Projet Fiches Explicatives – Santé numérique et IA

Sujets exposés sous forme de question : par ordre de présentation

1. Qu'est-ce que l'IA ? Comment cela marche ?
2. Quels sont les langages de l'IA ?
3. Qu'est-ce que le machine learning et le deep learning ?
4. Qu'est-ce que de la donnée personelle de santé ?
5. Comment protéger les données du patient ?
6. A qui appartiennent les données ?
7. Comment constituer un data set ?

Rédaction pour le GT ESIA : Collette Mavier, Fabien Taghon, Jeremy Corriger, Pascal Demoly, Julien Goret, Amir Guemari, Pol André Apoil

1. Qu'est-ce que l'IA ? Comment cela marche ?

Définition de l'intelligence artificielle

L'intelligence artificielle (IA) fait référence à la capacité d'une machine, généralement un ordinateur, à reproduire les fonctions cognitives humaines. Grâce au Machine Learning (ML), une sous-catégorie importante de l'IA, il est possible d'effectuer des tâches complexes plus rapidement et avec une plus grande précision que l'intelligence humaine, voire de la dépasser : résolution de problèmes, apprentissage, prise de décisions, reconnaissance de motifs… Le ML s'appuie sur des algorithmes alimentés par des bases de données modélisées pour en extraire l'information recherchée.

Dans le domaine de la santé, l'IA et le ML ont trouvé de nombreuses applications prometteuses pour améliorer les soins médicaux, la recherche et la gestion des ressources de santé.

En voici quelques exemples :

- Diagnostic médical : détection d'anomalies dans des images médicales telles que les radiographies, les IRM et les scanners
- Prédictions de maladies : identification de facteurs de risque de développer certaines maladies.
- Traitements : personnalisation selon les caractéristiques individuelles de chaque patient, recherche de nouveaux médicaments.
- Recherche médicale : analyse d'ensembles de données massives provenant de diverses sources, contribuant ainsi à la découverte de nouvelles connaissances médicales

Les sociétés savantes telles que la société européenne d'allergologie et d'immunologie clinique (EAACI) précisent que l'IA est un domaine émergent qui exploite de puissants algorithmes informatiques pour

effectuer des tâches difficiles qui dépassent l'intelligence humaine. Les modèles ML « entraînés » sur les données sont alors capables d'effectuer une série de tâches potentiellement très utiles, telles que l'estimation du risque d'un résultat d'intérêt pour un individu, la recherche de regroupements naturels au sein des données, ou l'extraction automatique de sens à partir du contenu d'images, de vidéos ou de textes. Ces puissantes approches analytiques modifient la manière dont les systèmes complexes et riches en données sont évalués, et le domaine médical ne fait pas exception.

2. Quels sont les langages de l'IA ?

Les langages de programmation jouent un rôle essentiel dans le domaine de l'intelligence artificielle. Voici quelques-uns des langages les plus couramment utilisés en IA :

Python est un langage de programmation populaire et polyvalent qui est largement utilisé en IA. Sa syntaxe claire et concise, ses bibliothèques puissantes telles que scikit-learn, TensorFlow et PyTorch, ainsi que sa grande communauté de développeurs en font un choix populaire pour les projets d'IA en santé.

R est un langage de programmation statistique largement utilisé dans l'analyse de données et la modélisation statistique. Il possède de nombreuses bibliothèques spécialisées dans l'IA, comme Caret et MLR, ce qui en fait un choix courant pour les chercheurs et les professionnels de la santé qui travaillent avec des données médicales.

Julia est un langage de programmation relativement nouveau, mais il gagne rapidement en popularité dans le domaine de l'IA. Il est apprécié pour sa syntaxe simple et sa grande efficacité dans le calcul numérique, ce qui en fait un choix attrayant pour les projets d'IA nécessitant des calculs intensifs.

MATLAB est un environnement de développement intégré qui offre des fonctionnalités puissantes pour le traitement du signal et l'analyse des données. Il est couramment utilisé en IA dans le domaine de la santé pour des tâches telles que l'imagerie médicale.

C++ est un langage de programmation polyvalent et rapide, souvent utilisé dans le développement de bibliothèques et de frameworks d'IA en santé. Il est apprécié pour sa capacité à gérer efficacement les opérations intensives en termes de calcul.

Ces langages de programmation ne sont qu'une sélection parmi de nombreux autres utilisés en IA. Le choix du langage dépend souvent des préférences individuelles, des exigences du projet et de la compatibilité avec les bibliothèques et les outils spécifiques utilisés dans le domaine médical.

