avec 300 patients qui, à ce jour, n'est malheureusement pas accessible au public.

Simulation des données

Dans le cadre de la thèse, nous avons travaillé avec la population d'adultes virtuels du simulateur qui ont été soumis au scénario suivant [161, 164] :

- Pour chaque jour de la simulation, un sujet prend 3 repas dont les quantités en glucides et les horaires sont randomisés. En particulier, les instants ont été échantillonnés sur des distributions gaussiennes avec une variance de 0.5 et des moyennes de 7h, 13h et 20h respectivement. Les quantités ont également été échantillonnées à partir de distributions normales avec une moyenne de 40g, 85g et 60g respectivement, et une variance de 0.5 multipliée par la quantité moyenne de glucides du repas. Chaque repas dure 15 minutes.
- Au début de chaque repas, un bolus d'insuline est pris. La valeur du bolus est prise uniformément entre 0.7 et 1.3 fois le bolus optimal du patient (calculé à partir du ratio glucides/insuline optimal du patient).
- Chaque patient est soumis à une injection d'insuline basale constante, optimisée par le simulateur.

La durée de la simulation a été de 8 semaines, soit égale à celle du jeu de données OhioT1DM. La randomisation

des horaires des repas, des quantités de glucides et des valeurs de bolus d'insuline permet de mieux représenter la variabilité des situations réelles (environnement non contrôlé). À la fin de la simulation, pour chaque patient, nous obtenons quatre séries temporelles différentes avec un échantillon toutes les minutes : heure de la journée (min), lectures de glucose (mg/dL), prises de glucides (g/min) et des bolus d'insuline (pmol/min).

3.4 Prétraitement des données

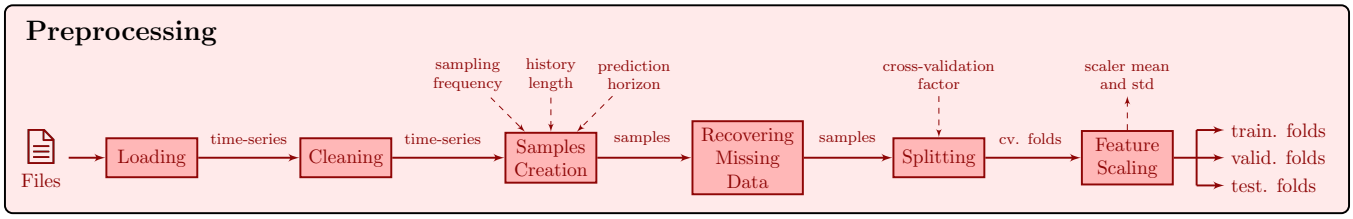


FIGURE 3.9: Étapes de prétraitement des données.

Afin d'utiliser les données pour entraîner des modèles basés sur l'apprentissage automatique, celles-ci doivent tout d'abord être prétraitées. Ce prétraitement s'organise en plusieurs étapes ayant pour objectif général de préparer les données à leur utilisation dans l'apprentissage des modèles prédictifs. Ces étapes sont décrites par la Figure 3.9 et sont détaillées dans les paragraphes suivants. De manière générale, ces étapes de prétraitement sont identiques pour toutes les études du doctorat, rendant les comparaisons entre études possibles.

3.4.1 Choix des données – *Loading*

Les trois jeux de données IDIAB, OhioT1DM et T1DMS sont composés d'un grand nombre de données de nature très variée. Les études faites dans ce doctorat étant centrées autour de l'utilisation de l'apprentissage profond pour la prédiction de la glycémie, nous avons préféré nous restreindre aux données dont l'intérêt est avéré dans l'état de l'art : la glycémie, les ingestions de glucides et les injections d'insuline. Par ailleurs, cette restriction nous permet de travailler sur des données de nature identique, quel que soit le jeu de données, rendant les comparaisons entre jeux de données plus pertinentes. L'exploration des données diverses collectées pendant la campagne mérite quant à elle une étude à part entière, ce qui sera fait à la suite de la thèse par notre groupe de recherche.

La Figure 3.10 donne une visualisation graphique des distributions des valeurs en glycémie, glucides et insuline pour les trois jeux de données. Les distributions du jeu de données T1DMS sont assez différentes des distributions des jeux IDIAB et OhioT1DM. Tout d'abord, nous pouvons noter que les patients réels des jeux IDIAB et OhioT1DM passent plus de temps en hyperglycémie que les patients virtuels T1DMS. Quant aux distributions de glucides et d'insuline, elles possèdent des échelles différentes de celles des jeux IDIAB et OhioT1DM. Cela est dû aux unités différentes que sont utilisées pour le jeu T1DMS. Tandis que les ingestions de glucides sont étalées pendant toute la durée du repas, l'insuline est mesurée en pmol au lieu d'unités pour les jeux IDIAB et OhioT1DM.

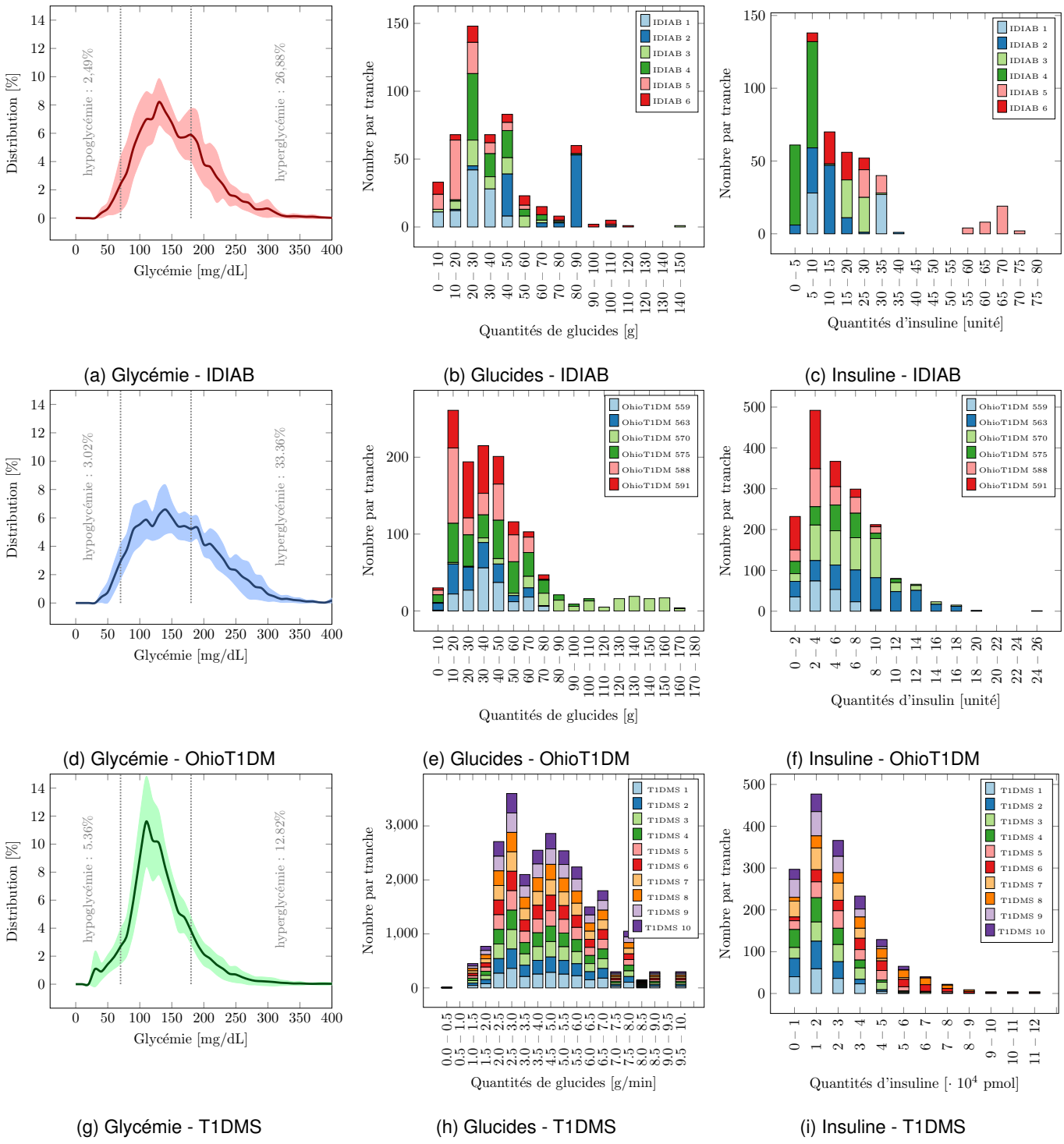


FIGURE 3.10: Distribution des valeurs de glycémie (gauche), de glucides (centre) et d'insuline (droite) pour les jeux de données IDIAB (haut), OhioT1DM (centre), T1DMS (droite). Pour les histogrammes des valeurs de glucides et d'insuline, le nombre par intervalle de tous les patients est empilé les uns sur les autres.

3.4.2 Nettoyage des données — *Cleaning*

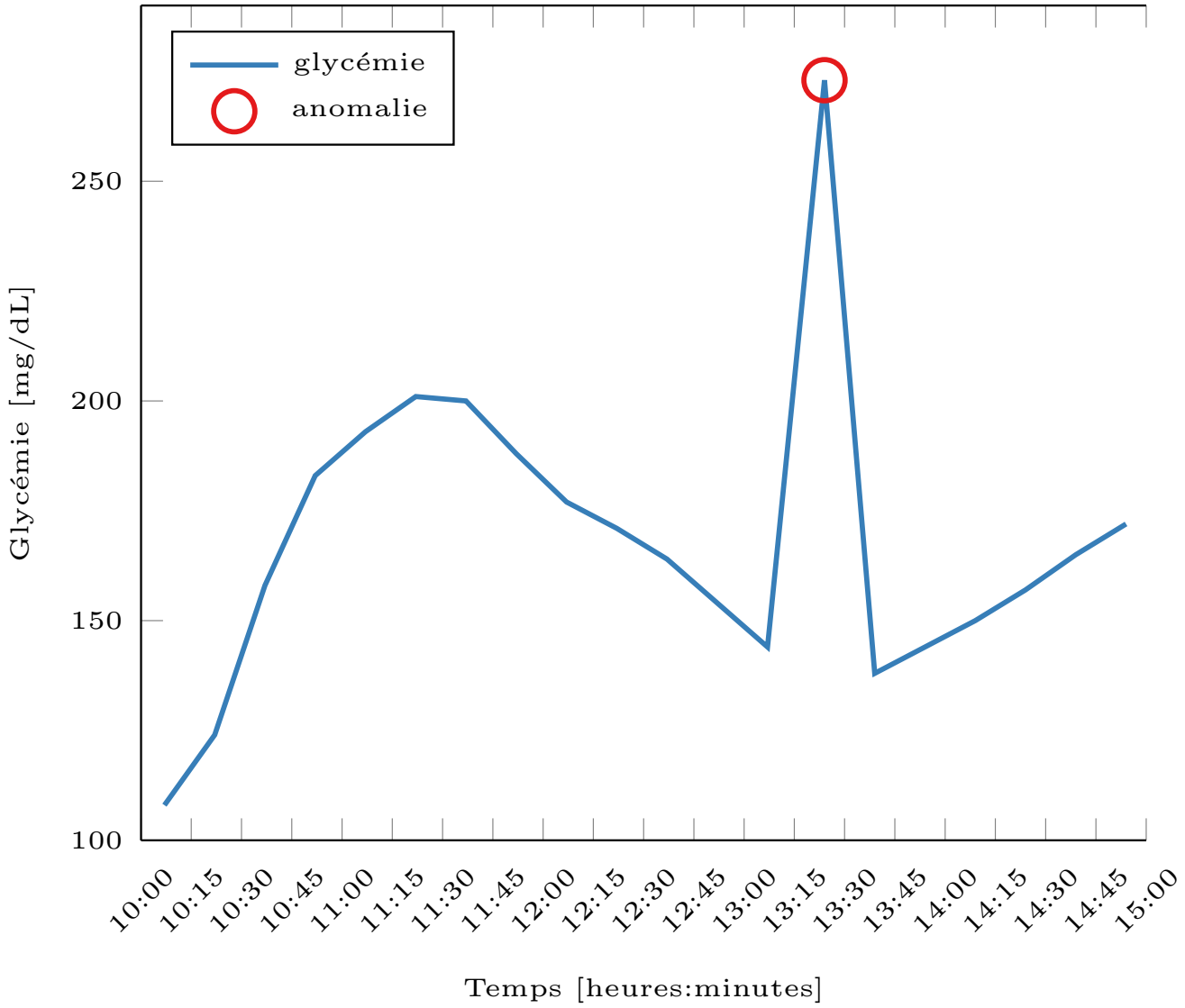


FIGURE 3.11: Glycémie du patient 1 du jeu IDIAB en fonction du temps. La valeur enregistrée à 13:24 est manifestement une anomalie, car elle est incohérente avec le reste du signal.

L'analyse graphique des données de glycémie des participants du jeu de données IDIAB fait ressortir des valeurs qui semblent erronées (voir la Figure 3.11). Ces valeurs se caractérisent par des pics de glycémie ne durant que l'espace d'un échantillon (contrairement à l'augmentation progressive de la glycémie suite à l'ingestion de glucides). La quantité de valeurs erronées varie d'un patient à un autre. Les conserver biaiserait l'entraînement des modèles prédictifs tout comme leur évaluation. Ainsi, nous avons choisi de les retirer du jeu de données. Pour rendre la tâche plus facile et pour être capable de les détecter en temps réel, nous proposons d'automatiser le processus.

Pour cela, nous devons caractériser la nature de ces erreurs. Ces valeurs ne sont pas des valeurs anormales en elles-mêmes, mais des valeurs anormales en les comparant aux valeurs avoisinantes. Nous appelons ces valeurs

ID	1	2	3	4	5	6
Anomalies (nombre)	16	29	21	20	24	26
Anomalies (%)	0.51	0.86	0.7	0.56	0.68	0.88

Tableau 3.2: Nombre de valeur de glycémie erronées supprimées automatiquement par l'algorithme proposé par participant IDIAB.

anormales des *anomalies contextuelles* [**?**]. Pour les détecter et les supprimer automatiquement, nous devons d'abord définir un contexte dans lequel ces valeurs sont des anomalies et un comportement dans le contexte les signalant comme des anomalies. Ici ces valeurs erronées se montrent être incohérentes avec leurs valeurs avoisinantes sous la forme d'un pic (très grande variation positive puis négative). Nous proposons la méthodologie suivante pour éliminer de la plupart d'entre elles :

1. *Attribut contextuel* : le contexte de chaque valeur est les deux valeurs qui l'entourent ;
2. *Attribut comportemental* : les deux variations de la première valeur à la deuxième et de la deuxième valeur à la troisième (toutes deux divisées par le temps écoulé entre les valeurs) ;
3. *Algorithme de détection d'anomalies* :
   (a) Hypothèse : les variations des valeurs de glycémie ont une distribution normale de la moyenne μ et de l'écart type σ (la Figure 3.12 nous montre que la plupart des variations sont petites et centrées sur zéro) ;
   (b) si les variations sont à plus de $n \cdot \sigma$ de μ ; et si les variations ont un signe opposé (un positif et un négatif) ;
   (c) alors la valeur du milieu est une anomalie.

En général, $n = 3$, car la région $\mu \pm 3 \cdot \sigma$ contient 99.7 % des échantillons de données. Cependant, ici, comme nous avons une condition supplémentaire (variations de signes opposés), nous pouvons diminuer la valeur n. Nous avons défini la valeur n à 2.5 (intervalle de confiance $\sim 99\%$).

L'algorithme de détection d'anomalies proposé ne couvre pas le cas où plusieurs valeurs consécutives sont des anomalies. Pour prendre en charge ce cas rare, nous pouvons exécuter l'algorithme plusieurs fois, chaque exécution supprimant une anomalie. Comme l'un des participant possède un nombre important de valeurs anormales, nous avons effectué l'algorithme proposé 5 fois. Le Tableau 3.2 donne le nombre de valeurs supprimées pour chaque participant.

3.4.3 Créations des échantillons de données — *Samples Creation*

Les jeux de données IDIAB, OhioT1DM et T1DMS ne sont pas échantillonnés à la même fréquence. En effet, tandis que le FreeStyle Libre du jeu IDIAB permet d'obtenir les données de glycémie une fois toutes les 15 minutes, le Medtronic Enlite du jeu OhioT1DM le permet une fois toutes les 5 minutes, et le similateur T1DMS une fois par minute. Afin de faciliter la comparaison entre jeux de données, nous avons uniformisé la fréquence d'échantillonnage

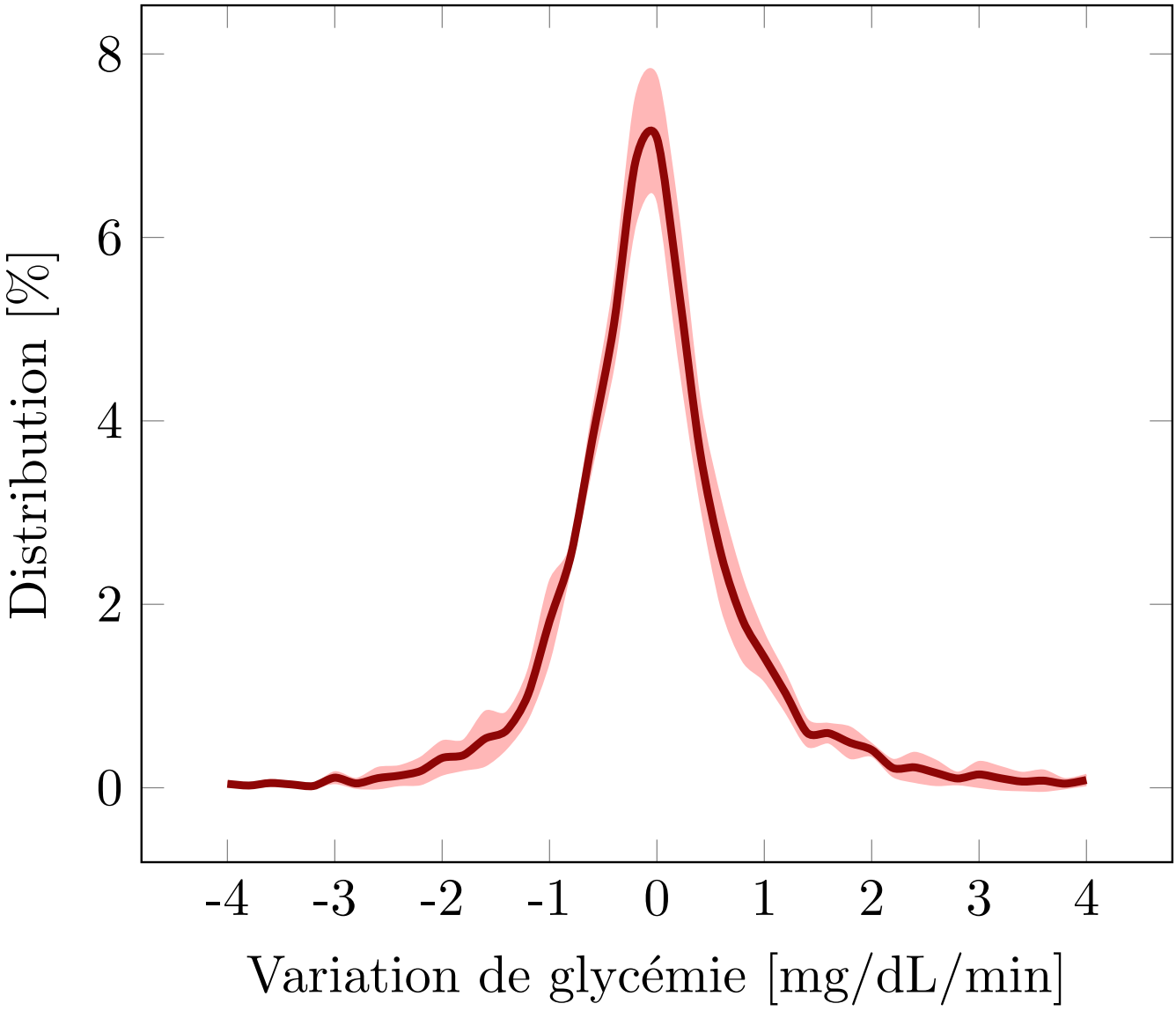


FIGURE 3.12: Distribution des variations de glycémie des patients du jeu de données IDIAB.

des trois jeux à un échantillon toutes les 5 minutes (celle du jeu OhioT1DM). La fréquence d'échantillonnage du jeu T1DMS n'a pas été retenue car, étant plus élevée, elle n'est pas représentative des capteurs de glycémie en continue que nous pouvons trouver dans le commerce.

Le jeu T1DMS a été sous-échantillonné avec la stratégie suivante : la glycémie a été moyennée, l'insuline et les glucides additionnés. Quant au jeu IDIAB, son sur-échantillonage fait apparaître un grand nombre de données manquantes. Ces valeurs vont pouvoir être récupérées artificiellement après avoir créé les échantillons d'apprentissage (voir Section 3.4.4).

Une fois les séries temporelles ré-échantillonnées, elles peuvent être utilisées pour créer les échantillons de données qui serviront à l'apprentissage des modèles. Un échantillon de données d'apprentissage est constitué des données d'entrée au modèle ainsi que de l'objectif de prédiction du modèle (voir Figure 3.13). Dans le cadre de la

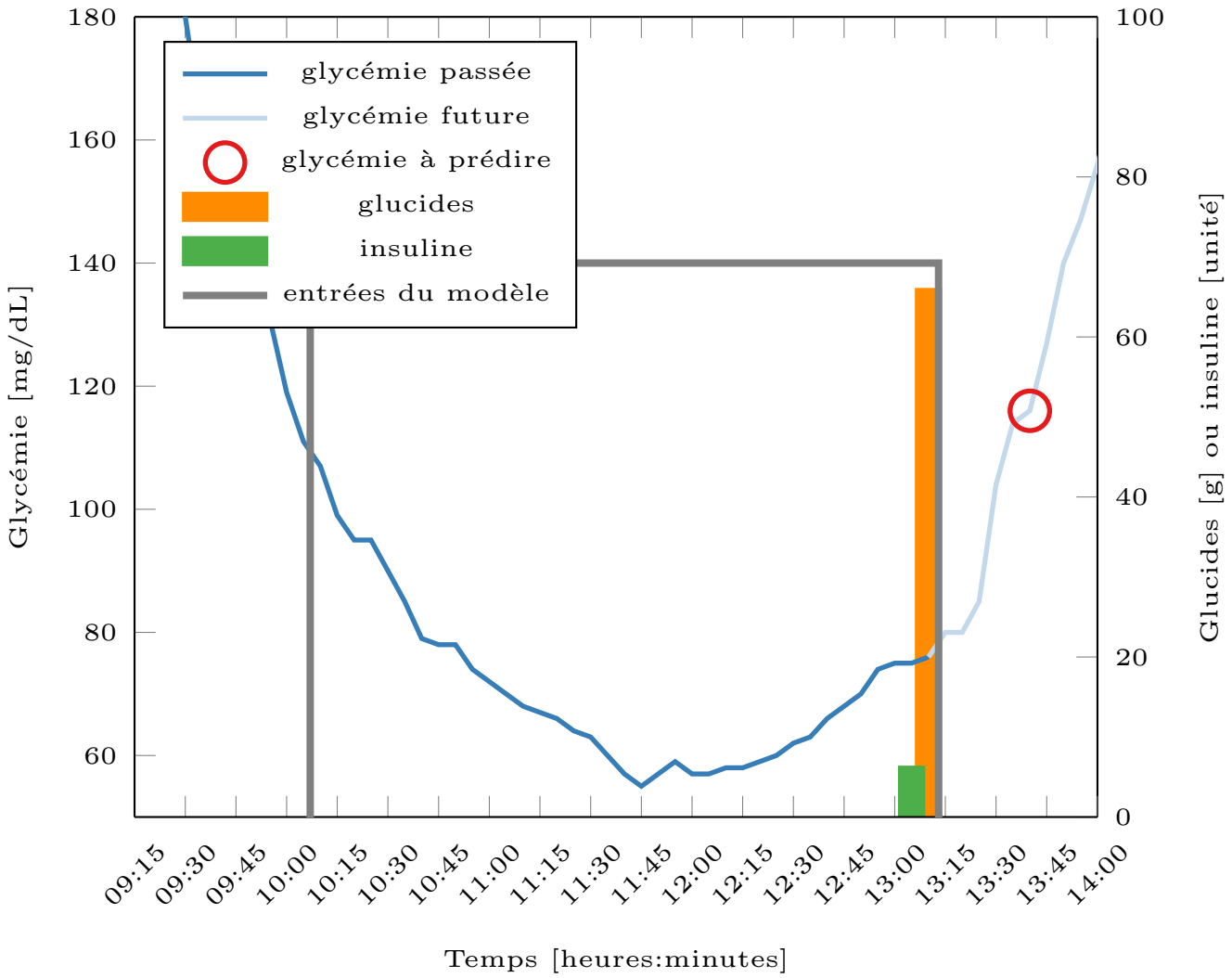


FIGURE 3.13: Création d'un échantillon de données contenant l'historique des trois dernières heures en données de glycémie, glucides et insuline ainsi que la glycémie à prédire 30 minutes dans le futur.

thèse, nous utilisons comme entrée les 36 dernières valeurs de glycémie, d'insuline et de glucides, ce qui correspond à l'historique des trois dernières heures. L'objectif de prédiction dépend l'étude et de l'horizon de prédiction analysé. Tandis que pour l'étude du Chapitre 4, nous utilisons des horizons de prédictions de 30, 60 et 120 minutes, pour les études des Chapitres 5, 6 et 7, nous utilisons seulement un horizon de prédiction de 30 minutes.

3.4.4 Récupération des données manquantes — *Recovering Missing Data*

Contrairement au jeu de données T1DMS, les échantillons des jeux OhioT1DM et IDIAB possèdent de nombreuses valeurs de glycémie manquantes. Ces valeurs manquantes proviennent soit de défauts de capteur (valeurs manquantes ou valeurs erronées nettoyées) ou du sur-échantillonnage du jeu de données IDIAB. Certaines de ces valeurs peuvent être artificiellement récupérées en suivant la stratégie suivante pour chaque échantillon :

1. interpoler linéairement l'historique de glycémie lorsque cela est possible (voir Équation 3.1, où f est le signal

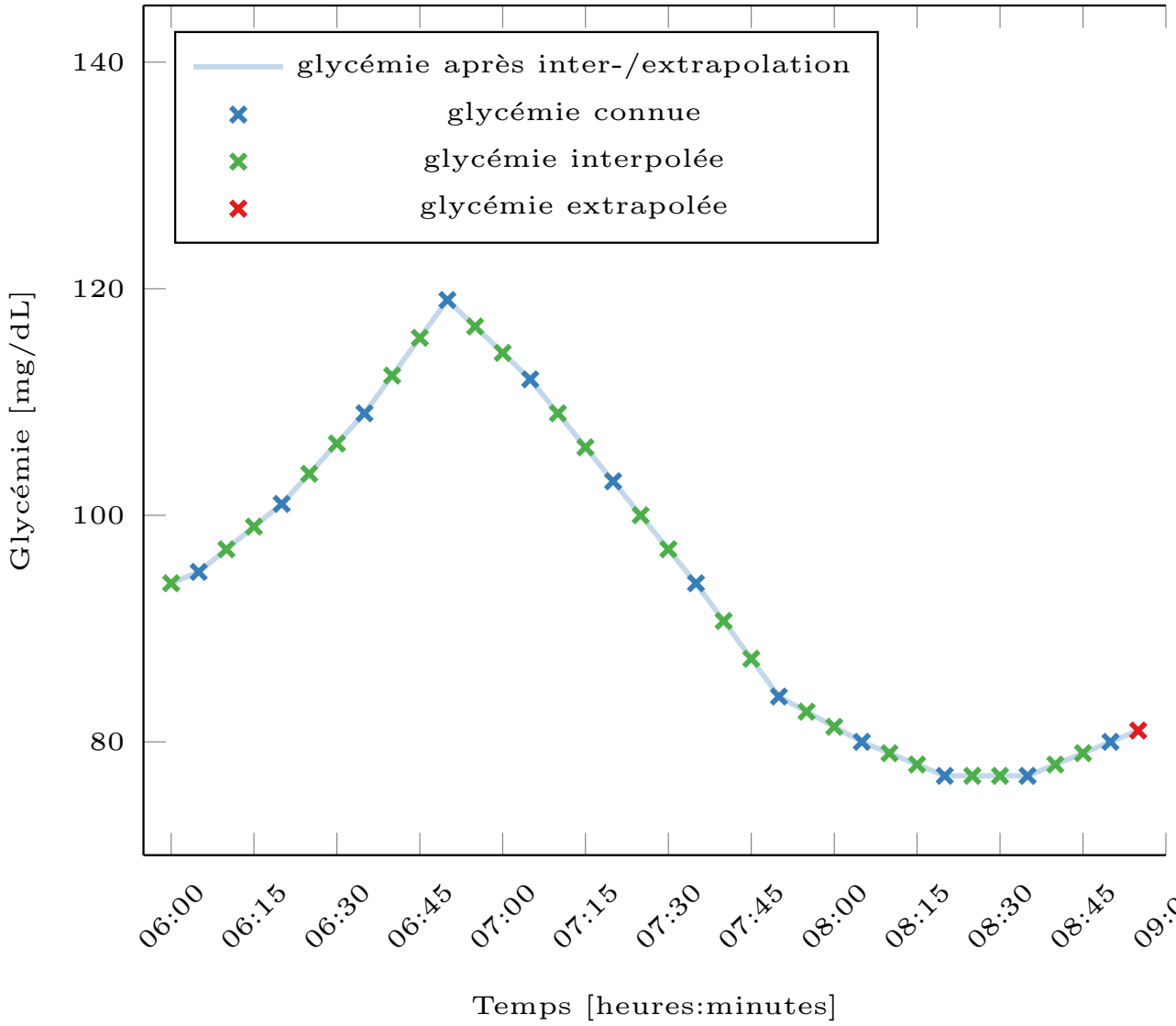


FIGURE 3.14: Historique de glycémie du patient 1 du jeu IDIAB donné en entrée au modèle prédictif après inter-/extrapolation des valeurs manquantes.

de glycémie, $\overline{f}$ est le signal de glycémie interpolé, et t_a, g_a et t_b, g_b caractérisant les deux valeurs de glycémie connues encadrant la valeur à l'instant t) ;

2. extrapoler linéairement dans le cas contraire, généralement lorsque la valeur de glycémie manquante est la donnée la plus récente (voir Équation 3.2, où $\overline{\overline{f}}$ est le signal de glycémie extrapolé, et t_a, g_a et t_b, g_b caractérisant les deux valeurs de glycémie connues précédant la valeur à l'instant t) ;

3. jeter des échantillons lorsque l'objectif de prédiction n'est pas connu pour empêcher d'entraîner les modèles sur des valeurs artificielles.

$$\overline{f}(t) = g_a + (t - t_a) \cdot \frac{g_a - g_b}{t_a - t_b}, \text{ avec } f(t_a) = g_a \text{ et } f(t_b) = g_b \tag{3.1}$$

$$\overline{\overline{f}}(t) = g_b + (t - t_b) \cdot \frac{g_b - g_a}{t_b - t_a}, \text{ avec } f(t_a) = g_a \text{ et } f(t_b) = g_b \tag{3.2}$$

Cette stratégie de nettoyage, dont la Figure 3.14 donne une représentation graphique, garantit que les données du futur, non disponibles en situation réelle, ne sont pas utilisées pour la récupération de données manquantes et que les modèles sont évalués sur des observations réelles et non artificielles.

3.4.5 Créations des ensembles d'apprentissage, de validation et de test — *Splitting*

Nous avons séparé les échantillons d'apprentissage de chaque patient en ensembles d'entraînement, de validation et de test, chaque ensemble ayant un but précis en apprentissage automatique. L'ensemble d'entraînement est utilisé pour entraîner les modèles. L'ensemble de validation est utilisé pour évaluer le modèle lors de l'optimisation de ses hyperparamètres, pour s'assurer que le modèle appris et les hyperparamètres fonctionnent correctement sur des données nouvelles. L'ensemble de test est utilisé pour l'évaluation finale des modèles.

Dans le contexte du *Blood Glucose Prediction Challenge*, le jeu de données OhioT1DM possède naturellement un ensemble de test [109]. Celui-ci est composé des 10 derniers jours de données disponibles pour chaque patient. Les données restantes sont utilisées pour l'entraînement et la validation des modèles, à la discrétion de l'utilisateur. Le jeu T1DMS possédant autant de données que le jeu OhioT1DM, nous l'avons séparé de la même manière. Le jeu IDIAB, quant à lui, ne possède qu'un nombre de jours environ deux fois plus faible. Utiliser 10 jours comme ensemble de test pour le jeu IDIAB serait amputer son jeu d'entraînement (et/ou de validation) d'un trop grand nombre de jours. Ainsi, pour le jeu IDIAB, nous avons choisi de ne réserver que les 5 derniers jours pour l'ensemble de test. Pour chaque jeu de données, les données restantes n'appartenant pas à l'ensemble de test sont divisés en ensembles d'entraînement et de validation suivant une distribution de 80%/20%, selon une évaluation en validation croisée à 5 plis (*5-fold cross-validation*).

Le Tableau 3.3 représente le nombre moyen d'échantillons d'apprentissage par ensemble et par jeu de données. Tout d'abord, nous pouvons remarquer que les jeux de données réels (IDIAB et OhioT1DM) possèdent des écarts types non négligeables, témoignant de la variabilité de la qualité des données collectées par participant. Les différences entre deux participants peuvent s'expliquer les quantités plus ou moins importantes de données de glycémie manquantes. D'autre part, le jeu de données IDIAB possède entre 5 et 6 fois moins d'échantillons d'apprentissage que les deux autres jeux. Cela s'explique par une collecte de données ayant duré 4 semaines au lieu de 8 pour les autres jeux, ainsi que par la fréquence d'échantillonnage d'une valeur de glycémie toutes les 15 minutes pour le jeu IDIAB. Cette dernière différence sera à garder en tête lors de l'analyse des résultats des modèles prédictifs.

		Ensemble d'apprentissage		
		entraînement	*validation*	*test*
Jeu de données	*IDIAB*	1941.2 (128.69)	485.3 (32.17)	480.5 (38.31)
	OhioT1DM	9088.13 (575.74)	2272.03 (143.18)	2661.67 (106.99)
	T1DMS	10368.0 (0.0)	2592.0 (0.0)	2880.0 (0.0)

Tableau 3.3: Nombre moyen (avec écart type) d'échantillons d'apprentissage à la fin du pré-traitement des données par patient pour les ensembles d'entraînement, de validation et de test des jeux de données IDIAB, OhioT1DM et T1DMS.

3.4.6 Standardisation des données — *Feature Scaling*

Certains algorithmes d'apprentissage automatique (e.g., réseaux de neurones, machines à vecteurs de support) ont besoin que les descripteurs d'entrées (i.e., glycémie, insuline, glucides) varient sur des plages de données normalisées pour s'entraîner correctement et efficacement. Ainsi, nous avons standardisé (moyenne de 0 et écart type de 1) les valeurs de glycémie, d'insuline et de glucide. L'Équation 3.3 représente ce procédé, où μ et σ représentent la moyenne et l'écart type du descripteur x, et x' sa valeur standardisée. La moyenne μ et l'écart type σ de chaque descripteur sont calculés sur les ensembles d'entraînement, afin de faire l'évaluation des modèles sur des données possédant la même distribution que les données utilisées pour leur apprentissage.

$$x' = \frac{x - \mu}{\sigma} \tag{3.3}$$

3.4.7 Étapes de prétraitement non utilisées

Bien qu'ils existent d'autres étapes de prétraitement pour la prédiction de la glycémie, comme l'extraction manuelle de descripteurs ou le filtrage des données de glycémie, nous avons choisi de ne pas les utiliser.

Premièrement, l'intérêt de l'utilisation de descripteurs extraits manuellement pour la prédiction de la glycémie est aujourd'hui mal connu, beaucoup d'études n'en utilisant pas. De plus, dans le cadre de l'apprentissage profond, les réseaux de neurones ont la capacité de faire l'extraction de caractéristiques pertinentes eux-mêmes. Par exemple, dans le domaine de la reconnaissance de la parole, un groupe de chercheurs a réussi à obtenir, sur des données brutes et en utilisant un apprentissage dit « bout en bout », de meilleures performances qu'un système de traitement standard comportant de l'extraction manuelle de descripteurs [70]. Toutefois, l'extraction de caractéristiques manuelles, comme cela est fait en utilisant des modèles physiologiques de consommation de glucides ou d'insuline par le corps, pourrait être utile dans une situation de manque de données.

Quant au filtrage des données de glycémie, bien qu'utiliser un filtre passe-bas atténue le bruit haute-fréquence

du signal et permet ainsi de faire ressortir les tendances du signal plus facilement, cela implique aussi un déphasage de celui-ci. Or, un signal déphasé perd l'information de ses valeurs les plus récentes, valeurs qui sont les plus pertinentes pour la prédiction de glycémie (la valeur la plus récente étant la valeur la plus représentative de la valeur à prédire). Pour éviter ce déphasage, il est possible de filtrer une première fois dans le sens du temps (déphasage positif), puis une seconde fois dans le sens inverse du temps (déphasage négatif, annulant le premier déphasage positif). Cependant, pour que ce type de filtre soit utile, il faudrait pouvoir connaître les données du futur, données qui ne sont bien évidemment pas accessibles pour une application réelle.

4 | GLYFE : une base de résultats de référence

Sommaire

4.1 Introduction

L'état de l'art du domaine de la prédiction de la glycémie de patients diabétiques nous a montré que celui-ci suscite de plus en plus d'intérêt. Une grande partie des dernières recherches se focalisent sur des modèles prédictifs complexes, relevant de l'apprentissage automatique ou profond, en opposition aux simples modèles autorégressifs utilisés historiquement. Cependant, à ce jour, la question de la prédiction de la glycémie reste ouverte. Les progrès dans le domaine sont entravés par plusieurs facteurs, l'un d'eux étant la disponibilité des données utilisées pour entraîner et évaluer les modèles. Premièrement, la collecte de données liées au diabète, telles que la glycémie en temps réel ou les bolus d'insuline pris par les patients, prend beaucoup de temps. Deuxièmement, ces données

ne peuvent pas être facilement partagées entre les chercheurs en raison de leur nature sensible. Cela conduit à l'utilisation de jeux de données qui sont petits et différents des études d'un groupe de recherche à un autre, rendant les comparaisons entre elles non pertinentes. Néanmoins, la situation a récemment changé avec notamment la démocratisation du logiciel de simulation T1DMS et de la publication du jeu de données réelles OhioT1DM. Ces nouveaux accès aux données d'intérêt nous donnent l'opportunité de réunir les chercheurs autour des mêmes données, facilitant et accélérant ainsi la recherche de modèles de prédiction de glycémie plus performants.

Dans ce chapitre nous proposons la création d'une base de résultats de référence, nommée GLYFE (*GLYcemia Forecasting Evaluation*), utilisant ces deux ensembles de données. Le premier objectif de ce benchmark est d'analyser selon une procédure standardisée les performances des modèles basés sur l'apprentissage automatique et profond qui ont suscité de l'intérêt ces dernières années. Nous avons publié l'ensemble de l'architecture du traitement des données (prétraitement, entraînement des modèles, évaluation) en open source afin de rendre les résultats réplicables et l'étude approfondie dans le futur. Du point de vue de la thèse, ce projet a permis de dégager différentes problématiques qui ont été ensuite étudiées pendant la thèse. À ce titre, bien que le jeu de données IDIAB n'est pas aujourd'hui rendu public, nous proposons tout de même d'analyser les résultats des modèles de référence obtenus sur ce troisième jeu de données.

Ce chapitre est organisé comme suit. Premièrement, nous détaillons l'architecture de traitement des données, de leur prétraitement jusqu'à l'évaluation des modèles prédictifs. Après avoir passé en revue les composantes du logiciel open source qui a été créé pour la réalisation de cette base de résultats de référence, nous analysons les résultats obtenus. Enfin, de ces résultats nous tirons des conclusions générales ainsi que des lignes directrices pour de futures recherches pour la communauté et pour la thèse.

4.2 Méthodologie

Dans cette section, nous détaillons l'ensemble de la chaîne de traitement des données, de leur prétraitement à l'entraînement et évaluation des modèles prédictifs. L'objectif est de standardiser la méthodologie générale afin qu'elle soit réutilisée par les autres chercheurs du domaine pour comparer leurs modèles.

4.2.1 Obtention et prétraitement des données

Pour la création de cette base de résultats de référence, nous utilisons les jeux de données OhioT1DM et T1DMS présentés au chapitre précédent. Nous avons décidé de ne pas officiellement (dans la publication liée à cette étude [34]) utiliser le jeu IDIAB, car celui-ci, étant en cours de création, n'a pas été rendu public. En effet, dans l'optique d'une utilisation future du benchmark par d'autres chercheurs, les données qui sont utilisées doivent être identiques, et donc accessibles. L'utilisation combinée des jeux OhioT1DM et T1DMS permet d'établir une comparaison entre les performances des modèles sur des données réelles et sur des données simulées. Néanmoins, du point de

vue des travaux de la thèse que nous détaillons dans les chapitres subséquents, nous proposons d'analyser les résultats obtenus sur le jeu IDIAB. Cela nous permet de généraliser nos conclusions ainsi que de posséder un ensemble de résultats de référence identique pour tous les jeux de données, résultats qui seront utilisés dans la suite de la thèse.

Avant de pouvoir être données en apprentissage aux modèles prédictifs, les données doivent être prétraitées. Les étapes de prétraitement suivies sont celles décrites dans le chapitre précédent et dont la Figure 4.1 en fait le rappel. Pour plus d'informations, se référer au Chapitre 3 pour plus d'information. Nous avons vu dans le Chapitre 2 que les performances des modèles ne sont pas seulement affectées par leur choix et entraînement, mais aussi par le choix des données et de leur prétraitement. Ainsi, dans le cadre d'une utilisation ultérieure du benchmark par d'autres chercheurs, ces étapes de prétraitement sont modifiables avec par exemple l'ajout de nouvelles données, ou l'extraction de descripteurs.

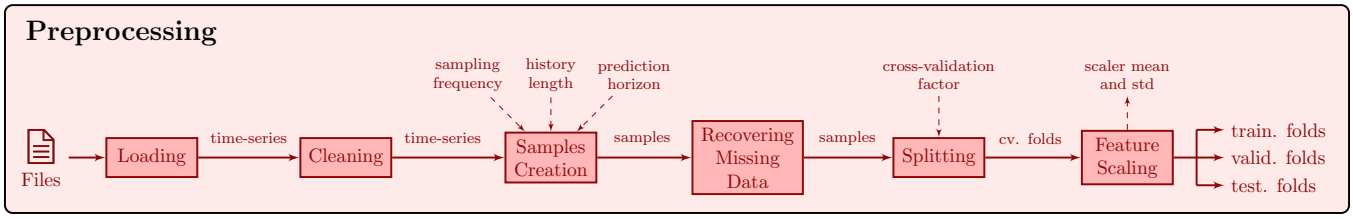


FIGURE 4.1: Étapes de prétraitement des données.

À la fin du prétraitement des données, nous obtenons, pour chaque patient, trois ensembles d'échantillons de données : celui d'entraînement, celui de validation et celui de test. Pour permettre l'évaluation des modèles par validation croisée, nous avons 5 permutations des ensembles d'entraînement et de validation. Tous les échantillons sont composés de l'historique de glycémie, d'insuline, et de glucides des 3 heures passées, avec une information toutes les 5 minutes.

4.2.2 Entraînement et optimisation des modèles

Les ensembles d'entraînement, de validation et de test ont tous un objectif différent, objectifs qui sont schématisés par la Figure 4.2. L'ensemble d'entraînement est utilisé pour entraîner les modèles. L'ensemble de validation est utilisé pour évaluer le modèle lors de l'optimisation de ses hyperparamètres, afin de s'assurer que le modèle et les hyperparamètres fonctionnent correctement lors de l'utilisation de données nouvelles. L'ensemble de test est utilisé pour faire les prédictions finales sur lesquelles l'évaluation des modèles est basée. Bien que l'utilisation d'un ensemble de validation soit commune dans la communauté d'apprentissage automatique, elle n'est que rarement mentionnée dans les études relevant de la prédiction de la glycémie [110, 114, 115, 130, 168, 169]. En absence d'ensemble de validation, l'optimisation des hyperparamètres du modèle doit se faire soit sur l'ensemble de test, soit sur celui d'entraînement. Dans le cas de l'utilisation de l'ensemble d'entraînement, il existe le risque d'obtenir

des modèles sous-optimaux, ne se généralisant pas bien à l'ensemble de test car surentraînés sur les données d'entraînement. Quant à l'utilisation des données de test pour l'optimisation des hyperparamètres, cela compromet l'évaluation finale, car les données d'évaluation ont été utilisées dans le processus d'apprentissage.

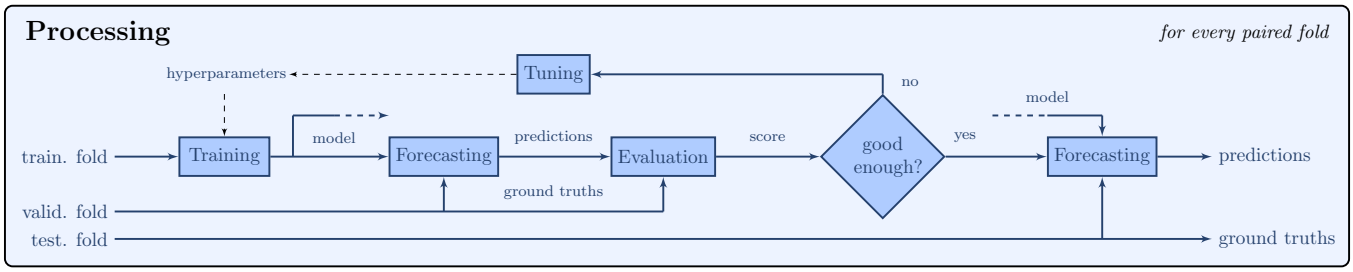


FIGURE 4.2: Étapes de traitement des données (entraînement, optimisation et utilisation du modèle).

La plupart des modèles testés dans cette étude ont déjà été utilisés dans le cadre de la prédiction de la glycémie. Bien qu'il serait possible d'entraîner des modèles globaux, nous avons choisi ici de personnaliser les modèles aux différents patients. Ainsi, pour chaque patient, un modèle différent est entraîné en l'utilisant que les données de ce même patient. Bien que la méthodologie d'entraînement varie d'un modèle à un autre, chaque modèle est entraîné à minimiser l'erreur quadratique moyenne (acsmse) des prédictions.

Quant au choix des hyperparamètres, il ne serait pas juste de les choisir arbitrairement sur l'état de l'art de la prédiction de glycémie, car les spécificités expérimentales sont différentes (e.g., les données utilisées ou les étapes de prétraitement associées). Pour obtenir les meilleures performances possibles pour chaque modèle, nous devons optimiser les hyperparamètres pour chaque patient et pour chaque combinaison d'ensembles d'entraînement et de validation. En raison du nombre élevé de modèles et de patients dans cette étude, il serait trop coûteux d'effectuer une recherche par grille (*grid search*) uniforme et détaillée pour tous les hyperparamètres. Ainsi, nous proposons l'automatisation de l'optimisation des hyperparamètres basée sur une méthodologie d'affinage progressif :

1. Sur la base de l'état de l'art, des limites grossières pour chaque espace de recherche d'hyperparamètres sont identifiées.
2. Chaque espace de recherche identifié basé sur l'état de l'art est affiné pour mieux correspondre aux données utilisées.
3. Une recherche par grille peu précise est effectuée sur l'espace de recherche identifié. En fonction de l'hyperparamètre, l'échelle de la recherche peut être linéaire ou logarithmique.
4. La meilleure valeur de chaque hyperparamètre de l'étape précédente est affinée par une recherche locale.

4.2.3 Modèles de référence

Cette section présente les neuf modèles qui constituent cette base de résultats de référence : le modèle Ref faisant une prédiction de glycémie naïve, trois modèles de régression traditionnels (Poly, AR, ARX), deux modèles

de régression non linéaire plus complexes (SVR, GP) et trois modèles basés sur des réseaux de neurones (ELM, FFNN, LSTM). Pour chaque modèle, lorsque l'un de ses hyperparamètres est dit être optimisé dans une plage donnée, il est optimisé selon la méthode d'affinage progressif présenté précédemment au sein de cette plage.

— Le modèle **Ref** [14, 110, 114, 115, 126] prédit une valeur de glucose égale à la valeur au moment où la prédiction est faite. Cette prédiction peut être modélisée par l'Équation 4.1, où y_i et $\hat{y}_i$ représentent respectivement une observation et une prédiction de glycémie à l'instant i, et où PH est l'horizon de prédiction). Ce modèle ne nécessite aucun entraînement ni optimisation. Il sert de modèle de référence pour comparer les performances des autres modèles.

$$\hat{y}_{t+PH} = y_t \tag{4.1}$$

— Le modèle **Poly** [139] est un modèle de régression polynomiale utilisant uniquement l'heure de la prédiction comme entrée dans le modèle. Il n'est pas attendu que ce modèle soit très performant, en raison de la grande variabilité quotidienne des individus (notamment concernant l'heure et la composition des repas). L'ordre du modèle est optimisé dans la plage $[10^0, 10^2]$.

— Les modèles **AR** [140, 56, 44, 5, 4, 125, 164, 167] et **ARX** [24, 77, 154, 105, 167, 95, 96, 169] proviennent tous deux de la famille des processus autorégressifs ARIMAX. Les modèles ARIMAX ont 3 hyperparamètres : l'ordre de régression endogène p, l'ordre d'intégration d, l'ordre de la moyenne mobile q. Pour les modèles AR et ARX, nous avons optimisé p dans la plage $[1, 36]$ car il représente l'historique du glucose disponible (celui de la dernière heure) pour le modèle, gardé d à 0 et fixé q à 0 puisque l'ajout d'une composante MA ne semblait pas améliorer les performances. Le modèle ARX diffère du modèle AR par l'utilisation d'entrées exogènes supplémentaires (prises de glucides et de bolus d'insuline). La quantité d'entrées exogènes données au modèle ARX correspond à l'ordre p du modèle.

— Le modèle **SVR** (machines à vecteurs de support pour la régression) [14, 59, 130, 4, 60, 63, 64, 69, 95, 96, 111, 114, 143, 169] a été implémenté à l'aide dun noyau gaussien qui a été utilisé pour transformer l'espace d'entrée. Nous avons optimisé le coefficient du noyau dans la plage $[10^{-5}, 10^{-2}]$. Dans un modèle SVR, seules les erreurs de prédictions supérieure à un seuil donné pénalisent le modèle dans son apprentissage. Alors que la pénalité en elle-même a été optimisée dans $[10^{-2}, 10^3]$, la seuil de pénalité a été optimisé dans $[10^{-3}, 10^0]$.

— Le modèle **GP** (processus gaussien) [148, 60] utilise un noyau linéaire (produit scalaire) couplé à un noyau à bruit blanc. Ce noyau peut être représenté par l'Équation 4.2 où c est la constante d'inhomogénéité du noyau et σ^2 une constante de bruit ajouté à la diagonale de la matrice de covariance du noyau. Nous avons choisi ce noyau car il a montré de meilleurs performances empiriques sur l'ensemble de validation que d'autres noyaux, notamment le noyau gaussien utilisé dans le modèle SVR. En particulier, l'ajout du bruit aux observations

permet de faciliter l'entraînement du modèle. Dans notre étude, le coefficient d'inhomogénéité à été fixé à 1 et le bruit a été optimisé dans la plage $[10^{-3}, 10^{2}]$.

$$k(x_n, x_m) = \langle x_n, x_m \rangle + c + \epsilon, \quad \text{avec} \quad \epsilon = \sigma^2 \quad \text{si} \quad x_n \equiv x_m \tag{4.2}$$

— Le modèle **ELM** (*Extreme Learning Machines*) [61, 77, 97, 4] n'a que 3 hyperparamètres : le nombre de neurones et leur fonction d'activation dans sa seule couche cachée, ainsi que la pénalité L2 appliquée aux poids pour la régularisation. Alors que le nombre de neurones a été optimisé dans la plage $[2000, 20000]$, nous avons choisi la fonction d'activation logistique pour ses performances sur l'ensemble de validation. Enfin, la pénalité L2 a été recherchée dans la plage $[10^{0}, 10^{3}]$.

— Le modèle **FFNN** (*Feed-forward Neural Network*) [164, 163, 60, 4, 5, 10, 63, 77, 95, 96, 111, 122, 125, 166, 169] est composé de 4 couches cachées de, respectivement, 128, 64, 32 et 16 neurones. La fonction d'activation des neurones est la *Scaled Exponential Linear Unit* (SELU), qui est l'ELU avec une valeur optimisée de α [81] et pondérée par le coefficient λ (voir Équation 4.3). En comparaison à d'autres fonctions d'activation standards comme la ReLU, elle procure la capacité au réseau de s'auto-régulariser. Nous l'avons choisie pour son gain en performance empirique. Le modèle est ensuite entraîné par *mini-batchs* (1500) en utilisant l'optimiseur Adam [80] avec l'erreur quadratique moyenne (MSE) comme fonction de coût et un taux d'apprentissage initial de 10^{-3}. Pour éviter que le modèle ne soit trop sur-ajusté à l'ensemble d'entraînement, nous avons arrêté l'entraînement du modèle après 100 époques successives sans amélioration sur l'ensemble de validation (arrêt anticipé, *early stopping*, avec une patience de 100 époques).

$$SELU(x) = \lambda \begin{cases} x & \text{si } x > 0 \\ \alpha e^{x} - \alpha & \text{si } x \leq 0 \end{cases}, \quad \text{avec} \quad \begin{cases} \alpha \approx 1.67326324235437728481704299167 17 \\ \lambda \approx 1.05070098735548049341933498529 46 \end{cases} \tag{4.3}$$

— Le modèle **LSTM** [5, 53, 114, 143, 95, 110, 113, 115] est composé de 2 couches cachées de 256 unités LSTM. Le modèle est entraîné avec l'optimiseur Adam et la fonction de coût MSE. Des *mini-batchs* de taille 50 et un taux d'apprentissage automatiquement recherché dans la grille $[10^{-4}, 10^{-3}]$ ont été utilisés. En ce qui la régularisation, nous avons appliqué des pénalités L2 aux poids (10^{-4}) et utilisé un arrêt anticipé de l'apprentissage (patience de 50).

4.2.4 Évaluation des modèles prédictifs

Une fois les modèles entraînés et leurs hyperparamètres optimisés, nous pouvons passer à la phase d'évaluation des modèles, dont la Figure 4.3 donne une représentation graphique.

Pour cela, après avoir calculé les prédictions sur les ensembles de test, nous devons tout d'abord effectuer une

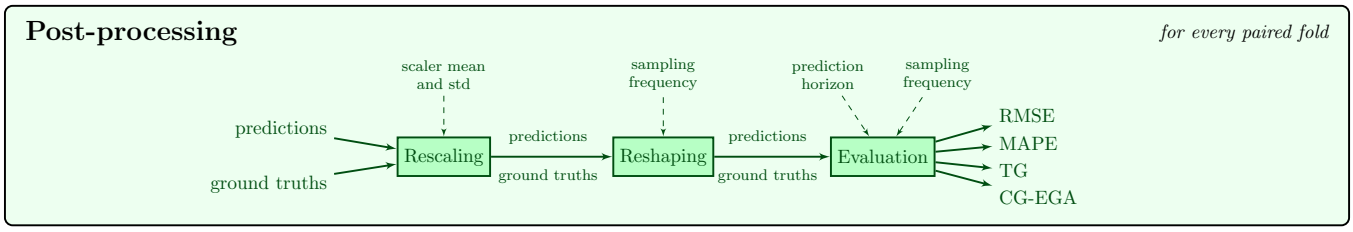


FIGURE 4.3: Étapes de post-traitement des prédictions et d'évaluation des modèles.

série d'étapes de post-traitement des prédictions :

— ***Rescaling*** : Pendant la phase de prétraitement, les valeurs de glucose ont été mises à l'échelle pour avoir une moyenne nulle et une variance unitaire. Pour évaluer correctement les prédictions, nous devons d'abord les ramener (*rescaling*) à leurs moyenne et variance d'origine.

