République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche scientifique

Université Dr Tahar Moulay de Saida

Faculté de Technologie

Département d'électronique

Mémoire de fin d'étude pour l'obtention d'un diplôme de Master

Spécialité : Génie Biomédical

Option : Instrumentation Biomédicale

Thème

Surveillance non-invasive de la glycémie utilisant des méthodes d'apprentissage automatique

---

Présenté par **-MEJDOUB Fatima Zohra**

-MEBARKI Marwa Manel

Membres de jury

Dr. DINE Khaled	President
Pr. DJELLOULI Bouazza	Examinateur
Pr. DAHANI Ameur	Encadreur

Année Universitaire 2024-2025

Remerciement

Il est difficile de trouver les mots justes pour exprimer toute la gratitude que je ressens aujourd'hui. Ce travail est l'aboutissement d'un parcours jalonné de doutes. D'efforts, mais surtout de belles rencontres et de soutiens inestimables.

À mes parents, vous qui avez toujours cru en moi, même lorsque la fatigue prenait le dessus. Merci pour votre amour inconditionnel, vos encouragements silencieux et vos sacrifices que je mesure un peu plus chaque jour. Votre présence a été ma plus grande force.

*Au **Pr. DAHANI Ameur**, je tiens à exprimer ma reconnaissance pour votre patience, votre écoute et vos conseils avisés. Votre accompagnement bienveillant a été essentiel dans la réalisation de ce travail et m'a permis de grandir autant sur le plan académique que personnel.*

Je remercie également ma famille et mes proches, pour leurs mots rassurants, leurs gestes réconfortants et leur indéfectible soutien dans les moments de Doute. Vous avez su rendre ce parcours plus doux.

Enfin, une pensée sincère a toutes les personnes qui, de près ou de loin, ont croisé mon chemin durant cette aventure. Chacun de vous a contribué, d'une manière ou d'une autre, l'accomplissement de ce projet.

Ce mémoire n'est pas seulement le fruit d'un travail personnel, mais aussi le reflet de tout l'amour, la confiance et l'accompagnement que j'ai reçus

Merci, du fond du cœur.

Dédicace

À ma chance, la lumière qui brille ma vie ma très chère mère quoique je fasse ou que je dise, je ne saurai point te remercier comme il se doit, ton affection me couvre, ta bienveillance me guide et ta présence à mes côtés a toujours été une source de force pour affronter les différents obstacles.

À ma vie et ma raison de vivre, mon très cher père qui a toujours été à mes côtés pour me soutenir et m'encourager, l'éducation que tu nous as donnée aussi est irréprochable. Que ce travail traduit la gratitude et mon affection.

À mes très chères sœurs ***SARA, HADJER*** *et mes frère* ***AYMEN NADIR*** *pour votre soutien qui a été d'une grande aide.*
À mon joli cadeau de dieu, ce qui est devenu la joie de notre maison ***ISAAC***

À mon âme sœur ***SOUNDOUSS*** *que je l'aime trop et compte beaucoup pour moi*

A *des personnes qui avec j'ai partagé mes moments de bonheur, de tristesse* ***INES, AMINA***
P*uisse Allah vous donner santé, bonheur et réussite*

Mebarki Marwa Manel

Dédicace

A mon père bien-aimé,

Aujourd'hui, je termine cette étape importante de ma vie, et je me rends compte que je ne serais pas ici sans toi. Même si tu n'es plus parmi nous, tu as toujours été présent dans mon cœur et dans mon esprit.

À ma chère mère,

Tu as été mon roc, ma force et ma lumière tout au long de ce voyage. Tu as été mon inspiration et ma source de soutien constant. Cette réussite est aussi la tienne, car tu as joué un rôle essentiel dans ma formation et mon épanouissement.

À mes chères sœurs et mon frère,

Je suis si reconnaissante d'avoir pu compter sur vous tout au long de mes études. Votre soutien, votre aide et vos encouragements m'ont donné la force de continuer lorsque les défis semblaient insurmontables.

À ma famille bien-aimée,

Je tiens à vous exprimer toute ma gratitude et mon amour profond pour chacun d'entre vous. Vous avez été mes piliers de soutien tout au long de mon parcours d'études.

À ma meilleure personne

A ma personne préférée, qui a été un soutien et a partagé ma fatigue du début jusqu'à la fin, qui a été ma force dans mes moments de faiblesse et mon encouragement malgré tout.

Mejdoub Fatima Zohra

Résumé

Cette étude présente le développement d'un système d'apprentissage automatique pour la prédiction non invasive de la glycémie, en exploitant le signal physiologique ECG et les paramètres cliniques collectés. Les signaux vitaux et les données cliniques de cette base de données sont recueillis à l'aide de l'appareil HealthyPi v3. Après prétraitement, une segmentation du signal a permis de localiser les ondes caractéristiques du cœur (P, Q, R, S, T), essentielles à l'extraction des caractéristiques physiologiques.

L'ensemble des caractéristiques extraites du signal ECG, combinées aux caractéristiques cliniques, a servi de variables indépendantes pour l'entraînement de trois algorithmes de régression supervisée : K-Nearest Neighbors (KNN), Decision Tree et Random Forest. Les performances des modèles ont été évaluées à l'aide des métriques MAE, MSE, RMSE et du coefficient de détermination R^2.

Les résultats obtenus montrent que le modèle Random Forest offre la meilleure précision par rapport aux autres modèles étudiés, avec un $R^2 = -0{,}015$. Les performances sont moins précises pour le modèle KNN, et plus faibles encore pour le modèle Decision Tree, ce qui suggère une sensibilité accrue à la sélection des hyperparamètres et des caractéristiques, et souligne la nécessité d'une base de données plus large, d'un nettoyage plus rigoureux des données, ainsi que l'exploration d'algorithmes plus avancés, notamment ceux issus de l'apprentissage profond.

Ce travail démontre la faisabilité d'une approche non invasive de la prédiction de la glycémie à partir du signal ECG, tout en mettant en évidence les défis à relever en termes de précision et de généralisation.

Mots-clés : Glycémie non invasive, ECG, apprentissage automatique, K-Nearest Neighbors, Random Forest, Decision Tree

Abstract

This study presents the development of a machine learning system for the non-invasive prediction of blood glucose by leveraging the physiological ECG signal and collected clinical parameters. The vital signs and clinical data used in this dataset were acquired using the HealthyPi v3 device. After preprocessing, signal segmentation enabled the localization of characteristic cardiac waves (P, Q, R, S, T), which are essential for extracting physiological features.

The set of features extracted from the ECG signal, combined with clinical parameters, served as independent variables for training three supervised regression algorithms: K-Nearest Neighbors (KNN), Decision Tree, and Random Forest. Model performances were evaluated using MAE, MSE, RMSE, and the coefficient of determination (R^2).

The results indicate that the Random Forest model offers the highest accuracy among the models tested, with an $R^2 = -0.015$. Performance was less accurate for KNN and weaker for Decision Tree, suggesting a high sensitivity to hyperparameter and feature selection. This highlights the need for a larger dataset, more thorough data cleaning, and potentially the use of more advanced algorithms such as deep learning methods.

This work demonstrates the feasibility of a non-invasive approach to blood glucose prediction using ECG, while also emphasizing the challenges related to accuracy and generalizability.

Keywords: Non-invasive blood glucose, ECG, machine learning, K-Nearest Neighbors, Random Forest, Decision Tree

الملخص

تقدم هذه الدراسة تطوير نظام التعلم الآلي للتنبؤ بمستوى الجلوكوز في الدم بطريقة غير جراحية، من خلال استغلال إشارة تخطيط القلب الفسيولوجية المجمعة والمعايير السريرية. يتم جمع العلامات الحيوية والبيانات السريرية في هذه القاعدة البيانات باستخدام جهاز HealthyPi v3. بعد المعالجة المسبقة، أتاح تقسيم الإشارة تحديد الموجات المميزة للقلب (P، Q، R، S، T)، وهي ضرورية لاستخراج الخصائص الفسيولوجية.

تم استخدام مجموعة الميزات المستخرجة من إشارة تخطيط القلب، جنبًا إلى جنب مع الميزات السريرية، كمتغيرات مستقلة لتدريب ثلاث خوارزميات انحدار خاضعة للإشراف: أقرب جيران KNN، وشجرة القرار، والغابة العشوائية. تم تقييم أداء النموذج باستخدام مقاييس MAE وMSE وRMSE ومعامل التحديد R^2.

وتظهر النتائج التي تم الحصول عليها أن نموذج الغابة العشوائية يقدم أفضل دقة مقارنة بالنماذج الأخرى المدروسة، مع R^2 = -0.015. الأداء أقل دقة بالنسبة لنموذج KNN، وأقل دقة بالنسبة لنموذج شجرة القرار، مما يشير إلى زيادة الحساسية لاختيار المعلمات الفائقة والميزات، وتسليط الضوء على الحاجة إلى قاعدة بيانات أكبر، وتنظيف البيانات بشكل أكثر صرامة، واستكشاف الخوارزميات الأكثر تقدمًا، بما في ذلك تلك المستمدة من التعلم العميق.

يوضح هذا العمل جدوى النهج غير الجراحي للتنبؤ بمستوى السكر في الدم من إشارة تخطيط كهربية القلب، مع تسليط الضوء على التحديات من حيث الدقة والتعميم.

الكلمات المفتاحية: قياس نسبة الجلوكوز في الدم غير الجراحي، تخطيط كهربية القلب، التعلم الآلي، أقرب جيران K، الغابة العشوائية، شجرة القرار

Sommaire

LISTE DES FEGURES

- ***chapitre I***

- **chapitre II**

- **chapitre III**

Introduction Générale

Le diabète est devenu un terme très familier dans la société actuelle et constitue un problème majeur de la santé publique aussi bien dans les pays développés que dans les pays en développement. Il s'agit d'une affection chronique causée par une élévation du taux de sucre dans le sang, principalement due soit à une production insuffisante ou nulle d'insuline par pancréas, soit à une résistance des cellules cible à l'insuline produite. Le glucose provenant des aliments peut être absorbé dans la circulation sanguine grâce à la sécrétion de l'hormone insuline par le pancréas. Le diabète est donc une condition dans laquelle la production d'insuline est insuffisante en raison d'un dysfonctionnement pancréatique.

Un diabète non contrôlé entraîne fréquemment une hyperglycémie, qui, avec le temps, endommage gravement de nombreux organes aboutissant à des complications sévères telles que le coma, l'insuffisance rénale et rétinienne, les dysfonctionnements cardiovasculaires et cérébro-vasculaires, les troubles vasculaires périphériques, ainsi que des effets pathogènes sur le système immunitaire.

Le nombre de personnes atteintes de diabète a augmenté dans le monde au cours des dix dernières années. On compte plus de 200 millions de personnes touchées, avec une augmentation annuelle de 7 % de la prévalence du diabète dans le monde . En 2017, on dénombrait 425 millions de diabétiques dans le monde, et une enquête menée par la Fédération Internationale du Diabète en 2017 prévoyait que ce nombre atteindra 625 millions d'ici l'an 2045.

La gestion efficace du diabète repose sur une surveillance régulière de la glycémie. Cependant, les méthodes conventionnelles, souvent utilisées, impliquent des prélèvements sanguins fréquents, ce qui peut être inconfortable et dissuasif pour les patients. Ces limitations ont stimulé le développement de méthodes non invasives, notamment basées sur les biocapteurs optiques et les signaux physiologiques, comme le photopléthysmogramme (PPG), l'électrocardiogramme (ECG), …etc.

Les biocapteurs ECG exploitent l'activité électrique du cœur pour enregistrer des signaux électrocardiographiques riches en informations physiologiques. Ces signaux permettent d'extraire des caractéristiques précieuses, telles que les intervalles PR, QRS et QT, ainsi que la variabilité de la fréquence cardiaque, qui peuvent être corrélées à divers paramètres biologiques, y compris les fluctuations du taux de glucose sanguin. Toutefois, l'analyse de ces données brutes requiert des techniques avancées de traitement du signal et d'apprentissage automatique afin de transformer les signaux complexes en estimations fiables des niveaux de glucose.

Ces derniers temps, de grands progrès et innovations technologiques ont marqué le domaine de l'intelligence artificielle (IA). Celle-ci a permis le développement de nouvelles méthodes non invasives, notamment basées sur les biocapteurs optiques et les signaux physiologiques tels que le photopléthysmogramme (PPG) et l'électrocardiogramme (ECG), entre autres. L'intégration de l'IA dans les technologies de biocapteurs ouvre la voie à une surveillance non invasive, en temps réel et continue de la glycémie. Ces innovations améliorent non seulement le confort et l'adhésion des patients, mais représentent également une avancée majeure dans le domaine de la santé personnalisée, rendant la gestion du diabète plus accessible et efficace.

Par conséquent, les applications des méthodes d'apprentissage automatique dans les capteurs optiques ont gagné en importance ces dernières années, en particulier dans la surveillance et l'amélioration de la précision de détection des capteurs pour des performances améliorées.

L'apprentissage automatique (machine learning) permet de prédire et de détecter le diabète à un stade précoce. L'apprentissage automatique est une branche de l'intelligence artificielle qui utilise l'analyse statistique et qui est reconnue comme un domaine prometteur, capable de contribuer à la classification des patients (diabétiques ou non diabétiques) ou à la prédiction de leurs glycémies sanguines à partir d'un ensemble de données d'apprentissage. Le principal avantage de ces méthodes réside dans la capacité des algorithmes à apprendre à partir des données physiologiques et des paramètres cliniques et d'appliquer cet apprentissage pour l'extraction des caractéristiques et des motifs permettant des prédictions futures de la glycémie .

Ces dernières années, la prédiction faite par des modèles de l'apprentissage automatique présente des performances de précision qui surpassent la majorité des modèles classiques de prédiction. Dans le cadre de cette problématique, on expose dans chapitre une étude détaillée de l'utilisation du signal physiologique ECG ainsi que certains paramètres cliniques pour la prédiction de la glycémie à l'aide d'algorithmes d'apprentissage automatique. On a utilisé les architectures neuronales suivantes, KNN (K-Plus Proches Voisins, en anglais K-Nearest Neiboghrs), RF (Foret aléatoire, en anglais Random Forest) et DT (Arbre de Décision, en anglais Random Forest), dans un modèle de capteurs de signaux physiologiques pour la prédiction de la glycémie. Pour l'entraînement de notre système, une base de données composée de 100 patients a été utilisée. Les signaux physiologiques et les paramètres cliniques de cette base ont été collectés à l'aide du moniteur HealthyPi v3, auprès de patients suivis à la maison du diabète de la daïra de Saïda. Le traitement des signaux ECG et le développement des algorithmes

d'apprentissage automatique ont été réalisés dans l'environnement Jupyter Notebook, en s'appuyant sur la plateforme Anaconda pour la gestion des dépendances et des bibliothèques Python.

L'évaluation des performances du système a été réalisée à l'aide de plusieurs indicateurs tels que la fonction de perte (Loss) pour mesurer l'erreur globale du modèle, L'erreur moyenne absolue MAE (Mean Absolute Error) pour quantifier l'erreur moyenne absolue, l'erreur quadratique moyenne MSE (Mean Squared Error) pour mesurer l'erreur quadratique moyenne, la racine carrée de l'erreur quadratique moyenne RMSE (Root Mean Squared Error) pour évaluer la racine carrée de l'erreur quadratique moyenne, et R^2 (coefficient de détermination) pour estimer la qualité d'ajustement du modèle aux données.

Afin d'aboutir à notre objectif de prédiction de la glycémie en utilisant les méthodes d'apprentissage automatique et l'exploitation du signal physiologique ECG, ce manuscrit est structuré en trois chapitres. Le premier chapitre présente des généralités sur le signal ECG ainsi que les méthodes utilisées pour son traitement. Le deuxième chapitre est consacré aux différentes techniques d'apprentissage automatique appliquées à la prédiction de la glycémie. Enfin, le troisième chapitre décrit en détail la méthodologie mise en œuvre, les résultats obtenus, leur interprétation, ainsi qu'une comparaison des modèles testés.

Ce manuscrit se termine par une conclusion générale synthétisant les résultats, identifiant les défis rencontrés et proposant des pistes d'amélioration et des perspectives de recherche.

Chapitre 01

Généralités sur l'électrocardiogramme

I. Introduction

L'ECG (électrocardiogramme) enregistre l'activité électrique du cœur, qui est contrôlée par des impulsions provenant du nœud sinusal dans l'oreillette droite. Il permet de détecter des anomalies de la conduction cardiaque, telles que des rythmes cardiaques anormaux (trop rapides, lents ou irréguliers).

L'ECG peut également révéler des signes de dommages au muscle cardiaque, causés par des affections comme l'infarctus, l'hypertension artérielle ou l'embolie pulmonaire. Cependant, il ne mesure pas la capacité du cœur à pomper.

II. Anatomie du cœur

Le cœur est un organe essentiellement musculaire tapissé en dedans par l'endocarde qui se continue par l'endothélium vasculaire.

Il est recouvert à sa surface par le péricarde viscéral ou épicarde.Son anatomie est divisée en deux côtés "en miroir", gauche et droite, qui soutiennent de différents systèmes circulatoires mais qui pompent de manière synchronisée et rythmée. Chaque côté du cœur se compose de deux chambres, l'oreillette où le sang entre et le ventricule où le sang est forcé à circuler plus loin.

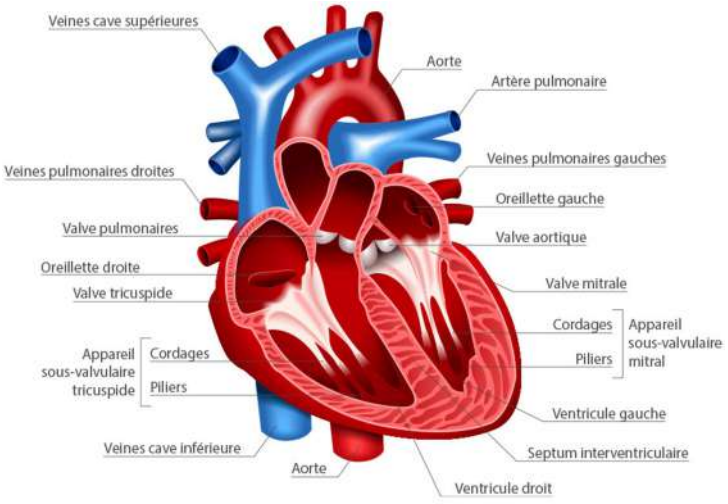

Fig. I.1: Anatomie du cœur

II.1 Activité mécanique cardiaque

Ces étapes comprennent la systole auriculaire, la systole ventriculaire et la diastole. Lors de la systole auriculaire, les oreillettes se contractent en propulsant le sang dans les ventricules. Après que le sang a été libéré, les valvules auriculo-ventriculaires, les disques de tissu fibreux entre les oreillettes et les ventricules, se ferment pour empêcher tout reflux du sang aux oreillettes.

Quant à la systole ventriculaire, elle concerne la contractilité des ventricules qui éjectent le sang hors du cœur pour le propulser dans le système circulatoire. Après cela, les deux valvules, soit pulmonaire côté droit et aortique côté gauche, se referment. Enfin, la diastole représente le repos simultané de toutes les parties du cœur, ce qui amène à l'emboitement passif des ventricules et l'approvisionnement sang nouveau.

Enfin, la diastole signifie la relaxation de toutes les parties du cœur et fournit le remplissage passif des ventricules et l'arrivée de nouveau sang. Les phases de contractions ordonnées des oreillettes et des ventricules sont régies par la propagation d'une impulsion électrique. De cette manière, lors de la modification de la fréquence du battement du cœur, la diastole est ou réduite ou étendue, tandis que la durée de la systole reste relativement constante.

II.2 Activité électrique cardiaque

Comme tous les muscles du corps, la contraction du muscle cardiaque est causée par l'impulsions électriques voyagent le long des fibres du myocarde polarisation les cellules musculaires. En fait, le cœur contient des cellules conductrices Cellules qui génèrent et propagent des impulsions électriques et répondent à ces impulsions lors d'une activité cardiaque normale, la stimulation. L'électricité myocardique est émise par le nœud sino-auriculaire.

L'impulsion électrique permet une courte pause afin que le sang puisse entrer dans le ventricule. Ensuite, il lui prêtera un plat, composé de deux branches principales. Chacun va dans un ventricule. Les fibres qui composent ce faisceau étalent l'impulsion électrique à plusieurs points du ventricule, permettant ainsi une dépolarisation immédiate de tout le muscle ventriculaire. Ensuite, les fibres musculaires reviennent à nouveau à leur état initial.

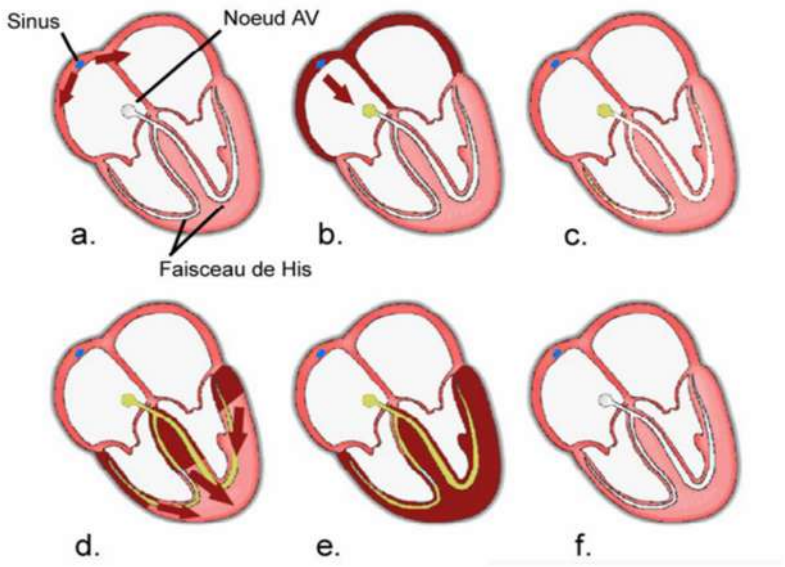

Fig. I.2: Propagation de l'impulsion électrique dans le muscle cardiaque

III. L'électrocardiographie

III.1 Principe

L'activité électrique du cœur peut être mesurée à la surface du corps en fixant un ensemble d'électrodes à la peau. Les électrodes doivent être disposées de manière que le changement spatio-temporel du champ électrique cardiaque soit suffisamment bien réfléchi. Pour un enregistrement ECG, qui détermine la différence de tension entre une paire d'électrodes, on appelle cela une dérivation.

Une convention internationale a standardisé l'ECG à 12 dérivations, comprenant les six dérivations frontales et les six précordiales. Les dérivations standard offrent divers sites d'observation qui fournissent une perspective tridimensionnelle de l'activité électrique du cœur. On note 3 types de dérivations

III.1.1 Dérivations bipolaires

Les trois dérivations des membres bipolaires sont notées I, II et III et sont obtenues en mesurant la différence de tension entre le bras gauche, le bras droit et la gauche jambe dans les combinaisons suivantes :

- D_I avec DI = aVL – aVR
- D_{II} avec DII =a VF – aVR
- D_{III} avec DIII = aVF – aVL

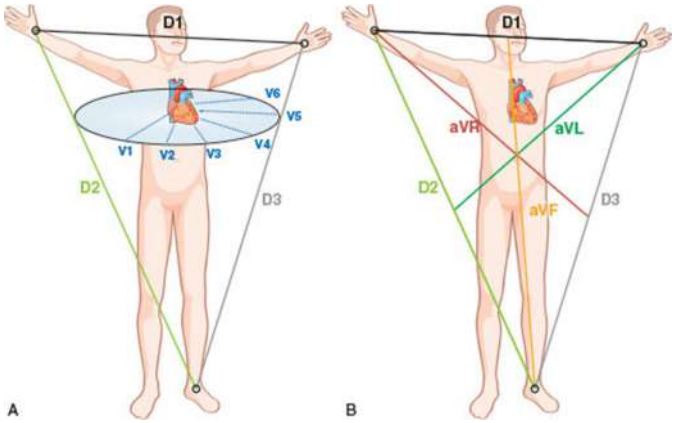

Fig. I.3 : Dérivations bipolaires

III.1.2 Dérivations unipolaires

Connues sous le nom de dérivations de Goldberger, elles font appel aux mêmes électrodes que celles utilisées par Einthoven. Chaque électrode est considérée comme positive, tandis que les deux autres sont considérées comme des références négatives.

- AVR : mesure unipolaire sur bras droit.
- AVL : mesure unipolaire sur bras gauche.
- AVF : mesure unipolaire sur jambe gauche.

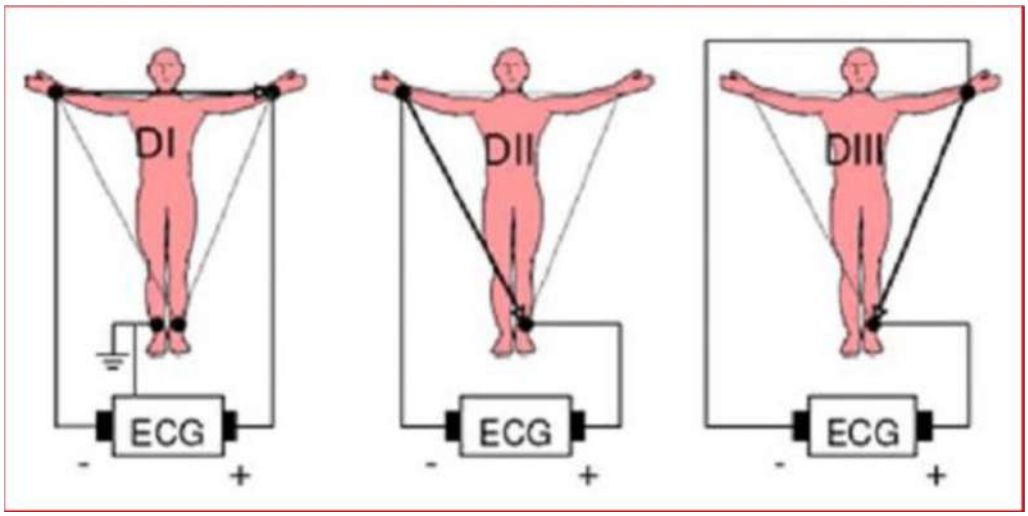

Fig. I.4: Dérivations unipolaires

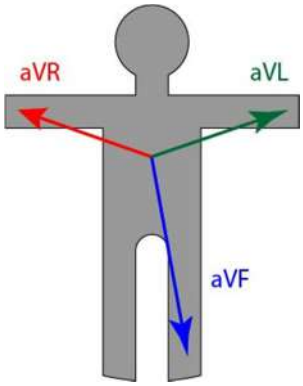

Fig. I.5: Dérivations unipolaires

III.1.3 Dérivations précordiales

Les dérivations précordiales sont placées successivement sur les flancs antérieur et gauche du thorax pour offrir une perspective plus précise du cœur par rapport aux dérivations des membres.