3. Qu'est-ce que le machine learning et le deep learning ?

Le ***Machine Learning*** **(ou apprentissage automatique)** est un champ d'étude de l'IA qui vise à permettre à un ordinateur "d'apprendre" à partir de données, simulant ainsi la capacité humaine d'acquérir des connaissances par l'expérience. Cette approche peut être **supervisée**, où l'ordinateur est entraîné à partir d'un ensemble de données et des résultats associés

(données dites “étiquetées”), ou **non supervisée**, où il découvre des structures et des relations dans les données par lui-même.

A partir de données d’entraînement, l’objectif est ainsi de générer et rendre le plus performant possible un modèle mathématique pouvant être utilisé pour **effectuer des prédictions** sur d’autres ensembles de données.

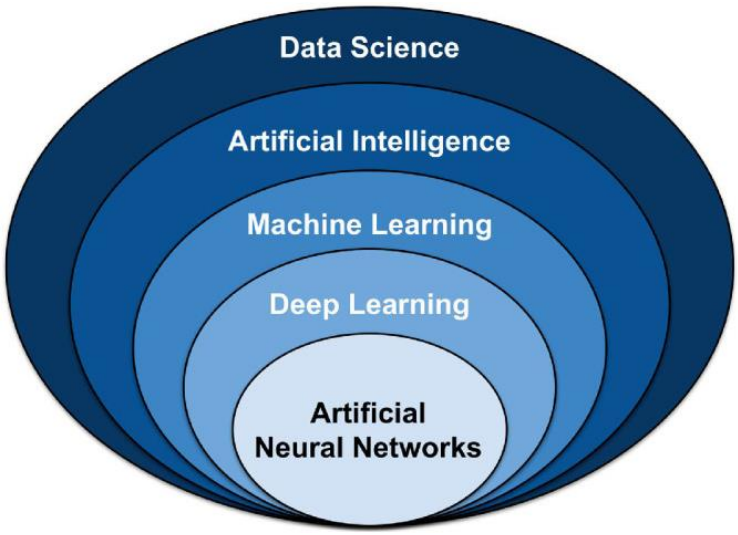

Figure tirée de (5).

Le ***Deep Learning*** **(ou apprentissage profond)** est une branche du *Machine Learning* qui imite le fonctionnement du cerveau humain grâce à des **réseaux de neurones artificiels** composés de plusieurs couches de traitement mathématique. Ces réseaux peuvent être entraînés en utilisant des **techniques de renforcement**, reposant sur des modèles d’essais et erreurs, plus proches de l’apprentissage humain.

Ces algorithmes possèdent souvent de **nombreux paramètres**, qui vont être **ajustés par la machine au fur et à mesure** de l’entraînement, afin de réduire ses erreurs et améliorer sa performance. Des **jeux de données très larges et de bonne qualité** sont essentiels pour entraîner efficacement de tels modèles.

4. https://www.cnil.fr/fr/definition/apprentissage-automatique
5. https://www.cnil.fr/fr/definition/apprentissage-profond-deep-learning
6. https://datascientest.com/machine-learning-tout-savoir
7. Y LeCun, Y Bengio & G Hinton, Nature 2015. doi:10.1038/nature14539
8. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Trans Vis Sci Tech. 2020;9(2):14. doi.org/10.1167/tvst.9.2.14

4. Qu’est-ce que de la donnée personnelle de santé ?

Le Règlement Général sur la Protection des Données (RGPD), entré en vigueur le 25 mai 2018, définit les données de santé comme « **toutes données à caractère personnel relatives à la santé physique ou mentale d'une personne ou à la prestation de services de santé à cette personne** » (1). Cette définition est finalement assez large, et est « à apprécier, au cas par cas, compte tenu de la nature des données recueillies » (2). Elles concernent un individu précis identifié de manière unique à des fins de santé.

Des informations qui peuvent avoir des incidences particulières critiques sur la vie privée d'une personne si elles étaient révélées sont dites sensibles. La Commission Nationale de l'Informatique et des Libertés (CNIL) définit les données personnelles sensibles comme « des informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique » (2).