— ***Reshaping*** : Certaines métriques opèrent sur des prédictions qui doivent être temporellement ordonnées. Ainsi, nous devons reconstruire la chronologie de prédiction en fonction de l'intervalle de prédiction. L'intervalle de prédiction est la plus grande valeur entre l'intervalle de temps entre chaque donnée d'entrée (ici 5 minutes) et l'intervalle d'échantillonnage des valeurs de glycémie par les capteurs utilisées (1, 5, et 15 minutes pour les jeux T1DMS, OhioT1DM et IDIAB respectivement). Ainsi, les jeux T1DMS et OhioT1DMS sont évalués toutes les 5 minutes, et le jeu IDIAB toutes les 15 minutes.

Pour évaluer des modèles, en se basant sur l'état de l'art du Chapitre 2, nous avons choisi la RMSE, la MARD, le TG, et la CG-EGA :

— **Racine carrée de l'erreur quadratique moyenne, *Root Mean Squared Error* (RMSE) :** La RMSE est la métrique standard de l'évaluation des modèles prédictifs de glycémie. Elle permet de mesurer la précision des modèles prédictifs.

— **Erreur absolue moyenne en pourcentages, *Mean Absolute Percentage Error* (MAPE) :** Aussi connue sous le nom de *mean absolute relative deviation* (MARD), la MAPE est la deuxième métrique la plus utilisée pour la prédiction de la glycémie. Bien qu'évaluant aussi la précision des modèles, elle est complémentaire de la RMSE. En effet, la MAPE est indépendante de l'échelle de prédiction (permettant la comparaison des performances entre différents patients d'être plus pertinente) et s'exprime en pourcentage. Il est considéré qu'un modèle avec une MAPE inférieure à 10% n'est pas assez précis [23].

— **Gain temporel, *Time Gain* (TG) :** Le TG permet d'estimer, en minutes, le temps gagné, pour le patient diabétique, grâce aux prédictions. Plus celui-ci est important, plus il permet d'anticiper les variations futures de glycémie. Cela est particulièrement intéressant dans l'optique d'éviter les évènements d'hypoglycémie ou d'hyperglycémie. Nous avons préféré sa forme *time gain* plutôt que celle du *time lag* qui nous semblait moins intuitive.

— ***Continuous Glucose-Error Grid Analysis* (CG-EGA) :** La CG-EGA permet de mesurer l'acceptabilité clinique des prédictions. Pour cela, elle classifie une prédiction comme étant cliniquement précise (AP), comme étant une erreur bénigne (BE) ou comme étant une erreur grave (EP). Nous préférons utiliser la CG-EGA plutôt que la P-EGA car cette première pénalise les erreurs cliniques de variations prédites en plus des erreurs cliniques de prédiction. En effet, ces erreurs cliniques témoignent d'un manque de cohérence entre les prédictions successives, incohérence étant potentiellement dangereuse pour le patient.

Pour plus d'informations, quant au calcul de ces différentes métriques, se référer au Chapitre 2, Section 2.4.2.

4.3 GLYFE, un logiciel open source

Dans le but de rendre les résultats reproductibles et de promouvoir son utilisation à l'avenir, nous avons publié le code source de GLYFE dans un dépôt GitHub [25]. La Figure 4.4 représente l'état et la structure du dépôt.

Pour utiliser le code source GLYFE, l'utilisateur doit d'abord obtenir les données OhioT1DM et T1DMS. Alors que le jeu de données OhioT1DM est accessible via [109], les données T1DMS doivent être simulées par l'utilisateur à l'aide du logiciel T1DMS (v3.2.1) dans Matlab (R2018a). Pour garantir que les données simulées sont identiques aux données utilisées dans le benchmark, nous fournissons dans le dépôt GitHub un didacticiel étape par étape qui comprend : le fichier du scénario de simulation, les paramètres du simulateur, la graine aléatoire (*random seed*) et la somme de contrôle SHA-256. En particulier, la graine aléatoire est mise deux fois à 1 : d'abord, avant de lancer le simulateur en exécutant la commande *rng(1,"twister")* dans la console Matlab, puis directement dans l'interface du simulateur avant de lancer la simulation.

Le benchmark a été écrit en Python 3.7 à l'aide de bibliothèques standards d'apprentissage automatique telles que Scikit-learn (modèles Poly, SVR, GP et ELM) [124], statsmodels (modèles AR et ARX) [112], PyTorch (modèles FFNN et LSTM) [123]. À l'instar des Figures 4.1, 4.2 et 4.3, le code source est composé de trois modules principaux appelés *preprocessing*, *processing* et *postprocessing*. Le module *preprocessing* contient les fonctions générales de prétraitement illustrées dans la Figure 4.1 ainsi que les fonctions spécifiques au jeu de données. Le module *processing* est composé de la boucle générale d'entraînement par validation croisée, ainsi que des implémentations des modèles avec leurs hyperparamètres respectifs. Enfin, le module *postprocessing* contient toutes les implémentations des métriques ainsi que des outils pratiques pour la visualisation des résultats.

Dans l'ensemble, la mise en œuvre de l'ensemble de la chaîne de traitement décrit par les Figures 4.1, 4.2 et 4.3, a été rendu flexible de telle sorte que de nouveaux modèles, de nouvelles métriques, de nouvelles étapes de prétraitement ou même de nouvelles données peuvent facilement être inclus tout en préservant l'intégrité du benchmark.

Du point de vue de la thèse, cette architecture a été reprise pour l'ensemble des projets dont elle a fait question rendant ainsi le développement de nouvelles idées très rapide.

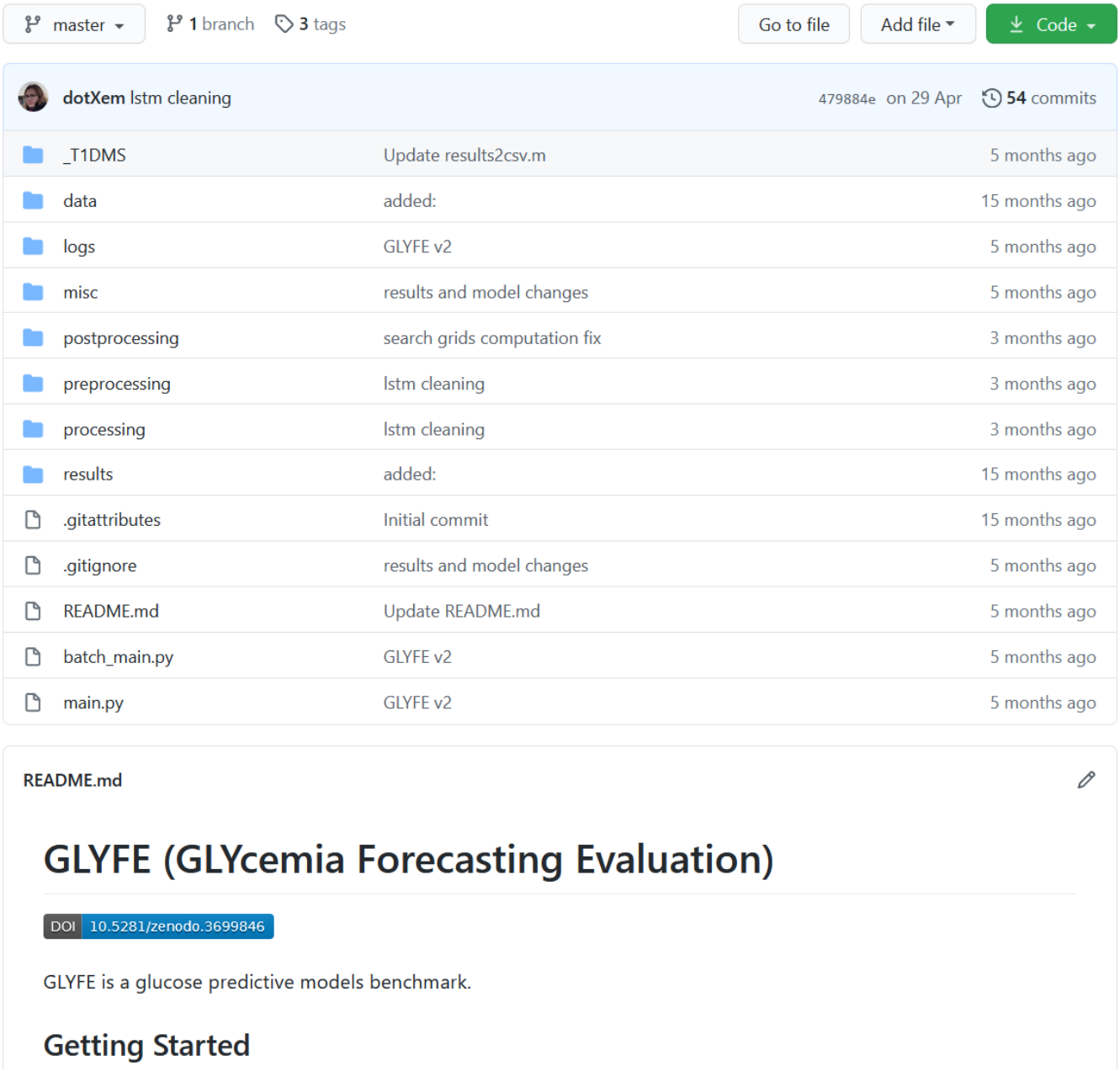


FIGURE 4.4: Dépôt GitHub de GLYFE [25].

4.4 Résultats expérimentaux

4.4.1 Présentation des résultats

Dans cette section, nous présentons et analysons les résultats du benchmark à travers les Tableaux 4.2, 4.3, 4.4 évaluant la résultats généraux des modèles (RMSE, MAPE, TG, ainsi que CG-EGA générale), ainsi qu'avec les Tableaux 4.5, 4.6 et 4.7 mesurant leur acceptabilité clinique détaillée (CG-EGA par région) [1] :

1. Étant en nombre important, les tableaux ont été positionnés à la fin du chapitre pour faciliter la lecture de celui-ci.

— **Poly** : Tout d'abord, le modèle Poly est le moins performant de tous les modèles avec une précision (RMSE et MAPE) et une acceptabilité clinique générale mauvaises (CG-EGA générale) pour tous les jeux de données et horizons de prédiction. Il est incapable de détecter les hypoglycémies (taux d'EP proche de 100%) et est dangereux en région d'hyperglycémie (haut taux d'EP). Toutefois, il est le meilleur modèle du point de vue du gain temporel TG ainsi que de l'acceptabilité clinique en région d'euglycémie.

— **AR et ARX** : Les résultats affichés par les modèles AR et ARX sont assez similaires, l'ARX étant légèrement meilleur. Cela montre l'intérêt d'utiliser des informations supplémentaires, telles que les bolus d'insuline ou les prises de glucides passés, pour prédire les futures valeurs de glycémie. Cependant, les deux modèles sont nettement surclassés par les modèles restants. Ils sont moins précis (RMSE et MAPE plus élevées), offrent moins d'anticipation (TG inférieur) et sont globalement moins sûrs (CG-EGA).

— **SVR et GP** : Dans l'ensemble, les modèles SVR et GP sont les modèles les plus précis (RMSE et MAPE) évalués dans cette étude. Bien que les modèles soient équivalents du point de vue de la précision, le modèle SVR possède une acceptabilité clinique supérieure avec un taux d'AP général et par région plus haut ainsi qu'un taux d'EP général et par région plus bas. Cela fait du modèle SVR le meilleur modèle étudié dans cette étude pour la prédiction de la glycémie.

— **ELM** : Le modèle ELM possède de mauvaises performances à court horizon de prédiction (30 minutes). Mis à part pour le jeu de données T1DMS, il possède une précision inférieure au modèle de référence Ref. Nous notons que les dégradations des performances du modèle avec l'augmentation de l'horizon de prédiction sont plus faibles que pour les autres modèles le rendant compétitif avec les autres modèles à un horizon de 120 minutes.

— **FFNN** : Le modèle FFNN montre de bonnes performances pour le jeu de données T1DMS, meilleures que les modèles autorégressifs AR et ARX mais moins bonnes que celles des modèles SVR et GP. Appliqué aux jeux de données IDIAB et OhioT1DM, ses performances sont sensiblement moins bonnes en comparaison avec les autres modèles. Comme pour le modèle ELM, nous notons que le modèle FFNN est relativement meilleur à des horizons de prédictions plus longs-termes.

— **LSTM** : Le modèle LSTM est le meilleur modèle basé sur l'apprentissage profond analysé dans cette étude. Ses performances sont comparables avec le modèle FFNN sur le jeu T1DMS et comparativement meilleures sur les jeux IDIAB et OhioT1DM. En particulier, il se hisse au niveau des modèles GP et SVR en termes de précision. Néanmoins, il possède une acceptabilité clinique légèrement en dessous de ces modèles pour l'ensemble des régions glycémiques.

4.4.2 Discussion

Analyse des performances entre horizons de prédiction

Pour l'évaluation des performances des capteurs de glycémie en continu, le critère d'avoir une MAPE sur l'erreur de mesure inférieure à 10% est souvent utilisé [23]. En considérant la prédiction de la glycémie comme une mesure de la glycémie future, nous pouvons appliquer ce critère pour analyser la performance des modèles. Selon ce critère, seuls les horizons de 30 minutes pour les trois jeux de données ainsi que celui de 60 pour le jeu T1DMS possèdent des prédictions suffisamment précises. Cela nous montre que la tâche de la prédiction de la glycémie est particulièrement difficile, notamment à des horizons moyens et longs-termes en raison des potentiels évènements pouvant avoir lieu dans l'intervalle de temps (e.g., un repas non anticipé).

Analyse des performances entre jeux de données

De manière générale, quelleque soit la métrique utilisée, les résultats sont meilleurs sur le jeu de données T1DMS que sur les jeux IDIAB et OhioT1DM. Cela s'explique par la plus grande simplicité des données virtuelles du simulateur T1DMS. En comparaison avec des données réelles, celles-ci ne prennent pas en compte d'autres facteurs importants pour la régulation de la glycémie comme l'activité physique ou l'état émotionnel du patient. Toutefois, dans l'ensemble, nous notons que les performances relatives des modèles au sein du jeu T1DMS sont comparables à celles des jeux IDIAB ou OhioT1DM. Cela suggère que, malgré la plus grande simplicité des données simulées du logiciel T1DMS, celles-ci peuvent être utilisées pour la recherche de modèles plus performants de prédiction de glycémie. Quant aux jeux IDIAB et OhioT1DM, tous les modèles montrent avoir des performances similaires sur les deux jeux de données. Ce résultat est particulièrement intéressant de la perspective des patients diabétiques de type 2 présents dans le jeu IDIAB. En effet, comme l'état de l'art a pu nous le montrer, les personnes de diabétiques de type 2 ne sont que très rarement utilisées pour l'évaluation des modèles prédictifs en raison de la plus grande inter variabilité de la population de type 2. Ces résultats nous montrent que les avancées faites sur le type 1 peuvent se transférer à la population de type 2, car les résultats sont comparativement similaires. Nous notons tout de même que le jeu IDIAB possède une meilleure acceptabilité clinique moyenne que le jeu OhioT1DM. Nous imputons cela à l'utilisation de capteurs de glycémie différents. Tandis que les patients du jeu IDIAB utilisent le FreeStyle Libre de Abbott faisant des mesures toutes les 15 minutes, le du jeu OhioT1DM fait des mesures toutes les 5 minutes. L'intervalle de mesure du FreeStyle Libre plus grand lui permet d'être moins impacté par le bruit de mesure, et ainsi d'être plus facile à prédire. Nous pouvons valider cette hypothèse en regardant les résultats du modèle de référence Ref pour les deux jeux. Alors que la précision (RMSE et MAPE) est sensiblement similaire, l'acceptabilité clinique (taux AP et EP généraux) est sensiblement meilleure pour le jeu IDIAB. Cependant, bien que l'acceptabilité clinique soit meilleure, le FreeStyle Libre ne permet pour le jeu IDIAB de ne faire qu'une prédiction toutes les 15 minutes, contre toutes les 5 minutes pour le jeu OhioT1DM.

Performances des modèles prédictifs

Parmi les modèles prédictifs, seuls les modèles Poly et ELM ne possèdent pas une précision suffisante pour la prédiction de la glycémie. En effet, soit ces modèles possèdent une MAPE supérieure au modèle Ref, soit supérieure à 10% [23]. Pour le modèle Poly, cela s'explique par sa grande simplicité, celui-ci n'utilisant qu'une information temporelle pour faire les prédictions. Cela montre que la variabilité quotidienne est trop importante pour un modèle simple seulement basé sur le temps. Cela montre également que les données simulées sont suffisamment irrégulières, grâce à la randomisation des quantités et horaires des repas, ainsi que des prises d'insuline. Quant aux performances du modèle ELM, nous imputons ses mauvaises performances au processus de randomisation des poids de la couche cachée du réseau. Toutefois, le modèle conserve l'avantage d'être très rapide à entraîner et à utiliser ce qui est un avantage dans le cadre de l'utilisation en temps réel des modèles prédictifs de glycémie. Cet avantage pourrait être mis à profit à travers une approche ensembliste pour améliorer les performances [133].

Bien que le réseau de neurones FFNN possède une meilleure précision que le modèle Ref ainsi qu'une MAPE inférieure à 10%, il fait partie des moins bons modèles de cette étude. En particulier, il est le seul à posséder une grande différence de performances entre les données simulées et les données réelles. Tandis qu'il possède de très bonnes performances sur le jeu T1DMS, en particulier à horizon de prédiction lointain (RMSE, MAPE et TG), ses performances sur les jeux IDIAB et OhioT1DM sont moins bonnes que les modèles linéaires AR et ARX. Ces différences de performances nous montrent que le modèle possède des difficultés avec des données plus bruitées et pour lesquelles il manque des informations explicatrices (e.g., activité physique pour les jeux réels).

En revanche, nous ne retrouvons pas ces difficultés chez le modèle LSTM qui possède de très bonnes performances générales, tant bien en précision qu'en acceptabilité clinique. Il reste toutefois moins performant que les modèles SVR et GP, soulignant la difficulté générale que peut avoir les réseaux de neurones pour le tâche de la prédiction de la glycémie. Cette limitation peut provenir d'un besoin plus important en données pour de tels modèles.

Enfin, les modèles SVR et GP basés sur l'utilisation de noyaux montrent les résultats les plus prometteurs de cette étude pour la prédiction de la glycémie. Relativement équivalents du point de vue de leur précision, le modèle GP possédant généralement la meilleure RMSE et le modèle SVR la meilleure MAPE, les deux modèles ne montrent pas la même acceptabilité clinique. En effet, pour tous les jeux de données, à un horizon de prédiction de 30 minutes, le modèle SVR a systématiquement un plus haut taux AP et plus bas d'EP dans toutes les régions de la CG-EGA. En excluant les modèles Ref, Poly et ELM pour leur incapacité à prédire la glycémie future de personnes diabétiques, le modèle SVR possède généralement la meilleure acceptabilité clinique de tous les modèles étudiés.

Analyse des métriques utilisées

Dans l'ensemble, nous notons que chaque modèle n'est pas nécessairement bon dans chaque métrique utilisée. La métrique du gain temporel TG est, sans doute, le cas le plus marquant de ce phénomène. En effet, le modèle

Poly est à la fois le modèle le moins précis, mais aussi le modèle possédant le meilleur gain temporel. Cela n'est pas étonnant du point de vue du calcul de la métrique TG qui ne s'intéresse qu'au déphasage du signal, sans tenir compte de la précision du modèle. À l'inverse, le modèle FFNN possède, pour un horizon de prédiction de 120 minutes sur le jeu T1DMS, à la fois une très bonne précision et un très bon gain temporel. Ainsi, la métrique TG doit être utilisée avec prudence dans l'évaluation des modèles prédictifs, un modèle possédant un meilleur TG n'étant pas nécessairement meilleur.

Du point de vue de l'acceptabilité clinique des modèles, nous notons que les modèles les plus précis ne sont pas nécessairement les plus cliniquement acceptables. Par exemple, le modèle GP possède une des meilleures précisions, mais aussi une acceptabilité clinique parfois moins bonne que d'autres modèles moins précis (e.g., modèle SVR à horizon de prédiction de 30 minutes pour le jeu IDIAB). Aussi, le modèle de référence Ref possède une très bonne acceptabilité clinique, souvent meilleure que celle des modèles entraînés sur les données réelles des jeux IDIAB et OhioT1DM. Cela s'explique par la nature de la CG-EGA qui possède des critères d'évaluation différents. Tout d'abord, la CG-EGA n'évalue pas seulement la précision des modèles, mais aussi leur cohérence à travers la précision des variations prédites. En effet, pour assurer la sécurité d'un patient diabétique utilisant un modèle prédictif de glycémie, celui-ci doit être à la fois précis, mais aussi être cohérent d'une prédiction à une autre pour ne pas perturber le patient dans sa compréhension des prédictions. De plus, la précision des prédictions n'est pas évaluée de la même manière par la CG-EGA et la RMSE ou MAPE. En effet, certaines erreurs de prédiction (e.g., prédictions en région d'hypoglycémie) sont plus graves que d'autres et ne sont ainsi pas prises en compte de la même manière dans la CG-EGA.

Variabilité des résultats

Avec l'aide de la figure 4.5, nous fournissons plus d'informations sur la variabilité intra/inter patient pour les trois ensembles de données T1DMS, OhioT1DM et IDIAB. Chaque jeu de données possède des patients pour lesquels il est plus ou moins facile de faire les prédictions de glycémie ainsi que des patients possédant une plus grande variabilité quotidienne. Néanmoins, nous pouvons voir que la variabilité inter et intra patients est plus importante pour les jeux de données réels IDIAB et OhioT1DM que pour le jeu simulé T1DMS. Cette différence de variabilité peut s'expliquer par plusieurs facteurs. Premièrement, les vrais signaux de glucose ne sont pas seulement affectés par les injections d'insuline et les apports en glucides (comme le sont les signaux de glycémie simulés), mais également par d'autres facteurs tels que l'activité physique, le sommeil ou les émotions. Deuxièmement, il y a beaucoup de données manquantes dans jeux OhioT1DM et IDIAB, ce qui rend les prédictions pour certains jours très difficiles à faire. Cela étant, puisque les performances relatives des modèles sont les mêmes pour tous les jeux de données, nous pouvons conclure que l'évaluation de la capacité de prédiction des modèles peut être effectuée sur les données simulées T1DMS même si ses données sont synthétiques et donc pas entièrement représentatives des données réelles.

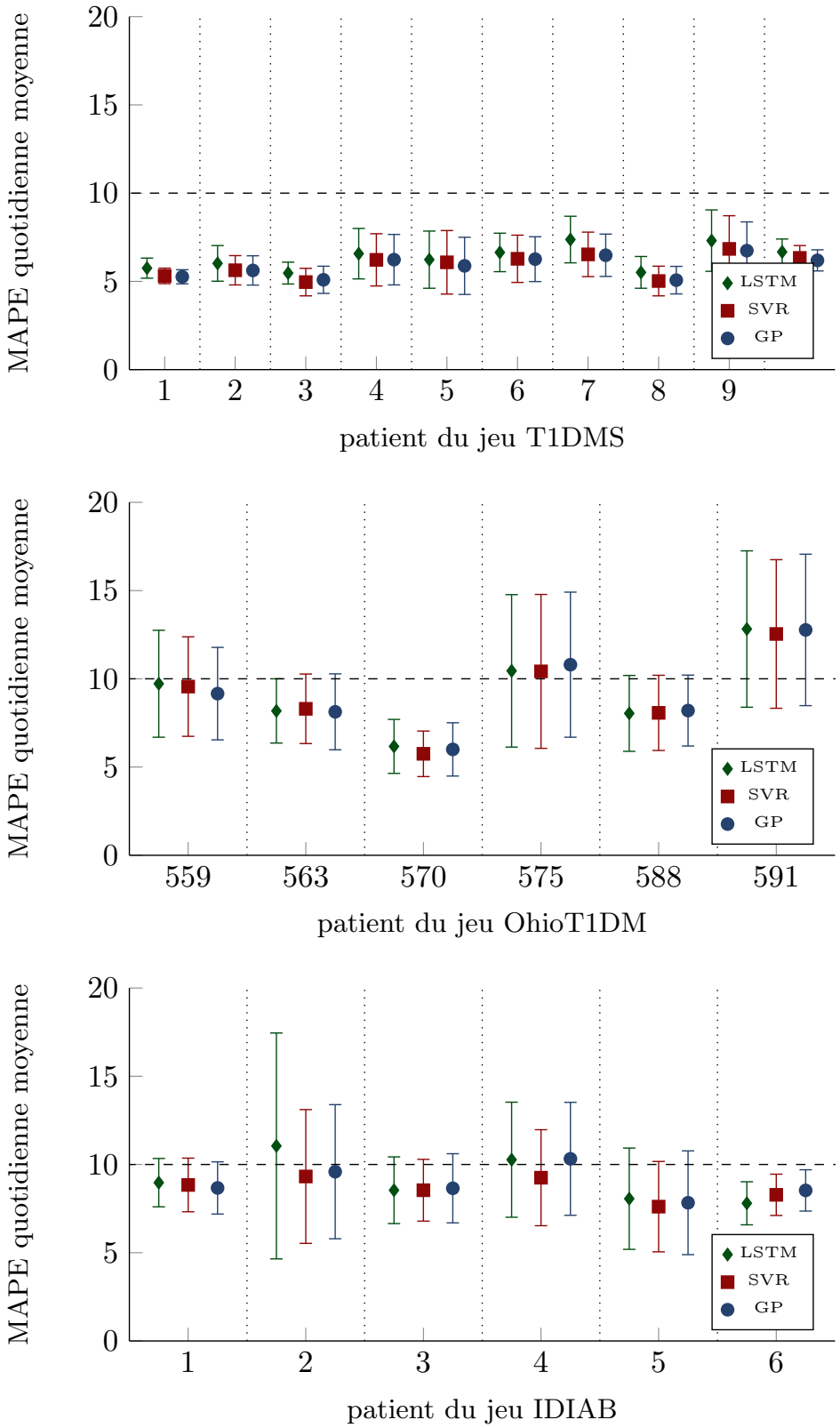


FIGURE 4.5: MAPE quotidienne moyenne (avec écart type) des prédictions à un horizon de 30 minutes par patient pour les jeux de données T1DMS (haut), OhioT1DM (milieu) et IDIAB (bas).

Comparaison avec l'état de l'art

Étude	Modèle	Données												RMSE (PH=30)
		Glucose	Insuline	Glucides	Activité Physique	Temps	Sommeil	Évènement	Humeur	Descripteurs Phys.	Descripteurs Stat.	*Over-sampling*	*Transfer Learning*	
Bertachi *et al.* [10]	ensemble FFNN	✓	✓	✓	✓					✓	✓			19.33 (2.24)
Contreras *et al.* [21]	Grammatical Evolution	✓	✓	✓	✓					✓				21.19 (1.63)
Jeon *et al.* [78]	SPI-GBM	✓	✓	✓	✓	✓	✓	✓	✓		✓			19.20 (2.59)
	LI-GBM	✓	✓	✓	✓	✓	✓	✓	✓		✓			20.86 (1.82)
	ensemble GBM	✓	✓	✓	✓	✓	✓	✓	✓		✓			19.59 (2.20)
Li and Zhu *et al.* [96, 168, 169]	*dilated* CNN	✓	✓	✓		✓							✓	19.28 (2.76)
	causal CNN	✓	✓	✓		✓						✓		22.48 (2.48)
	dilated RNN	✓	✓	✓		✓							✓	18.9
	FFNN	✓	✓	✓										22.93 (2.95)
	SVR	✓	✓	✓										21.75 (1.86)
	LVX	✓	✓	✓										21.70 (3.28)
	ARX	✓	✓	✓										21.54 (2.42)
Mirshekarian *et al.* [115]	LSTM	✓	✓	✓	✓	✓							✓	18.70
	LSTM	✓	✓	✓									✓	18.74
	ARIMA	✓	✓	✓										20.17
	Ref	✓												22.60
Cette étude	Ref	✓												23.40 (2.56)
	Poly				✓									57.27 (6.59)
	AR	✓												20.70 (2.23)
	ARX	✓	✓	✓										20.61 (2.20)
	SVR	✓	✓	✓										20.15 (2.33)
	GP	✓	✓	✓										20.02 (2.32)
	ELM	✓	✓	✓										25.30 (1.37)
	FFNN	✓	✓	✓										21.43 (2.07)
	LSTM	✓	✓	✓										20.46 (2.08)

Tableau 4.1: Comparaison de la RMSE obtenue pour un horizon de prédiction de 30 minutes sur le jeu OhioT1DM par différentes études.

Depuis sa mise à disposition en 2018, le jeu de données OhioT1DM a été utilisé dans plusieurs études. Comme ces études possèdent le même ensemble de test que celui utilisé dans ce benchmark, la plupart des résultats sont comparables. Dans le Tableau 4.1, nous répertorions plusieurs études effectuées sur le jeu OhioT1DM qui sont comparables avec la nôtre. Nous avons exclu les études n'utilisant pas la métrique de RMSE à un horizon de prédiction de 30 minutes [150] ou celles ayant modifié l'ensemble de test [110, 113, 126]. Grâce à la hétérogénéité de ces études, nous pouvons analyser les différences de performances induites par les modèles, par les données ou étapes de prétraitement spécifiques.

Tout d'abord, certaines études utilisent les mêmes modèles et phases de traitement que cette étude. Cela nous permet de comparer directement l'implémentation et optimisation de ces modèles. Dans leur étude, Zhu *et al.* utilisent comme modèles de comparaison les modèles FFNN, SVR et ARX [169]. En comparaison avec ceux que

nous présentons de ces études, leurs performances sont significativement plus faibles. Par ailleurs, Mirshekarian *et al.* utilisent, aussi comme modèles de comparaison, le modèle Ref et ARIMA. Bien que leur modèle Ref et le nôtre soient censés être rigoureusement identiques, nous observons une différence non négligeable (22.60 contre 23.40). Nous pensons que cette différence vient de l'utilisation de l'interpolation des valeurs manquantes lorsque la valeur manquante est la dernière valeur connue. Dans notre étude, nous utilisons l'extrapolation dans ce cas précis, car elle est compatible avec l'utilisation en ligne (données du futur non disponibles) d'un tel dispositif. Nous notons que l'utilisation mixe d'interpolation et d'extrapolation est bien respectée dans l'étude de Zhu *et al.*. Globalement, ces deux études soulignent à la fois la rigueur de notre travail, le besoin du processus automatique d'optimisation des hyperparamètres que nous avons présenté, ainsi que le besoin de manière générale d'utiliser des résultats fiables pour les comparaisons.

Certains des modèles présentés dans ces études montrent de meilleurs performances en RMSE que tous nos modèles. Bien que ces études présentent de nombreuses différences avec la nôtre, nous pouvons dégager quelques points communs et tendances utiles pour le développement de modèles prédictifs plus performants :

— L'étude de Bertachi *et al.*, utilisant un ensemble de plusieurs réseaux de type passe en avant (FFNN), chacun spécialisé dans une région glycémique, nous montre qu'il est possible d'utiliser efficacement ce modèle malgré ses faibles performances dans notre étude. Toutefois, il reste difficile de savoir si ce gain en performances provient de l'utilisation des données supplémentaires d'activité physique, de descripteurs physiologiques (e.g., insuline encore active dans le corps) ou statistiques (e.g., taux de variation de glycémie).

— L'étude de Jeon *et al*, analysant l'impact des différentes techniques d'imputation des données manquantes sur un modèle basé sur une les machines à boosting de gradient (GBM), suggère que l'interpolation Spline (SPI-GBM) est plus efficace que l'interpolation linéaire (LI-GBM) utilisée dans notre étude. Quant au modèle GBM en lui-même, celui-ci propose des résultats prometteurs, bien que nous ne pouvons savoir s'ils proviennent de l'ajout de données diverses (activité physique, temps, sommeil, etc.), des descripteurs statistiques, ou de la nature même du modèle.

— Parmi ces études, les meilleurs modèles sont des modèles basés sur l'apprentissage profond, et en particulier ceux utilisant des réseaux de neurones récurrents. En particulier, les réseaux LSTM de Mirshekarian *et al.* [115] ainsi que le réseau de neurones récurrents dilaté de Zhu *et al.* [169] possèdent une RMSE inférieure à 19. En comparaison avec notre modèle LSTM, possédant une performance significativement plus faible, leurs modèles utilisent une méthodologie d'apprentissage par transfert. L'apprentissage par transfert consiste en l'entraînement d'un modèle profond sur un ensemble de patients (modèle global), puis en son affinage et personnalisation au patient d'intérêt. L'apprentissage par transfert est une technique permettant de faciliter l'apprentissage en présence d'une quantité de données d'intérêt faible. Pour cela, des informations a priori sont apprises sur des données similaires provenant d'une source alternative (e.g., un autre patient), puis

réutilisées dans l'apprentissage sur les données d'intérêt. Ces études couplées à la nôtre suggèrent ainsi que les modèles profonds ne possèdent pas assez de données d'entraînement pour être efficaces, et que l'approche d'apprentissage par transfert est prometteuse pour combattre cette limitation.

4.5 Conclusion et Problématiques

Dans ce chapitre nous proposons la création d'une base de résultats de références pour les modèles prédictifs de glycémie. Ce benchmark répond au besoin d'études comparatives dans le milieu, afin d'évaluer les progrès effectués par les nouvelles architectures. En utilisant deux jeux de données publiquement accessibles et en mettant son code source à disposition, l'étude permet d'accélérer la recherche dans le domaine en réunissant les chercheurs autour de données et d'une architecture communes.

Les résultats obtenus par les neuf modèles utilisés dans cette étude, combiné aux études d'autres groupes de chercheurs, permettent de mettre en évidence plusieurs problématiques :

— Tout d'abord, nous mettons en évidence que la tâche de la prédiction de la glycémie est une tâche particulièrement difficile. En effet, celle-ci n'est efficace que pour des horizons courts (30 minutes), et les données provenant de patients réels comportent beaucoup de variabilité, à la fois entre chaque patient, mais aussi pour un même patient. Bien qu'il soit préférable d'évaluer les modèles prédictifs sur des données réelles, plutôt que simulées par le logiciel T1DMS, nous montrons que celui-ci permet tout de même d'estimer les performances relatives des modèles de manière fiable.

— En comparant les performances de nos modèles aux modèles d'autres études utilisant le jeu de données OhioT1DM, nous mettons en évidence un challenge autour des données utilisées pour l'entraînement des modèles. À cause du besoin en modèles personnalisés aux patients et du coût de récolte de telles données, les modèles, et en particulier les modèles profonds, sont sous-entraînés. Plusieurs pistes peuvent être envisagées pour contourner le fardeau de récolter davantage de données. La première est l'extraction de descripteurs experts permettant de donner plus d'informations aux modèles. Ces descripteurs peuvent à la fois décrire des comportements du corps humain (descripteurs physiologiques), ou bien des informations statistiques sur les données initiales. La seconde piste, seulement applicable aux modèles profonds basés sur les réseaux de neurones, et d'utiliser la méthodologie d'apprentissage par transfert pour faciliter l'entraînement des modèles personnalisés.

— De manière générale, les performances des modèles plaident en faveur de l'utilisation de modèles complexes comme les SVR ou les réseaux de neurones récurrents. Cependant, cette complexification accrue s'accompagne d'une baisse en interprétabilité des modèles. L'interprétabilité des prédictions faites par les modèles est très importante dans ce domaine, comme dans tous les domaines biomédicaux. Elle permet à la fois de comprendre les erreurs de prédictions faites par les modèles et de lever le voile sur leur raisonnement interne,

créant ainsi de la connaissance. Dans le cadre du diabète et de la prédiction de la glycémie, cela permettrait de construire des modèles plus performants, de collecter des données plus pertinentes et ciblées, ainsi que de servir d'outil pour l'éducation thérapeutique du patient.

— Enfin, les résultats nous montrent qu'une amélioration de la précision des prédictions ne se traduit pas nécessairement en une amélioration sur les métriques avec un intérêt plus terrain comme le gain temporel (TG) ou bien l'acceptabilité clinique (CG-EGA). Ces métriques sont plus complexes et n'incluent qu'en partie un critère de précision des prédictions. En particulier, la métrique TG, censée mesurer l'utilité des prédictions à travers la quantification du temps gagné par le mécanisme de prédiction, s'est avérée difficile d'utilisation. Par exemple, le modèle Poly possède une très bonne performance en TG alors qu'il est extrêmement peu précis et dangereux. Quant à elle, la CG-EGA n'évalue pas seulement la précision clinique des modèles, mais aussi la précision clinique des variations prédites. Ainsi, un modèle précis n'est pas garanti d'avoir une bonne acceptabilité clinique à cause de prédictions successives potentiellement incohérentes. Il serait bénéfique d'inclure des critères et contraintes permettant de tenir compte de l'acceptabilité clinique des modèles lors de leur entraînement.

Suite à ces constats, dans la suite de la thèse nous avons choisi de nous concentrer sur les prédictions à 30 minutes dans le futur, ainsi que sur les données réelles des jeux IDIAB et OhioT1DM. Aussi, nous avons décidé de ne pas utiliser la métrique du gain temporel TG qui s'est avérée peu représentative des performances des modèles. Enfin, nous tentons de répondre aux problématiques soulevées. Dans le Chapitre 5, pour améliorer l'acceptabilité clinique des modèles profonds, nous proposons une méthodologie incluant des critères cliniques basés sur la CG-EGA au sein de leur apprentissage. Dans le Chapitre 6, nous explorons l'utilisation de l'apprentissage par transfert pour la prédiction de la glycémie en en proposant une amélioration ainsi qu'en étudiant le transfert entre jeux de données différents. Enfin, dans le Chapitre 7, nous investiguons l'architecture neuronale RETAIN pour la prédiction de la glycémie. Celle-ci se caractérise par un double mécanisme d'attention lui permettant d'être interprétable.

Modèle	RMSE	MAPE	TG	CG-EGA (générale)		
				AP	BE	EP
Horizon de prédiction = 30 minutes						
Ref	17.98 (2.54)	9.77 (1.25)	0.00 (0.00)	95.24 (1.53)	3.48 (1.07)	1.28 (0.74)
Poly	42.92 (15.38)	25.63 (10.55)	**24.40 (2.62)**	91.71 (5.69)	**0.82 (0.44)**	7.47 (5.49)
AR	13.08 (1.13)	7.77 (0.83)	13.30 (2.15)	79.43 (3.29)	17.75 (2.59)	2.81 (0.82)
ARX	11.78 (0.87)	7.21 (0.78)	16.90 (3.73)	85.11 (3.48)	12.71 (2.90)	2.18 (0.82)
SVR	9.06 (0.51)	5.92 (0.62)	22.80 (2.36)	**96.92 (0.63)**	2.17 (0.27)	**0.91 (0.43)**
GP	**9.00 (0.54)**	**5.88 (0.56)**	22.50 (2.50)	94.12 (0.95)	4.65 (0.70)	1.23 (0.38)
ELM	13.20 (2.36)	7.27 (0.79)	23.90 (1.76)	93.81 (1.98)	4.78 (1.32)	1.41 (0.72)
FFNN	10.01 (0.71)	6.34 (0.65)	23.50 (2.11)	96.22 (0.85)	2.76 (0.53)	1.02 (0.49)
LSTM	10.14 (0.78)	6.35 (0.64)	18.60 (2.11)	93.22 (1.12)	5.43 (0.68)	1.35 (0.52)
Horizon de prédiction = 60 minutes						
Ref	29.54 (5.01)	15.96 (2.53)	0.00 (0.00)	92.37 (2.90)	4.51 (1.41)	3.12 (1.73)
Poly	42.92 (15.38)	25.63 (10.55)	**54.40 (2.62)**	91.71 (5.69)	**0.82 (0.44)**	7.47 (5.49)
AR	24.44 (3.32)	13.94 (1.98)	16.50 (2.29)	74.57 (7.10)	19.80 (4.80)	5.63 (2.83)
ARX	22.73 (2.69)	13.38 (2.02)	20.00 (5.20)	75.95 (7.20)	18.84 (4.98)	5.21 (2.80)
SVR	16.18 (1.49)	**9.49 (1.23)**	43.30 (2.24)	**94.53 (1.81)**	3.53 (0.76)	**1.95 (1.36)**
GP	16.32 (1.60)	9.87 (1.24)	42.20 (2.48)	92.04 (2.46)	5.23 (1.26)	2.73 (1.66)
ELM	18.76 (2.70)	10.73 (1.40)	45.30 (3.07)	92.60 (2.66)	5.26 (1.71)	2.14 (1.35)
FFNN	**16.10 (1.51)**	9.84 (1.23)	47.00 (2.10)	92.97 (1.95)	4.94 (0.95)	2.09 (1.30)
LSTM	18.66 (1.95)	11.34 (1.63)	37.70 (4.10)	89.80 (3.26)	7.29 (1.85)	2.91 (1.81)
Horizon de prédiction = 120 minutes						
Ref	44.42 (9.30)	25.27 (5.28)	0.00 (0.00)	89.04 (4.96)	5.26 (1.55)	5.70 (3.86)
Poly	42.92 (15.38)	25.63 (10.55)	**114.40 (2.62)**	**91.71 (5.69)**	**0.82 (0.44)**	7.47 (5.49)
AR	37.42 (7.43)	22.38 (4.87)	20.10 (5.94)	80.11 (11.06)	11.89 (6.63)	8.00 (5.15)
ARX	36.81 (7.02)	22.30 (4.77)	24.30 (7.54)	79.36 (10.64)	12.55 (6.41)	8.08 (4.97)
SVR	28.17 (4.53)	14.57 (2.32)	70.10 (8.19)	91.42 (3.80)	4.70 (1.19)	3.88 (2.79)
GP	28.45 (5.04)	16.48 (3.02)	69.50 (10.38)	88.79 (5.47)	6.18 (2.17)	5.03 (3.68)
ELM	30.15 (5.43)	17.16 (2.98)	87.50 (13.37)	89.36 (5.18)	6.56 (3.06)	4.08 (2.64)
FFNN	**24.00 (3.88)**	**13.69 (1.92)**	108.90 (4.81)	88.15 (4.18)	8.74 (2.75)	**3.11 (1.92)**
LSTM	32.78 (5.81)	20.21 (4.14)	56.00 (9.27)	84.77 (7.03)	9.49 (3.44)	5.74 (3.95)

Tableau 4.2: Précision statistique (RMSE et MAPE), gain temporel (TG) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeu de données T1DMS.

Modèle	RMSE	MAPE	TG	CG-EGA (générale)		
				AP	BE	EP
Horizon de prédiction = 30 minutes						
Ref	23.40 (2.56)	10.96 (2.39)	0.00 (0.00)	**88.94 (3.02)**	8.08 (1.98)	**2.98 (1.46)**
Poly	57.27 (6.59)	31.09 (6.71)	**25.50 (6.87)**	84.29 (3.32)	**5.66 (1.72)**	10.05 (3.41)
AR	20.70 (2.23)	9.62 (2.26)	5.17 (0.37)	81.72 (4.05)	13.50 (3.22)	4.78 (1.79)
ARX	20.61 (2.20)	9.59 (2.19)	4.17 (1.86)	81.49 (4.02)	13.68 (3.16)	4.84 (1.86)
SVR	20.15 (2.33)	**9.12 (2.11)**	5.83 (1.86)	83.35 (3.91)	12.38 (2.83)	4.28 (1.83)
GP	**20.02 (2.32)**	9.19 (2.15)	5.83 (1.86)	81.16 (4.34)	14.26 (3.18)	4.57 (1.85)
ELM	25.30 (1.37)	11.55 (2.46)	5.67 (4.15)	76.39 (4.27)	17.93 (2.83)	5.68 (2.36)
FFNN	21.43 (2.07)	9.82 (2.22)	5.67 (2.49)	76.34 (4.33)	17.95 (3.02)	5.71 (2.28)
LSTM	20.46 (2.08)	9.24 (2.10)	6.00 (1.83)	80.03 (4.17)	14.83 (2.88)	5.14 (2.11)
Horizon de prédiction = 60 minutes						
Ref	37.83 (3.56)	18.57 (3.85)	0.00 (0.00)	85.65 (3.94)	9.08 (1.60)	**5.26 (2.51)**
Poly	57.27 (6.58)	31.04 (6.67)	**56.00 (6.11)**	**84.41 (3.36)**	**5.66 (1.74)**	9.92 (3.45)
AR	33.20 (2.69)	16.73 (3.94)	6.67 (3.73)	78.06 (4.45)	14.69 (2.88)	7.25 (3.03)
ARX	33.43 (2.53)	16.73 (3.95)	5.83 (1.86)	76.95 (4.46)	15.62 (2.85)	7.44 (3.04)
SVR	32.25 (2.42)	**15.45 (3.41)**	9.00 (5.32)	79.10 (4.55)	14.35 (2.70)	6.55 (2.80)
GP	**31.99 (2.56)**	15.97 (3.79)	8.67 (5.37)	77.26 (5.11)	15.69 (3.12)	7.05 (2.94)
ELM	34.94 (1.80)	16.92 (3.72)	11.50 (10.24)	74.56 (4.50)	17.84 (2.32)	7.60 (3.02)
FFNN	33.01 (2.13)	16.05 (3.67)	12.00 (7.14)	72.10 (4.61)	20.10 (3.38)	7.81 (3.04)
LSTM	32.88 (2.66)	16.00 (3.57)	9.50 (4.11)	76.55 (5.31)	15.80 (2.83)	7.65 (3.37)
Horizon de prédiction = 120 minutes						
Ref	57.81 (5.79)	29.53 (5.54)	0.00 (0.00)	81.77 (4.66)	9.83 (1.85)	8.40 (3.15)
Poly	57.26 (6.59)	31.01 (6.64)	**115.33 (7.45)**	**84.38 (3.27)**	**5.67 (1.74)**	9.95 (3.39)
AR	47.48 (3.75)	25.69 (5.79)	9.83 (6.72)	79.23 (3.89)	12.28 (1.85)	8.48 (3.57)
ARX	47.56 (3.57)	25.11 (5.33)	8.33 (7.18)	79.58 (4.45)	11.75 (2.12)	8.67 (3.43)
SVR	47.20 (2.17)	24.28 (5.74)	21.33 (17.48)	79.13 (3.07)	12.68 (2.71)	**8.20 (2.65)**
GP	**46.28 (2.89)**	24.75 (5.71)	19.83 (16.04)	78.35 (4.09)	13.31 (1.95)	8.35 (3.25)
ELM	46.54 (3.04)	24.39 (5.29)	31.00 (24.33)	74.99 (4.43)	16.00 (1.84)	9.02 (3.30)
FFNN	46.85 (2.64)	**24.08 (5.11)**	39.67 (31.96)	70.41 (4.88)	20.34 (2.54)	9.25 (3.05)
LSTM	47.56 (3.05)	24.69 (5.34)	21.33 (15.22)	76.71 (4.43)	14.62 (2.15)	8.67 (3.47)

Tableau 4.3: Précision statistique (RMSE et MAPE), gain temporel (TG) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeu de données OhioT1DM.

Modèle	RMSE	MAPE	TG	CG-EGA (générale)		
				AP	BE	EP
Horizon de prédiction = 30 minutes						
Ref	24.50 (8.54)	10.75 (1.87)	0.00 (0.00)	**94.16 (4.76)**	3.84 (2.80)	2.00 (2.13)
Poly	53.71 (7.03)	28.70 (7.09)	**19.50 (12.93)**	89.13 (4.85)	**3.67 (3.39)**	7.20 (3.10)
AR	20.81 (7.41)	9.38 (1.39)	11.50 (5.59)	91.17 (4.11)	6.46 (2.88)	2.37 (1.85)
ARX	20.44 (6.49)	9.27 (1.18)	12.50 (5.59)	92.11 (3.75)	5.37 (2.01)	2.52 (1.91)
SVR	20.32 (6.02)	**8.66 (0.44)**	12.50 (5.59)	92.69 (2.81)	5.34 (2.06)	**1.97 (1.23)**
GP	**19.80 (5.92)**	8.92 (0.80)	12.50 (5.59)	91.92 (3.14)	5.80 (2.19)	2.29 (1.85)
ELM	26.25 (7.57)	11.82 (1.06)	10.50 (5.68)	91.21 (4.60)	5.59 (2.57)	3.20 (2.47)
FFNN	22.01 (6.26)	9.97 (0.89)	11.00 (5.10)	91.45 (3.77)	5.48 (2.60)	3.07 (2.23)
LSTM	19.85 (6.00)	9.04 (1.11)	12.00 (5.48)	92.20 (2.99)	5.05 (1.71)	2.76 (1.82)
Horizon de prédiction = 60 minutes						
Ref	39.98 (13.82)	18.60 (3.58)	0.00 (0.00)	**89.94 (6.97)**	5.09 (3.02)	**4.97 (4.17)**
Poly	54.15 (7.42)	29.01 (7.15)	**46.00 (16.94)**	88.67 (5.38)	**4.07 (3.67)**	7.26 (3.17)
AR	34.86 (11.26)	16.88 (2.67)	12.50 (5.59)	86.22 (5.93)	7.48 (2.22)	6.30 (3.95)
ARX	34.58 (10.62)	16.85 (2.56)	12.50 (5.59)	86.34 (5.82)	7.41 (1.82)	6.25 (4.07)
SVR	33.72 (8.42)	**15.79 (1.34)**	13.00 (5.92)	87.12 (4.74)	7.22 (1.85)	5.65 (3.26)
GP	**33.31 (9.00)**	16.24 (1.97)	14.50 (7.83)	86.55 (4.88)	7.59 (1.72)	5.87 (3.50)
ELM	35.63 (8.59)	17.32 (1.75)	14.00 (7.07)	88.35 (5.10)	6.24 (2.75)	5.41 (2.42)
FFNN	34.05 (8.01)	16.49 (1.49)	15.00 (7.75)	87.24 (4.43)	7.29 (1.92)	5.47 (2.87)
LSTM	33.70 (7.85)	16.21 (1.71)	16.00 (7.87)	87.88 (4.88)	6.94 (1.95)	5.18 (3.09)
Horizon de prédiction = 120 minutes						
Ref	58.72 (19.98)	30.10 (7.51)	0.00 (0.00)	84.08 (10.50)	6.73 (4.08)	9.19 (6.61)
Poly	54.27 (7.53)	29.11 (7.18)	**105.50 (16.92)**	**88.59 (5.44)**	**4.14 (3.75)**	7.27 (3.18)
AR	49.69 (13.61)	26.64 (5.39)	17.50 (10.45)	86.23 (8.38)	5.83 (3.63)	7.94 (4.94)
ARX	54.42 (14.88)	28.26 (6.88)	16.50 (10.50)	83.30 (8.26)	7.85 (5.41)	8.85 (4.35)
SVR	48.67 (12.43)	25.53 (5.52)	30.50 (23.01)	86.94 (5.60)	6.14 (3.77)	6.93 (3.27)
GP	48.20 (13.39)	25.66 (5.51)	23.00 (16.58)	84.89 (7.34)	7.58 (3.44)	7.53 (4.05)
ELM	**47.39 (10.83)**	**24.80 (4.06)**	25.50 (21.27)	87.62 (5.83)	4.83 (3.62)	7.55 (3.15)
FFNN	47.94 (11.80)	25.12 (4.37)	26.50 (19.93)	85.47 (6.29)	6.90 (3.81)	7.63 (3.26)
LSTM	48.45 (11.86)	25.52 (4.74)	25.50 (22.03)	87.86 (5.91)	5.26 (3.14)	**6.89 (3.35)**

Tableau 4.4: Précision statistique (RMSE et MAPE), gain temporel (TG) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeu de données IDIAB.