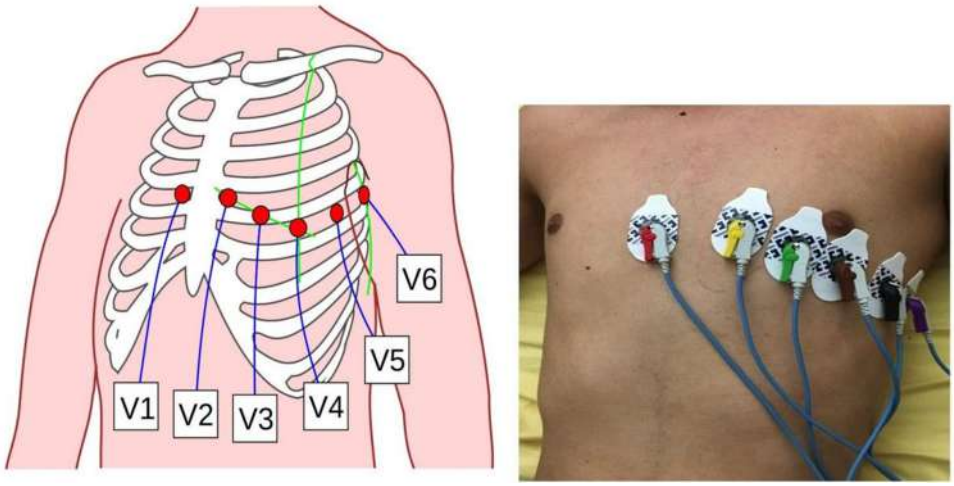

Fig. I.6: Dérivation précordiales

III.2 Le signale électrocardiogramme

L'activité électrique du cœur s'enregistre par le signal électrocardiographe ECG. Il s'agit du signal électrophysiologique en série d'ondes électriques de formes et de durées particulières que se répètent ces ondes à chaque cycle cardiaque. Il traduit les différents phénomènes mécaniques et électriques relatifs au parcours du potentiel d'action, dont les étapes sont successives. Cet enregistrement est illustré dans la figure ci-dessous :

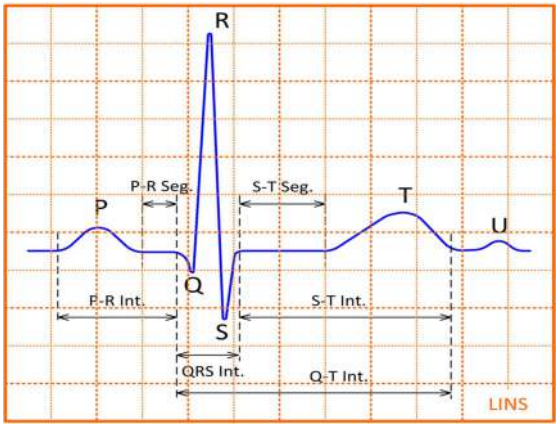

Fig. I.7: ondes du signale ECG

III.3 Analyse de l'ECG

III.3.1 L es ondes et les intervalles

a. Les ondes :

L'onde P: de première déflexion positive sur l'ECG représentant la dépolarisation auriculaire des oreillettes droite et gauche. Son amplitude est 0.2 mV et dure environ 90 ms.

Complexe QRS : C'est un ensemble de petites déflexions positives et négatives, qui provient de la contraction des ventricules. Il dure moins d'une seconde, une durée inférieure de 0.12 secondes, tandis que son amplitude varie de 5 à 20 mV. Il est constitué de trois ondes :

- L'onde Q : première déflexion négative
- L'onde R : première déflexion positive.
- L'onde S : défection négative qui suit l'onde R.

Sa forme est variable selon les dérivations utilisées (emplacement des électrodes) ou une arythmie donnée.

L'onde T: représente la conclusion de la repolarisation ventriculaire. Elle est généralement de faible amplitude et ne révèle aucun incident mécanique. Il s'agit d'un phénomène exclusivement électrique, durant lequel les ventricules retrouvent leur capacité à être stimulés. Elle présente généralement une dissymétrie.

L'onde U : elle peut parfois suivre l'onde T. Si elle est détectable, elle représente la repolarisation tardive de certaines zones du myocarde.

b. Les intervalles

- **Intervalle PR ou PQR :** correspond au temps de conduction auriculo-ventriculaire, sa durée doit être comprise entre 0.12s ET 0.2s. Ø
- **Intervalle ST ou RST :** il sépare la fin de la dépolarisation (fin du QRS) et le début de la repolarisation (début de l'onde T) Ø
- I**ntervalle QT :** il s'agit de distance entre le début du complexe QRS et la fin de l'onde T, englobant la dépolarisation et la repolarisation ventriculaires. Ø
- **Intervalle RR :** cet intervalle désigne le temps entre deux ondes R successives. Cet intervalle sert à mesurer la fréquence cardiaque.

III.3.2 Rythme cardiaque

Le rythme cardiaque se caractérise par deux propriétés fondamentales : la fréquence des ondes R, mesurée en nombre de battements par minute (bpm), et la régularité de ces battements. En l'absence de toute pathologie, le rythme est généralement régulier, avec une fréquence variante entre 60 et 100 bpm durant la journée et entre 40 et 80 bpm pendant la nuit. Si la fréquence cardiaque dépasse ces limites, il pourrait s'agir d'un trouble du rythme nécessitant une étude approfondie afin de déterminer la présence d'une éventuelle pathologie sous-jacente.

III.3.3 La fréquence cardiaque

Le rythme cardiaque correspond au nombre de battements du cœur par intervalle de temps (par minute). Sa vitesse est très élevée chez le nourrisson, rapide chez l'enfant et un peu plus lente chez le senior. En général, les sportifs ont un rythme cardiaque plus faible au repos comparativement à une personne qui s'exerce peu ou pas du tout.
On considère qu'un rythme cardiaque est normal s'il se situe en moyenne autour de 70 BPM chez l'adulte (dans la journée : entre 60 et 100 BPM, et durant la nuit : entre 40 et 80 BPM). En dehors de ces paramètres, on appelle bradycardie une fréquence cardiaque trop basse, et tachycardie lorsqu'elle est excessivement élevée.

III.3.4 Les caractéristiques fréquentielles de l'ECG

- Le domaine de l'ECG couvre une fréquence allant de 0.01 à 150 Hz.

- L'onde P et l'onde T présentent un spectre de basse fréquence, avec des composantes fréquentielles situées entre 0.5 Hz et 10 Hz.

La complexité du QRS englobe une fréquence nettement supérieure à celle des autres ondes de l'ECG. Ses éléments fréquentiels se trouvent essentiellement dans la plage de fréquences de 10 Hz à 150 Hz. C'est pour cette raison qu'il est couramment employé pour détecter les pulsations cardiaques et évaluer la fréquence du rythme cardiaque.

III.4 Diagnostic à partir des ondes

L'examen, en plus du rythme, de la forme des ondes de chaque battement a été rendu possible par la puissance des ordinateurs modernes et les nouvelles méthodes de traitement du signal. Actuellement, ce genre d'analyse est surtout circonscrit à la forme de l'onde R. L'examen individuel de chaque onde permet d'effectuer un pré-diagnostic réel. Ce diagnostic est réalisé en s'appuyant sur une expertise approfondie, par l'identification de la source du problème lorsque les battements de cœur, le complexe QRS et l'onde T présentent des anomalies.

Donc, les techniques que nous essayons de suggérer permettent une détection précise et constante de la majorité des ondes caractéristiques (Q, R.…) du battement. Elles devront faciliter l'identification précise des zones du signal susceptibles de refléter une anomalie cardiaque sur les 24 heures d'enregistrement.

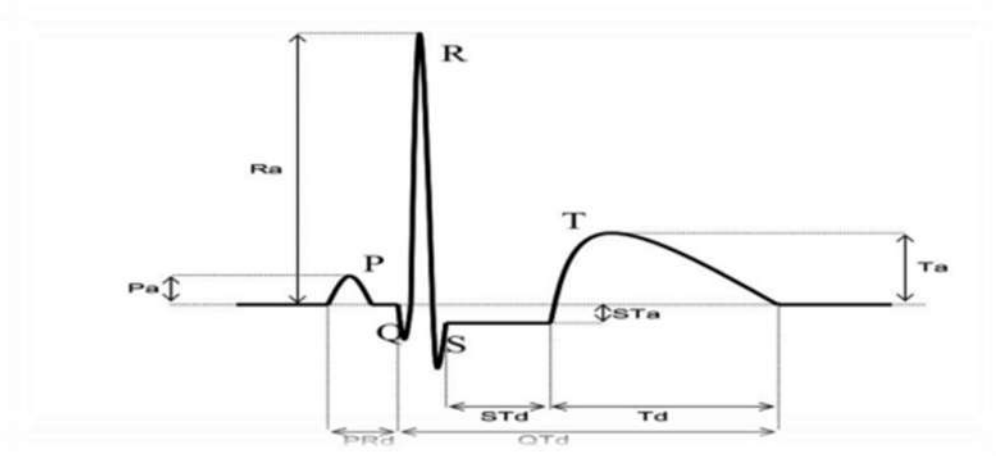

Fig. I.8: Paramètres d'intérêt pour la description d'un battement

Les tableaux ci-dessous présentent les valeurs habituelles des paramètres illustrés dans la figure (Figure 08) chez un adulte en santé

	Onde P	**Interv. PQ**	**Complexe QRS**	**Interv. ST**	**Interv . QT**	**Onde T**
Durée(s)	*0,08-0,1*	*0,12-0,2*	0,08	0,2	0,36	0,2
Amplitude	0,25	*Isoélectrique* 0	*Qa <0, Ra>0 Sa<0*	*Isoélectrique 0*	–	*Ta>0*

Type d'onde	**Origine**	**Amplitude (mV)**	**Durée (sec)**
L'onde P	Dépolarisation articulaire	≤ 0,2 mV	Intervalle P-R : 0,12-0,22
L'onde R	Repolarisation et dépolarisation Ventriculaire	1,60	0,07-0,1
L'onde T	Repolarisation de ventricules	0,1-0,5	Intervalle Q-T : 0, 37-0,44
Intervalle S-T	Contraction ventriculaire		Intervalle S-T : 0,015-0,5

IV. Enregistrement de l'électrocardiogramme

L'électrocardiogramme est le tracé qui est montré sur le papier de l'enregistreur ou affiché sur l'écran du moniteur. L'appareil d'enregistrement emploie du papier quadrillé, qui est divisé en grandes sections (5 mm × 5 mm) et petites sections (1 mm × 1 mm). Un écart de 2 grandes divisions ou de 10 mm équivaut à une tension d'1 mV.

La dimension temporelle est déterminée par la rapidité de mouvement du papier. La vitesse habituelle est de 25 mm/s, ce qui signifie que 0.2 s équivaut à une grande division et 0.04 s à une petite division. On utilise également une vitesse de 50 mm/s pour analyser plus précisément les complexes QRS lorsque la fréquence cardiaque est trop élevée ou lorsqu'il est nécessaire d'obtenir des détails spécifiques de l'électrocardiogramme.

La figure 09 montre le système d'enregistrement de l'électrocardiogramme qui comporte principalement les éléments suivants :

- Des électrodes de surface
- Un amplificateur

• Un enregistreur ou un moniteur (oscilloscope).

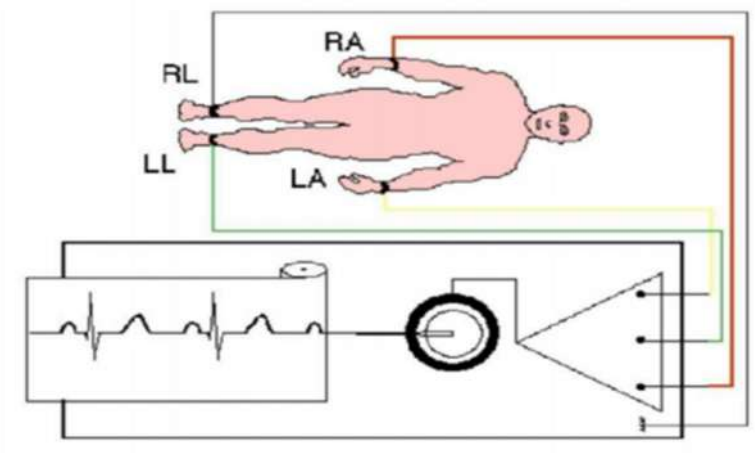

Fig. I.9: Système d'enregistrement de l'ECG

IV.1 Catégories des appareils d'enregistrement des ECG

Il existe différentes catégories d'appareils ECG. Leur domaine d'application est différent, ils Doivent donc se conformer à de différentes exigences. Les principales catégories d'appareils sont :

- ECG de diagnostic pour enregistrer les signaux au repos
- ECG d'effort
- ECG Holter
- Module ECG dans un moniteur patient de soins intensif
- Module ECG dans un défibrillateur automatique implantable
- ECG de télémétrie
- ECG fœtal

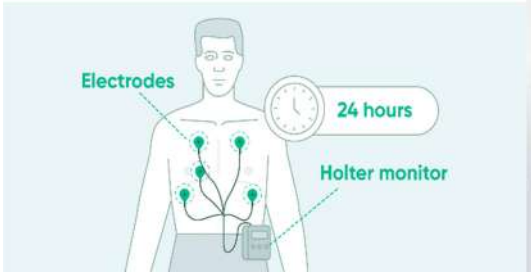

Fig. I.10: ECG Holter

Fig. I.11: ECG d'éffort

Fig. I.12: Module ECG -Moniteur patient de soins intensif- **Fig. I.13**: ECG de diagnostic au repos

IV.2 Artéfacts et bruits dans l'ECG

IV.2.1 Bruits techniques

a. Bruit dû au secteur

Le réseau de distribution électrique peut parfois brouiller le signal ECG avec une onde dont l'harmonique principale est à 50 Hz. La figure I.14 montre ce type de bruit qui apparait sur tout l'enregistrement et peut être assez fort mais il s'élimine facilement avec un filtre sélectif car c'est un bruit haute fréquence à bande étroite.

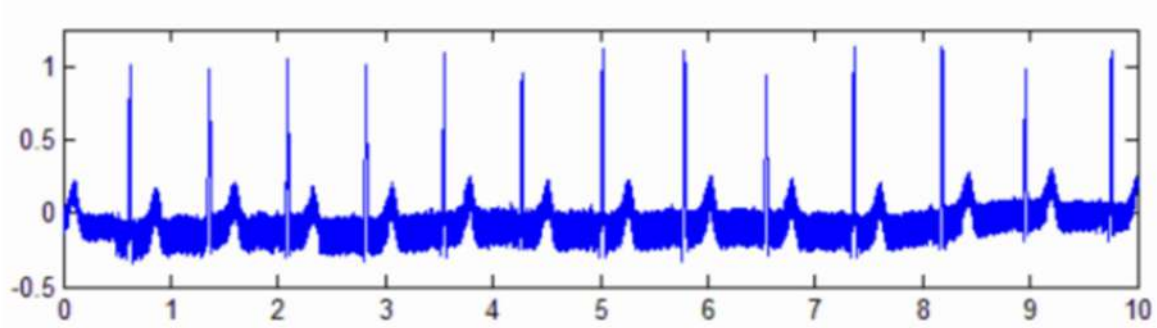

Fig. I.14: Signal électrocardiographe perturbé par le secteur

b. Bruit dû aux mouvements d'électrodes

Lorsque les électrodes sont connectées incorrectement, des sauts brusques de la ligne de base apparaissent. L'effet sur le tracé peut aller de la simple diminution d'amplitude à l'apparition de pics lorsque les électrodes sont en contact intermittent avec la peau. Ces pics peuvent parfois être confondus avec les ondes du tracé normal comme le montre la figure I.18. Ce type de bruit intermittent à bande spectrale large s'élimine difficilement car son énergie se trouve dans la même gamme de fréquence que le complexe QRS.

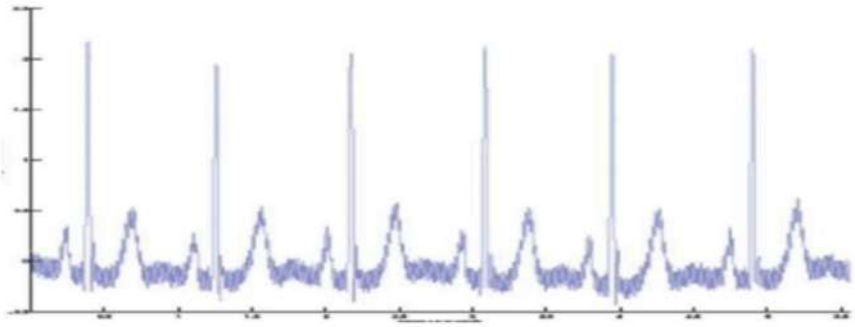

Fig. I.15: Bruit dû aux mouvements d'électrodes

c. Autres bruits courants

Parmi les bruits courants on peut citer les artefacts dus aux mouvements des câbles électriques, la saturation des instruments de mesure, les mauvais câblages, les artefacts dus au port de vêtements synthétiques, etc.

IV.2.2 Artefacts physiques

a. Mouvements de la ligne de base

Pendant l'enregistrement de l'électrocardiogramme, la ligne de base de l'ECG peut fluctuer en raison de l'activité respiratoire. Des perturbations supplémentaires peuvent occasionner un déplacement temporaire de la référence, comme par exemple, les mauvais contacts entre la peau et les électrodes.

Lors d'une surveillance de courte durée dans les établissements médicaux, les mouvements du patient sont restreints et donc, l'artefact lié au mouvement est rare. Toutefois, dans la surveillance ECG sur le long terme, les variations de la ligne de base sont fréquentes.

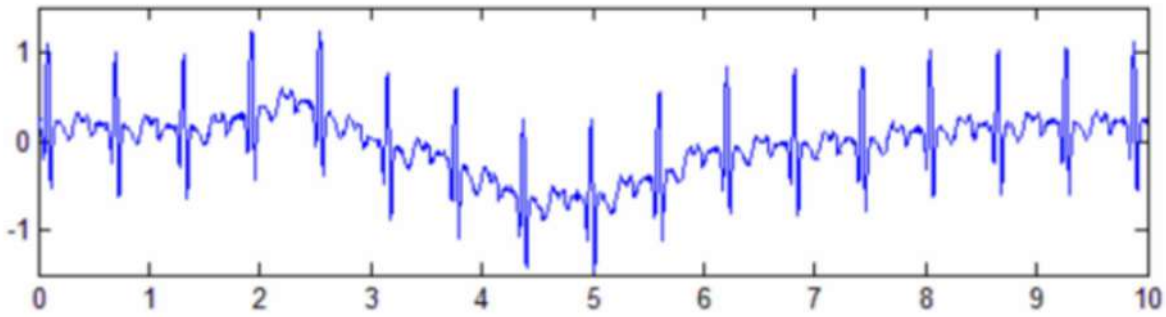

Fig. I.16: Mouvement de la ligne de base

b. Bruit myoélectrique ou tremblement somatique

L'activité musculaire est initiée par une dépolarisation des cellules musculaires. Bien que les appareils d'électrocardiographie soient principalement conçus pour détecter les rythmes cardiaques, l'ECG peut également capturer les contractions des muscles squelettiques. Comme l'indique la figure I.20, le phénomène le plus fréquent est une oscillation à haute fréquence associée à la tension musculaire d'un individu qui n'est pas convenablement étendu.

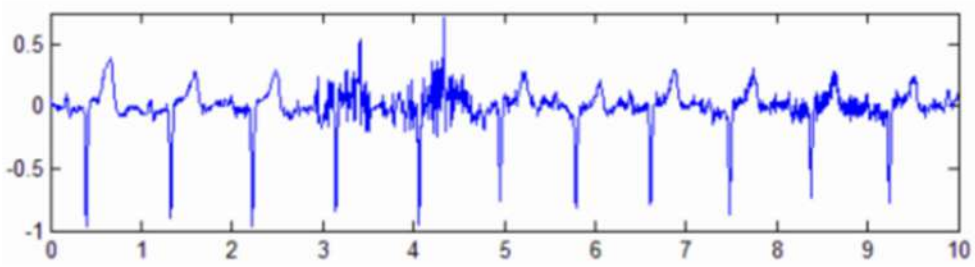

Fig. I.17: Bruit musculaire

V. Acquisition des signaux ECG

V.1 Structure d'une chaine d'acquisition ECG

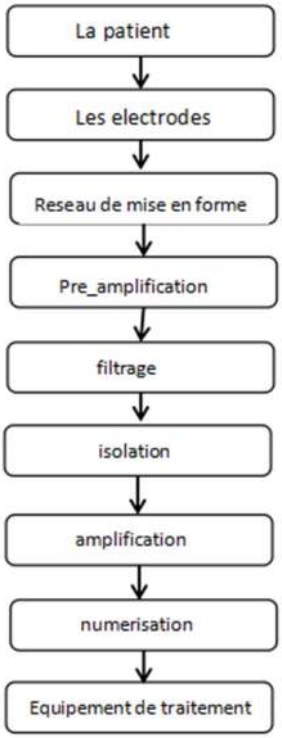

Fig. I.18: Schéma général d'une chaine acquis

V.2 Schéma général

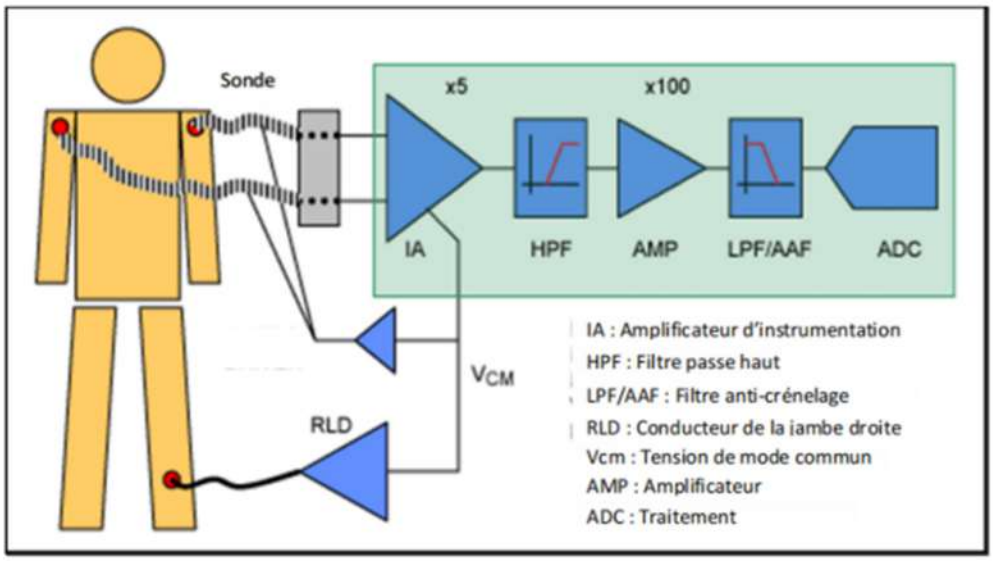

Fig. I.19: Chaine du traitement de l'ECG

V.3 Les électrodes

V.3.1 Le choix des électrodes

La sélection des électrodes représente un aspect crucial pour l'évaluation de l'activité électrodermale.

Par conséquent, les électrodes doivent se conformer à deux exigences majeures :

1. Une faible possibilité de polarisation entre les deux électrodes utilisées pour effectuer la mesure.
2. L'électrode ne doit pas être soumise à une polarisation due au passage du circuit. Pour mesurer l'activité électrodermale, on utilise généralement des électrodes réversibles faites d'Argent / Argent-chlorure (Ag/Ag Cl).

Fig. I.20: Les différentes électrodes utilisées pour l'enregistrement de l'ECG

V.3.2 La position des électrodes

La pose de toute électrode nécessite une préparation de la peut afin de réduire l'impédance naturelle de celle-ci. Il faut tout d'abord l'abraser en frottant avec un gel d'argile, pour éliminer les cellules mortes. Puis un dégraissage, à l'aide d'une solution d'éther alcool acétone, permet d'éliminer la couche de sécrétion protectrice de la peau (ainsi que les grains d'abrasif).

Nous avons opté pour la collecte de nos ECG à l'aide de deux électrodes positionnées à la base et à l'apex du cœur, dans le but de mesurer la différence de potentiel entre ces deux points, qui sont les plus adjacents au cœur. Cette position réduit le bruit associé aux muscles. Il est également envisageable d'utiliser des électrodes à pinces placées sur les poignets ou le pied pour obtenir l'ECG, mais leur position éloignée du cœur peut entraîner l'apparition de perturbations. Elles ne sont donc pas appropriées dans notre situation.

Nous n'avons eu recours qu'à deux électrodes, ce qui indique que nous avons effectué une unique dérivation. Cela est suffisant pour enregistrer le parcours électrique du courant et la progression temporelle de la différence de potentiel entre ces deux points.

V.3.3 Les caractéristiques des électrodes

- Une aptitude à capteur les basses amplitudes dans la gamme de 0,05 mV à 10 mV.
- Une impédance d'entrée très élevée
- Un courant d'entrée très bas, inférieur à 1Ma

V.4 Chaine d'acquisition

V.4.1 Capteur

Toutes information biologique doit se présent sous forme de signaux compréhensibles, enregistrables, et mesurables en valeurs normalisées. Le rôle joué par le capteur biomédical est l'un des aspects techniques que l'on rencontre.

Le capteur médical est constitué d'électrodes de recueil plates, métalliques, inoxydables et adhésives dont le contact d'électrique est assuré par une pâte conductrice.

V.4.2 Préamplificateur

Étant donné que le signal provenant des électrodes est généralement de 5mV pour un adulte et de 10µV pour un fœtus, une amplification préalable est nécessaire pour augmenter l'énergie du signal à un niveau convenable pour une utilisation ou un traitement ultérieur.

En plus de ses caractéristiques, le préamplificateur de l'ECG doit posséder une haute impédance d'entrée pour pouvoir isoler l'individu de la chaîne de mesures.

L'amplificateur d'instrumentation assure cette fonction grâce aux bénéfices qu'il offre pour une acquisition optimale du signal ECG. L'amplificateur d'instrumentation, basé sur trois (3) amplificateurs opérationnels, est illustré dans la figure 21.

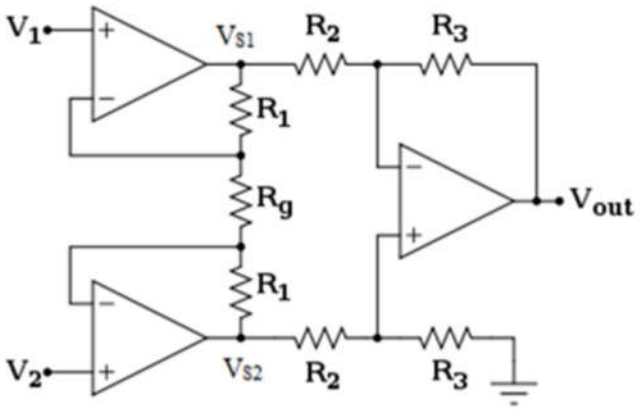

Fig. I.21: Amplificateur d'instrumentation à trois étages

V.4.3 Amplificateur d'isolation

L'amplificateur d'isolement est inclus pour des raisons de sécurité. Il isole le patient des autres circuits de traitement en limitant le flux de courant entre l'appareil d'enregistrement et le patient. Il nécessite que sa partie d'entrée soit alimenté par une batterie isolée galvaniquement. Le couplage peut être réalisé en utilisant une transmission en fibre de verre ou une transmission RF.