Le RGPD (1) d'abord interdit de recueillir ou d'utiliser ces données (dans son article 9), ensuite l'autorise dans de nombreuses situations dérogatoires (1, 2) :

- la personne concernée a donné **son consentement explicite** au traitement de ces données,
- le traitement est **nécessaire aux fins de l'exécution des obligations** et des droits propres au responsable du traitement ou à la personne concernée en matière de droit du travail, de la sécurité sociale et de la protection sociale,
- le traitement est **nécessaire à la sauvegarde des intérêts vitaux de la personne concernée**,
- le traitement est effectué, dans le cadre de leurs activités légitimes et moyennant les garanties appropriées, par une fondation, une association ou tout autre organisme à but non lucratif et poursuivant une finalité politique, philosophique, religieuse ou syndicale, à condition que les données à caractère personnel ne soient pas communiquées en dehors de cet organisme sans le consentement des personnes concernées,
- le traitement porte sur **des données à caractère personnel qui sont manifestement rendues publiques** par la personne concernée,
- le traitement est nécessaire à la constatation, à l'exercice ou à la défense d'un droit en justice,
- le traitement est nécessaire pour des motifs d'intérêt public important, sur la base du droit de l'Union ou du droit d'un 'État membre qui doit être proportionné à l'objectif poursuivi,
- le traitement est nécessaire aux fins de **la médecine préventive ou du travail**,
- le traitement est nécessaire pour **des motifs d'intérêt public** dans le domaine de la santé publique,

- le traitement est nécessaire **à des fins archivistiques dans l'intérêt public**, à des fins de recherche scientifique ou historique ou à des fins statistiques.

[1] https://donnees-rgpd.fr/definitions/ (consulté le 21.06.2023).

[2] https://www.cnil.fr/fr/definition/donnee-sensible (consulté le 21.06.2023).

**Cet article est largement inspiré (avec permission) du blog https://www.advicemedica.com/blog/donnees-de-patients (consulté le 21.06.2023).*

5. Comment protéger les données des patients ?

Quand il s'agit de protection des données des patients, la première solution envisageable est l'anonymisation de toute donnée susceptible d'identifier le patient. Une liste non exhaustive d'exemples tel que la date de naissance ou la région de résidence est alors essentiel à cet effet. Et en effet, bien qu'il s'agisse d'une étape importante, elle ne vient que compléter un ensemble de mesures visant à protéger les données, car avant l'anonymisation, celles-ci doivent être collectées. Une première étape est de limiter le nombre de personnes ayant accès aux données non anonymisées, notamment par la désignation d'une personne compétente qui sera seule au contact de ces données, c'est le cas des professionnels de santé. Ceux-ci s'engage à respecter des règles informatiques strictes (serveur sécurisé, mot de passe, pas de transfert, cryptage des données …)[1,2]. En effet, le nombre croissant de cyberattaques pouvant être à l'origine de fuite de données, l'organisme en charge de la base de données doit être en mesure d'assurer sa protection et son respect[3], notamment au travers de certification telle que celle délivrée par la CNIL (Commission Nationale de l'Informatique et des Libertés) qui garantit le respect du RGPD (Règlement Général sur la Protection des Données)[4].

1. Zarocostas, J. Health under cyberattack. *Lancet* **398**, 829–830 (2021).

2. Pinkham, D. W., Sala, I. M., Soisson, E. T., Wang, B. & Deeley, M. A. Are you ready for a cyberattack? *Journal of Applied Clinical Medical Physics* **22**, 4–7 (2021).

3. Article L1111-8 - Code de la santé publique - Légifrance. ttps://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000033862549.

4. Ce qu'il faut savoir sur la certification. https://www.cnil.fr/fr/ce-quil-faut-savoir-sur-la-certification.

6. A qui appartiennent les données ?

Le RGPD ne reconnaît pas un droit de propriété sur les données à caractère personnel. Les données ne peuvent donc faire l'objet d'appropriation. Toutefois, l'absence d'appropriation n'exclut pas toute protection. En effet, la personne dispose d'un droit de décider et de contrôler les utilisations faites des données à caractère personnel la concernant (cf. question 4). Cette logique porte le nom d'autodétermination informationnelle par opposition à un droit de propriété. Les patients doivent être informés des traitements de leurs données de santé et peuvent exercer des droits d'accès, de rectification, d'effacement, de limitation des traitements et de transfert (articles 15 à 21 RGPD). Le patient peut également décider qui peut collecter ses données de santé. Les données de santé identifiantes ne peuvent en aucun cas être vendues, que ce soit par le patient lui-même, ou par un tiers, avec ou sans l'accord du patient (article 1111-8 chap. VII du CSP).