Modèle	CG-EGA (par région)								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
Horizon de prédiction = 30 minutes									
Ref	81.92 (9.31)	1.98 (3.45)	16.09 (7.55)	96.62 (0.88)	2.90 (0.70)	0.48 (0.27)	89.82 (6.98)	8.54 (5.69)	1.64 (1.38)
Poly	0.11 (0.33)	**0.00 (0.00)**	99.89 (0.33)	**99.27 (0.33)**	**0.68 (0.31)**	**0.05 (0.06)**	79.88 (7.80)	**2.11 (1.35)**	18.02 (7.57)
AR	63.96 (10.62)	14.65 (7.21)	21.39 (9.03)	79.79 (3.41)	18.01 (2.89)	2.20 (0.59)	79.66 (4.35)	18.00 (3.51)	2.34 (1.00)
ARX	71.51 (7.51)	10.90 (5.84)	17.60 (6.59)	85.86 (3.90)	12.59 (3.60)	1.55 (0.43)	83.15 (2.05)	15.07 (1.66)	1.78 (0.76)
SVR	83.05 (7.11)	2.80 (3.80)	14.15 (6.12)	97.51 (0.47)	2.06 (0.34)	0.44 (0.14)	**96.37 (0.97)**	2.92 (0.70)	**0.71 (0.52)**
GP	78.78 (8.91)	4.70 (3.42)	16.53 (7.95)	94.49 (1.05)	4.74 (0.91)	0.77 (0.26)	94.47 (1.34)	4.57 (1.03)	0.97 (0.50)
ELM	76.74 (13.01)	3.42 (4.02)	19.84 (11.98)	95.23 (1.39)	4.11 (1.15)	0.66 (0.29)	88.95 (2.82)	9.19 (2.15)	1.86 (0.82)
FFNN	80.98 (14.91)	2.15 (2.98)	16.87 (15.72)	96.93 (0.54)	2.56 (0.44)	0.51 (0.17)	95.19 (1.39)	3.98 (1.10)	0.83 (0.39)
LSTM	**85.89 (9.37)**	3.84 (4.18)	**10.27 (5.63)**	93.77 (1.12)	5.32 (0.91)	0.91 (0.32)	91.90 (1.69)	6.77 (1.51)	1.33 (0.66)
Horizon de prédiction = 60 minutes									
Ref	47.43 (21.77)	1.03 (1.69)	51.54 (21.83)	95.79 (1.42)	3.51 (1.14)	0.70 (0.42)	82.17 (7.76)	11.97 (5.00)	5.87 (3.19)
Poly	0.11 (0.33)	**0.00 (0.00)**	99.89 (0.33)	**99.27 (0.33)**	**0.68 (0.31)**	**0.05 (0.06)**	79.88 (7.80)	**2.11 (1.35)**	18.02 (7.57)
AR	21.38 (15.05)	8.31 (6.39)	70.31 (20.77)	77.38 (6.28)	20.07 (5.48)	2.55 (0.87)	70.20 (7.40)	22.98 (4.98)	6.82 (2.58)
ARX	23.17 (16.21)	9.36 (5.82)	67.47 (20.74)	78.63 (6.64)	18.94 (5.79)	2.43 (0.93)	71.93 (6.17)	22.45 (4.39)	5.63 (1.92)
SVR	44.83 (22.51)	1.63 (1.91)	53.55 (23.42)	96.73 (0.74)	3.07 (0.62)	0.20 (0.13)	**91.94 (3.75)**	7.31 (3.05)	**0.75 (0.80)**
GP	35.87 (23.06)	1.97 (2.06)	62.17 (24.30)	94.39 (1.82)	4.97 (1.58)	0.64 (0.29)	89.58 (2.68)	8.87 (2.33)	1.55 (0.76)
ELM	46.14 (22.65)	2.29 (2.51)	51.57 (24.01)	94.91 (1.83)	4.66 (1.61)	0.42 (0.28)	88.73 (3.79)	9.93 (2.83)	1.34 (1.18)
FFNN	**47.85 (22.78)**	1.63 (1.64)	**50.53 (24.04)**	95.12 (0.96)	4.52 (0.85)	0.36 (0.13)	90.41 (2.99)	8.61 (2.42)	0.98 (0.64)
LSTM	43.31 (18.99)	3.54 (2.91)	53.16 (20.07)	92.21 (2.66)	6.89 (2.22)	0.90 (0.47)	85.34 (3.25)	11.96 (3.00)	2.70 (0.90)
Horizon de prédiction = 120 minutes									
Ref	21.47 (16.35)	0.46 (0.61)	78.07 (16.71)	94.69 (2.29)	4.44 (1.49)	0.87 (0.90)	74.96 (7.14)	12.03 (3.99)	13.01 (4.41)
Poly	0.11 (0.33)	**0.00** (0.00)	99.89 (0.33)	**99.27 (0.33)**	**0.68 (0.31)**	**0.05 (0.06)**	79.88 (7.80)	**2.11 (1.35)**	18.02 (7.57)
AR	1.92 (3.19)	1.29 (2.15)	96.79 (5.32)	85.52 (8.74)	12.49 (7.84)	1.99 (1.03)	71.38 (8.94)	14.13 (5.84)	14.50 (4.38)
ARX	2.43 (4.08)	1.33 (2.16)	96.24 (6.21)	84.65 (8.52)	13.22 (7.67)	2.14 (0.95)	70.72 (8.26)	14.60 (5.42)	14.67 (4.51)
SVR	37.34 (20.22)	1.76 (1.46)	60.90 (21.04)	95.45 (1.40)	4.20 (1.12)	0.35 (0.30)	**84.10 (4.71)**	8.91 (2.28)	6.99 (3.05)
GP	17.76 (13.89)	2.20 (2.05)	80.04 (14.74)	93.58 (2.88)	5.75 (2.44)	0.67 (0.45)	82.46 (4.79)	10.73 (2.19)	6.81 (3.22)
ELM	31.93 (17.35)	2.86 (3.13)	65.21 (18.80)	93.43 (3.38)	6.00 (2.88)	0.57 (0.51)	81.60 (6.02)	10.86 (3.47)	7.54 (3.10)
FFNN	**44.21 (12.00)**	6.00 (3.20)	**49.78 (13.63)**	91.09 (2.82)	8.22 (2.56)	0.70 (0.32)	83.06 (4.93)	12.85 (3.70)	**4.09 (1.63)**
LSTM	10.98 (10.94)	2.96 (2.49)	86.06 (12.68)	89.56 (4.90)	9.30 (4.30)	1.14 (0.65)	77.19 (5.20)	14.19 (2.25)	8.63 (3.33)

AP : *Accurate Prediction* ; BE : *Benign Error* ; EP : *Erroneous Prediction*

Tableau 4.5: Acceptabilité clinique (CG-EGA) par région moyenne (avec écart type) des prédictions de glycémie par modèle pour l'ensemble de la population du jeu T1DMS et pour les horizons de prédictions 30, 60 et 120 minutes.

Modèle	CG-EGA (par région)								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
Horizon de prédiction = 30 minutes									
Ref	47.23 (23.39)	3.99 (3.70)	48.77 (23.51)	91.37 (3.22)	6.63 (2.46)	2.01 (0.90)	**86.95 (3.35)**	10.27 (2.44)	**2.78 (1.97)**
Poly	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	**94.54 (1.74)**	**5.20 (1.84)**	**0.27 (0.55)**	75.71 (6.30)	**7.00 (2.82)**	17.29 (5.72)
AR	38.11 (21.40)	5.30 (3.87)	56.59 (22.30)	85.42 (5.40)	11.47 (4.22)	3.10 (1.32)	79.18 (2.98)	16.06 (3.16)	4.75 (1.67)
ARX	38.32 (23.33)	4.88 (3.92)	56.80 (23.69)	85.10 (5.41)	11.67 (4.25)	3.23 (1.34)	78.96 (2.91)	16.26 (3.00)	4.78 (1.69)
SVR	**49.71 (18.75)**	5.62 (4.02)	**44.67 (18.70)**	86.35 (4.24)	10.71 (3.26)	2.94 (1.23)	80.85 (3.24)	14.77 (3.01)	4.37 (1.84)
GP	45.09 (26.13)	6.94 (4.57)	47.97 (27.41)	84.61 (5.37)	12.21 (4.15)	3.17 (1.40)	78.29 (3.53)	16.87 (3.21)	4.84 (1.59)
ELM	30.75 (23.97)	3.93 (3.94)	65.32 (26.93)	79.43 (4.09)	16.93 (2.98)	3.64 (1.47)	74.11 (4.33)	19.98 (3.30)	5.91 (1.78)
FFNN	37.17 (18.07)	6.07 (3.42)	56.76 (17.33)	80.11 (4.57)	15.97 (3.56)	3.92 (1.47)	72.65 (4.21)	21.64 (3.28)	5.71 (1.79)
LSTM	38.37 (23.17)	3.97 (3.72)	57.67 (24.23)	83.78 (5.33)	12.70 (4.06)	3.52 (1.47)	76.86 (3.70)	17.87 (2.73)	5.27 (2.21)
Horizon de prédiction = 60 minutes									
Ref	**24.64 (17.15)**	2.12 (2.41)	**73.23 (19.09)**	89.18 (3.32)	7.69 (1.82)	3.12 (1.56)	**83.24 (4.74)**	11.60 (2.58)	**5.17 (2.30)**
Poly	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	**94.77 (1.85)**	**5.13 (1.90)**	**0.10 (0.18)**	75.87 (6.23)	**6.99 (2.86)**	17.14 (5.74)
AR	7.58 (7.38)	1.67 (2.09)	90.74 (8.74)	83.24 (5.01)	12.71 (3.84)	4.05 (1.34)	74.91 (4.36)	17.71 (2.81)	7.39 (3.07)
ARX	7.67 (7.42)	1.97 (2.30)	90.36 (8.90)	82.12 (4.92)	13.69 (3.81)	4.19 (1.28)	73.78 (4.64)	18.54 (3.10)	7.67 (3.09)
SVR	14.95 (12.97)	3.74 (3.43)	81.32 (14.39)	83.59 (4.38)	12.68 (3.33)	3.73 (1.21)	76.24 (4.55)	17.08 (2.69)	6.68 (2.62)
GP	11.43 (8.96)	2.22 (2.71)	86.35 (11.35)	82.12 (5.44)	13.81 (4.19)	4.08 (1.44)	74.22 (5.15)	18.63 (3.02)	7.16 (2.85)
ELM	9.29 (9.16)	3.21 (3.04)	87.51 (12.14)	78.06 (3.53)	17.40 (2.46)	4.54 (1.31)	73.08 (4.87)	19.38 (2.97)	7.54 (2.63)
FFNN	4.43 (4.64)	1.81 (1.50)	93.76 (5.87)	76.42 (4.24)	19.00 (3.70)	4.59 (1.15)	69.14 (5.31)	23.11 (3.66)	7.75 (2.65)
LSTM	3.48 (5.14)	1.07 (1.09)	95.45 (5.83)	82.40 (5.19)	13.59 (3.89)	4.01 (1.46)	72.33 (6.42)	19.43 (3.19)	8.24 (3.78)
Horizon de prédiction = 120 minutes									
Ref	7.24 (5.21)	0.93 (1.03)	91.83 (6.04)	85.57 (3.68)	8.85 (2.72)	5.57 (1.28)	**80.30 (4.55)**	11.71 (2.19)	**7.99 (2.55)**
Poly	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	**94.80 (1.83)**	**5.10 (1.87)**	**0.10 (0.15)**	75.88 (6.16)	**7.00 (2.89)**	17.12 (5.63)
AR	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	85.55 (3.22)	11.18 (2.57)	3.27 (0.99)	74.94 (5.61)	14.04 (1.86)	11.02 (5.12)
ARX	0.55 (1.23)	0.22 (0.48)	99.23 (1.72)	86.09 (3.54)	10.85 (2.88)	3.06 (0.85)	74.63 (6.55)	13.28 (2.37)	12.09 (5.02)
SVR	**9.27 (11.03)**	2.64 (4.58)	**88.09 (15.43)**	84.39 (3.29)	12.33 (3.55)	3.28 (0.59)	75.38 (4.47)	14.27 (2.59)	10.35 (3.60)
GP	2.28 (3.50)	0.38 (0.80)	97.34 (4.21)	83.90 (3.45)	12.55 (2.77)	3.55 (1.23)	74.89 (5.50)	14.74 (2.18)	10.37 (4.21)
ELM	0.02 (0.05)	0.30 (0.60)	99.68 (0.60)	79.34 (2.46)	16.83 (1.56)	3.84 (1.10)	72.71 (5.57)	15.82 (2.83)	11.47 (4.55)
FFNN	1.05 (1.22)	0.70 (1.33)	98.25 (2.51)	74.32 (2.79)	21.41 (2.50)	4.27 (1.24)	67.97 (6.93)	20.76 (3.32)	11.27 (3.79)
LSTM	1.06 (1.99)	0.91 (2.03)	98.04 (4.01)	82.39 (3.43)	13.92 (2.48)	3.68 (1.34)	72.89 (5.68)	16.29 (2.20)	10.82 (4.41)

AP : *Accurate Prediction* ; BE : *Benign Error* ; EP : *Erroneous Prediction*

Tableau 4.6: Acceptabilité clinique (CG-EGA) par région moyenne (avec écart type) des prédictions de glycémie par modèle pour l'ensemble de la population du jeu OhioT1DM et pour les horizons de prédictions 30, 60 et 120 minutes.

Modèle	CG-EGA (par région)								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
Horizon de prédiction = 30 minutes									
Ref	**76.15 (19.57)**	**0.00 (0.00)**	**23.85 (19.57)**	**97.35 (1.23)**	**2.29 (1.13)**	**0.36 (0.17)**	**90.04 (9.01)**	6.50 (4.91)	3.46 (4.25)
Poly	0.24 (0.47)	**0.00 (0.00)**	99.76 (0.47)	96.06 (3.29)	2.48 (2.23)	1.46 (1.75)	84.17 (9.41)	**5.37 (5.43)**	10.46 (9.02)
AR	53.26 (26.47)	**0.00 (0.00)**	46.74 (26.47)	94.34 (2.12)	5.10 (2.15)	0.56 (0.54)	87.73 (7.10)	9.24 (4.67)	**3.02 (2.89)**
ARX	54.85 (23.44)	**0.00 (0.00)**	45.15 (23.44)	95.09 (1.54)	4.34 (1.31)	0.57 (0.50)	89.16 (6.87)	7.46 (3.88)	3.38 (3.11)
SVR	69.39 (33.51)	0.35 (0.70)	30.27 (33.54)	95.17 (2.01)	4.33 (1.83)	0.50 (0.47)	89.51 (6.09)	7.43 (3.86)	3.06 (2.53)
GP	60.89 (22.05)	**0.00 (0.00)**	39.11 (22.05)	94.67 (2.33)	4.89 (2.26)	0.44 (0.51)	88.83 (5.93)	7.77 (3.53)	3.40 (3.17)
ELM	31.06 (40.09)	0.70 (1.39)	68.25 (39.57)	95.13 (2.28)	4.23 (2.09)	0.64 (0.70)	88.71 (7.29)	7.95 (4.29)	3.34 (3.27)
FFNN	33.88 (37.15)	0.17 (0.35)	65.95 (37.01)	95.09 (2.40)	4.36 (2.32)	0.56 (0.41)	89.26 (6.58)	7.54 (3.75)	3.20 (3.14)
LSTM	47.39 (31.03)	**0.00 (0.00)**	52.61 (31.03)	95.85 (1.27)	3.75 (1.34)	0.39 (0.34)	89.91 (5.08)	7.00 (2.75)	3.09 (2.46)
Horizon de prédiction = 60 minutes									
Ref	**39.43 (27.10)**	**0.00 (0.00)**	**60.57 (27.10)**	95.24 (1.90)	3.53 (1.29)	1.24 (0.73)	84.25 (13.12)	7.84 (5.58)	7.91 (7.73)
Poly	1.18 (2.35)	**0.00 (0.00)**	98.82 (2.35)	**95.47 (4.08)**	**2.97 (3.11)**	1.56 (1.76)	83.82 (9.67)	**5.62 (5.40)**	10.56 (9.11)
AR	6.74 (10.04)	**0.00 (0.00)**	93.26 (10.04)	92.40 (2.68)	6.37 (2.45)	1.24 (0.56)	81.48 (11.31)	9.85 (3.88)	8.67 (7.50)
ARX	9.54 (13.02)	**0.00 (0.00)**	90.46 (13.02)	92.43 (1.82)	6.22 (1.49)	1.35 (0.64)	81.50 (11.60)	9.81 (4.13)	8.69 (7.61)
SVR	17.07 (14.34)	1.27 (1.56)	81.66 (13.32)	92.62 (2.68)	6.00 (2.21)	1.38 (0.82)	82.40 (9.72)	10.09 (4.32)	7.52 (5.76)
GP	13.99 (13.53)	**0.00 (0.00)**	86.01 (13.53)	91.85 (3.21)	6.93 (3.07)	1.22 (0.47)	82.18 (9.38)	9.64 (3.65)	8.19 (6.04)
ELM	12.99 (14.92)	0.24 (0.47)	86.77 (15.24)	94.07 (1.95)	4.90 (1.56)	**1.02 (0.53)**	**84.75 (9.69)**	8.17 (5.63)	7.08 (5.00)
FFNN	13.63 (16.59)	**0.00 (0.00)**	86.37 (16.59)	92.72 (2.29)	5.91 (1.84)	1.36 (0.72)	83.41 (8.84)	9.82 (4.23)	6.77 (4.73)
LSTM	16.25 (20.76)	**0.00 (0.00)**	83.75 (20.76)	93.45 (1.93)	5.42 (1.61)	1.13 (0.54)	83.66 (8.92)	9.74 (4.06)	**6.59 (4.91)**
Horizon de prédiction = 120 minutes									
Ref	**12.89 (13.08)**	**0.00 (0.00)**	**87.11 (13.08)**	90.89 (5.58)	4.72 (2.01)	4.39 (3.67)	77.27 (16.08)	10.52 (6.20)	12.21 (10.44)
Poly	1.18 (2.35)	**0.00 (0.00)**	98.82 (2.35)	95.36 (4.26)	3.08 (3.33)	1.56 (1.76)	**83.81 (9.72)**	**5.62 (5.42)**	10.57 (9.23)
AR	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	94.04 (3.28)	4.49 (2.05)	1.47 (1.28)	79.39 (15.34)	8.10 (5.89)	12.50 (10.30)
ARX	0.00 (0.00)	0.80 (1.60)	99.20 (1.60)	91.25 (6.94)	6.21 (4.88)	2.54 (2.14)	76.68 (14.78)	10.44 (7.29)	12.88 (10.21)
SVR	2.33 (2.94)	0.58 (0.74)	97.08 (3.68)	93.65 (5.06)	4.94 (3.32)	1.40 (1.82)	81.63 (10.11)	7.90 (5.87)	10.47 (7.82)
GP	1.18 (2.35)	**0.00 (0.00)**	98.82 (2.35)	92.40 (4.18)	6.21 (3.30)	1.40 (1.15)	78.73 (14.19)	9.80 (6.33)	11.47 (8.33)
ELM	0.24 (0.47)	0.94 (1.88)	98.82 (2.35)	**95.95 (2.13)**	**3.08 (1.44)**	**0.97 (0.96)**	80.41 (11.81)	7.62 (6.28)	11.96 (8.43)
FFNN	1.35 (2.29)	0.24 (0.47)	98.41 (2.76)	92.82 (3.42)	5.70 (2.34)	1.48 (1.29)	79.79 (12.11)	8.65 (6.57)	11.56 (7.84)
LSTM	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	95.11 (3.35)	3.77 (1.85)	1.12 (1.57)	82.44 (11.51)	7.46 (5.81)	**10.10 (6.89)**

AP : *Accurate Prediction* ; BE : *Benign Error* ; EP : *Erroneous Prediction*

Tableau 4.7: Acceptabilité clinique (CG-EGA) par région moyenne (avec écart type) des prédictions de glycémie par modèle pour l'ensemble de la population du jeu IDIAB et pour les horizons de prédictions 30, 60 et 120 minutes.

5 | Amélioration de l'acceptabilité clinique des prédictions

Sommaire

5.1 Introduction

Dans le Chapitre 4, nous avons montré que les modèles avec la plus grande précision statistique (e.g., RMSE) n'ont pas nécessairement la meilleure acceptabilité clinique (CG-EGA). Cette acceptabilité clinique est très variable d'un modèle à un autre, mais aussi entre les différentes régions glycémiques (hypoglycémie, euglycémie, hyperglycémie). En particulier, les modèles semblent montrer des difficultés avec les prédictions en région d'hypoglycémie,

avec un taux AP particulièrement faible. Ces difficultés sont dues à l'entraînement des modèles n'incluant pas les critères cliniques de la CG-EGA. En effet, l'erreur quadratique moyenne (MSE), généralement utilisée comme fonction objectif à minimiser, ne représente cependant pas la réalité de la prédiction de la glycémie. Par exemple, les erreurs en région d'hypoglycémie et d'hyperglycémie sont plus dangereuses pour la personne diabétique, sans pour autant être plus importantes du point de vue de la MSE. Afin de donner plus d'importance aux prédictions jugées dangereuses pour le patient, Del Favero *et al.* ont proposé une nouvelle métrique appelée gMSE, qui augmente le poids des échantillons en région d'hypoglycémie et d'hyperglycémie lors du calcul de la MSE [40]. Cela permet de focaliser davantage l'apprentissage des modèles sur les zones critiques de la P-EGA. Cependant, bien que permettant d'obtenir une meilleure précision dans ces régions dangereuses, la gMSE ne tient pas compte de la R-EGA, seconde composante de la CG-EGA. Une prédiction peut être cliniquement dangereuse non seulement par sa faible précision, mais aussi par la faible précision des variations prédites. En effet, si le signal de glycémie prédit possède des fluctuations trop importantes, non représentatives des variations de glycémie observées, le patient peut être amené à prendre de mauvaises décisions et mettre sa vie en danger. Par exemple, le patient prend de l'insuline en pensant que sa glycémie est en train d'augmenter très vite, sans que ce soit le cas.

Plus généralement, leurs travaux ont montré que l'optimisation de l'acceptabilité clinique et de la précision des modèles sont deux objectifs différents et qu'ils ne peuvent pas se substituer l'un à l'autre. Ce constat serait d'autant plus vrai avec l'éventuelle prise en compte la R-EGA, qui fait intervenir la précision clinique des variations prédites de glycémie. Il n'existe généralement pas de solution simultanément optimale pour plusieurs objectifs, mais plutôt un ensemble de solutions dites *Pareto-optimales*, chacune représentant un compromis entre les différents objectifs [107]. La résolution d'un problème d'optimisation multi objectif se fait en deux étapes : identification de l'ensemble des solutions Pareto-optimales, puis sélection de l'une de ces solutions via un critère de sélection. Dans le contexte de la prédiction de glycémie pour les personnes diabétiques, la mise en pratique de ces deux étapes s'avère complexe. Tout d'abord, la recherche des solutions Pareto-optimale suppose l'entraînement d'un modèle pour chaque nouvelle solution testée. Cet entraînement est particulièrement coûteux dans le cadre de l'apprentissage profond. Par ailleurs, il n'existe aujourd'hui pas de critères cliniques concernant les modèles de prédiction de glycémie pour personnes diabétiques. Dans le futur, nous pouvons imaginer que ces critères cliniques pourraient prendre la forme de seuils minimaux en AP ou de seuils maximaux en EP basés sur la CG-EGA.

L'objectif de cette étude est de proposer une méthodologie efficace permettant d'inclure des contraintes d'acceptabilité clinique à l'entraînement des modèles tout en maximisant la précision des prédictions. Pour cela, nous analysons dans un premier temps les erreurs cliniques de prédiction faites par le modèle LSTM dans l'étude benchmark présentée au Chapitre 4. En se basant sur ces analyses, nous proposons une nouvelle fonction de coût du nom d'erreur quadratique moyenne cohérente (cMSE) qui pénalise le modèle pendant son entraînement non seulement sur les erreurs de prédictions, mais aussi sur les erreurs de variations prédites. Cela permet d'améliorer la stabilité du signal prédit, et ainsi d'améliorer l'acceptabilité clinique des prédictions à travers l'amélioration de la

R-EGA. Inspiré des travaux de Del Favero *et al.* et leur fonction de coût gMSE [40], nous personnalisons la cMSE au problème de la prédiction de la glycémie en incluant une pondération par zone suivant la P-EGA et la R-EGA. Cette nouvelle fonction de coût, du nom de gcMSE, permet d'opérer de meilleurs compromis entre précision et acceptabilité clinique des prédictions. Enfin, afin d'identifier et sélectionner la solution optimale maximisant la précision des prédictions tout en respectant les critères cliniques (hypothétiques), nous proposons l'algorithme d'Amélioration Progressive de l'Acceptabilité Clinique (APAC). Cet algorithme permet de relâcher petit à petit les contraintes en précision des prédictions pendant l'apprentissage, afin de mettre progressivement l'accent sur l'amélioration de l'acceptabilité clinique grâce à la fonction de coût gcMSE.

5.2 Analyse de l'acceptabilité clinique des prédictions

Afin de réduire le nombre d'erreurs de prédiction cliniquement mauvaises, et ainsi améliorer l'acceptabilité clinique des modèles, nous proposons tout d'abord de procéder à une analyse de ces erreurs. L'objectif est ainsi de comprendre plus en détail la nature des erreurs cliniques de prédiction, et de dégager des pistes de traitement visant à réduire leurs apparitions.

Procéder à l'analyse des erreurs directement sur les résultats présentés dans le Chapitre 4 ne serait pas rigoureux, car il s'agit des résultats obtenus sur les ensembles de test. Pour cela, nous utilisons les ensembles de validation, dont le Tableau 5.1 présente les performances en acceptabilité clinique (CG-EGA). Les résultats sont présentés pour le modèle LSTM, qui est le modèle d'apprentissage profond le plus performant du benchmark, pour un horizon de prédiction de 30 minutes, et pour chaque patient des jeux de données IDIAB et OhioT1DM.

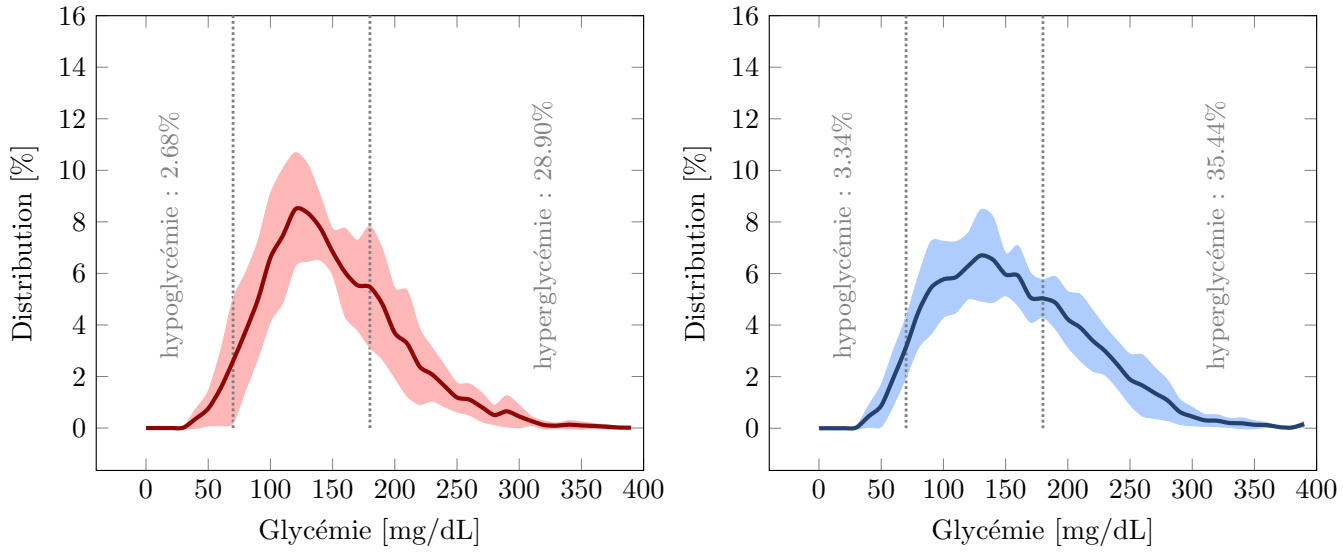
Comme nous avons pu le noter dans l'étude benchmark, du point de vue de l'acceptabilité clinique, les régions glycémiques les plus compliquées (taux supérieur d'erreurs graves de prédiction, EP, et d'erreurs bénignes de prédiction, BE) sont celles de l'hypoglycémie et de l'hyperglycémie. Cela est dû aux dynamiques complexes et rapides mises en jeu, ainsi que du manque de données dans ces régions (surtout dans le cas de l'hypoglycémie, voir Figure 5.1). Afin de pouvoir améliorer l'acceptabilité clinique des prédictions, revenant à réduire le taux de prédictions erronées EP et d'erreurs bénignes BE, nous devons nous intéresser au mécanisme mis en œuvre catégorisant les prédictions comme telles.

Le Tableau 5.2 nous donne la grille de classification des prédictions en AP, BE, ou EP. Cette acceptabilité clinique des prédictions est déterminée en fonction des notes obtenues selon les grilles P-EGA (notes de A à E) et R-EGA (notes de A à E, avec une séparation des zones C, D et E en deux). Tandis que la P-EGA évalue la précision clinique des prédictions, la R-EGA s'intéresse à la cohérence des prédictions successives avec les variations de glycémie observées. La Figure 5.2 représente la visualisation de l'attribution des notes aux prédictions par la P-EGA et la R-EGA pour les deux jeux IDIAB et OhioT1DM. Pour qu'une prédiction soit jugée cliniquement acceptable (AP), une prédiction doit obtenir, pour les deux grilles P-EGA et R-EGA, une note de A ou de B. À

Patient ID	CG-EGA								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
Jeu de données IDIAB									
1	82.43 (9.80)	0.00 (0.00)	17.57 (9.80)	96.74 (0.45)	2.74 (0.26)	0.53 (0.23)	91.68 (4.13)	7.27 (2.39)	1.05 (2.11)
2	5.00 (8.66)	0.00 (0.00)	95.00 (8.66)	94.12 (2.16)	4.84 (1.83)	1.03 (0.92)	93.48 (1.27)	5.13 (1.09)	1.39 (0.50)
3	57.88 (11.11)	0.00 (0.00)	42.12 (11.11)	94.87 (1.51)	3.98 (1.38)	1.14 (0.45)	88.35 (3.33)	7.80 (1.26)	3.85 (2.46)
4	49.21 (42.82)	0.00 (0.00)	50.79 (42.82)	92.52 (1.77)	5.79 (1.68)	1.69 (0.39)	80.48 (5.05)	15.24 (4.38)	4.28 (0.99)
5	20.00 (21.21)	0.00 (0.00)	80.00 (21.21)	93.82 (2.45)	4.65 (1.77)	1.53 (0.72)	92.04 (1.84)	6.34 (1.40)	1.63 (0.89)
6	24.55 (35.68)	0.00 (0.00)	75.45 (35.68)	94.26 (3.62)	4.47 (3.12)	1.28 (0.54)	90.33 (3.32)	7.50 (2.80)	2.17 (0.86)
moyenne	39.84 (26.04)	0.00 (0.00)	60.16 (26.04)	94.39 (1.27)	4.41 (0.93)	1.20 (0.37)	89.39 (4.29)	8.21 (3.27)	2.39 (1.23)
Jeu de données OhioT1DM									
559	39.58 (13.16)	2.76 (2.54)	57.66 (12.78)	78.56 (1.45)	16.49 (1.63)	4.95 (0.73)	69.85 (2.23)	21.13 (1.68)	9.02 (0.99)
563	44.61 (10.47)	3.88 (2.38)	51.51 (9.35)	85.44 (4.13)	12.26 (3.19)	2.29 (1.04)	75.72 (7.98)	19.68 (6.30)	4.61 (1.75)
570	64.31 (9.49)	2.50 (3.37)	33.19 (9.86)	84.40 (1.84)	12.10 (1.39)	3.50 (1.14)	81.78 (4.19)	14.78 (3.73)	3.44 (0.47)
575	69.68 (11.20)	4.81 (3.75)	25.51 (8.70)	82.12 (1.23)	13.80 (0.69)	4.08 (0.73)	72.82 (4.50)	19.91 (2.84)	7.27 (2.18)
588	10.88 (13.30)	1.60 (2.24)	87.52 (13.24)	79.48 (2.94)	15.59 (1.74)	4.93 (1.27)	76.08 (2.91)	18.66 (2.06)	5.26 (1.48)
591	31.54 (13.26)	5.65 (2.68)	62.80 (15.23)	73.31 (4.38)	19.71 (2.29)	6.98 (2.39)	70.88 (3.96)	21.11 (2.41)	8.01 (2.49)
moyenne	43.43 (19.76)	3.53 (1.39)	53.03 (20.25)	80.55 (4.06)	14.99 (2.65)	4.46 (1.45)	74.52 (3.97)	19.21 (2.16)	6.27 (1.98)

AP : *Accurate Prediction* ; BE : *Benign Error* ; EP : *Erroneous Prediction*

Tableau 5.1: Acceptabilité clinique (CG-EGA) par région moyenne (écart type) du modèle LSTM pour un horizon de prédiction de 30 minutes sur les ensembles de validation des jeux de données IDIAB et OhioT1DM.



(a) Jeu de données IDIAB

(b) Jeu de données OhioT1DM

FIGURE 5.1: Distribution de la glycémie des échantillons des ensembles d'entraînements pour les jeux de données IDIAB et OhioT1DM.

		P-EGA										
		Hypoglycémie			*Euglycémie*			*Hyperglycémie*				
		A	D	E	A	B	C	A	B	C	D	E
R-EGA	A	AP	EP	EP	AP	AP	EP	AP	AP	EP	EP	EP
	B	AP	EP	EP	AP	AP	EP	AP	AP	EP	EP	EP
	uC	BE	EP	EP	BE	BE	EP	BE	BE	EP	EP	EP
	lC	BE	EP	EP	BE	BE	EP	BE	BE	EP	EP	EP
	uD	EP	EP	EP	BE	BE	EP	BE	BE	EP	EP	EP
	lD	BE	EP	EP	BE	BE	EP	EP	EP	EP	EP	EP
	uE	EP	EP	EP	EP	EP	EP	EP	EP	EP	EP	EP
	lE	BE	EP	EP	EP	EP	EP	EP	EP	EP	EP	EP

AP : *Accurate Prediction* ; BE : *Benign Error* ; EP : *Erroneous Prediction*

Tableau 5.2: Classification des prédictions de glycémie opérée par la CG-EGA. En fonction des scores obtenus sur la P-EGA et la R-EGA, une prédiction est classifiée comme étant une prédiction cliniquement précise (AP), une erreur benigne (BE) ou une erreur dangereuse (EP).

l'inverse, une prédiction est jugée être une erreur grave (EP) lorsque sa précision est faible (zone C à E dans la P-EGA), ou que les variations prédites ne reflètent pas les variations fortes observées. Une prédiction est jugée erreur bénigne (BE) lorsque sa précision clinique est bonne (A ou B), mais pas son taux de variation. En fonction des régions glycémiques (hypoglycémie, euglycémie ou hyperglycémie), la gravité de cette erreur de variation est différente. Par exemple, en hypoglycémie, les zones lD et lE sont jugées moins dangereuses que les zones uD et uE car les variations prédites sont négatives, prévoyant ainsi que le patient va rester en hypoglycémie. Du point de vue comportemental pour le patient, cela aura pour conséquence de le pousser à ingérer des glucides afin de faire remonter sa glycémie rapidement. Cette action, bien que potentiellement problématique, car pouvant entraîner une hyperglycémie, n'est pas dangereuse pour le patient dans l'immédiat. En revanche, l'absence de détection de variations négatives des zones uD et uE est extrêmement dangereuse : l'hypoglycémie est en train de s'aggraver fortement, pouvant résulter en des conséquences telles que le coma ou même la mort.

Les Tableaux 5.3 et 5.4 donnent, en détail, la distribution des notes obtenues par les patients des deux jeux sur la P-EGA et la R-EGA respectivement. Avec 99.09% et 98.34% des prédictions obtenant une note de A ou B (A+B) pour les jeux IDIAB et OhioT1DM, les résultats de la P-EGA nous montrent que les prédictions ont globalement une très bonne précision clinique. En revanche, les variations prédites sont nettement moins bonnes, comme nous le montre les taux de 93.23% et 78.99% pour les notes A ou B (A+B) pour les jeux IDIAB et OhioT1DM sur la R-EGA. Cela nous permet de supposer la grande part de responsabilité de la R-EGA dans l'attribution du label EP aux prédictions, et ainsi au manque de cohérence entre les prédictions successives.

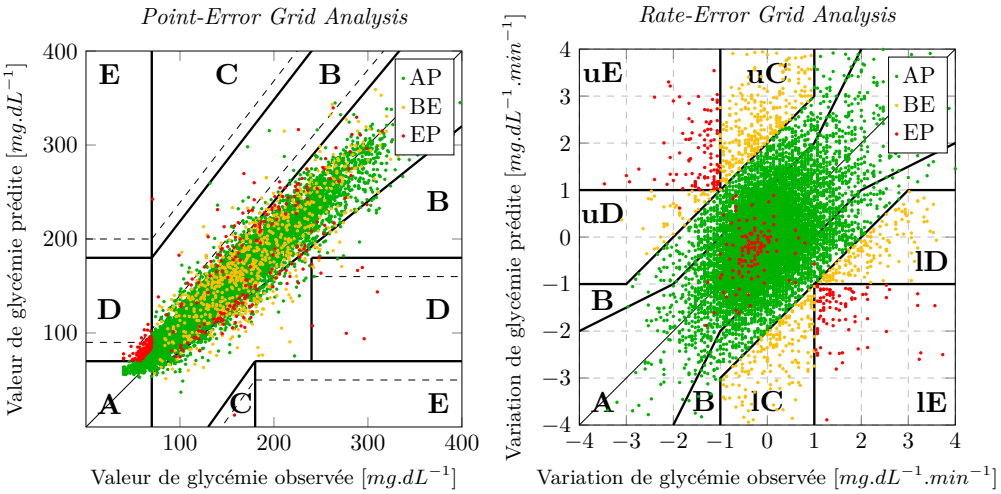
Le Tableau 5.5, nous permet de quantifier ce sentiment en représentant la distribution de la cause d'attribution du label EP pour les différentes régions de glycémie. Tandis qu'une prédiction jugée EP a pour cause la P-EGA lorsque celle-ci est mauvaise (C, D ou E) et que la R-EGA est bonne (A ou B), à l'inverse, elle a pour cause la R-EGA lorsque celle-ci est mauvaise (uD, lD, uE, lE) et que la P-EGA est bonne (A ou B). Pour les prédictions possédant

Patient ID	P-EGA					
	A+B	A	B	C	D	E
Jeu de données IDIAB						
1	98.99 (0.67)	95.43 (1.30)	3.56 (0.95)	0.04 (0.09)	0.96 (0.67)	0.00 (0.00)
2	99.12 (0.74)	96.05 (1.48)	3.07 (1.04)	0.00 (0.00)	0.88 (0.74)	0.00 (0.00)
3	98.31 (0.99)	94.52 (2.64)	3.79 (1.88)	0.00 (0.00)	1.69 (0.99)	0.00 (0.00)
4	99.34 (0.53)	93.09 (2.55)	6.25 (2.26)	0.04 (0.09)	0.62 (0.46)	0.00 (0.00)
5	99.34 (0.44)	96.25 (1.59)	3.09 (1.23)	0.00 (0.00)	0.66 (0.44)	0.00 (0.00)
6	99.42 (0.44)	95.09 (1.24)	4.33 (1.26)	0.00 (0.00)	0.58 (0.44)	0.00 (0.00)
moyenne	99.09 (0.38)	95.07 (1.06)	4.01 (1.09)	0.01 (0.02)	0.90 (0.38)	0.00 (0.00)
Jeu de données OhioT1DM						
559	97.84 (1.12)	89.57 (2.08)	8.27 (1.10)	0.03 (0.05)	2.13 (1.12)	0.00 (0.00)
563	98.72 (0.41)	94.30 (1.26)	4.42 (1.04)	0.01 (0.02)	1.25 (0.38)	0.02 (0.05)
570	99.43 (0.26)	95.35 (0.64)	4.08 (0.66)	0.00 (0.00)	0.56 (0.27)	0.01 (0.02)
575	97.74 (1.20)	90.74 (2.29)	7.00 (1.51)	0.05 (0.11)	2.17 (1.22)	0.03 (0.07)
588	98.99 (0.49)	91.42 (2.48)	7.57 (2.24)	0.02 (0.03)	0.97 (0.47)	0.02 (0.02)
591	97.30 (1.53)	86.76 (2.84)	10.55 (2.08)	0.05 (0.07)	2.63 (1.57)	0.02 (0.03)
moyenne	98.34 (0.76)	91.36 (2.87)	6.98 (2.23)	0.03 (0.02)	1.62 (0.74)	0.02 (0.01)

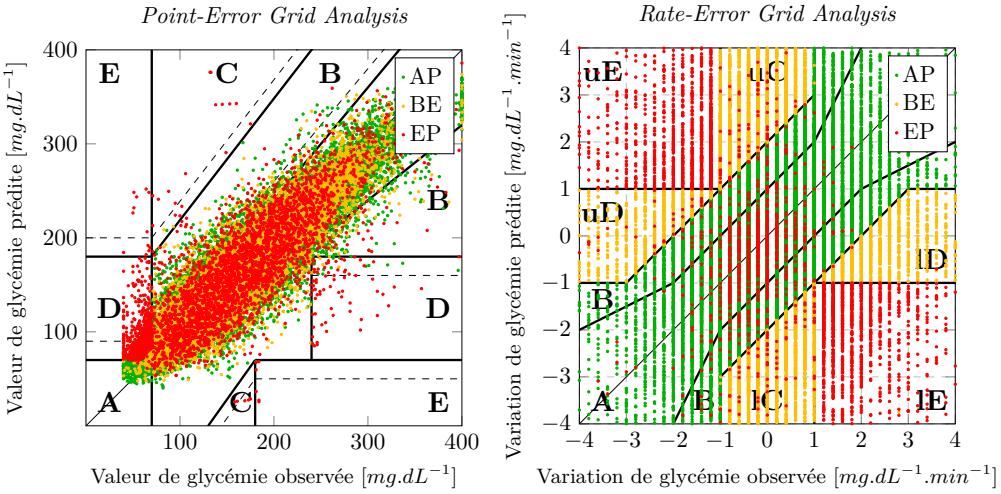
Tableau 5.3: Distribution des erreurs cliniques de prédiction de la P-EGA du modèle LSTM sur l'ensemble de validation des jeux de données IDIAB et OhioT1DM, pour un horizon de prédiction de 30 minutes.

Patient ID	R-EGA								
	A+B	A	B	lC	uC	lD	uD	lE	uE
Jeu de données IDIAB									
1	96.79 (0.59)	81.66 (3.91)	15.13 (3.37)	0.76 (0.32)	1.51 (0.22)	0.42 (0.29)	0.04 (0.09)	0.28 (0.08)	1.51 (0.22)
2	93.97 (1.28)	76.60 (2.57)	17.37 (1.60)	1.49 (0.54)	2.02 (0.49)	0.91 (0.44)	0.39 (0.16)	0.70 (0.37)	2.02 (0.49)
3	93.76 (1.82)	74.55 (6.27)	19.21 (4.60)	1.29 (0.45)	2.27 (0.43)	0.92 (0.49)	0.09 (0.11)	0.71 (0.27)	2.27 (0.43)
4	88.49 (1.98)	72.66 (3.33)	15.83 (2.24)	1.98 (0.63)	3.49 (0.58)	2.72 (0.89)	0.80 (0.41)	0.93 (0.23)	3.49 (0.58)
5	92.96 (1.94)	71.66 (5.87)	21.30 (4.85)	1.56 (0.66)	2.59 (0.54)	0.91 (0.42)	0.46 (0.34)	0.82 (0.31)	2.59 (0.54)
6	93.40 (3.48)	73.14 (6.19)	20.26 (3.15)	1.38 (1.14)	2.71 (1.13)	0.69 (0.38)	0.43 (0.49)	0.69 (0.44)	2.71 (1.13)
moyenne	93.23 (2.45)	75.04 (3.35)	18.18 (2.26)	1.41 (0.36)	2.43 (0.61)	1.09 (0.75)	0.37 (0.25)	0.69 (0.20)	2.43 (0.61)
Jeu de données OhioT1DM									
559	75.56 (1.46)	51.03 (2.61)	24.53 (1.43)	6.67 (0.83)	7.33 (0.84)	2.33 (0.39)	1.76 (0.07)	3.32 (0.51)	7.33 (0.84)
563	83.46 (4.29)	59.23 (5.69)	24.23 (1.73)	5.36 (1.43)	6.19 (1.60)	1.30 (0.17)	0.90 (0.29)	1.41 (0.48)	6.19 (1.60)
570	83.16 (2.23)	59.05 (2.58)	24.11 (2.59)	5.04 (1.21)	5.68 (0.98)	1.54 (0.17)	1.17 (0.08)	1.96 (0.30)	5.68 (0.98)
575	80.86 (1.47)	55.71 (1.79)	25.15 (1.02)	5.51 (0.39)	6.35 (0.38)	1.59 (0.22)	1.27 (0.28)	2.16 (0.49)	6.35 (0.38)
588	78.21 (2.67)	54.10 (2.68)	24.11 (0.42)	5.77 (0.47)	6.34 (0.76)	2.80 (0.39)	1.88 (0.55)	2.53 (0.56)	6.34 (0.76)
591	72.72 (3.97)	46.92 (3.97)	25.80 (0.58)	8.09 (0.79)	8.49 (0.94)	2.02 (0.24)	1.61 (0.43)	3.48 (0.94)	8.49 (0.94)
moyenne	78.99 (3.93)	54.34 (4.36)	24.65 (0.63)	6.07 (1.03)	6.73 (0.92)	1.93 (0.51)	1.43 (0.35)	2.48 (0.73)	6.73 (0.92)

Tableau 5.4: Distribution des erreurs cliniques de variations prédites de la R-EGA du modèle LSTM sur l'ensemble de validation des jeux de données IDIAB et OhioT1DM, pour un horizon de prédiction de 30 minutes.



(a) P-EGA et R-EGA du jeu IDIAB



(b) P-EGA et R-EGA du jeu OhioT1DM

FIGURE 5.2: P-EGA (gauche) et R-EGA (droite) du modèle LSTM sur les sous-ensembles de validation des jeux de données IDIAB (haut) et OhioT1DM (bas).

un mauvais score P-EGA et R-EGA, la responsabilité du label EP est attribuée aux deux grilles simultanément. Les résultats nous montrent que la P-EGA est responsable de la plupart des EP en région d'hypoglycémie, et que la R-EGA l'est pour les régions d'euglycémie et d'hyperglycémie. Ainsi, la région d'hypoglycémie étant globalement très peu présente dans les deux jeux, les erreurs de R-EGA sont la principale cause de prédictions EP.

En conclusion, dans l'optique d'améliorer l'acceptabilité clinique des prédictions, il est nécessaire d'améliorer la cohérence des prédictions successives afin qu'elles représentent mieux le véritable sens de variation de la glycémie. Cependant, nous devons tout de même veiller à réduire les prédictions peu précises, notamment en région d'hypoglycémie. Bien qu'elles ne représentent que peu de prédictions dans l'ensemble, celles-ci restent néanmoins particulièrement dangereuses pour les personnes diabétiques.

Patient ID	CG-EGA								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	P	R	P/R	P	R	P/R	P	R	P/R
Jeu de données IDIAB									
1	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	0.07 (0.13)	**0.93** (0.13)	0.00 (0.00)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)
2	**0.92** (0.14)	0.00 (0.00)	0.08 (0.14)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)
3	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)	0.12 (0.22)	**0.88** (0.22)	0.00 (0.00)
4	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**0.97** (0.06)	0.03 (0.06)	0.00 (0.00)	**0.94** (0.08)	0.06 (0.08)
5	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)	0.03 (0.07)	**0.97** (0.07)	0.00 (0.00)
6	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)	0.00 (0.00)	**0.80** (0.24)	0.20 (0.24)
moyenne	**0.99** (0.03)	0.00 (0.00)	0.01 (0.03)	0.01 (0.03)	**0.98** (0.03)	0.00 (0.01)	0.03 (0.05)	**0.93** (0.08)	0.04 (0.08)
Jeu de données OhioT1DM									
559	**0.88** (0.10)	0.02 (0.02)	0.10 (0.09)	0.00 (0.01)	**0.99** (0.02)	0.01 (0.01)	0.02 (0.01)	**0.95** (0.04)	0.03 (0.03)
563	**0.86** (0.12)	0.01 (0.01)	0.13 (0.12)	0.01 (0.01)	**0.99** (0.01)	0.00 (0.00)	0.01 (0.02)	**0.98** (0.04)	0.01 (0.01)
570	**0.76** (0.14)	0.03 (0.06)	0.21 (0.16)	0.00 (0.00)	**1.00** (0.00)	0.00 (0.00)	0.02 (0.02)	**0.98** (0.02)	0.01 (0.01)
575	**0.77** (0.05)	0.04 (0.04)	0.19 (0.05)	0.02 (0.04)	**0.98** (0.04)	0.00 (0.00)	0.02 (0.03)	**0.97** (0.04)	0.01 (0.02)
588	**0.78** (0.17)	0.01 (0.01)	0.21 (0.16)	0.00 (0.01)	**0.99** (0.01)	0.00 (0.00)	0.02 (0.02)	**0.98** (0.02)	0.00 (0.00)
591	**0.71** (0.13)	0.01 (0.01)	0.28 (0.12)	0.01 (0.02)	**0.99** (0.02)	0.00 (0.00)	0.03 (0.03)	**0.95** (0.05)	0.03 (0.03)
moyenne	**0.79** (0.06)	0.02 (0.01)	0.19 (0.06)	0.01 (0.01)	**0.99** (0.01)	0.00 (0.00)	0.02 (0.01)	**0.97** (0.02)	0.01 (0.01)

Tableau 5.5: Attribution de la responsabilité des erreurs cliniques graves de prédiction (EP) aux grilles de la CG-EGA faites par le modèle LSTM sur les ensembles de validation et un horizon de prédiction de 30 minutes pour les jeux de données IDIAB et OhioT1DM. La responsabilité de l'erreur clinique est attribuée soit à la P-EGA (P), soit à la R-EGA (R), ou soit aux deux grilles simultanément (P/R).

5.3 Intégration de critère d'acceptabilité clinique dans l'apprentissage des réseaux de neurones

Dans cette section nous proposons une méthode pour intégrer les critères cliniques de la CG-EGA au sein de l'apprentissage des modèles profonds. Dans un premier temps nous présentons une nouvelle fonction de coût, l'erreur quadratique moyenne cohérente, permettant de pénaliser non seulement les erreurs de prédictions, mais aussi les erreurs de variations prédites. Puis, nous proposons une personnalisation de cette nouvelle fonction de coût à la prédiction de la glycémie en pondérant différemment les erreurs en fonction des notes obtenues par la CG-EGA. Enfin, nous proposons une méthodologie incluant des critères cliniques experts son utilisation.

5.3.1 Erreur quadratique moyenne cohérente

Dans l'apprentissage profond, les prédictions faites par les modèles sont issues du processus itératif d'apprentissage par rétropropagation du gradient d'erreur. Afin d'améliorer l'acceptabilité clinique des prédictions, deux approches différentes peuvent être envisagées. La première consiste à modifier a posteriori les prédictions, juste après leur calcul par le modèle. L'analyse des erreurs cliniques de prédiction a montré que le signal de glycémie prédit est caractérisé par des variations de glycémie ne reflétant pas les variations observées. Ces fluctuations

peuvent être vues comme du bruit pouvant être atténué en lissant le signal de glycémie. Ce lissage peut se faire par l'utilisation de filtres numériques passe-bas qui atténuent les composantes hautes-fréquences du signal. Cependant, dû aux contraintes temps réel d'un système de prédiction de glycémie ne permettant pas l'utilisation de valeur du futur (filtre causal), le processus de filtrage introduit nécessairement un déphasage temporel du signal [83]. Dans le contexte de la prédiction de la glycémie de personnes diabétiques, ce déphasage temporel réduit considérablement l'intérêt du système prédictif. Par ailleurs, il existe des techniques dérivées des filtres passe-bas permettant de lisser les signaux temporels, comme le lissage par moyenne mobile ou de lissage exponentiel [157]. Toutefois, en utilisant de telles méthodes, il est nécessaire de ne pas opérer un lissage excessif qui aurait pour conséquence de déphaser grandement le signal de glycémie prédit, le rendant ainsi trop imprécis.

La seconde approche que nous pouvons envisager est de s'intéresser directement au processus d'apprentissage du modèle. En apprentissage automatique, l'entraînement d'un modèle prédictif implique l'optimisation d'un objectif représenté par la minimisation (ou maximisation) d'une fonction objectif (aussi appelée *fonction de coût*). Dans l'apprentissage profond, l'apprentissage du modèle se fait par l'application du gradient de la fonction de coût aux connexions du réseau de neurones. Ainsi, en modifiant la fonction objectif, il est possible de modifier le comportement prédictif du modèle. De fait, nous retrouvons un grand nombre de fonctions de coût dans la littérature. Les plus connues sont l'entropie croisée (*cross-entropy*) utilisée pour des problèmes de classification et l'erreur quadratique moyenne (MSE) pour les problèmes de régression. L'Équation 5.1 décrit la MSE comme la différence quadratique entre la glycémie observée $\boldsymbol{y}$ et prédite $\hat{\boldsymbol{y}}$, moyennée sur N échantillons. La prédiction de la glycémie étant une tâche de régression, les modèles profonds du milieu utilisent la MSE dans l'apprentissage des modèles. Dans la présente étude, nous proposons d'apporter des modifications à la fonction de coût MSE dans l'optique d'améliorer l'acceptabilité clinique des prédictions.

$$MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 \tag{5.1}$$

Tout d'abord, comme nous avons pu le voir en analysant les erreurs cliniques, il est primordial de pénaliser les erreurs de variation de glycémie prédites en plus des erreurs de prédiction. Pour cela, nous proposons une nouvelle fonction de coût, l'erreur quadratique moyenne cohérente (cMSE) décrite comme la MSE des prédictions pondérée par la MSE des variations prédites de glycémie. L'Équation 5.2 décrit la cMSE, avec $\Delta\boldsymbol{y}$ et $\Delta\hat{\boldsymbol{y}}$ représentant, respectivement, les variations observées et prédites de glycémie. Le facteur de cohérence c représente l'importance relative que nous donnons à la précision des variations prédites vis-à-vis de la précision des prédictions.

$$\begin{aligned} cMSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) + c \cdot MSE(\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}) \\ &= \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 + c \cdot (\Delta y_n - \Delta \hat{y}_n)^2 \end{aligned} \tag{5.2}$$

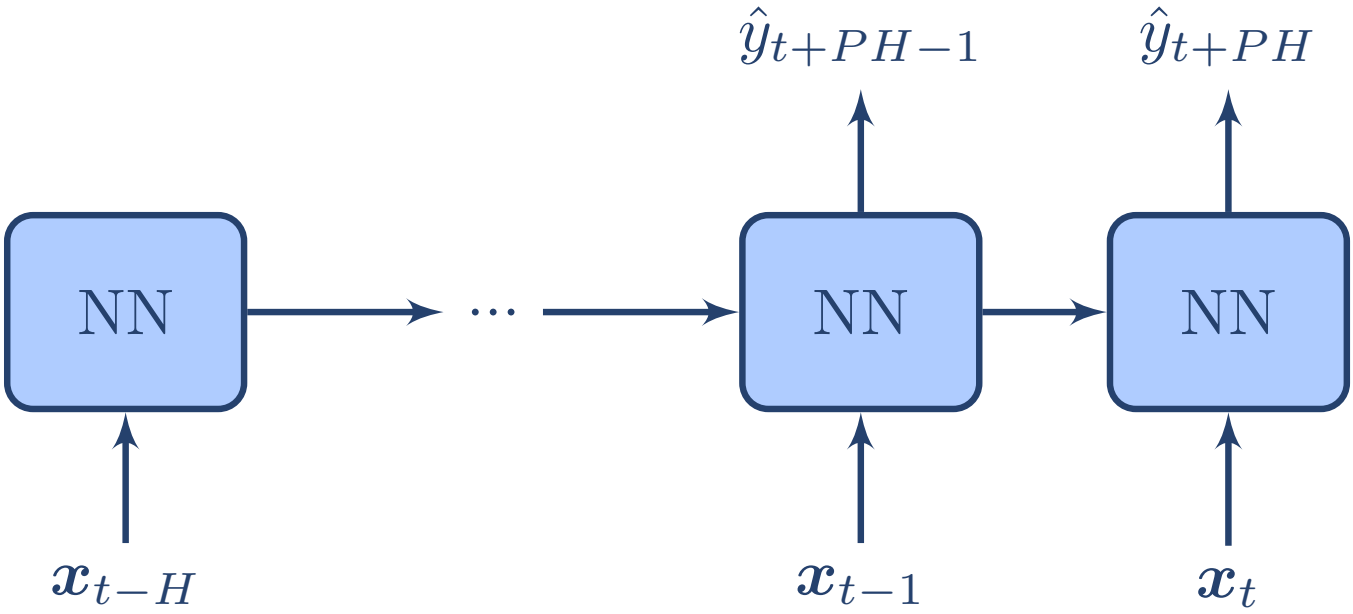


FIGURE 5.3: Architecture générale d'un réseau de neurones récurrents à deux sorties qui a été déroulé H fois, H étant la longueur de l'historique des données d'entrée au modèle. $\boldsymbol{x}_t$ sont les données d'entrées au modèle à l'instant t (e.g., glucose, insuline et glucides à l'instant t), et $\hat{y}_{t+PH}$ est la prédiction du modèle (e.g., prédiction de glycémie) à $t+PH$, PH étant l'horizon de prédiction.

