UNIVERSITÉ TOULOUSE III – PAUL SABATIER
FACULTÉ DE SANTÉ – DÉPARTEMENT D'ODONTOLOGIE

ANNÉE 2024 2024 TOU3 3003

THÈSE

POUR LE DIPLÔME D'ÉTAT DE DOCTEUR EN CHIRURGIE DENTAIRE

Présentée et soutenue publiquement

par

Julien MAY

le 26 janvier 2024

PROPOSITION D'UNE MÉTHODOLOGIE D'ÉVALUATION DE L'INTÉRÊT DE LA XAI POUR LE CHIRURGIEN-DENTISTE

Directeur de thèse : Pr Paul MONSARRAT

JURY

Président : Pr Paul MONSARRAT

1er assesseur : Dr Géromine FOURNIER

2e assesseur : Dr Julien DELRIEU

3e assesseur : Dr Julien ALIGON

Faculté de santé
Département d'Odontologie

➔ DIRECTION

Doyen de la Faculté de Santé
M. Philippe POMAR

Vice Doyenne de la Faculté de Santé
Directrice du Département d'Odontologie
Mme Sara DALICIEUX-LAURENCIN

Directeurs Adjoints
Mme Sarah COUSTY
M. Florent DESTRUHAUT

Directrice Administrative
Mme Muriel VERDAGUER

Présidente du Comité Scientifique
Mme Cathy NABET

➔ HONORARIAT

Doyens honoraires
M. Jean LAGARRIGUE +
M. Jean-Philippe LODTER +
M. Gérard PALOUDIER
M. Michel SIXOU
M. Henri SOULET

Chargés de mission
M. Karim NASR (*Innovation Pédagogique*)
M. Olivier HAMEL *(Maillage Territorial)*
M. Franck DIEMER *(Formation Continue)*
M. Philippe KEMOUN *(Stratégie Immobilière)*
M. Paul MONSARRAT *(Intelligence Artificielle)*

➔ PERSONNEL ENSEIGNANT

Section CNU 56 : Développement, Croissance et Prévention

56.01 ODONTOLOGIE PEDIATRIQUE et ORTHOPEDIE DENTO-FACIALE (Mme Isabelle BAILLEUL-FORESTIER)

ODONTOLOGIE PEDIATRIQUE

Professeurs d'Université : Mme Isabelle BAILLEUL-FORESTIER, M. Frédéric VAYSSE
Maîtres de Conférences : Mme Marie- Cécile VALERA, M. Mathieu MARTY
Assistants : Mme Anne GICQUEL, M. Robin BENETAH
Adjoints d'Enseignement : M. Sébastien DOMINE, M. Mathieu TESTE, M. Daniel BANDON

ORTHOPEDIE DENTO-FACIALE

Maîtres de Conférences : M. Pascal BARON, M. Maxime ROTENBERG
Assistants : M. Vincent VIDAL-ROSSET, Mme Carole VARGAS JOULIA, Mme Chahrazed BELAILI
Adjoints d'Enseignement : Mme. Isabelle ARAGON

56.02 PRÉVENTION, ÉPIDÉMIOLOGIE, ÉCONOMIE DE LA SANTÉ, ODONTOLOGIE LÉGALE (Mme Catherine NABET)

Professeurs d'Université : M. Michel SIXOU, Mme Catherine NABET, M. Olivier HAMEL, M. Jean-Noël VERGNES
Maîtres de Conférences : Mme Géromine FOURNIER
Adjoints d'Enseignement : M. Alain DURAND, Mlle. Sacha BARON, M. Romain LAGARD, M. Jean-Philippe GATIGNOL
Mme Carole KANJ, Mme Mylène VINCENT-BERTHOUMIEUX, M. Christophe BEDOS

Section CNU 57 : Chirurgie Orale, Parodontologie, Biologie Orale

57.01 CHIRURGIE ORALE, PARODONTOLOGIE, BIOLOGIE ORALE (M. Philippe KEMOUN)

PARODONTOLOGIE

Professeurs d'Université : Mme Sara LAURENCIN- DALICIEUX,
Maîtres de Conférences Mme Alexia VINEL, Mme. Charlotte THOMAS
Assistants: M. Joffrey DURAN, M. Antoine AL HALABI
Adjoints d'Enseignement : M. Loïc CALVO, M. Antoine SANCIER, M. Ronan BARRE , Mme Myriam KADDECH,
M. Mathieu RIMBERT

CHIRURGIE ORALE

Professeur d'Université : Mme Sarah COUSTY
Maîtres de Conférences : M. Philippe CAMPAN, M. Bruno COURTOIS
Assistants : M. Antoine DUBUC
Adjoints d'Enseignement : M. Gabriel FAUXPOINT, M. Arnaud L'HOMME, Mme Marie-Pierre LABADIE, M. Jérôme SALEFRANQUE, M. Clément CAMBRONNE

BIOLOGIE ORALE

Professeurs d'Université : M. Philippe KEMOUN, M Vincent BLASCO-BAQUE
Maîtres de Conférences : M. Pierre-Pascal POULET, M. Matthieu MINTY
Assistants : Mme Chiara CECCHIN-ALBERTONI, M. Maxime LUIS, Mme Valentine BAYLET GALY-CASSIT, Mme Sylvie LE
Adjoints d'Enseignement : M. Mathieu FRANC, M. Hugo BARRAGUE, Mme Inessa TIMOFEEVA-JOSSINET

Section CNU 58 : Réhabilitation Orale

58.01 DENTISTERIE RESTAURATRICE, ENDODONTIE, PROTHESES, FONCTIONS-DYSFONCTIONS, IMAGERIE, BIOMATERIAUX (M. Franck DIEMER)

DENTISTERIE RESTAURATRICE, ENDODONTIE

Professeur d'Université : M. Franck DIEMER
Maîtres de Conférences : M. Philippe GUIGNES, Mme Marie GURGEL-GEORGELIN, Mme Delphine MARET-COMTESSE
Assistants : M. Nicolas ALAUX, M. Vincent SUAREZ, M. Lorris BOIVIN, M. Thibault DECAMPS, Mme Emma STURARO, Mme Anouk FESQUET
Adjoints d'Enseignement : M. Eric BALGUERIE, M. Jean- Philippe MALLET, M. Rami HAMDAN, M. Romain DUCASSE, Mme Lucie RAPP, Mme Marion CASTAING-FOURIER

PROTHÈSES

Professeurs d'Université : M. Philippe POMAR, M. Florent DESTRUHAUT,
Maîtres de Conférences : M. Antoine GALIBOURG, M. Julien DELRIEU
Assistants: Mme Coralie BATAILLE, Mme Mathilde HOURSET, Mme Constance CUNY, M. Anthony LEBON
Adjoints d'Enseignement : M. Christophe GHRENASSIA, Mme Marie-Hélène LACOSTE-FERRE, M. Olivier LE GAC, M. Luc RAYNALDY, M. Jean-Claude COMBADAZOU, M. Bertrand ARCAUTE, M. Fabien LEMAGNER, M. Eric SOLYOM, M. Michel KNAFO, M. Victor EMONET-DENAND, M. Thierry DENIS, M. Thibault YAGUE, M. Antonin HENNEQUIN, M. Bertrand CHAMPION, M. Steven CECCAREL

FONCTIONS-DYSFONCTIONS , IMAGERIE, BIOMATERIAUX

Professeur d'Université : Mr. Paul MONSARRAT
Maîtres de Conférences : Mme Sabine JONIOT, M. Karim NASR, M. Thibault CANCEILL,
Assistants : M. Olivier DENY, Mme Laura PASCALIN, Mme Alison PROSPER
Adjoints d'Enseignement : Mme Sylvie MAGNE, M. Thierry VERGÉ, M. Damien OSTROWSKI

Mise à jour pour le 11 janvier 2024

Au président du jury et directeur de thèse,

Monsieur le Professeur Paul MONSARRAT

- Professeur des Universités, Praticien Hospitalier d'Odontologie
- Docteur de l'Université Paul Sabatier - Spécialité Physiopathologie
- Diplôme Universitaire d'Imagerie 3D maxillo-faciale
- Diplôme universitaire de Recherche Clinique en Odontologie
- Habilitation à Diriger les Recherches
- Lauréat de la faculté de Médecine Rangueil et de Chirurgie Dentaire de l'Université Paul Sabatier

Un grand merci pour m'avoir fourni cette ouverture à l'usage de l'intelligence artificielle en santé et m'avoir fait découvrir l'explicabilité, qui est un sujet touffu et épineux pour celui qui, comme moi, le découvre de zéro, mais passionnant !

Également, merci pour l'encadrement clinique qui s'est toujours fait dans la bonne humeur et un engagement de qualité pour nous permettre d'apprendre au mieux notre métier, ainsi que pour les cours et travaux pratiques qui nous ont permis de nous engager dans cet externat préparés et sereins.

Au jury de thèse,

Madame le Docteur Géromine FOURNIER

- Maître de Conférences des Universités, Praticien Hospitalier d'Odontologie
- Docteur en Chirurgie Dentaire
- Docteur en anthropologie
- Lauréate de l'Université Paul Sabatier
- DU Odontologie Légale et Ethique
- DU Méthode et pratique en identification Oro Faciale
- Expert judiciaire en identification Odontologique près de la Cour d'Appel de Toulouse

Un grand merci pour nous avoir aussi bien encadré durant notre externat et supporté nos accessoires de bureau avec le sourire. Merci également d'être un modèle humain inspirant de par les soins apportés à toute personne, quelle que soit sa condition ; ainsi que de par les cours dispensés, engagés, notamment sur la sensibilisation, le dépistage et la prise en charge à notre échelle des violences domestiques. C'est un modèle que je m'efforcerai de suivre, avec l'espoir de contribuer également à rendre le monde un petit peu meilleur à travers ma pratique.

Au jury de thèse,

Monsieur le Docteur Julien DELRIEU

- Maître de Conférences des Universités, Praticien Hospitalier d'Odontologie
- Docteur en Chirurgie Dentaire
- CES de Prothèse Fixée
- Master 1 de Santé Publique
- Master 2 Anthropobiologie intégrative

Un grand merci pour l'encadrement clinique durant tout cet externat et pour être venu de nombreuses fois en heureux et inopiné soutien lorsque nous en avions le plus besoin. Merci également pour l'engagement dans les cours dispensés qui, bien qu'annonçant des sujets manquant parfois d'attrait, s'en sont toujours trouvés bien plus intéressants que de prime abord. Je pense que les méthyl méthacrylates ont une dette éternelle à votre égard.

Au jury de thèse,

Monsieur le Docteur Julien ALIGON

- Maître de conférence en informatique – Université Toulouse Capitole
- Docteur de l'université de Tours – discipline informatique

Merci pour la bonne humeur durant les réunions à Restore ainsi que pour avoir contribué, via ces réunions, à me faire découvrir l'explicabilité en intelligence artificielle.

Merci au Dr. Delphine EHRHARDT sans qui je n’aurais même pas eu l’idée de m’engager dans ces études. Merci donc de m’avoir fait découvrir ce formidable métier. Merci également pour, avec le Dr. Olivier GIRAULT, m’avoir permis de réaliser un remplacement en tant qu’assistant qui m’a énormément appris ainsi que pour ce tout premier remplacement en tant que praticien qui fut une confirmation évidente de m’être engagé dans la bonne voie.

Merci au Dr. Pascal BRU pour, avec le Dr. Delphine EHRHARDT, m’avoir permis de rester motivé durant les premières années où le manque de pratique clinique me manquait cruellement.

Un immense merci aux Dr. Paul SABAHI, Bruno SOUCHE, Sophie DI LUCCI et Margaux BROUTIN pour ce stage actif qui m’a tant appris sur bien des plans. J’en retire également des leçons de vie qui, je pense, me serviront encore longtemps, ainsi qu’une passion décuplée pour les lambeaux et autres points de suture.

Merci à mes camarades de promotion, sans qui ces trois années d’externat à l’Hôtel Dieu n’auraient clairement pas eu la même saveur. Ce sont là des amitiés forgées dans les lapins, le sang et les bris radiculaires qui laissent une empreinte indélébile.

Merci également à tous les gens qui permettent à la faculté et aux deux centres de soins, et plus particulièrement à l’Hôtel Dieu, de fonctionner au quotidien, ainsi qu’à Mme Batoul BASHOUN dont la réactivité s’est révélée d’une grande aide durant la finalisation de ce manuscrit et les démarches administratives associées.