V.4.4 Filtrage

Le signal cardiaque (ECG) peut contenir des composantes fréquentielles allant jusqu'à environ 100–150 Hz dans des cas particuliers (comme les pics très abrupts), mais la majorité de l'information cliniquement utile se situe généralement en dessous de 40 à 100 Hz. Ainsi, un échantillonnage à 250 Hz ou plus est généralement suffisant pour capter l'essentiel des informations utiles du signal ECG.

Pour cela l'utilisation d'un filtrage passe-bas nous limitera l'étendu d'étude, ce qui a fait qu'on a choisi d'un filtre de fréquence de coupure basse de 0.5 Hz et une fréquence de de coupure haute égale à 45 Hz.

Le filtrage choisi est un filtrage passe bande. Le filtre utilisé est un filtre numérique de type Butterworth d'ordre quatre. Sa structure est celle de Sallen Key, car elle est la plus adaptée au filtrage passe bas ou passe bande.

L'amplificateur opérationnel est au cœur de la conception du filtre, où il est installé en tant qu'amplificateur non inverseur. On y ajoute des cellules RC.

Nous allons cascader deux filtres du second ordre pour obtenir un filtre de quatrième ordre.

De plus, afin de maintenir la linéarité de la réponse Butterworth (et éviter une transition vers une réponse Bessel ou Tchebychev), nous avons opté pour des gains en boucle fermée.

La fréquence de coupure du filtre est donnée par : $fc = \frac{1}{2\pi\sqrt{R1R2C1C2}}$

V.4.5 Amplificateur

L'amplificateur a pour rôle d'augmenter l'amplitude du signal mesuré, souvent très faible (de l'ordre de quelques microvolts à millivolts), afin de le rendre exploitable pour la numérisation et l'analyse, tout en améliorant le rapport signal sur bruit.

Dans une chaîne d'acquisition biomédicale, on utilise généralement un amplificateur d'instrumentation, spécialement conçu pour :

- Extraire les faibles signaux différentiels issus du capteur (par exemple entre deux électrodes ECG),
- Rejeter le bruit de mode commun (interférences électriques comme le 50/60 Hz secteur),
- Fournir un gain stable, précis et ajustable,
- Offrir une excellente impédance d'entrée (pour ne pas charger le capteur).

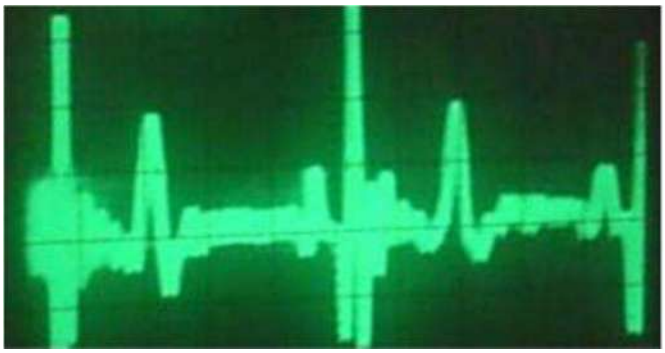

Fig. I.22: Signal ECG après amplification

Chapitre II

Apprentissage Automatique

II.1 Introduction

Depuis un demi-siècle, les chercheurs en intelligence artificielle travaillent à programmer des machines capables d'effectuer des tâches qui requièrent de l'intelligence. Cette intelligence peut nous offrir une aide à la décision telle que l'aide au diagnostic médical, la reconnaissance de formes, la reconnaissance de la parole ou la vision artificielle, la conduite de robots, l'exploration de grandes bases de données. Ce chapitre dresse des notions fondamentales du domaine de l'apprentissage automatique et sa contribution dans le développement technologique moderne.

On commence par définir l'intelligence artificielle et donner la distinction entre intelligence artificielle, apprentissage automatique (machine learning), et l'apprentissage profond (deep learning). On donne ensuite un aperçu sur les différents types d'apprentissage utilisés (supervisé, Semi-supervisé et non supervisé) et on finira par citer quelques modèles d'apprentissage automatique que nous avons utilisé dans le cadre de cette étude.

II.2 Intelligence artificielle

II.2.1 Définition de l'intelligence artificielle

L'intelligence artificielle (IA) désigne un domaine de l'informatique visant à créer des systèmes capables de réaliser des tâches normalement requérant l'intelligence humaine (Fig. II.1). Ces tâches incluent des activités comme la reconnaissance vocale, la prise de décision, la compréhension du langage naturel, la perception visuelle, et la résolution de problèmes complexes.

L'objectif de l'IA est de permettre aux machines de simuler certaines fonctions cognitives humaines, telles que l'apprentissage (via l'apprentissage automatique ou machine Learning), la compréhension contextuelle, et l'adaptation à des situations nouvelles.

En résumé, l'intelligence artificielle est une branche qui cherche à doter les machines de la capacité à exécuter des tâches humaines de manière autonome, voire intelligente.[8]

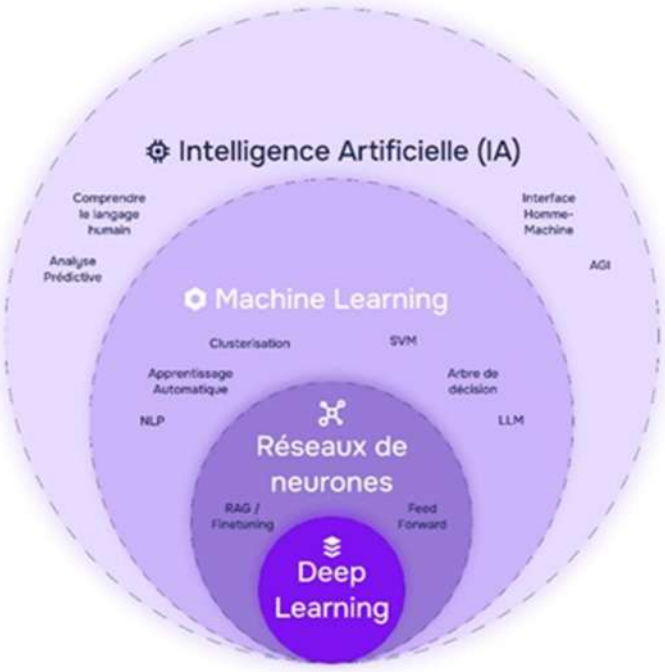

Fig. II.1 : Schéma de fonctionnement d'une IA

II.2.2 Axes de l'intelligence artificielle

II.2.2.1 l'approche humaine

- L'appareil doit se comporter de la même façon que les êtres humains :

Alan Turing est le premier à avoir pensé à la création d'une machine intelligente. Pour qu'un appareil soit jugé intelligent, il doit réussir ce test avec brio. Sans entrer dans les détails de ce test, il établit une connexion entre un interrogateur (humain) et un participant (humain ou virtuel). Si, après le test, l'interrogateur n'arrive pas à déterminer si le participant était un homme ou une machine, alors le test est considéré comme réussi.

- La machine doit être capable de penser de manière similaire aux humains :

Pour que les scientifiques puissent concevoir une machine qui raisonne comme un être humain, ils doivent d'abord comprendre la manière dont l'humain pense. Ils ont donc identifié trois méthodes pour essayer de déchiffrer le processus de pensée :

- L'introspection (l'examen de ses propres pensées).
- La psychologie (analyser un individu à travers ses actions).
- L'imagerie du cerveau.

Une fois que l'on a acquis une compréhension relativement précise de l'esprit, notamment par le biais des nombreuses études menées sur ce thème, cette théorie dépeint le cerveau comme un programme informatique. Il suffirait donc d'identifier le code de notre cerveau pour pouvoir l'implémenter sur un ordinateur.[9]

II.2.2.2 L'approche rationnelle

- La machine doit se comporter de manière rationnelle :

Pour qu'une machine soit considérée comme rationnelle, elle doit en théorie être capable de :

1. opérer de manière autonome.
2. Percevoir son entourage.
3. Maintenir sur une durée étendue.
4. S'ajuster aux modifications.
5. Suivre des objectifs.

- La machine doit faire preuve de rationalité dans ses pensées :

On appelle cette méthode la « loi de la pensée logique ». Aristote a été l'un des premiers à reconnaître que certaines vérités demeurent toujours valides. Il a donc souhaité formaliser le « bien penser », c'est-à-dire les méthodes de raisonnement logiques. On décrit ce système de pensée comme étant « logique ». Ces principes de la pensée étaient supposés gouverner l'ensemble des opérations de l'esprit humain.

C'est à partir de cette observation que les premiers programmes informatiques capables de résoudre des problèmes logiques ont vu le jour. On les qualifie déjà de « intelligents », ces systèmes qui aident l'homme dans des tâches nécessitant une réflexion rationnelle.

- La machine doit faire preuve de rationalité dans ses pensées :

On appelle cette méthode la « loi de la pensée ». Aristote a été l'un des premiers à reconnaître que certaines vérités demeurent toujours. Il a donc souhaité formaliser le « bien penser », c'est-à-dire les méthodes de raisonnement indiscutables. Je reprends son illustration : « Socrate est un homme, tous les hommes sont mortels, donc Socrate est mortel ». On décrit ce système de pensée comme étant « logique ». Ces principes de la pensée étaient supposés gouverner l'ensemble des opérations de l'esprit humain.

C'est à partir de cette observation que les premiers programmes informatiques capables de résoudre des problèmes logiques ont vu le jour. On les qualifie déjà de « intelligents », ces systèmes qui aident l'homme dans des tâches nécessitant une réflexion rationnelle.[9]

II.3 Apprentissage automatique (machine Learning)

II.3.1 Définition

Machine Learning est une branche de l'intelligence artificielle, se référant à la capacité des systèmes informatiques à découvrir de manière autonome des solutions aux problèmes en identifiant les motifs présents dans les bases de données. Autrement dit : l'apprentissage automatique (AA) permet aux systèmes informatiques d'identifier des modèles à partir d'algorithmes et de jeux de données existants, et de développer des idées pour des solutions appropriées.

Ainsi, dans le domaine de l'apprentissage automatique, la connaissance artificielle est produite à partir de l'expérience.[10]

II.3.2 Modélisation

L’apprentissage automatique consiste à permettre à une machine d’améliorer sa performance sur un ensemble donné de tâches, à partir d’expériences antérieures ou de données d’entraînement (Fig. II.2). Plus formellement, un système apprend lorsqu’il est capable, en fonction d’un ensemble d’expériences, d’augmenter son efficacité selon une mesure de performance définie. L’ensemble d’expériences peut être fourni dès le départ sous forme de données existantes, ou s’enrichir progressivement au cours du fonctionnement du système.

L’objectif fondamental de l’apprentissage automatique est donc de construire des modèles capables de généraliser à partir des données d’entraînement, afin d’obtenir de bonnes performances sur de nouvelles données.

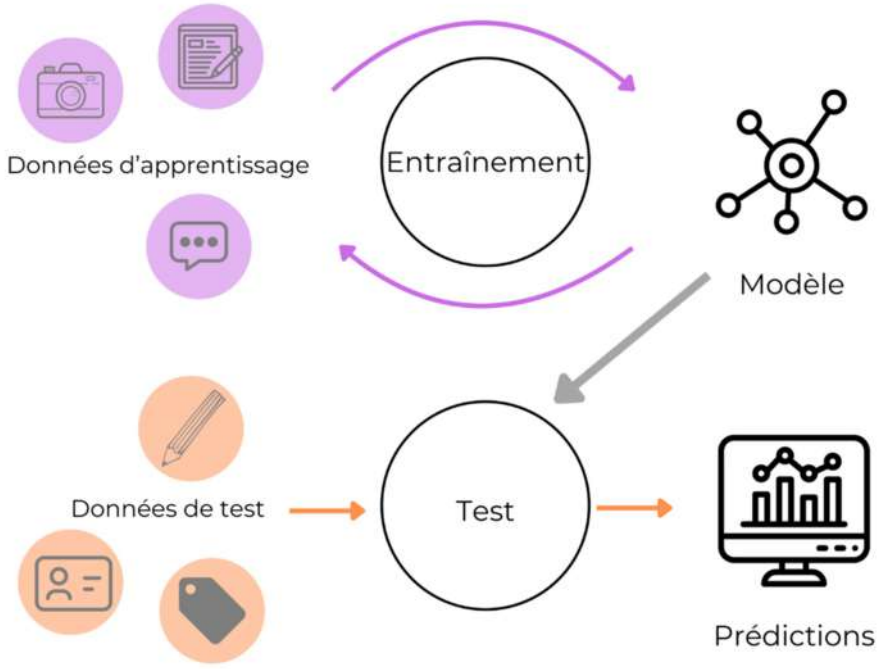

Fig. II.2 : Les bases de données de l'entrainement d'un modèle de machine Learning

II.3.3 Domaines d'applications de l'apprentissage automatique

L'apprentissage automatique est pertinent pour de nombreuses actions humaines et est particulièrement adapté à la question de l'automatisation du processus décisionnel. Par exemple, cela pourrait impliquer :

- **Bio-informatique et diagnostic médical**

  Ces dernières années, l'émergence de nouvelles technologies en biologie ainsi que les avancées en informatique ont augmenté non seulement la quantité des données biologiques mais aussi leur complexité. Les scientifiques font alors appel aux technologies de l'informatique et déployés des techniques d'apprentissage automatique qui leur permettent de transformer ces données en information et de résoudre ainsi des problèmes biologiques et médicaux, de classer et de mieux comprendre diverses maladies. Ces approches devraient également aider à diagnostiquer la maladie en identifiant les segments de la population qui sont les plus à risque de certaines maladies.[11]

- **Automatisation**

  L'apprentissage automatique, qui fonctionne de manière entièrement autonome dans n'importe quel domaine sans aucune intervention humaine. Par exemple, des robots exécutant les étapes essentielles du processus dans les usines de fabrication.

- **Traitement du langage naturel**

  Le traitement du langage nature est un domaine de la linguistique, de l'informatique qui s'intéresse à l'interaction entre ordinateur et langage humains (naturels). Il fait partie des techniques d'intelligence artificielle. Les algorithmes d'intelligence artificielle ont pour rôle d'identifier et d'extraire les règles du langage naturel, de convertir les données de langage non-structuré sous une forme que les ordinateurs pourront le comprendre. La plupart des techniques de traitement naturel du langage reposent sur l'apprentissage profond.[11]

- **Organisation gouvernementale**

  Certains gouvernements utilisent le ML pour gérer la sécurité publique et les services publics. En Chine, le gouvernement utilise l'intelligence artificielle « reconnaissance faciale massive » pour empêcher les piétons de traverser en dehors des passages autorisés (jaywalking).

- **Moteur de recherche**

  Un moteur de recherche est une application informatique permettant de rechercher des ressources, des contenues, des documents (page web, d'images, de vidéos, d'actualités, de fichiers, etc.…), à partir de mots clés.[11]

II.4 Types d'apprentissage automatique

II.4.1 L'apprentissage supervisé

On parle d'apprentissage supervisé lorsque l'on dispose de données d'entraînement étiquetées, c'est à dire dont on connaît la sortie voulue (Fig. II.3). En notant les N entrées x_i et les sorties cibles associées y_i, on dispose de l'ensemble de données $D = \{x_i, y_i\}$ $i \in [1, N]$. L'objectif est d'entraîner le modèle choisi pour qu'il puisse prédire correctement la sortie pour des entrées non étiquetées.[12]

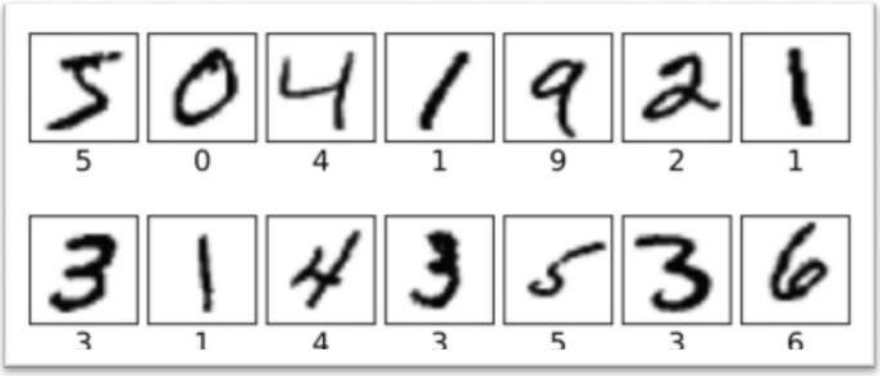

Fig. II.3 : Ensemble de données étiquetées Chaque image d'entrée est associée avec la prédiction voulue

L'apprentissage supervisé est généralement utilisé pour de la régression ou de la classification :

II.4.1.1 Classification

Cela consiste à pouvoir associer une donnée complexe telle qu'une image ou un profil d'utilisateur à une classe d'objets, les différentes classes possibles étant fournies au préalable par le concepteur. La classification utilise un jeu de données d'entrainement associé à des descriptifs (les classes) pour la détermination d'un modèle. Ceci produit un modèle qui a la capacité d'anticiper la catégorie d'un nouvel élément introduit en entrée. Dans les exemples classiques on peut citer la reconnaissance d'un simple chiffre sur une image, l'attribution d'un client à un segment spécifique ou à un type particulier de clientèle (insatisfaits, susceptibles de se désabonner d'un service, etc.), ou encore l'identification d'un virus basée sur le comportement ou les caractéristiques d'un logiciel.[13]

II.4.1.2 Régression

La régression permet de prédire une valeur numérique y en fonction d'une valeur x, grâce à un ensemble d'apprentissage composé de paires de données (x, y).

On peut par exemple prédire la valeur d'un bien immobilier ou d'une société en fonction de divers paramètres les décrivant.

Les figures II.4 et II.5 ci-dessous représentent ce principe qui se base uniquement sur une donnée d'entrée et une donnée de sortie. En pratique, les régressions font appel à plusieurs paramètres en entrée.[13]

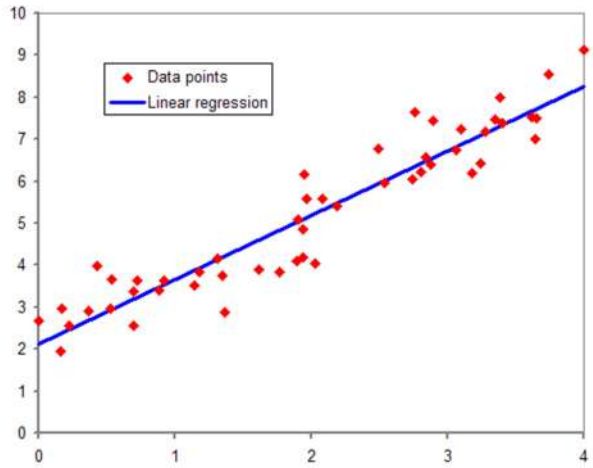

Fig. II.4 : Régression linéaire

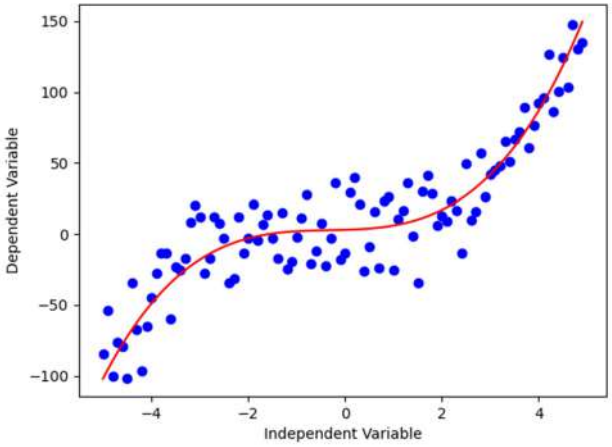

Fig. II.5 : Régression non linéaire

II.4.2 Apprentissage non-supervisé

On parle d'apprentissage non supervisé si les données ne sont pas étiquetées. On dispose donc de données d'entrée dont on ne connaît pas la sortie associée (Fig. II.6). L'ensemble de données est donc $D = \{x_i\}\ i \in [1, N]$ et l'objectif du système est d'identifier des caractéristiques communes aux données d'entraînement.

Fig. II.6 : Ensemble de données non-étiquetées Les images d'entrées, extraites de la base de données MNIST Fashion, ne sont pas associées à une sortie cible

Le clustering, ou segmentation automatique, est une méthode d'apprentissage non supervisé qui permet, à partir d'un jeu de données non labellisé, d'identifier des groupes de données similaires appelés clusters. L'objectif est de regrouper les données de manière à ce que les éléments d'un même cluster soient plus proches les uns des autres, selon une certaine mesure de similarité, que des éléments appartenant à d'autres clusters.

Parmi les différentes méthodes de clustering, l'algorithme des k-moyennes (k-means) est l'un des plus largement utilisés. Il consiste à partitionner les données en k clusters en minimisant la variance intra-cluster, c'est-à-dire la distance entre les points d'un cluster et son centroïde.[12]

II.4.3 L'apprentissage semi-supervisé

Qu'il soit réalisé de manière probabiliste ou non, le clustering a pour objectif de révéler la distribution sous-jacente des exemples dans l'espace descriptif. Il est utilisé lorsque les données ne sont pas étiquetées (absence de labels). Le modèle exploite ainsi des exemples non labellisés qui peuvent néanmoins contenir des informations utiles à la compréhension ou à la structuration des données.

À titre d'exemple, dans le domaine médical, le clustering peut assister le diagnostic en identifiant des sous-groupes de patients présentant des caractéristiques similaires, ou aider à sélectionner les protocoles de tests diagnostiques les plus appropriés et les moins coûteux.[14]

II.4.4 L'apprentissage par renforcement

Ce type d'apprentissage automatique repose sur le renforcement. La machine apprend par essais et erreurs dans différents contextes, en évaluant les conséquences de ses actions (Fig. II.7).

Que les résultats soient connus ou non à l'avance, la machine ne dispose pas initialement de la stratégie optimale pour atteindre les meilleurs résultats.

Au fil des interactions avec l'environnement, l'algorithme associe progressivement certaines actions aux situations rencontrées, en fonction des récompenses ou pénalités reçues. C'est par la répétition de ces expériences et l'accumulation de retours qu'il améliore sa stratégie de décision et tend vers des performances de plus en plus optimales.

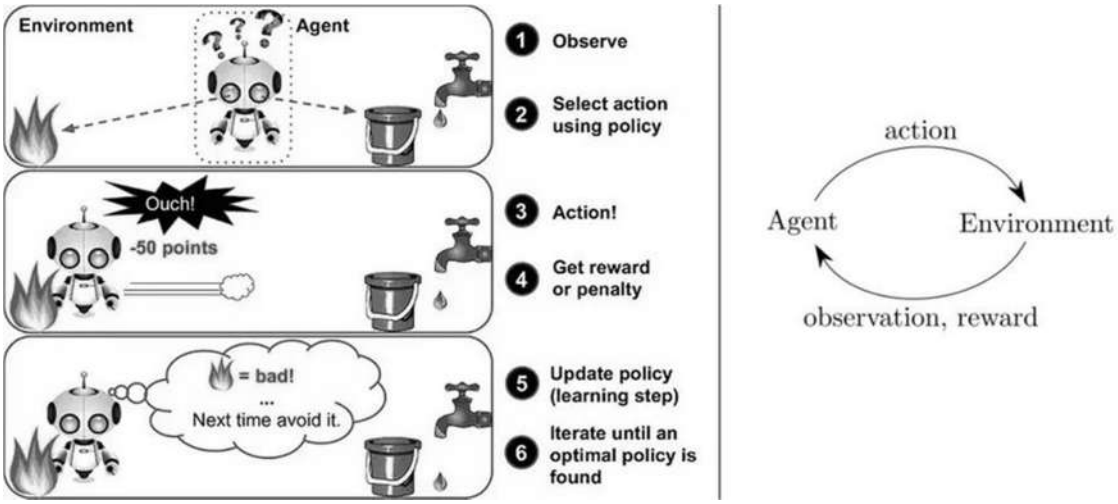

Fig. II.7 : Algorithme des méthodes d'apprentissage par renforcement

III.5 Quelques exemples d'algorithmes d'apprentissage automatique

III.5.1 Réseaux de neurones artificiels

Les réseaux de neurones artificiels sont des modèles mathématiques inspirés de la biologie. La brique de base de ces réseaux, le neurone artificiel, était issue au départ d'une volonté de modélisation du fonctionnement d'un neurone biologique.

Le cerveau humain se compose d'environ 1012 milliards (mille milliards) de neurones interconnectés, avec 1000 à 10000 synapses (connexions) par neurone (Fig.II.8). Les neurones ne sont pas tous identiques, leur forme et certaines caractéristiques permettent de les répartir en quelques grandes classes :

- Perceptron simple (single layer perceptron)
- Perceptron multicouche (multilayer perceptron - MLP)
- Réseaux convolutifs (Convolutional Neural Networks-CNN)

- Réseaux récurrents (Recurent Neural Networks-RNN)
- LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit)
- et d'autres (Auto-encodeurs, Transformers, …)

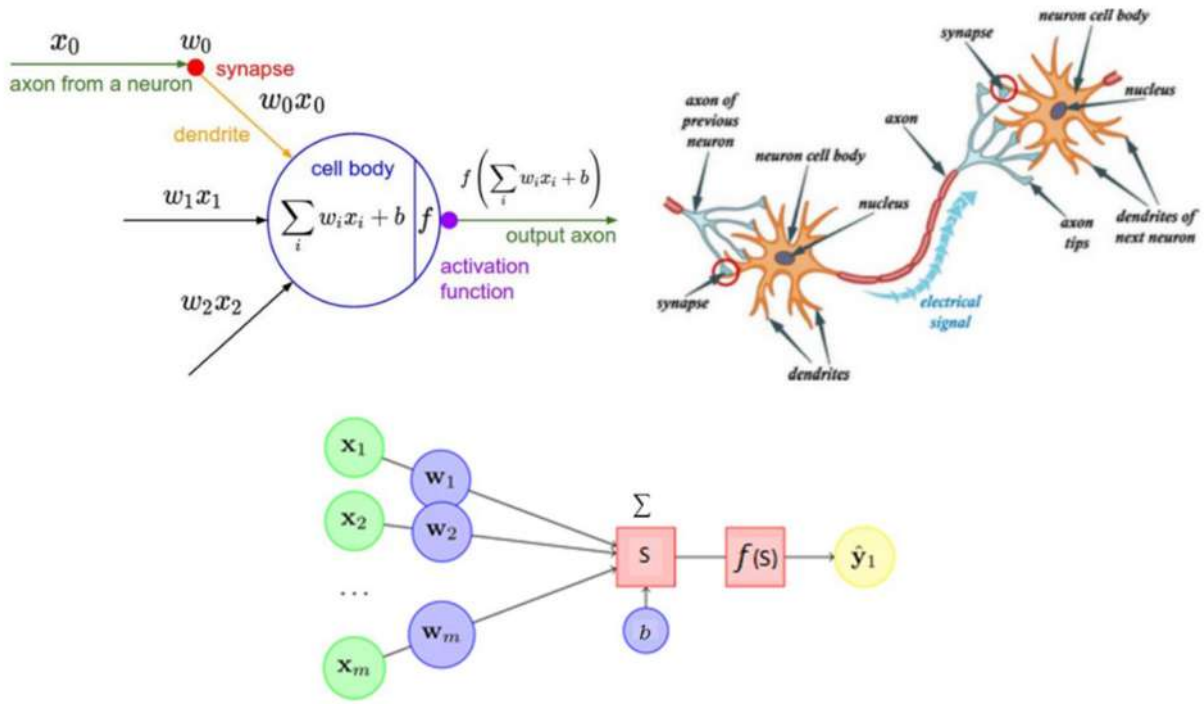

Fig. II.8 : Réseaux de neurones

Tels que :

- (x_1, x_2, …., x_m) : sont les entrées du neurone (signaux qui lui parviennent)
- (w_1, w_2, …., w_m) : sont les poids associés à chaque connexion.
- b : le seuil d'activation
- S : la somme pondérée des entrées (potentiel d'activation)

$$S = \sum_{i=1}^{n} w_i x_i + b$$

$\hat{y} = f(S)$: la sortie du neurone (réponse du neurone « activé $\hat{y} = 1$ ou non activé $\hat{y} = 0$ » (Fig. II.9)

$$\hat{y} = \begin{cases} 1 & si \quad S > 0 \\ 0 & si \quad S < 0 \end{cases}$$

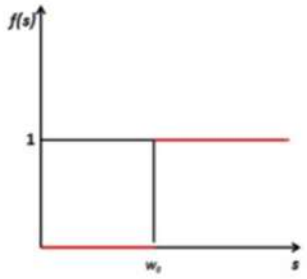

Fig. II.9 : Fonction d'activation d'un neurone artificiel

A partir de ce modèle ont été définis divers modèles de neurones et avec d'autres fonctions d'activations. La figure suivante montre d'autres fonctions d'activation (Fig. II.10).