Le droit des personnes concernées ne porte que sur les données à caractère personnel à l'exception des données anonymisées (cf. question 5) (avis du G29 du 10 avril 2014). Le RGPD n'est pas applicable aux données anonymisées traitées. Le procédé d'anonymisation permet l'utilisation sans frein des données dans le cadre notamment de la recherche scientifique ou d'outils d'intelligence artificielle. Le traitement des données mène à la création d'une base de données qui permet de retrouver des données structurées ou des données brutes qui seront utilisées pour des statistiques et des analyses et au final à un entrepôt de données pour stocker définitivement les informations. La titularité du droit des bases de données va au producteur au sens de personne qui donne les moyens pour les produire (moyens financiers, techniques, humains, support, sécurisation, etc.). Ainsi, le titulaire des droits (et non pas le propriétaire) ou le producteur au sens juridique est l'établissement, peu importe qui concourt à la production concrète des données. C'est le cas par exemple d'un CHU qui sera le titulaire d'une base de données anonymisées contenant des données des patients de l'établissement collectées par un agent : le droit d'auteur de ce dernier est automatiquement cédé à sa tutelle.

- Avis 06/2014 sur la notion d'intérêt légitime poursuivi par le responsable du traitement des données au sens de l'article 7 de la directive 95/46/CE

7. Comment constituer un data set ?

<u>Il faut vérifier la qualité des données recueillies en répondant aux questions suivantes :</u>

1- Part de l'erreur humaine lors de la saisie des données ?

Les données de santé complexes, comprenant à la fois des données cliniques, paracliniques et biologiques, seront le plus souvent collectées par des individus, lesquels peuvent ou non être ceux qui ont généré ces données. Il est utile d'analyser en détail un sous-ensemble du dataset afin d'estimer le taux d'erreur de saisie.

2- Quelle est la fréquence des données non disponibles ?

3- Les données recueillies sont-elles celles requises pour répondre aux questions posées ?

4- Le *dataset* est-il équilibré ?

Si l'on souhaite, par exemple, générer un modèle qui identifie les patients présentant une pathologie donnée au sein d'un groupe plus large de patients non atteints, le dataset doit comprendre une proportion suffisante de patients atteints.

Formatage des données. Le format des données doit être constant, en particulier si celles-ci proviennent de plusieurs sources, ont été générées sur plusieurs sites, ou bien que plusieurs personnes soient intervenues pour vérifier la consistance du *dataset*. Ces tâches gagnent à être automatisées *via* des outils logiciel. Des anomalies fréquentes consistent en l'utilisation de points ou de virgules comme séparateurs numériques, ou la présence de données qui sont en dehors de l'échelle permise (ex. seules des valeurs entières entre 1 et 10 sont autorisées et un enregistrement contient la valeur 10,46…).

Réduction de la complexité du dataset. Bien que terme *big data* suggère qu'il est préférable d'inclure le plus de données possibles, il est préférable d'éliminer les données redondantes, qui sont fortement corrélées entre elles.

« Nettoyage » de la base de données. Traitement des données manquantes : remplacées par une valeur nulle, par la valeur moyenne d'une variable numérique, ou par l'item le plus fréquemment présent d'une variable catégorielle.

Création de variables supplémentaires à partir de celles recueillies. Par exemple, ajouter une définition des individus asthmatiques *via* la combination de plusieurs autres variables : « crise d'asthme au cours de l'année qui précède OU diagnostic posé par un pneumologue au cours des 5 années précédentes OU VEMS < XXX »

Normalisation des données. Si des unités numériques très différentes sont utilisées, par exemple, si une variable A doit se situer dans une échelle de 0 à 10, alors qu'une autre variable

B peut prendre une valeur entre -10 000 et + 10 000), la variable B peut avoir un « poids » trop important par rapport à A. Une méthode simple consiste à convertir toutes les variables continues pour qu'elles soient exprimées dans une échelle identique (ex. 0 à 100).

Convertir une variable continue en variable discrète. Ceci peut être avantageux si l'on s'attend à un effet important à partir d'une certaine valeur de cette variable. Par exemple, la variable continue « âge » peut être convertie en une variable discontinue du type « groupe d'âge ».

Données manquantes. Il existe aujourd'hui de nombreux modèles mathématiques et statistiques qui permettent de compléter les données manquantes et d'évaluer leur complétion dans un data set.