Enfin, et le plus important : à mes parents pour m’avoir fourni l’ouverture intellectuelle nécessaire et la possibilité concrète de m’engager dans ces études, et plus globalement à toute ma famille pour me supporter et supporter, depuis, mon humour contaminé par le domaine médical, merci du fond du cœur.

ABRÉVIATIONS

IA : Intelligence Artificielle

LIME : *Local Interpretable Model-agnostic Explanations*

ML : *Machine Learning*

SHAP : *Shapley Additive exPlanations*

XAI : *eXplainable Artificial Intelligence*

SOMMAIRE

I. L’explicabilité en intelligence artificielle

1. Introduction à l’intelligence artificielle

a. Naissance et champs d’application de l’intelligence artificielle [2, 3, 4]

Le concept d’intelligence artificielle (IA) est décrit pour la première fois par Alan Turing en 1950 avant d’être cité pour la première fois sous ce nom six ans plus tard par John McCarthy lors de la conférence de Darmouth où l’IA s’affirme comme domaine de recherche à part entière. Les années 70 voient sa première application au domaine médical avec les modèles MYCIN et CASNET notamment, cependant il faudra attendre les années 90 et 2000 pour voir la puissance de calcul sans cesse croissante des ordinateurs permettre la réalisation d’IA en nombre sans cesse croissant.

Dans le domaine médical, les premiers modèles d’IA ont été développés dans le but de faciliter la prescription d’antibiotiques les plus adaptés possibles (MYCIN) ou de conseiller les médecins sur la gestion des patients atteints de glaucomes (CASNET). Le modèle *Arterys* (qui analyse des IRM cardiaques), quant à lui, est le premier modèle d’apprentissage profond (i.e., deep learning) basé sur le cloud approuvé par la *U.S. Food and Drug Administration* en 2017 avant de se développer et de se diversifier plus encore.

En effet, la quantité phénoménale d’informations capable d’être traitée par un modèle de machine learning (ML) à partir de données radiologiques, pathologiques, ou encore biochimiques surpasse la capacité du cerveau humain et montre des perspectives de monitoring de patients ou de diagnostic de cas complexes intéressantes.

Ses applications peuvent s’étendre à de nombreux champs médicaux : l’analyse d’examens complémentaires, la chirurgie, la modélisation en 3D de modèles complexes, le développement et la production de médicaments, la recherche, la gestion médicale ou encore l’enseignement.

Il est possible de classer les IA médicales en deux catégories : les IA physiques (comme les bras robotiques) et les IA virtuelles (comme les logiciels d’aide à la décision).

Dans le domaine de l’odontologie on retrouve actuellement essentiellement des IA virtuelles notamment :

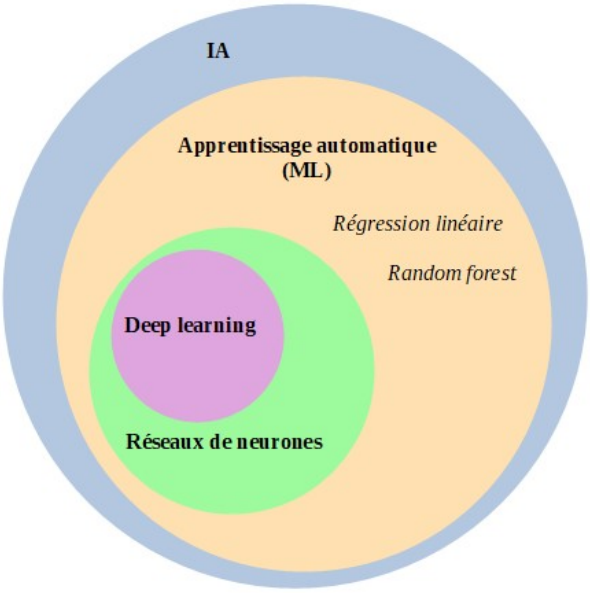
– dans l'aide au diagnostic (*par exemple Yilmaz et al. 2017 obtiennent 94 % d'accuracy dans la différenciation de lésions périapicales avec des lésions kératokystiques sur des images CBCT*) ;
– dans le traitement de pathologies (*par exemple Fu et al. 2017 améliorent l'intelligibilité orale des patients post-chirurgie grâce à la conversion via ML de murmures non-audibles en paroles compréhensibles*) ;
– ou encore dans la prédiction de pathologies (*comme l'identification d'individus sujets à la survenue de caries radiculaires en fonction de facteurs nutritionnels, cliniques et d'habitudes de vie par Hung et al. 2019*).

b. Comment entraîner et utiliser un algorithme de machine learning [5, 6]

Le développement d'une IA nécessite une base de données importante en premier lieu. Cette base de données est divisée en plusieurs parties, la première étant le jeu de données d'entraînement qui est fourni à un algorithme qui va apprendre à prédire un diagnostic avec (pour prendre un exemple du domaine médical). On ajuste ensuite les hyperparamètres de l'algorithme à l'aide d'un second jeu de données, le jeu de validation, afin de le rendre le plus performant possible avant de le tester sur le jeu de données de test qui permet d'évaluer la performance de l'apprentissage réalisé par l'algorithme. À ce niveau-là peuvent se poser un problème de sous-apprentissage (l'algorithme n'a pas réussi à tirer des règles de prédiction suffisamment performantes du jeu de données) ou de sur-apprentissage (les règles sont « trop » performantes, comme un étudiant qui a appris par cœur les réponses pour un examen mais est incapable de les comprendre et de s'adapter si les questions diffèrent un peu de celles qu'il a appris).

Une fois que l'algorithme est testé et obtient des résultats jugés satisfaisants, il est désormais capable dans la théorie de prédire dans la population d'où est extraite la base de données (pour peu qu'elle soit représentative de cette population) des diagnostics avec la même précision que celle observée sur le jeu de données de test.

Il existe de nombreux modèles d'algorithmes différents, comme la régression linéaire, les arbres de décision, les *random forests*, les réseaux de neurones… comme on peut le voir sur le schéma suivant.



Exemples de différents modèles d'algorithmes catégorisés en tant qu'IA

Cependant, la complexité et la précision des algorithmes augmentent de paire, au détriment de la compréhension de son mécanisme de décision comme on peut le voir sur le schéma ci-dessous. Ainsi, à partir d'un certain niveau, on peut parler de véritable boîte noire dans laquelle on rentre des données et de laquelle il sort une prédiction sans que l'on sache comment l'algorithme est passé de l'un à l'autre. C'est là qu'entre en jeu le domaine de l'explicabilité, ou *eXplainable Artificial Intelligence* (XAI), mais avant de poursuivre plus loin il est important de s'arrêter quelques instants afin de définir et comprendre quelques termes et notions qui nous serviront pour la suite.

Figure 1.10 illustrates the accuracy-interpretability trade-off of popular ML algorithms:

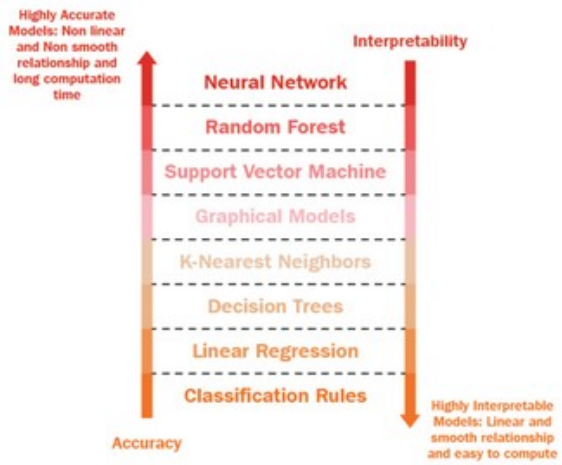


Figure 1.8 – Accuracy-interpretability trade-off diagram

Illustration issue du livre *Applied Machine Learning Explainability Techniques* de Aditya Bhattacharya [8] illustrant la diminution de l'interprétabilité des modèles en fonction de l'augmentation de leur complexité et de leurs performances

2. Qu’est-ce que l’explicabilité ? [8, 10, 14]

a. Le problème de la boîte noire

Qu’est-ce qu’une boîte noire, ou *black-box* en anglais ? Dans un algorithme hors ML, le programmeur donne des instructions spécifiques à la machine qui les exécute. Dans un algorithme de ML, le programmeur nourrit l’algorithme avec les données du jeu d’entraînement et l’algorithme en tire ses propres règles qui lui servent ensuite à faire des prédictions quand on lui donne de nouvelles données. Seulement, on ne sait pas quelles sont ces règles et le schéma ci-dessous résume cette problématique. Plus généralement une boîte noire, ou boîte opaque, est un système dont on ne connaît pas le fonctionnement interne.

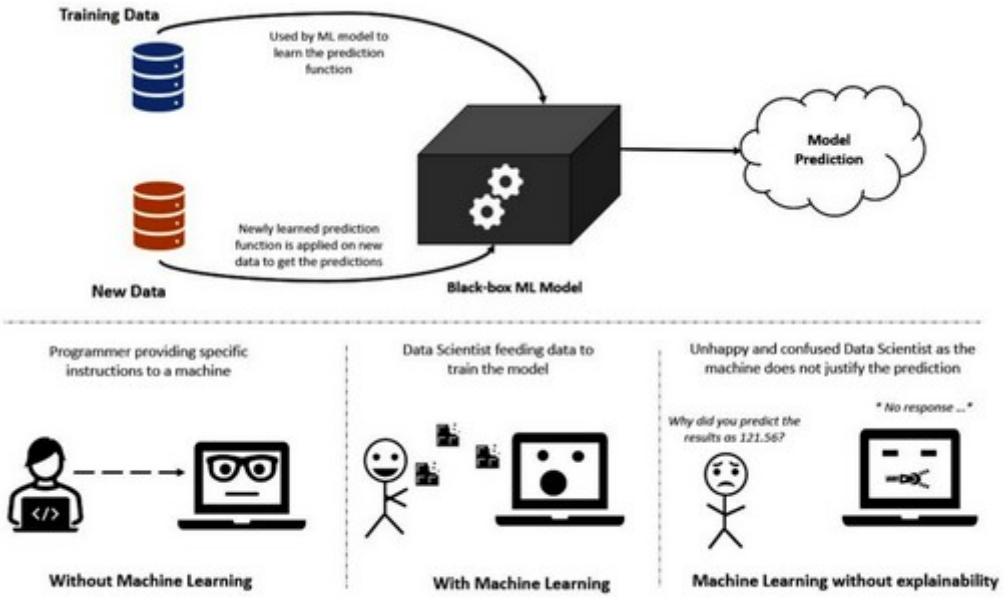


Figure 1.1 – Conventionally, black-box models do not provide any rationale behind predictions

Illustration issue du livre *Applied Machine Learning Explainability Techniques* de Aditya Bhattacharya [8] illustrant la méthode d’entraînement d’un modèle de ML et le problème de boîte noire qu’elle engendre

Du fait de la méconnaissance de ces règles, des biais dans la prise de décision peuvent passer inaperçu. Ces biais peuvent avoir différentes origines :

- de l’algorithme lui-même ;
- du jeu de données d’entraînement ;
- du *data drift* (évolution progressive des données entrantes dans le modèle et de leur répartition), du *concept drift* (évolution progressive des relations entre les données entrantes et la donnée cible à prédire) ;
- du sur-apprentissage (*overfitting*) du modèle ou de son sous-apprentissage (*underfitting*).

L’explicabilité permet donc de :

- vérifier et débugger les modèles de ML ;
- améliorer les modèles de ML avec une approche centrée sur l’utilisateur ;
- découvrir de nouveaux liens d’associations entre variables *(*grâce à des liens potentiellement mis en évidence par le ML) ;
- être en conformité avec la législation (déjà en place ou future) ;
- augmenter la confiance de l’utilisateur dans la prédiction du modèle.

En somme, on cherche à obtenir de la transparence sur une boîte opaque.

Mais qu’est-ce que l’explicabilité ? L’explicabilité c’est comprendre pourquoi un modèle prédit un résultat, à différencier de l’interprétabilité qui est comprendre comment un modèle prédit un résultat bien que souvent ces deux termes soient mélangés et interchangés. Les schémas suivants illustrent où se place l’explicabilité : dans le processus d’entraînement continu d’un modèle de ML ainsi que de la vérification de ses résultats pour le premier, et dans le processus d’adoption de l’IA par la communauté scientifique ainsi que le grand public dans le second (où l’explicabilité est considérée comme appartenant à la base de ce processus).