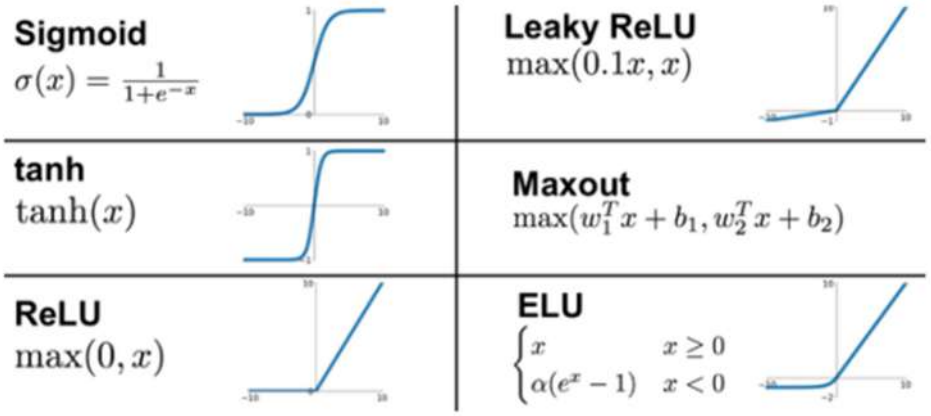

Fig. II.10 : Fonctions d'activation possibles d'un neurone formel

Schématiquement, le fonctionnement d'un neurone artificiel est le suivant :

- Les neurones sont organisés en couches successives : chaque neurone reçoit une information (une entrée) issue des neurones de la couche qui précède.
- Chacune de ces informations est pondérée : elle est multipliée par une valeur qui lui confère un "poids" wi particulier.
- Les entrées ainsi pondérées sont additionnées.

- Elles sont traitées par une fonction objective dont le but est d'adapter la valeur de sortie à une plage de valeurs.
- La valeur de sortie issue de cette fonction constitue l'entrée de l'ensemble des neurones de la couche suivante.

Le fonctionnement d'un réseau neuronal est composé de deux phases distinctes :

1. Une phase d'apprentissage pendant laquelle les poids des connexions sont mis à jour de manière dynamique ;
2. Une phase d'exécution pendant laquelle le réseau est effectivement opérationnel.

Il existe un grand nombre de modèles ou d'architectures de réseau de neurones qui peuvent être distingués par :

- La règle ou fonction d'activation locale à chaque neurone qui est la fonction des valeurs d'entrée reçues d'autres neurones.
- - La règle d'apprentissage qui permet de modifier les poids des connexions entre neurones.
- - La topologie décrivant la manière dont les neurones sont connectés entre eux.

III.5.2 Random Forest

L'algorithme d'apprentissage automatique supervisé Random Forest est un algorithme polyvalent et puissant qui combine différents arbres de décision pour créer une « forêt » (Fig. II.11).

Le langage de programmation R et Python sont employés pour résoudre des problèmes de classification et de régression. Les Data Scientists ont une grande préférence pour cette méthode de machine learning en raison de ses multiples bénéfices par rapport aux autres algorithmes de données. Son interprétation est simple, sa stabilité est généralement satisfaisante et elle peut être employée pour des tâches de régression ou de classification, ce qui permet de traiter une grande diversité de problèmes en Machine Learning.

Dans Random Forest, le mot « Forest » suggère clairement que cet algorithme utilise des arbres de décision, également connus sous le nom d'arbres décisionnels.[15]

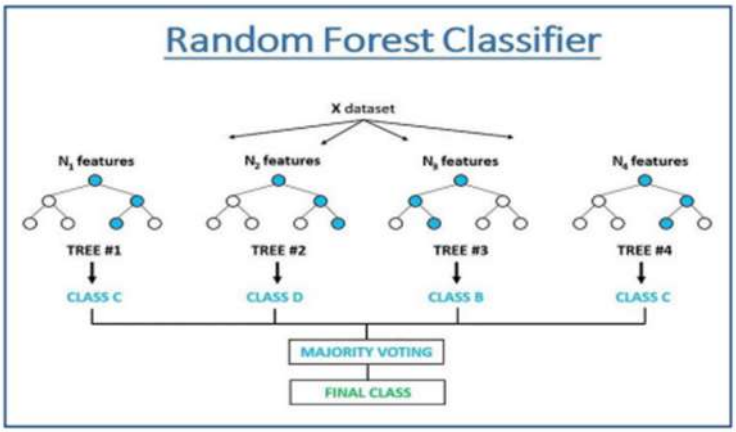

Figure 11 : Exemple d'arbre de Décision [16]

III.5.2 Arbre de décision

L'arbre de décision est l'outil le plus puissant et le plus populaire pour la classification et la prédiction. Un arbre de décision est un organigramme semblable a une structure arborescente, ou chaque nœud interne (nœud de décision) désigne un test sur un attribut, chaque branche représente un résultat du test et chaque nœud feuille (nœud terminal) est repéré par sa position (liste des numéros des arcs qui permettent d'y accéder en partant de la racine), et étiquetées par une classe (Fig. II.12). Lors de l'apprentissage d'un arbre, les données source sont divisées en sous-ensembles en fonction d'un test de valeur d'attribut, qui est répété récursivement sur chacun des sous-ensembles dérivés. Une fois que le sous-ensemble d'un nœud a la valeur équivalente à sa valeur cible, le processus sera terminé. La profondeur de l'arbre et la présence de nombreux nœuds améliorent les performances du modèle [15]

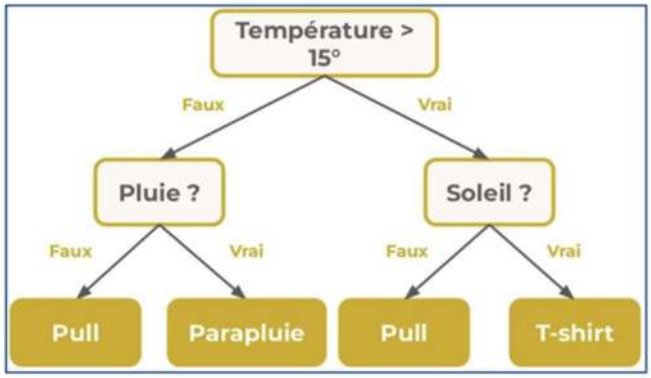

Fig. II.12 : Exemple d'un arbre de décision [17]

III.5.3 K plus proches voisin (KNN)

L'algorithme KNN (k-Nearest Neighbors) figure parmi les plus simples algorithmes d'apprentissage artificiel. Dans un contexte de classification d'une nouvelle observation x, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation. La classe de x est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation x (Fig. II.13). Donc la méthode du plus proche voisin est une méthode non paramétrique ou une nouvelle observation est classée dans la classe d'appartenance de l'observation de l'échantillon d'apprentissage qui lui est la plus proche, au regard des covariables utilisées. La détermination de leur similarité est basée sur des mesures de distance.[18]

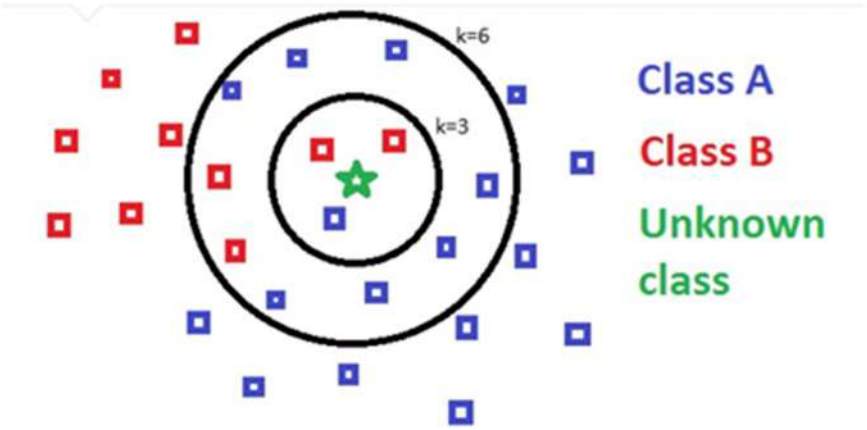

Fig. II.13 : Illustration du K plus proches voisin.[19]

III.5.4 Machine à vecteurs de support (SVM)

Les machines à vecteurs de support ou séparateurs à marge étendue (en anglais, Support Vector Machine, SVM) constituent une approche d'apprentissage supervisé performante dans de nombreux domaines tels que la bio-informatique, la recherche d'information, la vision par ordinateur et la finance. L'idée des SVMs est de rechercher un hyperplan (droite dans le cas de deux dimensions) qui sépare le mieux ces deux classes. Si un tel hyperplan existe, c'est-à-dire si les données sont linéairement séparables, on parle d'une machine à vecteur support à marge dure (Hard margin). (Fig. II.14) [18]

Leur efficacité repose sur la recherche de l'hyperplan optimal séparant les classes d'exemples d'apprentissage avec la plus grande marge possible.

Mathématiquement, L'hyperplan séparateur est défini comme l'ensemble des points x satisfaisant l'équation suivante :

$$H(x) = w^T.x + b$$

où :

- $w \in R^n$ est le vecteur normal à l'hyperplan, dimension m, détermine l'orientation de l'hyperplan
- $b \in R$ est le biais, permet le déplacement parallèle de l'hyperplan.

La fonction de décision associée est :

$$h(x) = sign(w.x + b)$$

Cette fonction de décision, pour un exemple x, peut être exprimée comme suit :

$$\begin{cases} Classe = 1 \ \ si\ H(x) > 0 \\ Classe = -1 \ \ si\ H(x) < 0 \end{cases}$$

Puisque les deux classes sont linéairement séparables, il n'existe aucun exemple qui se situe sur l'hyperplan, c-à-d qui satisfait H(x) = 0. Il convient alors d'utiliser la fonction de décisions suivante :

$$\begin{cases} Classe = 1 \ \ si\ H(x) > 1 \\ Classe = -1 \ \ si\ H(x) < -1 \end{cases}$$

Les valeurs +1 et -1 à droite des inégalités peuvent être des constantes quelconques +a et -a, mais en divisant les deux parties des inégalités par a, on trouve les inégalités précédentes qui sont équivalentes à l'équation suivante :

$$y_i(w^T.x_i + b) \geq 1, i = 1..n$$

Les exemples situés au plus proche de l'hyperplan, appelés vecteurs de support, déterminent sa position définitive.

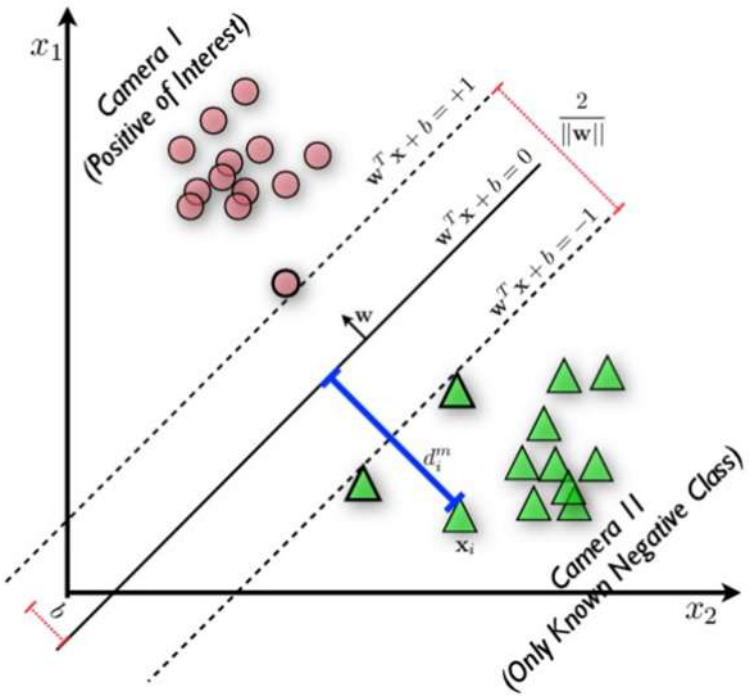

Fig. II. 14 : SVM binaire à marge dure

L'hyperplan $\mathrm{w}^T.\mathrm{x}+\mathrm{b}=0$ représente un hyperplan séparateur des deux classes, et la distance entre cet hyperplan et l'exemple le plus proche s'appelle la marge. La région qui se trouve entre les deux hyperplans $\mathrm{w}^T.\mathrm{x}+\mathrm{b}=-1$ et $\mathrm{w}^T.\mathrm{x}+\mathrm{b}=1$ est appelée la région de généralisation de la machine d'apprentissage. La maximisation de cette région est l'objectif de la phase d'entrainement qui consiste, pour la méthode SVM, à rechercher l'hyperplan qui maximise la région de généralisation c-à-d la marge. Un tel hyperplan est appelé "hyperplan de séparation optimale". En supposant que les données d'apprentissage ne contiennent pas des données bruitées (mal-étiquetées) et que les données de test suivent la même probabilité que celle des données d'entraînement, l'hyperplan de marge maximale va certainement maximiser la capacité de généralisation de la machine d'apprentissage.

Lorsque les données ne sont pas linéairement séparables, les SVM peuvent utiliser des fonctions noyaux (kernel functions) pour projeter les données dans un espace de dimension supérieure, permettant ainsi une séparation linéaire dans cet espace transformé.

Parmi les noyaux couramment utilisés figurent : le noyau linéaire, le noyau polynomiale, le noyau gaussien (RBF).

III.5.5 Régressions Lasso

La régression Lasso (Least Absolute Shrinkage and Selection Operator) est une méthode de régression linéaire régulière introduite par Robert Tibshirani en 1996, qui permet à la fois de réduire la complexité du modèle et de sélectionner les variables les plus pertinentes. Elle repose sur l'ajout d'une pénalité de type L1 à la fonction de coût, ce qui conduit à l'annulation de certains coefficients. Ainsi, le Lasso réalise automatiquement une sélection de variables, en forçant certains coefficients à devenir exactement nuls.

La fonction à minimiser dans la régression Lasso est donnée par :

$$\frac{min}{\beta}\left[\frac{1}{2n} \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1\right]$$

Où :

- y est le vecteur des valeurs cibles,
- X est la matrice des variables explicatives,
- β est le vecteur des coefficients à estimer,
- $\lambda \geq 0$ est un paramètre de régularisation qui contrôle la force de la pénalisation L1.

Plus la valeur de λ est élevée, plus la pénalisation est forte, et plus de coefficients sont contraints à zéro. Cette propriété rend le Lasso particulièrement utile dans les contextes de haute dimension, où il est important de réduire le nombre de variables utilisées dans le modèle.[20]

III.5.6 LightGBM (Light Gradient Boosting Machine)

LightGBM est un algorithme de machine learning de type gradient boosting développé par Microsoft, optimisé pour la performance et l'efficacité. Il fait partie des algorithmes dits ensemble, basés sur la construction séquentielle d'un ensemble de modèles faibles (souvent des arbres de décision) pour corriger les erreurs des modèles précédents.

LightGBM se distingue par deux principales innovations :

- Leaf-wise growth (au lieu de level-wise comme dans XGBoost), ce qui permet une meilleure réduction de l'erreur, bien qu'avec un risque accru d'overfitting.
- Histogram-based splitting, qui accélère considérablement l'entraînement en réduisant la complexité du calcul et la consommation mémoire.

LightGBM est particulièrement adapté aux grands volumes de données et aux jeux de données à haute dimension, tout en conservant une grande précision.[21]

Avantages :

- Temps d'entraînement rapide,
- Faible consommation mémoire,
- Prise en charge du traitement de données catégorielles sans encodage préalable,
- Bonne précision et capacité de généralisation.

III.5.7 XGBoost (Extreme Gradient Boosting)

XGBoost, abréviation de *Extreme Gradient Boosting*, est un algorithme d'apprentissage supervisé basé sur la méthode du gradient boosting. Il a été proposé par Tianqi Chen en 2016 et s'est rapidement imposé comme une référence dans les compétitions de data science en raison de ses performances élevées, de sa robustesse et de sa rapidité.

XGBoost construit un ensemble d'arbres de décision de manière séquentielle, où chaque nouvel arbre vise à corriger les erreurs commises par les arbres précédents. Sa force réside dans plusieurs optimisations techniques :

- Utilisation d'une fonction objective régulière intégrant la pénalisation de la complexité des arbres (via la régularisation L1/L2),
- Prise en charge du traitement parallèle lors de la construction des arbres,
- Gestion efficace des valeurs manquantes,
- Réduction du surapprentissage grâce à des mécanismes comme la shrinkage (apprentissage lent) et le subsampling (échantillonnage aléatoire).

Ces caractéristiques font de XGBoost un algorithme particulièrement efficace pour les problèmes de classification et de régression, notamment sur des jeux de données volumineux ou avec des relations complexes entre les variables.

L'ElasticNet est une méthode de régularisation utilisée en régression linéaire qui combine les pénalités des méthodes Lasso (L1) et Ridge (L2). Elle a été introduite pour surmonter les limitations de ces deux approches, notamment en présence de variables explicatives fortement corrélées ou lorsque le nombre de variables est supérieur au nombre d'observations.

La fonction objective de l'ElasticNet ajoute une double pénalisation à la somme des carrés des résidus :

Où :

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \alpha \left(\rho \|\beta\|_1 + \frac{1-\rho}{2} \|\beta\|_2^2 \right) \right\}$$

- y est le vecteur des réponses,
- X est la matrice des prédicteurs,
- β est le vecteur des coefficients à estimer,
- $A \geq 0$ contrôle l'intensité globale de la régularisation,
- $\rho \in [0,1]$ équilibre la contribution entre les pénalités L1 (Lasso) et L2 (Ridge).

L'ElasticNet bénéficie ainsi des avantages du Lasso (sélection de variables) tout en conservant la stabilité offerte par la régularisation Ridge. Cette méthode est particulièrement adaptée aux jeux de données haute dimension, où plusieurs variables peuvent être redondantes.[22]

II.6 Critères de performance

II.6.1 Erreur quadratique moyenne RMSE - Root Mean Square Error)

L'erreur quadratique moyenne est une mesure de la différence entre les valeurs prédites par un modèle et les valeurs observées réelles. Elle pénalise davantage les grandes erreurs que les petites.

$$RMSE = \sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2 \right)}$$

- y_i : valeur réelle
- x_i: valeur prédite
- n : nombre total d'observations

Avantage : Plus sensible aux erreurs importantes, ce qui est utile quand les grandes erreurs sont coûteuses.

II.6.2 L'erreur absolue moyenne (MAE - Mean Absolute Error)

L'erreur absolue moyenne mesure la moyenne des écarts absolus entre les prédictions du modèle et les valeurs réelles.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - x_i |$$

Avantage : Moins sensible aux valeurs aberrantes que le RMSE. Plus interprétable en unités de la variable de sortie.

II.6.3 coefficient de détermination (R^2 - Coefficient of Determination)

Le coefficient de détermination mesure la proportion de la variance des données qui est expliquée par le modèle. Sa valeur est comprise entre 0 et 1 (ou négative si le modèle est pire qu'une moyenne constante).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - x')^2}$$

- Y' : moyenne des valeurs réelles

Interprétation :

- R2=1 : prédiction parfaite.
- *R2=0* : le modèle n'explique aucune variance.
- *R2<0* : modèle pire que la moyenne.

Chapitre III

Prédiction glycémique par les Méthodes d'apprentissage Automatique

III.1 Introduction

Un diabète non contrôlé entraîne fréquemment une hyperglycémie, qui, avec le temps, endommage gravement de nombreux organes aboutissant à des complications sévères telles que le coma, l'insuffisance rénale et rétinienne, les dysfonctionnements cardiovasculaires et cérébro-vasculaires, les troubles vasculaires périphériques, ainsi que des effets pathogènes sur le système immunitaire [23,24].

La gestion efficace du diabète repose sur une surveillance régulière de la glycémie. Cependant, les méthodes conventionnelles, souvent utilisées, impliquent des prélèvements sanguins fréquents, ce qui peut être inconfortable et dissuasif pour les patients. Ces limitations ont stimulé le développement de méthodes non invasives, notamment basées sur les biocapteurs optiques et les signaux physiologiques, comme le photopléthysmogramme (PPG), l'électrocardiogramme (ECG), …etc.

Les biocapteurs ECG exploitent l'activité électrique du cœur pour enregistrer des signaux électrocardiographiques riches en informations physiologiques. Ces signaux permettent d'extraire des caractéristiques précieuses, telles que les intervalles PR, QRS et QT, ainsi que la variabilité de la fréquence cardiaque, qui peuvent être corrélées à divers paramètres biologiques, y compris les fluctuations du taux de glucose sanguin. Toutefois, l'analyse de ces données brutes requiert des techniques avancées de traitement du signal et d'apprentissage automatique afin de transformer les signaux complexes en estimations fiables des niveaux de glucose.

Par conséquent, les applications des méthodes d'apprentissage automatique dans les capteurs optiques ont gagné en importance ces dernières années, en particulier dans la surveillance et l'amélioration de la précision de détection des capteurs pour des performances optimisées [25,26].

L'apprentissage automatique, branche de l'intelligence artificielle, est reconnue comme un domaine prometteur, capable de contribuer à la classification des patients (diabétiques ou non diabétiques) ou à la prédiction de leurs glycémies sanguines à partir d'un ensemble de données d'apprentissage. Le principal avantage de ces méthodes réside dans la capacité des algorithmes à

apprendre à partir des données physiologiques et des paramètres cliniques et d'appliquer cet apprentissage pour l'extraction des caractéristiques et des motifs permettant des prédictions futures de la glycémie [24].

Dans le cadre de cette problématique, ce chapitre présente une étude détaillée de l'utilisation du signal physiologique ECG, ainsi que de certains paramètres cliniques, pour la prédiction du taux de glycémie à l'aide d'algorithmes d'apprentissage automatique. Nous avons utilisé les architectures suivantes : KNN (K-Nearest Neighbors, K-Plus Proches Voisins), RF (Random Forest, Forêt Aléatoire) et DT (Decision Tree, Arbre de Décision), dans le cadre d'un modèle basé sur des capteurs de signaux physiologiques pour estimer la glycémie.

Pour l'entraînement du système, une base de données composée de 100 patients a été exploitée. Les signaux physiologiques et les paramètres cliniques, de cette base, ont été collectés à l'aide du moniteur HealthyPi v3, auprès de patients suivis à la maison du diabète de la Daïra de Saïda. Le traitement des signaux ECG et l'entrainement des algorithmes d'apprentissage automatique ont été réalisés via l'environnement Jupyter Notebook, en s'appuyant sur la plateforme Anaconda pour la gestion des dépendances et des bibliothèques Python.

L'évaluation des performances du système a été réalisée à l'aide de plusieurs indicateurs tels que la fonction de perte (Loss) pour mesurer l'erreur globale du modèle, L'erreur moyenne absolue MAE (Mean Absolute Error) pour quantifier l'erreur moyenne absolue, l'erreur quadratique moyenne MSE (Mean Squared Error) pour mesurer l'erreur quadratique moyenne, la racine carrée de l'erreur quadratique moyenne RMSE (Root Mean Squared Error) pour évaluer la racine carrée de l'erreur quadratique moyenne, et R^2 (coefficient de détermination) pour estimer la qualité d'ajustement du modèle aux données.

III. 2 Méthodologie de l'étude

III.2.1 Description de la base de données

La première étape d'un algorithme de Supervised Learning consiste donc à importer un Dataset qui contient les exemples que la machine doit étudier.

Ce Dataset inclut toujours 2 types de variables :

1. Une variable objective (target) y.
2. Une ou plusieurs variables caractéristiques (features) x.

Pour la mise en œuvre des différents algorithmes d'apprentissage automatique dans cette étude, une base de données formée de 100 patients a été utilisée.

Dans cette base de données, les patients concernés sont des hommes, femmes et enfants de différents âges. Les données collectées contiennent des informations sur 100 patientes et 13 caractéristiques physiologiques et cliniques de chaque chacun. Ces caractéristiques utilisées pour l'extraction des caractéristiques du modèle pour l'apprentissage et la génération du motif de prédiction de la glycémie sont les suivants (Tableau III.1) :

- Signal physiologique ECG,
- SpO2,
- Age,
- Pression artérielle systolique,
- Pression artérielle diastolique,
- Poids,
- Glycémie,
- Fréquence cardiaque,
- Fréquence respiratoire,
- Température corporelle,
- Pression artérielle moyenne (PAM).

Tableau III.1 : Caractéristiques cliniques et physiologique collectées auprès d'un patient en utilisant le moniteur moniteur HealthyPi v3

	ECG	PPG	Resp	Spo2	Age	sys	dias	Poids	Gl	Fc	Fr	Tempr
1	-0.00038	-2.4E+07	-1.09755	98	50	9	5	67	0.7	65	34	35.48
2	0.001642	-2.3E+07	-1.08561	98	50	9	5	67	0.7	65	34	35.48
3	0.001656	-2.1E+07	-1.07268	98	50	9	5	67	0.7	65	34	35.48
4	0.004665	-2E+07	-1.06175	98	50	9	5	67	0.7	65	34	35.48
5	0.001677	-1.8E+07	-1.04584	98	50	9	5	67	0.7	65	34	35.48
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
4001	0.028959	12196864	-0.18032	98	50	9	5	67	0.7	65	34	35.48

Chaque fichier CSV correspond à un patient unique, et contient un ensemble de mesures collectées pendant une session d'enregistrement.