Pour pouvoir utiliser la cMSE, nous proposons l'utilisation d'un réseau de neurones récurrents (e.g., LSTM) à deux sorties (voir Figure 5.3). Les deux sorties représentent les prédictions de glycémies à horizon PH (e.g., 30 minutes) et PH-1 (e.g., 25 minutes avec une prédiction toutes les 5 minutes). Elles permettent de calculer les variations de glycémie prédites (voir Équation 5.3, où ΔT est l'intervalle de temps entre deux prédictions). L'architecture des réseaux de neurones récurrents est particulièrement adaptée à cette tâche puisque tous les sous-modules du réseau déplié (voir Figure 5.3) partagent les mêmes poids.

$$\Delta\hat{y}_{t+PH} = \frac{\hat{y}_{t+PH} - \hat{y}_{t+PH-\Delta T}}{\Delta T} \tag{5.3}$$

5.3.2 Personnalisation de la cMSE pour la tâche de la prédiction de la glycémie

Nous avons pu voir précédemment que les erreurs de prédiction de glycémie ou de variations prédites n'ont pas la même importance clinique en fonction des zones auxquelles elles appartiennent (voir Tableau 5.2). Bien que généralement de magnitude plus importante, ces erreurs cliniques restent rares et ne représentent qu'une faible portion du gradient de mise à jour des poids du réseau pendant l'entraînement. Par conséquent, minimiser la MSE (ou, de manière équivalente, la cMSE) ne permet pas directement de réduire le nombre d'erreurs cliniques de prédiction. En effet, la plus grande partie de la mise à jour des poids se fait en faveur de l'amélioration de la précision de prédictions possédant déjà une bonne acceptabilité clinique. Dans le domaine de la classification multi-classes, il est très courant de pondérer les échantillons des classes sous-représentées de telle sorte à augmenter

artificiellement leur représentation au sein du jeu d'entraînement [15]. Dans leurs travaux sur la reconnaissance d'objets au sein d'images, Lin *et al.* ont proposé de pondérer dynamiquement les échantillons d'apprentissages en fonction de leur facilité [98]. Un échantillon est jugé facile lorsque la probabilité de la classe correspondante est très élevée, démontrant un haut degré de confiance de la part du modèle. En réduisant le poids des échantillons jugés faciles, l'entraînement du modèle se focalise sur les échantillons pour lesquels il a le plus de difficultés. Enfin, comme évoqué précédemment, Del Favero *et al.* ont proposé, dans le cadre de la prédiction de glycémie, de modifier le MSE pour mieux tenir compte des régions dangereuses de la P-EGA. En particulier, ils ont proposé d'apporter une pondération supérieure aux échantillons dont la glycémie observée se trouve en hypoglycémie ou en hyperglycémie. Bien que ces travaux aient été évalués sur des modèles autorégressifs de type ARX et des patients virtuels du logiciel T1DMS, leurs résultats ont montré que cette nouvelle fonction de coût permet de réduire le nombre de prédictions en zone D et E de la grille P-EGA. En prenant inspiration dans ces travaux, nous proposons de pénaliser dynamiquement les erreurs de prédiction ainsi que les erreurs de variation prédites. Cette nouvelle fonction de coût, nommée erreur glycémique quadratique moyenne cohérente (gcMSE), pénalise différemment les prédictions en fonction des régions de la P-EGA et de la R-EGA dans lesquelles elles se trouvent (voir Équation 5.4). Dans l'Équation 5.4a, P_X et p_x, $X \in \{A, B, uC, lC, uD, lD, uE, lE\}$ et $x \in \{a, b, uc, lc, ud, ld, ue, le\}$ représentent respectivement les régions de la grille P-EGA et leurs poids respectifs. Contrairement à la P-EGA originelle, nous avons segmenté les régions C, D et E en deux, comme cela est déjà le cas pour la R-EGA. Cela permet d'apporter plus de flexibilité dans l'assignation des pondérations. De manière équivalente, dans l'Équation 5.4a, R_X et r_x, $X \in \{A, B, uC, lC, uD, lD, uE, lE\}$ et $x \in \{a, b, uc, lc, ud, ld, ue, le\}$ représentent les régions et poids de la R-EGA.

$$gcMSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = P(\boldsymbol{y}, \hat{\boldsymbol{y}}) \cdot MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) + c \cdot R(\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}) \cdot MSE(\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}) \tag{5.4}$$

avec,

$$P(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \begin{cases} p_a, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_A \\ p_b, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_B \\ p_{uc}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_{uC} \\ p_{lc}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_{lC} \\ p_{ud}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_{uD} \\ p_{ld}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_{lD} \\ p_{ue}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_{uE} \\ p_{le}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in P_{lE} \end{cases}, \qquad R(\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}) = \begin{cases} r_a, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_A \\ r_b, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_B \\ r_{uc}, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_{uC} \\ r_{lc}, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_{lC} \\ r_{ud}, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_{uD} \\ r_{ld}, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_{lD} \\ r_{ue}, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_{uE} \\ r_{le}, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in R_{lE} \end{cases} \tag{5.4a}$$

L'utilisation de la gcMSE à la place la MSE standard introduit 17 nouveaux hyperparamètres à optimiser : le

facteur de cohérence c, ainsi que les poids associés aux régions de la P-EGA et la R-EGA. Cette tâche s'avérant particulièrement laborieuse, nous proposons quelques simplifications réduisant ainsi le nombre d'hyperparamètres de la gcMSE :

— Tout d'abord, il n'est pas intéressant d'améliorer la précision des variations prédites dans les zones A et B. En effet, toutes les prédictions appartenant à ces zones sont cliniquement suffisamment précises. Ainsi, nous pouvons fixer $r_a = r_b = 0$.

— Du point de vue du taux d'AP et de son éventuelle maximisation, les prédictions BE et EP peuvent être vues comme d'égale importance. Cela nous permet de fixer la plupart des zones C, D et E à une valeur identique. De plus, le facteur de cohérence c permet à lui seul de pondérer l'importance que nous donnons entre la précision des prédictions et la précision de variations prédites. Ainsi, nous pouvons décider de fixer tous ces poids à 1.

— Seules les régions D et E hypoglycémiques de la P-EGA (P_{uD} et P_{uE}) demandent un traitement particulier afin d'augmenter l'importance des échantillons en région d'hypoglycémie. Nous dénotons le poids associé à ces zones par p_{hypo}.

L'Équation 5.5 résume ces simplifications de conception, permettant à la fonction de coût gcMSE de n'avoir plus que 3 hyperparamètres : p_{ab}, p_{hypo}, et c. Le choix de ces hyperparamètre dépend à la fois des objectifs de l'apprentissage et des conditions expérimentales. Le facteur de cohérence c doit être choisi en fonction de magnitude relative de la fonction de coût $MSE(\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}})$ par rapport à celle de $MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}$. Le choix du coefficient p_{hypo} doit se faire en fonction de la taille de nos jeux de données. Lorsque peu d'échantillons hypoglycémiques sont disponibles, il est possible de donner une valeur $p_{hypo} > 1$. Quant à p_{ab}, il représente la contrainte de précision que nous donnons pendant l'entraînement. Plus sa valeur est faible, plus l'entraînement du modèle se focalise sur l'amélioration de son acceptabilité clinique au détriment de sa précision.

$$P(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \begin{cases} p_{ab}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in \{P_A, P_B\} \\ p_{hypo}, & \text{si } \{\boldsymbol{y}, \hat{\boldsymbol{y}}\} \in \{P_{uD}, P_{uE}\} \\ 1, & \text{sinon} \end{cases}, \qquad R(\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}) = \begin{cases} 0, & \text{si } \{\boldsymbol{\Delta y}, \boldsymbol{\Delta \hat{y}}\} \in \{R_A, R_B\} \\ 1, & \text{sinon} \end{cases} \tag{5.5}$$

5.3.3 Amélioration progressive de l'acceptabilité clinique

Afin de pouvoir utiliser la fonction de coût gcMSE, nous devons formuler l'objectif d'apprentissage, et en particulier l'importance relative que l'on porte à l'amélioration de l'acceptabilité clinique. En effet, comme montré dans les travaux de Del Favero *et al.*, toute amélioration de l'acceptabilité clinique s'accompagne d'une détérioration de la précision statistique [40].

Les travaux dans le domaine de l'optimisation multi objectif mettent en évidence le besoin d'utiliser des critères de sélection, pouvant prendre la forme d'une pondération entre les différents objectifs ou de seuils minimaux (ou maximaux) pour les différents objectifs [107]. Bien qu'ils n'existent pas aujourd'hui de critères cliniques standards pour les modèles de la prédiction de la glycémie, nous proposons de nous projeter en supposant leur existence. Ces critères cliniques pourraient prendre la forme de seuils minimaux en AP et/ou maximaux en EP, selon la CG-EGA (e.g., minimum 95% de prédictions obtenant le score d'AP selon la CG-EGA). Notre objectif d'apprentissage consisterait dans ce cas à maximiser la précision des prédictions tout en respectant ces critères cliniques.

Pour parvenir à réaliser cet objectif, nous devons tester un grand nombre d'architectures de modèles (d'hyperparamètres) différents, chaque test impliquant l'entraînement d'un réseau de neurones. Cet entraînement est très coûteux dans le cadre de l'apprentissage profond. Il convient ainsi d'utiliser une méthodologie d'entraînement efficace afin de parvenir à la solution optimale. Les méthodologies généralement utilisées pour répondre aux problèmes d'optimisation multi objectifs reposent sur des méthodes génétiques (comme le NSGA-II [39]). Bien que plus rapides qu'une simple recherche par grille, ces algorithmes font intervenir une part d'aléatoire dans les changements apportés aux différents tests.

Afin de contourner ce problème, nous proposons une méthodologie d'Amélioration Progressive de l'Acceptabilité Clinique (APAC). En partant d'une solution maximisant la précision du modèle sans tenir compte de son acceptabilité clinique, les contraintes sur la précision du modèle sont peu à peu relâchées au profit de celles sur l'acceptabilité clinique du modèle. Cela a pour conséquence de dégrader petit à petit la précision statistique du modèle, dégradation qui s'accompagne d'une amélioration progressive de l'acceptabilité clinique.

Algorithme 1 : Amélioration Progressive de l'Acceptabilité Clinique (APAC)

Données : critère(s) clinique(s) C, modèle M, coefficient de mise-à-jour α, coefficient de lissage β, ensemble d'entraînement $train\_set$, ensemble de validation $valid\_set$
Résultat : Modèle maximisant la précision et respectant le(s) critère(s) C ou -1

$i \leftarrow 0$
$M_0 \leftarrow$ entraîner(MSE,$train\_set$,$valid\_set$)
$\boldsymbol{y}_0, \hat{\boldsymbol{y}}_0 \leftarrow$ prédire(M_0,$valid\_set$)
$\hat{\boldsymbol{y}}_0^* \leftarrow$ lisser($\hat{\boldsymbol{y}}_0, \beta$)
tant que $C(M_i) = Faux$ *et MASE*$(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i^*) < 1$ **faire**
  $i \leftarrow i + 1$
  gcMSE$_i \leftarrow$ gcMSE avec $p_{ab} \leftarrow \alpha^{i-1}$
  $M_i \leftarrow$ affiner(M_0, cMSE$_i$,$train\_set$,$valid\_set$)
  $\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i \leftarrow$ prédire(M_i,$valid\_set$)
  $\hat{\boldsymbol{y}}_i^* \leftarrow$ lisser($\hat{\boldsymbol{y}}_i, \beta$)
si *MASE*$(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i^*) < 1$ **alors**
  retourner M_i
sinon
  retourner -1

L'Algorithme 1 donne une description des étapes de cette méthodologie. Dans celui-ci, la loi de mise à jour des poids p_{ab} représentant les contraintes en précision statistiques est à choisir en fonction des conditions expérimen-

tales. Dans cette étude, nous utilisons la loi définie par l'Équation 5.6 (avec $\alpha \in [0, 1]$ représentant la vitesse de l'assouplissement des contraintes de précision). Quant à elle, la métrique *Mean Absolute Scaled Error* (MASE), proposée par Hyndman *et al.* [76]), sert de critère d'arrêt lorsque les critères cliniques ne sont pas atteignables pour le patient en question. L'algorithme s'arrête lorsque la MASE dépasse 1, signifiant qu'un modèle de prédiction naïf (dont la prédiction est égale à la dernière observation connue, équivalent au modèle Baseline de l'étude benchmark) est plus précis que le modèle présent. La section 5.4.3 donne une description plus détaillée de la MASE. Enfin, nous faisons intervenir un lissage exponentiel des prédictions. Ce lissage permet d'atténuer les fluctuations importantes des prédictions présentes dans les premières étapes de l'algorithme. En étant faible, il permet un gain non négligeable en acceptabilité clinique, en contrepartie d'une perte de précision minime. Pour plus de détails quant au lissage exponentiel, se référer à la section sur les étapes de post-traitement des prédictions en section 5.4.3.

$$p_{ab} = \alpha^{i-1} \tag{5.6}$$

L'algorithme APAC permet de ne pas faire d'itérations inutiles, chaque itération rapprochant le modèle de plus en plus de son objectif. De plus, au lieu d'être entraîné depuis son état initial, le modèle est affiné à partir du premier modèle, entraîné avec la MSE standard. Cet affinage nécessite bien moins d'itération qu'un entraînement complet, et permet ainsi à l'algorithme de s'exécuter plus rapidement. Une autre approche aurait été d'affiner le modèle à partir de l'itération précédente. Cependant, en pratique, nous avons été confronté à un problème d'optimum local, empêchant le modèle de trouver une meilleure solution avec la fonction de coût mise à jour.

5.4 Méthodologie

Cette section a pour objectif de donner tous les détails méthodologiques suivis tout au long des expérimentations dont la section suivante présente les résultats. Le code lié à cette étude, notamment l'analyse des erreurs, l'implémentation des modèles et de l'algorithme, a été mis à disposition en accès libre sur GitHub [27].

5.4.1 Données expérimentales

Dans cette étude, nous utilisons les deux jeux de données réels IDIAB et OhioT1DM. En effet, les données virtuelles du jeu T1DMS étant plus simples, en témoignent les meilleurs résultats statistiques et cliniques, nous avons préféré nous concentrer sur les véritables données terrain.

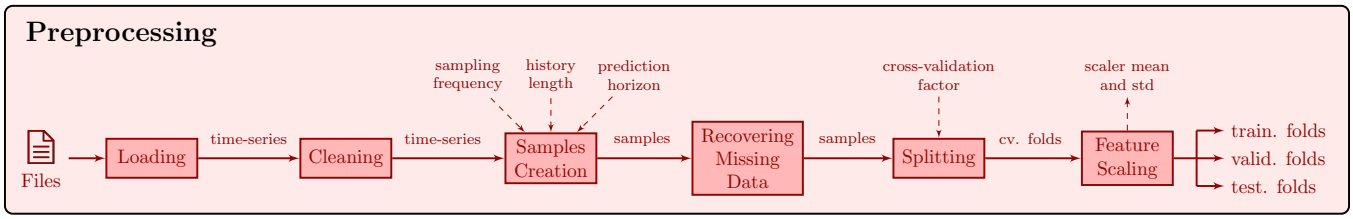


FIGURE 5.4: Étapes de prétraitement des données.

5.4.2 Prétraitement

Dans les grandes lignes, les étapes de prétraitement des données, dont la Figure 5.4 en fait le rappel, sont identiques à celles présentées dans le Chapitre 3.

Nous pouvons néanmoins noter une différence, liée à l'utilisation de la fonction de coût gcMSE présentée précédemment. A l'étape de création des échantillons d'apprentissage (*samples creation* sur la Figure 5.4), pour calculer la fonction de coût gcMSE, nous avons besoin de connaître l'observation de glycémie à l'horizon de prédiction précédent (i.e., y_{t+PH-1} en plus de y_{t+PH} pour pouvoir calculer Δy_{t+PH}). Avec un horizon de prédiction de 30 minutes, et un réseau LSTM déroulé à un intervalle de 5 minutes, nous avons donc besoin de rajouter aux échantillons d'apprentissage la valeur y_{t+25}. Tandis qu'il n'y a aucun souci pratique avec le jeu de données OhioT1DM, cette valeur n'existe généralement pas dans le jeu IDIAB. En effet, la fréquence d'échantillonnage du capteur de glycémie FreeStyle Libre, utilisé dans la collecte du jeu IDIAB, n'est que de 15 minutes. Ainsi, pour obtenir la valeur y_{t+25}, nous avons choisi d'interpoler linéairement le signal de glycémie entre la valeur y_{t+30} et la valeur précédente, y_{t+15}.

5.4.3 Évaluation des modèles prédictifs

L'évaluation des modèles prédictifs suit le même schéma global que l'étude benchmark du Chapitre 4. Nous y rajoutons ici une étape de post-traitement optionnel des prédictions : le lissage des prédictions par lissage exponentiel (*smoothing*). Quant aux métriques d'évaluation, nous avons décidé de substituer la métrique de gain temporel TG par la métrique MASE. La Figure 5.5 donne une représentation graphiques de ces étapes de traitement permettant l'évaluation des modèles prédictifs.

Dans cette étude, nous nous intéressons qu'aux prédictions de glycémie à un horizon de 30 minutes, horizon pour lequel les modèles ont déjà des difficultés.

Lissage exponentiel — *Smoothing*

L'algorithme APAC fait intervenir un lissage des prédictions à chaque itération. L'objectif de ce lissage est de réduire les fluctuations excessives du signal de glycémie prédit. Ces fluctuations ne sont pas représentatives des variations réelles de glycémie, et ainsi dangereuses pour le patient.

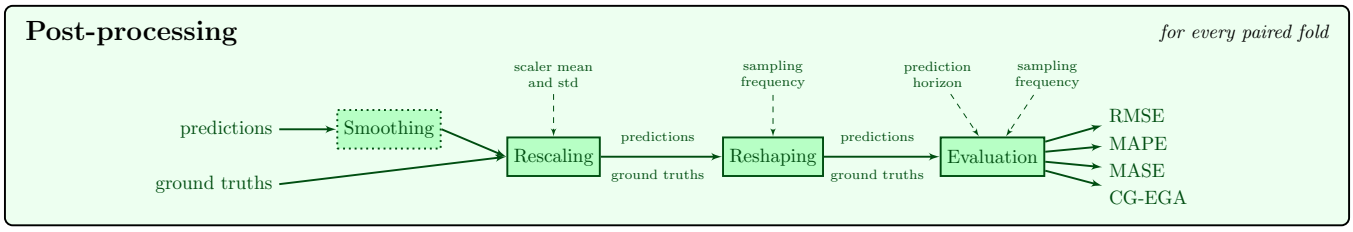


FIGURE 5.5: Post-traitement et évaluation des prédictions de glycémie. En comparaison avec le schéma du Chapitre 4 - GLYFE : une base de résultats de référence, l'étape optionnelle du lissage des données ainsi que la métrique MASE ont été rajoutées.

Nous avons choisi la technique du lissage exponentiel plutôt que celle par moyenne mobile car elle permet de donner plus de poids aux prédictions récentes. Le lissage exponentiel peut être défini comme un lissage récursif, chaque valeur du signal lissé étant égale à une pondération entre la valeur du signal original et la valeur précédente du signal lissé. L'Équation 5.7 donne la définition du lissage exponentielle, avec $\hat{y}_t^*$ représentant la valeur lissée de la prédiction de glycémie $\hat{y}_t$ et β, le coefficient de lissage [13].

$$\hat{y}_t^* = \begin{cases} \hat{y}_0, & \text{si } t = 0 \\ \beta \cdot \hat{y}_t + (1-\beta) \cdot \hat{y}_{t-1}^*, & \text{sinon} \end{cases} \tag{5.7}$$

Plus β est élevé, plus le poids attribué au signal original est fort, et moins le signal est lissé. Le choix du coefficient de lissage $\beta \in [0, 1]$ doit être fait avec soin. En effet, un lissage trop agressif aura pour conséquence de déphaser temporellement le signal. Dans le cadre de la prédiction de la glycémie, cela aura pour conséquence de diminuer grandement la précision du modèle, et donc son intérêt.

A notre connaissance, bien que commun en traitement du signal (e.g., prédiction de consommation en électricité [145]), aucun lissage à posteriori n'a été fait dans la littérature de la prédiction de glycémie. Nous pouvons néanmoins noter l'utilisation ponctuelle de filtre passe-bas (dont l'action est similaire à la technique de lissage exponentiel) sur le signal d'entrée [140, 125].

Métriques d'évaluation

Afin d'évaluer les améliorations faites en acceptabilité clinique, mais aussi des compromis faits sur la précision des modèles, nous devons utiliser à la fois des métriques cliniques ainsi que des métriques statistiques. Nous reprenons les métriques d'évaluation de l'étude benchmark, à l'exception de la métrique TG que nous avons remplacé par la métrique MASE.

La métrique *Mean Absolute Scaled Error* (MASE), calculée comme le ratio de l'erreur absolue moyenne de la glycémie prédite par le modèle et celle d'un modèle naïf [76, 55]. Un modèle naïf est un modèle prédisant une valeur de glycémie égale à la dernière observation connue (voir Équation 5.8). Ainsi, comparer notre modèle à un modèle naïf permet de mesurer l'intérêt du modèle prédictif. Plus la valeur s'approche de 0, plus le modèle a une bonne

capacité de prédiction. En revanche, une valeur supérieure à 1 signifie que le modèle est inutile, un modèle naïf étant plus précis. Grâce à cela, la MASE est utilisée dans l'algorithme APAC comme critère d'arrêt de l'algorithme : lorsque la précision est dégradée au-delà d'une MASE de 1, alors le modèle n'a plus d'intérêt pour la personne diabétique.

$$MASE(\boldsymbol{y}, \hat{\boldsymbol{y}}, PH) = \frac{\frac{1}{N} \cdot \sum_{n=1}^{N} |g_n - \hat{y}_n|}{\frac{1}{N-PH} \cdot \sum_{n=PH}^{N} |g_n - g_{n-PH}|} \tag{5.8}$$

5.4.4 Présentation des modèles prédictifs

L'objectif de l'étude est d'améliorer l'acceptabilité clinique des modèles profonds. Dans ce sens, nous avons tout d'abord proposé la nouvelle fonction de coût cMSE permettant de pénaliser l'apprentissage du modèle non seulement sur les erreurs de prédictions, mais aussi sur les erreurs de variations prédites. Nous avons ensuite proposé la gcMSE qui est la cMSE personnalisée à la prédiction de glycémie. Son objectif est de permettre d'inclure des critères d'acceptabilité clinique dans l'apprentissage des modèles. Enfin, nous avons proposé l'algorithme APAC permettant d'améliorer progressivement l'acceptabilité clinique des modèles via l'utilisation de la fonction gcMSE. Les modèles que nous présentons ici ont pour but de permettre l'évaluation de ces différentes propositions.

Nous utilisons comme modèles de référence les modèles SVR et LSTM tirés de l'étude benchmark du Chapitre 4. Le SVR représente les meilleurs résultats de l'étude benchmark en étant à la fois très précis du point de vue statistique, mais aussi proposant une des meilleures acceptabilités cliniques. De son côté, le modèle LSTM s'est avéré être le meilleur modèle basé sur l'apprentissage profond. Les différences de prétraitement de données évoquées plus tôt n'impactent pas l'évaluation des modèles. Ainsi, nous pouvons réutiliser à l'identique les architectures et résultats des modèles SVR et LSTM de l'étude benchmark.

Dans un premier temps, pour analyser le potentiel d'amélioration de l'acceptabilité clinique des fonctions de coût cMSE et gcMSE, nous pouvons évaluer deux modèles, pcLSTM et gpcLSTM respectivement. Ces deux modèles sont basés sur un réseau LSTM à deux sorties, qui, mis à part la présence des deux sorties, est identique au modèle LSTM de l'étude benchmark. Ils sont respectivement entraînés à minimiser la cMSE et la gcMSE avec un facteur de cohérence c fixé à 8 pour le jeu IDIAB et 2 pour le jeu OhioT1DM. Cette différence entre les deux jeux s'explique par une MSE des variations prédites étant environ 4 fois plus importantes pour le jeu OhioT1DM. Quant aux coefficients p_{ab} et p_{hypo} de la gcMSE, nous les avons fixé à 1 et 10 respectivement. Ces coefficients sont identiques à ceux de la première utilisation de l'algorithme APAC. Par ailleurs, nous proposons d'évaluer une variante supplémentaire de la gcMSE dont le coefficient p_{ab} est fixé à 0. Ce modèle, dénoté gpcLSTM$_{\text{AC}}$, est un modèle qui vise à maximiser l'acceptabilité clinique, sans tenir compte de la précision du modèle au-delà des besoins en acceptabilité clinique.

Pour pouvoir évaluer pleinement l'impact des fonctions de coût et de l'algorithme APAC, nous utilisons la technique du lissage exponentiel sur tous les modèles présentés dans cette étude. La variante lissée de chaque modèle

est représentée par un astérisque en exposant (e.g., LSTM*, pcLSTM*, gpcLSTM$^*_{\text{AC}}$). Tous ces modèles utilisent un coefficient de lissage de 0.85.

L'algorithme APAC permet d'obtenir un compromis entre les résultats obtenus par le modèle gpcLSTM* et ceux obtenus par le modèle gpcLSTM$^*_{\text{AC}}$. L'accent de ce compromis est mis progressivement, au fil des itérations de l'algorithme, sur l'acceptabilité clinique. Cependant, la contrainte en précision, à travers le coefficient p_{ab}, n'est jamais égale à 0 (modèle gpcLSTM$^*_{\text{AC}}$). En effet, un tel modèle posséderait une précision bien trop faible pour être utile pour le patient diabétique. C'est pourquoi l'algorithme APAC s'arrête lorsque la MASE (voir la Section 5.4.3) dépasse la valeur de 1 sur l'ensemble de validation. Nous représentons par le modèle gpcLSTM$^*_{\text{APAC}}$ les résultats obtenus au moment de l'arrêt de l'algorithme APAC. Ces résultats présentent les bornes maximales en acceptabilité clinique conservant une précision utile. Dans l'algorithme APAC, nous utilisons la loi de mise à jour du coefficient p_{ab} présentée par l'Équation 5.6. Elle met en jeu le coefficient α de dégradation de la contrainte en précision, coefficient qui a été fixé à 0.9 dans cette étude. Un coefficient plus élevé permet d'avoir un meilleur contrôle sur le compromis final, en contrepartie d'un temps d'exécution plus lent. L'algorithme APAC utilise le lissage exponentiel sur les prédictions des modèles afin d'accroître la stabilité du signal prédit. Le coefficient de lissage, β a été choisi à 0.85 permettant de ne dégrader que faiblement la précision du signal prédit.

5.5 Résultats expérimentaux

5.5.1 Présentation des résultats

Nous présentons dans cette section les résultats expérimentaux de cette étude. Ces résultats sont représentés sous la forme de deux tableaux : Tableau 5.6 et 5.7. Tandis que le Tableau 5.6 décrit les résultats généraux des différents modèles en termes de RMSE, MAPE, MASE et CG-EGA générale, le Tableau 5.7 donne une description plus détaillée, par région, de la CG-EGA.

Au sein de nos deux modèles de référence, SVR et LSTM, le modèle SVR est le modèle présentant la meilleure acceptabilité clinique (CG-EGA générale ou par région) pour une précision comparable. En particulier, le modèle SVR possède une des meilleures acceptabilités cliniques en région d'hypoglycémie (69.39% et 49.71% d'AP pour les jeux de données IDIAB et OhioT1DM respectivement). Le lissage exponentiel permet d'améliorer l'acceptabilité clinique du modèle SVR (modèle SVR*) de -12.79%[1] de taux d'AP pour une augmentation de +0.90% en RMSE (baisse de précision). Le modèle LSTM* est sujet à des changements similaires avec -11.44% d'AP et +0.98% de RMSE. Le Tableau 5.7 montre que ces améliorations de l'acceptabilité cliniques ont lieu en région d'euglycémie ou d'hyperglycémie, et non en région d'hypoglycémie (faible baisse du taux d'AP).

Le modèle pcLSTM et sa variante lissée pcLSTM*, utilisant la fonction de coût cMSE ainsi que l'architecture à

1. On représente ici la baisse, en %, de ce qui est améliorable du point de vue de la métrique. Pour l'AP, qui a un maximum de 100%, le ratio de changement est calculé comme $(100 - AP_1)/(100 - AP_2)$.

Modèle	RMSE	MAPE	MASE	CG-EGA (générale) AP	BE	EP
Jeu de données IDIAB						
SVR	20.32 (6.02)	**8.66 (0.44)**	0.85 (0.15)	92.69 (2.81)	5.34 (2.06)	1.97 (1.23)
LSTM	**19.85 (6.00)**	9.04 (1.11)	**0.85 (0.10)**	92.20 (2.99)	5.05 (1.71)	2.76 (1.82)
SVR*	20.67 (6.20)	8.86 (0.44)	0.88 (0.15)	93.62 (2.57)	4.47 (1.69)	1.92 (1.35)
LSTM*	20.27 (6.30)	9.25 (1.21)	0.87 (0.09)	93.16 (3.13)	4.16 (1.75)	2.68 (2.00)
pcLSTM	21.89 (5.68)	10.28 (1.34)	0.96 (0.11)	94.04 (3.26)	3.20 (1.66)	2.76 (2.07)
pcLSTM*	22.63 (6.04)	10.64 (1.40)	1.00 (0.11)	94.24 (3.35)	**2.94 (1.73)**	2.82 (2.07)
gpcLSTM	21.21 (5.64)	9.35 (0.92)	0.91 (0.13)	94.03 (2.66)	3.91 (1.48)	2.06 (1.54)
gpcLSTM*	21.86 (5.94)	9.66 (0.95)	0.94 (0.13)	94.53 (2.84)	3.38 (1.55)	2.08 (1.57)
gpcLSTM$_{AC}$	40.68 (11.20)	18.14 (5.55)	1.91 (0.55)	95.34 (2.76)	3.29 (2.56)	**1.37 (0.91)**
gpcLSTM*$_{AC}$	41.15 (11.18)	18.36 (5.47)	1.93 (0.54)	**95.35 (2.87)**	3.20 (2.61)	1.45 (0.92)
gpcLSTM*$_{APAC}$	24.03 (7.15)	10.43 (1.18)	1.03 (0.09)	95.00 (2.74)	3.38 (1.99)	1.61 (1.22)
Jeu de données OhioT1DM						
SVR	**20.15 (2.33)**	**9.12 (2.11)**	**0.85 (0.02)**	83.35 (3.91)	12.38 (2.83)	4.28 (1.83)
LSTM	20.46 (2.08)	9.24 (2.10)	0.86 (0.02)	80.03 (4.17)	14.83 (2.88)	5.14 (2.11)
SVR*	20.17 (2.30)	9.18 (2.12)	**0.85 (0.02)**	85.00 (4.05)	10.97 (2.72)	4.03 (1.90)
LSTM*	20.43 (2.03)	9.26 (2.10)	0.86 (0.02)	82.14 (3.94)	13.06 (2.51)	4.81 (2.04)
pcLSTM	21.53 (2.23)	10.07 (2.32)	0.93 (0.03)	87.45 (3.76)	8.46 (2.05)	4.09 (2.14)
pcLSTM*	21.71 (2.22)	10.19 (2.35)	0.94 (0.03)	87.89 (3.61)	8.15 (1.94)	3.96 (2.12)
gpcLSTM	21.66 (2.69)	9.65 (2.14)	0.92 (0.03)	86.97 (3.63)	9.50 (2.52)	3.53 (1.48)
gpcLSTM*	21.82 (2.69)	9.76 (2.16)	0.93 (0.03)	87.59 (3.45)	9.01 (2.31)	3.41 (1.49)
gpcLSTM$_{AC}$	47.70 (6.31)	22.43 (2.76)	2.37 (0.53)	90.46 (2.85)	7.16 (1.66)	**2.37 (1.28)**
gpcLSTM*$_{AC}$	47.82 (6.27)	22.47 (2.76)	2.37 (0.53)	**90.51 (2.88)**	**7.12 (1.64)**	2.37 (1.30)
gpcLSTM*$_{APAC}$	23.50 (2.49)	10.46 (2.09)	1.01 (0.03)	88.72 (3.59)	8.20 (2.23)	3.08 (1.64)

Tableau 5.6: Précision statistique (RMSE, MAPE et MASE) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour les jeux de données IDIAB et OhioT1DM.

deux sorties du réseau LSTM, montrent des résultats améliorant et détériorant l'acceptabilité clinique et la précision respectivement. En particulier, le modèle pcLSTM* par rapport au modèle LSTM* possède -24.18% d'AP, et +8.95% de RMSE. L'amélioration en acceptabilité clinique est plus importante pour le jeu de données OhioT1DM (-32.19% d'AP) que pour le jeu IDIAB (-16.16% d'AP). Pour une baisse en précision comparable, le jeu OhioT1DM profite donc plus de la fonction de coût cMSE que le jeu IDIAB. Par ailleurs, le modèle pcLSTM* possède parmi les meilleurs scores d'acceptabilité clinique en région d'euglycémie et d'hyperglycémie. Cependant, en comparaison avec les modèles LSTM ou LSTM*, l'acceptabilité clinique en région d'hypoglycémie est détériorée, et ce particulièrement pour le jeu de données OhioT1DM.

Les modèles gpcLSTM et gpcLSTM*, utilisant la fonction de coût gcMSE, cMSE personnalisée à la prédiction de glycémie, affichent une dégradation de la RMSE et une amélioration du taux d'AP similaire aux modèles pcLSTM et pcLSTM*. Cependant, les modèles gpcLSTM et gpcLSTM* possèdent un taux d'EP plus faible (-19.53% et -20.07% respectivement), permettant de conclure à une amélioration de l'acceptabilité clinique. Le Tableau 5.7 montre que cette amélioration se fait principalement en région d'hypoglycémie avec des taux d'EP beaucoup plus faibles.

Les modèles gpcLSTM$_{AC}$ et gpcLSTM*$_{AC}$ utilisent une fonction de gcMSE avec le coefficient p_{ab} de 0. Ainsi, ces

Modèle	CG-EGA (par région)								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
Jeu de données IDIAB									
SVR	69.39 (33.51)	0.35 (0.70)	30.27 (33.54)	95.17 (2.01)	4.33 (1.83)	0.50 (0.47)	89.51 (6.09)	7.43 (3.86)	3.06 (2.53)
LSTM	40.94 (30.73)	**0.00 (0.00)**	59.06 (30.73)	95.78 (1.48)	3.83 (1.55)	0.39 (0.38)	89.55 (5.60)	7.35 (3.21)	3.10 (2.45)
SVR*	66.37 (31.47)	0.17 (0.35)	33.45 (31.51)	96.13 (1.81)	3.49 (1.66)	0.39 (0.36)	90.61 (5.67)	6.60 (3.23)	2.79 (2.79)
LSTM*	37.99 (31.22)	**0.00 (0.00)**	62.01 (31.22)	96.71 (1.35)	2.95 (1.46)	0.33 (0.38)	91.02 (6.04)	6.18 (3.67)	2.80 (2.58)
pcLSTM	34.59 (29.27)	**0.00 (0.00)**	65.41 (29.27)	97.58 (0.90)	2.13 (0.82)	0.29 (0.20)	92.60 (5.81)	4.94 (3.18)	2.46 (2.80)
pcLSTM*	32.20 (27.83)	**0.00 (0.00)**	67.80 (27.83)	**97.96 (0.98)**	**1.81 (0.91)**	**0.23 (0.11)**	92.81 (6.25)	**4.68 (3.48)**	2.51 (2.85)
gpcLSTM	64.79 (24.95)	**0.00 (0.00)**	35.21 (24.95)	96.60 (1.11)	3.03 (0.99)	0.37 (0.26)	92.06 (5.12)	5.42 (2.83)	2.51 (2.46)
gpcLSTM*	61.87 (25.17)	**0.00 (0.00)**	38.13 (25.17)	97.23 (1.17)	2.46 (1.02)	0.31 (0.22)	92.65 (5.60)	4.85 (3.09)	2.50 (2.68)
gpcLSTM$_{AC}$	**87.95 (9.58)**	1.71 (3.43)	**10.34 (8.15)**	97.37 (1.36)	2.12 (1.03)	0.51 (0.40)	92.17 (4.46)	5.11 (4.52)	2.72 (2.39)
gpcLSTM*$_{AC}$	87.77 (9.53)	1.71 (3.43)	10.51 (8.13)	97.50 (1.32)	1.97 (0.97)	0.52 (0.44)	92.10 (4.69)	5.03 (4.70)	2.87 (2.33)
gpcLSTM*$_{APAC}$	68.49 (27.85)	0.57 (1.14)	30.94 (28.22)	97.35 (1.18)	2.32 (1.08)	0.33 (0.15)	**93.16 (4.84)**	5.08 (3.53)	**1.76 (1.49)**
Jeu de données OhioT1DM									
SVR	49.71 (18.75)	5.62 (4.02)	44.67 (18.70)	86.35 (4.24)	10.71 (3.26)	2.94 (1.23)	80.85 (3.24)	14.77 (3.01)	4.37 (1.84)
LSTM	38.37 (23.17)	3.97 (3.72)	57.67 (24.23)	83.78 (5.33)	12.70 (4.06)	3.52 (1.47)	76.86 (3.70)	17.87 (2.73)	5.27 (2.21)
SVR*	46.95 (21.11)	5.97 (4.05)	47.09 (21.65)	87.83 (4.22)	9.46 (3.21)	2.71 (1.22)	82.81 (3.43)	13.12 (2.98)	4.07 (2.00)
LSTM*	37.34 (23.50)	4.11 (4.15)	58.56 (24.17)	85.71 (4.83)	11.10 (3.58)	3.19 (1.37)	79.27 (3.55)	15.85 (2.40)	4.88 (2.24)
pcLSTM	25.28 (19.11)	3.64 (3.73)	71.08 (19.35)	90.79 (3.43)	6.93 (2.53)	2.28 (1.01)	85.78 (3.64)	10.83 (2.55)	3.40 (2.03)
pcLSTM*	23.82 (18.23)	3.72 (3.48)	72.45 (18.55)	91.20 (3.17)	6.67 (2.35)	2.13 (0.96)	86.33 (3.54)	10.44 (2.50)	3.23 (1.96)
gpcLSTM	53.66 (22.59)	4.34 (3.83)	42.00 (22.86)	89.39 (3.91)	7.99 (2.90)	2.63 (1.12)	84.61 (3.84)	11.79 (3.20)	3.61 (2.01)
gpcLSTM*	52.37 (22.06)	4.32 (3.15)	43.30 (22.42)	90.02 (3.69)	7.47 (2.77)	2.52 (1.04)	85.27 (3.69)	11.31 (2.95)	3.42 (2.02)
gpcLSTM$_{AC}$	**91.17 (8.50)**	1.26 (2.08)	**7.57 (8.01)**	91.61 (2.03)	6.62 (1.39)	1.77 (0.74)	**87.97 (5.00)**	**8.67 (2.64)**	**3.36 (2.63)**
gpcLSTM*$_{AC}$	91.02 (8.49)	**1.21 (1.97)**	7.77 (8.00)	**91.71 (2.02)**	**6.55 (1.34)**	**1.75 (0.77)**	87.95 (5.05)	8.69 (2.69)	**3.36 (2.62)**
gpcLSTM*$_{APAC}$	61.30 (20.12)	2.92 (2.38)	35.79 (20.23)	90.84 (3.57)	7.04 (2.57)	2.11 (1.07)	86.48 (3.95)	10.07 (2.66)	3.45 (2.31)

Tableau 5.7: Acceptabilité clinique (CG-EGA) par région moyenne (écart type) pour un horizon de prédiction de 30 minutes pour les jeux de données IDIAB et OhioT1DM.

modèles se focalisent sur l'amélioration de l'acceptabilité clinique uniquement. Ne cherchant pas à améliorer la précision des prédictions au-delà de la précision clinique nécessaire (zone B de la P-EGA), ces modèles possèdent une très mauvaise RMSE, acsmape et MASE. Ils possèdent néanmoins la meilleure acceptabilité clinique, avec le plus haut taux d'AP et le plus faible taux d'EP. L'amélioration est particulièrement importante en région d'hypoglycémie, comme peut en témoigner le Tableau 5.7.

Le modèle gpcLSTM$^*_{APAC}$ représente la dernière itération de l'algorithme APAC dont la MASE sur l'ensemble de validation est inférieure à 1. Ce modèle se veut maximiser l'acceptabilité clinique, tout en gardant une contrainte raisonnable sur la précision du modèle (MASE inférieure à 1). En comparaison avec le modèle gpcLSTM$^*_{AC}$, il possède une acceptabilité clinique légèrement inférieure (mais meilleure que tous les autres modèles, grâce notamment à son faible taux d'EP), mais avec une précision restant acceptable.

5.5.2 Discussion

Les résultats nous montrent que lissage exponentiel permet, en réduisant l'amplitude des variations entre prédictions successives, de baisser le taux d'erreurs bénignes (BE) au profit d'un meilleur taux d'AP. Cette amélioration

est valable pour la majeure partie des modèles et a pour contrepartie une baisse assez faible de la précision générale des modèles. Ainsi, le lissage exponentiel, utilisé de manière douce (coefficient β de 0.85) est une méthode efficace pour améliorer la stabilité du signal de glycémie prédit, le rendant plus sûr pour le patient diabétique. Il reste cependant inutile en région d'hypoglycémie où la majeure partie des erreurs cliniques de prédictions ont pour faute une mauvaise précision.

Les effets de l'utilisation de la fonction de coût cMSE sur les prédictions de glycémie sont similaires : les prédictions successives de glycémie sont plus cohérentes les unes avec les autres, résultant en une grande réduction du taux de BE. Les effets sont plus importants pour le jeu de données OhioT1DM qui voit son taux d'EP diminuer par la même occasion. Nous pouvons expliquer cela par un bruit plus important dans le signal de glycémie prédit du jeu OhioT1DM, bruit associé à sa fréquence de prédiction de 5 minutes (contre 15 minutes pour le jeu IDIAB). La cMSE permet d'avoir des prédictions successives dont le taux de variation tient mieux compte du taux de variation réel et permet ainsi d'améliorer son acceptabilité clinique. Toutefois, tout comme le lissage exponentiel, les améliorations de l'acceptabilité clinique ne se généralisent pas à l'ensemble des régions glycémiques. En particulier, la région d'hypoglycémie semble souffrir de l'utilisation de la cMSE avec une augmentation de son taux EP, notamment pour le jeu de données OhioT1DM.

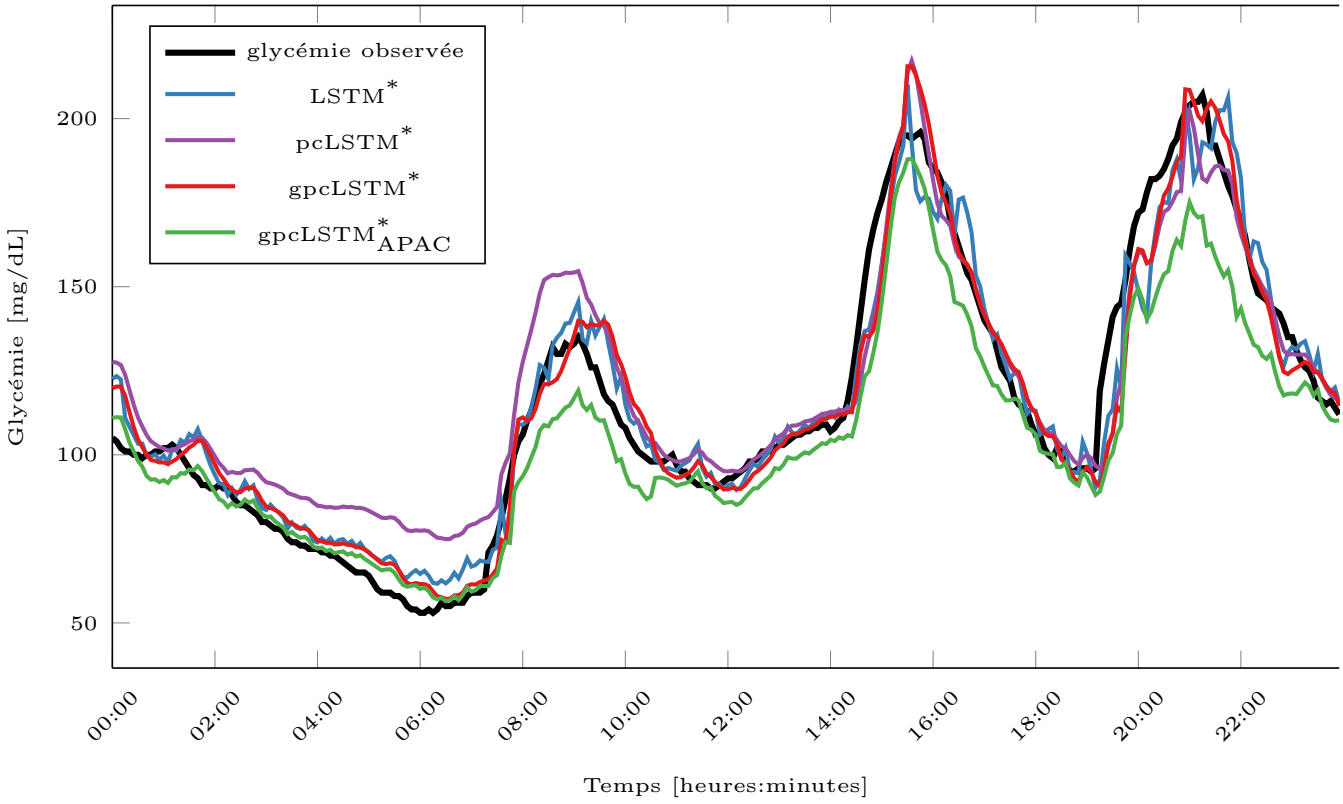


FIGURE 5.6: Prédictions des modèles LSTM, pcLSTM*, gpcLSTM* et gpcLSTM*$_{\text{APAC}}$ pour le patient 575 du jeu de données OhioT1DM pour un jour donné.

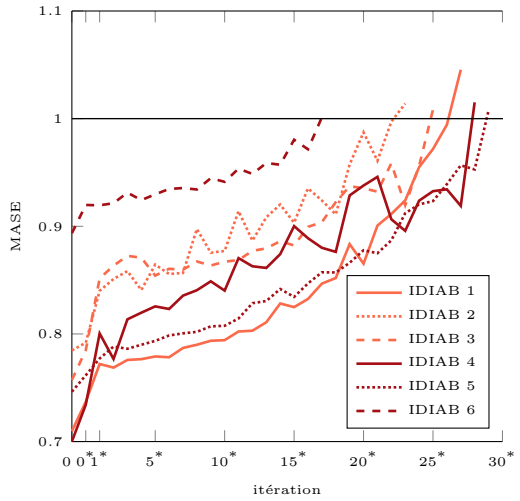
De son côté, l'action de la gcMSE se focalise plus sur la baisse du taux d'EP, comme le montrent les mo-

Critère Clinique		Jeu de données	
AP ($\geq$)	EP ($\leq$)	IDIAB	Ohio
80	-	6	6
90	-	6	3
95	-	4	0
97	-	3	0
-	7	6	6
-	5	6	4
-	3	6	3
-	1	4	0
80	7	6	6
90	5	6	3
95	3	4	0
97	1	2	0

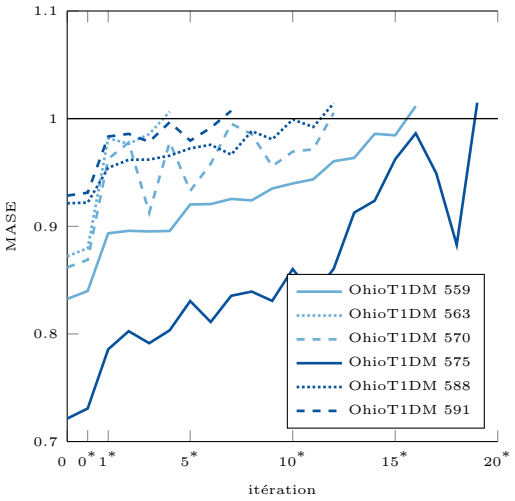
Tableau 5.8: Nombre de patients, par jeu de données, pouvant respecter, à travers l'utilisation de l'algorithme itératif d'amélioration de l'acceptabilité clinique, différents critères cliniques comme un seuil de prédiction AP minimal ou d'EP maximal.

dèles gpcLSTM, gpcLSTM$_{AC}$, gpcLSTM$^{*}_{APAC}$. Contrairement au lissage exponentiel et à la fonction de coût cMSE, la gcMSE permet d'améliorer l'ensemble des régions glycémiques, et en particulier celle de l'hypoglycémie. De plus, ces améliorations permettent au réseau de neurones de type LSTM de dépasser, en acceptabilité clinique, le modèle SVR qui est le meilleur modèle de l'étude benchmark. La Figure 5.6 permet d'apprécier les différences de prédiction des différents modèles. Premièrement, nous pouvons apercevoir les variations importantes ainsi que le bruit présents dans le signal de glycémie prédit du modèle LSTM*. Ces oscillations sont réduites pour les autres modèles, devenant ainsi plus fidèles du signal de glycémie observé. Cependant, dans le cadre de l'utilisation de la fonction de coût cMSE (signal pcLSTM* en violet), cela se fait au coup d'une grosse perte de précision dans la région d'hypoglycémie (entre 4h00 et 8h00). Tandis que le signal gpcLSTM$^{*}_{APAC}$ se montre être très fidèle au signal observé en région d'hypoglycémie, cela se fait au prix d'une perte de précision globale. Enfin, gpcLSTM*, est un compromis entre les deux.

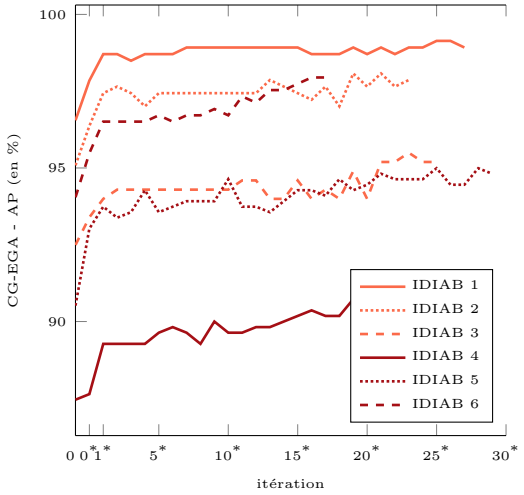
Bien que nous pouvons conclure sur l'intérêt de l'utilisation de la fonction de coût gcMSE dans l'apprentissage des modèles profond prédisant la glycémie future de personnes diabétiques, les différents résultats nous montrent l'existence de nombreux compromis possibles entre précision et acceptabilité clinique des prédictions. L'algorithme APAC proposé dans cette étude a pour objectif de permettre de sélectionner le meilleur compromis entre précision et acceptabilité clinique en fonction de critères de sélection. La Figure 5.7 donne une représentation graphique des changements, itération après itération, en MASE, du taux général AP et du taux général EP. Comme discuté précédemment, il n'existe aujourd'hui pas encore de critères cliniques pour les modèles prédictifs de glycémie, ainsi le seul critère d'arrêt de l'algorithme a été ici la MASE dépassant 1. La figure nous montre tout d'abord que le nombre d'itérations avant arrêt de l'algorithme est variable d'un jeu de donnée à l'autre, et aussi d'un patient à un autre (25.0 $\pm$ 3.96 pour le jeu IDIAB, et 11.66 $\pm$ 5.06 pour le jeu OhioT1DM). Cela s'explique tout d'abord



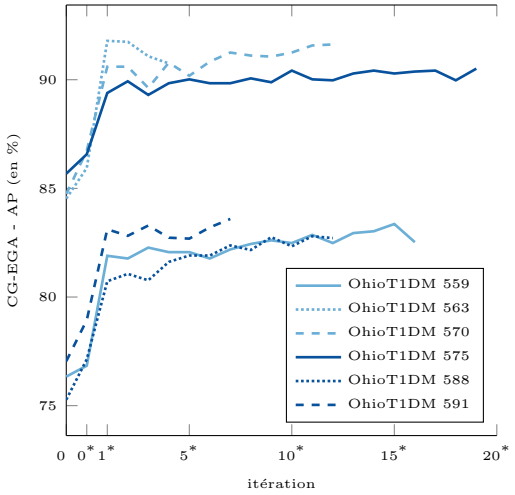
(a) Évolution de la MASE des patients IDIAB



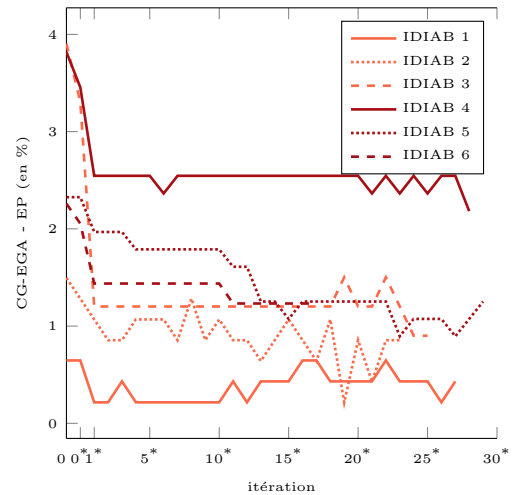
(b) Évolution de la MASE des patients OhioT1DM



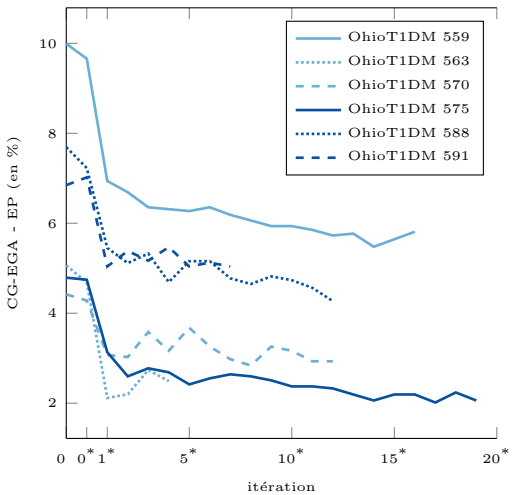
(c) Évolution du taux CG-EGA AP des patients IDIAB



(d) Évolution du taux CG-EGA AP des patients OhioT1DM



(e) Évolution du taux CG-EGA EP des patients IDIAB



(f) Évolution du taux CG-EGA EP des patients OhioT1DM

FIGURE 5.7: Évolution des métriques MASE et CG-EGA (AP et EP) tout au long de l'algorithme d'amélioration progressive de l'acceptabilité clinique pour les jeux de données IDIAB et OhioT1DM. Les itérations 0 et 0^* représentent respectivement les résultats du modèle entraîné avec la fonction de coût MSE avant et après lissage des prédictions.

par la précision initiale variable que les différents patients obtiennent, certains patients étant plus faciles à prédire que d'autres (voir itération 0 sur les Figures 5.7a et 5.7b). Comme nous avons pu l'observer à travers l'analyse du Tableau 5.6, les principales améliorations de l'acceptabilité clinique se font à la première itération (itération 1) de l'algorithme lors de l'introduction de la fonction de coût gcMSE et du lissage exponentiel. Néanmoins, tout au long de l'algorithme, l'acceptabilité clinique s'améliore progressivement, au détriment de la précision. Nous pouvons constater que le rythme de détérioration et d'amélioration est différent d'un patient à un autre, soulignant la très grande variabilité inter patient de la population diabétique.

Bien qu'il n'existe pas encore de critères cliniques pour les modèles de prédiction de glycémie, nous pouvons d'analyser l'utilisation de deux critères à travers le Tableau 5.8 : un taux d'AP minimal, et un taux d'EP maximal. Comme attendu, plus les critères cliniques sont élevés (seuil plus élevé et/ou combinaison de plusieurs critères), plus le nombre patients réussissant le test clinique est faible. Seul 1 patient du jeu de données IDIAB arrive à avoir simultanément plus de 97% d'AP et moins de 1% d'EP. Par ailleurs, nous pouvons noter un plus grand succès des patients IDIAB sur ces tests cliniques, par rapport aux patients OhioT1DM. Comme nous avons pu le souligner précédemment, ces différences de performances cliniques prennent leur source dans la différence de système expérimental. En effet, à cause des fréquences d'échantillonnage respectives des deux jeux de données, l'évaluation finale des patients OhioT1DM se fait toute les 5 minutes. De plus, le signal de glycémie des patients IDIAB est dans l'ensemble moins bruité, et donc plus stable et facile à prédire. Ainsi, dans l'optique de l'utilisation de tels critères cliniques dans le futur, ceux-ci, tout comme leur utilisation, doivent être rigoureusement standardisés.