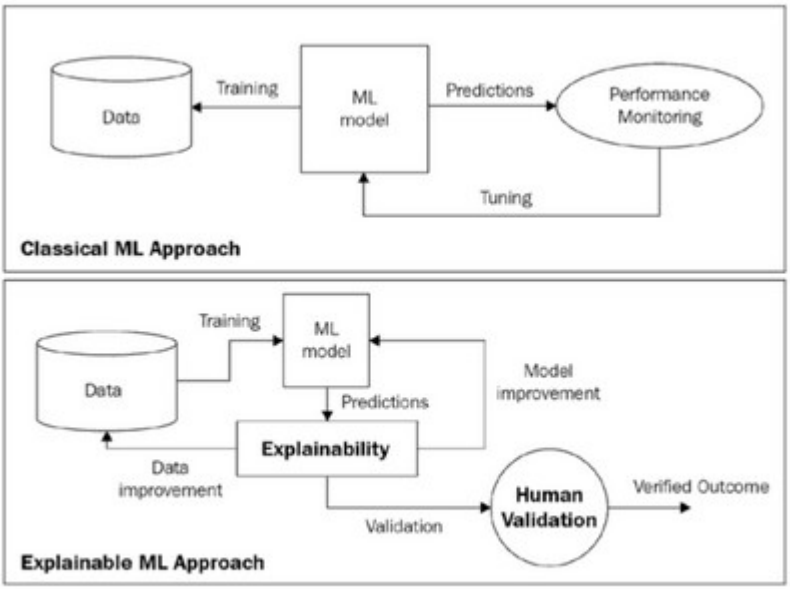


Figure 1.4 – Comparison between classical ML and explainable ML approach

Illustration issue du livre *Applied Machine Learning Explainability Techniques* de Aditya Bhattacharya [8] illustrant la place de l’explicabilité dans le fonctionnement d’un algorithme de ML

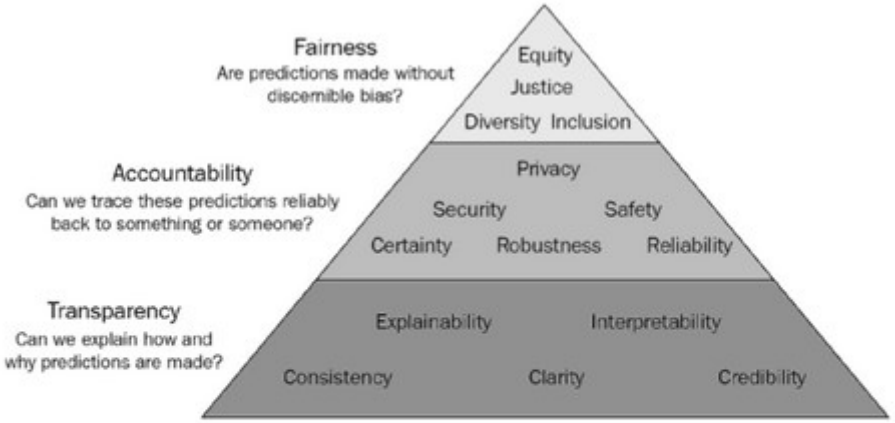


Figure 1.6 – FAT model of explainable ML (from Interpretable Machine Learning with Python by Serg Masís)

Figure 1.6 shows the pyramid that forms the FAT model of explainable ML system for increasing AI adoption. Let us discuss about defining explanation methods and approaches in the next section.

Illustration issue du livre *Applied Machine Learning Explainability Techniques* de Aditya Bhattacharya [8] illustrant la place de l'explicabilité dans le processus d'adoption de l'IA

b. Concepts et définitions liés à l'explicabilité

L'explicabilité amène donc avec elle son lot de concepts et définitions :

Explicabilité locale : Explicabilité pour une seule instance (un patient) du jeu de données pour comprendre comment les variables influent sur la prédiction relative à cette instance.

Explicabilité globale : L'explicabilité globale permet de déterminer quelles variables sont utilisées et comment le modèle prend ses décisions et ainsi expliquer son fonctionnement global.

Explicabilité intrinsèque : Quand il est possible de comprendre clairement la logique permettant de passer d'une entrée (input) à une sortie (output) d'un modèle (exemples : modèle linéaire, arbre de décision simple).

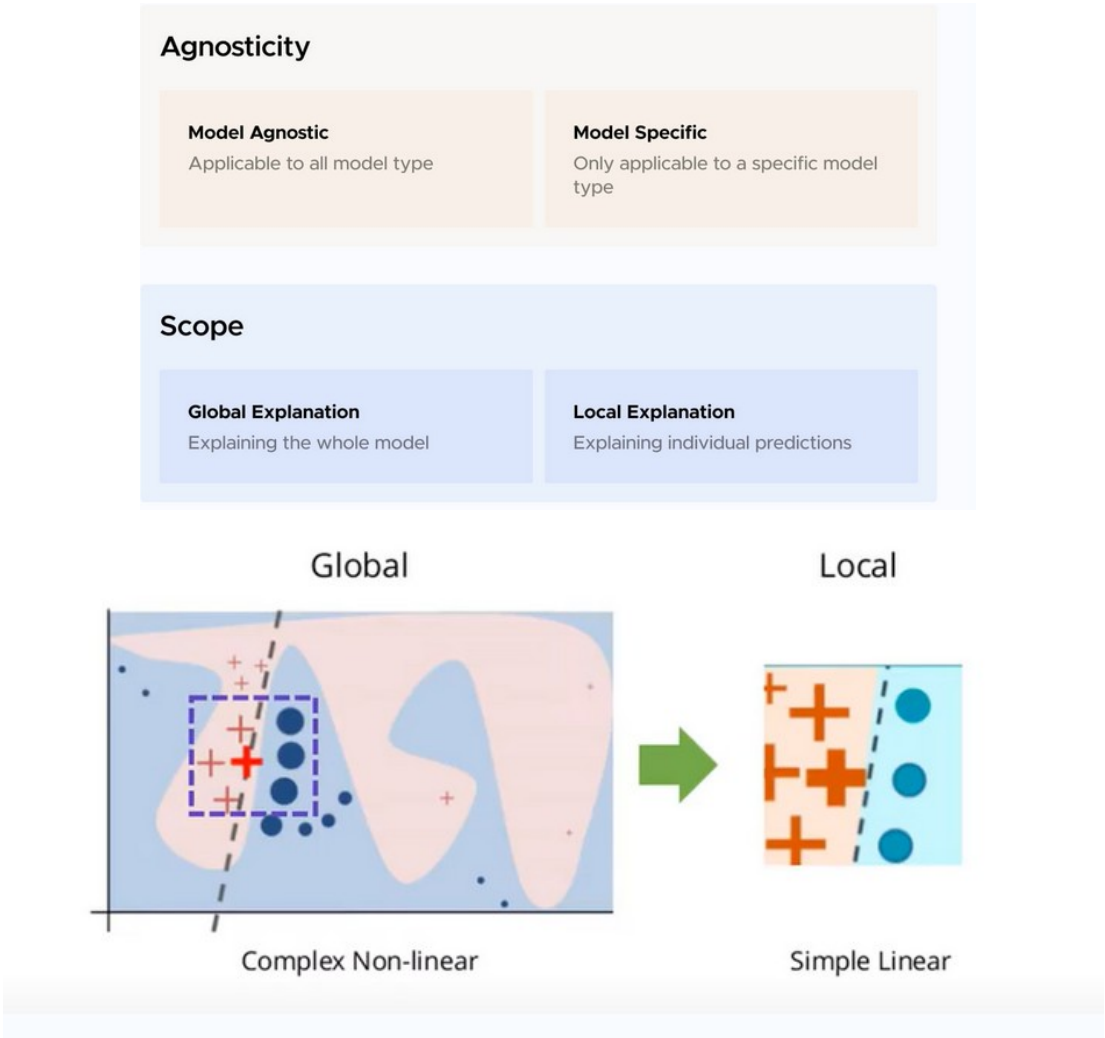
Explicabilité extrinsèque ou *explicabilité post hoc* : Quand il n'est pas possible de comprendre clairement la logique sus-mentionnée et qu'il est nécessaire d'utiliser une technique d'explicabilité séparée après la prédiction du modèle pour la comprendre.

Explicabilité modèle-spécifique (Model-specific XAI) : Technique de XAI utilisable pour un modèle spécifique d'algorithme (ex : visualisation de l'arbre pour un modèle *decision tree* applicable uniquement aux modèles *decision tree*).

Explicabilité modèle-agnostique (Model-agnostic XAI) : Technique de XAI utilisable sur différents types de modèles d'algorithme.

Explicabilité basée sur le modèle (*model-centric explainability*) : Explication de comment les entrées sont utilisées par le modèle pour parvenir aux sorties.

Explicabilité basée sur les données (*data-centric explainability*) : Explication des données pour comprendre si elles sont cohérentes, bien organisées et bien adaptées.



Illustrations issues du blog *Censius AI Observability* [9]. La première résume les concepts d'agnosticité/spécificité et de globalité/localité des techniques de XAI. La seconde illustre le processus d'explication locale.

Une explication, et plus globalement l'utilisation de la XAI en support à l'IA seule, pour être utile et fiable doit répondre à de nombreux critères, elle doit être :

- cohérente avec les connaissances actuelles ;
- fidèle au fonctionnement du modèle expliqué (une technique d'explication pas assez fidèle au fonctionnement du modèle reviendra à peindre la boîte noire en blanc et non à la rendre transparente, apportant son lot de biais) ;
- concise et résumée pour permettre une compréhension claire par l'utilisateur, avec un design remplissant les critères d'une bonne interface humain-machine ;
- contrastée, c'est-à-dire que par contraste il doit être possible de comprendre pourquoi deux instances avec deux prédictions différentes ont des prédictions différentes ;
- capable d'expliquer les cas rares et sortant de la norme ;
- non-biaisée ;
- capable de mettre en avant les liens de causalité ;
- robuste (exemple : si de petites modifications des variables provoquent de grandes modifications des prédictions, alors le modèle n'est pas robuste. Une explication robuste est une explication stable, fidèle, cohérente, et adaptable qui conserve sa validité dans différentes situations ou conditions) ;
- capable d'améliorer la confiance dans la prédiction en diminuant l'effet de boîte noire.

Nous verrons donc ici deux méthodes d'explicabilité très largement utilisées aujourd'hui afin de comprendre leur fonctionnement : SHAP et LIME.

c. Technique d'explicabilité : SHAP [1, 7, 8, 9, 11, 12, 13]

SHapley Additive exPlanations, ou SHAP, est une méthode d'explication basée sur la théorie du jeu. Si chaque variable est un joueur et la sortie le résultat du match, la méthode SHAP consiste à rejouer le match en retirant à chaque fois un joueur différent (et en gardant tous les autres) afin de comparer le résultat du match avec et sans ce joueur pour déterminer son impact dans le match.

C'est une technique d'explicabilité qui est :

- post-hoc ;
- model-agnostique ;

- basée sur le modèle et modèle-dépendante : avec un même jeu de données, les résultats seront différents en fonction de l'algorithme de ML qui a traité ce jeu de données (l'ordre d'importance des variables pour le diagnostic peut par exemple changer entre deux modèles différents) ;

- capable de réaliser de l'explication locale et de l'explication globale ;

- une technique très utilisée et éprouvée donnant des résultats fiables.

Cependant la méthode SHAP présente des défauts et limites :

- du fait de la nécessité de rejouer la prédiction pour chaque variable de chaque instance, c'est une méthode qui nécessite un temps de computation important ;

- un utilisateur qui n'est pas habitué à l'interprétation des scores SHAP pourrait commettre des erreurs d'interprétation car ces scores n'ont pas la même unité que la variable à laquelle ils sont associés;

- en cas de biais de classification du jeu de données, la méthode SHAP peut générer des explications irréalistes et ne pas révéler les biais sous-jacents. Plus globalement, si le jeu de données est biaisé alors ces biais ne seront pas montrés par la méthode SHAP ;

- la méthode SHAP part du principe que les variables sont indépendantes et non corrélées. Les corrélations potentielles ne sont donc pas prises en compte et certaines variables peuvent se voir attribuer un faible score SHAP malgré une association significative avec la prédiction finale car leur colinéarité avec une variable plus importante qui a du poids dans la prise de décision les fait passer inaperçu.

Pour pallier ce dernier point, une méthode a été développée pour évaluer la stabilité de la liste notamment en cas de variables colinéaires. Cette méthode tend à calculer le NMR (Normalized Movement Rate) en supprimant itérativement les variables les plus importantes et en observant ainsi si l'ordre des autres variables subit un quelconque changement. Un NMR faible est alors synonyme de liste stable.