Chaque fichier CSV contient 4000 lignes (echantillons) pour chaque signal physiologique ECG, PPG, Resp, associée aux paramètres cliniques suivants :

1. ECG – Signal électrocardiographique (µV)
2. PPG – Signal photopléthysmographique (non normalisé)
3. Resp – Signal respiratoire
4. SpO2 – Saturation en oxygène du sang (%)
5. Age – Âge (ans)
6. Sys – Pression artérielle systolique (mm Hg)
7. Dias – Pression artérielle diastolique (mm Hg)
8. Poids – Poids (kg)
9. Gl – Glycémie (valeur cible pour prédiction)
10. Fc – Fréquence cardiaque (bpm)
11. Fr – Fréquence respiratoire (rpm)
12. Tempr – Température corporelle (°C)
13. PAM–pression artificielle moyenne

Les colonnes Age, Sys, Dias, Poids, Fc, Fr, Tempr sont constantes et peuvent limiter l'apprentissage, cependant les colonnes des signaux ECG, PPG, et Respiratoire sont dynamiques (signaux variant en fonction du temps) et sont très utiles pour l'entraînement du système.

Un autre paramètre qui est la pression artérielle moyenne (PAM) est ajouté aux paramètres cliniques. Il est déterminé selon la formule suivante :

$$PAM = Sys + \frac{1}{3}(Sys - Dias)$$

La caractéristique 'glycémie, Gl' est considéré comme la variable dépendante (ou variable cible=Label), tandis que les autres caractéristiques en associées avec les caractéristiques extraites de l'ECG sont considérées comme des variables indépendantes (ou variables explicatives, en anglais, features).

Les informations d'entrée 'Attributs' sont utilisées par un modèle d'apprentissage automatique pour prédire une variable cible (ou target, en anglais).

III.2.1.1 Acquisition des données cliniques et physiologiques

L'acquisition des signaux physiologiques tels le signal ECG, PPG et Respiratoire (Resp) et certains paramètres cliniques tels que la saturation en oxygène SPO_2, La fréquence cardiaque, La

fréquence respiratoire, et la température du corps ont été effectué par l'utilisation du moniteur HealthyPi v3.[27]

a. Moniteur HealthyPi v3

HealthyPi v3 est un moniteur de signes vitaux capable de mesurer surveiller en direct les signaux ECG, PPG et Resp ainsi que la SpO2, le rythme cardiaque, le rythme respiratoire ainsi que la température corporelle (Fig. III. 1).

Fig. III.1: Photo du moniteur healthy Pi V3 et ses accessoires

Ce dispositif se présente sous forme d'un HAT (Hardware Attached on Top) open-source, multi-paramètre et complet pour la surveillance des signes vitaux du corps humain, conçu pour une utilisation avec le Raspberry Pi ainsi qu'en mode autonome, permettant ainsi de transformer un micro-ordinateur en un moniteur physiologique portable (Fig. III. 2).

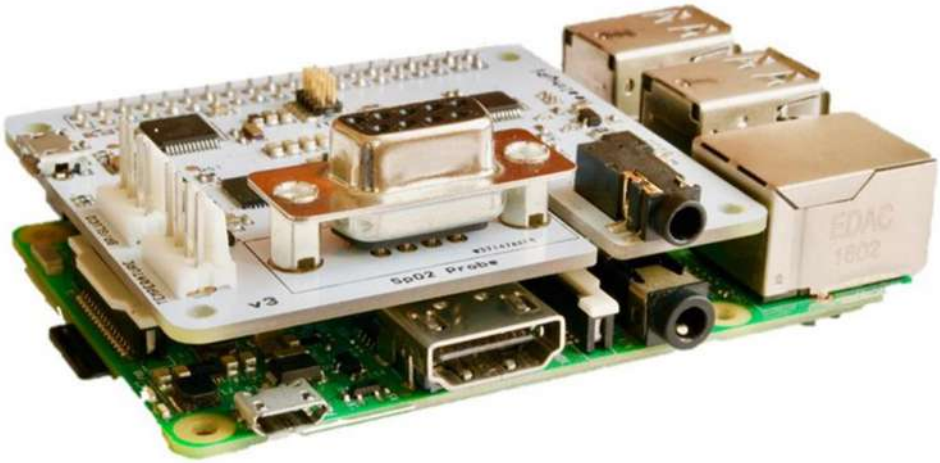

Fig. III. 2 : HealthyPi connecté au Raspberry Pi

Il permet une acquisition en continu des données via USB ou UART, ainsi qu'une visualisation temps réel, une intégration IoT (via MQTT) et la possibilité de programmer ou modifier facilement le firmware grâce à l'environnement Arduino/ATSAMD21

Le HealthyPi v3 mesure plusieurs paramètres vitaux, notamment :

- L'électrocardiogramme (ECG)
- Le Photopléthysmogramme (PPG)
- Le signal respiratoire (Resp)
- La saturation en oxygène (SpO_2)
- La température corporelle
- La fréquence respiratoire
- Photopléthysmographie
- La fréquence cardiaque

Les signaux vitaux doivent être enregistrés en environnement contrôlé puis exporté en format exploitable (.csv/.txt) pour leur traitement ultérieur.

Avantages : Portable, abordable, compatible avec des logiciels comme Python pour le traitement des données.

Limites : Sensibilité au bruit, nécessité de pré-traitement, ou contraintes liées à l'utilisation en conditions réelles.

b. Configuration Healthy Pi v3 pour collecter les signaux

- **Connexion matérielle**

Le module HealthyPi v3 est installé directement sur un Raspberry Pi 3 à travers le port GPIO (HAT). Le Raspberry Pi fournit lui-même l'alimentation électrique de ce module. Les données sont transmises par le biais de :

  - Port série UART pour l'envoi continu des données vers le Raspberry.
  - Enregistrement des signaux dans un fichier CSV via un script Python.

- **Placement des électrodes ECG**

L'enregistrement ECG a été effectué en utilisant trois électrodes disposées conformément à la configuration Lead I : sur le bras droit (RA, couleur rouge), le bras gauche (LA, couleur bleue) et la jambe droite ou la partie inférieure du ventre (RL – référence, couleur noire) (Fig. III. 3). Pour garantir un contact optimal, des électrodes médicales contenant du gel conducteur ont été employées. Avant l'application, la peau a été purifiée pour assurer un signal clair et constant.

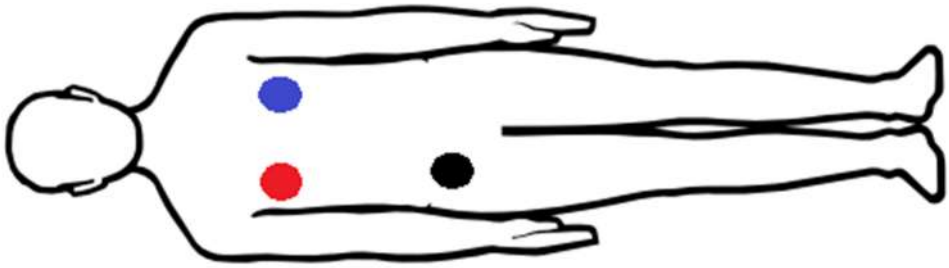

Fig. III. 3 : Emplacement des électrodes ECG

- **Environnement de test**

Les acquisitions ont été effectuées dans un environnement calme et tempéré, avec peu de perturbations électromagnétiques. Le sujet examiné était allongé et au repos, dans le but d'acquérir un signal stable et utilisable pour le traitement.

- **Configuration logicielle**

L'environnement logiciel utilisé comprend :

  - Le firmware fourni par ProtoCentral (HealthyPi v3) chargé sur la carte.
  - Un script Python utilisant pyserial pour écouter et enregistrer les données.

- Visualisation possible via la GUI officielle (Processing) ou logiciel tiers.

Les signaux ECG ont été extraits à une fréquence d'échantillonnage de 125 Hz, suffisante pour capturer les détails du cycle cardiaque

- **Calibration et vérification**

Bien que le module HealthyPi v3 ne nécessite pas de calibration manuelle fréquente, les étapes suivantes ont été respectées pour garantir la qualité du signal :

- Test à vide (électrodes déconnectées) pour vérifier l'absence de bruit parasite.
- Vérification visuelle du tracé ECG (onde P, complexe QRS, onde T).
- Réglage du **gain logiciel** si nécessaire via le GUI ou script de configuration.

- **Format de sauvegarde**

Les données brutes ECG sont stockées dans des fichiers .csv, avec des échantillons sur plusieurs secondes/minutes selon la session. Chaque ligne correspond à un instant d'acquisition.

II.2.1.2 Méthode de mesure de la glycémie

Pour la plupart des diabétiques, il n'est pas très agréable de contrôler régulièrement leur glycémie. Les dispositifs conventionnels de surveillance du glucose utilisent la méthode électrochimique, qui nécessite le prélèvement d'une petite quantité de sang dans le corps, soit par une piqûre au doigt, soit par l'implantation d'une fine lancette sous-cutanée (Fig. III.4). La différence entre les deux méthodes est que la première ne fournit qu'un instantané du niveau de glucose à un moment précis et ne nécessite pas d'assistance professionnelle, c'est pourquoi on l'appelle dispositif d'auto-surveillance de la glycémie (SMBG). Le second fournit une surveillance continue, c'est pourquoi il est appelé dispositif de surveillance continue de la glycémie (CGM).

Cependant, ces deux méthodes provoquent non seulement une gêne et une douleur après une utilisation répétée, mais présentent également des risques d'infection et de lésions tissulaires, ce qui entraîne une mauvaise observance des patients pour les mesures quotidiennes qui leur sont assignées.

Les mesures de la glycémie, variable cible (target), utilisées dans cette base de données (dataset) ont été prélevées par la méthode conventionnelle (invasive) couramment pratiquée dans le domaine médical. Elle consiste à prélever une goutte de sang au bout du doigt à l'aide d'une petite lancette à usage unique. Une fois la goutte obtenue, elle est dépose sur une bandelette

réactive, et insérée ensuite dans un glucomètre pour obtenir la valeur de la glycémie (Fig. III.4). C'est cette méthode qui nous a permis de collecter la valeur de référence à comparer avec les prédictions faites à partir du signal ECG.

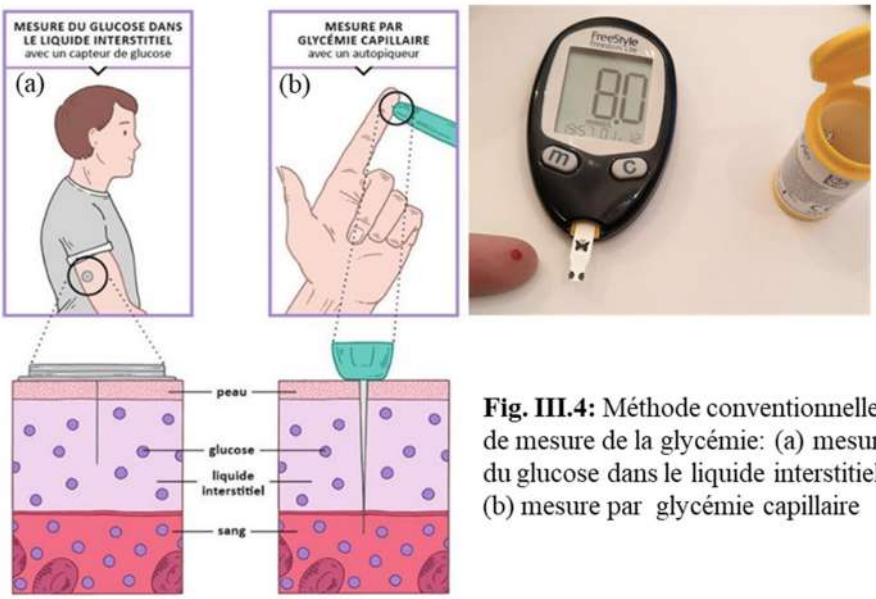

Fig. III.4: Méthode conventionnelles de mesure de la glycémie: (a) mesure du glucose dans le liquide interstitiel, (b) mesure par glycémie capillaire

III.3 Système proposé pour la prédiction non invasive de la glycémie

III.3.1 Schéma synoptique – Système de prédiction non invasive de la glycémie

La figure III.5 schématise les différentes étapes du système d'apprentissage automatique (Machine learning) proposé pour la prédiction non invasive de la glycémie. Ces étapes seront traitées en détail dans ce qui suivra.

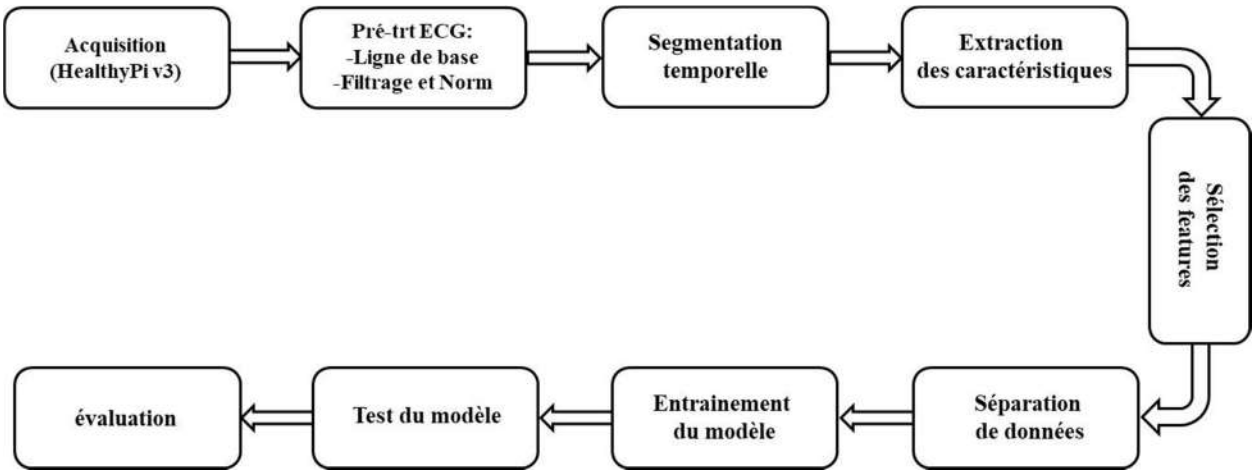

Fig. III.5 : Schéma synoptique du modèle proposé pour le système de prédiction non invasive de la glycémie par l'apprentissage automatique (Machine learning)

III.3.2 Prétraitement des signaux

Les signaux physiologiques, tels que l'électrocardiogramme (ECG), le photopléthysmogramme (PPG) ou encore les signaux de respiration, sont des données biologiques précieuses permettant d'analyser l'état de santé d'un individu. Cependant, ces signaux sont souvent exposés à divers bruits et artefacts provenant de sources internes (mouvements corporels, activité musculaire) ou externes (interférences électriques, bruit thermique des capteurs). La figure III.6 montre le signal ECG brut après enregistrement et avant tout prétraitement, exposé au bruit et présentant une ligne de base déformée. Ces perturbations peuvent masquer ou altérer les informations essentielles contenues dans le signal d'origine, rendant l'interprétation clinique ou l'analyse automatique peu fiable. Ainsi, il est indispensable d'appliquer un prétraitement rigoureux, notamment à travers des techniques de filtrage, pour atténuer ces bruits sans déformer les composantes physiologiques importantes. Le filtrage permet, par exemple, de supprimer le bruit de fréquence industrielle (50/60 Hz), d'éliminer les dérives de la ligne de base ou de réduire les interférences de hautes fréquences dues à l'activité musculaire. Ce processus est donc une étape cruciale pour garantir la qualité des signaux avant toute extraction de caractéristiques, classification ou interprétation médicale.

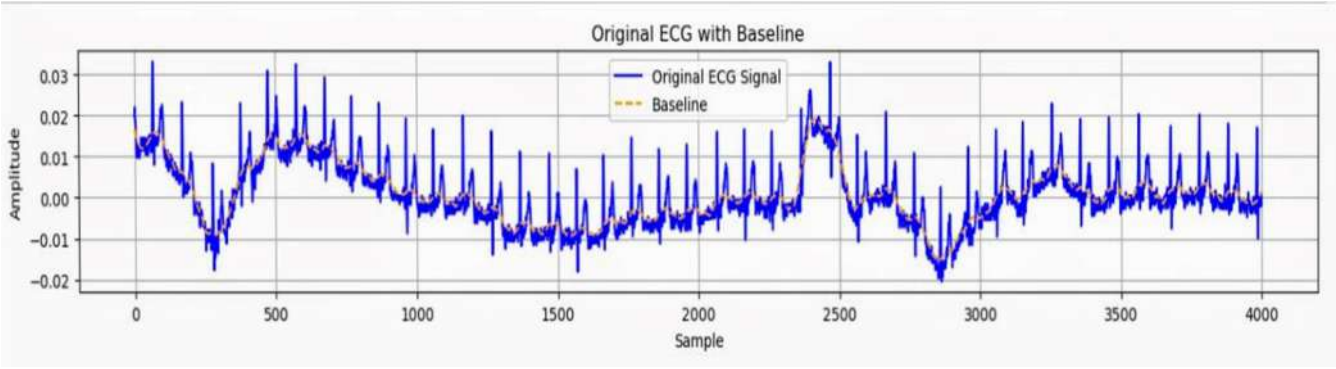

Fig. III.6: Signal ECG brut avant prétraitement

III.3.2 Lecture des données

Avant tout traitement, il faut d'abord lire les donnes du signal ECG à partir du fichier csv du patient par la commande suivante :

```
# Liste pour stocker les lignes de tous les patients
all_patients_data = []

# Boucle sur les fichiers de 1 à 101
for i in range(1, 101):
    filename = f"features{i}.csv"
    if not os.path.exists(filename):
        print(f"⚠ Fichier {filename} introuvable.")
        continue

    df = pd.read_csv(filename, delimiter=';')
```

III.3.2.1 Réglage de base de ligne

Pour optimiser la qualité du signal ECG et supprimer les variations lentes (bruit à basse fréquence) causées par le mouvement ou la respiration. Une rectification de la ligne de base a été mise en œuvre. Ainsi, on a fait appel à une fonction nommée baseline_correction_ecg. Cette fonction débute par l'évaluation de la référence du signal grâce au filtre de Savitzky-Golay, un filtre numérique qui permet d'adoucir les données sans modifier significativement les propriétés du signal. Ce filtre est mis en œuvre en utilisant une fenêtre de 50 échantillons et un polynôme d'ordre 3.

Par la suite, la référence estimée est déduite du signal ECG initial afin de produire un signal ajusté qui oscille autour de zéro (Fig. III.7). La fonction retourne à la fois le signal corrigé et la ligne de base estimée.

- **Filtre de Savitzky-Golay**

Le filtre de Savitzky-Golay est un filtre numérique utilisé pour lisser des signaux bruités tout en préservant la forme et les caractéristiques importantes du signal, comme les pics, les pentes et les minima/maxima. Contrairement aux filtres passe-bas classiques qui peuvent atténuer ou déformer les signaux rapides (comme les ondes du signal ECG), le filtre de Savitzky-Golay applique une régression polynomiale locale à une fenêtre glissante de points.

Plus précisément, à chaque position de la fenêtre, il ajuste un polynôme d'un certain ordre (ex. : 2 ou 3) sur les points de données de cette fenêtre, puis remplace le point central par la valeur du polynôme ajusté. Ce processus est répété tout au long du signal, ce qui permet d'obtenir un signal lissé, sans supprimer les détails importants.[28]

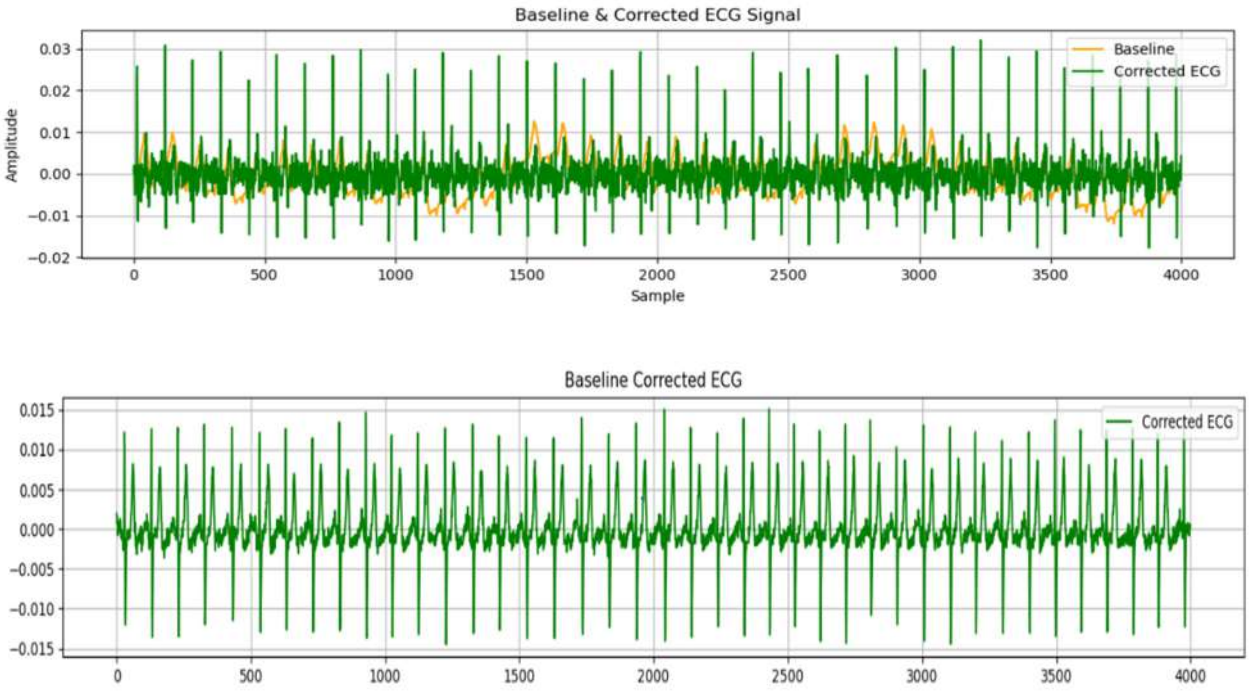

Fig. III.7: Correction de la ligne de base

III.3.2.2 Filtrage

Pour améliorer la qualité du signal ECG et réduire les bruits indésirables, on applique un filtrage passe-bande. Cette opération est réalisée par la fonction `filter_ecg`, qui utilise un filtre de Butterworth du deuxième ordre. Le filtre est conçu pour laisser passer les fréquences comprises entre 0.5 Hz et 45 Hz, ce qui correspond à la bande utile pour l'analyse du signal ECG (les composantes physiologiques du cœur se trouvent généralement dans cet intervalle). Les fréquences inférieures à 0.5 Hz (comme la dérive de la ligne de base) et supérieures à 45 Hz (comme les interférences électriques ou le bruit musculaire) sont ainsi atténuées.

Le signal est filtré à l'aide de la fonction `filtfilt`, qui applique le filtre dans les deux sens (aller-retour), ce qui permet de préserver la phase du signal. Le paramètre `fs` correspond à la fréquence d'échantillonnage, ici fixée à 125 Hz, et il est utilisé pour normaliser les fréquences de coupure selon le critère de Nyquist.

- **Le filtre de Butterworth**

Le filtre de Butterworth est un filtre passe-bas, passe-haut, ou passe-bande très utilisé en traitement du signal. Il a été conçu par Stephen Butterworth en 1930 pour fournir une réponse en fréquence aussi plate que possible dans la bande passante (c'est-à-dire sans ondulation). Cela signifie qu'il ne déforme pas l'amplitude des fréquences qu'il laisse passer, ce qui le rend idéal

pour des applications où la fidélité du signal est importante, comme le traitement des signaux biomédicaux (ECG, EEG, etc.).

Ce filtre est caractérisé par un ordre (le nombre de pôles) qui détermine la raideur de la pente de coupure. Plus l'ordre est élevé, plus la transition entre la bande passante et la bande atténuée est abrupte.[29]

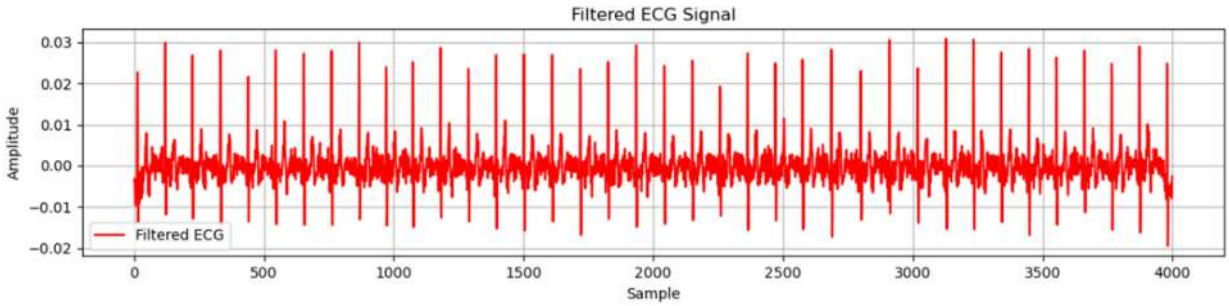

Fig. III.8 : signal ECG filtré

III. 3.3 Matrice de corrélation

Après avoir vérifié qu'il n'y a aucune valeur manquante ou nulle dans notre base de données et que l'ensemble des données a été normalisé dans l'intervalle de 0 à 1, nous avons effectué une sélection des caractéristiques en examinant la matrice de corrélation des données afin d'identifier les attributs les plus significatifs. La matrice de corrélation nous aide à calculer le coefficient de corrélation entre les attributs indépendants et dépendants pour déterminer l'interdépendance entre tous les attributs.

Les valeurs supérieures à 0,5 ou inférieures à −0,5 indiquent généralement une forte corrélation entre deux attributs, tandis que les valeurs proches de 0 montrent peu ou pas de corrélation entre eux.

Dans notre étude, en analysant les valeurs du coefficient de corrélation dans la Fig. III.9 pour les attributs par rapport à l'attribut « Glycémie » et en utilisant 0,01 comme valeur seuil, nous avons conclu que toutes les attributs sont significatifs, car ils présentent un coefficient de corrélation supérieur à 0.01 avec l'attribut « glycémie», indiquant ainsi une forte dépendance entre ces attributs et l'attribut cible.