Enfin, nous notons que la MASE sur l'ensemble de test (celle qui est reportée dans les Tableaux 5.6 et 5.7), est légèrement supérieure à 1 (1.03 et 1.01 pour les jeux de données IDIAB et OhioT1DM). En arrêtant l'algorithme lorsque la MASE sur l'ensemble de validation dépasse 1, nous pourrions supposer que la MASE finale sur l'ensemble de test soit, elle aussi, inférieure à 1. Cela s'explique par le fait que l'ensemble de test n'est pas totalement représentatif de l'ensemble de validation. Cela est dû à une faible quantité de données dans l'ensemble, impactant négativement la représentativité de ces ensembles. Nous notons d'ailleurs que l'écart type pour le jeu IDIAB est plus élevé, montrant que la valeur finale de la MASE est très variable en fonction du sujet. Ainsi, la précision de l'algorithme APAC serait améliorée en utilisant plus de données (ce qui améliorerait aussi les performances des modèles en général).

5.6 Conclusion

Dans ce chapitre, nous nous sommes intéressés à comment améliorer l'acceptabilité clinique, représentée par la métrique CG-EGA, des modèles basés sur l'apprentissage profond, et en particulier le modèle LSTM.

Tout d'abord, nous avons procédé à l'analyse des erreurs cliniques du modèle LSTM dans le cadre du benchmark présenté au Chapitre 4. Elle nous a montré que la majeure partie des erreurs cliniques provient d'un manque

de cohérence entre les prédictions successives. En effet, les variations de glycémie prédites sont généralement excessives, ce qui est dangereux pour le patient. Toutefois, les erreurs cliniques en région d'hypoglycémie sont dues à une précision trop faible. En effet, lorsqu'une hypoglycémie est observée, si le modèle ne prédit pas la présence d'une hypoglycémie (soit une glycémie en dessous de 70 mg/dL), alors l'erreur sera nécessairement grave. Ainsi, pour améliorer l'acceptabilité clinique des modèles, il faut à la fois améliorer la cohérence des prédictions successives, tout en améliorant la précision des prédictions en hypoglycémie.

Pour cela, dans un premier temps, nous avons proposé une première fonction de coût appelée erreur quadratique moyenne cohérente (cMSE). Utilisée avec un modèle LSTM à deux sorties, elle permet de pénaliser l'apprentissage du réseau de neurones non seulement sur sa précision, mais aussi sur la précision des variations prédites (différences entre les deux sorties du modèle). Puis, nous proposons de personnaliser cette fonction de coût à la prédiction de la glycémie à travers la gcMSE : chaque région des grilles P-EGA et R-EGA, responsables de l'acceptabilité clinique des modèles, se voit assigner un coefficient de pondération en fonction de leur importance relative. L'utilisation de la cMSE et de la gcMSE induit le choix d'un compromis entre précision et acceptabilité clinique du modèle. Afin de faciliter l'identification du compromis idéal, symbolisé par le modèle maximisant sa précision tout en respectant les critères cliniques, nous proposons l'algorithme APAC d'amélioration progressive de l'acceptabilité clinique. En partant d'une solution à précision maximale (minimisation de la MSE), les contraintes en précision sont petit à petit réduites afin de se focaliser sur l'acceptabilité clinique (au moyen de la gcMSE et d'un lissage exponentiel a posteriori des prédictions).

Les résultats expérimentaux, obtenus sur les jeux de données IDIAB et OhioT1DM, nous ont montré que la fonction de coût gcMSE permet bien d'améliorer l'acceptabilité clinique des prédictions. Ce gain se fait au prix d'une baisse de précision des prédictions. Nous avons montré que l'algorithme APAC permet d'utiliser des contraintes cliniques (sous la forme de seuils minimaux sur la CG-EGA par exemple), afin d'obtenir le modèle avec la meilleure précision tout en satisfaisant ces contraintes. Nous avons ensuite exploré différentes contraintes cliniques possibles, sous la forme de seuil minimal en prédictions AP et maximal en prédictions EP. Cette analyse nous a montré que la satisfaction de ces contraintes par les patients dépend à la fois du patient en lui-même (certains patients étant plus simple à prédire que d'autres), mais aussi du jeu de données (différent système expérimental, impliquant des données de nature différente ainsi qu'une évaluation elle aussi différente). Dans l'optique d'une future réglementation sur la validité clinique de dispositifs de prédiction de glycémie, ces différents éléments devront être pris en compte dans la création des standards.

6 | Apprentissage en situation de manque de données

Sommaire

6.1 Introduction

Dans le Chapitre 4, nous avons montré que les modèles basés sur l'apprentissage profond, bien que plus performants que les modèles linéaires classiques, le sont moins que d'autres modèles d'apprentissage automatique standards comme les machines à vecteur de support. Nous supposons que cela est dû au manque de données

d'apprentissage, et en particulier les évènements rares liés aux prises d'insuline ou de glucides. Ce manque de données, induisant le surapprentissage des modèles profonds, s'explique par la difficulté de récolter des données biomédicales sensibles. Ceci est accentué par le besoin d'avoir des modèles personnalisés au patient pour tenir compte de la grande inter et intra variabilité de la population diabétique.

Cette situation n'est pas spécifique à la prédiction de la glycémie et peut se généraliser à l'ensemble du secteur biomédical [16]. Premièrement, les données sont coûteuses à obtenir, car elles sont sensibles et demandent souvent des connaissances expertes pour être étiquetées (e.g., étiquetage de radiographies dans la détection de cancers, diagnostics de médecins lors de consultations successives pour la prédiction de troubles futurs). Ces données sont aussi potentiellement disponibles en trop petites quantités (e.g., détection de maladies rares). Enfin, ces données sont hétérogènes compliquant leur utilisation simultanée (e.g., données provenant de plusieurs hôpitaux différents) ou l'interopérabilité des modèles. Cette hétérogénéité s'explique notamment par une différence de matériels expérimentaux, de phénotypes ou de normes.

Pour pallier le manque de données disponibles, différentes stratégies peuvent être envisagées. Premièrement, les données originales peuvent être artificiellement augmentées en opérant des transformations (e.g., rotations d'images) ou en simulant artificiellement de nouvelles données [42]. Alternativement, l'efficacité des données peut être améliorée, avec par exemple des méthodologies d'apprentissage à partir de très peu d'exemples (*few-shot learning*) [6]). Enfin, des connaissances a priori peuvent être insérées dans les modèles profonds afin de réduire la quantité de données nécessaires à leur entraînement. Ces connaissances peuvent être des connaissances spécifiques à un domaine précis ou des connaissances expertes [72]. Elles peuvent prendre la forme de descripteurs calculés à partir d'équations mathématiques modélisant la physiologie du corps humain. Cependant, ces connaissances a priori sont souvent difficiles à obtenir et ne sont ni nécessairement adaptées à la tâche rencontrée, ni suffisamment précises. Par exemple, dans le domaine de la prédiction de la glycémie, il existe des modèles physiologiques décrivant les mécanismes mis en œuvre dans la régulation de la glycémie (ingestion de glucides, diffusion de l'insuline dans le sang, etc.) [74, 22]. Bien que certains groupes de chercheurs utilisent ces connaissances a priori en tant que données d'entrées aux modèles [165, 10], la méthode n'est pas généralisée au sein de la communauté [119]. Par ailleurs, il est possible d'obtenir des connaissances a priori plus adéquates à la tâche d'intérêt en entraînant dans un premier temps le modèle sur d'autres données plus ou moins similaires, puis en l'affinant avec les données d'intérêt. Cette dernière méthode est connue sous le nom d'apprentissage par transfert [9].

L'apprentissage par transfert (*transfer learning* en anglais) est particulièrement intéressant dans le secteur biomédical en raison de la grande variété des sources disponibles. Le transfert peut être effectué entre deux dossiers de santé électroniques (*electronic health records*) provenant d'hôpitaux différents, entre deux tâches ou données étroitement liées, ou entre deux patients dans le cadre de la médecine personnalisée. Plusieurs sources peuvent être combinées afin de faciliter l'extraction de connaissances (e.g., données provenant de plusieurs hôpitaux, transférées vers un nouvel hôpital) [18]. Cette technique a pour nom l'apprentissage par transfert multi-sources. L'utili-

sation de plusieurs sources permet l'utilisation de plus grandes quantités de données qui sont difficiles à obtenir pour chaque source. En outre, elle offre la possibilité de rendre les connaissances extraites plus générales et donc plus facilement transférables. Cependant, l'efficacité dépend fortement de la similarité entre les tâches sources et la tâche cible. Si les tâches sont trop différentes entre elles, cela peut donner lieu à un transfert négatif, impactant négativement les performances finales [158]. Enfin, il existe aussi le risque que le modèle apprenne à discriminer les données de différentes sources. Les connaissances extraites seraient ainsi surspécialisées aux domaines sources, et donc difficilement transférables vers le domaine cible.

Ganin *et al.* ont abordé ce problème dans le contexte de l'adaptation de domaine en proposant une méthodologie d'entraînement adverse sur deux domaines simultanés [57]. L'adaptation de domaine est un sous-secteur de l'apprentissage par transfert. Il diffère de l'apprentissage par transfert standard (et donc de l'apprentissage par transfert multi-sources) par un entraînement du modèle se faisant simultanément sur les données sources (vérité terrain connue) et les données cibles (vérité terrain inconnue). Dans ce cadre, l'entraînement adverse implémente un module qui classifie l'origine (source ou cible) des caractéristiques intermédiaires extraites, et un module d'extraction de caractéristiques travaillant contre cet objectif. Cela a pour conséquence de favoriser l'apprentissage d'une représentation des données commune entre la source et la cible (et, en ce sens, *adapter* la représentation des données sources aux données cibles).

Dans ce chapitre, nous proposons de transposer cette idée à l'apprentissage par transfert multi-sources, avec pour objectif d'apprendre une représentation de caractéristiques la plus agnostique possible des sources. Cette représentation de caractéristiques serait alors plus générale et ainsi plus adaptée au transfert vers une cible inconnue. Dans le cadre de cette thèse, plusieurs données sources peuvent être envisagées. En effet, nous possédons trois jeux de données hautement dissimilaires (différent type de diabète, différents capteurs ou conditions expérimentales, données réelles ou simulées) et chacun de ces jeux possède entre 6 et 10 patients. Cette grande variabilité peut être utilisée afin d'apprendre des connaissances a priori plus générales dans l'optique de faciliter l'apprentissage des modèles de prédiction de glycémie personnalisés.

À ce jour, peu d'études se sont intéressées à l'application de l'apprentissage par transfert pour la prédiction de la glycémie. Toutefois, de très récentes études comme celles de Zhu *et al.* [169, 168] ou de Mirshekarian *et al.* [115] suggèrent qu'il permettrait de significativement améliorer les performances des modèles. En effet, en les comparant aux modèles du Chapitre 4, leurs modèles basés sur les réseaux de neurones récurrents ont montré une meilleure précision que notre modèle LSTM sur le jeu OhioT1DM à un horizon de prédiction de 30 minutes. En pratique, l'apprentissage par transfert consiste en l'apprentissage d'un premier modèle sur plusieurs patients sources, puis en l'affinage de celui-ci sur le patient cible. En présence de quantité de données insuffisantes sur le patient cible, le modèle peut alors utiliser ces connaissances a priori pour améliorer sa précision. Toutefois, ces études ne se focalisent pas sur l'utilisation de l'apprentissage par transfert pour la prédiction de la glycémie. Ainsi, l'intérêt de celui-ci n'est aujourd'hui pas quantifié.

Ainsi, plusieurs questions peuvent être soulevées quant à l'utilisation de l'apprentissage par transfert pour la tâche de la prédiction de la glycémie :

1. Peut-on effectuer un transfert entre patients diabétiques réels, étant donné la forte variabilité inter/intra patient de la maladie ?
2. L'apprentissage par transfert est-il utile pour les modèles basés sur l'apprentissage profond ?
3. Pouvons-nous transférer de la connaissance entre patients dont les données ont été collectées dans différents contextes expérimentaux (e.g., différents capteurs, environnements) ?
4. Peut-on effectuer un transfert entre les patients diabétiques de type 1 et de type 2 ?
5. Les données synthétiques peuvent-elles être utilisées pour le transfert vers des données réelles ?

Ayant pour objectif de répondre à ces questions, ce chapitre est structuré comme suit. Dans un premier temps, nous énonçons le cadre théorique et applicatif de l'apprentissage multi-sources adverse. Après avoir présenté la méthodologie de l'étude et ses spécificités, nous démontrons l'efficacité de l'apprentissage par transfert multi-sources adverse. En outre, nous proposons une analyse en profondeur du nouveau mécanisme d'extraction de connaissances. En améliorant de manière significative les meilleurs résultats dans le domaine, notre apprentissage par transfert multi-sources adverse se montre être efficace pour répondre au problème de manque de données dans le domaine de la prédiction de la glycémie, et dans le domaine biomédical de manière plus générale.

6.2 Fondements de l'apprentissage par transfert multi-sources adverse

Le but de cette section est de donner une définition formelle de l'apprentissage par transfert multi-sources, de mettre en évidence les défis de son application dans la santé, et enfin de décrire la méthodologie d'apprentissage par transfert adverse proposée qui vise à les résoudre.

6.2.1 Apprentissage par transfert

Un *domaine* $\mathcal{D}$ est défini par un espace de caractéristiques $\mathcal{X}$ et une distribution de probabilité marginale $P(X)$, où $X \in \mathcal{X}$. Une *tâche* $\mathcal{T}$ consiste en un espace objectif $\mathcal{Y}$ et une fonction prédictive de l'objective $f(\cdot)$. $f(\cdot)$ est inconnu mais peut être apprise à partir d'échantillons de données $\{x_n, y_n\}$, où $x_n \in X$, $y_n \in \mathcal{Y}$ et $n \in [1, N]$.

Pan et Yang ont défini l'apprentissage par transfert comme suit [120] : étant donné un domaine source $\mathcal{D}_S$ et une tâche d'apprentissage $\mathcal{T}_S$, un domaine cible $\mathcal{D}_T$ et la tâche d'apprentissage $\mathcal{T}_T$, l'*apprentissage par transfert* $\{\mathcal{D}_S, \mathcal{T}_S\} \rightarrow \{\mathcal{D}_T, \mathcal{T}_T\}$ vise à améliorer l'apprentissage de la fonction prédictive cible $f_T(\cdot)$ dans $\mathcal{D}_T$ sur $\mathcal{T}_T$ en utilisant les connaissances dans $\mathcal{D}_S$ sur $\mathcal{T}_S$, avec $\mathcal{D}_S \neq \mathcal{D}_T$ ou $\mathcal{T}_S \neq \mathcal{T}_T$ [1].

1. Les lettres S et T font référence aux mots *source* et *target* en anglais.

Dans cette étude, nous nous concentrons sur l'apprentissage par transfert *inductif*, qui est le type d'apprentissage par transfert le plus courant. L'apprentissage par transfert inductif se caractérise par des tâches source et cible étroitement liées, et par la présence d'un nombre faible de données étiquetées dans le domaine cible et suffisamment de données étiquetées dans le domaine source. En utilisant des modèles basés sur l'apprentissage profond, ce type de transfert s'effectue généralement en entraînant un premier modèle sur le domaine source, puis en l'affinant (dans sa totalité ou en partie) sur le domaine cible.

6.2.2 Apprentissage multi-sources standard

Dans l'apprentissage par transfert *multi-sources*, les connaissances que nous visons à transférer sont apprises conjointement sur plusieurs paires sources $\{\mathcal{D}_{S_n}, \mathcal{T}_{S_n}\}$, chacune d'elles étant différente l'une de l'autre et étant différente de la cible $\{\mathcal{D}_T, \mathcal{T}_T\}$. Entraîner un modèle sur plusieurs sources permet de répondre à un potentiel problème de manque de données au sein des sources prises individuellement.

Toutefois, la nature des domaines sources peut varier considérablement. Cela nous oblige à être prudents lors de leur sélection, afin d'éviter que le transfert nuise à l'apprentissage de la tâche cible dans le domaine cible (transfert négatif). Cette prudence doit être exacerbée dans le secteur médical où les données sont hétérogènes, ayant des distributions de probabilité différentes en raison de leur origine (e.g., différents patients, capteurs, environnements expérimentaux), ayant différents formats (e.g., résolution d'image, fréquence d'échantillonnage), ayant différentes échelles (e.g., échelle de couleurs pour les images, unités pour les mesures physiques ou physiologiques), ou simplement étant présentes en quantités variables [153].

Pour rendre un transfert positif possible, ces différences doivent être traitées avant ou pendant l'entraînement du modèle que nous voulons transférer. Les différences d'échelle ou de format peuvent être facilement éliminées grâce à la transformation préalable des données (redimensionnement, changement d'échelle, standardisation). Le problème de quantités de données inégales entre les domaines peut être résolu de la même manière que les jeux de données déséquilibrés pour des tâches de classification avec, par exemple, les techniques de pondération d'échantillons ou d'augmentation des données [84]. Cependant, ces techniques présentes des limitations. Pondérer les échantillons en fonction de leur rareté ne reflète pas nécessairement la réalité, car crée une surreprésentation des données les moins fréquentes. Quant à l'augmentation des données, celle-ci nécessite des connaissances expertes qui ne sont pas nécessairement disponibles.

6.2.3 Apprentissage multi-sources adverse

La différence dans les distributions de probabilité des différentes sources pourrait être saine pour la construction d'un modèle général, en tirant parti de la diversité des données. Cependant, il existe un risque, pour le modèle, d'apprendre à discriminer les différents domaines et en apprenant une représentation distincte des caractéristiques

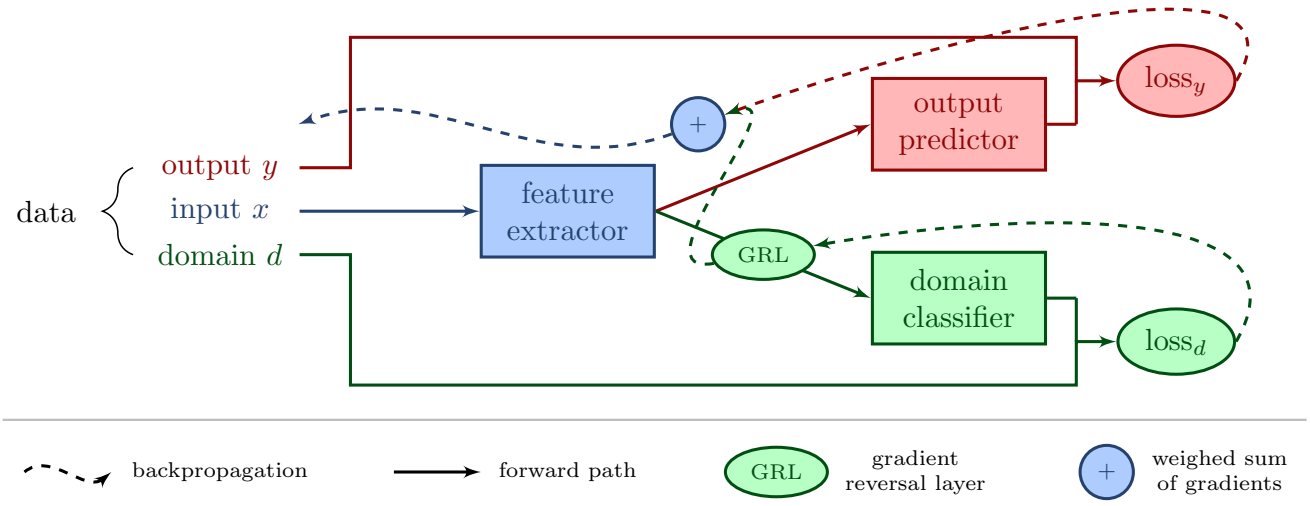


FIGURE 6.1: Représentation générale d'un modèle basé sur l'apprentissage profond, entraîné avec la méthodologie d'apprentissage adverse dans le cadre de l'apprentissage par transfert multi-sources. Le modèle est composé de trois parties différentes : un *extracteur de caractéristiques*, un *prédicteur de sortie* et un *classificateur de domaine*. La nature adverse de l'apprentissage est obtenue en faisant en sorte que l'extracteur de caractéristiques entre en concurrence avec le classificateur de domaine grâce à l'inversion de loss$_d$ (multiplication par -1) du gradient lors de sa rétropropagation.

pour chacun d'eux. Ce type de représentation des caractéristiques serait moins générale en étant fortement spécifique à ces domaines individuels. En conséquence, cela nuirait à son futur transfert vers le domaine cible. Pour résoudre ce problème, inspiré par les travaux de Ganin *et al.* dans le secteur de l'adaptation de domaine, nous proposons la méthodologie d'apprentissage par transfert multi-sources adverse.

La Figure 6.1 fournit une représentation graphique d'un modèle utilisant la méthodologie d'entraînement adverse dans un environnement d'apprentissage par transfert multi-sources. Dans un premier temps, la représentation cachée des variables d'entrées est calculée par le module d'extraction de caractéristiques. Cette représentation cachée est utilisée par le prédicteur de sortie pour effectuer les prédictions de la tâche en question. Parallèlement, elle est aussi utilisée par le classificateur de domaine qui prédit le domaine d'origine des échantillons des données. Le prédicteur de sortie et le classificateur de domaine sont tous les deux entraînés de manière classique en rétropropageant les pénalités de leur fonction de coût respective. Tandis que la fonction de coût du prédicteur de sortie dépend du problème que le modèle vise à résoudre (i.e., problème de régression ou de classification), la fonction de coût du classificateur de domaine est l'entropie croisée multi-classes. Ici, chaque classe représente un domaine source particulier. Afin d'apprendre une représentation cachée générale à l'ensemble des domaines sources, les pénalités associées au classificateur de domaine sont inversées (multipliées par -1) au niveau du module d'extraction de caractéristiques. Cela pousse ce dernier à apprendre une représentation cachée à la fois utile pour la prédiction finale et depuis laquelle les domaines d'origines sont indifférenciables. Le compromis biais-variance de la généralisation peut être modifié en ajustant le coefficient λ qui pondère la magnitude du gradient lié au classificateur de domaine. Chacune des sources utilisées n'a pas nécessairement la même importance. Celle-ci

dépend de la proximité de la source avec le domaine cible. Aussi, certains domaines peuvent être sous-représentés à cause du faible nombre de données d'entraînement sur ceux-ci. Afin de favoriser davantage le transfert, il est possible d'assigner une pondération différente à chacun des domaines lors de leur prédiction par le classificateur de domaine.

En pratique, pour réaliser le transfert, une fois le modèle de la Figure 6.1 entraîné sur les domaines sources, celui-ci peut être affiné sur le domaine cible. Pendant l'affinage, le classificateur de domaine n'a plus d'utilité et peut ainsi être retiré (soit en transférant les poids de l'extracteur de caractéristiques et du prédicteur de sortie dans un nouveau modèle qui n'a pas de classificateur de domaine, soit en fixant λ à 0).

6.3 Méthodologie

Dans cette section, nous détaillons la méthodologie suivie pour l'évaluation de l'apprentissage par transfert multi-sources adverse pour la prédiction de la glycémie de patients diabétiques. Nous avons mis à disposition l'ensemble du code Python de la chaîne de traitement de données sur GitHub [29].

6.3.1 Données expérimentales

Dans cette étude, nous utilisons les trois jeux de données IDIAB (I), OhioT1DM (O) et T1DMS (T). Dans le cadre de l'apprentissage par transfert, tandis que les patients de chaque jeu sont utilisés comme patients sources, seulement les patients réels provenant de IDIAB et OhioT1DM sont utilisés comme patients cibles.

6.3.2 Prétraitement

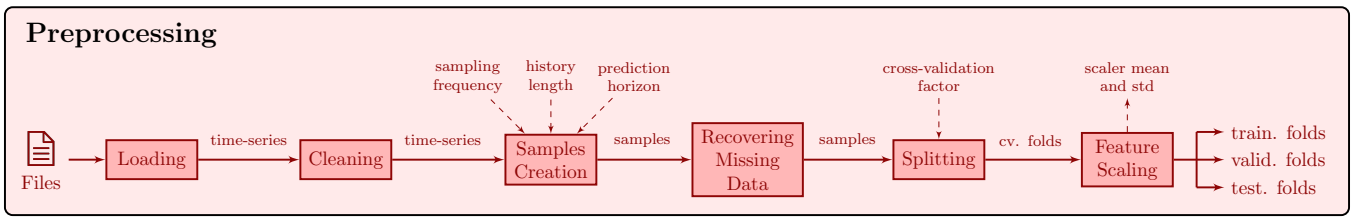


FIGURE 6.2: Étapes de prétraitement des données.

En comparaison avec les étapes de prétraitement des données des patients présentées au Chapitre 3, dont la Figure 6.2 en fait le rappel, nous avons apporté quelques légères modifications visant à faciliter l'apprentissage et le transfert des modèles sources.

Premièrement, pour qu'un modèle soit interopérable entre les différents ensembles de données, les données de chaque ensemble doivent avoir le même format et la même échelle. Le jeu T1DMS diffère des jeux OhioT1DM et IDIAB en mesurant les prises d'insuline en pmol au lieu d'unités et par la répartition des ingestions de glucides sur

l'ensemble de la durée du repas. Pour remettre les valeurs d'insuline à la même échelle (en unités), nous avons effectué la transformation décrite par l'équation 6.1. Quant à la gestion des données de glucides, elles ont été accumulées sur la durée du repas (15 minutes, soit 3 échantillons avec un échantillon toutes les 5 minutes) à son commencement.

$$insuline_t \leftarrow insuline_t / 6000 \tag{6.1}$$

Par ailleurs, les données des patients sources ne sont pas utilisées pour l'évaluation des modèles qui se fait sur les données du patient cible, après transfert. Ainsi, il n'est pas utile de posséder des ensembles de test pour les données des patients sources. En conséquence, les ensembles de tests de chaque patient source ont été considérés comme des ensembles de validation, et les ensembles de validation originaux ont été ajoutés aux ensembles d'entraînement. Cela a pour conséquence d'augmenter la quantité de données d'entraînement disponible, et ainsi d'améliorer la performances des modèles. Enfin, la standardisation des échantillons de données a été faite sur les ensembles de chaque patient pris individuellement. Cela permet notamment d'avoir une moyenne uniforme sur l'ensemble des patients, réduisant ainsi la variabilité entre patients.

6.3.3 Présentation des modèles prédictifs

Nous présentons ici les modèles qui ont été utilisés dans cette étude, la plupart basée sur des réseaux de neurones entièrement convolutifs, *Fully Convolutional Neural Networks* (FCN). Bien que non-utilisés dans le chapitre 4, ceux-ci ont montré des résultats prometteurs dans la littérature récente de la prédiction de la glycémie [96, 168] ains que dans nos expérimentations préléminaires. De plus, ce sont les modèles utilisés originellement dans l'étude de Ganin *et al.* proposant la méthodologie d'apprentissage adverse dans le contexte de l'adaptation de domaine. Les poids des modèles globaux pré-affinage sont disponibles sur GitHub [29].

Modèles de références

Dans cette étude, nous utilisons trois modèles de référence : un modèle SVR et deux modèles FCN, à savoir FCN #1 et FCN #2. Les trois modèles sont uniquement entraînés et évalués sur les patients cibles.

Le modèle SVR est basé sur l'étude benchmark présentée au Chapitre 4, dans laquelle il se distingue des autres modèles en étant le meilleur modèle pour la prédiction de la glycémie.

Quant aux modèles FCN #1 et #2, bien qu'ils partagent la même architecture, ils diffèrent dans leurs hyperparamètres d'entraînement. Ils sont constitués de 2 modules : un extracteur de caractéristiques et un prédicteur de glycémie. Le module d'extraction de caractéristiques est composé de 3 couches (convolution unidimensionnelle de taille 3 $\rightarrow$ fonction d'activation ReLU $\rightarrow$ *batch normalization* $\rightarrow$ *dropout*) possédant 64, 128 et 64 canaux respec-

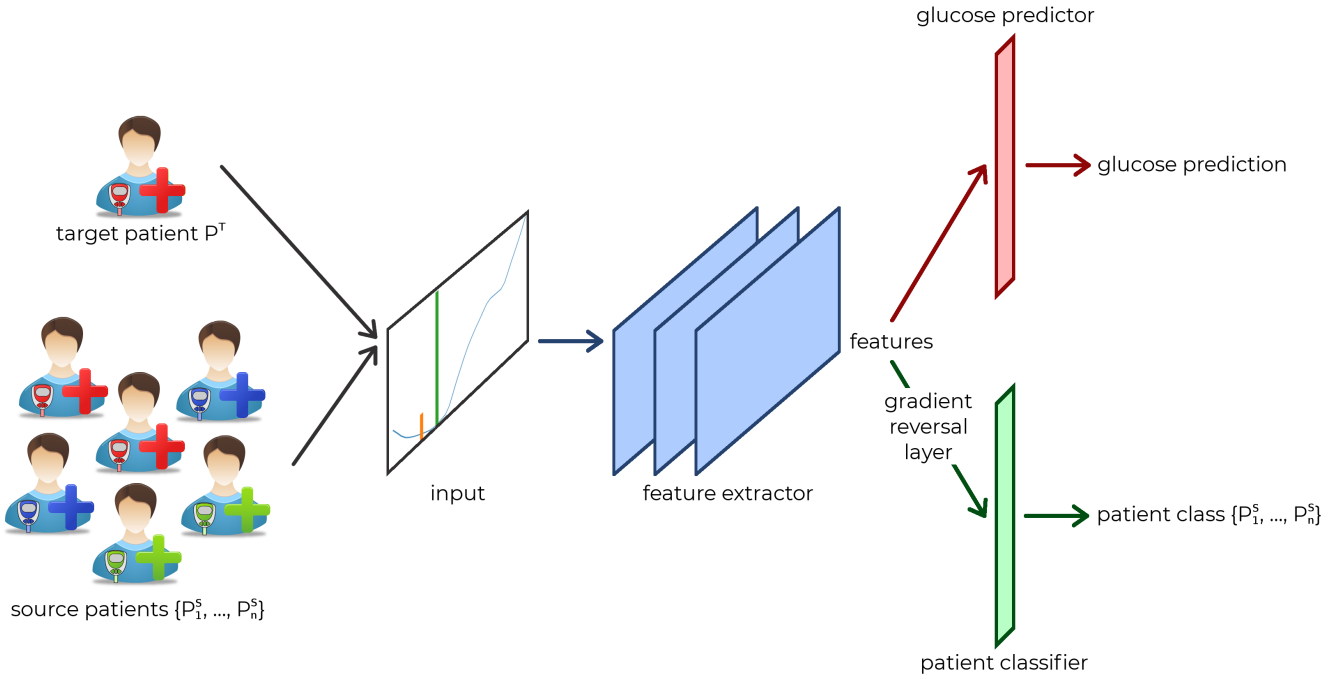


FIGURE 6.3: Apprentissage par transfert multi-sources adverse, basé sur un réseau de neurones convolutif pour la prédiction de la glycémie de personnes diabétiques. Un premier modèle est entraîné sur les patients sources qui peuvent provenir de différents ensembles de données et est ensuite affiné sur le patient cible. Grâce à la couche d'inversion de gradient (*gradient reversal layer*), le module de classification de patients permet à l'extracteur de caractéristiques d'apprendre une représentation de caractéristiques commune aux patients sources et donc plus générale.

tivement. Le module prédicteur est lui constitué d'une couche dense de 2048 neurones (implémentée comme une convolution de taille 30 et de 2048 canaux). Concernant leur entraînement, bien qu'ils utilisent tous deux l'optimiseur Adam pour minimiser la fonction de coût MSE avec une patience d'arrêt anticipé de 250 époques, FCN #1 utilise un taux d'apprentissage de 10^{-4} et un taux de *dropout* de 50%, et FCN #2 utilise un taux d'apprentissage de 10^{-3} et un taux de *dropout* de 90%.

Nous avons choisi d'utiliser ces deux réseaux convolutifs car FCN #1, bien qu'identique aux modèles utilisés pour le transfert (voir ci-dessous), ne présentait pas de performances satisfaisantes. Cela est dû à la faible quantité de données allouées aux modèles de référence, aspect auquel les réseaux convolutifs sont particulièrement vulnérables. FCN #2 combat ce manque de données en implémentant une régularisation extrême (taux de *dropout* à 90%). En utilisant FCN #1 et FCN #2 comme modèles convolutifs de référence, nous assurons la pertinence de l'évaluation des modèles.

Modèles de transfert standard

Nous appelons les modèles d'apprentissage par transfert standard (TL), les modèles qui ont été entraînés sur les patients source (TL global) puis affinés au patient cible (TL affiné). Tous les modèles TL sont identiques dans leur architecture à FCN #1. Pendant l'affinage sur le patient cible, le taux d'apprentissage et la patience d'arrêt

anticipé ont été réduits à 10^{-5} et 50 respectivement.

Modèles de transfert adverse

Les modèles issus de l'apprentissage par transfert adverse (ATL, global et affiné) et les modèles TL partagent la même architecture et la même méthodologie d'entraînement, à l'exception de la présence du module de classificateur de domaine (de patient) et sa couche d'inversion de gradient. La Figure 6.3 donne une représentation graphique d'un réseau de neurones convolutifs utilisant la méthodologie d'apprentissage par transfert multi-sources adverse.

Le classificateur de domaine est constitué d'une seule convolution de taille 30 et de 2048 canaux (identique au module prédicteur de glycémie). Il minimise l'entropie croisée multi-classes pondérée par $\lambda = 10^{-0.75}$ avec l'optimiseur Adam. La valeur de λ a été choisie parmi 9 valeurs comprises entre 10^{-2} et 10^{1} dans une échelle logarithmique [57]. Pour tenir compte de la représentation déséquilibrée des patients dans les ensembles d'entraînement (et en particulier de la sous-représentation des patients IDIAB), les pénalités associées aux échantillons d'un patient donné sont pondérées inversement proportionnellement au nombre total d'échantillons de ce patient.

6.3.4 Évaluation des modèles prédictifs

Comme pour les autres études présentées dans cette thèse, l'évaluation des modèles prédictifs suit les étapes présentées dans l'étude benchmark du Chapitre 4 dont la Figure 6.4 en fait le rappel. Dans cette étude, nous nous intéressons qu'à l'horizon de prédiction court-terme de 30 minutes, horizon pour lequel les modèles prédictifs présentent déjà des difficultés.

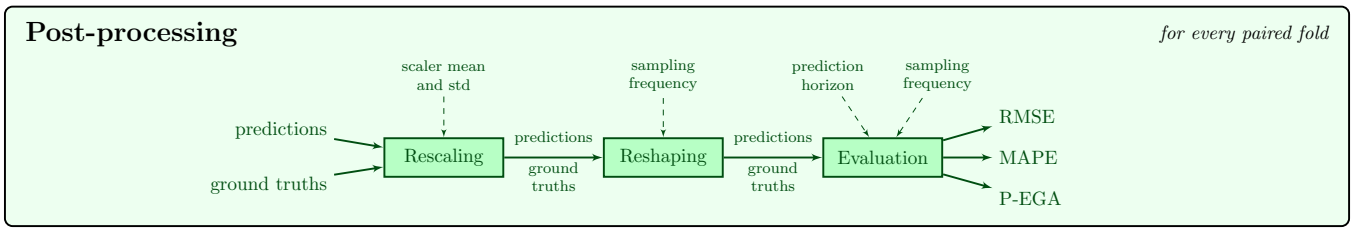


FIGURE 6.4: Post-traitement et évaluation des prédictions de glycémie. En comparaison avec le schéma du Chapitre 4, la métrique du gain temporel TG a été retirée, et celle de l'acceptabilité clinique CG-EGA a été remplacée par la P-EGA.

Métriques d'évaluation

L'objectif de l'apprentissage par transfert est ici d'optimiser la précision des prédictions de glycémie, ainsi nous utilisons principalement les métriques RMSE et acsmape pour évaluer les modèles. La CG-EGA, qui donne une mesure de l'acceptabilité clinique des modèles, se scinde en deux grilles : la première (P-EGA) évaluant l'acceptabilité des prédictions via leur précision et la seconde (R-EGA) évaluant l'acceptabilité des variations des prédictions.

Afin de mieux visualiser l'impact du transfert sur l'acceptabilité clinique, nous ne reportons ici que la grille liée à la précision (P-EGA).

Pour un patient cible donné, les performances sont moyennées sur les permutations des ensembles d'entraînement et de validation du patient cible. Puis, les performances sur un jeu de données (IDIAB ou OhioT1DM) sont obtenues en faisant la moyenne des performances de la population.

Pour l'analyse des résultats, nous considérons tous les différents scénarios de transfert possibles, comme le scénario IDIAB $\rightarrow$ IDIAB, écrit de manière équivalente I$\rightarrow$I, ou OhioT1DM $\rightarrow$ IDIAB, O$\rightarrow$I. Ces différents scénarios peuvent être regroupés en différentes catégories afin de faciliter l'analyse des résultats :

— **intra** : les patients sources et cibles appartiennent au même jeu de données (I$\rightarrow$I et O$\rightarrow$O) ;

— **inter** : les patients sources et cibles n'appartiennent pas au même jeu de données, et les patients sont réels (O$\rightarrow$I et I$\rightarrow$O) ;

— **synth** : les patients sources sont virtuels (par exemple, T$\rightarrow$I, T$\rightarrow$O) ;

— ainsi que toutes les combinaisons de ces trois scénarios de transfert (e.g., le scénario intra+inter regroupant à la fois IO$\rightarrow$I et IO$\rightarrow$O).

Analyse de la significativité statistiques des résultats

Pour analyser la pertinence des résultats obtenus, nous proposons d'analyser la significativité de l'amélioration de la acsmape des modèles. Nous avons choisi la acsmape, plutôt que la RMSE, car elle est indépendante de l'échelle des prédictions. La significativité de l'amélioration des performances d'une condition expérimentale par rapport à une autre selon une métrique donnée se calcule à partir du ratio appairé des performances et s'exprime sous la forme d'un intervalle de confiance. Si l'intervalle contient la valeur 1, alors l'hypothèse H_0 d'absence de différence significative ne peut pas être rejetée et nous ne pouvons pas conclure. Si la borne supérieure de l'intervalle est inférieure à 1, alors nous observons une amélioration significative (si nous cherchons à minimiser la métrique choisie, ce qui est le cas avec la acsmape). Enfin, si la borne inférieure de l'intervalle est supérieure à 1, nous pouvons conclure à une détérioration significative des performances.

Nous préférons utiliser les intervalles de confiance plutôt que des valeurs-*p* car ces dernières sont plus délicates à interpréter. Contrairement aux valeurs-*p* qui apportent une information binaire, les intervalles de confiance nuancent l'information en quantifiant l'effet étudié et le degré de certitude des estimations [11].

Nous avons choisi de nousintéresser aux intervalles de confiance à 99% qui nous permettent de réduire l'impact du hasard dans le processus d'entraînement des modèles (notamment lié à l'initialisation des poids). La loi de probabilité utilisée est la loi de Student. La taille des échantillons sur laquelle nous calculons le ratio appairé de acsmape est de 60 (i.e., nombre de patients fois le nombre de plis dans la validation croisée, soit $12 \times 5 = 60$).

Jeu de données	Modèle	RMSE	MAPE
IDIAB (I)	SVR *	**20.32 (6.02)**	**8.66 (0.44)**
	FCN #1	21.06 (5.14)	9.66 (1.00)
	FCN #2	20.64 (5.20)	9.62 (1.27)
OhioT1DM (O)	SVR *	**20.15 (2.33)**	**9.12 (2.11)**
	FCN #1	21.51 (1.89)	9.82 (2.08)
	FCN #2	20.61 (2.09)	9.34 (2.07)

* Ces résultats sont ceux présentés dans le Chapitre 4

Tableau 6.1: Précision (RMSE et MAPE) moyenne (avec écart type) des prédictions de glycémie par modèle de référence pour les jeux de données IDIAB et OhioT1DM.

6.4 Résultats expérimentaux

Dans cette section, nous nous intéresserons tout d'abord aux résultats obtenus par les modèles de référence, puis aux résultats des modèles utilisant les différentes méthodologies de transfert, et enfin au comportement de l'apprentissage des connaissances par le réseau multi-sources adverse en comparaison avec le réseau multi-sources standard.

6.4.1 Résultats de références

Le Tableau 6.1 présente la précision des trois modèles de référence SVR, FCN #1 et FCN #2 pour les deux jeux de données IDIAB et OhioT1DM. Ces résultats de référence nous montrent tout d'abord que les réseaux convolutifs, non utilisés dans notre étude benchmark, n'arrivent pas à obtenir de meilleures performances que le modèle SVR. Cette difficulté, partagée avec les autres modèles basés sur les réseaux de neurones, s'explique par leur surapprentissage aux données d'entraînements, surapprentissage qui est en partie surmonté par FCN #2 par une régularisation accrue (*dropout* à 90% au lieu de 50% pour FCN #1).

6.4.2 Résultats par scénario et type de transfert

Le Tableau 6.2 résume les résultats obtenus par le transfert standard et le transfert adverse pour les différents scénarios de transfert. Les résultats sont représentés avant affinage (transfert global) et après affinage (transfert affiné) au patient cible.

Transfert standard global

Les performances des modèles TL globaux sont variables en fonction à la fois des patients sources et des patients cibles (sur lesquels le modèle non personnalisé est testé). Dans le domaine de la prédiction de la glycémie, les modèles globaux sont supposés être moins bons que les modèles personnalisés, car ne tenant pas compte de

Jeu de données		Transfert Standard (global)		Transfert Adverse (global)		Transfert Standard (affiné)		Transfert Adverse (affiné)	
S	**C**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
I	I	21.47 (7.50)	9.67 (1.48)	**19.61 (6.27)**	8.95 (1.00)	20.25 (5.02)	8.96 (1.50)	**18.51 (5.48)**	**8.44 (1.07)**
O		21.70 (5.75)	10.22 (1.85)	19.87 (6.01)	9.01 (1.52)	19.26 (4.97)	9.13 (1.26)	18.84 (5.75)	8.57 (1.11)
T		25.47 (6.00)	11.11 (1.60)	29.57 (6.01)	14.53 (3.14)	20.08 (4.94)	9.26 (0.85)	19.50 (5.14)	9.02 (1.16)
IO		*20.20 (5.90)*	9.51 (1.49)	19.66 (6.48)	**8.90 (1.52)**	*19.10 (5.04)*	*8.95 (1.00)*	18.75 (6.01)	8.50 (1.23)
IT		22.25 (8.28)	10.61 (2.91)	22.96 (8.22)	10.52 (1.98)	19.45 (5.08)	9.03 (1.16)	19.70 (6.21)	8.93 (1.10)
OT		22.25 (8.28)	10.61 (2.91)	23.16 (6.44)	10.20 (1.61)	19.45 (5.31)	9.04 (1.20)	19.47 (6.60)	8.79 (1.13)
IOT		20.72 (6.34)	*9.45 (2.02)*	22.46 (9.40)	9.87 (2.07)	19.44 (5.24)	9.02 (1.07)	18.79 (5.82)	8.59 (1.05)
I	O	24.01 (2.24)	11.62 (1.85)	21.45 (1.50)	10.15 (1.70)	20.52 (2.08)	9.49 (2.18)	19.74 (2.13)	8.96 (2.02)
O		21.95 (1.98)	10.20 (2.10)	20.22 (1.48)	9.18 (1.83)	19.92 (2.02)	9.09 (2.14)	19.27 (1.78)	8.68 (1.97)
T		30.17 (4.64)	14.18 (4.30)	36.63 (7.99)	18.16 (5.35)	20.20 (1.99)	9.20 (2.03)	19.93 (1.74)	9.13 (1.87)
IO		21.17 (2.16)	9.81 (2.04)	19.58 (1.65)	**9.04 (2.10)**	19.91 (2.01)	9.06 (2.08)	**18.94 (1.66)**	**8.50 (1.87)**
IT		23.46 (2.60)	11.58 (2.75)	26.44 (5.25)	13.37 (2.86)	20.03 (1.88)	9.25 (2.08)	19.57 (2.02)	8.81 (1.85)
OT		21.39 (2.16)	9.72 (2.16)	19.88 (1.26)	9.36 (1.55)	19.72 (2.04)	8.97 (2.18)	19.16 (1.73)	8.64 (1.94)
IOT		*20.68 (2.12)*	*9.58 (2.14)*	**19.45 (1.78)**	9.19 (1.91)	*19.57 (2.02)*	*8.93 (2.13)*	18.99 (1.72)	8.56 (1.89)

Tableau 6.2: Précision (RMSE et MAPE) moyenne (avec écart type) des modèles après transfert, avec ou sans affinage, pour chaque combinaison de jeu de données Source (S) et Cible (C).

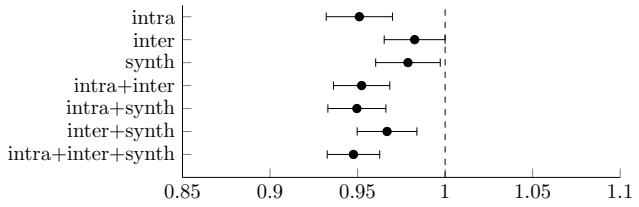
la grande inter variabilité de la population diabétique. Pourtant, certains scénarios de transfert global réussissent à égaler les performances du modèle FCN #2 (i.e., IO→I, IOT→I, O→O, IO→O, IOT→O). Cela montre, une fois de plus, que les modèles personnalisés ne possèdent pas assez de données d'entraînement.

Transfert adverse global

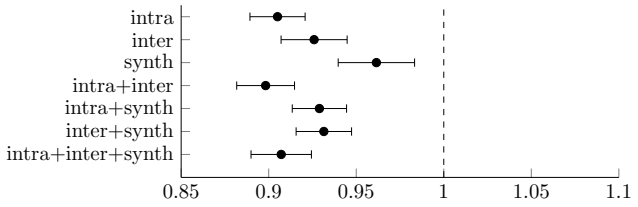
L'utilisation du module adverse sur les modèles globaux permet d'améliorer généralement ces premiers résultats. Elle permet en particulier de faire fonctionner le transfert I→I, ce que le transfert standard ne permettait pas. De plus, le scénario *inter+exter* (IO→I ou IO→O) se démarque en ayant de meilleurs résultats que FCN #2, ainsi que des résultats équivalents au modèle SVR. Toutefois, nous pouvons remarquer que l'efficacité du transfert global dépend beaucoup de la provenance des patients sources et du patient cible. Si aucun patient *intra* n'est utilisé comme source, alors le transfert n'est que peu efficace. En particulier, le transfert global n'utilisant que des données synthétiques ne fonctionne pas.

Transfert standard affiné

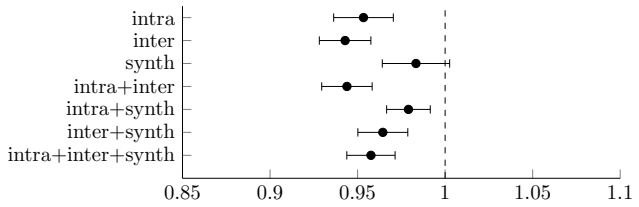
En comparaison avec les résultats du transfert standard global, ceux du transfert standard affiné montrent de nettes améliorations pour tous les scénarios de transfert. De plus ces performances sont significativement supérieures à celles obtenues par notre modèle de référence FCN #2 (voir Figure 6.5a). En personnalisant le modèle au patient cible, nous permettons en outre au transfert depuis des données synthétiques de fonctionner. Globalement, les meilleurs scénarios de transfert sont ceux qui utilisent des données *intra* avec ou sans données *inter* ou *synth*. Ces résultats montrent l'intérêt de procéder au transfert de connaissances entre patients pour la prédiction de glycémie et que ce transfert peut fonctionner même lorsque les données ont une nature différente (type 1 et type 2,



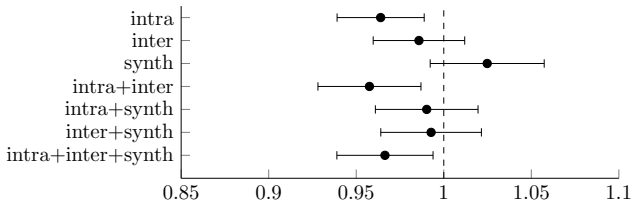
(a) Transfert Standard contre FCN #2



(b) Transfert Adverse contre FCN #2



(c) Transfert Adverse contre Transfer Standard



(d) Transfert Adverse contre SVR

FIGURE 6.5: Intervalles de confiance à 99% des performances appairées en acsmape d'un modèle par rapport à un autre pour tous les groupes de scénario de transfert possible.

différents capteurs de glycémie). Nous notons tout de même que certains résultats du transfert adverse non affiné sont meilleurs ou équivalents que ceux du transfert standard affiné (e.g., I→I, IO→O), suggérant une potentielle utilisation de modèles de prédiction de glycémie globaux en utilisant la méthodologie adverse.

Transfert adverse affiné

Les résultats du transfert standard affiné sont eux-mêmes améliorés par le transfert adverse affiné, et ce pour la quasi-totalité des scénarios de transfert (certains scénarios, comme IT ou OT→I, montrant des résultats équivalents). Comme nous le montre la Figure 6.5c, ces améliorations liées à l'utilisation de la méthodologie adverse sont statistiquement significatives et sont principalement efficaces pour l'utilisation de données qui ne sont pas *intra*. Bien que l'intervalle de confiance du scénario *synth* ne nous permet pas de conclure à travers cette Figure, nous pouvons tout de même voir l'amélioration de la significativité statistique en comparant avec les résultats de FCN #2 avec les Figures 6.5a et 6.5b. Enfin, la Figure 6.5d indique que ces résultats sont significativement meilleurs que le modèle SVR de référence pour les scénarios de transfert *intra*, *intra+inter* et *intra+inter+synth*.

Analyses des scénarios de transfert

Bien que le transfert, qu'il soit adverse ou non, fonctionne pour tous les scénarios étudiés, son efficacité est variable. Au sein des transferts utilisant un unique jeu de données source (*intra*, *inter* ou *synth*), la configuration présentant les meilleures performances est, sans surprise, le transfert *intra*. L'ajout de nouvelles données sources (par exemple : *intra+inter* contre *intra*, ou *inter+synth* contre *inter*) ne parvient pas toujours à améliorer les résultats. Tandis que la combinaison de plusieurs jeux sources a l'air d'être profitable dans le cadre d'un transfert vers le jeu OhioT1DM, les meilleurs résultats pour un transfert vers IDIAB restent ceux du scénario *intra*. Lorsque les

données du jeu IDIAB sont combinées avec les données d'un autre dataset, les données IDIAB deviennent sous-représentées. Nous supposons que cet ajout de données supplémentaire n'est pas efficace, car le modèle entraîné sur ces données sources combinées est davantage spécialisé au domaine qui est le plus représenté, et donc plus IDIAB. Nous notons que cette spécialisation se fait malgré la pondération des domaines inversement proportionnelle à leur représentation dans le jeu.

Acceptabilité clinique des résultats

Le Tableau 6.3 nous permet de nous intéresser à l'acceptabilité clinique des résultats. Afin de simplifier la lecture de ceux-ci, nous avons choisi de ne représenter seulement l'acceptabilité clinique des modèles affinés présentant la meilleure acsmape dans le Tableau 6.2. Sans étonnement, les résultats obtenus par l'apprentissage par transfert adverse montrent une nette amélioration en comparaison avec les modèles de référence FCN #1 et #2 ainsi qu'avec un apprentissage par transfert standard. Seul le modèle SVR semble rester compétitif du point de vue de l'acceptabilité clinique. En effet, il possède des taux de prédictions dans les zones dangereuses (E, D) particulièrement bas, en particulier pour le jeu de données IDIAB.

Synthèse des résultats

Nous pouvons ainsi conclure sur l'importance de l'apprentissage par transfert, et en particulier de celle de l'apprentissage par transfert adverse pour l'entraînement de modèles de prédiction de glycémie basés sur l'apprentissage profond. Bien que ce transfert est efficace avec des données provenant d'origines différentes (données réelles ou simulées, données de diabétiques de type 1 ou 2, ou provenant de conditions expérimentales diverses), il est préférable d'utiliser des données *intra*, données qui peuvent être augmentées au besoin avec des données *inter*. Bien que le scénario de transfert *synth* soit le moins performant ici, il reste intéressant dans une situation d'absence totale ou quasi totale de données réelles.

6.4.3 Analyse du comportement du réseau adverse

Après avoir analysé les performances des différents modèles et types de transfert, nous proposons de nous intéresser au comportement de l'apprentissage des connaissances apprises sur les patients sources. En particulier, nous cherchons à comprendre plus en détail comment le caractère adverse parvient à rendre le transfert vers le patient plus efficace.

Pour cela, nous pouvons analyser les caractéristiques en sortie du module d'extraction de caractéristiques (voir la Figure 6.3). La Figure 6.6 montre la visualisation par t-SNE de ces caractéristiques dans le cadre d'un transfert standard ou adverse et en utilisant un seul ou plusieurs jeux de données. La visualisation t-SNE effectue une projection à 2 dimensions de l'espace des caractéristiques en accentuant les relations de distance entre caractéristiques :

Jeu de données	Modèle	*Point Error-Grid Analysis* (P-EGA)					
		A+B	**A**	**B**	**C**	**D**	**E**
IDIAB (I)	SVR	**99.18 (0.43)**	94.80 (1.49)	4.38 (1.33)	0.04 (0.10)	**0.78 (0.49)**	**0.00 (0.00)**
	FCN #1	98.43 (1.58)	92.12 (2.58)	6.32 (1.62)	**0.00 (0.00)**	1.57 (1.58)	**0.00 (0.00)**
	FCN #2	98.21 (1.71)	92.59 (3.37)	5.61 (2.09)	**0.00 (0.00)**	1.79 (1.71)	**0.00 (0.00)**
	TL *	98.57 (1.18)	93.73 (2.68)	4.84 (1.81)	**0.00 (0.00)**	1.43 (1.18)	**0.00 (0.00)**
	ATL †	98.77 (1.30)	**94.96 (2.59)**	**3.81 (1.67)**	**0.00 (0.00)**	1.23 (1.30)	**0.00 (0.00)**
OhioT1DM (O)	SVR	99.10 (0.88)	93.48 (3.14)	5.62 (2.33)	0.01 (0.02)	0.88 (0.87)	0.01 (0.03)
	FCN #1	98.67 (1.18)	91.67 (3.47)	7.01 (2.38)	**0.00 (0.00)**	1.33 (1.18)	**0.00 (0.00)**
	FCN #2	98.67 (1.07)	92.71 (3.32)	5.96 (2.29)	**0.00 (0.00)**	1.33 (1.07)	**0.00 (0.00)**
	TL **	98.89 (1.13)	93.69 (3.63)	5.20 (2.52)	0.00 (0.01)	1.11 (1.12)	0.00 (0.01)
	ATL ‡	**99.20 (0.76)**	**94.44 (2.83)**	**4.77 (2.12)**	**0.00 (0.00)**	**0.80 (0.76)**	**0.00 (0.00)**

scénarios : * IO→I, ** IOT→I, † I→I, ‡ IO→O

Tableau 6.3: Acceptabilité clinique (P-EGA) moyenne (avec écart type) des modèles de référence ainsi que des meilleurs modèles affinés après transfert.