Cette méthode permet également de comparer deux modèles de ML différents en comparant leur NMR une fois l'explication fournie par SHAP.

d. Technique d’explicabilité : LIME [1, 7, 8, 9, 10, 11, 12, 13, 15]

Local Interpretable Model-agnostic Explanations, ou LIME, est une méthode qui convertit, pour une instance, le modèle en un modèle linéaire local duquel il ressort des coefficients qui représentent le poids des variables dans la prédiction. En somme, la méthode propose des modèles de substitution pour se rapprocher localement du modèle de la boîte noire, trop complexe dans son esemble. C’est une approximation locale. Dans un modèle de classification, il montre aussi la probabilité pour un sujet d’appartenir à chaque classe.

C’est une technique d’explicabilité qui est :

– post-hoc ;
– model-agnostic ;
– model-centric ;
– modèle-dependante ;
– capable de faire de l’explication locale ;
– capable de faire de l’explicabilité sur des types de données très variées ;
– facile d’utilisation ;
– rapide à computer ;
– fonctionne avec un grand nombre de types de données différents (valeurs, images, texte…)

Cependant c’est une méthode qui :

– ne permet pas de faire d’explication globale ;
– ne permet pas de mettre en avant des relations colinéaires car elle observe l’impact des variations d’une variable sur le résultat tandis que les autres caractéristiques demeurent constantes, ce qui est très éloigné de la réalité biologique ;
– si le jeu de données est biaisé, les résultats de l’explication par LIME ne montreront pas ces biais ;
– les explications peuvent parfois être instables, c’est-à-dire que les explications de deux cas relativement similaires peuvent être significativement différentes ;
– nécessite le réglage de beaucoup d’hyperparamètres dont le choix va fortement influencer les résultats de l’explication.

e. Quelle technique d'explicabilité choisir ? [1, 11, 12, 13, 15]

SHAP	LIME
Explicabilité locale et globale	Explicabilité uniquement locale
Pas de mise en évidence des relations colinéaires MAIS existence de la méthode NMR pour y pallier	Pas de mise en évidence des relations colinéaires
Temps de computation important	Prend en charge des types de données variés
Attention à l'interprétation des scores (se focaliser sur l'ordre de contribution des variables)	

Les deux techniques, assez similaires sur le type d'explicabilité qu'elles sont capables de fournir (modèle-centrique/dépendante/agnostique…) présentent leurs avantages et leurs inconvénients. Cependant, à travers sa capacité à faire de l'explicabilité locale et de la méthode du NMR pour mettre en évidence les relations colinéaires, la méthode SHAP semble présenter un plus grand intérêt pour l'analyse de données médicales qui présentent souvent des relations colinéaires comme, par exemple, une association entre un fort taux de cholestérol et un indice de masse corporelle élevée [1] ainsi que la possibilité de faire de l'explication globale : en biologie on étudie généralement des populations importantes afin de pouvoir en tirer une norme biologique et observer ce qui s'éloigne de cette norme, rendant l'utilisation de l'explicabilité globale adéquate voir nécessaire.

De plus, on peut prendre en compte les articles qui comparent directement les deux techniques sur différents éléments :

Adversarial attacks

Un article de Slack et al. [11] met en lumière un problème majeur dans le domaine de la XAI. Il souligne que des individus mal intentionnés peuvent manipuler délibérément des ensembles de données de manière à ce que des biais importants restent invisibles lors de l'application de techniques d'explication populaires comme LIME et SHAP.

Les décideurs (les professionnels de santé dans notre cas) ne peuvent prendre des décisions éclairées à l'aide de l'IA qu'à condition de comprendre et de faire confiance aux explications fournies par la XAI. Pour ce faire, ils doivent être en mesure de diagnostiquer

les erreurs et les biais du modèle. Cependant les auteurs de l'article montrent que des jeux de données extrêmement biaisés créés à l'aide de leur méthodologie sont capables de tromper ces techniques d'explication en générant des explications contrôlées arbitrairement qui ne révèlent pas les biais sous-jacents. Cette manipulation peut conduire à la création délibérée d'ensembles de données truqués, ce qui peut être exploité par des groupes d'intérêt (lobbies) pour prouver des affirmations erronées par exemple.

L'article conclut que les techniques d'explication post hoc existantes ne sont pas suffisamment robustes pour évaluer le comportement discriminatoire des jeux de données dans des applications sensibles. Cependant, les auteurs ont également constaté que SHAP est moins trompé que LIME dans leur évaluation sur trois jeux de données différents. Cette différence résiderait dans la manière dont certains paramètres sont choisis. Dans LIME, ces fonctions sont définies de manière heuristique, tandis que SHAP s'appuie sur des principes de théorie des jeux pour garantir que les explications respectent certaines propriétés souhaitées.

Qualité des graphiques

Un article de Duell et al. [12] met également en avant l'efficacité des illustrations de SHAP par rapport à LIME pour communiquer des explications, apportant une clarté supplémentaire. Cependant cet argument est à pondérer, car tous les auteurs ne sont pas d'accord dessus.

Sparsité des données

Une étude de Roberts et al. [13] met en lumière l'impact de la sparsité des données sur les performances de LIME et SHAP. SHAP semble mieux adapté aux données éparses en raison de sa gestion du compromis entre biais et variance (SHAP semble présenter un biais plus élevé mais une variance plus faible par rapport à LIME, en particulier dans les environnements de données à forte sparsité), et l'introduction de la contrainte de complétude dans LIME (CLIMB) améliore également sa performance dans ces contextes tout en restant rapide à calculer et simple à utiliser. Ces résultats suggèrent que le choix de la méthode XAI dépend du contexte spécifique de l'application et de la densité des données.

Or dans notre cas, avec l'utilisation de moins de dix variables pour notre expérimentation, nous nous trouvons dans un cas de sparsité de données.

Dans un article de Sheu et Pardeshi [10] LIME est qualifié comme « *populaire pour mettre en évidence les caractéristiques importantes et fournir des explications basées sur ses coefficients, mais* [est hasardeux] *lors de l'étape d'échantillonnage, ce qui le rend inacceptable pour des applications médicales.* ».

En somme, ce cumul d'éléments fait que pour la proposition de protocole qui suit, SHAP semble être la méthode de XAI la plus indiquée des deux que nous venons de voir.

3. Contexte et objectifs de l'étude

a. Observations conduisant à l'imagination et l'élaboration d'un protocole expérimental [16, 17, 18, 19, 20, 21, 22]

Au cours de mes recherches sur Google Scholar, je me suis aperçu qu'un bon nombre d'articles comparant les différences de performances entre une IA et des experts recrutaient peu d'experts pour leur étude. À tel point que ces experts peuvent se compter sur les doigts d'une, voir deux mains pour certaines études. Cependant il y a certaines exceptions, comme une étude recrutant vingt-huit experts [18]. Néanmoins pour atteindre ce chiffre, les chercheurs ont recruté des étudiants.

De plus, toujours au cours de ces recherches, il semble apparaître que dans les études des cinq dernières années, l'IA atteigne des performances de diagnostic proches de celles de praticiens. Pour la bibliographie en lien avec cette partie, on retrouve une similarité de performance plus ou moins marquée pour certains articles [16, 17, 20, 21], une meilleure performance de l'IA pour un article [18] tandis que pour un autre article [19] elle présente de meilleures performances face à une lecture de la mammographie par un seul expert mais dans le cas d'une double lecture par deux experts (comme c'est le cas dans le système de santé britannique) l'utilisation de l'IA pour la première lecture permet au second lecteur (un expert) d'avoir des performances non-inférieures à une double lecture par deux experts avec une charge de travail réduite de 88 %.

Si les publications sur le ML sont désormais très nombreuses, la XAI, elle, est en plein essor et les expérimentations sur son apport ou non sur les performances d'un expert sont plus rares.

b. Élaboration du plan global de l'étude

Partant de ce constat, nous avons imaginé une expérimentation permettant d'évaluer l'intérêt et les bénéfices potentiels que pourrait apporter la XAI au professionnel de santé, et plus précisément au chirurgien-dentiste dans notre cas, par rapport à l'apport de l'IA seule.

L'idée est donc de recruter des chirurgiens dentiste (en tentant de recruter un plus grand nombre d'entre eux comparé aux études sus-mentionnées) afin de leur faire remplir un questionnaire automatisé sur un support informatique. Chaque praticien devra répondre à une série de questions leur présentant des cas de parodontite sous trois formes différentes :

– sous forme d'une liste réduite de variables en lien avec la parodontite ;

– sous forme de cette même liste associée à une prédiction de perte d'attache issue du ML ;

– sous forme de cette même liste associée à cette même prédiction, cette fois complétée par des résultats de XAI issus de la méthode SHAP.

Les praticiens devront, à chaque fois, prédire la survenue d'une perte d'attache supérieure à 2 mm en fonction des données fournies. Le temps de réponse sera également mesuré.

De plus, pour les deux premières formes de questions il leur sera demandé de noter les variables présentées en fonction de l'importance qu'ils leur donnent pour poser leur diagnostic pour le cas associé. Cela permettra de comparer la hiérarchie des variables pour le diagnostic entre les praticiens et la méthode SHAP.

c. Objectif principal de l'étude

L'hypothèse de ce travail est que l'emploi de la XAI améliore la justesse et la rapidité du diagnostic pour le chirurgien-dentiste par rapport à l'utilisation des données seules ou associées au machine learning.

d. Objectifs secondaires

De plus, plusieurs objectifs secondaires peuvent être dors et déjà définis sous la forme des questions suivantes auxquelles nous tenterons de répondre :

– Retrouve-t-on (ou non) une amélioration de la justesse et de la rapidité quelle que soit l’expertise du praticien ?

– Retrouve-t-on une amélioration de la justesse et de la rapidité du diagnostic avec l’utilisation des données associées au ML par rapport à l’utilisation des données seules ?

– Le praticien et la méthode SHAP donnent-ils le même ordre d’importance aux variables pour la réalisation de leur diagnostic ?

II – Matériel et méthodes pour la réalisation d'un essai randomisé sur l'intérêt de la XAI pour le chirurgien-dentiste

Ce protocole a été élaboré en collaboration avec Elodie Escriva.

1. Défis et biais attendus

Lors d'une première réflexion autour de la conception du protocole, un certain nombre de défis et de biais potentiels ressort :

- Comment recruter des professionnels de santé avec un emploi du temps déjà chargé ?
- Comment avoir, pour chaque praticien interrogé, le plus de réponses (et donc de données) possible sans que le questionnaire ne soit abandonné en cours de route ou réalisé sans le soin nécessaire en cause d'une durée trop longue ?
- Comment ne pas avoir de réponses erronées à cause d'une mauvaise compréhension ou utilisation de l'interface informatique ou de la présentation ou formulation de la question ?
- Une augmentation des performances du praticien pourrait être observée à force de répondre aux questions dues à un apprentissage et une habituation, ce qui ne serait pas lié à un quelconque apport du ML ou de la XAI. Une habituation peut également être observée pour un patient en particulier si les différentes étapes des tests sont réalisées à la suite, toujours dans le même ordre – dans le cas par exemple d'un affichage des données seules, puis des résultats ML puis des résultats XAI toujours dans le même ordre. Cette habituation peut engendre un impact sur l'apport de chaque méthode – ML et XAI – pour les utilisateurs.

Une partie des biais mentionnés peuvent être résolus via les trois moyens suivants. Afin de maximiser le nombre et la qualité des réponses simultanément, nous pouvons limiter le temps de l'expérience en limitant le nombre de cas étudiés par chaque utilisateur. Si nous considérons que le temps nécessaire pour évaluer un cas est de 1min au maximum, et que nous proposons 10 patients différents à chaque utilisateur, alors l'expérimentation se limite à une durée de 10min. Nous privilégions ainsi la qualité des résultats. Cette solution permet aussi de recruter plus facilement des personnes avec un emploi du temps chargé, puisque la durée du test resterait relativement courte. Il faut cependant avoir un ratio entre le nombre d'utilisateurs et le nombre de cas suffisant afin de garantir une exploitabilité des

résultats lors de l'analyse statistique. Enfin, puisque la durée du test est courte, le risque d'adaptation des utilisateurs au test est réduit.

Deuxièmement, dans le cas d'une habituation des praticiens à chaque patient, il est possible de séparer les étapes de tests afin que les utilisateurs n'étudient pas, pour tous les patients, toutes les données dans le même ordre. Une répartition aléatoire de chaque étape peut être mise en place afin de limiter ce biais. En séparant les phases, nous limitons le risque que les praticiens ne reconnaissent leurs patients grâce à leurs données et ne soient influencés pour la pose de leur diagnostic par un diagnostic précédemment posé lors d'une autre phase. Ce stratagème permet aussi de réduire la durée des tests et donc possiblement, d'améliorer la qualité des réponses des praticiens.