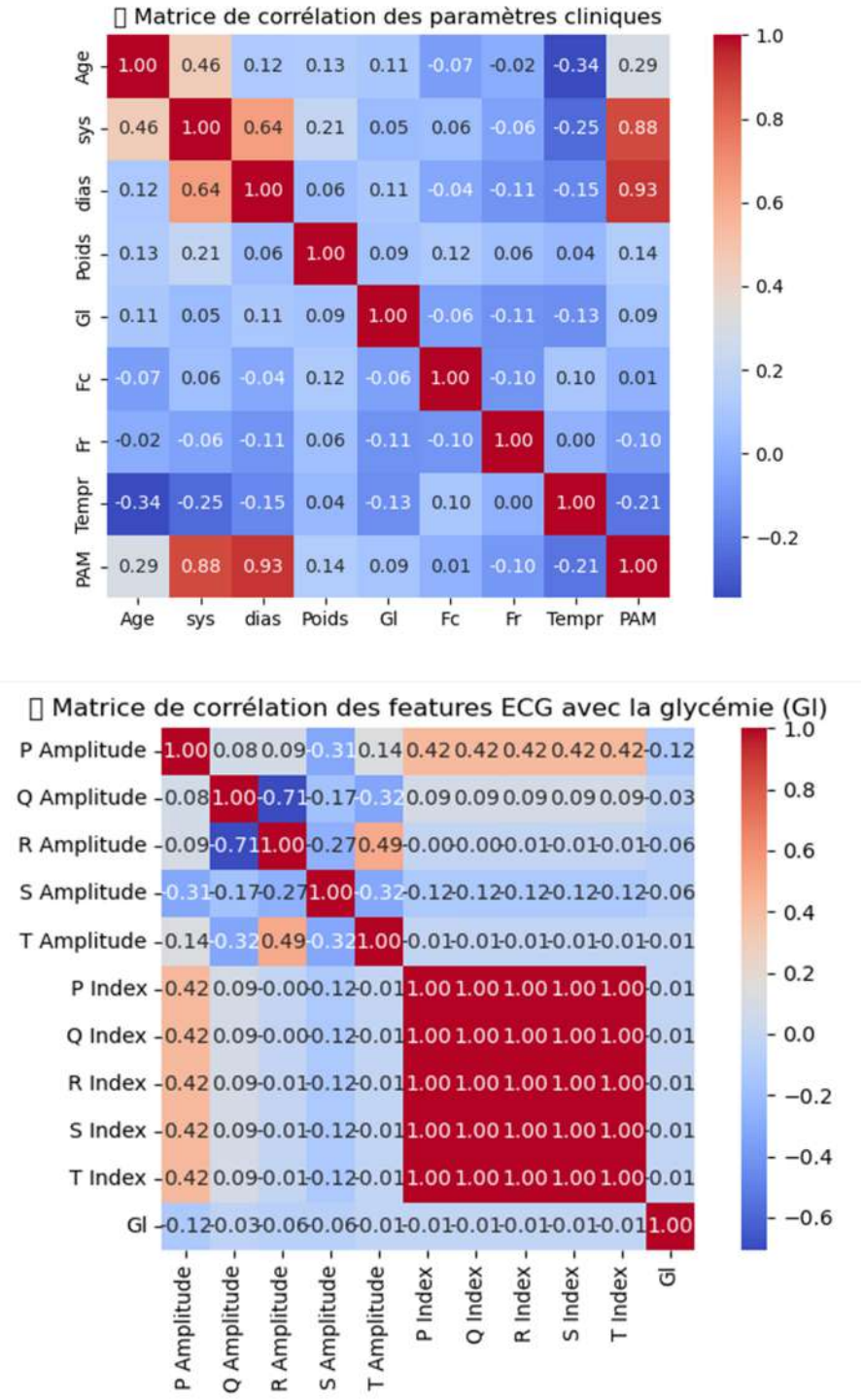

Fig. III. 9 : Matrices de corrélation

III.3.4 Segmentation du signal ECG

La segmentation du signal ECG est une étape essentielle dans le traitement. La division du signal ECG en segments constitue une phase cruciale dans le processus de traitement, car elle

facilite l'identification de parties significatives et utiles du signal pour des fins d'analyse ou d'apprentissage automatique.

Dans notre projet, cette segmentation consiste à diviser le signal ECG en plusieurs fenêtres de durée fixe, afin d'en extraire des caractéristiques (features) pour chaque segment indépendamment.

Le but de la segmentation est :

- Diviser les signaux continus de grande taille en segments plus petits et plus facilement analysables.
- Pour détecter des événements spécifiques (comme les ondes P, QRS, T) dans chaque segment.
- Pour faciliter l'extraction de caractéristiques statistiques ou temporelles localisées.

III.3.5 Extraction de caractéristiques

L'extraction de caractéristiques est une phase déterminante dans le traitement du signal ECG pour l'apprentissage automatique. Elle vise à convertir le signal brut en une série de valeurs numériques représentatives (connues sous le nom de features), qui permettent de saisir les éléments clés de l'activité cardiaque et d'optimiser la prédiction du taux de sucre dans le sang.

La fonction '`detect_ecg_waves`' permet de détecter les différentes ondes caractéristiques du cycle cardiaque dans un signal ECG, à savoir les ondes P, Q, R, S et T (Fig. III. 10).

Pour chaque onde détectée, la fonction enregistre sa position (indice dans le signal) et son amplitude. Le résultat est un dictionnaire contenant les coordonnées des cinq ondes ECG typiques pour chaque battement cardiaque, ce qui permet d'extraire des caractéristiques cliniquement pertinentes pour l'analyse ou la prédiction.

Après la segmentation et la détection des ondes ECG (P, Q, R, S, T), plusieurs types de caractéristiques ont été extraites pour chaque segment du signal :

III.3.5.1 Caractéristiques temporelles

Ces caractéristiques sont calculées à partir des distances entre les pics des ondes ECG :

- **Intervalle PR** : temps entre les ondes P et R

- **Durée du complexe QRS** : distance entre Q et S
- **Intervalle QT** : distance entre Q et T
- **Fréquence cardiaque** : inverse du temps entre deux pics R successifs (RR interval)

III.3.5.2 Caractéristiques statistiques

Elles sont calculées sur les segments de signal :

- Moyenne, médiane, écart-type.
- Valeur maximale et minimale.

III.3.5.3 Caractéristiques morphologiques

Ce sont des mesures relatives aux amplitudes des ondes ECG :

- Amplitudes des ondes P, Q, R, S, T.
- Rapport R/S ou R/Q.
- Surface sous le complexe QRS.

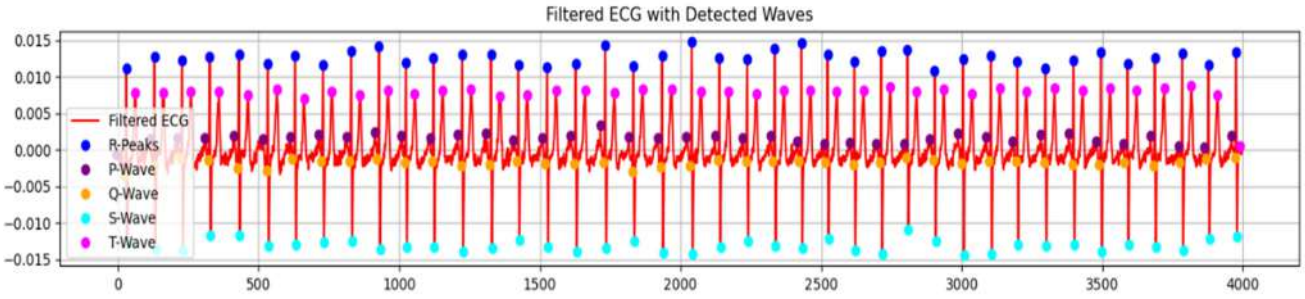

Fig. III.10: Signal ECG filtre avec détection des pics P, Q, R, S et T

III.3.6 Sélection des caractéristiques

Une fois les caractéristiques extraites à partir du signal ECG, nous avons procédé à une sélection des caractéristiques (features) les plus pertinentes pour améliorer la performance du modèle de prédiction de la glycémie. En effet, toutes les caractéristiques extraites ne sont pas forcément utiles : certaines peuvent être redondantes, corrélées entre elles, ou même bruitées. Pour cela, nous avons utilisé des méthodes de sélection classiques comme la variance (pour éliminer les features peu variables) ou des tests statistiques (comme le test ANOVA ou la corrélation de Pearson) afin d'évaluer l'influence de chaque caractéristique sur la variable cible (le taux de glycémie).

L'objectif est de réduire la dimensionnalité des données, d'accélérer l'apprentissage et d'éviter le sur-apprentissage (overfitting) tout en gardant un maximum d'informations utiles pour la prédiction.

L'analyse des 5 premières lignes du fichier csv d'un patient donne :

```
[33]: df.head()
```

[33]:

	ECG	PPG	Resp	Spo2	Age	sys	dias	Poids	Gl	Fc	...	P Amplitude	Q Index	Q Amplitude	R Index	R Amplitude	S Index	S Amplitude	T Index	T Amplitude
0	-0.002880	-16871104.0	-0.279638	94	36	11	7	59	2.14	80	...	0.002199	-1.0	-0.000279	19.0	0.008746	22.0	-0.008617	56.0	0.004072
1	0.000114	-16128320.0	-0.280191	94	36	11	7	59	2.14	80	...	0.002611	280.0	-0.001052	289.0	0.005740	292.0	-0.008959	326.0	0.003330
2	-0.000893	-15331456.0	-0.278748	94	36	11	7	59	2.14	80	...	0.002825	418.0	-0.001078	428.0	0.005655	431.0	-0.009082	466.0	0.003501
3	-0.000898	-14687616.0	-0.279310	94	36	11	7	59	2.14	80	...	0.001516	563.0	-0.000451	566.0	0.005871	570.0	-0.010927	604.0	0.003370
4	0.000095	-13911168.0	-0.278877	94	36	11	7	59	2.14	80	...	0.001729	700.0	-0.001116	704.0	0.005722	708.0	-0.008678	741.0	0.003118

5 rows × 23 columns

Il est clair que chaque colonne représente une caractéristique, il y a donc 23 caractéristiques pour chaque patient y compris la variable cible.

```
df.shape
```

```
(4000, 23)
```

Chaque colonne ECG est composée de 4000 échantillons (fréquence d'échantillonnage utilisée est fs=125 Hz).

III.3.7 Séparation du jeu de données

Lorsqu'on développe un modèle d'apprentissage supervisé, il est essentiel de séparer les données disponibles en deux parties principales :

- Ensemble d'entraînement (training set) : utilisé pour entraîner le modèle, c'est-à-dire pour ajuster ses paramètres internes à partir des exemples connus (généralement 80 % des données).
- Ensemble de test (test set) : utilisé pour évaluer les performances du modèle sur des données qu'il n'a jamais vues (généralement 20 % des données).

La répartition 80/20 est un standard empirique bien accepté dans la communauté scientifique. Elle permet :

- D'avoir assez de données pour bien entraîner le modèle.
- D'avoir un échantillon indépendant suffisamment représentatif pour estimer ses performances réelles.

III.4 Apprentissage automatique

L'apprentissage automatique (machine learning en anglais) est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés pour chaque tâche. L'idée principale est de fournir à un algorithme un grand nombre d'exemples (appelés données d'entraînement), afin qu'il identifie automatiquement des relations, des motifs ou des comportements cachés dans ces données.

Une fois entraîné, le modèle peut ensuite faire des prédictions ou prendre des décisions sur de nouvelles données jamais vues auparavant. Dans le cadre de notre projet, nous avons utilisé l'apprentissage automatique pour prédire le taux de glycémie à partir des caractéristiques extraites des signaux ECG, sans utiliser de capteur invasif.

III.4.1 Apprentissage supervisée

Dans ce type d'apprentissage, le modèle apprend à partir d'un ensemble de données étiquetées, c'est-à-dire que chaque exemple comporte des entrées (features) et une valeur de sortie connue (étiquette ou cible).

L'objectif est de prédire la sortie pour de nouvelles entrées.

Exemples : régression (prédiction d'une valeur numérique), classification (prédiction d'une catégorie)

Dans notre projet, nous utilisons clairement l'apprentissage supervisé, car :

- Nous disposons d'un ensemble de données étiquetées : chaque enregistrement ECG est associé à une valeur réelle de glycémie, que nous avons mesurée manuellement à l'aide d'un glucomètre.
- Notre objectif est de prédire une valeur continue, à savoir le taux de glycémie, ce qui relève d'une tâche de régression supervisée.

III.4.2 Modèles utilisés

III.4.2.1 K plus proches voisins KNN (K-Nearest Neighbors)

KNN est un algorithme non paramétrique, ce qui indique qu'il ne suppose rien sur la distribution des données. Ce dernier est fréquemment appliqué dans les missions de régression et de classification, du fait de sa facilité de mise en pratique et de son principe basé exclusivement sur la similitude entre les échantillons.

Le modèle KNN fonctionne pas à pas pour un problème de régression comme dans le cas de notre étude (prédiction du taux de glycémie à partir d'un signal ECG) :

1. **Distance :** Pour une observation inédite (comme un cas de patient), le KNN détermine la distance (souvent euclidienne) entre cette dernière et toutes les autres données dans l'ensemble d'entraînement.

   Exemple :

$$d = \sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2) + (\dots\dots)}$$

   Ou x_i sont les caractéristiques (features) extraites du signal ECG.

2. **Recherche des K voisins les plus proches**

   Une fois les distances calculées, l'algorithme sélectionne les K échantillons les plus proches (c'est-à-dire les plus similaires en termes de caractéristiques).

3. **Prédiction**

   Dans ce cas de régression, il prend la moyenne des valeurs de glycémie de ces K voisins pour faire une estimation du taux de glycémie du nouvel échantillon.

$$y' = \frac{1}{k}\sum_{i=1}^{k} yi$$

Le modèle KNN est caractérisé par :

- Simplicité : Facile à implémenter et à comprendre.
- Pas besoin d'entraînement complexe : Le modèle apprend "à la volée".
- Bonne performance sur de petites bases de données ou des données bien normalisées.

Par contre, il présente certaines limites

- Lent avec beaucoup de données, car il doit comparer à tous les points d'entraînement.
- Sensible au bruit et aux données non normalisées.
- Choix de K : Trop petit, le modèle est instable ; trop grand, il devient trop général.

III.4.2.2 Random Forest

L'algorithme de Forêt Aléatoire (Random Forest) est une méthode d'apprentissage supervisé qui repose sur le concept des arbres de décision. C'est une technique d'ensemble qui

fusionne plusieurs arbres de décision afin d'améliorer la précision, la robustesse et la capacité de généralisation du modèle.

Pour faire simple, Random Forest génère une « forêt » d'arbres, chaque arbre apprenant différemment à partir de segments aléatoires des données, puis moyennant leurs prévisions (dans le cas de la régression) ou via un vote (pour la classification).[30]

1. **Création d'échantillons (Bagging)**
   - À partir de la base de données d'origine, Random Forest génère plusieurs sous-échantillons aléatoires avec remplacement (technique appelée bootstrap).
   - Chaque sous-échantillon sert à entraîner un arbre de décision différent.
2. **Construction d'arbres de décision**
   - Pour chaque arbre, à chaque noeud de décision, l'algorithme ne teste qu'un sous-ensemble aléatoire de variables (features).
   - Cela introduit encore plus d'aléa, ce qui réduit la corrélation entre les arbres.
3. **Agrégation du résultat**
   - En régression (comme dans cette étude), Random Forest moyenne les prédictions de tous les arbres :

$$y' = \frac{1}{N}\sum_{i=1}^{N} y_i^{(tree)}$$

   - En classification, il effectue un vote majoritaire entre les arbres.

Le modèle Random forest est caractérisé par :

- Très bonne performance sur des données complexes et bruitées.
- Résistant au surapprentissage (grâce au bagging et à l'aléa introduit).
- Donne une estimation de l'importance des variables (utile pour la sélection de caractéristiques).
- Prend en charge à la fois la classification et la régression.

D'autre part, il présente certaines limites tels que :

- Plus lent que les modèles simples comme KNN ou régression linéaire.
- Moins interprétable qu'un arbre unique.
- Nécessite plus de ressources (mémoire et calcul).

Dans notre contexte, Random Forest est utilisé pour :

- ✓ Prédire la glycémie à partir des caractéristiques extraites des signaux ECG.
- ✓ Évaluer l'importance de chaque caractéristique pour déterminer lesquelles influencent le plus la glycémie.
- ✓ Comparer ses résultats avec ceux du modèle KNN.[31]

III.4.2.3 Decision Tree

Cet algorithme de classification et de régression supervisé est basé sur l'arbre de décision. Il illustre les décisions à travers un arbre, dans lequel :

- Chaque nœud interne représente un test sur une caractéristique.
- Chaque branche représente un résultat de ce test.
- Chaque feuille produit une estimation (catégorie ou valeur numérique).

1. **Diviser les données**

L'algorithme commence à la racine de l'arbre et cherche la meilleure caractéristique (feature) pour diviser les données en deux groupes afin d'obtenir les sous-groupes les plus "purs" possibles (c'est-à-dire homogènes en termes de sortie attendue).

- En classification, la pureté est mesurée avec des critères comme :
  - L'entropie (information gain),
  - L'indice de Gini.
- En régression, on utilise :
  - L'erreur quadratique moyenne (MSE) Où
  - L'erreur absolue moyenne (MAE)

2. **Répéter la division**

L'algorithme répète ce processus récursivement pour chaque sous-ensemble obtenu, créant ainsi un arbre binaire.

3. **Condition d'arrêt**

L'arbre s'arrête :

- Soit quand toutes les données dans un nœud sont identiques (pures),
- Soit lorsqu'un critère de profondeur maximale ou de taille minimale d'échantillon est atteint (pour éviter le surapprentissage).

Les avantages de ce modèle sont :

- Facile à interpréter et à visualiser.
- Peut fonctionner avec peu de prétraitement des données (pas besoin de normalisation).
- Utilise des règles logiques simples et compréhensibles.

Ses limites peuvent être énumérées comme suite :

- Risque élevé de sur-apprentissage (overfitting), surtout si l'arbre est trop profond.
- Moins performant que des méthodes d'ensemble comme Random Forest ou XGBoost, sauf sur de très petits jeux de données.

Dans cette étude, ce modèle d'arbre de décision est utilisé pour :

- ✓ Prédire la valeur de la glycémie à partir des caractéristiques du signal ECG (durée, amplitude, etc.).
- ✓ Analyser les règles de décision que le modèle apprend, ce qui est utile pour l'interprétation médicale.[32]

III.4.3 Les métriques d'évaluation

III.4.3.1 Coefficient de Détermination R² (Coefficient of Détermination)

Il est définit par :

$$R^2 = 1 - \left(\sum_{i=1}^{n}(y_i - y_i')^2\right)\frac{1}{\sum_{i=1}^{n}(y_i - y'')^2}$$

- y_i : valeur réelle (glycémie mesurée)
- y_i' : valeur prédite par le modèle
- y'': moyenne des valeurs réelles
- n est le nombre total d'échantillons

- $R^2 = 1$: prédiction parfaite
- $R^2 = 0$: le modèle n'explique rien (équivalent à la moyenne)
- $R^2 < 0$: le modèle est pire qu'un modèle constant

III.4.3.2 Erreur Moyenne Absolue MAE (Mean Absolute Error)

MAE mesure la moyenne des écarts absolus entre les prédictions et les vraies valeurs.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - y^{i'} |$$

- Facile à comprendre : elle donne l'erreur moyenne en unités réelles (par exemple, g/L pour la glycémie).
- Plus robuste que MSE aux valeurs aberrantes (outliers).

III.4.3.2 Erreur Quadratique Moyenne MSE (Mean Squared Error)

L'erreur quadratique moyenne MSE est une métrique d'évaluation utilisée pour mesurer la qualité d'un modèle de régression. Elle représente la moyenne des carrés des écarts entre les valeurs réelles (vraies) et les valeurs prédites par le modèle.

Autrement dit, elle mesure l'écart global entre ce que le modèle prédit et la réalité. Plus cette erreur est faible, plus le modèle est précis.

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_i} - \mathrm{y_i'}\right)^2$$

- Une valeur de MSE faible indique que les prédictions du modèle sont proches des valeurs réelles.
- Une valeur élevée signifie que les prédictions sont souvent éloignées de la réalité.
- Comme l'erreur est élevée au carré, MSE pénalise fortement les grandes erreurs, ce qui est utile pour détecter des modèles imprécis.

III.4.3.2 Racine de l'Erreur Quadratique Moyenne RMSE (Root Mean Squared Error)

RMSE est une mesure statistique qui évalue la différence moyenne entre les valeurs prédites par un modèle et les valeurs réelles, en tenant compte de la racine carrée de la moyenne des carrés des erreurs.

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} \left(y_i - y_i'\right)^2\right)}$$

- RMSE donne une indication directe de l'erreur moyenne absolue entre les prédictions et les vraies valeurs, dans la même unité que la variable cible.

- Plus RMSE est faible, plus le modèle est précis.
- MSE donne une erreur en unités au carré, alors que RMSE ramène l'erreur à l'échelle d'origine des données.

Le tableau ci-dessous, donne une comparaison entre ces principales métriques utilisées pour l'évaluation de nos modèles d'apprentissage automatiques :

Tableau III.1 : Comparaison des différentes métriques utilisées

Métrique	Utilité	Avantage	Inconvénient
R^2	Proportion de variance expliquée	Interprétation intuitive	Sensible aux mauvais ajustements
MAE	Erreur Moyenne Absolue	Simple et robuste	Ne punit pas fortement les grandes erreurs
MSE	Erreur quadratique moyenne	Punit les grandes erreurs	Plus sensible aux grandes valeurs aberrantes (outliers)
RMSE	Racine Erreur quadratique moyenne	Ramène l'erreur à l'échelle d'origine des données	Plus sensible aux grandes valeurs aberrantes (outliers)

III.4.4 Implémentation, Résultats et validations

III.4.4.1 Architecture proposé

Dans cette partie, nous proposons l'organigramme de l'architecture globale des différentes approches effectuées pour l'évaluation des performances des modèles d'apprentissage supervisé appliqués à notre base de données pour la prédiction non invasive du taux de glucose (Fig. III. 11). Nous avons testé et évalué trois algorithmes : K-Nearest Neighbors (KNN), Decision Tree et Random Forest.

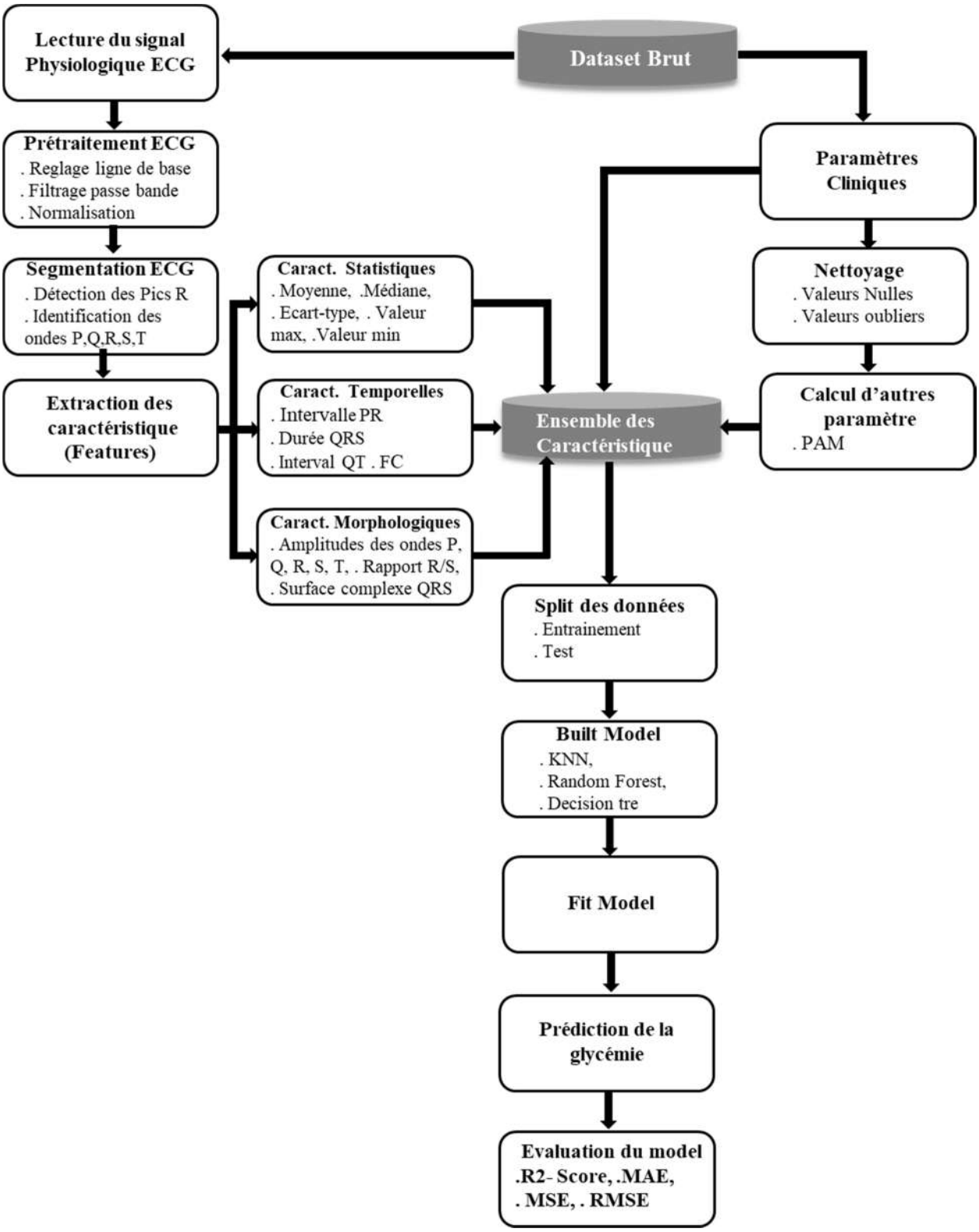

Fig. III.11 : Organigramme de l'architecture globale des différentes approches effectuées

III.4.4.1 Modèle KNN

Le modèle utilisé dans cette étude est KNeighborsRegressor avec un nombre de voisins fixé à k=27. Les données ont été divisées en un jeu d'entraînement (80 %) et un jeu de test (20 %) de manière aléatoire et reproductible (random_state=42). Les colonnes contenant des séries de valeurs amplitudes, indices des ondes ECG ont été réduites à leur valeur moyenne afin de constituer un vecteur de caractéristiques compact pour chaque patient.

Le paramètre principal k=n_neighbors=27, est le meilleur nombre de plus proches voisins donnant la meilleure prédiction de la glycémie basée sur la moyenne des 27 plus proches voisins dans l'espace des caractéristiques ECG.

```
# Modèle KNN de base
knn = KNeighborsRegressor()

# Recherche des meilleurs hyperparamètres (nombre de voisins K)
param_grid = {'n_neighbors': list(range(1, 31))}
grid_search = GridSearchCV(knn, param_grid, cv=5, scoring='neg_mean_squared_error')
grid_search.fit(X_train, y_train)

best_k = grid_search.best_params_['n_neighbors']
print(f"Meilleur nombre de voisins (K) trouvé : {best_k}")

# Entraîner le modèle avec le meilleur K
best_knn = KNeighborsRegressor(n_neighbors=best_k)
best_knn.fit(X_train, y_train)
```

Les résultats de la prédiction glycémique du modèle KNN comparés aux valeurs réelles sont donnés à la figure III.12 pour K=27.