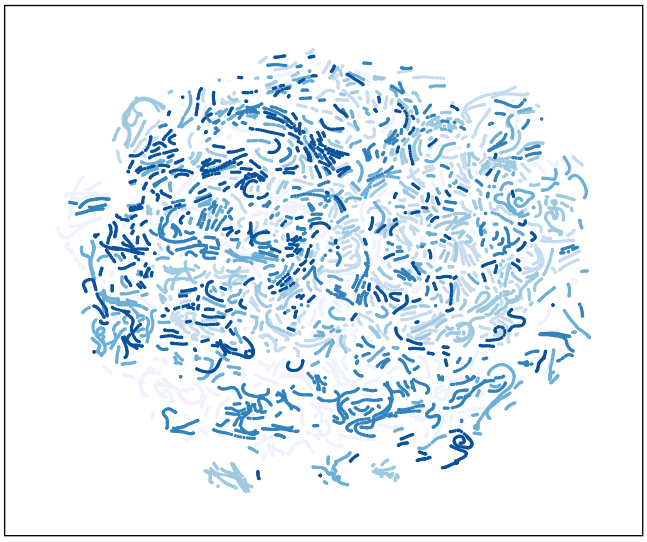
tandis que des caractéristiques proches dans leur espace d'origine sont regroupées dans une même région de l'espace 2D, des caractéristiques éloignées sont davantage écartées les unes des autres.

Au sein de la Figure 6.6a, nous pouvons remarquer que les caractéristiques d'un même patient (d'une même couleur) sont souvent regroupées en clusters. Cela montre que le modèle parvient à discriminer les patients les uns des autres en apprenant des caractéristiques très spécifiques au patient qu'il identifie. À l'inverse, la représentation t-SNE des caractéristiques issues d'un transfert adverse de la Figure 6.6b est beaucoup plus diffuse indiquant des caractéristiques plus générales au sein des patients sources.

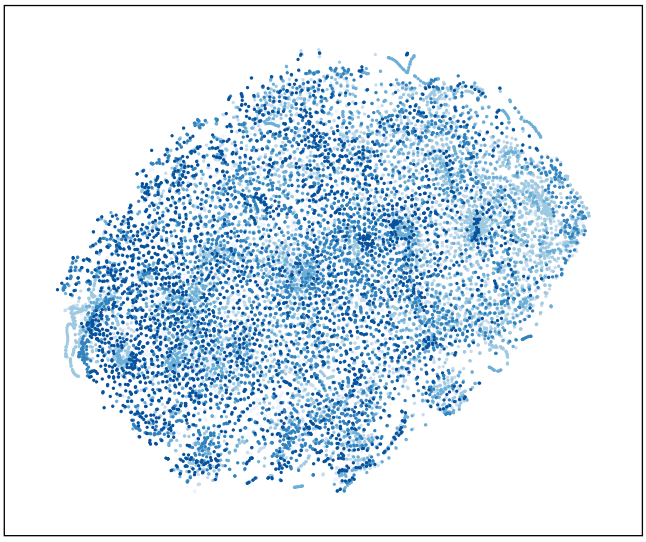
Quant à eux, les transferts avec plusieurs jeux de données sources à travers les Figures 6.6c et 6.6d, montrent le même comportement qui est, cette fois-ci, plus prononcé. Alors que, dans le cadre d'un transfert standard, les trois jeux de données occupent des régions très délimitées dans l'espace 2D, ils deviennent entrelacés en utilisant le transfert adverse. Cela montre que les caractéristiques extraites sont bien plus générales à la fois aux patients d'un même jeu de données, mais aussi plus générales aux jeux de données utilisés. En étant ainsi plus générales dans l'ensemble, ces caractéristiques sont donc plus facilement transférables et personnalisables à un nouveau patient.

Afin de comparer la généralisation des caractéristiques extraites avec les deux modes de transfert et de quantifier, sur les caractéristiques brutes, les observations faites sur les représentations t-SNE, nous proposons une nouvelle métrique que nous nommons perplexité de domaine locale, *Local Domain Perplexity* (LDP). La LDP mesure, entre 0 et 1, l'uniformité moyenne de la distribution de domaine (ou classe) des caractéristiques dans leur proche voisinage. Tandis qu'une LDP élevée indique que les caractéristiques se généralisent à l'ensemble des domaines étudiés, une LDP faible indique que les caractéristiques sont très spécialisées aux domaines. L'Algorithme 2 donne le pseudo-code du calcul de la LDP :

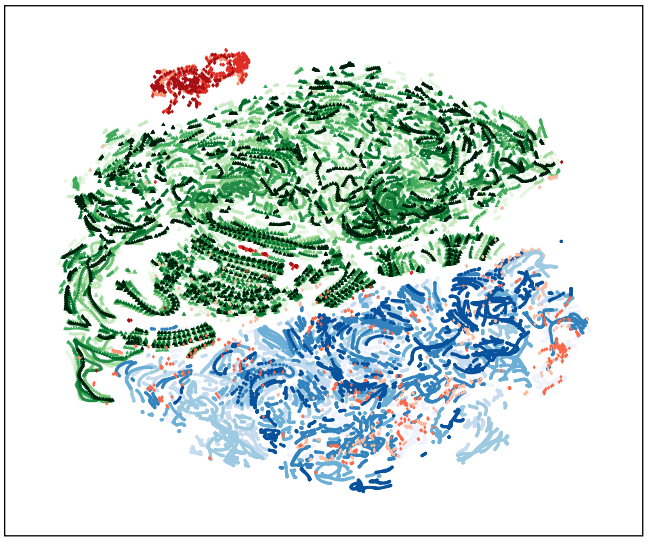
1. La distance de chaque échantillon de caractéristiques aux autres échantillons de caractéristiques est calculée ;



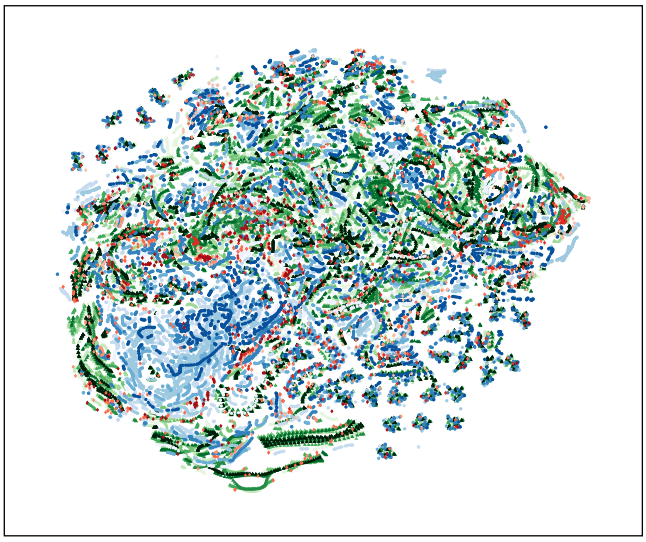
(a) O→I : Transfert Standard



(b) O→I : Transfert Adverse



(c) IOT→I : Transfert Standard



(d) IOT→I : Transfert Adverse

FIGURE 6.6: Visualisation t-SNE des caractéristiques pour les scénarios de transfert O→I (en haut) IOT→I (en bas) et pour un type de transfert standard (à gauche) et adverse (à droite). Les caractéristiques représentées sont celles des patients sources, le patient IDIAB #1 ayant été choisi arbitrairement comme patient cible. La représentation en deux dimensions des caractéristiques a été obtenue par t-SNE en réduisant d'abord la dimension à 50 par analyse par composantes principales (ACP). Chaque couleur représente un jeu de données (rouge pour le jeu IDIAB, bleu pour OhioT1DM et vert pour T1DMS), et chaque nuance d'une couleur représente un patient unique appartenant au jeu de données. Chacune de ces nuances de couleurs a été choisie grâce aux travaux de Cynthia A. Brewer sur la création de palettes de couleurs pour la cartographie [12].

2. Les N plus proches échantillons dans l'espace des caractéristiques sont considérés comme les voisins, $voisins_i$, $i \in [1, N]$, de l'échantillon considéré (dans cette étude, $N = 50$) ;
3. Pour chaque échantillon, la distribution de probabilité des voisins est calculée : $P(d) = 1/N \cdot \sum_1^N \text{count}(voisins_i = d)$;
4. La LDP est calculée comme la perplexité, $2^{\sum_d P(d) \cdot \log_2(P(d))}$, normalisée entre 0 et 1 et moyennée sur l'ensemble des échantillons.

La perplexité est calculée comme l'exponentiation de l'entropie (voir Équation 6.2, où $H(p)$ est l'entropie de la distribution de probabilités p et d est le domaine) et est très largement utilisée dans la recherche en traitement automatique des langues pour mesurer la distribution des mots prédits au sein des phrases [66]. Tandis que dans le traitement automatique des langues, nous cherchons à minimiser la perplexité, ici nous cherchons à la maximiser : plus la perplexité est grande, moins il est possible de prédire le domaine à partir des caractéristiques extraites.

$$Perplexité = 2^{H(p)} = 2^{\sum_d p(d) \log_2(p(d))} \tag{6.2}$$

Algorithme 2 : Calcul de la métrique LDP

Données : caractéristiques C et domaine de N échantillons, taille du voisinage V, nombre de domaines D
Résultat : Perplexité de Domaine Locale (LDP)

pour $i = 1 \rightarrow N$ **faire**
  $C_i \leftarrow$ caractéristiques de l'échantillon i;
  $voisins_i \leftarrow$ les V échantillons qui minimisent la distance Euclidienne avec les caractéristiques de l'échantillon C_i;
  $P(d), d \in [1..D] \leftarrow$ probabilité de distribution des domaines dans le voisinage de l'échantillon i;
  $LDP_i \leftarrow 1/D \cdot 2^{\sum_d P(d) \cdot \log_2(P(d))}$;
$LDP \leftarrow 1/S \cdot \sum_{i=1}^{N} LDP_i$;
retourner LDP;

La Tableau 6.4 fournit les mesures LDP pour chaque scénario de transfert des modèles globaux. L'utilisation de la méthodologie adverse montre une augmentation de la LDP pour tous les scénarios de transfert. Les améliorations sont les plus fortes pour les scénarios *inter*, *intra+inter* et *intra+inter+synth* et les plus faibles pour le scénario *synth*, conformément aux améliorations de précision visibles sur la la Figure 6.5c.

6.5 Conclusion

Dans cette étude, nous avons analysé l'utilisation novatrice de l'apprentissage par transfert pour la prédiction de la glycémie dans le but de combattre le manque de données d'apprentissage utilisées par les modèles prédictifs. Cette situation est causée par la difficulté de récolter des données biomédicales sensibles en grandes quantités et par le besoin d'avoir des modèles prédictifs personnalisés au patient pour tenir compte de la grande variabilité de la population diabétique.

Notre hypothèse de départ a été qu'il est possible d'apprendre des connaissances a priori sur plusieurs patients diabétiques et de transférer ces connaissances pour faciliter l'apprentissage des modèles sur de nouveaux patients. Afin de favoriser la généralisation des connaissances a priori apprises sur les patients dits *sources*, nous avons proposé la méthodologie d'apprentissage par transfert multi-sources adverse. Inspirée des travaux de Ganin *et al.* [57], la méthodologie utilise un module apprenant à discriminer les patients à partir des connaissances apprises par

Jeu de données		Transfert Standard (global)	Transfert Adverse (global)	Variation relative (en %)
S	**C**			
I	I	0.72 (0.04)	**0.79 (0.01)**	+11.06 (6.58)
O		0.50 (0.02)	**0.68 (0.01)**	+36.93 (5.20)
T		0.72 (0.01)	**0.73 (0.01)**	+1.15 (2.09)
IO		0.34 (0.01)	**0.53 (0.01)**	+54.28 (7.00)
IT		0.47 (0.00)	**0.51 (0.01)**	+10.29 (2.58)
OT		0.38 (0.00)	**0.46 (0.02)**	+21.53 (5.70)
IOT		0.29 (0.00)	**0.40 (0.01)**	+38.99 (3.03)
I	O	0.67 (0.01)	**0.78 (0.01)**	+15.48 (1.71)
O		0.67 (0.04)	**0.71 (0.02)**	+7.26 (8.03)
T		0.72 (0.00)	**0.72 (0.01)**	+0.28 (2.01)
IO		0.35 (0.02)	**0.52 (0.01)**	+47.51 (7.02)
IT		0.43 (0.00)	**0.49 (0.01)**	+13.84 (3.10)
OT		0.41 (0.00)	**0.49 (0.03)**	+21.80 (7.86)
IOT		0.29 (0.01)	**0.40 (0.01)**	+35.63 (4.92)

Tableau 6.4: Perplexité de Domaine Locale moyenne (avec écart type) des modèles globaux avant affinage, pour chaque combinaison de jeu de données Source (S) et Cible (C).

le modèle. Pendant l'apprentissage, ce module est en mis en compétition avec le module d'extraction de caractéristiques qui apprend à extraire des caractéristiques à la fois utiles pour prédire la glycémie, mais aussi qui ne permettent pas de discriminer les patients les uns des autres. Notre seconde hypothèse est que ces connaissances sont ainsi plus générales et se transfèrent mieux à un nouveau patient.

Dans cette thèse, nous avons à notre disposition trois jeux de données qui sont différents tant bien dans la nature des patients (type 1 ou type 2, patients virtuels), que dans la nature des signaux (systèmes expérimentaux différents). Afin de rendre les modèles interopérables et favoriser le transfert, nous avons uniformisé la représentation des données au sein des différents jeux. La diversité des jeux de données nous a donné l'opportunité d'explorer la qualité du transfert en fonction des jeux sources et cibles utilisés. Nous avons ainsi envisagé plusieurs scénarios de transferts possibles : le transfert *intra*, lorsque les patients sources proviennent du même jeu que le patient cible, le transfert *inter* lorsque les patients sources proviennent d'un jeu réel, mais différent de celui du patient cible, le transfert *synth* se faisant à partir des patients virtuels du jeu T1DMS, et enfin toutes les combinaisons possibles de ces différents scénarios.

L'étude s'est faite à la lueur de réseaux entièrement convolutifs, modèles utilisés à l'origine dans les travaux de Ganin *et al.* dont l'apprentissage par transfert multi-sources adverse tire son inspiration. Originellement conçus pour la reconnaissance d'images, les réseaux convolutifs sont applicables aux séries temporelles qui peuvent être vues comme des images à une unique dimension. Dans un premier temps, un premier modèle général (implémentant le module adverse ou non) est appris sur les patients sources. Le transfert s'effectue en personnalisant ce modèle général au patient cible en l'affinant avec ses données.

Dans un premier temps, nous avons présenté de premiers résultats en utilisant directement les modèles gé-

néraux non affinés sur les patients cibles. Les performances des modèles globaux, non affinés, et évalués sur les patients cibles ont montré qu'elles dépendent grandement de la source utilisée et requièrent principalement des données *intra* pour être bonnes. Cette efficacité est améliorée par la méthodologie adverse, dont les meilleurs résultats parviennent à égaler, voire surpasser, les résultats de référence.

En affinant les modèles globaux aux patient cibles, nous observons une grande amélioration des performances soulignant la nécessité d'avoir des modèles personnalisés au patient. Contrairement aux modèles globaux, ces performances sont meilleures que les modèles convolutifs de référence pour chaque scénario validant l'hypothèse du transfert de connaissance pour améliorer la prédiction de la glycémie de personnes diabétiques. Les résultats montrent, en outre, qu'il est possible de transférer des connaissances entre des personnes diabétiques de type 1 ou de type 2, depuis des personnes diabétiques virtuelles et entre conditions expérimentales différentes. De plus, ce transfert est significativement amélioré par l'utilisation de la méthodologie d'apprentissage adverse pour tous les scénarios de transfert. En particulier il permet, dans le cadre d'un transfert utilisant des données *intra* de surpasser significativement le modèle SVR qui est le meilleur modèle de référence identifié dans l'étude benchmark du Chapitre 4.

Enfin, nous avons analysé le comportement de l'apprentissage adverse à travers l'observation des caractéristiques apprises par les modèles globaux. À cet effet, nous avons proposé l'utilisation d'une nouvelle métrique, la Perplexité de Domaine Locale (LDP), qui quantifie l'uniformité moyenne de la distribution des caractéristiques des différentes sources dans leur voisinage proche. Appuyés par une visualisation des caractéristiques via la méthode t-SNE, les résultats confirment que l'apprentissage adverse permet bien d'apprendre des connaissances se généralisant mieux aux différentes sources, ce qui facilite l'affinage des modèles sur une cible nouvelle.

Dans l'ensemble, cette étude montre que l'apprentissage par transfert, et en particulier sa variante multi-sources et adverse, est très prometteur pour concevoir des modèles de prédictions de glycémie à la fois précis et sûr pour le patient. Dans la continuité de ces travaux, nous pouvons entrevoir de nombreuses pistes d'améliorations. Premièrement, l'étude se focalise sur les réseaux de neurones convolutifs, or cette thèse montre qu'ils existent de nombreux modèles basés sur l'apprentissage profond qui pourraient, eux aussi, tirer parti du transfert de connaissances (comme les réseaux récurrents LSTM pour lesquels nous utilisons la méthodologie de transfert multi-sources adverse dans le Chapitre 7). Par ailleurs, il serait possible de travailler sur la qualité et la sélection des échantillons de données utilisés dans l'apprentissage des modèles globaux sur les patients sources. En particulier, dans ce but précis, le simulateur T1DMS pourrait être spécifiquement utilisé pour générer des données particulièrement riches et diverses.

7 | Interprétabilité des modèles prédictifs

Sommaire

7.1 Introduction

L'un des freins majeurs à l'adoption de l'apprentissage profond dans le domaine médical est le manque d'interprétabilité des modèles, souvent caractérisés comme des « boîtes noires » [16]. Ce besoin en interprétabilité peut s'expliquer par plusieurs raisons distinctes [1]. La raison principale est sans doute le besoin de comprendre les décisions pour pouvoir leur faire confiance. Ce besoin est accentué lorsque les prédictions sont inattendues et

mettent en jeu la vie du patient. En effet, bien que statistiquement vérifiée, une décision peut être le fruit d'un biais du modèle sur les données d'entraînement. Par exemple, en entraînant un modèle prédisant la probabilité de décès de patients par pneumonie, Ba *et al.* ont montré que leur modèle associait les patients asthmatiques à une faible probabilité de décès [7]. Cette association erronée faite par le modèle prend sa source dans un biais contenu dans les données d'entraînement. En effet, les patients asthmatiques sont souvent traités en priorité dans les hôpitaux, résultant en une faible mortalité statistique. La seconde raison derrière le besoin d'avoir des modèles interprétables est que ces derniers peuvent permettre d'améliorer nos connaissances générales des pathologies.

De manière générale, les modèles interprétables, comme la régression linéaire ou les arbres de décision, sont peu performants en comparaison avec des modèles plus complexes. Cette complexité accrue, associée à un gain en performances, entraîne bien souvent une grande baisse en interprétabilité (e.g., réseaux de neurones profonds, forêts aléatoires). Ainsi, de nombreux efforts ont été faits ces dernières années afin de tenter d'interpréter les modèles complexes, et en particulier les modèles profonds. Parmi ces efforts, nous pouvons dégager deux axes d'études. La première approche vise à mesurer, visualiser, l'importance des données d'entrées sur les prédictions. Par exemple, Simonyan *et al.* ont proposé la construction d'une carte de saillance permettant d'identifier les pixels importants pour la classification d'images [137]. Ces cartes de saillances ont été utilisées par Ma *et al.* pour analyser la nature d'attaques adverses (*adversarial attacks*) sur un réseau de neurones convolutifs entraîné sur des images médicales [104]. Dans leurs travaux, Lundberg *et al.* proposent un framework visant à mesurer l'importance de chaque descripteur d'entrée sur les prédictions [102], avec son application pour la prévention d'hypoxémies pendant les chirurgies [103].

Plutôt que de proposer des méthodes pour interpréter les boîtes noires, un grand nombre de chercheurs se penchent sur des modifications d'architecture rendant les modèles profonds directement plus interprétables. Parmi ces nouvelles architectures, les plus notables reposent sur le principe novateur d'attention. Celui-ci a été introduit par Bahdanau *et al.* dans le domaine de la traduction automatique [8] et a été repris dans l'architecture Transformer [149]. Les modèles basés sur l'architecture Transformer sont aujourd'hui les modèles obtenant les meilleurs résultats pour toutes tâches relevant du traitement automatique des langues. Construit pour des modèles utilisant des données séquentielles, le principe d'attention permet au modèle de se concentrer sur une ou plusieurs parties de la séquence afin de faire sa prédiction. Permettant d'obtenir des performances égales, voire supérieures dans certains domaines, l'attention que porte le modèle sur les différentes étapes temporelles lui permet d'améliorer son interprétabilité. Le principe d'attention général a eu beaucoup de dérivés, comme celle de l'attention multi-têtes du modèle Transformer ou de celle à deux niveaux du modèle RETAIN. Ce dernier a été proposé par Choi *et al.* afin de traiter et d'analyser les dossiers électroniques de santé [17]. Son attention temporelle couplée à son attention à la variable lui permet de quantifier directement la contribution de chaque variable, pour chaque instant, à la prédiction finale.

Dans ce chapitre, nous explorons l'utilisation de l'architecture RETAIN et du principe d'attention pour la prédic-

tion de la glycémie future de personnes diabétiques. Pouvoir expliquer les prédictions faites par le modèle présente plusieurs intérêts. Premièrement, cela permet au patient de prendre des décisions plus éclairées suite aux prédictions du modèle (e.g., sur quoi se base le modèle pour prédire cette hypoglycémie future ?). Aussi cela peut permettre, à lui ainsi qu'à son médecin, de mieux comprendre sa maladie et ses spécificités individuelles. Enfin, cela peut aider le scientifique dans la construction du modèle, que ce soit dans son architecture ou dans la nature des données utilisées. Afin d'utiliser l'architecture RETAIN pour la prédiction de la glycémie de personnes diabétiques, nous avons apporté plusieurs modifications pour la rendre compatible avec des tâches de régression. En outre, nous proposons une nouvelle mesure de la contribution des variables aux prédictions. En comparaison avec la mesure de la contribution proposée par les auteurs, celle-ci facilite l'analyse statistique et visuelle de la contribution des variables. Enfin, nous démontrons l'intérêt de la métrique et de l'architecture RETAIN plus généralement par l'analyse graphique de la mesure de la contribution des variables dans la prédiction de la glycémie.

Ce chapitre est organisé comme suit. Tout d'abord, nous présentons le principe d'attention, l'architecture RETAIN ainsi que le processus d'interprétation de ses prédictions. Après avoir détaillé la méthodologie générale qui a été suivie dans cette étude, nous présentons les résultats expérimentaux des modèles étudiés. Enfin, nous démontrons empiriquement l'intérêt de l'architecture RETAIN à travers divers outils de visualisation pour la prédiction de glycémie.

7.2 Architecture RETAIN et principe d'attention

Dans cette section, nous introduisons le principe d'attention générale pour les tâches de régression. Puis, nous développons l'architecture RETAIN ainsi que les modifications que nous avons apportées. Enfin, nous expliquons comment le modèle RETAIN mesure la contribution de chaque variable, à chaque instant, sur les prédictions.

7.2.1 Principe d'attention

Avant de décrire l'architecture RETAIN, nous proposons de poser les fondements du principe d'attention appliqué à des tâches de régression. Celui-ci a initialement été proposé dans le cadre de la traduction automatique [8]. Cette tâche se caractérise par l'utilisation d'architectures séquentielles à plusieurs entrées et sorties, les données d'entrée et de sortie étant représentées par des vecteurs de mots formant des phrases (e.g., traduction d'une phrase en français vers une phrase en anglais). De leur côté, les tâches de régression ne possèdent généralement qu'une seule sortie, résultant en une simplification de l'architecture mettant en œuvre le principe d'attention [127].

Dans les paragraphes qui suivent, à l'aide de la Figure 7.1, nous procédons à la description de l'architecture d'un réseau de neurones récurrents implémentant le mécanisme d'attention standard. Ce modèle cherche à prédire la valeur y_{t+PH} à partir des données d'entrée $\boldsymbol{x}_{t-H}, ..., \boldsymbol{x}_t$, où $\boldsymbol{x}_i \in \mathcal{R}^r$, $i \in [t-H, t]$ et H représente la longueur de la séquence d'entrée (représentant la longueur de l'historique dans les travaux de prédiction de glycémie).

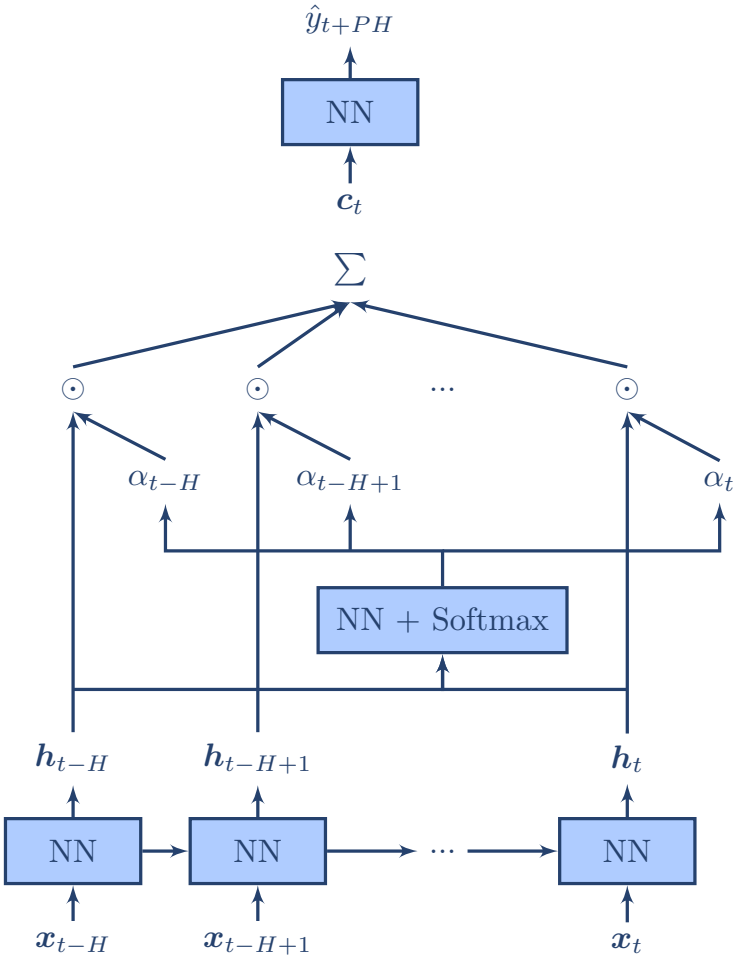


FIGURE 7.1: Réseau de neurones récurrents utilisant le principe d'attention standard.

Tout d'abord, selon l'Équation 7.1, un réseau de neurones récurrents RNN transforme les données d'entrée en représentations cachées $\boldsymbol{h}_i \in \mathcal{R}^p$, où p est le nombre de neurones (ou d'unités LSTM) du réseau de neurones récurrents.

$$\boldsymbol{h}_{t-H}, ..., \boldsymbol{h}_t = \text{RNN}(\boldsymbol{x}_{t-H}, ..., \boldsymbol{x}_t) \tag{7.1}$$

À partir des représentations cachées $\boldsymbol{h}_i$, les poids d'attention $\alpha_i \in \mathcal{R}$ peuvent être calculés suivant l'Équation 7.2. L'Équation 7.2a met en œuvre une couche dense de neurones (poids $\boldsymbol{w}_\alpha \in \mathcal{R}^p$ et biais $b_\alpha \in \mathcal{R}$) pour calculer l'attention relative $e_i \in \mathcal{R}$ de chaque représentation cachée. Ces poids d'attention sont ensuite normalisés en α_i à travers l'opération Softmax décrite par l'Équation 7.2b. Cette normalisation permet de garantir des poids d'attention

entre 0 et 1 et dont la somme fait 1.

$$e_i = \boldsymbol{w}_\alpha^T \boldsymbol{h}_i + b \tag{7.2a}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=t-H}^{t} \exp(e_j)} \tag{7.2b}$$

Puis, selon l'Équation 7.3, le vecteur de contexte $\boldsymbol{c}_t \in \mathcal{R}^p$ est calculé comme la somme des représentations cachées $\boldsymbol{h}_i$ pondérées par leur poids d'attention respectif α_i.

$$\boldsymbol{c}_t = \sum_{i=t-H}^{t} \alpha_i \boldsymbol{h}_i \tag{7.3}$$

Enfin, la prédiction du modèle peut être calculée par une couche dense de neurones (poids $\boldsymbol{w}_{out} \in \mathcal{R}^p$ et biais $b_{out} \in \mathcal{R}$) prenant en entrée le vecteur de contexte $\boldsymbol{c}_t$ selon l'Équation 7.4.

$$\hat{y}_{t+PH} = \boldsymbol{w}_{out}^T \boldsymbol{c}_t + b_{out} \tag{7.4}$$

En comparaison avec une architecture basée sur un réseau de neurones récurrents classique, cette architecture pondère les représentations cachées $\boldsymbol{h}_i$ des poids d'attention α_i. Cela permet d'inciter la dernière couche cachée à hiérarchiser les instants temporels en fonction de leur importance. Il est possible d'analyser les poids d'attention afin d'identifier les instants temporels importants dans le mécanisme de prédiction. Cette particularité permet au modèle attentionnel d'être plus interprétable qu'un modèle standard.

7.2.2 Architecture RETAIN

Bien que l'architecture standard à base d'attention permette une certaine interprétabilité des prédictions, celle-ci reste limitée. En effet, il n'est pas possible d'évaluer l'importance des variables d'entrée au modèle au sein d'un instant temporel précis. Cela est dû au fait que la représentation cachée est calculée par le réseau de neurones récurrents, non-interprétable à cause de sa structure non-linéaire. Pour palier cette limitation du mécanisme d'attention standard, Choi *et al.* ont proposé l'architecture RETAIN [17]. Celle-ci sépare le calcul de l'attention de celui de la représentation cachée. Tandis que le calcul des poids d'attention se fait avec un réseau de neurones récurrents, celui de la représentation cachée se fait au moyen d'une couche dense linéaire. De plus, l'architecture RETAIN se voit ajouter un second réseau de neurones récurrents permettant de calculer un second niveau d'attention. Cette nouvelle attention se fait à la variable, permettant ainsi au modèle de se focaliser sur certaines variables d'entrée au sein d'un instant temporel précis. Une fois les poids d'attention déterminés, le calcul des prédictions se fait de manière linéaire. Cela permet de remonter à la contribution à la prédiction des variables d'entrée à chaque instant.

Le modèle RETAIN n'en reste pas moins un modèle non linéaire grâce au calcul des poids d'attention se faisant de manière non linéaire à travers les réseaux de neurones récurrents. La mesure de la contribution de chaque variable, à chaque instant temporel, rend l'architecture RETAIN bien plus interprétable que celle implémentant le principe d'attention standard.

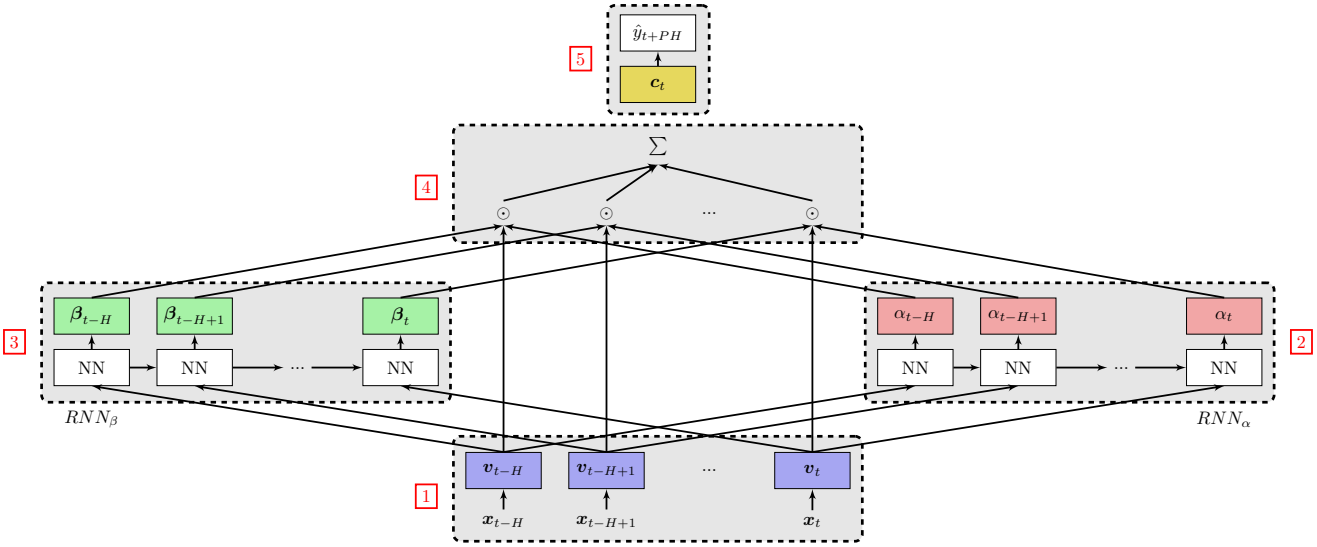


FIGURE 7.2: Représentation graphique du modèle RETAIN. **Étape 1** : Les signaux d'entrée sont transformés en représentations cachées. **Étape 2** : les poids liés à l'attention temporelle sont calculés à partir des caractéristiques. **Étape 3** : les poids liés à l'attention à la variable sont aussi calculés à partir des caractéristiques. **Étape 4** : Le vecteur de contexte est calculé à partir des caractéristiques pondérées des poids d'attention. **Étape 5** : La prédiction est faite à partir du vecteur de contexte.

Les prédictions du modèle RETAIN sont faites en 5 étapes dont la Figure 7.2 en donne la représentation graphique. Comme précédemment, $\boldsymbol{x}_t \in \mathcal{R}^r$ représente le vecteur des r variables d'entrées à l'instant t. L'ensemble des données d'entrées sont représentées par $\boldsymbol{x}_{t-H}, ..., \boldsymbol{x}_t$ où H correspond à la longueur de l'historique connu par le modèle.

Étape 1 : Tout d'abord, pour chaque instant temporel $i, i \in [t, t-H]$, des représentation cachées [1] ou caractéristiques $\boldsymbol{v}_i \in \mathcal{R}^m$ sont calculée par l'opération linéaire décrite par l' Étape 1 à partir des données d'entrées. Tandis que m représente la taille des représentations cachées, $\boldsymbol{W}_{emb} \in \mathcal{R}^{m \times r}$ est la matrice permettant leur calcul.

$$\boldsymbol{v}_i = \boldsymbol{W}_{emb}\boldsymbol{x}_i \qquad \text{(Étape 1)}$$

Étape 2 : Ces caractéristiques sont données en entrée à un premier réseau de neurones récurrents RNN_α à p neurones (Étape 2.1, avec $g_i \in \mathcal{R}^p$), suivi d'une couche linéaire (Étape 2.2, avec $\boldsymbol{w}_\alpha \in \mathcal{R}^p$ et $b_\alpha \in \mathcal{R}$) et d'une normalisation par Softmax (Étape 2.3) pour calculer les poids d'attention temporelle $\alpha_i \in \mathcal{R}$. Ces poids représentent la pondération (positive, entre 0 et 1) que le modèle va donner à chaque instant i de l'historique des représentations

1. Dans le papier original du modèle RETAIN, ces représentations cachées sont référées sous le nom de *plongements*, en référence au domaine du traitement automatique des langues qui utilise le terme de *plongements lexicaux*, ou *word embeddings* en anglais.

cachées des variables d'entrée. Plus la pondération est importante, plus l'instant en question sera pris en compte dans le calcul final de la prédiction.

$$\boldsymbol{g}_{t-H}, ..., \boldsymbol{g}_t = \text{RNN}_\alpha(\boldsymbol{v}_{t-H}, ..., \boldsymbol{v}_t) \quad \text{(Étape 2.1)}$$

$$e_i = \boldsymbol{w}_\alpha^T \boldsymbol{g}_i + b_\alpha \quad \text{(Étape 2.2)}$$

$$\alpha_{t-H}, ..., \alpha_t = \text{Softmax}(e_{t-H}, ..., e_t) \quad \text{(Étape 2.3)}$$

Étape 3 : Simultanément, les caractéristiques extraites à l'Étape 1 sont aussi données en entrée à un second réseau de neurones récurrents RNN_β à q neurones (Étape 3.1). Sa sortie, $\boldsymbol{h}_i \in \mathcal{R}^q$, sert à calculer les poids d'attention à la variable $\boldsymbol{\beta}_i \in \mathcal{R}^m$ (Étape 3.2, avec $\boldsymbol{W}_\beta \in \mathcal{R}^{m \times q}$ et $\boldsymbol{b}_\beta \in \mathcal{R}^m$). L'utilisation de la fonction d'activation $\tanh$ permet de pondérer positivement et négativement, entre -1 et 1, l'impact des différentes caractéristiques d'un instant donné sur la prédiction finale. Bien que les poids d'attention à la variable $\boldsymbol{\beta}_i$ soient reliés aux représentations cachées $\boldsymbol{v}_i$ (voir la Figure 7.2), nous pouvons inférer l'attention aux variables $\boldsymbol{x}_i$ grâce à la linéarité du calcul des représentation cachées $\boldsymbol{v}_i$.

$$\boldsymbol{h}_{t-H}, ..., \boldsymbol{h}_t = \text{RNN}_\beta(\boldsymbol{v}_{t-H}, ..., \boldsymbol{v}_t) \quad \text{(Étape 3.1)}$$

$$\boldsymbol{\beta}_i = \tanh(\boldsymbol{W}_\beta \boldsymbol{h}_i + \boldsymbol{b}_\beta) \quad \text{(Étape 3.2)}$$

Étape 4 : Le vecteur de contexte $\boldsymbol{c}_t \in \mathcal{R}^m$ est calculé comme la somme, sur l'axe temporel, des caractéristiques $\boldsymbol{v}_i$ pondérées par leur attention temporelle α_i et leurs attentions à la variable $\boldsymbol{\beta}_i$ (voir Étape 4).

$$\boldsymbol{c}_t = \sum_{i=t-H}^{t} \alpha_i \boldsymbol{\beta}_i \odot \boldsymbol{v}_i \quad \text{(Étape 4)}$$

Étape 5 : Enfin, la prédiction $\hat{y}_{t+PH}$ est calculée par une couche dense linéaire selon l'Étape 5, où $\boldsymbol{w}_{out} \in \mathcal{R}^m$ et $b_{out} \in \mathcal{R}$. Après le calcul des prédictions, comme tout réseau de neurones, le modèle peut ajuster ses différents poids ($\boldsymbol{W}_{emb}$, RNN_α, RNN_β, $\boldsymbol{w}_\alpha$, $\boldsymbol{W}_\beta$ et $\boldsymbol{w}_{out}$) par rétropropagation du gradient d'erreur (e.g., erreur quadratique moyenne).

$$\hat{y}_{t+PH} = \boldsymbol{w}_{out}^T \boldsymbol{c}_t + b_{out} \quad \text{(Étape 5)}$$

Différences avec le modèle RETAIN de Choi *et al.* : Le modèle RETAIN a été initialement proposé pour

des tâches de classification (e.g., détection de crise cardiaque), tâches différentes de celles de régression (e.g., prédiction de la glycémie). Ainsi, nous l'avons adapté, à travers l'Étape 5, aux tâches régression.

Par ailleurs, dans sa version publiée par Choi *et al.*, les réseaux de neurones récurrents RNN_α et RNN_β traitent les instants temporels $t-H$ à t dans le sens inverse du temps. D'après les auteurs, cela permet au modèle d' imiter l'analyse des médecins se basant en premier lieu sur les consultations récentes. De notre côté, nos expérimentations n'ont pas témoigné de bénéficier de ce traitement dans le sens inverse du temps. Ainsi, pour la prédiction de la glycémie, il n'y aurait pas de préférence pour le sens de calcul des poids d'attention. Nous avons donc conservé un traitement standard dans le sens temporel [2].

7.2.3 Comment interpréter les prédictions faites par le modèle RETAIN

Contribution des variables d'entrée sur la prédiction

Les coefficients α_i et $\boldsymbol{\beta}_i$ représentent respectivement les poids des instants temporels passés et ceux des caractéristiques extraites $\boldsymbol{v}_i$ dans le calcul de la prédiction finale. Grâce à sa structure majoritairement linéaire, nous pouvons calculer la contribution de chaque variable d'entrée, à chaque instant, à la prédiction faite par l'architecture RETAIN. L'Équation 7.7 permet d'exprimer, à partir de l'Étape 5, le calcul de la prédiction finale $\hat{y}_{t+PH}$. Elle s'exprime en fonction des variables d'entrées $\boldsymbol{x}_i$ des attentions α_i et $\boldsymbol{\beta}_i$, ainsi que de la matrice $\boldsymbol{W}_{emb}$ calculant les caractéristiques $\boldsymbol{v}_i$, la matrice $\boldsymbol{w}_{out}$ et son biais associé b_{out} calculant la prédiction finale.

$$\hat{y}_{t+PH} = \boldsymbol{w}_{out}^T \boldsymbol{c}_t + b_{out} \tag{7.7a}$$

$$= \boldsymbol{w}_{out}^T \Big(\sum_{i=t-H}^{t} \alpha_i \boldsymbol{\beta}_i \odot \boldsymbol{v}_i \Big) + b_{out} \tag{7.7b}$$

$$= \boldsymbol{w}_{out}^T \Big(\sum_{i=t-H}^{t} \alpha_i \boldsymbol{\beta}_i \odot (\boldsymbol{W}_{emb} \boldsymbol{x}_i) \Big) + b_{out} \tag{7.7c}$$

L'Équation 7.8 donne la réécriture des caractéristiques $\boldsymbol{v}_i$ comme la somme sur j des variables d'entrées $x_{i,j}, j \in [1, r]$ pondérées par la j-ème colonne de la matrice $\boldsymbol{W}_{emb}$, $\boldsymbol{W}_{emb}[:, j]$.

$$\boldsymbol{v}_i = \boldsymbol{W}_{emb} \boldsymbol{x}_i \tag{7.8a}$$

$$= \sum_{j=1}^{r} x_{i,j} \boldsymbol{W}_{emb}[:, j] \tag{7.8b}$$

2. Le nom de RETAIN signifie *REverse Time AttentIoN*. Dans nos travaux, nous ne calculons pas les poids d'attention dans le sens inverse du temps. Bien que le nom de RETAIN ne soit ainsi plus très adéquat, nous l'avons conservé pour donner crédit aux auteurs.

En partant de l'Équation 7.7c, le calcul de la prédiction finale $\hat{y}_{t+PH}$ peut ainsi être réarrangé selon l'Équation 7.9.

$$\hat{y}_{t+PH} = \boldsymbol{w}_{out}^T(\sum_{i=t-H}^{t} \alpha_i \boldsymbol{\beta}_i \odot (\sum_{j=1}^{r} x_{i,j} \boldsymbol{W}_{emb}[:,j])) + b_{out} \tag{7.9a}$$

$$= \sum_{i=t-H}^{t} \sum_{j=1}^{r} \boldsymbol{w}_{out}^T(\alpha_i \boldsymbol{\beta}_i \odot (x_{i,j} \boldsymbol{W}_{emb}[:,j])) + b_{out} \tag{7.9b}$$

$$= \sum_{i=t-H}^{t} \sum_{j=1}^{r} x_{i,j} \alpha_i \boldsymbol{w}_{out}^T(\boldsymbol{\beta}_i \odot \boldsymbol{W}_{emb}[:,j]) + b_{out} \tag{7.9c}$$

Ce réarrangement permet de mettre en évidence la nature du calcul la prédiction finale $\hat{y}_{t+PH}$, s'exprimant comme la combinaison linéaire des variables d'entrées $x_{i,j}$. Ainsi, l'Équation 7.10 nous permet de donner une définition de la contribution, $\omega(\hat{y}_{t+PH}, x_{i,j})$, de j-ème variable à l'instant i à la prédiction $\hat{y}_{t+PH}$.

$$\omega(\hat{y}_{t+PH}, x_{i,j}) = \underbrace{\alpha_i \boldsymbol{w}_{out}^T(\boldsymbol{\beta}_i \odot \boldsymbol{W}_{emb}[:,j])}_{\text{coefficient de contribution}} \underbrace{x_{i,j}}_{\text{variable d'entrée}} \tag{7.10}$$

Contribution absolue normalisée

La contribution $w(\hat{y}_{t+PH}, x_{i,j})$ de la variable d'entrée $x_{i,j}$ sur la prédiction $\hat{y}_{t+PH}$ permet d'analyser en détail le mécanisme du modèle permettant d'aboutir à la prédiction. Cependant, cette valeur n'est pas pratique pour faire des analyses statistiques du comportement moyen du modèle. Premièrement, une variable peut avoir une contribution négative ou positive en fonction de la situation. Ainsi, le calcul de la contribution moyenne d'une telle variable risque de ne pas d'être représentatif de son réel impact sur les prédictions. Aussi, la contribution d'une variable d'entrée dépend de l'amplitude de la prédiction. Par conséquent, le calcul de la contribution *moyenne* d'une variable d'entrée ne donne pas autant d'importance à toutes les prédictions. En particulier, les analyses sont biaisées en faveur des prédictions à forte amplitude. Pour pallier ces limites, nous proposons dans cette étude d'utiliser la contribution absolue normalisée, $\omega_{AN}(\hat{y}_{t+PH}, x_{i,j})$, de la variable d'entrée $x_{i,j}$ sur la prédiction $\hat{y}_{t+PH}$. Décrite par l'Équation 7.11, elle permet de mesurer, entre 0 et 1, l'amplitude absolue de la contribution de la variable d'entrée $x_{i,j}$ sur la prédiction $\hat{y}_{t+PH}$.

$$\omega_{AN}(\hat{y}_{t+PH}, x_{i,j}) = \frac{|\omega(\hat{y}_{t+PH}, x_{i,j})|}{\sum_{k=t-H}^{t} \sum_{l=1}^{r} |\omega(\hat{y}_{t+PH}, x_{k,l})|} \tag{7.11}$$

7.3 Méthodologie

Dans le Chapitre 6, nous avons vu que l'apprentissage par transfert, et en particulier sa variante adverse, permet de grandement améliorer les performances des modèles. L'architecture RETAIN, étant un réseau de neurones, peut bénéficier du transfert de connaissances. Ainsi, dans ce chapitre, nous reprenons, dans l'ensemble, les étapes de traitement des données présentées au Chapitre 6 autour de l'apprentissage par transfert.

Comme pour les études des chapitres précédents, nous avons mis en accès libre l'ensemble de l'implémentation en Python de l'étude sur GitHub [28].

7.3.1 Données expérimentales

Dans cette étude, nous n'utilisons pas les données provenant du jeu T1DMS. En effet, ces données présentent moins d'intérêt que les données réelles des jeux IDIAB et OhioT1DM quant à l'évaluation des performances des modèles prédictifs. Aussi, dans le cadre de l'apprentissage par transfert, l'utilisation des données T1DMS comme sources du transfert s'est avérée moins intéressante que celle de données réelles.

7.3.2 Prétraitement des données

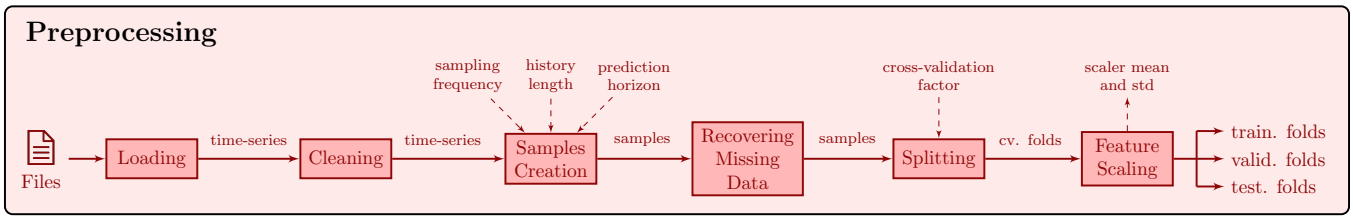


FIGURE 7.3: Étapes de prétraitement des données.

Les étapes de prétraitement des données, dont la Figure 7.3 en fait le rappel, sont similaires à celles présentées dans les chapitres précédents.

Afin d'opérer la méthodologie d'apprentissage par transfert multi-sources adverse présentée dans le Chapitre 6, nous reprenons les étapes de prétraitement suivantes :

— **Standardisation individuelle** : Pour l'apprentissage des modèles sur les patients sources du transfert, les données des différents patients ont été standardisées individuellement puis concaténées pour former un unique ensemble de données.

— **Concaténation des ensembles d'entraînement et de validation** : L'apprentissage sur les patients sources ne nécessite pas d'évaluation. Ainsi, les ensembles de test ont été utilisés comme ensembles de validation, et les ensembles de validation ont été concaténés aux ensembles d'entraînement.

Dans l'étude sur l'apprentissage par transfert, nous avons exploré un grand nombre de scénarios de transfert (e.g., entre patients d'un même jeu de données, provenant de jeux de données différents, etc.). Ici, nous avons choisi de nous restreindre au transfert *intra*. Celui-ci se caractérise par l'utilisation de patients sources provenant du même jeu de données que le patient cible (IDIAB vers IDIAB, I→I, et OhioT1DM vers OhioT1DM, O→O). Ce transfert a l'avantage d'être simple à opérer et d'avoir présenté de bonnes performances générales sur les réseaux de neurones convolutifs.

7.3.3 Présentation des modèles prédictifs

Nous présentons dans cette section les différents modèles prédictifs de glycémie utilisés dans cette étude : le modèle RETAIN, les deux modèles profonds de référence LSTM et FCN, ainsi que trois modèles de référence basés sur les arbres de décision DT, RF et GBM.

Modèle RETAIN

L'architecture RETAIN possède trois différents éléments à paramétrer : la dimension des caractéristiques extraites $\boldsymbol{v}_i$ et les tailles et natures des réseaux de neurones récurrents RNN_α et RNN_β. Après une recherche par grille sur l'ensemble de validation, nous avons choisi une dimension de caractéristiques de 64 ainsi que des réseaux récurrents de nature LSTM possédant une unique couche de 128 unités.

Afin d'implémenter la méthodologie d'apprentissage par transfert adverse, comme pour les modèles convolutifs du Chapitre 6, nous avons ajouté à l'architecture RETAIN un module classificateur de patients. Celui-ci a été positionné après le calcul du vecteur de contexte $\boldsymbol{c}_t$ qui représente les descripteurs finaux utilisés pour la prédiction. Symétriquement au calcul de la prédiction de la glycémie se faisant via une couche dense, la prédiction du patient d'origine des échantillons se fait avec une couche dense suivie d'une normalisation Softmax. Cela permet au module classificateur de patients d'être entraîné à minimiser l'entropie croisée multi-classes, dont le gradient d'erreur est inversé en arrivant au calcul du vecteur de contexte (Étape 4). Pour rappel, l'inversion du gradient d'erreur, combinée au classificateur de patient, permet l'extraction de caractéristiques qui soient à la fois utile pour la tâche finale de prédiction de glycémie, mais aussi agnostiques du patient d'origine.

L'entraînement du modèle RETAIN a été fait en utilisant l'optimiseur Adam et des *mini-batchs* de 50 échantillons. L'entraînement utilise la fonction de coût combinée décrite par l'Équation 7.12 où $\boldsymbol{y}^g$,$\boldsymbol{y}^p$,$\hat{\boldsymbol{y}}^g$,$\hat{\boldsymbol{y}}^p$ représentent respectivement la valeur observée de glycémie, le patient d'origine de l'échantillon, la valeur de glycémie prédite, et le patient prédit. Le coefficient λ permet de donner une importance relative différente aux deux objectifs. Le taux d'apprentissage global a été de 10^{-3} pendant l'entraînement sur les patients sources, puis de 10^{-4} pendant son affinage sur le patient cible. Pour éviter le surapprentissage du modèle aux données d'entraînement, la méthodologie d'arrêt anticipé de l'entraînement a été utilisée avec une patience de 100 époques pour l'apprentissage sur les patients sources et de 25 époques pour l'affinage du modèle sur le patient cible. Enfin, le coefficient λ a été de

$10^{-2.5}$, permettant d'obtenir une précision maximale après transfert sur l'ensemble de validation du patient cible.

$$Loss = MSE(\boldsymbol{y}^g, \hat{\boldsymbol{y}}^g) + \lambda \cdot Cross - Entropy(\boldsymbol{y}^p, \hat{\boldsymbol{y}}^p) \tag{7.12}$$

Modèles profonds de référence

Le modèle RETAIN utilise des réseaux de neurones récurrents LSTM afin de calculer les poids d'attention. Afin d'évaluer les performances liées à cette utilisation particulière du réseau LSTM, nous pouvons reprendre le modèle LSTM de l'étude benchmark du Chapitre 4.

Tout comme le modèle RETAIN, le modèle LSTM peut utiliser la méthodologie d'apprentissage adverse présentée au Chapitre 6. Pour cela, nous pouvons relier la représentation cachée du réseau, habituellement reliée à une couche dense pour faire la prédiction, à une deuxième couche dense parallèle effectuant la classification de patients. De manière similaire, le classificateur de patients est entraîné à minimiser l'entropie croisée multi-classes, dont le gradient d'erreur est inversé en arrivant au réseau LSTM.

Dans cette étude, nous reprenons l'architecture ainsi que les paramètres d'entraînement du modèle LSTM de l'étude benchmark du Chapitre 4. Il se compose de deux couches de 256 unités LSTM. Il est entraîné avec l'optimiseur Adam par *mini-batchs* de 50 échantillons avec un taux d'apprentissage de 10^{-3} pendant l'entraînement sur les patients sources, et de 10^{-4} pendant l'affinage sur le patient cible. Une régularisation L2 de 10^{-4} ainsi que la méthodologie d'arrêt anticipé (patience de 100 époques pendant l'apprentissage sur les patients sources, puis de 25 sur le patient cible) ont été utilisées pour limiter le surapprentissage du modèle aux données d'entraînement. Enfin, comme pour le modèle RETAIN, le gradient d'erreur lié à l'entropie croisée multi-classes du classificateur de patient est pondéré par $\lambda = 10^{-2.5}$.

Par ailleurs, nous reprenons les résultats des modèles convolutifs du Chapitre 6. En effet, ceux-ci ont montré de grandes améliorations de la précision des prédictions grâce à l'utilisation de la méthodologie d'apprentissage par transfert multi-sources adverse. En particulier, nous reprenons les modèles FCN issus du transfert adverse affiné, pour les scénarios *intra* (IDIAB vers IDIAB, I→I, et OhioT1DM vers OhioT1DM, O→O).

Modèles de référence basés sur les arbres de décision

Afin d'évaluer les performances du modèle RETAIN, nous avons choisi de les comparer aussi à un modèle simple, mais interprétable, DT (*decision tree*), basé sur les arbres de décision. Nous complétons ce modèle par deux autres modèles, RF (*random forest*) et GBM (*gradient boosting machine*), tous deux basés sur des ensembles d'arbres de décisions. Ces derniers modèles sont généralement plus performants que les arbres de décision simples en raison de leur complexité. Bien que ce gain en performances s'accompagne d'une baisse en interprétabilité, ces modèles sont plus interprétables que la plupart des modèles de l'apprentissage automatique grâce à l'importance Gini. Pour un arbre seul, l'importance Gini d'une variable d'entrée est calculée comme la baisse en impureté par le

noeud opérant la décision sur cette variable, pondérée par la probabilité d'atteindre le noeud. Pour un modèle RF ou GBM, l'importance des variables est moyennée sur l'ensemble de la forêt.

Le modèle DT (*decision tree*) est un arbre de décision standard. Bien que de nature simple, les arbres de décision ont été utilisés plusieurs fois pour la tâche de la prédiction de la glycémie [111]. Lors de la création de l'arbre, afin de réduire l'impact du surapprentissage aux données d'entraînement, nous pouvons contraindre la séparation de branches à avoir un minimum d'échantillons d'apprentissage supportant cette séparation. Ce nombre a été fixé à 100 pour le jeu de données IDIAB et 500 pour le jeu de données OhioT1DM. Cette différence s'explique par un nombre total d'échantillons d'apprentissage plus importants pour le jeu OhioT1DM permettant ainsi d'avoir une contrainte plus forte sur la séparation de branche.