Finalement, afin de limiter les réponses erronées en cas de mauvaise compréhension, des exemples d'entraînement peuvent être mis en place avant la réalisation du test afin que l'utilisateur se familiarise avec la tâche demandée. Ces exemples ne doivent cependant pas être trop longs afin de ne pas impacter négativement le test.

2. Recrutement et critères d'éligibilité

La population cible de l'expérimentation est celle des soignants, et plus particulièrement des chirurgiens-dentistes, afin d'évaluer l'intérêt potentiel pour elle de l'utilisation des explications en corrélation avec le ML.

Cependant, avoir une population représentative de la population cible dans cette étude n'est pas évident. En cause notamment, la surcharge de travail du personnel de santé français dû à son sous-effectif. Nous tenterons donc de recruter le plus de chirurgiens-dentistes libéraux possible, mais aussi des chirurgiens-dentistes hospitaliers et des enseignants-chercheurs de la faculté dentaire, idéalement experts en parodontologie, ainsi que des étudiants en fin de cycle avec déjà deux à trois années de pratique clinique derrière eux pour augmenter nos effectifs. Ainsi, les critères d'éligibilité pour être recruté sont les suivants :

– être un chirurgien-dentiste formé ou en cours de formation ;

– avoir des connaissances en parodontologie ;

– avoir au minimum deux années d'expérience clinique.

Les critères de recrutements amènent à une population avec un niveau d'expertise hétéroclite. Pour corriger un potentiel biais lié à cette hétérogénéité, les experts sont donc regroupés en clusters selon deux critères :

– un critère subjectif quant à leur niveau d'expertise en parodontie (sur une échelle de 1 à 10) ;

– un critère objectif quant à leur statut (chirurgien-dentiste libéral, chirurgien-dentiste fraîchement diplômé, étudiant de 6e année, enseignant-chercheurs en faculté dentaire).

Ces clusters peuvent être comparés avec le groupe entier d'experts afin de mesurer une augmentation ou diminution des critères observés par rapport au groupe entier. Ce procédé permet ainsi de mesurer l'impact des explications par rapport au niveau d'expertise des utilisateurs.

De plus, il n'y a pas de groupe témoin séparé : chaque expert sert de témoin à travers les questions sous le mode 1 (avec les données seules) et d'expérimentateur à travers les questions sous les modes 2 et 3 (avec les données + le ML ± la XAI, cf *7. Déroulé de l'expérimentation*).

En terme de quantité, une centaine d'experts semble être un nombre minimum intéressant. En effet, cela reste objectif atteignable tout en permettant d'obtenir 2700 réponses à propos de 90 patients (cf *7. Déroulé de l'expérimentation*).

3. Jeu de données et pré-traitement

La recherche du jeu de données à utiliser dans l'objectif de présenter des résultats de ML et de XAI aux expérimentateurs s'est révélée complexe.

J'ai trouvé plusieurs banques de données dentaires en accès libre sur le *National Center for Health Statistics* (qui ne correspondaient pas à ce que je cherchais) et le *National Dental Practice-Based Research Network* sur lequel j'ai pu trouver plusieurs jeux de données intéressants, dont un sur les fêlures dentaires (*Cracked Tooth Registery)*.

Après des premiers résultats tests issus du ML, nous nous sommes rendu compte que le jeu de données n'était pas adapté : pour les variables cibles, le ML présentait une précision

trop faible. Après ré-évaluation, j'ai pu redéfinir mes critères de recherche de jeu de données comme suit :

– le jeu de données doit présenter une variable que le chirurgien dentiste pourra prédire à l'aide d'autres variables et qui servira d'étiquette pour l'entraînement du ML ;

– le jeu de données doit présenter un petit nombre de variables utiles pour la prédiction de la variable cible et de variables inutiles à incorporer dans les variables présentées au chirurgien-dentiste pour qu'il puisse les classer par ordre d'importance dans sa prédiction, avec un nombre de variable qui soit compatible avec une durée restreinte pour le remplissage du questionnaire ;

– ce petit nombre de variables sus-mentionnées doit permettre au ML d'obtenir une bonne précision ;

– il ne doit pas y avoir de lien de corrélation trop important ou exclusif entre une variable et la cible (par exemple, dans le *Cracked Tooth Registery,* une réponse positive à la variable « sensibilité à la pression » entraînait forcément une réponse positive à la variable « dent symptomatique ») ;

– il doit y avoir suffisamment de données de bonne qualité dans les variables sélectionnées pour entraîner le ML (par exemple, dans nombre de jeux de données, certaines variables avaient un trop fort taux d'absence de réponse par rapport à l'ensemble des participants impliqués dans le recueil de données pour pouvoir être utilisées dans le cadre du ML).

Finalement, après une nouvelle recherche de jeux de données infructueuse, le choix s'est porté sur un jeu de données provenant de l'étude transversale NHANES dans lequel se trouvent des données médicales, socio-démographiques et dentaires de plusieurs milliers de sujets [81].

Pour des raisons de temps de préparation du jeu de données, nous n'avons pas pu mettre en œuvre les étapes suivantes.

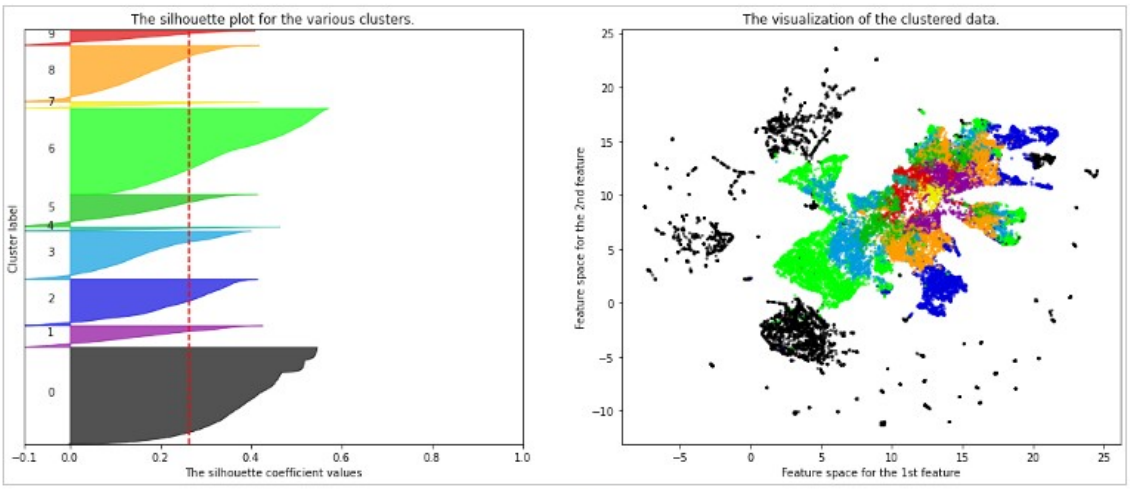
Néanmoins voici le déroulé théorique. Une fois le jeu de données sélectionné, il reste encore l'étape du pre-processing pour le préparer à être utilisé. En effet, pour éviter certaines erreurs lors de la phase d'apprentissage, il faut le rendre uniforme et cohérent en modifiant les données manquantes ou incomplètes. Plusieurs types de modifications sont possibles, pour n'en donner que quelques exemples on peut supprimer les variables pour

lesquelles les données sont trop incomplètes ou remplacer les données manquantes par une valeur nulle, une constante, une moyenne [10]…

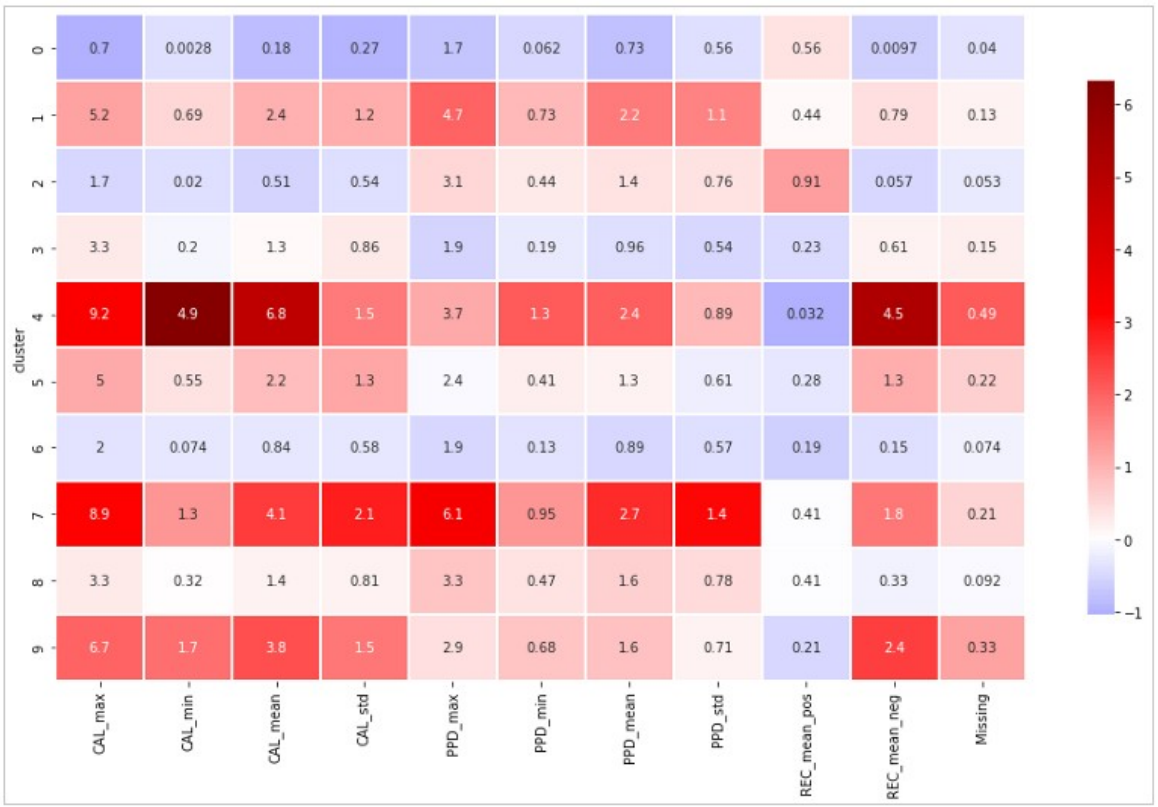
En parallèle de cela, il faut également sélectionner les variables qui serviront à l'apprentissage du modèle. En effet, le jeu de données contient nombre de variables qui ne sont pas parlantes cliniquement pour l'établissement d'un diagnostic de parodontite : rappelons que c'est un jeu de données destiné à la recherche. Un premier pré-tri manuel permet donc de sélectionner uniquement les variables parlantes d'un point de vue clinique ainsi que quelques variables qui ne sont pas parlantes pour voir comment les praticiens réagissent face à ces variables. Ensuite on réalise un premier entraînement du modèle et on analyse quelles sont les variables les plus utiles et les plus robustes [10] pour la prédiction de la récession pour le modèle, celles qui vont lui permettre d'obtenir les meilleures performances. Enfin, on réalise une dernière sélection manuelle parmi ces variables les plus utiles, en rajoutant une variable non-parlante pour arriver à une sélection finale de 6 variables.

Une fois tout cela fait, il ne reste plus qu'à séparer le jeu de données en plusieurs jeux comme nous l'avons vu plus tôt : jeu d'entraînement, jeu de validation, jeu de test.