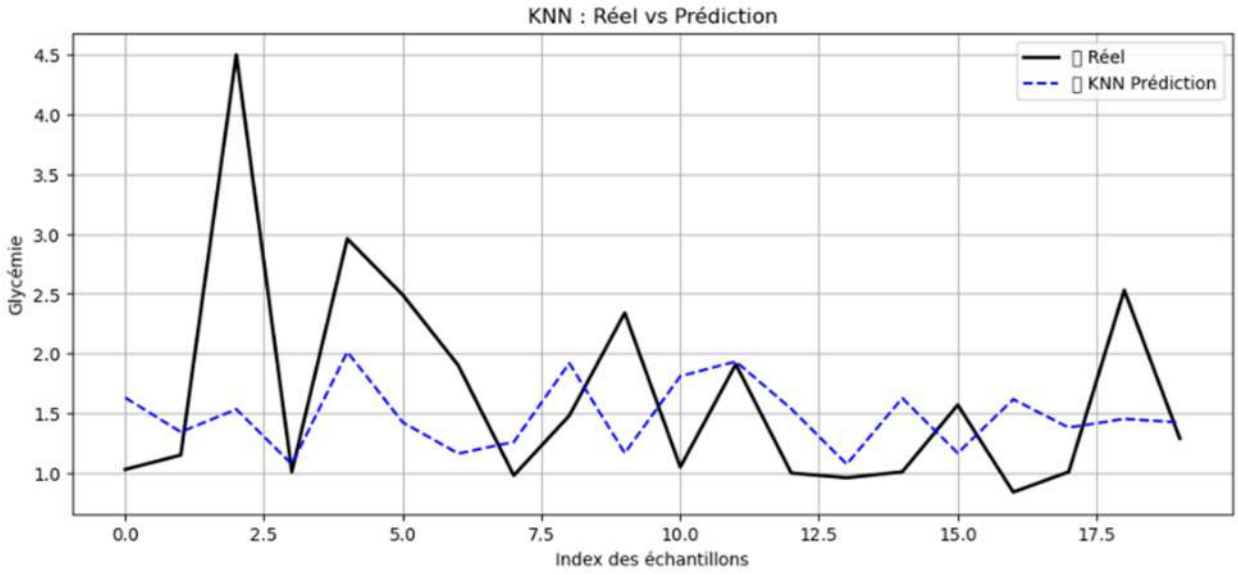

Fig. III.12 : Comparaison des Prédictions du modèle KNN

L'évaluation du modèle KNN basée sur les métriques citées ci-dessus a donné les résultats suivants :

```
# Affichage des résultats
print(f"📉 Mean Absolute Error (MAE): {mae:.2f}")
print(f"🔍 Mean Squared Error (MSE): {mse:.2f}")
print(f"📐 Root Mean Squared Error (RMSE): {rmse:.2f}")
print(f"📈 R² Score: {r2:.2f}")
```

```
📉 Mean Absolute Error (MAE): 0.44
🔍 Mean Squared Error (MSE): 0.44
📐 Root Mean Squared Error (RMSE): 0.67
📈 R² Score: 0.06
```

- ✓ Les résultats indiquent une amélioration notable par rapport aux essais précédents, bien que des marges de progression subsistent.
- ✓ MAE = 0.44 : L'erreur absolue moyenne est relativement faible, ce qui reflète une meilleure précision globale dans les prédictions du modèle.
- ✓ MSE = 0.44 : L'écart quadratique moyen est modéré, confirmant que les erreurs importantes sont bien maîtrisées.

- ✓ RMSE = 0.67 : L'erreur moyenne quadratique est la plus basse parmi les tests effectués, ce qui traduit une meilleure robustesse du modèle face aux écarts extrêmes.
- ✓ R^2 = 0.06 : Bien que la valeur reste faible, elle est positive, ce qui signifie que le modèle explique une petite part de la variance des données. Cela montre une tendance à l'amélioration, même si la capacité explicative reste limitée.

Les résultats montrent que le modèle KNN présente des performances modérées et légèrement inférieures à celles du Random Forest :

- MAE = 0.44 : L'erreur moyenne absolue montre que les prédictions du modèle s'écartent en moyenne de 0.44 g/L des valeurs réelles de glycémie. Cela indique une précision acceptable dans un contexte médical, où des marges d'erreur réduites sont importantes.
- MSE = 0.44 : L'erreur quadratique moyenne reflète une dispersion relativement faible des erreurs de prédiction. Cela signifie que le modèle parvient à limiter les erreurs importantes et fournit des résultats globalement cohérents.
- RMSE = 0.67 : La racine de l'erreur quadratique moyenne traduit une erreur moyenne modérée. Cette métrique est utile pour interpréter les écarts de manière intuitive, dans les mêmes unités que la glycémie.
- R^2 = 0.06 : Le coefficient de détermination indique que le modèle explique environ 6 % de la variance observée dans les valeurs de glycémie. Bien que cette proportion reste faible, elle montre que certaines relations entre les caractéristiques d'entrée et la glycémie sont captées

III.4.4.2 Random Forest

Le modèle RandomForestRegressor a été utilisé avec 100 arbres (`n_estimators=100`), pour construire un modèle robuste de prédiction de la glycémie à partir de signaux ECG. Les caractéristiques extraites (amplitudes et indices des ondes P, Q, R, S, T) ont été transformées en moyennes et utilisées comme entrées.

```
# Créer le modèle Random Forest
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
```

n_estimators=100 **:** Le modèle construit 100 arbres de décision. La prédiction finale est la moyenne des prédictions de ces arbres. Plus il y a d'arbres, plus le modèle est robuste (mais plus lent).

random_state=42 : Permet la reproductibilité des résultats (mêmes arbres construits à chaque exécution).

Les résultats de la prédiction glycémique du modèle random Forest comparés aux valeurs réelles sont donnés à la figure III.13.

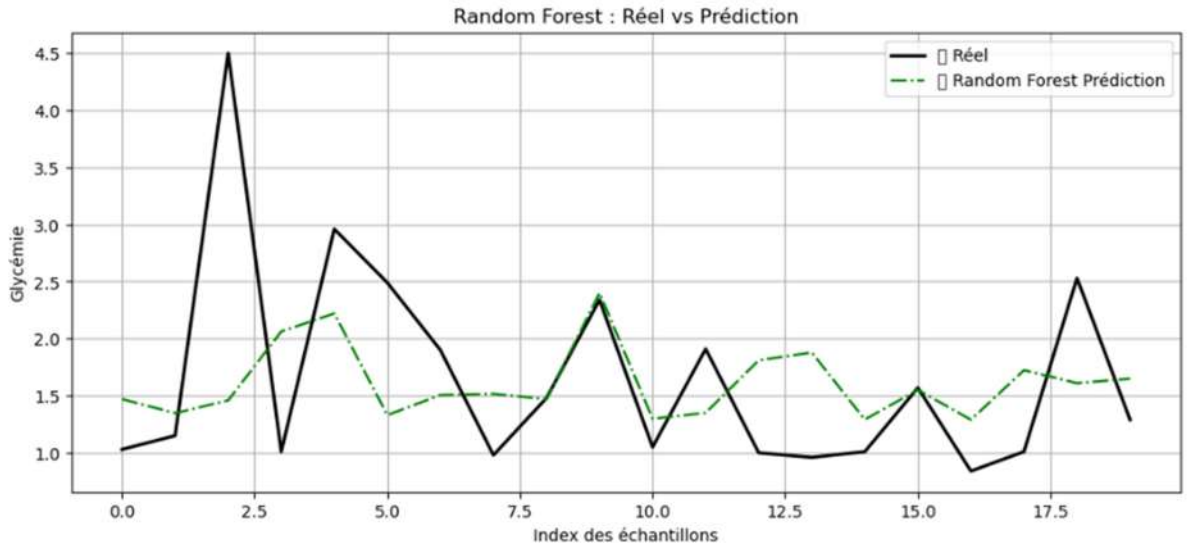

Fig. III.13 : Comparaison des Prédictions de modèle Random Forest

L'évaluation du modèle Random Forest basé sur les métriques citées ci-dessus a donné les résultats suivant :

```
# Affichage des résultats
print(f"📉 Mean Absolute Error (MAE): {mae:.2f}")
print(f"📏 Mean Squared Error (MSE): {mse:.2f}")
print(f"📐 Root Mean Squared Error (RMSE): {rmse:.2f}")
print(f"📊 R² Score: {r2:.2f}")
```

```
📉 Mean Absolute Error (MAE): 0.54
📏 Mean Squared Error (MSE): 0.58
📐 Root Mean Squared Error (RMSE): 0.76
📊 R² Score: -0.23
```

- Parmi les trois modèles testés, ce modèle affiche des performances légèrement meilleures, bien que celles-ci demeurent relativement faibles en termes de généralisation.
- MAE = 0.54 : L'erreur absolue moyenne est modérée, indiquant que, en moyenne, les prédictions du modèle s'écartent de 0.54 unités de la valeur réelle de la glycémie.
- MSE = 0.58 : L'erreur quadratique moyenne est encore significative, ce qui reflète la présence de quelques écarts importants entre les prédictions et les valeurs réelles.
- RMSE = 0.76 : L'erreur moyenne reste élevée en valeur absolue, ce qui suggère que le modèle n'est pas encore parfaitement calibré pour faire des prédictions précises.
- R^2 = -0.23 : Le coefficient de détermination négatif montre que le modèle explique moins de variance que la moyenne des données, ce qui indique un mauvais ajustement global. Cela pourrait être dû à un manque de complexité du modèle, à une variabilité élevée dans les données ou à un bruit important dans les mesures.

Les performances obtenues montrent que, parmi les trois modèles testés, Random Forest fournit les meilleurs résultats relatifs, bien que la qualité globale du modèle reste insatisfaisante :

- MAE = 0.54

  signifie que, en moyenne, les prédictions du modèle s'écartent de 0.54 unités par rapport aux valeurs réelles de glycémie.L'erreur est modérée, mais légèrement plus élevée qu'un bon modèle attendu pour ce type de tâche.

- MSE = 0.58

  L'erreur quadratique moyenne indique que le modèle fait parfois des erreurs plus importantes, car les écarts sont pénalisés de manière quadratique.La valeur reste raisonnable, mais elle confirme que le modèle ne capte pas bien toute la structure des données.

- RMSE = 0.76

  C'est une forme plus interprétable de l'erreur moyenne (comparable aux unités de glycémie). Une valeur inférieure à 1 est correcte, mais elle reste élevée si on vise une précision clinique.

- R^2 = -0.23

  Un R^2 négatif signifie que le modèle fait pire qu'une moyenne constante (autrement dit, prédire la valeur moyenne donnerait de meilleurs résultats).Cela indique que le modèle ne parvient pas à généraliser et n'explique pas la variance des données.

III.4.4.3 Arbre de décision DT (Decision Tree)

Le modèle DecisionTreeRegressor a été utilisé avec ses paramètres par défaut, à l'exception de random_state=42 pour assurer la reproductibilité. Les données d'entrée sont constituées des moyennes des amplitudes et indices des ondes ECG.

```
# 🌳 Création et entraînement du modèle Decision Tree
dt_model = DecisionTreeRegressor(random_state=42)
dt_model.fit(X_train, y_train)
```

random_state=42 : Il permet de garantir la reproductibilité des résultats (les mêmes découpages et décisions sont utilisés à chaque exécution).

criterion='squared_error' : l'arbre cherche à minimiser l'erreur quadratique moyenne (MSE) pour choisir les divisions.

max_depth=None : l'arbre peut croître jusqu'à ce que toutes les feuilles soient pures ou contiennent moins de 2 échantillons.

min_samples_split=2, min_samples_leaf=1 : contrôle la taille minimale des sous-ensembles pour éviter le sur-apprentissage.

Les résultats de la prédiction glycémique du modèle Arbre de décision comparés aux valeurs réelles sont donnés à la figure III.14.

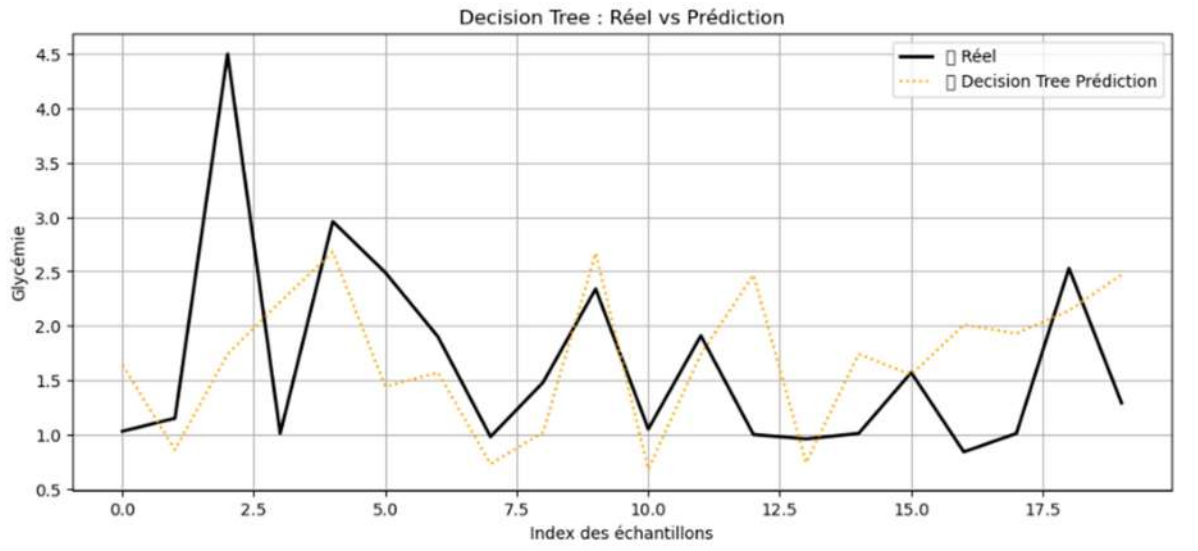

Fig. III.14 : Comparaison des Prédictions de modèle Decision Tree

L'évaluation du modèle Decision Tree basé sur les métriques citées ci-dessus a donné les résultats suivant :

```
# 📢 Affichage des résultats
print(f"🌳 Decision Tree - MAE  : {mae_dt:.2f}")
print(f"📉 Decision Tree - MSE  : {mse_dt:.2f}")
print(f"📐 Decision Tree - RMSE : {rmse_dt:.2f}")
print(f"📈 Decision Tree - R²   : {r2_dt:.2f}")
```

```
🌳 Decision Tree - MAE  : 0.77
📉 Decision Tree - MSE  : 1.15
📐 Decision Tree - RMSE : 1.07
📈 Decision Tree - R²   : -1.46
```

- MAE = 0.77 : Il s'agit de l'erreur absolue moyenne la plus élevée parmi les modèles testés, ce qui indique que les prédictions s'éloignent davantage des vraies valeurs en moyenne.
- MSE = 1.15 : L'écart quadratique moyen est aussi le plus élevé, signalant que le modèle fait parfois des erreurs très importantes.
- RMSE = 1.07 : L'erreur quadratique moyenne montre une précision faible, avec des prédictions souvent éloignées de la réalité (en unités de glycémie).
- R^2 = -1.46 : Ce score très négatif indique que le modèle explique très mal la variance des données. Il fait pire qu'un modèle naïf qui prédirait simplement la moyenne. Cela reflète un très mauvais ajustement, probablement lié à un sur-apprentissage (overfitting sur l'entraînement mais échec en test) ou à des données mal adaptées à cet algorithme.

Le modèle Decision Tree montre les performances les plus faibles parmi les trois modèles testés, ce qui soulève des limites importantes en termes de fiabilité prédictive :

MAE = 0.77

Cela signifie qu'en moyenne, le modèle se trompe de 0.77 unités de glycémie par prédiction. C'est l'erreur moyenne la plus élevée parmi les modèles testés, ce qui traduit une faible précision du modèle.

MSE = 1.15

Le MSE punit davantage les grandes erreurs. Une valeur de 1.15 indique que le modèle fait des erreurs importantes et fréquentes dans ses prédictions.

RMSE = 1.07

L'erreur quadratique moyenne (exprimée dans la même unité que la glycémie) est relativement élevée, ce qui confirme que le modèle n'est pas fiable pour prédire précisément la glycémie.

R^2 = -1.46

Ce score indique que le modèle est pire qu'un simple modèle naïf qui prédirait la moyenne de la glycémie pour tous les patients.

Un R^2 négatif aussi bas signifie :

Le modèle ne capture pas du tout la relation entre les variables d'entrée (features ECG) et la glycémie.

Il pourrait y avoir un sur-apprentissage, une variabilité excessive dans les données ou un manque de qualité/information dans les variables utilisées.

La figure III.15 donne une étude comparative des trois modèles de prédiction

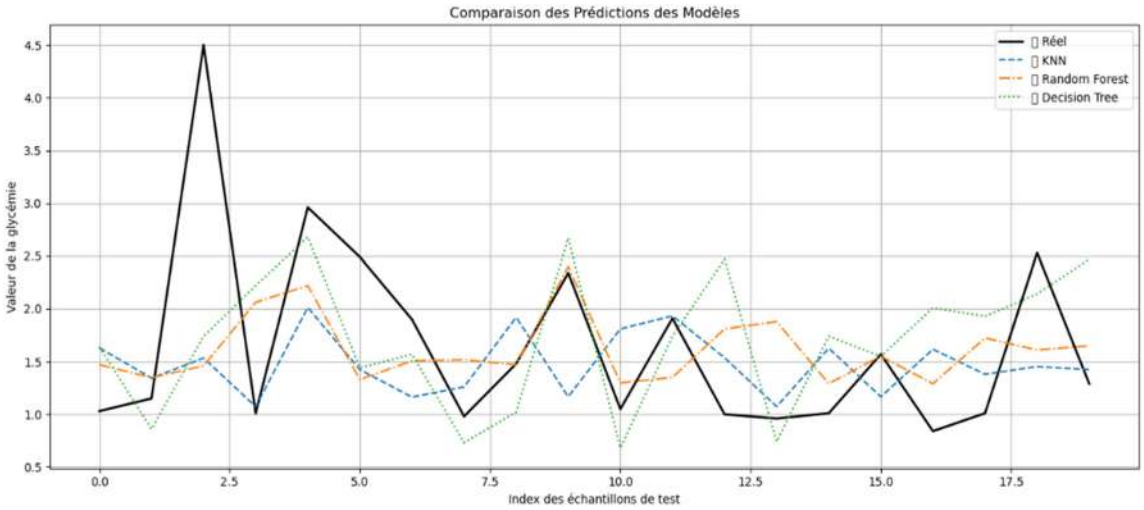

Fig. III.15 : comparaison des prédictions des modèles

Ces résultats suggèrent que le signal ECG seul (ou les caractéristiques actuellement extraites) ne contient pas suffisamment d'information pour prédire correctement la glycémie avec les modèles classiques utilisés.

La comparaison des résultats de l'évaluation des modèles est donnée ci-dessous :

📊 Comparaison des performances des modèles :

	Modèle	MAE	MSE	RMSE	R^2
0	KNN	0.438	0.444	0.666	0.055
1	Random Forest	0.537	0.576	0.759	-0.225
2	Decision Tree	0.775	1.154	1.074	-1.456

III.4.4.4 Comparaison et interprétation

Le modèle Random Forest affiche des performances globalement modérées. Avec un R^2 de -0.23, il montre une faible capacité explicative de la variance de la glycémie. Toutefois, son erreur moyenne absolue (MAE = 0.54) et sa racine de l'erreur quadratique moyenne (RMSE = 0.76) restent relativement faibles, accompagnées d'un MSE de 0.58. Ce modèle fournit donc des prédictions plus proches des valeurs réelles, malgré une faiblesse dans la généralisation.

Le modèle KNN montre des résultats légèrement meilleurs en termes d'erreur moyenne, avec une MAE de 0.44, un MSE de 0.44 et une RMSE de 0.67, ce qui indique une bonne précision globale. Son R^2 de 0.06 reste faible, mais il s'agit du seul modèle à avoir un score positif, indiquant une très légère capacité explicative de la variance.

Enfin, le modèle Decision Tree présente les performances les plus faibles. Avec une MAE de 0.77, un MSE de 1.15, un RMSE de 1.07, et un R^2 de -1.46, il démontre une mauvaise capacité de généralisation et une forte sensibilité au bruit, ce qui le rend peu adapté à ce type de prédiction.

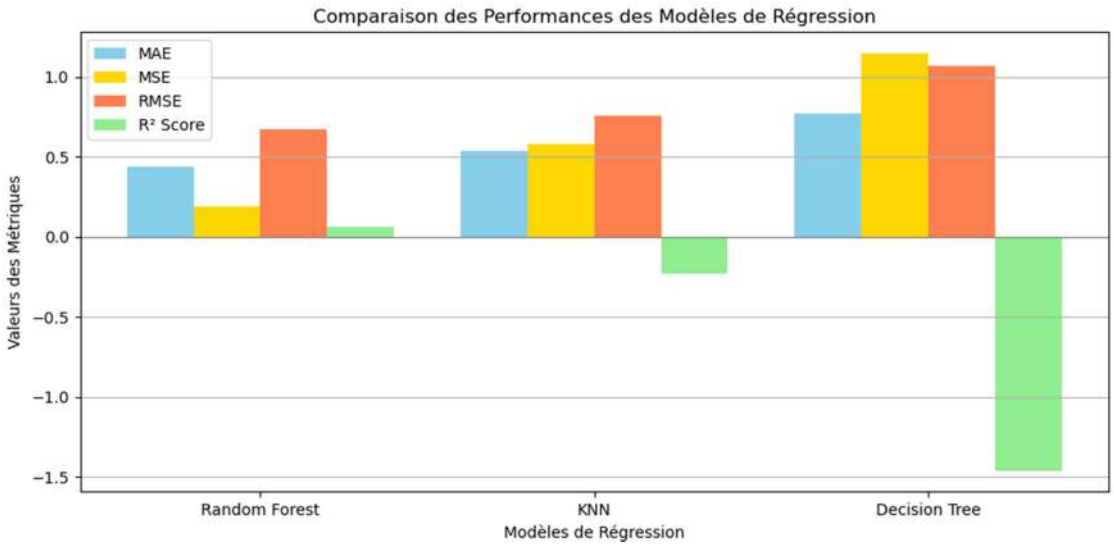

Fig. III.16: comparaison des performances des modèles de régression

III.5 Importances des caractéristiques

III.5.1 Importances des caractéristiques modèle Decision Tree

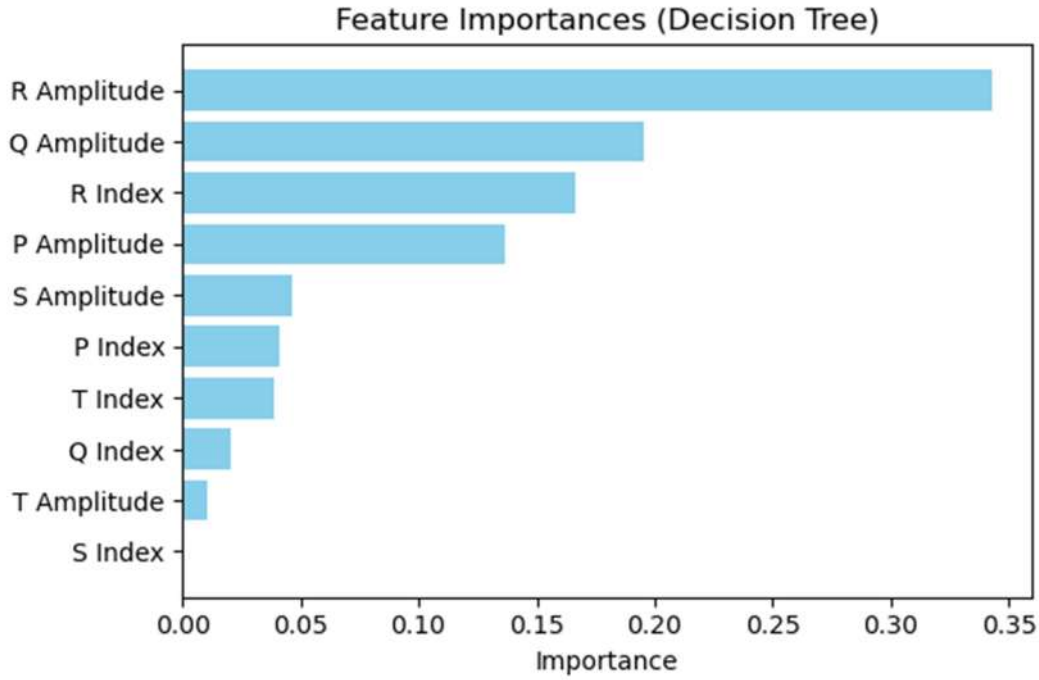

Fig. III. 17 : Les caractéristiques les plus importances pour le modèle Decision Tree

Dans la figure III.17, on observe l'importance des caractéristiques ECG utilisées pour la prédiction de la glycémie via le modèle Decision Tree. Il apparaît que l'amplitude de l'onde R est la caractéristique la plus influente, suivie de l'amplitude de l'onde Q et de l'indice de l'onde R. Ces variables représentent des points clés du cycle cardiaque qui sont fortement liés à l'activité électrique du cœur, et semblent fournir des informations pertinentes pour estimer le taux de glucose sanguin.

Les caractéristiques comme l'amplitude de l'onde P, S et les indices T, P, Q ont une importance plus modérée. Cela suggère que certaines phases du cycle ECG (en particulier les ondes R et Q) sont davantage liées aux variations de glycémie que d'autres. Les ondes S et T, bien qu'elles fassent partie de l'analyse ECG complète, semblent avoir un impact plus faible dans ce modèle.

En résumé, le modèle Decision Tree extrait principalement l'information de l'amplitude et de la position des ondes QRS, qui sont au cœur du complexe ventriculaire de l'ECG, pour prédire la glycémie. Ces résultats sont cohérents avec la physiologie cardiaque et les effets systémiques de l'hyperglycémie sur l'activité électrique du cœur.