Le modèle RF (*random forest*) est une forêt aléatoire. Il s'agit d'un modèle ensembliste basé sur les arbres de décision. Il est composé d'un grand nombre d'arbres de décision, chaque arbre étant différent des autres grâce à un processus de randomisation utilisé lors de leur création [132]. Cette randomisation affecte à la fois les descripteurs utilisés lors de la création de nouvelles branches, mais aussi la sélection des échantillons utilisés pour la création de l'arbre. Cette randomisation encourage la diversité au sein des arbres de décision. Cela permet à la prédiction finale, calculée comme la moyenne des décisions individuelles, d'être plus précise. Plus performants que les arbres de décisions classiques, les forêts aléatoires voient une utilisation croissante pour la tâche de la prédiction de la glycémie [113, 78, 111, 58]. Dans cette étude, nous avons utilisé une forêt de 100 arbres. Comme pour le modèle DT, nous avons optimisé par recherche par grille la contrainte du minimum d'échantillons pour opérer une séparation de branche. Cette valeur a été fixée à 50 et 250 échantillons pour les jeux IDIAB et OhioT1DM respectivement. Nous notons que ces valeurs sont plus faibles que pour le modèle DT. Cela s'explique intuitivement par un besoin moindre en régularisation, régularisation qui est déjà en partie effectuée par le mécanisme de création de la forêt.

Le modèle GBM (*gradient boosting machine*) est construit autour de la technique du boosting de gradient. De manière itérative, des arbres de décision sont créés, chaque arbre ayant pour objectif de réduire les erreurs des arbres précédemment créés. Cette méthode diffère des forêts aléatoires pour lesquels les arbres sont créés simultanément. Tout comme les forêts aléatoires, les modèles basés sur le boosting de gradient (e.g., GBM, LightGBM, XGBoost) sont eux aussi utilisés de plus en plus fréquemment dans le domaine de la prédiction de la glycémie [113, 78, 111]. Comme pour les modèles DT et RF, nous avons optimisé le nombre minimal d'échantillons nécessaires pour la création de nouvelles branches à 250 et 2000 pour les jeux IDIAB et OhioT1DM. Ces valeurs sont sensiblement plus élevées induisant des arbres moins profonds, ce qui est commun pour les modèles GBM. Lors de la création itérative des arbres, la contribution de chacun des arbres à la prédiction finale est diminuée par un coefficient portant le nom de taux d'apprentissage. Dans cette étude, nous avons optimisé le taux d'apprentissage à une valeur de 10^{-1}. Aussi, nous avons arrêté l'apprentissage au bout de 10 itérations sans amélioration des performances sur l'ensemble de validation. Cette méthode est similaire à celle de l'arrêt anticipé pratiquée dans l'apprentissage profond.

Contrairement aux réseaux de neurones, les modèles basés sur les arbres de décision ne peuvent pas utiliser la méthodologie d'apprentissage par transfert. Ainsi, les modèles DT, RF et GBM ont directement été entraînés sur les données du patient d'intérêt.

7.3.4 Évaluation des modèles prédictifs

L'évaluation des modèles prédictifs suit les étapes classiques présentées dans le Chapitre 4 et utilisées dans l'ensemble des travaux réalisés dans la thèse. Ces étapes sont rappelées par la Figure 7.4.

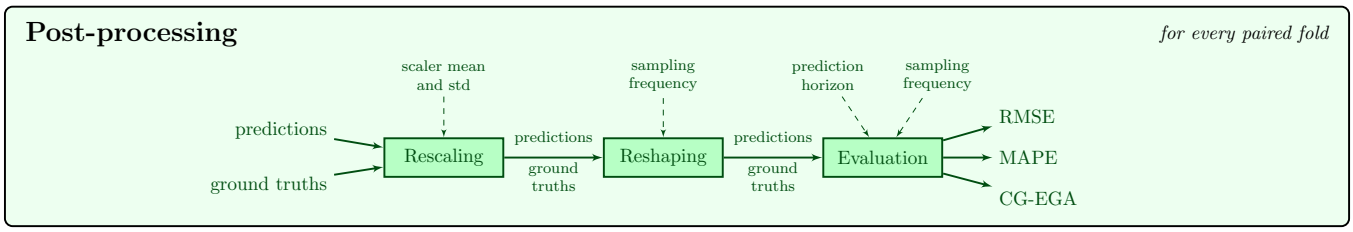


FIGURE 7.4: Post-traitement et évaluation des prédictions de glycémie.

Comme pour les études précédentes, nous nous intéressons uniquement à l'horizon de prédiction de 30 minutes. Après remise à l'échelle et réarrangement temporel des prédictions, l'évaluation des modèles se fait à travers trois métriques : la RMSE, la acsmape et la CG-EGA. Tandis que la RMSE et la acsmape donnent une mesure statistique de la précision des modèles, la CG-EGA évalue leur acceptabilité clinique. Pour chaque métrique, les performances sont moyennées sur les 5 ensembles de test de chaque patient lié à la validation croisée à 5 permutations, puis sur les patients d'un même jeu de données.

7.4 Résultats expérimentaux

Dans cette section, nous présentons tout d'abord les résultats statistiques et cliniques des différents modèles présentés précédemment. Puis, nous procéderons à l'analyse et l'interprétation des prédictions faites par le modèle RETAIN.

7.4.1 Présentation des résultats

Le Tableau 7.1 présente la précision moyenne (RMSE, acsmape) ainsi que l'acceptabilité clinique générale (CG-EGA générale) des modèles DT, RF, GBM, LSTM, FCN et RETAIN pour les jeux de données IDIAB et OhioT1DM. De son côté, le Tableau 7.2 donne les détails, par région, de l'acceptabilité clinique des modèles (CG-EGA par région).

Au sein des modèles de référence basés sur les arbres de décision (DT, RF et GBM), nous pouvons tout d'abord constater la faible précision et acceptabilité clinique du modèle DT en comparaison avec les modèles RF et GBM.

Modèle	RMSE	MAPE	CG-EGA (générale)		
			AP	BE	EP
Jeu de données IDIAB					
DT	24.45 (6.69)	11.44 (1.58)	88.18 (4.87)	8.38 (2.77)	3.44 (2.38)
RF	22.35 (6.33)	10.33 (1.50)	92.15 (4.51)	**4.76 (2.70)**	3.09 (2.12)
GBM	21.97 (6.13)	10.13 (1.60)	91.80 (4.29)	5.05 (2.53)	3.15 (2.13)
LSTM	19.27 (5.93)	8.66 (1.00)	92.12 (2.90)	5.57 (1.56)	**2.31 (1.69)**
FCN	**18.51 (5.48)**	**8.44 (1.07)**	92.23 (3.57)	5.27 (2.09)	2.50 (2.00)
RETAIN	19.49 (5.69)	8.71 (0.75)	**92.41 (2.94)**	5.15 (1.60)	2.43 (1.58)
Jeu de données OhioT1DM					
DT	23.87 (2.28)	11.22 (2.54)	79.07 (3.92)	16.81 (2.40)	**4.12 (2.13)**
RF	22.03 (2.41)	10.14 (2.38)	**83.67 (4.01)**	**11.89 (2.22)**	4.44 (2.28)
GBM	21.43 (2.35)	9.78 (2.48)	83.09 (3.85)	12.07 (1.82)	4.84 (2.38)
LSTM	19.68 (2.45)	8.81 (2.23)	79.37 (4.51)	15.61 (3.33)	5.02 (1.96)
FCN	**19.27 (1.78)**	**8.68 (1.97)**	78.73 (4.59)	15.96 (3.04)	5.31 (2.17)
RETAIN	20.29 (2.40)	9.16 (2.24)	80.98 (4.84)	14.28 (3.22)	4.74 (2.17)

Tableau 7.1: Précision statistique (RMSE et MAPE) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour les jeux de données IDIAB et OhioT1DM.

Modèle	CG-EGA (par région)								
	Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
Jeu de données IDIAB									
DT	36.49 (27.43)	0.57 (1.14)	62.94 (27.77)	92.47 (1.98)	6.65 (1.36)	0.88 (0.63)	85.07 (8.13)	11.62 (4.79)	3.30 (3.54)
RF	33.10 (29.94)	**0.00 (0.00)**	66.90 (29.94)	**96.38 (1.54)**	**3.10 (1.33)**	0.52 (0.37)	89.45 (7.38)	7.53 (4.69)	3.02 (2.78)
GBM	31.81 (29.18)	1.14 (2.29)	67.05 (28.58)	95.86 (1.60)	3.58 (1.39)	0.56 (0.28)	88.96 (6.81)	7.84 (4.08)	3.21 (2.85)
LSTM	52.02 (30.67)	**0.00 (0.00)**	47.98 (30.67)	95.17 (1.41)	4.45 (1.46)	**0.37 (0.35)**	**89.63 (5.60)**	7.65 (3.25)	2.72 (2.49)
FCN	51.84 (30.57)	**0.00 (0.00)**	48.16 (30.57)	95.87 (1.27)	3.62 (1.15)	0.51 (0.57)	88.82 (5.99)	8.38 (3.91)	**2.81 (2.64)**
RETAIN	**57.09 (33.07)**	**0.00 (0.00)**	**42.91 (33.07)**	95.63 (1.42)	3.94 (1.47)	0.43 (0.52)	89.09 (5.39)	**7.40 (3.03)**	3.51 (2.65)
Jeu de données OhioT1DM									
DT	23.67 (13.57)	3.61 (2.05)	72.72 (14.98)	80.96 (4.11)	16.60 (3.06)	**2.44 (1.15)**	79.51 (2.71)	17.65 (2.06)	**2.84 (1.27)**
RF	25.51 (17.82)	**1.42 (1.57)**	73.07 (18.34)	86.61 (3.72)	10.82 (2.79)	2.57 (1.11)	**82.53 (3.26)**	**13.92 (2.37)**	3.55 (1.71)
GBM	26.60 (19.79)	1.74 (1.87)	71.65 (20.89)	**86.73 (3.43)**	**10.33 (2.49)**	2.93 (1.15)	80.69 (4.16)	15.00 (2.66)	4.31 (2.01)
LSTM	**46.31 (24.61)**	2.43 (3.62)	**51.25 (25.13)**	83.02 (5.57)	13.48 (4.49)	3.50 (1.28)	75.96 (4.03)	18.74 (3.38)	5.30 (1.89)
FCN	44.98 (30.20)	2.83 (2.75)	52.19 (30.09)	82.29 (5.59)	13.99 (4.25)	3.71 (1.48)	75.35 (3.89)	18.86 (3.11)	5.78 (1.90)
RETAIN	44.08 (23.77)	2.89 (2.91)	53.03 (24.80)	84.11 (6.14)	12.57 (4.64)	3.33 (1.66)	78.81 (3.10)	16.58 (2.46)	4.61 (1.78)

Tableau 7.2: Acceptabilité clinique (CG-EGA) par région moyenne (écart type) pour un horizon de prédiction de 30 minutes pour les jeux de données IDIAB et OhioT1DM.

Cela n'est pas surprenant et s'explique par la simplicité trop importante d'un simple arbre de décision. Entre les modèles RF et GBM, le modèle GBM possède une meilleure précision statistique (RMSE, acsmape), mais aussi une moins bonne acceptabilité clinique (scores AP, BE et EP pour toutes les régions de la CG-EGA). Globalement, les résultats des modèles DT, RF et GBM sont similaires pour les deux jeux de données IDIAB et OhioT1DM. Les très bons scores en pourcentages d'erreurs bénignes BE du modèle RF montrent que ce dernier est capable de produire des prédictions successives cohérentes entre elles. En effet, comme nous avons pu le voir dans le Chapitre 5, une prédiction est caractérisée comme BE lorsqu'elle est cliniquement suffisamment précise, mais que la variation depuis la prédiction précédente ne l'est pas. Un modèle avec un fort taux de BE est généralement un modèle démontrant de fortes oscillations dans ses prédictions successives.

Quant aux modèles profonds de référence, LSTM et FCN, ceux-ci montrent des performances (précision et acceptabilité clinique) nettement supérieures aux modèles basés sur les arbres de décision. Seule l'acceptabilité clinique en région d'euglycémie et d'hyperglycémie pour le jeu OhioT1DM est moins bonne que celle des modèles RF et GBM (taux AP plus faible et taux EP plus élevé). Nous notons que les performances du modèle LSTM dans cette étude sont meilleures que celles du modèle LSTM de l'étude benchmark du Chapitre 4. Par exemple, le modèle LSTM du Chapitre 4 possède 19.46 et 20.85 en RMSE pour les jeux IDIAB et OhioT1DM respectivement contre 19.27 et 19.68 dans cette étude. Cette amélioration de la précision provient de l'utilisation de la méthodologie d'apprentissage par transfert multi-sources adverse. Toutefois, nous pouvons voir que le modèle FCN utilisé dans le Chapitre 6 reste le modèle le plus précis.

De son côté, le modèle RETAIN montre proposer un compromis entre précision et interprétabilité. En effet, ce dernier est nettement plus précis que les modèles basés sur les arbres de décision tout en restant interprétable. Sa précision reste toutefois légèrement inférieure à celle du modèle LSTM ou FCN. Nous pouvons attribuer cette différence à la simplicité du calcul de prédiction du modèle RETAIN. Dans l'architecture RETAIN, la non-linéarité du calcul réside seulement dans le calcul des poids d'attention. Cela force ainsi les descripteurs extraits à garder une certaine simplicité. Du point de vue de l'acceptabilité clinique, le modèle RETAIN est équivalent, voire légèrement meilleur, que les modèles FCN et LSTM.

7.4.2 Discussion

La plus grande force du modèle RETAIN réside dans son caractère interprétable. En effet, à travers la mesure de la contribution, il est possible de quantifier l'impact de chaque variable et ainsi de lever le voile sur le raisonnement du modèle derrière chaque prédiction. La Figure 7.5 donne un exemple de cette capacité. Dans cet exemple, nous voyons que les variables ayant le plus d'impact sur la prédiction sont les valeurs récentes de glycémie. Celles-ci présentent une contribution importante jusqu'à 1 heure dans le passé (historique de 1 heure). Quant aux apports en glucides et injections d'insuline, nous pouvons apercevoir des pics de contributions lorsqu'ils apparaissent. À ces mêmes instants, la contribution de la glycémie est proche de zéro. Ceci est permis par l'attention à la variable β_i

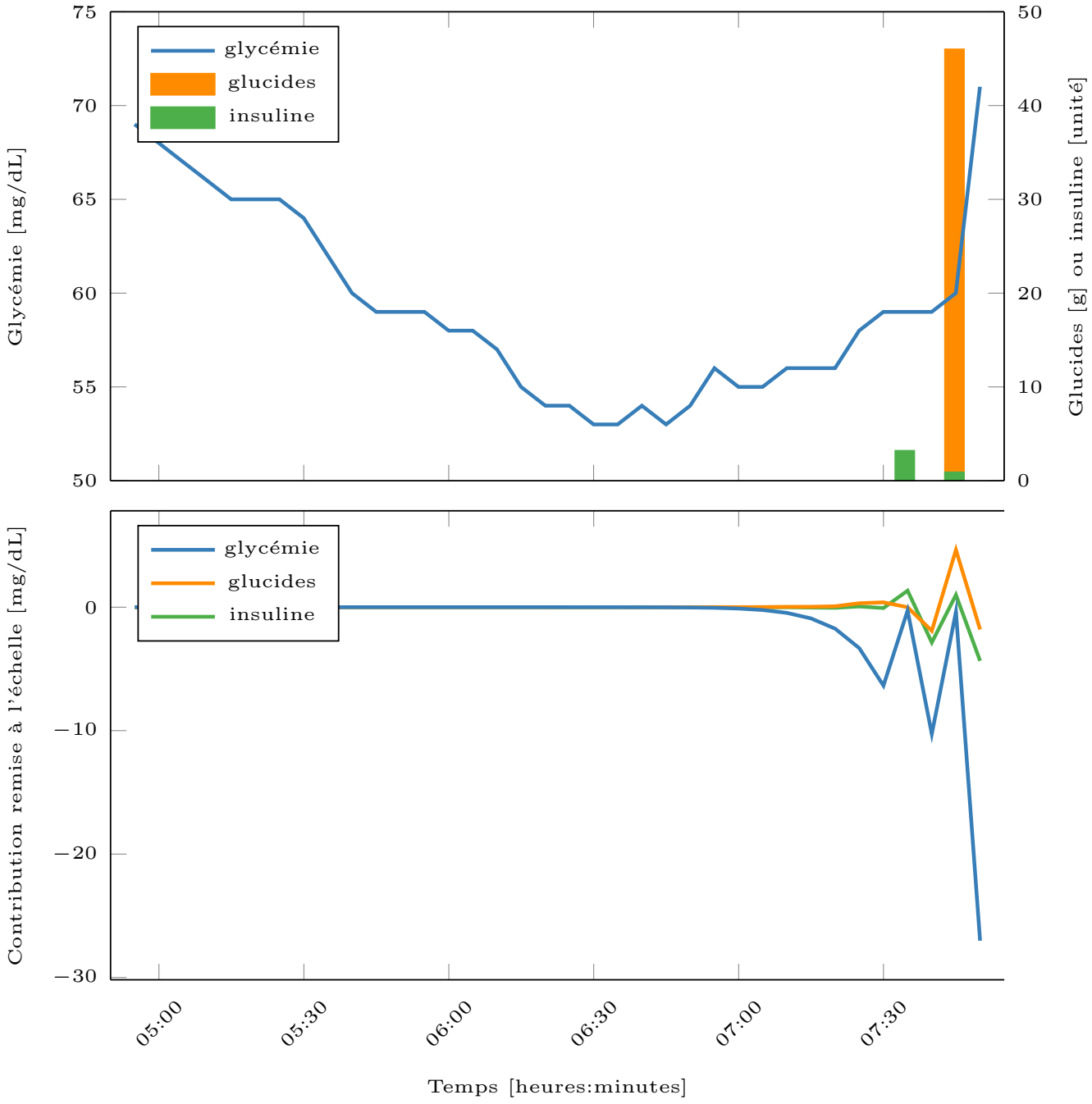
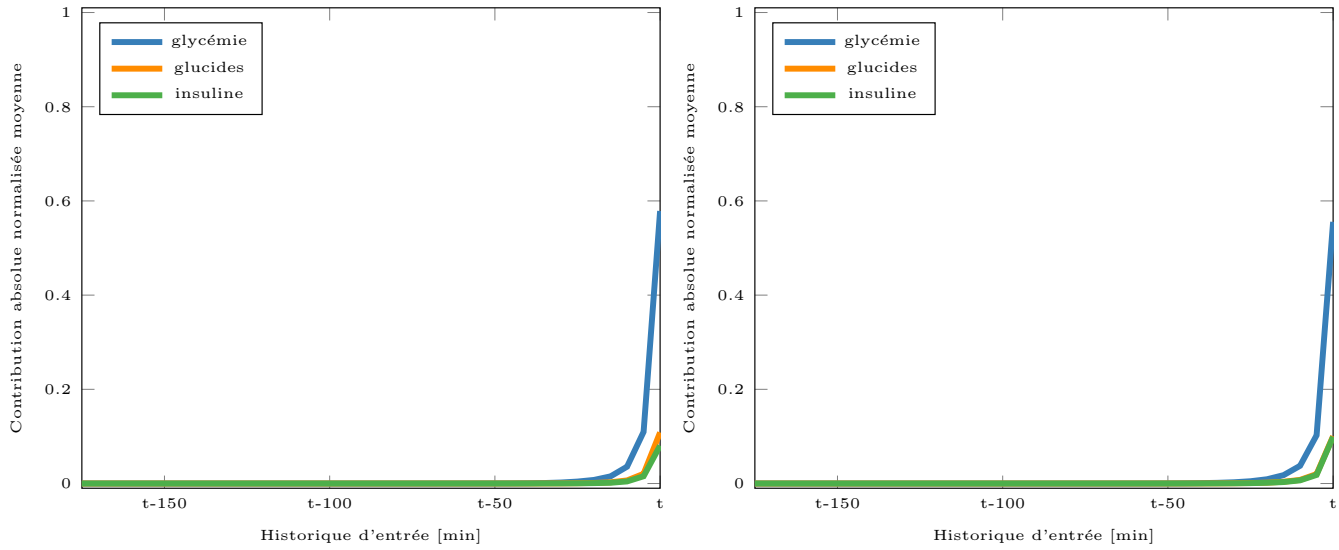


FIGURE 7.5: Variables d'entrée d'un échantillon de test du patient 575 du jeu OhioT1DM (haut) et contribution des variables à la prédiction faite par le modèle RETAIN (bas).

calculée par RNN_β. En effet, la seule présence de l'attention temporelle α_i calculée par RNN_α n'aurait pas permis d'attribuer une forte contribution à l'insuline/glucide et simultanément une faible contribution de la glycémie à cet instant. Enfin, sur cet exemple, nous pouvons noter que la contribution des variables d'une ancienneté supérieure à une heure est quasi nulle.

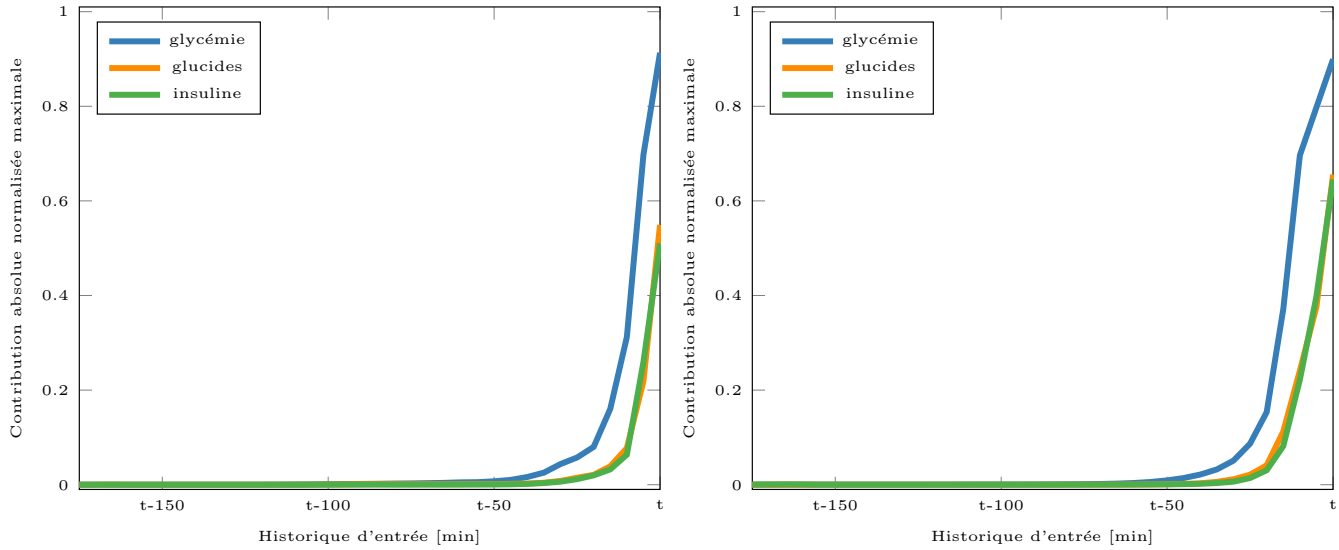
Nous pouvons utiliser la contribution absolue normalisée moyenne et maximale de chaque variable à chaque



(a) Jeu de données IDIAB

(b) Jeu de données OhioT1DM

FIGURE 7.6: Contribution absolue normalisée moyenne des variables d'entrées pour les patients des jeux de données IDIAB (gauche) et OhioT1DM (droite).



(a) Jeu de données IDIAB

(b) Jeu de données OhioT1DM

FIGURE 7.7: Contribution absolue normalisée maximale des variables d'entrées moyennée sur patients des jeux de données IDIAB (gauche) et OhioT1DM (droite).

instant afin d'évaluer son utilité pour la prédiction de la glycémie. Tandis que la contribution moyenne permet d'analyser le comportement moyen, la contribution maximale permet d'évaluer si une variable a été utile au moins une fois pour l'ensemble des échantillons de test. En effet, si une variable a été utile au moins une fois, alors sa contribution absolue normalisée maximale sera élevée (à la hauteur de son utilité). À l'inverse, si une variable n'est pas utilisée par le modèle pour calculer les prédictions, alors sa contribution sera proche de zéro. Les Figures 7.6 et 7.7 représentent respectivement la contribution absolue normalisée moyenne et maximale de chaque variable pour les jeux

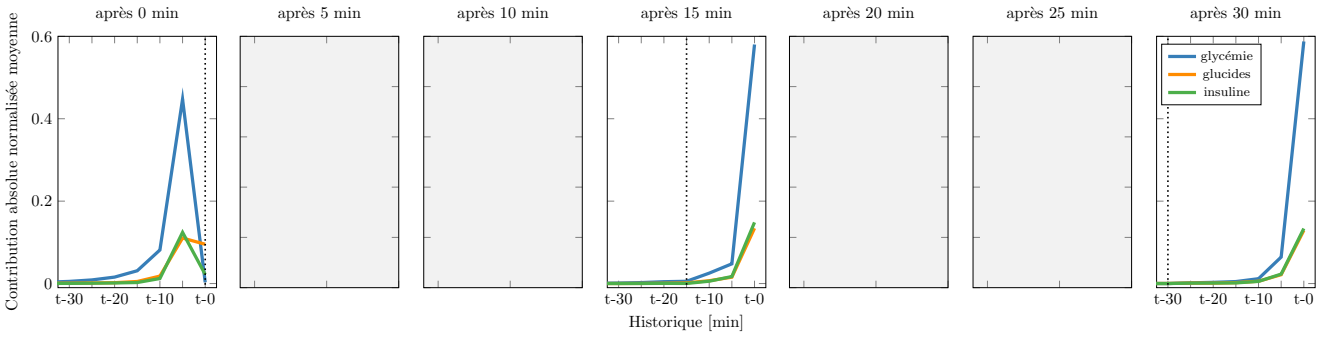
de données IDIAB et OhioT1DM. Premièrement, nous pouvons voir que l'intérêt de chaque signal décroît avec son ancienneté. Plus une variable est vieille, moins elle contribue aux prédictions. Cette décroissance est plus rapide pour les signaux d'ingestion de glucides ou d'injection d'insuline que pour le signal de glycémie. Tandis les signaux de glucides et d'insuline ne sont plus intéressants après 40 minutes environ, le signal de glycémie continue d'avoir un impact sur les prédictions jusqu'à 60 minutes. Au-delà de 60 minutes d'ancienneté, aucune variable ne montre avoir d'intérêt pour la prédiction de la glycémie. Pourtant, les autres modèles étudiés dans la thèse (e.g., SVR, GP, LSTM, FCN), ont montré bénéficier d'un historique long de plus d'une heure. Cela laisse supposer que le modèle RETAIN n'est pas capable d'utiliser efficacement un historique aussi long. Cette limitation expliquerait les performances légèrement moins bonnes du modèle par rapport aux modèles LSTM et FCN. Par ailleurs, la comparaison de la contribution absolue normalisée moyenne des deux jeux de données IDIAB et OhioT1DM à travers la Figure 7.6 montre que les variables des deux jeux, malgré leurs différences intrinsèques (types de diabète, matériels et protocoles expérimentaux différents), se comportent de manière similaire à l'intérieur du modèle RETAIN.

Pour aller plus loin dans l'analyse de la contribution de chaque variable, nous pouvons filtrer les échantillons présentant un intérêt particulier. Par exemple, à travers la Figure 7.8, nous nous intéressons à l'évolution de la contribution des variables après l'arrivée d'un évènement comme une ingestion de glucides ou une injection d'insuline. Le comportement du modèle est similaire à la fois pour les deux types d'évènements (glucides ou insuline) et pour les deux jeux de données (IDIAB et OhioT1DM). À l'instant précis de l'évènement, toutes les variables, excepté celle correspondant à l'évènement en question, possèdent une contribution presque nulle. À l'inverse, il s'agit plutôt de l'instant précédant l'évènement qui a une forte contribution sur la prédiction de la glycémie. Au fil du temps, bien que décroissante, la contribution de l'instant précédant l'évènement reste forte. Cela suggère que, lors de l'arrivée d'un évènement lié à la prise d'insuline ou de glucides, évènement modifiant considérablement la régulation de la glycémie du patient, le modèle tient particulièrement compte de l'état du patient avant la survenue de l'évènement. Après une trentaine de minutes, la contribution des variables liées à l'évènement devient nulle, indiquant que l'information de l'évènement n'est plus utilisée par le modèle pour faire ses prédictions.

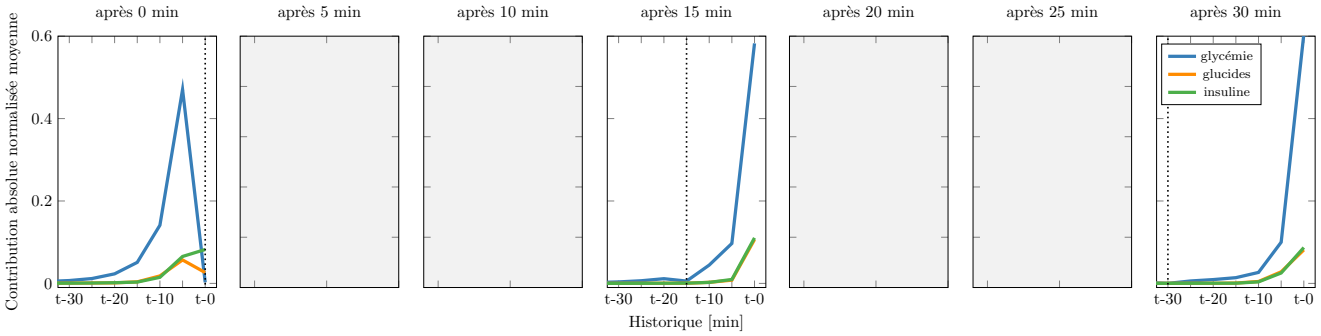
7.5 Conclusion

Dans cette étude, nous avons analysé et adapté l'architecture RETAIN proposée par Choi *et al.* pour la prédiction de la glycémie future de personnes diabétiques. Basé sur des réseaux de neurones, il implémente un principe d'attention double lui permettant d'être interprétable. Cette capacité le rend particulièrement intéressant pour des tâches biomédicales, et en particulier pour celle de la prédiction de la glycémie.

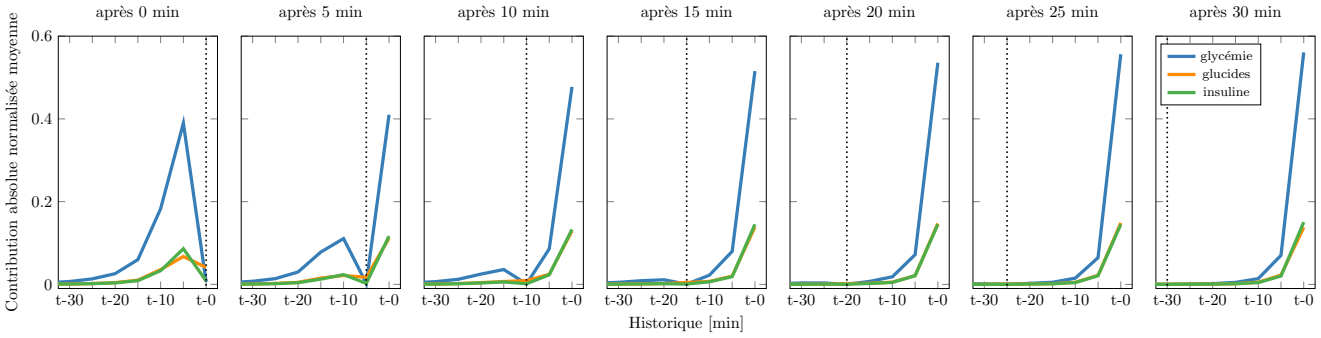
L'architecture RETAIN originelle a été conçue dans le cadre de l'analyse de dossiers électroniques de santé. Ces données sont caractérisées par des mesures physiologiques récurrentes dans le temps, au fil des consultations du patient. RETAIN utilise deux réseaux de neurones récurrents pour calculer des poids d'attention temporelle



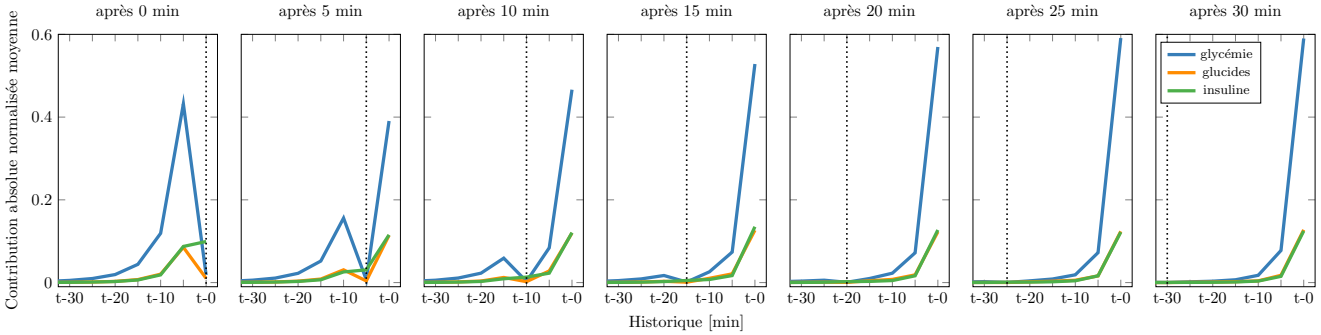
(a) IDIAB - ingestion de glucides



(b) IDIAB - injection d'insuline



(c) OhioT1DM - ingestion de glucides



(d) OhioT1DM - injection d'insuline

FIGURE 7.8: Évolution de la contribution absolue normalisée des variables après un évènement (ingestion de glucides ou injection d'insuline) moyennée pour les jeux IDIAB et OhioT1DM. Les contributions des variables du jeu IDIAB ne sont disponibles que toutes les 15 minutes en raison de la fréquence d'échantillonnage de son capteur de glycémie FreeStyle Libre.

et d'attention à la variable. Tandis que l'attention temporelle indique au modèle sur quel instant temporel (i.e., quelle consultation) se focaliser, l'attention à la variable indique, au sein de chaque instant temporel, quelle donnée d'entrée (i.e., signal physiologique) a de l'importance pour la tâche de prédiction en question. Son architecture linéaire à l'exception du calcul des poids d'attention lui permet d'être interprétable. Cette interprétation se fait à partir de la mesure de la contribution de chaque variable d'entrée à la prédiction finale.

RETAIN a initialement été conçu pour des tâches de classification (e.g., détection de crise cardiaque). Nous avons modifié sa couche de sortie afin de permettre son application à la prédiction de la glycémie qui est une tâche de régression. Les auteurs proposent de calculer les poids d'attention dans le sens inverse du temps afin d'imiter les médecins se basant en priorité sur les informations récentes. N'ayant pas mesuré empiriquement sur la prédiction de glycémie l'intérêt du calcul dans le sens inverse du temps, nous avons conservé une utilisation des réseaux récurrents dans le sens temporel standard. Quant à la mesure de la contribution des variables aux prédictions, nous proposons de prendre sa valeur absolue normalisée afin de la rendre plus adéquate à des analyses statistiques.

Nous avons évalué les performances statistiques (RMSE et acsmape) et cliniques (CG-EGA) du modèle RETAIN en le comparant à la fois à des modèles basés sur les arbres de décision et à des modèles issus de l'apprentissage profond. En effet, bien que les modèles DT, RF ou GBM soient généralement prisés pour leur meilleure interprétabilité, ils sont aussi souvent surclassés par des modèles plus complexes comme des réseaux de neurones. En particulier, nous avons utilisé des modèles de référence basés sur des arbres de décision standards, des forêts aléatoires ou forêts utilisant le boosting de gradient. Quant aux modèles profonds de référence, nous réutilisons les résultats obtenus par les modèles convolutifs FCN du Chapitre 6 sur un scénario de transfert *intra* ainsi qu'un modèle LSTM implémentant l'apprentissage par transfert multi-sources adverse. Comme le modèle LSTM et FCN, le modèle RETAIN a été entraîné en utilisant la méthodologie par transfert multi-sources adverse.

Les résultats nous ont montré que les modèles basés sur les arbres de décision sont largement surclassés par les modèles profonds, et notamment par le modèle RETAIN. En comparaison avec les modèles LSTM et FCN, le modèle RETAIN montre posséder une précision légèrement moins élevée, mais une meilleure, voire équivalente, acceptabilité clinique. Toutefois, la réelle force du modèle RETAIN réside dans son interprétabilité. Grâce à celle-ci, nous avons procédé à une analyse de l'importance des signaux de glycémie, de glucides et d'insuline en fonction de l'ancienneté des valeurs. Cette analyse nous a montré que les valeurs des signaux d'entrées vieilles de plus d'une heure (historique supérieur à une heure) ne sont pas utilisées par le modèle RETAIN. De leur côté, les modèles FCN et LSTM réussissent à utiliser efficacement ces valeurs anciennes expliquant ainsi leur meilleure précision statistique. Nous supposons que cette limitation vient de la quasi-linéarité du calcul des descripteurs fait par le modèle RETAIN. Par la suite, nous avons procédé à l'analyse de la contribution des variables d'entrée en présence d'un évènement de prise d'insuline ou de glucides. Suite à de tels évènements, le modèle RETAIN adopte un comportement différent en tenant fortement compte de l'instant précédant l'évènement. 30 minutes après la survenue de l'évènement, le modèle RETAIN retourne à son comportement standard.

Dans l'ensemble, le modèle RETAIN se révèle être prometteur pour un usage biomédical, et en particulier pour la prédiction de la glycémie future de personnes diabétiques. Son caractère interprétable est particulièrement intéressant à la fois pour le patient, mais aussi pour les praticiens et scientifiques derrière la création du modèle. Tout d'abord, un tel modèle peut être utile pour l'éducation thérapeutique du patient, lui expliquant l'incidence des variables sur la régulation de sa glycémie. Par ailleurs, le patient peut aussi comprendre le raisonnement derrière les décisions du modèle, et adapter en fonction son comportement. Enfin, comme nous avons pu le voir dans cette étude, l'analyse de l'importance des variables d'entrée peut se révéler capitale dans le design de nouvelles architectures plus performantes. Ces nouvelles architectures peuvent inclure de nouvelles données d'origines variées, comme des données d'activité physique ou de sommeil. Ces nouvelles architectures peuvent aussi être complexifiées, notamment à travers des méthodes d'extraction de descripteurs plus sophistiqués tout en restant interprétables. Dans la publication originale du RETAIN, les auteurs évoquent l'utilisation de perceptrons multicouches pour l'extraction de meilleurs descripteurs [17, 46, 90]. Toutefois, cette complexification de l'architecture doit permettre le calcul des contributions des variables à chaque instant à la prédiction finale pour ne pas perdre en interprétabilité.

8 | Conclusion et perspectives

Sommaire

8.1 Synthèse de la démarche scientifique

Dans cette thèse, nous nous intéressons à une innovation technologique prometteuse pour les personnes diabétiques. En utilisant les informations passées obtenues par le capteur de glycémie en continu, mais aussi des informations sur les repas et prises d'insuline, nous pouvons tenter de prédire leur glycémie future. De telles prédictions ont beaucoup de valeurs auprès de ces personnes, car elles leur permettraient d'anticiper les variations futures de glycémie et ainsi d'éviter les hypoglycémies ou hyperglycémies. Voici les étapes que nous avons suivies dans cette thèse :

1. Dans un premier temps, nous avons construit le jeu de données IDIAB en partenariat avec l'association Revesdiab. Celui-ci, composé de données de personnes diabétiques de type 2, combiné au jeu OhioT1DM, nous permet d'évaluer nos modèles prédictifs sur les deux types de diabète principaux.
2. Nous avons procédé à une analyse approfondie de l'état de l'art à travers la construction d'une base de résultats de référence. Cette analyse nous a permis d'identifier plusieurs problématiques, généralisables au secteur biomédical, auxquelles nous tentons d'apporter des réponses dans cette thèse.

3. L'entraînement des modèles n'inclut pas nécessairement les critères cliniques sur lesquels leur évaluation se porte. Afin de rendre les prédictions moins dangereuses pour le patient, nous nous sommes intéressé à comment améliorer l'acceptabilité clinique des modèles prédictifs. En particulier, nous avons exploré l'utilisation de nouvelles fonctions de coût dans l'apprentissage des modèles.
4. Bien que prometteuse, l'apprentissage profond pour la prédiction de la glycémie souffre du manque de données utilisées dans l'apprentissage des modèles. Pour combattre ce manque, nous avons étudié l'usage de l'apprentissage par transfert. Celui-ci permet de réutiliser les connaissances apprises sur plusieurs patients diabétiques lors de l'entraînement d'un modèle sur un nouveau patient.
5. La complexité accrue des modèles prédictifs de glycémie réduit leur interprétabilité. Celle-ci est particulièrement importante car elle permet de renforcer la confiance dans les prédictions. Nous nous intéressons à l'amélioration de l'interprétabilité des modèles profonds à travers le principe d'attention permettant de savoir sur quelles variables les prédictions sont basées.

8.2 Contributions

8.2.1 Construction et utilisation du corpus IDIAB

Afin de permettre l'évaluation des modèles prédictifs sur un ensemble de patients plus représentatifs de l'ensemble de la population diabétique, nous avons mené une campagne de collecte de données. Celle-ci, effectuée en collaboration avec le réseau pour personnes diabétiques Revesdiab, a eu pour objectif de collecter, auprès de patients diabétiques de type 2, des données variées utiles à la prédiction de la glycémie. Parmi ces données, nous comptons la glycémie en continu des patients, les ingestions de glucides, les prises d'insuline, leur activité physique, ainsi que leur humeur autoévaluée et les évènements du quotidien. Ces données ont été récoltées pour 6 patients diabétiques de type 2 en conditions réelles sur une durée de 4 semaines. L'utilisation de ce jeu de données dans l'apprentissage et évaluation des modèles prédictifs, et notamment sa comparaison à l'ensemble OhioT1DM, a montré que les modèles prédictifs fonctionnent de manière équivalente sur différents types de diabète ou dispositifs expérimentaux. Ceci est permis par la personnalisation des modèles aux patients permettant ainsi de prendre en compte la grande variabilité de la population diabétique.

8.2.2 Création d'une base de résultats de référence

À cause de la grande disparité dans les données utilisées dans chaque étude ainsi que dans leur processus d'évaluation, les comparaisons entre études ne peuvent pas être pertinentes. Toutefois, depuis la création du logiciel de simulation T1DMS ainsi que de la mise à disposition du jeu de données OhioT1DM, il est aujourd'hui possible d'entraîner et d'évaluer les modèles prédictifs sur des données similaires. Afin de faciliter et d'accélérer la recherche

dans le domaine de la prédiction de la glycémie, nous avons créé la base de résultats de référence GLYFE, présentée dans le Chapitre 4. En utilisant une méthodologie d'entraînement et d'évaluation standardisée, elle permet d'établir une comparaison juste entre les modèles prédictifs de glycémie. Afin de promouvoir son utilisation dans la communauté de la prédiction de la glycémie, nous avons mis à disposition, en accès libre, le code source du benchmark. À travers ce benchmark, nous avons évalué les performances de plusieurs modèles de référence utilisés dans la littérature. Ces résultats ont mis en évidence la difficulté de la tâche de prédiction de glycémie qui n'est efficace que pour un horizon de prédiction court de 30 minutes. Bien que confirmant globalement l'intérêt de l'utilisation de modèles prédictifs complexes, les performances statistiques et cliniques des modèles sont variables. En effet, au sein des modèles profonds étudiés, seul le modèle LSTM possède des performances compétitives. Aussi, les modèles les plus précis ne sont pas nécessairement les plus cliniquement acceptables. En particulier, les modèles ont montré posséder généralement une mauvaise acceptabilité clinique en région d'hypoglycémie. Enfin, au sein de cette étude, le modèle SVR s'est démarqué des autres en ayant à la fois une des meilleures précisions et une des meilleures acceptabilités cliniques.

La méthodologie proposée dans le benchmark a été utilisée pour l'ensemble des études présentées dans cette thèse. Cela permet à l'ensemble des résultats d'être totalement comparables les uns avec les autres. Nous avons regroupé l'ensemble de ces résultats au sein de l'Appendix A.

8.2.3 Inclusion de critères cliniques au sein de l'apprentissage profond

Les résultats de références ont montré qu'avoir un modèle précis n'était pas suffisant pour qu'il soit cliniquement acceptable, sans danger pour le patient. Nous avons montré que les erreurs cliniques de prédiction de la CG-EGA viennent du manque de cohérence dans les prédictions successives, mais aussi de prédictions peu précises en région d'hypoglycémie. Dans le Chapitre 5, dans l'optique d'améliorer la cohérence des prédictions successives, nous avons proposé une nouvelle fonction de coût, la cMSE. Celle-ci pénalise l'apprentissage du modèle non seulement sur les erreurs de prédiction de glycémie, mais aussi sur les erreurs de variations prédites. Implémentée en utilisant un réseau LSTM à deux sorties permettant de calculer les variations de glycémie prédites, la cMSE permet d'augmenter le taux de prédictions cliniquement précises AP. Pour aller plus loin, nous avons proposé une personnalisation de la cMSE à la prédiction de la glycémie, la gcMSE. Celle-ci donne une importance supérieure, pendant l'entraînement, aux prédictions obtenant un mauvais score selon les grilles P-EGA et R-EGA, composantes de la CG-EGA. Empiriquement, nous montrons que la gcMSE permet d'améliorer davantage le taux de prédictions AP, en réduisant les taux BE et EP. Toutefois, nous montrons aussi que ce gain en acceptabilité clinique s'accompagne d'une dégradation en précision statistique représentée par la RMSE.

En effet, en accentuant le poids de certaines prédictions ainsi qu'en ajoutant la nouvelle pénalité liée à la cohérence des prédictions successives, l'apprentissage du modèle se focalise moins sur l'objectif initial de précision moyenne et statistique des prédictions. La présence d'un double objectif à optimiser soulève des problèmes du point

de vue pratique pour l'utilisation de la gcMSE. En effet, celle-ci nécessite de faire un choix dans le compromis entre précision et acceptabilité clinique. Nous faisons l'hypothèse que ce choix peut se faire en pratique à partir de seuils minimaux en acceptabilité clinique (e.g., minimum 90% d'AP), seuils étant fixés par les autorités de santé. Afin d'inclure de tels critères dans l'apprentissage des modèles, nous proposons l'algorithme d'Amélioration Progressive de l'Acceptabilité Clinique (APAC). En accentuant progressivement les contraintes en acceptabilité clinique, l'algorithme permet d'obtenir le modèle ayant une précision maximale tout en respectant les critères médicaux. Toujours sur un réseau LSTM à deux sorties, nous montrons expérimentalement l'intérêt de l'algorithme qui permet de faire un compromis optimal entre précision et acceptabilité clinique.

8.2.4 Étude de l'apprentissage par transfert pour combattre le manque de données

Les performances relativement faibles des modèles profonds obtenus dans l'étude benchmark, couplées à d'autres résultats obtenus dans la littérature, suggèrent que les modèles profonds ne sont pas entraînés avec suffisamment de données et que l'apprentissage par transfert serait une réponse à ce problème. Dans le cadre de la prédiction de la glycémie, l'apprentissage par transfert est multi-sources et consiste en l'apprentissage d'un premier modèle sur plusieurs patients, puis en son affinage sur le patient d'intérêt. Dans le Chapitre 6, nous avons étudié l'utilisation de l'apprentissage par transfert pour la prédiction de la glycémie. Afin de lutter contre la surspécialisation des modèles aux différents patients sources entravant leur affinage sur les patients cibles, nous proposons la méthodologie d'apprentissage par transfert multi-sources adverse. Grâce à un module classifiant les patients à partir de la représentation cachée du réseau, et grâce à un mécanisme d'inversion de gradient, le modèle apprend une représentation cachée à la fois utile pour la tâche finale de prédiction de glycémie, mais aussi agnostique des patients. En étant plus générale, la représentation cachée est ainsi plus facilement transférable au patient cible.

À la lumière de réseaux totalement convolutifs (FCN), nous avons montré que l'apprentissage par transfert standard permet bien d'améliorer les performances des modèles. Nous montrons que la méthodologie adverse proposée permet d'améliorer davantage et significativement les performances. Dans cette thèse nous utilisons plusieurs jeux de données de natures très différentes : T1DMS, constitué de patients virtuels de type 1, IDIAB, composé de patients réels de type 2, et OhioT1DM fait de patients réels de type 1. En outre, les deux jeux IDIAB et OhioT1DM ont été obtenus avec des systèmes expérimentaux différents. Nous avons exploré la transférabilité des modèles en fonction de l'origine des patients sources par rapport au patient cible (e.g., depuis le même jeu de données, depuis des données simulées). Les résultats ont montré que le transfert multi-sources adverse est positif, quel que soit le scénario de transfert, y compris lorsque les patients sources ne proviennent pas du même jeu de données ou qu'ils sont virtuels. Les scénarios de transfert les plus efficaces restent toutefois les scénarios utilisant des patients provenant du même jeu que le patient cible, avec ou sans ajout de patients provenant d'autres jeux. Enfin, nous avons analysé le comportement du réseau appris par transfert adverse. Nous montrons visuellement et empiriquement que le transfert adverse permet d'améliorer la généralisation de la représentation cachée apprise

sur les patients sources. Cette généralisation accrue permet de faciliter l'affinage du modèle sur un nouveau patient cible.

8.2.5 Amélioration l'interprétabilité des modèles profonds

Les modèles les plus performants de l'état de l'art ainsi que ceux étudiés dans la thèse sont des modèles complexes basés sur des réseaux de neurones. Ces bonnes performances s'obtiennent en contrepartie d'une baisse en interprétabilité des prédictions. Pourtant celle-ci est particulièrement importante pour la prédiction de la glycémie, comme toutes tâches dans le domaine de la santé. En effet, celle-ci permet de faire confiance au modèle, car celui-ci peut expliciter le raisonnement derrière ses prédictions. Dans le Chapitre 7, nous avons étudié l'architecture RETAIN pour la prédiction de la glycémie. RETAIN, grâce à son double mécanisme d'attention, à la fois temporelle et à la variable, permet de calculer la contribution de chaque variable à la prédiction finale, le rendant ainsi interprétable.

Tirant parti de l'apprentissage par transfert, bien que légèrement moins bon que les autres réseaux de neurones comme les réseaux LSTM ou FCN, le modèle RETAIN a montré des performances prometteuses. Toutefois, le réel intérêt de l'architecture RETAIN réside dans son interprétabilité. Nous avons analysé l'importance des signaux d'entrée en fonction de leur ancienneté dans le mécanisme de prédiction. Nous montrons ainsi que les valeurs les plus anciennes que ce soit de glycémie, d'insuline ou de glucides, ne sont pas utilisées par le modèle RETAIN. Aussi, après une prise d'insuline ou de glucides, nous montrons que le modèle modifie son comportement en gardant en mémoire l'instant précédant l'évènement.

8.3 Limites et perspectives

Nous avons apporté des solutions afin d'améliorer la précision, l'acceptabilité clinique, ainsi que l'interprétabilité des modèles profonds pour la prédiction de la glycémie. Toutefois, nous pouvons identifier plusieurs limitations et opportunités :

— Bien que nos travaux soient parmi les seuls s'intéressant simultanément aux deux types du diabète, il serait bénéfique d'inclure davantage de patients diabétiques aux études pour mieux représenter la grande variabilité de la population diabétique. En particulier, la collecte de données faite avec le réseau Revesdiab dans le cadre de cette thèse pourrait être continuée afin d'augmenter le corpus de personnes diabétiques de type 2. Ces données pourraient alors être mises à disposition auprès de la communauté scientifique, à l'instar du jeu OhioT1DM. Quant à ce dernier, nous pourrions utiliser les 6 nouveaux patients mis à disposition avec la dernière mise à jour du corpus datant de juin 2020 [108].

— L'apprentissage par transfert permet d'apporter une réponse au problème du manque de données d'apprentissage en quantité suffisante. Toutefois, d'autres pistes complémentaires peuvent être envisagées. Par exemple,

celle de la génération de nouvelles données artificielles mérite d'être explorée, notamment à travers les réseaux antagonistes génératifs (GAN) [67]. En mettant en compétition un modèle générant de nouvelles données, et un second modèle discriminant les données artificielles des données réelles, les données générées artificiellement sont très proches des données réelles. Pour aller plus loin, la variante StyleGAN pourrait être utilisée sur les données déjà existantes pour les rendre plus semblables aux données du patient cible [79]. Cette méthodologie pourrait même être utilisée sur les données artificielles provenant du simulateur T1DMS afin de les rendre similaires aux données réelles. L'avantage de l'utilisation des données virtuelles provenant de T1DMS est qu'elles peuvent être simulées en très grandes quantités.

— La régulation de la glycémie est connue pour être influencée par beaucoup d'autres facteurs que les prises d'insuline ou les ingestions de glucides. Parmi ces facteurs nous retrouvons l'activité physique [73], le sommeil [82] ou l'humeur [138]. Bien que les jeux de données IDIAB et OhioT1DM possèdent des données décrivant ces différents facteurs, nous ne les avons pas utilisées dans les travaux présentés dans cette thèse. Une piste de recherche future serait d'étudier l'inclusion de ces nouvelles informations. Celle-ci peut se faire à travers l'architecture RETAIN présentée au Chapitre 7. En effet, grâce à son caractère interprétable, nous pourrions quantifier directement l'intérêt de l'ajout de telles données.

— Malgré l'intérêt de son caractère interprétable, l'architecture RETAIN a montré avoir une capacité prédictive inférieure aux autres modèles profonds comme les modèles FCN ou LSTM. Des modifications d'architecture autour du principe d'attention pourraient être recherchées afin d'améliorer cette capacité prédictive tout en veillant à ne pas perdre en interprétabilité. Par ailleurs, de nombreuses approches différentes ont été proposées ces dernières années pour lever le voile sur le raisonnement des modèles profonds. Parmi elles, nous pouvons citer l'utilisation de modèles de substitution interprétables (*surrogate models*), ou la méthode *Integrated Gradients* permettant d'identifier les variables d'entrées responsables de la prédiction finale [144]. Ces méthodes ont pour avantage d'être agnostiques du modèle utilisé.

A | Appendix A

Cet appendix résume les performances de l'ensemble des modèles étudiés dans la thèse. En particulier, les Tableaux A.1 et A.2 reprennent les performances générales (RMSE, MAPE, CG-EGA générale) et d'acceptabilité clinique détaille (CG-EGA par région) pour le jeu de données IDIAB. Les Tableaux A.3 et A.4 en font de même pour le jeu de données OhioT1DM.