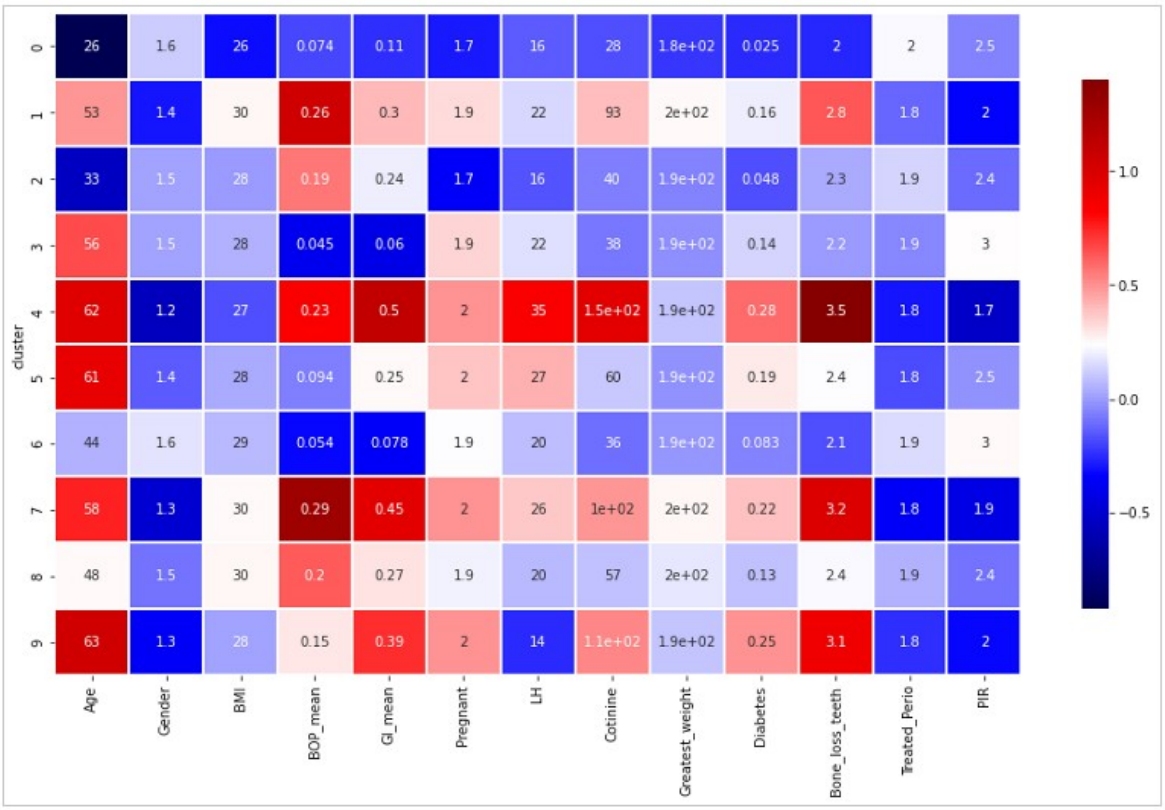
Pour en revenir à notre jeu de données issu de l'étude NHANES, il contient des données cliniques du parodonte (perte d'attache, profondeur de poche, récession, saignement au sondage), ainsi que des données biologiques (bilan biologique avec numération de formule sanguine, VS, CRP …) et socio-démographiques (âge, sexe, hygiène alimentaire, hygiène orale, revenus …). Suivant le nombre et le type de variables choisies le nombre de sujets va varier. Celui-ci est en moyenne de 22.000 sujets. Ci-dessous suivent trois graphiques concernant la répartition de ce jeu de données en fonctions de diverses variables :



Clusterisation des données parodontales des sujets du jeu de données.



Caractérisation parodontale de ces différents clusters.



Caractérisation des différents clusters selon différentes variables cliniques, médicales et socio-démographiques.

4. Modèle de ML et technique d'explication (SHAP)

Pour des raisons de performance et de nécessité d'explication post-hoc, nous avons choisi d'utiliser un modèle de forêt d'arbres boosté [82]. Après apprentissage, il sera nécessaire de présenter les métriques de performance sur le jeu de données d'apprentissage/validation et de test (accuracy, recall, matrice de confusion…).

La technique de XAI ensuite utilisée sera SHAP, que nous avons vu plus tôt [82]. Dans le graphique suivant, nous pouvons voir un exemple de variables classées par ordre d'importance (la variable la plus en haut étant la plus importante) dans la prédiction d'une parodontite. Chaque point représente un patient, sa couleur représente sa valeur pour la variable indiquée (bleu : faible et rouge : élevée) tandis que son placement sur l'axe des abscisses représente sa valeur de SHAP. On peut ainsi voir que pour ce modèle et ce jeu de données, plus l'âge augmente, plus un patient est à risque de présenter une parodontite tandis que l'absence de pathologies associées a plutôt tendance à être un facteur protecteur de la parodontite (SHAP-value négative pour la majorité des points bleus de la variable « pathologies »). Cela permet notamment de vérifier (ou non) que les variables sur lesquelles se base en priorité le modèle pour faire sa prédiction correspondent aux données acquises de la science et à notre sens clinique.

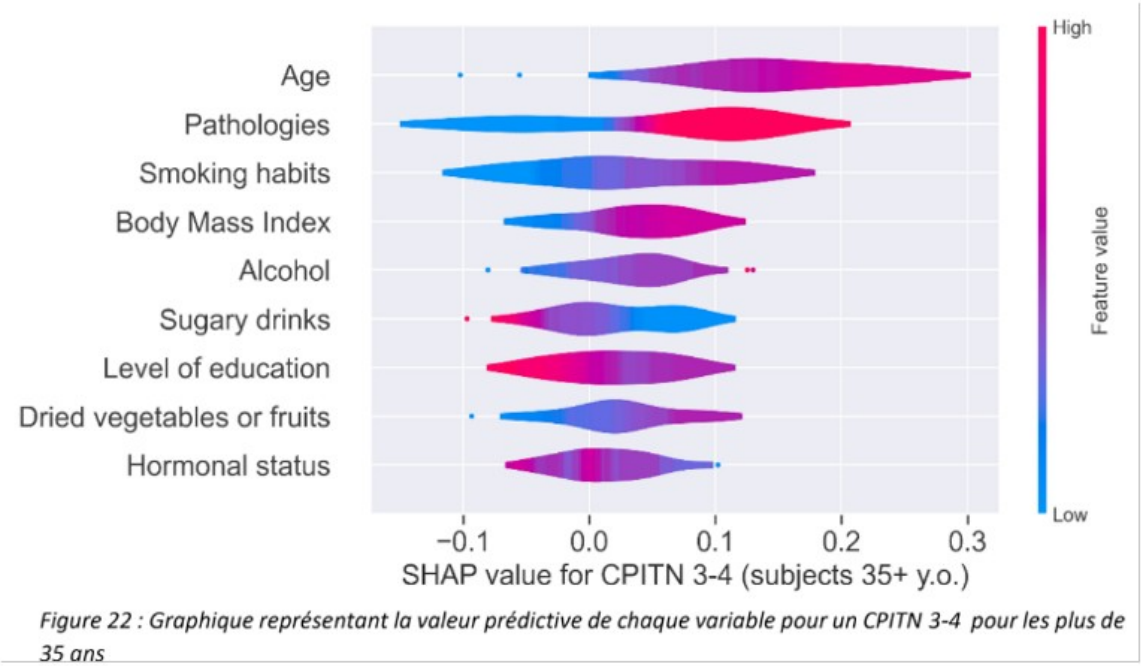
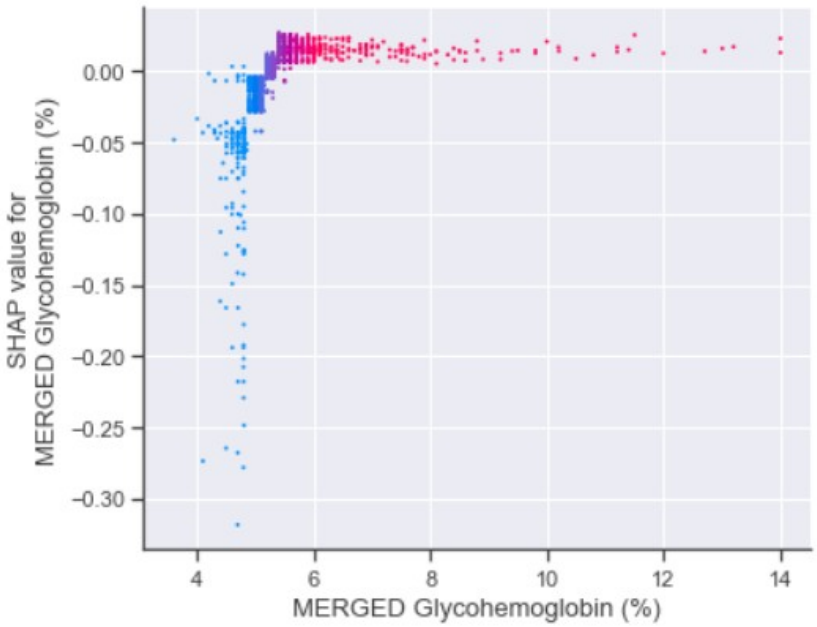


Figure 22 : Graphique représentant la valeur prédictive de chaque variable pour un CPITN 3-4 pour les plus de 35 ans

Exemple de graphique issu de l'utilisation de la technique de XAI SHAP.

Le graphique suivant, quant à lui, illustre la répartition des patients pour une variable (ici : le pourcentage d'hémoglobine glycquée). Chaque point représente toujours un patient, et sa couleur représente toujours sa valeur pour la variable. On observe ainsi une répartition très nette de l'influence de la variable pour la prédiction du modèle : un taux d'hémoglobine glycquée inférieur à environ 5,5 % est un facteur protecteur dans la prédiction (SHAP-value inférieure à zéro pour pratiquement tous les patients correspondants) au contraire d'un taux supérieur à environ 5,5 % (SHAP-value supérieure à zéro pour pratiquement tous les patients correspondants).



Graphique illustrant la répartition d'une variable individuelle après l'utilisation de la technique de XAI SHAP.

5. Sélection des cas

Afin de maximiser à la fois le nombre de réponses et leur qualité, le nombre de cas nécessaire doit être calculé en fonction du nombre de professionnels recrutés et du nombre souhaité de visionnage de chaque cas par l'ensemble des utilisateurs.

- en se basant sur 100 utilisateurs, pour que chaque patient soit vu par 9 praticiens et que chaque praticien voit 9 patients, alors le nombre de cas à sélectionner est : 100
- choisir des cas divers qui représente l'ensemble des comportements du jeu de données : des patients prototypes et des patients avec des comportements particuliers ou incohérent pour étudier comment les utilisateurs réagissent.

6. Déroulé de l'expérimentation

L'objectif est donc de recruter 100 experts et de tirer de manière semi-aléatoire (cf ci-dessus : tirage aléatoire + cas particuliers) 100 patients du jeu de données. Chaque expert se verra attribuer aléatoirement 9 patients propres (3 du groupe A, 3 du B et 3 du C – voir plus bas), anonymisés, qu'il verra durant chacune des trois phases d'expérimentation. Chaque patient sera donc ainsi suivi par 9 praticiens.

Chaque présentation de cas (un cas = un patient) sera subdivisé en trois modes :

– présentation du cas avec les données seules (mode 1) ;

– présentation du cas avec les données et le résultat du ML (mode 2) ;

– présentation du cas avec les données, le résultat du ML et les explications associées (mode 3).

Les cas seront divisés aléatoirement en trois groupes (que nous appelleront groupe A, B et C) et les expérimentations en 3 phases. Durant la première phase, les cas du groupe A seront présentés selon le mode 1, ceux du groupe B selon le mode 2 et ceux du C selon le mode 3. Durant la deuxième phase, les cas du groupe A seront présentés selon le mode 2, les cas du groupe B selon le mode 3 et les cas du groupe C selon le mode 1. Durant la troisième phase, chaque groupe sera présenté selon le mode sous lequel il n'a pas encore été présenté.

Ainsi, durant chaque phase, chaque praticien verra 3 cas selon le mode 1, 3 selon le mode 2 et 3 selon le mode 3. Il reverra ses mêmes 9 cas à chaque phase, chacun présenté sous un mode différent par rapport aux autres phases. À la fin des trois phases, chaque praticien aura vu chaque patient selon chaque mode avec un délai entre chaque phase suffisant pour que les données d'un mode n'influencent pas les autres modes.

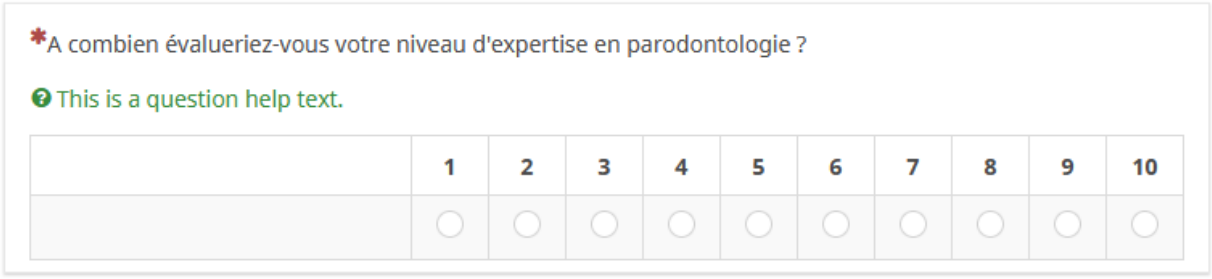
Les phases seront espacées de quelques semaines, et ce pour éviter que les praticiens ne reconnaissent leurs patients grâce à leurs données et ne soient influencés pour la pose de leur diagnostic par un diagnostic précédemment posé lors d'une autre phase.