III.5.2 Importances des caractéristiques modèle KNN

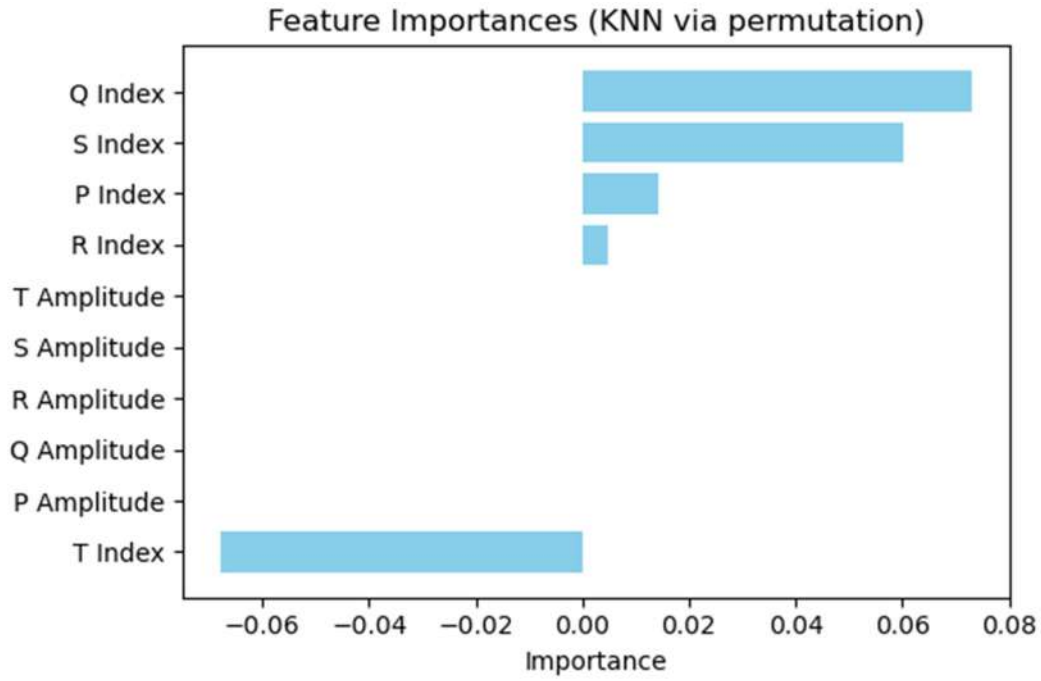

Fig. III. 18 : Les caractéristiques les plus importances pour KNN

Dans la figure III. 18, on illustre l'importance des différentes caractéristiques extraites du signal ECG pour la prédiction de la glycémie en utilisant la méthode de permutation avec le modèle KNN. Il ressort de cette analyse que les indices Q Index et S Index sont les plus déterminants dans la performance prédictive du modèle, suggérant que la position temporelle de ces ondes joue un rôle clé dans la relation entre l'activité électrique cardiaque et le taux de glucose.

En revanche, des caractéristiques telles que T Index présentent une importance négative, ce qui indique qu'elles pourraient introduire du bruit ou perturber légèrement la qualité de la prédiction. Les amplitudes des ondes ECG (P, Q, R, S, T) semblent avoir un impact négligeable sur le modèle KNN, leur importance étant proche de zéro.

Cela suggère que, dans le cas du modèle KNN appliqué aux signaux ECG, les indices temporels des ondes PQRST sont plus pertinents que leurs amplitudes pour estimer le taux de glycémie.

III.5.3 Importances des caractéristiques modèle Random Forest

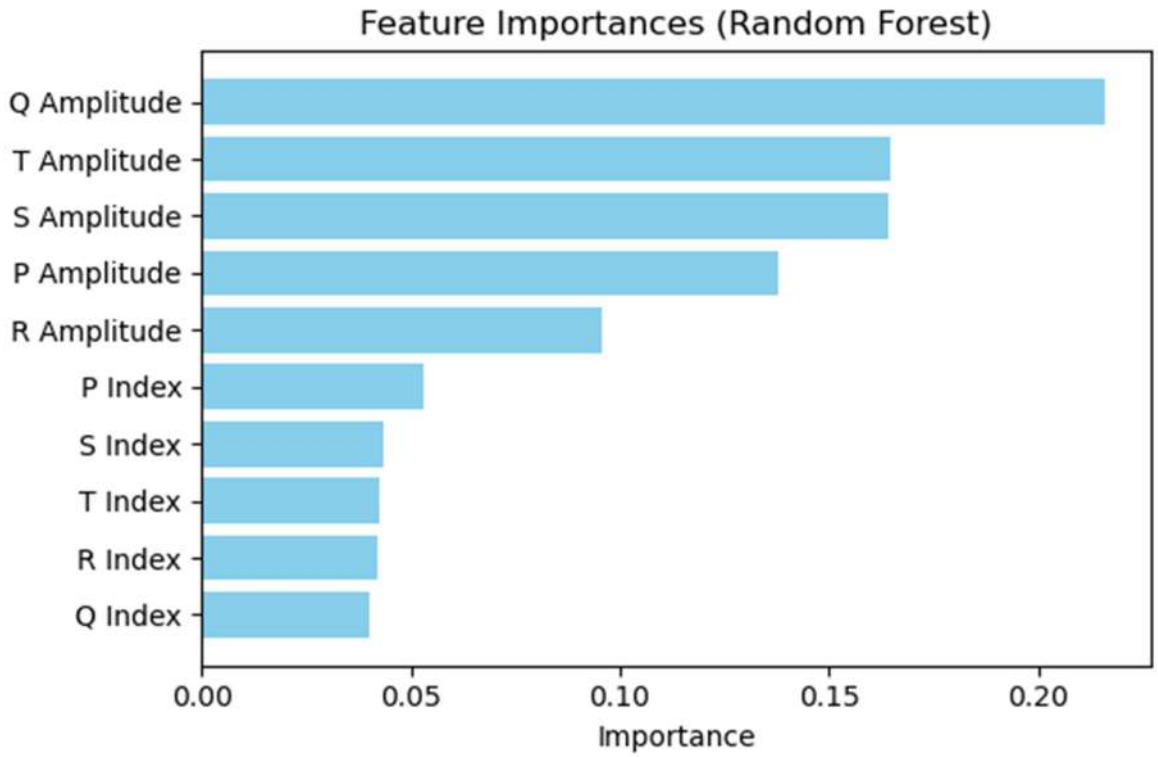

Fig. III. 19 : Les caractéristiques les plus importances pour Random Forest

Dans la figure III.19, on représente l'importance des caractéristiques extraites du signal ECG pour la prédiction de la glycémie à l'aide du modèle Random Forest. Il apparaît clairement que les amplitudes des ondes Q, T, S, P et R sont les plus contributives à la prédiction de la glycémie, avec une importance notable pour l'amplitude de l'onde Q. En revanche, les indices temporels des différentes ondes (P, Q, R, S, T) présentent une importance moindre dans cette tâche. Cela suggère que la morphologie des ondes ECG, notamment leur amplitude, joue un rôle plus significatif dans la prédiction du taux de glucose sanguin par le modèle Random Forest que leur position temporelle dans le cycle cardiaque

III.5.4 Comparaison de l'importance des caractéristiques ECG entre modèles

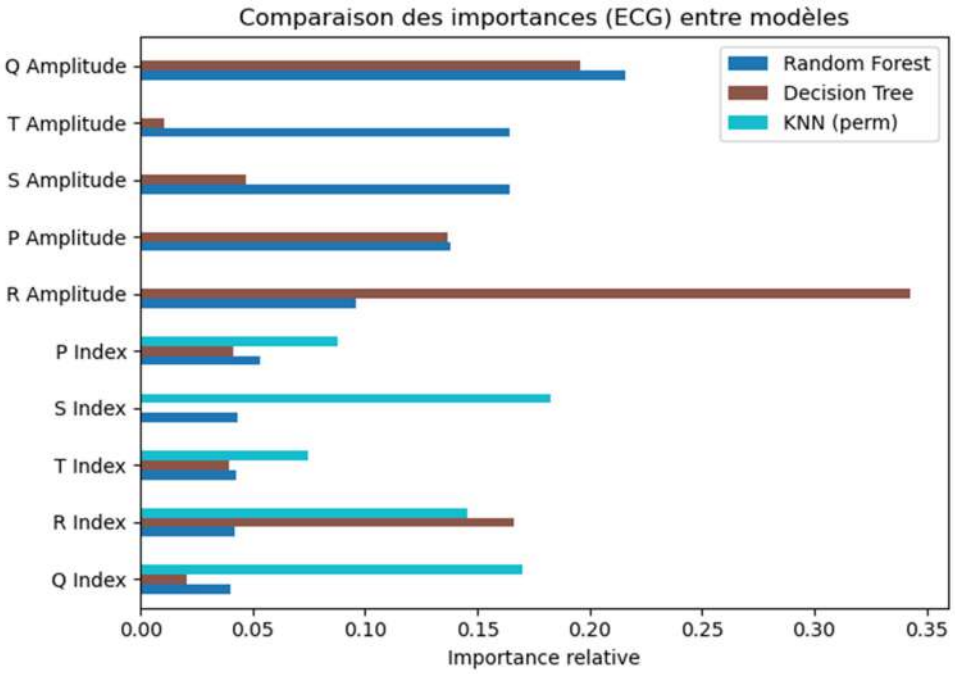

Fig. III. 20 : Comparaison de l'importance feactures ECG entre modèles

III.5.5 Importances des paramètres cliniques

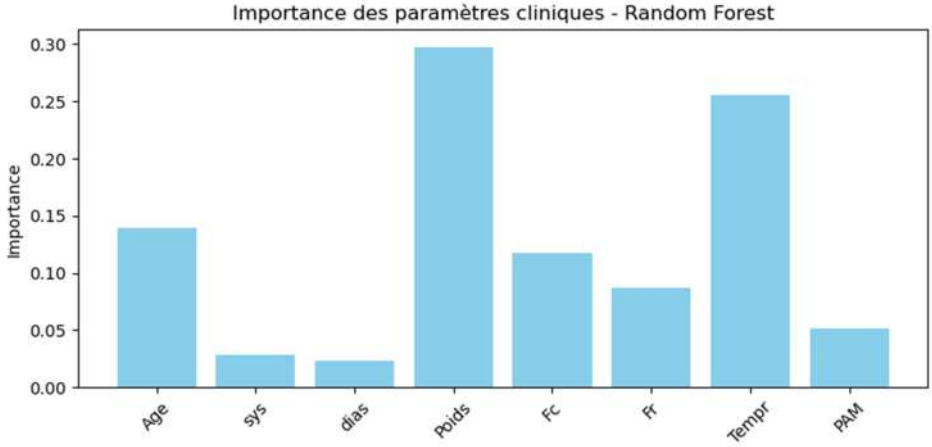

Fig. III. 21 : Importance des paramètres cliniques pour Random Forest

Dans la figure III.21, on représente l'importance des paramètres cliniques pour la prédiction de la glycémie à l'aide du modèle Random Forest. Il ressort de cette figure que le poids, la température corporelle, l'âge et la fréquence cardiaque (Fc) sont les paramètres les plus influents dans le processus de prédiction. Le poids apparaît comme le facteur prédominant, suivi de près par la température. À l'inverse, la pression artérielle systolique (sys), la pression artérielle diastolique (dias) et la pression artérielle moyenne (PAM) présentent une contribution relativement faible. Ces résultats indiquent que certains paramètres physiologiques, notamment

le poids et la température, ont un lien plus fort avec la variation du taux de glucose sanguin dans le contexte du modèle Random Forest.

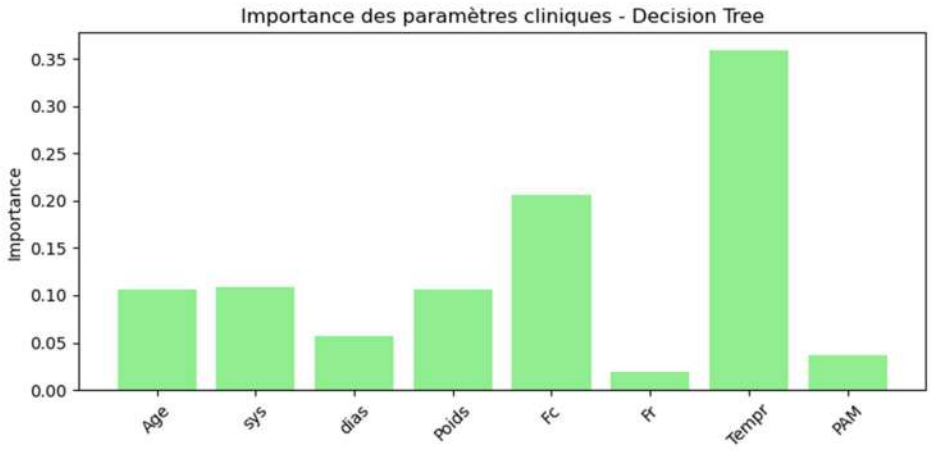

Fig. III. 22 : Importance des paramètres cliniques (Decision Tree)

Dans la figure III.22, on représente l'importance des caractéristiques cliniques pour la prédiction de la glycémie à partir de la méthode Decision Tree appliquée au signal ECG. Il apparaît de l'histogramme que la température corporelle (Tempr), la fréquence cardiaque (Fc), ainsi que les caractéristiques Age, poids et pression artérielle systolique (Sys) ont une importance relativement élevée dans la prédiction de la glycémie. En revanche, la fréquence respiratoire (Fr), la pression artérielle diastolique (dias) et la pression artérielle moyenne (PAM) présentent une contribution moins significative selon ce modèle.

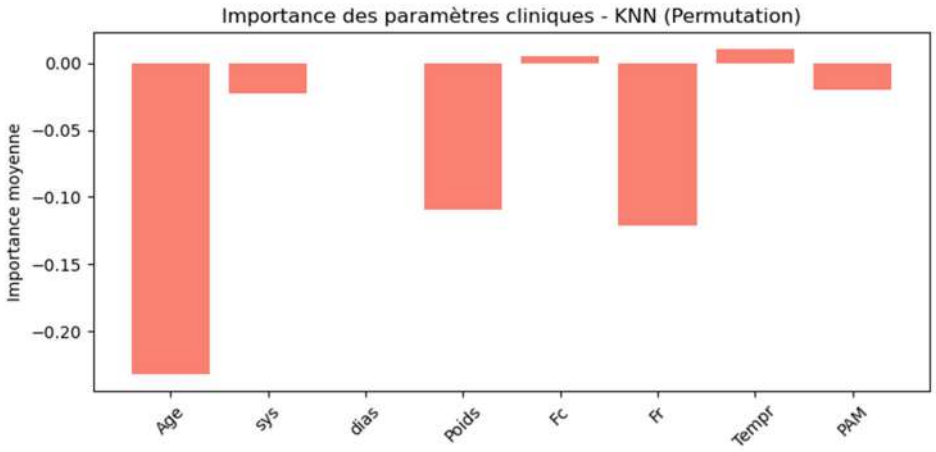

Fig. III. 23 : Importance des paramètres cliniques (KNN)

Dans la figure III.23, on présente l'importance des paramètres cliniques dans la prédiction de la glycémie à l'aide du modèle KNN, basée sur la méthode de permutation. Il

ressort de cette figure que la majorité des paramètres ont une influence faible, voire négative, sur la performance du modèle. L'âge, la fréquence respiratoire (Fr) et le poids présentent les valeurs négatives les plus marquées, suggérant qu'ils dégradent légèrement les performances du modèle lorsqu'ils sont permutés. À l'inverse, les autres paramètres, tels que la température, la fréquence cardiaque (Fc), la pression artérielle systolique (sys), diastolique (dias) et moyenne (PAM), ont une importance moyenne proche de zéro, indiquant qu'ils ont peu d'effet sur la prédiction du taux de glucose dans ce modèle. Ces résultats révèlent une faible sensibilité du modèle KNN aux variations des paramètres cliniques par rapport aux autres approches.

Le KNN ne donne pas d'importance de features par défaut, car ce n'est pas un modèle basé sur des poids mais sur la distance.

III.5.6 Comparaison de l'importance des paramètres cliniques pour les trois modèles

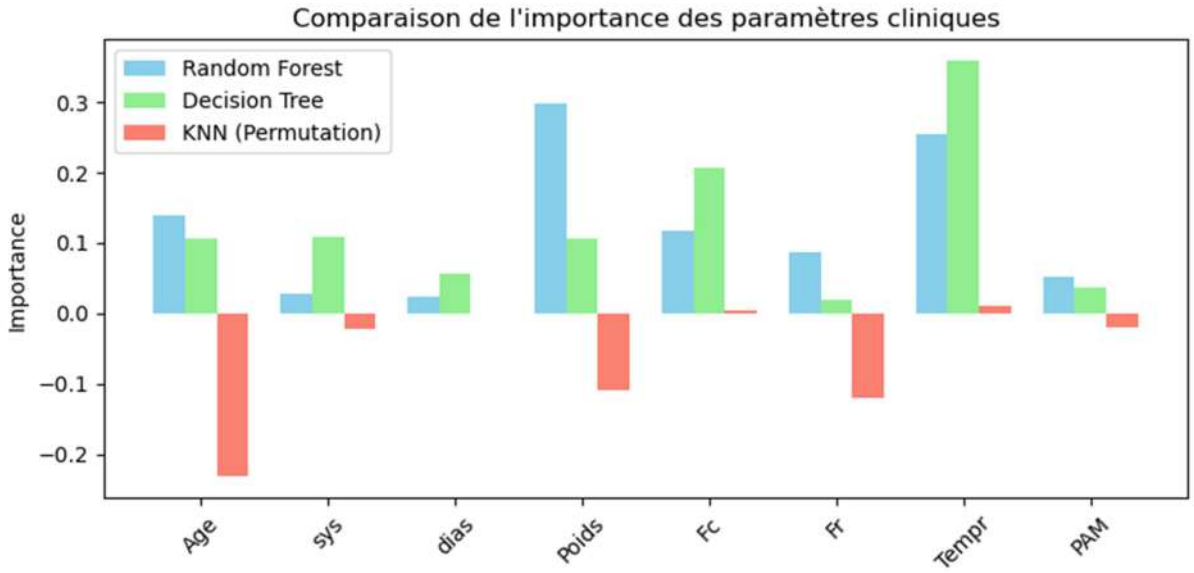

Fig. III. 24 : Comparaison de l'importance des paramètres cliniques pour les trois (3) modèles

III.5.7 Recommandations pour améliorer les performances

- ✓ Améliorer la qualité ou la diversité des caractéristiques extraites du signal ECG :
- ✓ Extraire plus de features temporelles, fréquentielles ou non linéaires.
- ✓ Utiliser des méthodes avancées comme la décomposition en ondelettes ou l'analyse PCA.
- ✓ Ajuster les hyperparamètres via une recherche systématique (GridSearchCV par exemple).
- ✓ Augmenter la taille ou la qualité de la base de données :
- ✓ Éviter les valeurs bruitées ou les enregistrements imprécis.

- ✓ Avoir plus de diversité dans les profils de patients.
- ✓ Tester d'autres modèles non linéaires plus puissants : SVR, XGBoost, ou réseaux de neurones.

III.6 Conclusion

Les résultats obtenus montrent que, dans l'état actuel, les caractéristiques extraites des signaux ECG ne permettent pas une prédiction fiable de la glycémie, quel que soit le modèle utilisé. Bien que Random Forest montre un léger avantage, les valeurs négatives de R^2 indiquent que les modèles testés ne parviennent pas à surpasser une simple moyenne comme prédiction.

Pour améliorer ces performances, des pistes possibles incluent :

- L'extraction de nouvelles caractéristiques plus représentatives ;
- L'augmentation du volume et de la qualité des données ;
- L'utilisation de modèles plus sophistiqués ou non linéaires comme XGBoost ou des réseaux de neurones.

Conclusion générale

Dans le contexte de ce projet, nous avons étudié une méthode non invasive pour prédire le niveau de glucose à partir des signaux ECG, en faisant appel à des techniques d'apprentissage automatique. Trois modèles de régression ont été examinés : le modèle de la Forêt Aléatoire RF (Random Forest), le modèle de k plus proche voisin KNN (K-Near Neighbors) et le modèle de l'Arbre de Décision DT (Decision Tree).

Pour l'entraînement de nos systèmes, une base de données composée de 100 patients a été mise en œuvre. Les signaux physiologiques ECG et les paramètres cliniques de cette base ont été collectés à l'aide du moniteur HealthyPi v3, auprès de patients suivis à la maison du diabète de la daïra de Saïda. Après le prétraitement des données (ECG et paramètres cliniques), l'ensemble des caractéristiques extraites, principalement, à partir du signal ECG (caractéristiques temporelles, caractéristiques statistiques, caractéristiques morphologiques) en association avec les caractéristiques cliniques ont été utilisés comme variables indépendantes pour l'apprentissage et le test de nos systèmes.

Les résultats obtenus ont montré des performances remarquables en général. Le modèle KNN a montré sa supériorité par rapport aux autres modèles étudiés (RF, DT) avec les métriques MAE évalué à 0.44, RMSE à 0.67 et un coefficient de détermination R^2 de 0.06, ce qui traduit une capacité modeste à expliquer la variance du taux de sucre dans le sang. Dans le contexte de cette recherche, ce modèle apparaît comme le plus prometteur, bien qu'il subsiste d'importantes possibilités d'amélioration.

Le modèle Random Forest a présenté des performances légèrement moins bonnes, donnant une MAE de 0.54, une RMSE de 0.76 et un R^2 de -0.23, ce qui indique une prédiction assez stable mais une aptitude limitée à la généralisation. En ce qui concerne le modèle Decision Tree, il a apparait moins performant que KNN et RF, avec MAE évaluée à 0.77, MSE à 1.15 et RMSE à 1.07 et un R^2 de -1.46 et, mettant en évidence sa vulnérabilité face aux données d'une base de données de taille limité.

Ces conclusions mettent en évidence que, malgré l'intérêt et le caractère novateur de la prédiction de la valeur de la glycémie à partir des signaux ECG, elle requiert encore des perfectionnements méthodologiques pour s'avérer fiable dans un contexte clinique. L'incorporation d'un plus grand volume de données, une sélection des caractéristiques plus minutieuse, et l'exploration de modèles plus robustes (réseaux de neurones, apprentissage

profond, ou modèles hybrides) pourraient constituer des perspectives pertinentes pour les travaux futurs.

Bien que les résultats actuels montrent une faisabilité partielle, ils ouvrent la voie à des recherches plus approfondies. La prédiction non invasive de la glycémie par ECG constitue une approche innovante, prometteuse à long terme, à condition d'y associer des techniques avancées et des bases de données plus étendues.

Pour améliorer la précision et la capacité de généralisation des modèles, plusieurs pistes sont à envisager :

- Augmentation de la taille et la diversité des données : Un plus grand nombre de patients, avec des profils physiologiques variés, permettrait de renforcer la robustesse du modèle.
- Sélection avancée des caractéristiques (feature selection) : Utiliser des méthodes statistiques ou algorithmiques (PCA, Lasso, Recursive Feature Elimination) pour ne retenir que les variables les plus pertinentes, et éviter le bruit.
- Fusion de signaux multimodaux : Combiner ECG avec d'autres signaux comme le PPG, la température corporelle ou la fréquence respiratoire peut enrichir la base d'information et améliorer la prédiction.
- Utilisation de modèles plus avancés : Explorer les réseaux de neurones profonds, les modèles LSTM pour les données temporelles, ou les approches hybrides (empiriques + ML) pourrait significativement améliorer les performances.
- Optimisation des hyperparamètres : Une recherche plus fine des paramètres des modèles (par GridSearchCV ou RandomizedSearchCV) pourrait conduire à un meilleur ajustement.

Références

[1] M.L. TALBI, ‹‹ Analyse et traitement de signal électrocardiographique (ECG) ››, Thèse de Doctorat, Université Mentouri de Constantine, 2011.

[2] de Jager, J., L. Wallis, and D. Maritz, ECG interpretation skills of South African Emergency Medicine residents. International journal of emergency medicine, 2010. 3(4) : p. 309-314.

[3] MESSIOUD M, Classification des signaux ECG en utilisant les réseaux de neurones .2019. Page (26,27)

[4] Hamadou El Mehdi, Bendehnoun.A, ‹‹ Développement et réalisation pratique d'un électrocardiographe ECG ››, Thèse de master, Centre Universitaire d'Ain Témouchent

[5] KRICHANE.N et TAZBOUDJT.S ‘’Classification des signaux ECG par les réseaux de neurone probabiliste 2015.

[6] K.Si yahia et M.kaddour, « conception et réalisation d'un dispositif d'exploration fonctionnelle cardiovasculaire », mémoire Master, Université Abou Bekr Belkaid Tlemcen, 2015/2016.

[7] M. BELMEKHFI, « Mise au point d'un système de mesure de paramètres physiologiques à base d'un Smartphone Androïde », thèse de Magister, Université Mouloud Mammri Tizi Ouzou.

[8] EZRATTY, Olivier. Les usages de l'intelligence artificielle. Olivier Ezratty, 2018

[9] Cleuet, valentin, « l'intelligence artificielle et ses axes de rechereches» 2 jullet 2019 [en

Ligne].available : https ://www.cloud-temple.com/intelligence-artificielle-axes-recherches. [accès le 20 mars 2020].

[10] CNIL. Apprentissage automatique. Consulté le 15 juin 2025, sur CNIL.

[11] DeepAI,» [En ligne]. Available : https ://deepai.org/machine-learning glossary-terms/machine-learning. [Access le 17 Mars 2020].

[12] EZRATTY, Olivier. Les usages de l'intelligence artificielle. Olivier Ezratty, 2018.

[13] SMOLYAKOV, Vadim. Ensemble learning to improve machine learning results. Stats et Bots,2017.

[14] Pegliasco,Gael ,« IInitiation au Machine Learning avec Python - La th´eorie,» 31 janvier2019 [en ligne].available : https ://makina-corpus.com/blog/metier/2017/initiation-au-machine-learning-avec-python. [accès le 20 mars 2020].

[15] Harfi, R. (2020) Amélioration des forêts aléatoires pour la classification des données médicales. : Mémoire Présenté En Vue De L'obtention Du Diplôme De Master Intelligence Artificielle Et Traitement De l'Information. Algérie : université Badji Mokhtar -Annaba. p 118.

[16] Khushaktov, M.F. (2023) ‘Introduction Random Forest Classification By Example, Medium. https://medium.com/@mrmaster907/introduction-random-forest-classification-byexample-6983d95c7b91.

[17] Keldenich, T. (2022) ‘Arbre de Décision Comment l’Utiliser - Meilleur Tutoriel’. [En ligne] (Page consultée le 22/05/2024) https://inside-machinelearning.com/arbre-decision/.

[18] CAPPONI, Cecile. Arbres de d´ecision. M2 mass, Université Aix-Marseille. Cité, p. 34.

[19] anjay,M,«to words data science»,26 Octobre 2018.[En ligne]. Available : https://towardsdatascience.com/.[Accès le 29 Juillet 2020].

[20] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://www.jstor.org/stable/2346178

[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems (NeurIPS 2017), 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[22] Chen, T., & Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System.Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 785–794. https://doi.org/10.1145/2939672.2939785

[23] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857

[24] Z. Tafa, N. Pervetica and B. Karahoda, "An intelligent system for diabetes prediction", 2015 4th Mediterranean Conference on Embedded Computing (MECO), 2015, pp. 378-382, doi: 10.1109/MECO.2015.7181948.

[25] Agarwal R., Gao G., DesRoches C., Jha A.K. Research commentary—The digital transformation of healthcare: Current status and the road ahead. Inf. Syst. Res. 2010;21:796–809. https://doi.org/10.1287/isre.1100.0327

[26] Pollard, T., Johnson, A., Raffa, J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 5, 180178 (2018). https://doi.org/10.1038/sdata.2018.178

[27] ProtoCentral. HealthyPi v3 – Open-Source Vital Sign Monitor. [En ligne]. Disponible sur : https://github.com/Protocentral/HealthyPi-v3

[28] Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, 36(8), 1627–1639. https://doi.org/10.1021/ac60214a047

[29] Smith, Steven W. (1997). The Scientist and Engineer's Guide to Digital Signal Processing. California Technical Publishing.

[30] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

[31] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[32] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.