Les résultats d'un grand nombre de modèles ont été reportés tout au long de la thèse. Afin de faciliter la lecture de ces tableaux, certains résultats n'ont pas été repris. En particulier, nous ne reportons que les meilleurs (en RMSE) scénarios de transfert du Chapitre 6. Aussi, pour différencier les modèles les uns des autres, nous utilisons les indices et exposants suivants :

— * : indique que les prédictions finales ont été lissées par lissage exponentiel (Chapitre 5) ;

— † : indique que le modèle a été entraîné avec la méthodologie d'apprentissage par transfert standard (Chapitre 6) ;

— ‡ : indique que le modèle a été entraîné avec la méthodologie d'apprentissage par transfert adverse (Chapitres 6 et 7) ;

— $_{G}$: indique qu'il s'agit d'un modèle global, non personnalisé au patient d'intérêt (Chapitre 6) ;

	Modèle	RMSE	MAPE	CG-EGA (générale)		
				AP	BE	EP
Chapitre 4	**Ref**	24.50 (8.54)	10.75 (1.87)	94.16 (4.76)	3.84 (2.80)	2.00 (2.13)
	Poly	53.71 (7.03)	28.70 (7.09)	89.13 (4.85)	3.67 (3.39)	7.20 (3.10)
	AR	20.81 (7.41)	9.38 (1.39)	91.17 (4.11)	6.46 (2.88)	2.37 (1.85)
	ARX	20.44 (6.49)	9.27 (1.18)	92.11 (3.75)	5.37 (2.01)	2.52 (1.91)
	SVR	20.32 (6.02)	8.66 (0.44)	92.69 (2.81)	5.34 (2.06)	1.97 (1.23)
	GP	19.80 (5.92)	8.92 (0.80)	91.92 (3.14)	5.80 (2.19)	2.29 (1.85)
	ELM	26.25 (7.57)	11.82 (1.06)	91.21 (4.60)	5.59 (2.57)	3.20 (2.47)
	FFNN	22.01 (6.26)	9.97 (0.89)	91.45 (3.77)	5.48 (2.60)	3.07 (2.23)
	LSTM	19.85 (6.00)	9.04 (1.11)	92.20 (2.99)	5.05 (1.71)	2.76 (1.82)
Chapitre 5	**SVR***	20.67 (6.20)	8.86 (0.44)	93.62 (2.57)	4.47 (1.69)	1.92 (1.35)
	LSTM*	20.27 (6.30)	9.25 (1.21)	93.16 (3.13)	4.16 (1.75)	2.68 (2.00)
	pcLSTM	21.89 (5.68)	10.28 (1.34)	94.04 (3.26)	3.20 (1.66)	2.76 (2.07)
	pcLSTM*	22.63 (6.04)	10.64 (1.40)	94.24 (3.35)	**2.94 (1.73)**	2.82 (2.07)
	gpcLSTM	21.21 (5.64)	9.35 (0.92)	94.03 (2.66)	3.91 (1.48)	2.06 (1.54)
	gpcLSTM*	21.86 (5.94)	9.66 (0.95)	94.53 (2.84)	3.38 (1.55)	2.08 (1.57)
	gpcLSTM$_{\mathbf{AC}}$	40.68 (11.20)	18.14 (5.55)	95.34 (2.76)	3.29 (2.56)	**1.37 (0.91)**
	gpcLSTM$^{\star}_{\mathbf{AC}}$	41.15 (11.18)	18.36 (5.47)	**95.35 (2.87)**	3.20 (2.61)	1.45 (0.92)
	gpcLSTM$^{\star}_{\mathbf{APAC}}$	24.03 (7.15)	10.43 (1.18)	95.00 (2.74)	3.38 (1.99)	1.61 (1.22)
Chapitre 6	**FCN #1**	21.06 (5.14)	9.66 (1.00)	92.80 (3.40)	4.44 (1.75)	2.76 (2.17)
	FCN #2	20.64 (5.20)	9.62 (1.27)	93.26 (4.24)	3.93 (2.47)	2.82 (2.15)
	FCN$^{\dagger}_{\mathbf{G}}$	20.20 (5.90)	9.51 (1.49)	91.44 (4.96)	5.39 (2.70)	3.16 (2.43)
	FCN†	19.10 (5.04)	8.95 (1.00)	92.28 (4.03)	4.84 (2.12)	2.88 (2.16)
	FCN$^{\ddagger}_{\mathbf{G}}$	19.61 (6.27)	8.95 (1.00)	91.80 (3.75)	5.75 (2.63)	2.45 (1.42)
	FCN‡	**18.51 (5.48)**	**8.44 (1.07)**	92.23 (3.57)	5.27 (2.09)	2.50 (2.00)
Chapitre 7	**DT**	24.45 (6.69)	11.44 (1.58)	88.18 (4.87)	8.38 (2.77)	3.44 (2.38)
	RF	22.35 (6.33)	10.33 (1.50)	92.15 (4.51)	4.76 (2.70)	3.09 (2.12)
	GBM	21.97 (6.13)	10.13 (1.60)	91.80 (4.29)	5.05 (2.53)	3.15 (2.13)
	LSTM‡	19.27 (5.93)	8.66 (1.00)	92.12 (2.90)	5.57 (1.56)	2.31 (1.69)
	RETAIN‡	19.49 (5.69)	8.71 (0.75)	92.41 (2.94)	5.15 (1.60)	2.43 (1.58)

Tableau A.1: Précision statistique (RMSE et MAPE) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeu de données IDIAB.

	Modèle	CG-EGA (par région)								
		Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
		AP	BE	EP	AP	BE	EP	AP	BE	EP
Chapitre 4	**Ref**	76.15 (19.57)	**0.00 (0.00)**	23.85 (19.57)	97.35 (1.23)	2.29 (1.13)	0.36 (0.17)	90.04 (9.01)	6.50 (4.91)	3.46 (4.25)
	Poly	0.24 (0.47)	**0.00 (0.00)**	99.76 (0.47)	96.06 (3.29)	2.48 (2.23)	1.46 (1.75)	84.17 (9.41)	5.37 (5.43)	10.46 (9.02)
	AR	53.26 (26.47)	**0.00 (0.00)**	46.74 (26.47)	94.34 (2.12)	5.10 (2.15)	0.56 (0.54)	87.73 (7.10)	9.24 (4.67)	3.02 (2.89)
	ARX	54.85 (23.44)	**0.00 (0.00)**	45.15 (23.44)	95.09 (1.54)	4.34 (1.31)	0.57 (0.50)	89.16 (6.87)	7.46 (3.88)	3.38 (3.11)
	SVR	69.39 (33.51)	0.35 (0.70)	30.27 (33.54)	95.17 (2.01)	4.33 (1.83)	0.50 (0.47)	89.51 (6.09)	7.43 (3.86)	3.06 (2.53)
	GP	60.89 (22.05)	**0.00 (0.00)**	39.11 (22.05)	94.67 (2.33)	4.89 (2.26)	0.44 (0.51)	88.83 (5.93)	7.77 (3.53)	3.40 (3.17)
	ELM	31.06 (40.09)	0.70 (1.39)	68.25 (39.57)	95.13 (2.28)	4.23 (2.09)	0.64 (0.70)	88.71 (7.29)	7.95 (4.29)	3.34 (3.27)
	FFNN	33.88 (37.15)	0.17 (0.35)	65.95 (37.01)	95.09 (2.40)	4.36 (2.32)	0.56 (0.41)	89.26 (6.58)	7.54 (3.75)	3.20 (3.14)
	LSTM	47.39 (31.03)	**0.00 (0.00)**	52.61 (31.03)	95.85 (1.27)	3.75 (1.34)	0.39 (0.34)	89.91 (5.08)	7.00 (2.75)	3.09 (2.46)
Chapitre 5	**SVR***	66.37 (31.47)	0.17 (0.35)	33.45 (31.51)	96.13 (1.81)	3.49 (1.66)	0.39 (0.36)	90.61 (5.67)	6.60 (3.23)	2.79 (2.79)
	LSTM*	37.99 (31.22)	**0.00 (0.00)**	62.01 (31.22)	96.71 (1.35)	2.95 (1.46)	0.33 (0.38)	91.02 (6.04)	6.18 (3.67)	2.80 (2.58)
	pcLSTM	34.59 (29.27)	**0.00 (0.00)**	65.41 (29.27)	97.58 (0.90)	2.13 (0.82)	0.29 (0.20)	92.60 (5.81)	4.94 (3.18)	2.46 (2.80)
	pcLSTM*	32.20 (27.83)	**0.00 (0.00)**	67.80 (27.83)	**97.96 (0.98)**	**1.81 (0.91)**	**0.23 (0.11)**	92.81 (6.25)	**4.68 (3.48)**	2.51 (2.85)
	gpcLSTM	64.79 (24.95)	**0.00 (0.00)**	35.21 (24.95)	96.60 (1.11)	3.03 (0.99)	0.37 (0.26)	92.06 (5.12)	5.42 (2.83)	2.51 (2.46)
	gpcLSTM*	61.87 (25.17)	**0.00 (0.00)**	38.13 (25.17)	97.23 (1.17)	2.46 (1.02)	0.31 (0.22)	92.65 (5.60)	4.85 (3.09)	2.50 (2.68)
	gpcLSTM$_{AC}$	**87.95 (9.58)**	1.71 (3.43)	**10.34 (8.15)**	97.37 (1.36)	2.12 (1.03)	0.51 (0.40)	92.17 (4.46)	5.11 (4.52)	2.72 (2.39)
	gpcLSTM*$_{AC}$	87.77 (9.53)	1.71 (3.43)	10.51 (8.13)	97.50 (1.32)	1.97 (0.97)	0.52 (0.44)	92.10 (4.69)	5.03 (4.70)	2.87 (2.33)
	gpcLSTM*$_{APAC}$	68.49 (27.85)	0.57 (1.14)	30.94 (28.22)	97.35 (1.18)	2.32 (1.08)	0.33 (0.15)	**93.16 (4.84)**	5.08 (3.53)	**1.76 (1.49)**
Chapitre 6	**FCN #1**	40.55 (33.71)	**0.00 (0.00)**	59.45 (33.71)	96.85 (0.89)	2.77 (1.02)	0.38 (0.34)	90.09 (6.73)	7.20 (3.94)	2.71 (3.09)
	FCN #2	30.55 (34.32)	**0.00 (0.00)**	69.45 (34.32)	97.57 (0.95)	2.11 (0.97)	0.32 (0.43)	90.67 (6.91)	6.89 (4.72)	2.44 (2.38)
	FCN$_G^{\dagger}$	40.24 (33.75)	**0.00 (0.00)**	59.76 (33.75)	95.42 (2.39)	4.13 (2.20)	0.45 (0.54)	88.55 (7.76)	7.91 (4.06)	3.55 (3.98)
	FCN†	45.14 (22.19)	**0.00 (0.00)**	54.86 (22.19)	95.81 (1.79)	3.80 (1.71)	0.39 (0.53)	89.66 (6.52)	6.92 (3.45)	3.42 (3.47)
	FCN$_G^{\ddagger}$	51.95 (31.37)	**0.00 (0.00)**	48.05 (31.37)	95.26 (1.69)	4.23 (1.46)	0.51 (0.51)	88.88 (6.95)	8.56 (4.54)	2.56 (2.64)
	FCN‡	51.84 (30.57)	**0.00 (0.00)**	48.16 (30.57)	95.87 (1.27)	3.62 (1.15)	0.51 (0.57)	88.82 (5.99)	8.38 (3.91)	2.81 (2.64)
Chapitre 7	**DT**	36.49 (27.43)	0.57 (1.14)	62.94 (27.77)	92.47 (1.98)	6.65 (1.36)	0.88 (0.63)	85.07 (8.13)	11.62 (4.79)	3.30 (3.54)
	RF	33.10 (29.94)	**0.00 (0.00)**	66.90 (29.94)	96.38 (1.54)	3.10 (1.33)	0.52 (0.37)	89.45 (7.38)	7.53 (4.69)	3.02 (2.78)
	GBM	31.81 (29.18)	1.14 (2.29)	67.05 (28.58)	95.86 (1.60)	3.58 (1.39)	0.56 (0.28)	88.96 (6.81)	7.84 (4.08)	3.21 (2.85)
	LSTM‡	52.02 (30.67)	**0.00 (0.00)**	47.98 (30.67)	95.17 (1.41)	4.45 (1.46)	0.37 (0.35)	89.63 (5.60)	7.65 (3.25)	2.72 (2.49)
	RETAIN‡	57.09 (33.07)	**0.00 (0.00)**	42.91 (33.07)	95.63 (1.42)	3.94 (1.47)	0.43 (0.52)	89.09 (5.39)	7.40 (3.03)	3.51 (2.65)

Tableau A.2: Acceptabilité clinique (CG-EGA) par région moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeux de données IDIAB.

	Modèle	RMSE	MAPE	CG-EGA (générale)		
				AP	BE	EP
Chapitre 4	**Ref**	23.40 (2.56)	10.96 (2.39)	88.94 (3.02)	8.08 (1.98)	2.98 (1.46)
	Poly	57.27 (6.59)	31.09 (6.71)	84.29 (3.32)	**5.66 (1.72)**	10.05 (3.41)
	AR	20.70 (2.23)	9.62 (2.26)	81.72 (4.05)	13.50 (3.22)	4.78 (1.79)
	ARX	20.61 (2.20)	9.59 (2.19)	81.49 (4.02)	13.68 (3.16)	4.84 (1.86)
	SVR	20.15 (2.33)	9.12 (2.11)	83.35 (3.91)	12.38 (2.83)	4.28 (1.83)
	GP	20.02 (2.32)	9.19 (2.15)	81.16 (4.34)	14.26 (3.18)	4.57 (1.85)
	ELM	25.30 (1.37)	11.55 (2.46)	76.39 (4.27)	17.93 (2.83)	5.68 (2.36)
	FFNN	21.43 (2.07)	9.82 (2.22)	76.34 (4.33)	17.95 (3.02)	5.71 (2.28)
	LSTM	20.46 (2.08)	9.24 (2.10)	80.03 (4.17)	14.83 (2.88)	5.14 (2.11)
Chapitre 5	**SVR***	20.17 (2.30)	9.18 (2.12)	85.00 (4.05)	10.97 (2.72)	4.03 (1.90)
	LSTM*	20.43 (2.03)	9.26 (2.10)	82.14 (3.94)	13.06 (2.51)	4.81 (2.04)
	pcLSTM	21.53 (2.23)	10.07 (2.32)	87.45 (3.76)	8.46 (2.05)	4.09 (2.14)
	pcLSTM*	21.71 (2.22)	10.19 (2.35)	87.89 (3.61)	8.15 (1.94)	3.96 (2.12)
	gpcLSTM	21.66 (2.69)	9.65 (2.14)	86.97 (3.63)	9.50 (2.52)	3.53 (1.48)
	gpcLSTM*	21.82 (2.69)	9.76 (2.16)	87.59 (3.45)	9.01 (2.31)	3.41 (1.49)
	gpcLSTM$_{AC}$	47.70 (6.31)	22.43 (2.76)	90.46 (2.85)	7.16 (1.66)	**2.37 (1.28)**
	gpcLSTM$^{*}_{AC}$	47.82 (6.27)	22.47 (2.76)	**90.51 (2.88)**	7.12 (1.64)	2.37 (1.30)
	gpcLSTM$^{*}_{APAC}$	23.50 (2.49)	10.46 (2.09)	88.72 (3.59)	8.20 (2.23)	3.08 (1.64)
Chapitre 6	**FCN #1**	21.51 (1.89)	9.82 (2.08)	83.59 (3.98)	11.73 (2.31)	4.68 (2.12)
	FCN #2	20.61 (2.09)	9.34 (2.07)	84.21 (4.53)	11.23 (2.81)	4.56 (2.23)
	FCN$^{\dagger}_{G}$	20.68 (2.12)	9.58 (2.14)	78.37 (2.78)	15.94 (1.88)	5.69 (2.00)
	FCN†	19.57 (2.02)	8.93 (2.13)	78.80 (4.71)	15.85 (2.93)	5.34 (2.31)
	FCN$^{\ddagger}_{G}$	19.45 (1.78)	9.19 (1.91)	78.12 (3.85)	16.67 (2.97)	5.21 (1.65)
	FCN‡	**18.94 (1.66)**	**8.50 (1.87)**	79.68 (4.11)	15.34 (2.73)	4.98 (1.87)
Chapitre 7	**DT**	23.87 (2.28)	11.22 (2.54)	79.07 (3.92)	16.81 (2.40)	4.12 (2.13)
	RF	22.03 (2.41)	10.14 (2.38)	83.67 (4.01)	11.89 (2.22)	4.44 (2.28)
	GBM	21.43 (2.35)	9.78 (2.48)	83.09 (3.85)	12.07 (1.82)	4.84 (2.38)
	LSTM‡	19.68 (2.45)	8.81 (2.23)	79.37 (4.51)	15.61 (3.33)	5.02 (1.96)
	RETAIN‡	20.29 (2.40)	9.16 (2.24)	80.98 (4.84)	14.28 (3.22)	4.74 (2.17)

Tableau A.3: Précision statistique (RMSE et MAPE) et acceptabilité clinique (CG-EGA) générale moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeu de données OhioT1DM.

	Modèle	CG-EGA (par région)								
		Hypoglycémie			*Euglycémie*			*Hyperglycémie*		
		AP	BE	EP	AP	BE	EP	AP	BE	EP
Chapitre 4	**Ref**	47.23 (23.39)	3.99 (3.70)	48.77 (23.51)	91.37 (3.22)	6.63 (2.46)	2.01 (0.90)	86.95 (3.35)	10.27 (2.44)	2.78 (1.97)
	Poly	0.00 (0.00)	**0.00 (0.00)**	100.00 (0.00)	**94.54 (1.74)**	**5.20 (1.84)**	**0.27 (0.55)**	75.71 (6.30)	**7.00 (2.82)**	17.29 (5.72)
	AR	38.11 (21.40)	5.30 (3.87)	56.59 (22.30)	85.42 (5.40)	11.47 (4.22)	3.10 (1.32)	79.18 (2.98)	16.06 (3.16)	4.75 (1.67)
	ARX	38.32 (23.33)	4.88 (3.92)	56.80 (23.69)	85.10 (5.41)	11.67 (4.25)	3.23 (1.34)	78.96 (2.91)	16.26 (3.00)	4.78 (1.69)
	SVR	49.71 (18.75)	5.62 (4.02)	44.67 (18.70)	86.35 (4.24)	10.71 (3.26)	2.94 (1.23)	80.85 (3.24)	14.77 (3.01)	4.37 (1.84)
	GP	45.09 (26.13)	6.94 (4.57)	47.97 (27.41)	84.61 (5.37)	12.21 (4.15)	3.17 (1.40)	78.29 (3.53)	16.87 (3.21)	4.84 (1.59)
	ELM	30.75 (23.97)	3.93 (3.94)	65.32 (26.93)	79.43 (4.09)	16.93 (2.98)	3.64 (1.47)	74.11 (4.33)	19.98 (3.30)	5.91 (1.78)
	FFNN	37.17 (18.07)	6.07 (3.42)	56.76 (17.33)	80.11 (4.57)	15.97 (3.56)	3.92 (1.47)	72.65 (4.21)	21.64 (3.28)	5.71 (1.79)
	LSTM	38.37 (23.17)	3.97 (3.72)	57.67 (24.23)	83.78 (5.33)	12.70 (4.06)	3.52 (1.47)	76.86 (3.70)	17.87 (2.73)	5.27 (2.21)
Chapitre 5	**SVR***	46.95 (21.11)	5.97 (4.05)	47.09 (21.65)	87.83 (4.22)	9.46 (3.21)	2.71 (1.22)	82.81 (3.43)	13.12 (2.98)	4.07 (2.00)
	LSTM*	37.34 (23.50)	4.11 (4.15)	58.56 (24.17)	85.71 (4.83)	11.10 (3.58)	3.19 (1.37)	79.27 (3.55)	15.85 (2.40)	4.88 (2.24)
	pcLSTM	25.28 (19.11)	3.64 (3.73)	71.08 (19.35)	90.79 (3.43)	6.93 (2.53)	2.28 (1.01)	85.78 (3.64)	10.83 (2.55)	3.40 (2.03)
	pcLSTM*	23.82 (18.23)	3.72 (3.48)	72.45 (18.55)	91.20 (3.17)	6.67 (2.35)	2.13 (0.96)	86.33 (3.54)	10.44 (2.50)	3.23 (1.96)
	gpcLSTM	53.66 (22.59)	4.34 (3.83)	42.00 (22.86)	89.39 (3.91)	7.99 (2.90)	2.63 (1.12)	84.61 (3.84)	11.79 (3.20)	3.61 (2.01)
	gpcLSTM*	52.37 (22.06)	4.32 (3.15)	43.30 (22.42)	90.02 (3.69)	7.47 (2.77)	2.52 (1.04)	85.27 (3.69)	11.31 (2.95)	3.42 (2.02)
	gpcLSTM$_{AC}$	**91.17 (8.50)**	1.26 (2.08)	**7.57 (8.01)**	91.61 (2.03)	6.62 (1.39)	1.77 (0.74)	**87.97 (5.00)**	8.67 (2.64)	3.36 (2.63)
	gpcLSTM*$_{AC}$	91.02 (8.49)	1.21 (1.97)	7.77 (8.00)	91.71 (2.02)	6.55 (1.34)	1.75 (0.77)	87.95 (5.05)	8.69 (2.69)	3.36 (2.62)
	gpcLSTM*$_{APAC}$	61.30 (20.12)	2.92 (2.38)	35.79 (20.23)	90.84 (3.57)	7.04 (2.57)	2.11 (1.07)	86.48 (3.95)	10.07 (2.66)	3.45 (2.31)
Chapitre 6	**FCN #1**	35.20 (22.71)	3.42 (3.68)	61.38 (23.78)	87.09 (4.32)	10.03 (3.15)	2.88 (1.29)	80.98 (3.64)	14.37 (2.66)	4.65 (1.92)
	FCN #2	30.40 (22.45)	1.01 (1.66)	68.58 (23.09)	86.51 (4.84)	10.50 (3.65)	2.99 (1.37)	83.17 (4.16)	12.64 (3.07)	4.19 (2.24)
	FCN$_G^{\dagger}$	34.35 (29.53)	3.18 (2.86)	62.47 (30.41)	82.82 (4.95)	13.48 (3.66)	3.70 (1.39)	74.97 (2.16)	19.02 (2.09)	6.00 (1.83)
	FCN†	38.43 (28.48)	2.50 (2.33)	59.07 (29.42)	82.62 (5.86)	13.82 (4.24)	3.57 (1.71)	75.60 (3.86)	18.78 (3.07)	5.62 (1.79)
	FCN$_G^{\ddagger}$	51.01 (24.96)	3.19 (3.24)	45.80 (24.52)	82.80 (5.77)	13.57 (4.43)	3.63 (1.46)	73.69 (3.64)	20.52 (3.32)	5.79 (1.42)
	FCN‡	47.37 (27.13)	2.90 (2.93)	49.73 (27.16)	83.24 (4.92)	13.23 (3.48)	3.53 (1.53)	75.88 (3.92)	18.67 (3.28)	5.45 (1.68)
Chapitre 7	**DT**	23.67 (13.57)	3.61 (2.05)	72.72 (14.98)	80.96 (4.11)	16.60 (3.06)	2.44 (1.15)	79.51 (2.71)	17.65 (2.06)	**2.84 (1.27)**
	RF	25.51 (17.82)	1.42 (1.57)	73.07 (18.34)	86.61 (3.72)	10.82 (2.79)	2.57 (1.11)	82.53 (3.26)	13.92 (2.37)	3.55 (1.71)
	GBM	26.60 (19.79)	1.74 (1.87)	71.65 (20.89)	86.73 (3.43)	10.33 (2.49)	2.93 (1.15)	80.69 (4.16)	15.00 (2.66)	4.31 (2.01)
	LSTM‡	46.31 (24.61)	2.43 (3.62)	51.25 (25.13)	83.02 (5.57)	13.48 (4.49)	3.50 (1.28)	75.96 (4.03)	18.74 (3.38)	5.30 (1.89)
	RETAIN‡	44.08 (23.77)	2.89 (2.91)	53.03 (24.80)	84.11 (6.14)	12.57 (4.64)	3.33 (1.66)	78.81 (3.10)	16.58 (2.46)	4.61 (1.78)

Tableau A.4: Acceptabilité clinique (CG-EGA) par région moyenne (écart type) pour un horizon de prédiction de 30 minutes pour le jeu de données OhioT1DM.

Bibliographie

[1] A. Adadi and M. Berrada. Peeking inside the black-box : A survey on explainable artificial intelligence (xai). *IEEE Access*, 6 :52138–52160, 2018.

[2] M. Akbari and R. Chunara. Using contextual information to improve blood glucose prediction. *arXiv preprint arXiv :1909.01735*, 2019.

[3] E. Albaum, E. Quinn, S. Sedaghatkish, P. Singh, A. Watkins, K. Musselman, and J. Williams. Accuracy of the actigraph wgt3x-bt for step counting during inpatient spinal cord rehabilitation. *Spinal cord*, 57(7) :571, 2019.

[4] J. B. Ali, T. Hamdi, N. Fnaiech, V. Di Costanzo, F. Fnaiech, and J.-M. Ginoux. Continuous blood glucose level prediction of type 1 diabetes based on artificial neural network. *Biocybernetics and Biomedical Engineering*, 38(4) :828–840, 2018.

[5] A. Aliberti, I. Pupillo, S. Terna, E. Macii, S. Di Cataldo, E. Patti, and A. Acquaviva. A multi-patient data-driven approach to blood glucose prediction. *IEEE Access*, 7 :69311–69325, 2019.

[6] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4) :283–293, 2017.

[7] J. Ba and R. Caruana. Do deep nets really need to be deep ? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[8] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014.

[9] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan. Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging 2015 : Computer-Aided Diagnosis*, volume 9414, page 94140V. International Society for Optics and Photonics, 2015.

[10] A. Bertachi, L. Biagi, I. Contreras, N. Luo, and J. Vehí. Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks. In *KHD@ IJCAI*, pages 85–90, 2018.

[11] L. Besançon and P. Dragicevic. La différence significative entre valeurs p et intervalles de confiance (the significant difference between p-values and confidence intervals). In *Conférence Francophone sur l'Interaction Homme-Machine*, 2017.

[12] C. A. Brewer. Color use guidelines for mapping. *Visualization in modern cartography*, 1994 :123–148, 1994.

[13] R. G. Brown. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation, 2004.

[14] R. Bunescu, N. Struble, C. Marling, J. Shubrook, and F. Schwartz. Blood glucose level prediction using physiological models and support vector regression. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 135–140. IEEE, 2013.

[15] N. V. Chawla, N. Japkowicz, and A. Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1) :1–6, 2004.

[16] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141) :20170387, 2018.

[17] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain : An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[18] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou. Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE journal of biomedical and health informatics*, 21(1) :76–84, 2016.

[19] C. Cobelli, E. Renard, and B. Kovatchev. Artificial pancreas : past, present, future. *Diabetes*, 60(11) :2672–2682, 2011.

[20] I. Contreras, S. Oviedo, M. Vettoretti, R. Visentin, and J. Vehí. Personalized blood glucose prediction : A hybrid approach using grammatical evolution and physiological models. *PloS one*, 12(11) :e0187754, 2017.

[21] I. Contreras, A. Bertachi, L. Biagi, J. Vehí, and S. Oviedo. Using grammatical evolution to generate short-term blood glucose prediction models. In *KHD@ IJCAI*, pages 91–96, 2018.

[22] C. Dalla Man, R. A. Rizza, and C. Cobelli. Meal simulation model of the glucose-insulin system. *IEEE Transactions on biomedical engineering*, 54(10) :1740–1749, 2007.

[23] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, S. Garg, L. Heinemann, I. Hirsch, S. A. Amiel, et al. International consensus on use of continuous glucose monitoring. *Diabetes care*, 40(12) : 1631–1640, 2017.

[24] E. Daskalaki, K. Nørgaard, T. Züger, A. Prountzou, P. Diem, and S. Mougiakakou. An early warning system for hypoglycemic/hyperglycemic events based on fusion of adaptive prediction models. *Journal of diabetes science and technology*, 7(3) :689–698, 2013.

[25] M. De Bois. Glyfe, 2019. URL `https://github.com/dotXem/GLYFE`. doi : `10.5281/zenodo.3234605`.

[26] M. De Bois. Cg-ega python implementation, 2019. URL `https://github.com/dotXem/CG-EGA`. doi : `10.5281/zenodo.3459485`.

[27] M. De Bois. Integration of clinical criteria into the training of deep models : Application to glucose prediction for diabetic people, 2020. URL `https://github.com/dotXem/DeepClinicalGlucosePrediction`. doi : `10.5281/zenodo.3904234`.

[28] M. De Bois. Interpreting deep glucose predictive models through the retain architecture, 2020. URL `https://github.com/dotXem/DeepInterpretableGlucosePrediction`. doi : `10.5281/zenodo.3951702`.

[29] M. De Bois. Multi-source adversarial transfer learning in glucose prediction for type-2 diabetic patients, 2020. URL `https://github.com/dotXem/GlucosePredictionATL`. doi : `10.5281/zenodo.3699846`.

[30] M. De Bois and M. A. E. Yacoubi. Integration of clinical criteria into the training of deep models : Application to glucose prediction for diabetic people. *arXiv preprint arXiv :2009.10514*, 2020.

[31] M. de Bois, M. El Yacoubi, and M. Ammi. Model fusion to enhance the clinical acceptability of long-term glucose predictions. In *BIBE 2019 : 19th International Conference on Bioinformatics and Bioengineering*, pages 258–264. IEEE Computer Society, 2019.

[32] M. De Bois, M. A. El Yacoubi, and M. Ammi. Prediction-coherent lstm-based recurrent neural network for safer glucose predictions in diabetic people. In *International Conference on Neural Information Processing*, pages 510–521. Springer, 2019.

[33] M. De Bois, M. A. El Yacoubi, and M. Ammi. Study of short-term personalized glucose predictive models on type-1 diabetic children. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[34] M. De Bois, M. Ammi, and M. A. E. Yacoubi. Glyfe : Review and benchmark of personalized glucose predictive models in type-1 diabetes. *arXiv preprint arXiv :2006.15946*, 2020.

[35] M. De Bois, M. El Yacoubi, and M. Ammi. Adversarial multi-source transfer learning in healthcare : Application to glucose prediction for diabetic people. *Computer Methods and Programs in Biomedicine*, 199 :105874–105874, 2020.

[36] M. De Bois, M. A. El Yacoubi, and M. Ammi. Interpreting deep glucose predictive models for diabetic people using retain. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 685–694. Springer, 2020.

[37] M. De Bois, M. A. E. Yacoubi, and M. Ammi. Enhancing the interpretability of deep models in heathcare through attention : Application to glucose forecasting for diabetic people. *arXiv preprint arXiv :2009.10514*, 2020.

[38] M. de Zambotti, A. Goldstone, S. Claudatos, I. M. Colrain, and F. C. Baker. A validation study of fitbit charge 2^{TM} compared with polysomnography in adults. *Chronobiology international*, 35(4) :465–476, 2018.

[39] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization : Nsga-ii. In *International conference on parallel problem solving from nature*, pages 849–858. Springer, 2000.

[40] S. Del Favero, A. Facchinetti, and C. Cobelli. A glucose-specific metric to assess predictors and identify models. *IEEE transactions on biomedical engineering*, 59(5) :1281–1290, 2012.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[42] N. Dhungel, G. Carneiro, and A. P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical image analysis*, 37 :114–128, 2017.

[43] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[44] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith. Estimation of future glucose concentrations with subject-specific recursive linear models. *Diabetes technology & therapeutics*, 11(4) :243–253, 2009.

[45] M. Eren-Oruklu, A. Cinar, D. K. Rollins, and L. Quinn. Adaptive system identification for estimating future glucose concentrations and hypoglycemia alarms. *Automatica*, 48(8) :1892–1897, 2012.

[46] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3) :1, 2009.

[47] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639) :115–118, 2017.

[48] A. Facchinetti, G. Sparacino, E. Trifoglio, and C. Cobelli. A new index to optimally design and compare continuous glucose monitoring glucose prediction algorithms. *Diabetes technology & therapeutics*, 13(2) : 111–119, 2011.

[49] I. D. Federation. Atlas du diabete de la fid huitième édition 2017, 2017. URL https://www.federationdesdiabetiques.org/.

[50] I. D. Federation. Atlas du diabete de la fid neuvième édition 2019, 2019. URL https://www.federationdesdiabetiques.org/.

[51] J. Feng, K. Turksoy, and A. Cinar. Performance assessment of model-based artificial pancreas control systems. In *Prediction Methods for Blood Glucose Concentration*, pages 243–265. Springer, 2016.

[52] A. Fiat and D. Pechyony. Decision trees : More theoretical justification for practical algorithms. In *International Conference on Algorithmic Learning Theory*, pages 156–170. Springer, 2004.

[53] S. Fiorini, C. Martini, D. Malpassi, R. Cordera, D. Maggi, A. Verri, and A. Barla. Data-driven strategies for robust forecast of continuous glucose monitoring time-series. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 1680–1683. IEEE, 2017.

[54] R. D. C. France. Gluci-chek. URL https://www.accu-chek.fr/sous-sites/gluci-chek/.

[55] P. H. Franses. A note on the mean absolute scaled error. *International Journal of Forecasting*, 32(1) :20–22, 2016.

[56] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman. Predicting subcutaneous glucose concentration in humans : data-driven glucose modeling. *IEEE Transactions on Biomedical Engineering*, 56(2) :246, 2009.

[57] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1) :2096–2030, 2016.

[58] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis. A predictive model of subcutaneous glucose concentration in type 1 diabetes based on random forests. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2889–2892. IEEE, 2012.

[59] E. I. Georga, V. C. Protopappas, D. Ardigò, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis. Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. *IEEE journal of biomedical and health informatics*, 17(1) :71–81, 2013.

[60] E. I. Georga, V. C. Protopappas, D. Ardigò, D. Polyzos, and D. I. Fotiadis. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. *Diabetes technology & therapeutics*, 15(8) :634–643, 2013.

[61] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis. Online prediction of glucose concentration in type 1 diabetes using extreme learning machines. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 3262–3265. IEEE, 2015.

[62] E. I. Georga, J. C. Principe, D. Polyzos, and D. I. Fotiadis. Non-linear dynamic modeling of glucose in type 1 diabetes with kernel adaptive filters. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 5897–5900. IEEE, 2016.

[63] E. I. Georga, J. C. Príncipe, E. C. Rizos, and D. I. Fotiadis. Kernel-based adaptive learning improves accuracy of glucose predictive modelling in type 1 diabetes : A proof-of-concept study. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 2765–2768. IEEE, 2017.

[64] E. I. Georga, J. C. Príncipe, and D. I. Fotiadis. Short-term prediction of glucose in type 1 diabetes using kernel adaptive filters. *Medical & biological engineering & computing*, 57(1) :27–46, 2019.

[65] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget : Continual prediction with lstm. 1999.

[66] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv :1602.06291*, 2016.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[68] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22) :2402–2410, 2016.

[69] T. Hamdi, J. B. Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, and J.-M. Ginoux. Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. *Biocybernetics and Biomedical Engineering*, 38(2) :362–372, 2018.

[70] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech : Scaling up end-to-end speech recognition. *arXiv preprint arXiv :1412.5567*, 2014.

[71] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8) :1735–1780, 1997.

[72] A. Holzinger. Interactive machine learning for health informatics : when do we need the human-in-the-loop ? *Brain Informatics*, 3(2) :119–131, 2016.

[73] A. T. Høstmark, G. S. Ekeland, A. C. Beckstrøm, and H. D. Meen. Postprandial light physical activity blunts the blood glucose increase. *Preventive medicine*, 42(5) :369–371, 2006.

[74] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological measurement*, 25(4) :905, 2004.

[75] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine : a new learning scheme of feedforward neural networks. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pages 985–990. IEEE, 2004.

[76] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4) :679–688, 2006.

[77] M. V. Jankovic, S. Mosimann, L. Bally, C. Stettler, and S. Mougiakakou. Deep prediction model : The case of online adaptive prediction of subcutaneous glucose. pages 1–5, 2016.

[78] J. Jeon, P. J. Leimbigler, G. Baruah, M. H. Li, Y. Fossat, and A. J. Whitehead. Predicting glycaemia in type 1 diabetes patients : Experiments in feature engineering and data imputation. *Journal of Healthcare Informatics Research*, pages 1–20, 2019.

[79] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[80] D. P. Kingma and J. Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

[81] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980, 2017.

[82] K. L. Knutson. Impact of sleep and sleep loss on glucose homeostasis and appetite regulation. *Sleep medicine clinics*, 2(2) :187–197, 2007.

[83] J. Kormylo and V. Jain. Two-pass recursive digital filter with zero phase shift. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(5) :384–387, 1974.

[84] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al. Handling imbalanced datasets : A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1) :25–36, 2006.

[85] B. P. Kovatchev, L. A. Gonder-Frederick, D. J. Cox, and W. L. Clarke. Evaluating the accuracy of continuous glucose-monitoring sensors : continuous glucose–error grid analysis illustrated by therasense freestyle navigator data. *Diabetes Care*, 27(8) :1922–1928, 2004.

[86] B. P. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli. In silico preclinical trials : a proof of concept in closed-loop control of type 1 diabetes. *Journal of Diabetes Science and Technology*, 2009.

[87] B. P. Kovatchev, D. Shields, and M. Breton. Graphical and numerical evaluation of continuous glucose sensing time lag. *Diabetes technology & therapeutics*, 11(3) :139–143, 2009.

[88] A. J. Laguna Sanz, F. J. Doyle III, and E. Dassau. An enhanced model predictive control for the artificial pancreas using a confidence index based on residual analysis of past predictions. *Journal of diabetes science and technology*, 11(3) :537–544, 2017.

[89] L. L. Law, R. N. Rol, S. A. Schultz, R. J. Dougherty, D. F. Edwards, R. L. Koscik, C. L. Gallagher, C. M. Carlsson, B. B. Bendlin, H. Zetterberg, et al. Moderate intensity physical activity associates with csf biomarkers in a cohort at risk for alzheimer's disease. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, 10(C) :188–195, 2018.

[90] Q. V. Le. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8595–8598. IEEE, 2013.

[91] S. I. Lee, B. Mortazavi, H. A. Hoffman, D. S. Lu, C. Li, B. H. Paak, J. H. Garst, M. Razaghy, M. Espinal, E. Park, et al. A prediction model for functional outcomes in spinal cord disorder patients using gaussian process regression. *IEEE journal of biomedical and health informatics*, 20(1) :91–99, 2014.

[92] E. Lehmann, T. Deutsch, E. Carson, and P. Sönksen. Aida : an interactive diabetes advisor. *Computer methods and programs in biomedicine*, 41(3-4) :183–203, 1994.

[93] E. D. Lehmann and T. Deutsch. Aida freeware diabetes software simulator program of glucose - insulin action. URL `http://www.2aida.org/online/`.

[94] C. Li, C. Zhao, H. Zhao, and C. Yu. Blood glucose control based on rapid model identification with particle swarm optimization method. In *Control And Decision Conference (CCDC), 2017 29th Chinese*, pages 947–952. IEEE, 2017.

[95] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas, and P. Georgiou. Convolutional recurrent neural networks for glucose prediction. *IEEE journal of biomedical and health informatics*, 2019.

[96] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou. Glunet : A deep learning framework for accurate glucose forecasting. *IEEE journal of biomedical and health informatics*, 2019.

[97] N. Li, J. Tuo, and Y. Wang. Chaotic time series analysis approach for prediction blood glucose concentration based on echo state networks. In *2018 Chinese Control And Decision Conference (CCDC)*, pages 2017–2022. IEEE, 2018.

[98] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[99] C. Liu, J. Vehi, N. Oliver, P. Georgiou, and P. Herrero. Enhancing blood glucose prediction with meal absorption and physical exercise information. *arXiv preprint arXiv :1901.07467*, 2018.

[100] W. Liu, J. C. Principe, and S. Haykin. *Kernel adaptive filtering : a comprehensive introduction*, volume 57. John Wiley & Sons, 2011.

[101] W.-Y. Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(1) :14–23, 2011.

[102] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[103] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10) :749–760, 2018.

[104] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, page 107332, 2020.

[105] M. Macas, L. Lhotska, K. Stechova, P. Pithova, and K. Saiti. Particle swarm optimization based adaptable predictor of glycemia values. In *Cybernetics (CYBCONF), 2017 3rd IEEE International Conference on*, pages 1–6. IEEE, 2017.

[106] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli. The uva/padova type 1 diabetes simulator : new features. *Journal of diabetes science and technology*, 8(1) :26–34, 2014.

[107] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6) :369–395, 2004.

[108] C. Marling and R. Bunescu. The ohiot1dm dataset for blood glucose level prediction : Update 2020. *KHD@ IJCAI*, 2020.

[109] C. Marling and R. C. Bunescu. The ohiot1dm dataset for blood glucose level prediction. In *KHD@ IJCAI*, pages 60–63, 2018.

[110] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren. Blood glucose prediction with variance estimation using recurrent neural networks. *Journal of Healthcare Informatics Research*, pages 1–18, 2019.

[111] M. Mayo, L. Chepulis, and R. G. Paul. Glycemic-aware metrics and oversampling techniques for predicting blood glucose levels using machine learning. *Plos one*, 14(12) :e0225613, 2019.

[112] W. McKinney, J. Perktold, and S. Seabold. Time series analysis in python with statsmodels. *Jarrodmillman. Com*, pages 96–102, 2011.

[113] C. Midroni, P. J. Leimbigler, G. Baruah, M. Kolla, A. J. Whitehead, and Y. Fossat. Predicting glycemia in type 1 diabetes patients : experiments with xgboost. *heart*, 60(90) :120, 2018.

[114] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz. Using lstms to learn physiological models of blood glucose behavior. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 2887–2891. IEEE, 2017.

[115] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling. Lstms and neural attention models for blood glucose prediction : Comparative experiments on real and synthetic data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 706–712. IEEE, 2019.

[116] E. Montaser, J.-L. Díez, and J. Bondia. Stochastic seasonal models for glucose prediction in the artificial pancreas. *Journal of diabetes science and technology*, 11(6) :1124–1131, 2017.

[117] mySugr Team. mysugr.com. URL `https://mysugr.com/`.

[118] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[119] S. Oviedo, J. Vehí, R. Calm, and J. Armengol. A review of personalized blood glucose prediction strategies for t1dm patients. *International journal for numerical methods in biomedical engineering*, 33(6) :e2833, 2017.

[120] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10) :1345–1359, 2009.

[121] Y.-H. Pao, G.-H. Park, and D. J. Sobajic. Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*, 6(2) :163–180, 1994.

[122] S. M. Pappada, B. D. Cameron, P. M. Rosman, R. E. Bourey, T. J. Papadimos, W. Olorunto, and M. J. Borst. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes technology & therapeutics*, 13(2) :135–141, 2011.

[123] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[124] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn : Machine learning in python. *Journal of machine learning research*, 12(Oct) :2825–2830, 2011.

[125] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gómez, M. Rigla, A. de Leiva, and M. Hernando. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes technology & therapeutics*, 12(1) :81–88, 2010.

[126] R. Phadke, V. Prasad, H. Nagaraj, and A. Bhograj. Univariate data-driven models for glucose level prediction of cgm sensor dataset for t1dm management. *Sādhanā*, 45(1) :46, 2020.

[127] X. Ran, Z. Shan, Y. Fang, and C. Lin. An lstm-based method with attention mechanism for travel time prediction. *Sensors*, 19(4) :861, 2019.

[128] C. E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[129] Revesdiab. Revesdiab, réseau de santé diabète. URL `https://www.revesdiab.fr/`.

[130] M. P. Reymann, E. Dorschky, B. H. Groh, C. Martindale, P. Blank, and B. M. Eskofier. Blood glucose level prediction based on support vector regression using mobile platforms. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 2990–2993. IEEE, 2016.

[131] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088) :533–536, 1986.

[132] O. Sagi and L. Rokach. Ensemble learning : A survey. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 8(4) :e1249, 2018.

[133] A. Samat, P. Du, S. Liu, J. Li, and L. Cheng. e^2lms : Ensemble extreme learning machines for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7 (4) :1060–1069, 2014.

[134] W. Sandham, D. Nikoletou, D. Hamilton, K. Paterson, A. Japp, and C. MacGregor. Blood glucose prediction for diabetes therapy using a recurrent artificial neural network. In *Signal Processing Conference (EUSIPCO 1998), 9th European*, pages 1–4. IEEE, 1998.

[135] B. Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

[136] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[137] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks : Visualising image classification models and saliency maps. *arXiv preprint arXiv :1312.6034*, 2013.

[138] M. M. Skaff, J. T. Mullan, D. M. Almeida, L. Hoffman, U. Masharani, D. Mohr, and L. Fisher. Daily negative mood affects fasting glucose in type 2 diabetes. *Health Psychology*, 28(3) :265, 2009.

[139] G. Sparacino, F. Zanderigo, A. Maran, and C. Cobelli. Continuous glucose monitoring and hypo/hyperglycaemia prediction. *Diabetes Research and Clinical Practice*, 74 :S160–S163, 2006.

[140] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Transactions on biomedical engineering*, 54(5) :931–937, 2007.

[141] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1) :1929–1958, 2014.

[142] F. Ståhl and R. Johansson. Diabetes mellitus modeling and short-term prediction based on blood glucose measurements. *Mathematical biosciences*, 217(2) :101–117, 2009.

[143] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou. Predicting blood glucose with an lstm and bi-lstm based deep neural network. In *2018 14th Symposium on Neural Networks and Applications (NEUREL)*, pages 1–5, Nov 2018. doi : 10.1109/NEUREL.2018.8586990.

[144] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv :1703.01365*, 2017.

[145] J. W. Taylor and P. E. McSharry. Short-term load forecasting methods : An evaluation based on european data. *IEEE Transactions on Power Systems*, 22(4) :2213–2219, 2007.

[146] J. M. Tomczak. Gaussian process regression with categorical inputs for predicting the blood glucose level. pages 98–108, 2016.

[147] K. Turksoy, S. Samadi, J. Feng, E. Littlejohn, L. Quinn, and A. Cinar. Meal detection in patients with type 1 diabetes : a new module for the multivariable adaptive artificial pancreas control system. *IEEE journal of biomedical and health informatics*, 20(1) :47–54, 2016.

[148] J. J. Valletta, A. J. Chipperfield, and C. D. Byrne. Gaussian process modelling of blood glucose response to free-living physical activity data in people with type 1 diabetes. Institute of Electrical and Electronics Engineers, 2009.

[149] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[150] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi. Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. *Health informatics journal*, page 1460458219850682, 2019.

[151] R. Visentin, E. Campos-Náñez, M. Schiavon, D. Lv, M. Vettoretti, M. Breton, B. P. Kovatchev, C. Dalla Man, and C. Cobelli. The uva/padova type 1 diabetes simulator goes from single meal to single day. *Journal of diabetes science and technology*, 12(2) :273–281, 2018.

[152] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv :1606.05718*, 2016.

[153] F. Wang, L. P. Casalino, and D. Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3) :293–294, 2019.

[154] Wang, Qian, P. Molenaar, S. Harsh, K. Freeman, J. Xie, C. Gold, M. Rovine, and J. Ulbrecht. Personalized state-space modeling of glucose dynamics for type 1 diabetes using continuously monitored glucose, insulin dose, and meal intake : An extended kalman filter approach. *Journal of diabetes science and technology*, 8 (2) :331–345, 2014.

[155] M. E. Whelan, A. P. Kingsnorth, M. W. Orme, L. B. Sherar, and D. W. Esliger. Sensing interstitial glucose to nudge active lifestyles (signal) : feasibility of combining novel self-monitoring technologies for persuasive behaviour change. *BMJ open*, 7(10), 2017.

[156] A. Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, and G. Hartvigsen. Data-driven modeling and prediction of blood glucose dynamics : Machine learning applications in type 1 diabetes. *Artificial intelligence in medicine*, 98 :109–134, 2019.

[157] R. R. Yager. Time series smoothing and owa aggregation. *IEEE Transactions on Fuzzy Systems*, 16(4) : 994–1007, 2008.

[158] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks ? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[159] X. Yu, M. Rashid, J. Feng, N. Hobbs, I. Hajizadeh, S. Samadi, M. Sevil, C. Lazaro, Z. Maloney, E. Littlejohn, et al. Online glucose prediction using computationally efficient sparse kernel filtering algorithms in type-1 diabetes. *IEEE Transactions on Control Systems Technology*, (99) :1–13, 2018.

[160] X. Yu, K. Turksoy, M. Rashid, J. Feng, N. Hobbs, I. Hajizadeh, S. Samadi, M. Sevil, C. Lazaro, Z. Maloney, et al. Model-fusion-based online glucose concentration predictions in people with type 1 diabetes. *Control engineering practice*, 71 :129–141, 2018.

[161] K. Zarkogianni, A. Vazeou, S. G. Mougiakakou, A. Prountzou, and K. S. Nikita. An insulin infusion advisory system based on autotuning nonlinear model-predictive control. *IEEE Transactions on Biomedical Engineering*, 58(9) :2467–2477, 2011.

[162] K. Zarkogianni, E. Litsa, K. Mitsis, P.-Y. Wu, C. D. Kaddi, C.-W. Cheng, M. D. Wang, and K. S. Nikita. A review of emerging technologies for the management of diabetes mellitus. *IEEE Transactions on Biomedical Engineering*, 62(12) :2735–2749, 2015.

[163] K. Zarkogianni, K. Mitsis, E. Litsa, M.-T. Arredondo, G. Fico, A. Fioravanti, and K. S. Nikita. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Medical & biological engineering & computing*, 53(12) :1333–1343, 2015.

[164] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli. Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. *IEEE Transactions on Biomedical Engineering*, 59(6) :1550–1560, 2012.

[165] C. Zecchin, A. Facchinetti, G. Sparacino, C. Dalla Man, C. Manohar, J. A. Levine, A. Basu, Y. C. Kudva, and C. Cobelli. Physical activity measured by physical activity monitoring system correlates with glucose trends reconstructed from continuous glucose monitoring. *Diabetes technology & therapeutics*, 15(10) :836–844, 2013.

[166] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli. Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information. *Computer methods and programs in biomedicine*, 113(1) :144–152, 2014.

[167] C. Zhao, E. Dassau, L. Jovanovič, H. C. Zisser, F. J. Doyle III, and D. E. Seborg. Predicting subcutaneous glucose concentration using a latent-variable-based statistical method for type 1 diabetes mellitus. *Journal of diabetes science and technology*, 6(3) :617–633, 2012.

[168] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou. A deep learning algorithm for personalized blood glucose prediction. In *KHD@ IJCAI*, pages 64–78, 2018.

[169] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *Journal of Healthcare Informatics Research*, pages 1–17, 2020.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage profond sous contraintes biomédicales pour la prédiction de la glycémie future de patients diabétiques

Mots clés : Apprentissage profond, Diabète, Prédiction de glycémie, Apprentissage par transfert, Acceptabilité clinique, Interprétabilité

Résumé : Malgré ses récents succès en vision assistée par ordinateur ou en traduction automatique, l'utilisation de l'apprentissage profond dans le secteur biomédical fait face à de nombreux challenges. Parmi eux, nous comptons l'accès difficile à des données en quantité et qualité suffisantes, ainsi que le besoin en l'interopérabilité et en l'interprétabilité des modèles. Dans cette thèse, nous nous intéressons à ces différentes problématiques à la lueur de la création de modèles prédisant la glycémie future de patients diabétiques. De tels modèles permettraient aux patients d'anticiper les variations de leur glycémie au quotidien, les aidant ainsi à mieux la réguler afin d'éviter les états d'hypoglycémie et d'hyperglycémie.

Pour cela, nous utilisons trois ensembles de données. Tandis que le premier a été récolté à l'occasion de cette thèse sur plusieurs patients diabétiques de type 2, les deux autres sont composés de patients diabétiques de type 1, à la fois réels et virtuels. Dans l'ensemble des études, nous utilisons les données passées de glycémie, d'insuline et de glucides de chaque patient pour construire des modèles personnalisés prédisant la glycémie du patient 30 minutes dans le futur.

Dans un premier temps, nous faisons une analyse détaillée de l'état de l'art en construisant une base de résultats de référence open source de modèles prédictifs de glycémie. Bien que prometteurs, nous mettons en évidence la difficulté qu'ont les modèles profonds à effectuer des prédictions qui soient à la fois précises et sans danger pour le patient.

Afin d'améliorer l'acceptabilité clinique des modèles, nous proposons d'intégrer des contraintes cliniques au sein de l'apprentissage des modèles. À cet effet, nous proposons de nouvelles fonctions de coût permettant d'améliorer la cohérence des prédictions et de se focaliser davantage sur les erreurs de prédictions cliniquement dangereuses. Nous explorons son utilisation pratique à travers un algorithme permettant d'obtenir un modèle maximisant la précision des prédictions tout en respectant des contraintes cliniques fixées au préalable.

Puis, nous étudions la piste de l'apprentissage par transfert pour améliorer les performances des modèles prédictifs de glycémie. Celui-ci permet de faciliter l'apprentissage des modèles personnalisés aux patients en réutilisant les connaissances apprises sur d'autres patients. En particulier nous proposons le cadre de l'apprentissage par transfert multi-sources adverse. Celui-ci permet de significativement améliorer les performances des modèles en permettant l'apprentissage de connaissances a priori qui sont plus générales, car agnostiques des patients sources du transfert. Nous investiguons différents scénarios de transfert à travers l'utilisation de nos trois jeux de données. Nous montrons qu'il est possible d'effectuer un transfert de connaissance à partir de données provenant de dispositifs expérimentaux différents, de patients de types de diabète différents, mais aussi à partir de patients virtuels.

Enfin, nous nous intéressons à l'amélioration de l'interprétabilité des modèles profonds à travers le principe d'attention. En particulier, nous explorons l'utilisation d'un modèle profond et interprétable pour la prédiction de la glycémie. Celui-ci implémente un double mécanisme d'attention lui permettant d'estimer la contribution de chaque variable en entrée au modèle à la prédiction finale. Nous montrons empiriquement l'intérêt d'un tel modèle pour la prédiction de glycémie en analysant son comportement dans le calcul de ses prédictions.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Deep Learning under Biomedical Constraints for the Forecasting of Glucose of Diabetic Patients

Keywords: Deep learning, Diabetes, Glucose prediction, Transfer learning, Clinical acceptability, Interpretability

Abstract: Despite its recent successes in computer vision or machine translation, the use of deep learning in the biomedical field faces many challenges. Among them, we have the difficult access to data in sufficient quantity and quality, as well as the need of having interoperable and the interpretable models. In this thesis, we are interested in these different issues from the perspective of the creation of models predicting future glucose values of diabetic patients. Such models would allow patients to anticipate daily glucose variations, helping its regulation in order to avoid states of hypoglycemia or hyperglycemia.

To this end, we use three datasets. While the first was collected during this thesis on several type-2 diabetic patients, the other two are composed of type-1 diabetic patients, both real and virtual. Across the studies, we use each patient's past glucose, insulin, and carbohydrate data to build personalized models that predict the patient's glucose values 30 minutes into the future.

First, we do a detailed state-of-the-art analysis by building an open-source benchmark of glucose-predictive models. While promising, we highlight the difficulty deep models have in making predictions that are at the same time accurate and safe for the patient.

In order to improve the clinical acceptability of the models, we investigate the integration of clinical constraints within the training of the models. We propose new cost functions enhancing the coherence of successive predictions. In addition, they enable the training to focus on clinically dangerous errors. We explore its practical use through an algorithm that enables the training of a model maximizing the precision of the predictions while respecting the clinical constraints set beforehand.

Then, we study the use of transfer learning to improve the performance of glucose-predictive models. It eases the learning of personalized models by reusing the knowledge learned on other patients. In particular, we propose the adversarial multi-source transfer learning framework. It significantly improves the performance of the models by allowing the learning of a priori knowledge which is more general, by being agnostic of the patients that are the source of the transfer. We investigate different transfer scenarios through the use of our three datasets. We show that it is possible to transfer knowledge using data coming from different experimental devices, from patients of different types of diabetes, but also from virtual patients.

Finally, we are interested in improving the interpretability of deep models through the attention mechanism. In particular, we explore the use of a deep and interpretable model for the prediction of glucose. It implements a double attention mechanism enabling the estimation of the contribution of each input variable to the model to the final prediction. We empirically show the value of such a model for the prediction of glucose by analyzing its behavior in the computation of its predictions.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France