De plus, l'évaluation de la douleur à travers une échelle de valeur analogique montre de très bons résultats en clinique pour évaluer objectivement quelque chose de subjectif.

Reprenant ce principe, à chaque fois qu'une donnée subjective sera demandée au praticien, elle sera demandée sous la forme d'une échelle de 1 à 10, où 1 représente une faible valeur concernant le ressenti demandé (par exemple : la confiance dans son propre diagnostic) et 10 représente une forte valeur concernant ce même ressenti.

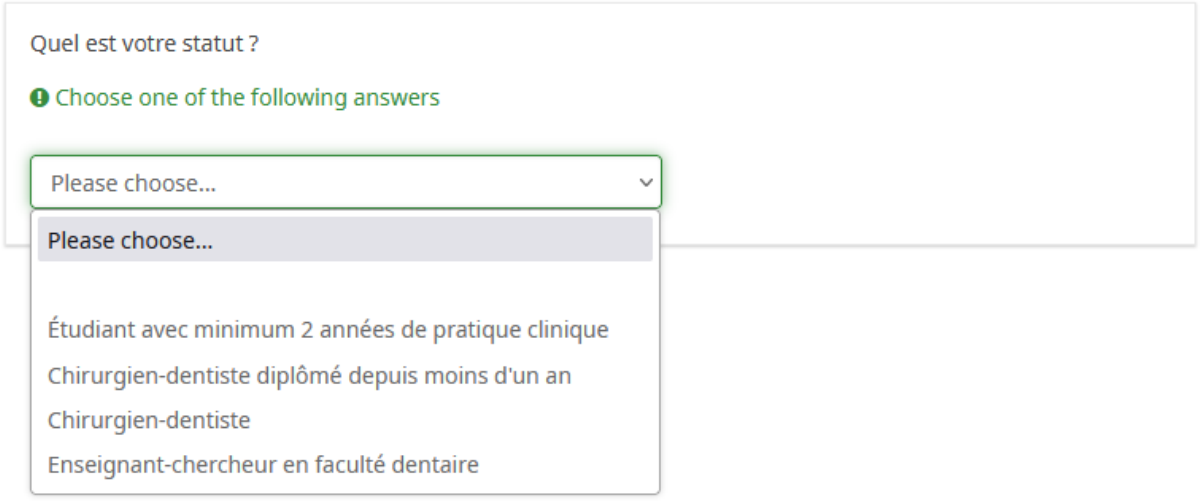
Lors de chaque phase expérimentale, les praticiens se verront questionnés de la sorte :

1/ Auto-évaluation du niveau d'expertise du professionnel de santé en parodontologie sur une échelle de 1 à 10



Visuel de la première question conçue sur la plateforme LimeSurvey

2/ Question sur la situation du professionnel de santé (dentiste libéral, étudiant en 6e année, dentiste tout juste diplômé, enseignant-chercheur, autre)

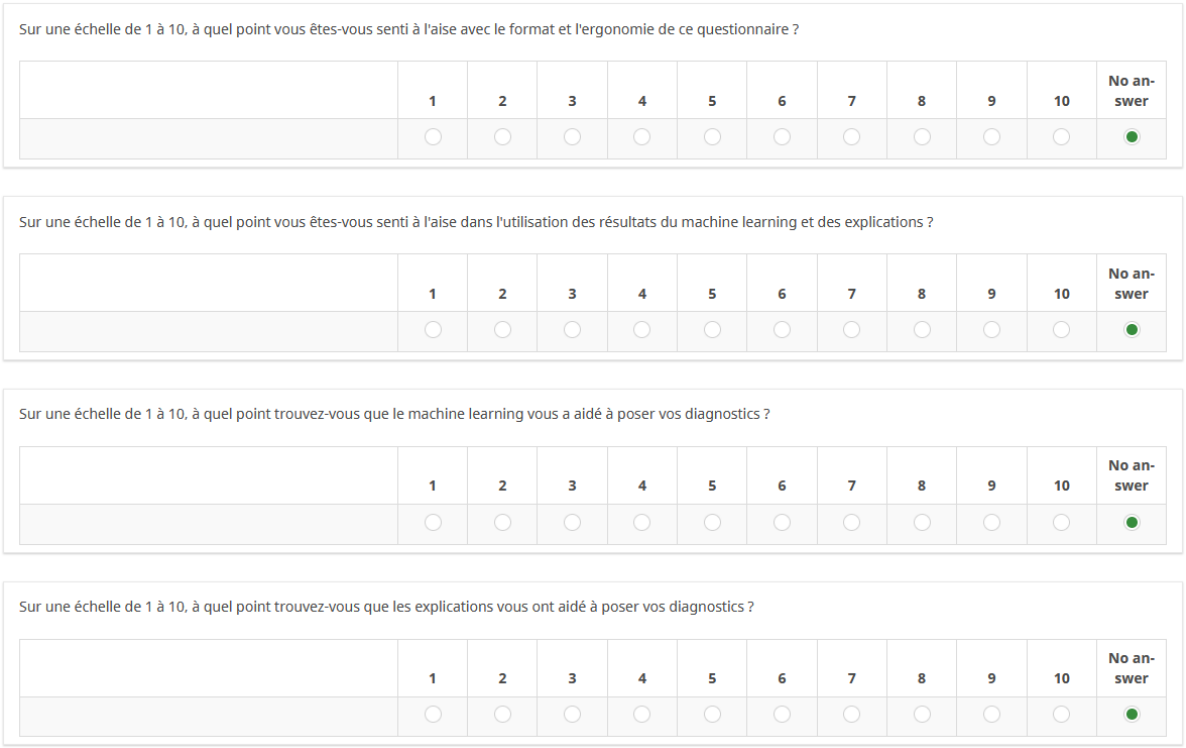


Visuel de la seconde question conçue sur la plateforme LimeSurvey

3/ Un cas d'entraînement selon le mode 1, un cas d'entraînement selon le mode 2 et un cas d'entraînement selon le mode 3. Nous aborderons en détail le contenu de ces modes ci-dessous.

4/ Présentation de leurs 9 patients (3 selon le mode 1, 3 selon le mode 2 et 3 selon le mode 3) ordonnés aléatoirement, les uns à la suite des autres.

5/ Questions sur le ressenti par rapport au questionnaire et à l'apport de la XAI



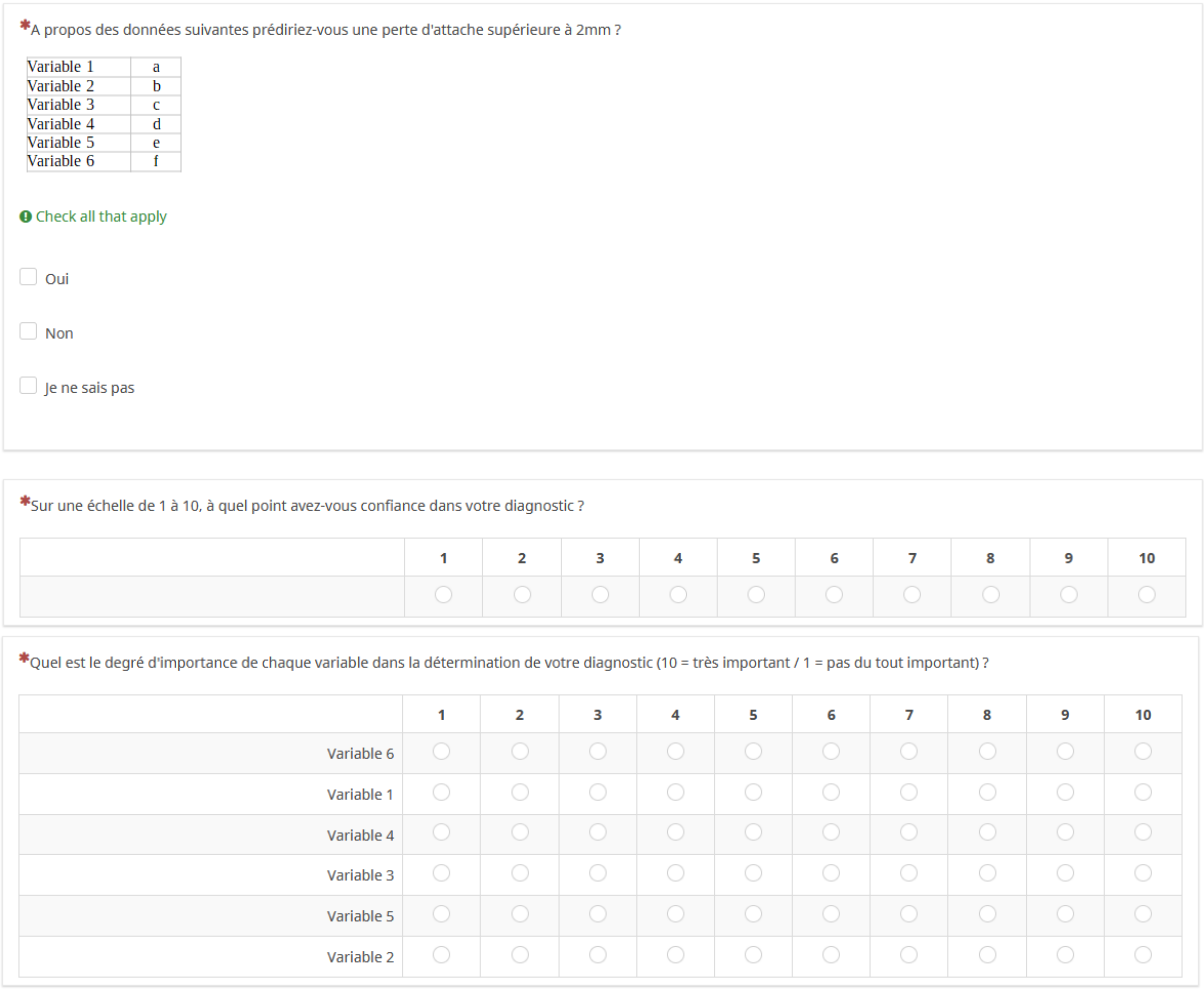
Visuel des questions finales sur le ressenti, conçues sur la plateforme LimeSurvey

Contenu détaillé du mode 1 :

1/ On propose tout d'abord au praticien les variables concernant un patient et on lui demande, à partir de ces variables, de prédire ou non une perte d'attache supérieure à 2 mm.

2/ On demande ensuite au praticien le niveau de confiance qu'il place dans son diagnostic.

3/ Enfin, on demande au praticien d'attribuer à chaque variable un score allant de 1 à 10 en fonction du degré d'utilité qu'a eu cette variable pour poser son diagnostic selon lui. Cela permettra ensuite d'établir un ordre d'importance des variables pour chaque diagnostic comparable à l'ordre d'importance des variables fournit par SHAP.



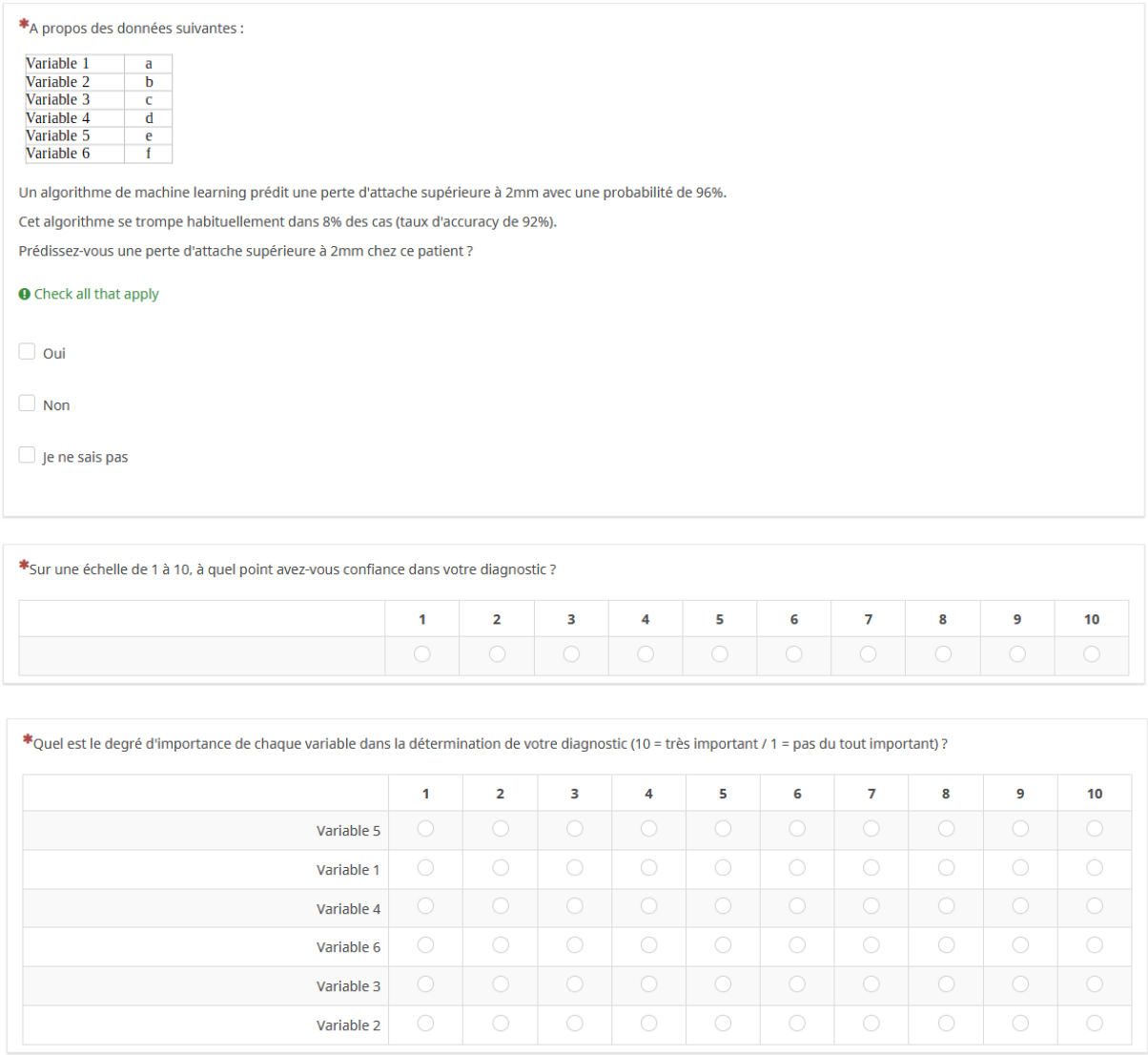
Visuel des questions du mode 1, conçues sur la plateforme LimeSurvey

Contenu détaillé du mode 2 :

1/ On propose également au praticien les variables concernant un patient, cette fois-ci accompagnée de la prédiction fournie par le machine learning pour ce patient et du taux d'accuracy de l'algorithme et on lui demande, à partir de ces informations, de prédire ou non une perte d'attache supérieure à 2 mm.

2/ On demande ensuite au praticien le niveau de confiance qu'il place dans son diagnostic, comme pour le mode 1.

3/ Enfin, on demande au praticien d'attribuer à chaque variable un score allant de 1 à 10 en fonction du degré d'utilité qu'a eu cette variable pour poser son diagnostic selon lui, comme pour le mode 1.



Visuel des questions du mode 2, conçues sur la plateforme LimeSurvey

Contenu détaillé du mode 3 :

1/ On propose au praticien les mêmes informations que dans le mode 2 et on y ajoute cette fois une explication globale du modèle ainsi qu'une explication locale à propos du patient et on lui demande, à partir de ces informations, de prédire ou non une perte d'attache supérieure à 2 mm.

2/ On demande ensuite au praticien le niveau de confiance qu'il place dans son diagnostic, comme pour les modes 1 et 2.

3/ Enfin, on demande au praticien d'attribuer à chaque variable un score allant de 1 à 10 en fonction du degré d'utilité qu'a eu cette variable pour poser son diagnostic selon lui, comme pour les modes 1 et 2.

*Sur une échelle de 1 à 10, à quel point avez-vous confiance dans votre diagnostic ?

	1	2	3	4	5	6	7	8	9	10
	○	○	○	○	○	○	○	○	○	○

*Quel est le degré d'importance de chaque variable dans la détermination de votre diagnostic (10 = très important / 1 = pas du tout important) ?

	1	2	3	4	5	6	7	8	9	10
Variable 2	○	○	○	○	○	○	○	○	○	○
Variable 5	○	○	○	○	○	○	○	○	○	○
Variable 3	○	○	○	○	○	○	○	○	○	○
Variable 4	○	○	○	○	○	○	○	○	○	○
Variable 1	○	○	○	○	○	○	○	○	○	○
Variable 6	○	○	○	○	○	○	○	○	○	○

Visuel des questions du mode 3, conçues sur la plateforme LimeSurvey

7. Métriques d’évaluation des résultats

Partant de notre objectif principal et de nos objectifs secondaires ainsi que de la la conception du présent protocole, nous souhaitons évaluer principalement la précision, la vitesse de réponse ainsi que la confiance associée et les sentiments des praticiens à l’égard des explications.

Pour notre objectif principal, nous comparerons donc pour chaque mode le rapport (précision)/(temps de réponse) des praticiens dans leur prédiction de perte d’attache. Nous évaluerons comment ce ratio moyen évolue en fonction du mode et, si il y a des différences, si elles sont significatives. Nous évaluerons également ce ratio au sein des différents sous-groupes d’utilisateurs répartis en fonction de leur niveau d’expertise.

De plus, les comparaisons de performances entre chaque mode peuvent être réalisées selon plusieurs axes :

- Comparaison axée sur le patient : pour chaque patient indépendamment, comment les performances des utilisateurs varient-elles ?
- Comparaison axée sur l’utilisateur : pour chaque utilisateur, comment varient leurs performances entre les modes ?
- Comparaison axée sur l’expertise : pour chaque groupe d’expertise, comment l’expertise impacte-t-elle les performances ?
- Comparaison axée sur le mode : pour chaque mode, comment les performances varient-elles entre les utilisateurs ?
- Comparaison axée sur la classification : comment varient les performances des utilisateurs entre chaque mode en fonction de la justesse de la prédiction des utilisateurs avec les données seules ?

Pour chaque axe, des statistiques sur le ratio (précision)/(temps de réponse) moyen seront calculées, ainsi que des valeurs p lorsque cela est pertinent. Ces valeurs p seront utilisées pour valider ou rejeter H0 et H1 en se basant sur le seuil alpha que nous définirons à 0,05. Nous calculerons également combien de fois les explications améliorent les performances des utilisateurs et leur proportion parmi tous les résultats.

Nous calculerons également les statistiques sur la confiance accordée par les praticiens dans leur diagnostic de la même manière, en étudiant les différences entre les différents

groupes d’expertise et les différents modes. Nous relierons ensuite ces résultats à ceux concernant le ratio (précision)/(temps de réponse) afin d’étudier un possible lien entre les deux.

Enfin, pour le classement des variables en fonction de leur utilité pour l’utilisateur, nous comparerons d’abord comment les réponses des utilisateurs diffèrent pour chaque mode pour le même patient. Nous voulons évaluer dans quelle mesure les utilisateurs changent leur classement pour le même patient en fonction des différents modes, en particulier entre les modes 1 et 2, étant donné qu’aucune information supplémentaire sur l’importance des variables n’est fournie aux praticiens. Ainsi, en se basant sur les classements pour les modes 1 et 2, nous pourrons mieux évaluer les différences avec le classement du mode 3 et l’impact des explications, c’est-à-dire l’importance locale des variables. Pour cette évaluation, nous utiliserons la distance de Kendall-Tau pour comparer les listes classées [80]. Pour chaque patient et chaque utilisateur, nous comparerons d’abord les classements des modes 1 et 2 pour évaluer l’inter-variabilité, puis comparerons le classement du mode 3 aux deux premiers modes et aux influences pour évaluer l’impact des explications. Enfin, nous comparerons les résultats en fonction du niveau d’expertise des utilisateurs.

III. Discussion autour de la XAI

Bien que nous n'ayons pas pu, à ce jour, mettre en application ce protocole afin d'évaluer l'intérêt de l'explicabilité pour le chirurgien-dentiste, il existe dans la littérature nombre de travaux qui abordent cet intérêt de différents points de vue. Au cours de cette partie, nous parlerons donc de quelques-uns de ces points de vue : quelles sont les attentes qui pèsent sur la XAI et l'IA de la part des personnels soignants, d'un point de vue juridique et éthique, et quels sont leurs intérêts pratiques pour les praticiens.

1. Quelles sont les attentes vis-à-vis de l'IA et de la XAI ?

a. Les attentes des praticiens

Afin de concevoir un nouvel outil, il est primordial de se renseigner auprès des futurs utilisateurs de cet outil : quels sont leurs besoins, leurs attentes, etc.

Tout d'abord, voyons donc voir ce qu'il en est déjà de l'utilisation actuelle des premiers systèmes à base d'IA mis sur le marché. Becker & al[26] ont réalisé une étude auprès des membres de la Société Européenne de Radiologie (ou ESR pour European Society of Radiology). Les radiologues ayant répondu à l'étude (690 en tout) provenaient de 44 pays différents et représentaient 2,5 % de l'ensemble de ses membres. Seulement, uniquement 276 d'entre eux avait une expérience clinique pratique avec de l'IA, ce qui représente donc 1 % de l'ESR et le cœur de cette étude. Les auteurs admettent eux-mêmes que ces membres ne sont pas représentatifs des radiologues en Europe car seuls ceux qui s'intéressaient un minimum à l'IA ont répondu à leur questionnaire. De plus, les auteurs ont également évoqué deux autres limites à leur enquête, que nous garderons en tête : ils jugent que certaines de leurs questions mériteraient une analyse approfondie là où leur enquête ne permet qu'une vision générale, et l'enquête a été réalisée en 2022, aussi tiennent-ils à faire remarquer que dans ce domaine en évolution rapide qu'est l'IA, les résultats et les opinions peuvent évoluer tout aussi rapidement. Cette enquête est d'intérêt pour le chirurgien-dentiste car nous sommes confrontés à de forts volumes quotidiens de données radiologiques à interpréter avec la mobilisation de compétences spécifiques à la radiologie, communes entre les deux professions.

Voici donc, point par point, ce qui ressort de cette enquête.

- A propos de la difficulté d'intégration des outils basés sur l'IA : Globalement, les radiologues ont répondu par la négative mais on observe tout de même un fort taux d'absence de réponse.

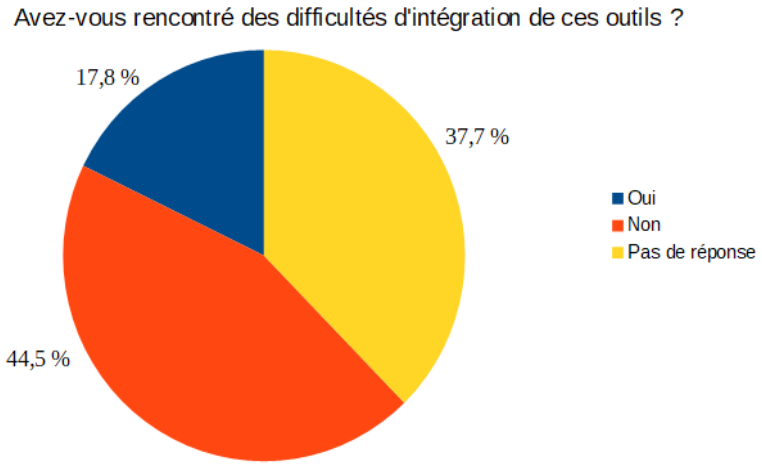


Diagramme illustrant les réponses à la question portant sur les difficultés d'intégration de l'IA

- A propos de la fiabilité : La majorité des participants ont estimé que les résultats fournis par les algorithmes étaient fiables, mais pour autant ils ne placeraient pas leur confiance en eux pour une utilisation complètement autonome. De plus, un grand nombre d'entre eux mettent en place des mécanismes d'assurance qualité pour évaluer les performances diagnostiques des algorithmes, afin de s'assurer de leur fiabilité.

- A propos de la charge de travail : Seule une minorité des praticiens interrogés ont constaté une réduction de leur charge de travail, tandis que la plupart n'ont pas remarqué d'effet significatif. Lors d'une précédente enquête menée par l'European Society of Radiology en 2018[31], les praticiens avaient exprimés des attentes à propos de la réduction de la charge de travail qui ne sont donc pas atteintes ici. Une analyse récente a révélé que malgré la présence de bénéfices pour les patients, la charge de travail des praticiens n'a guère diminué (seulement pour 4 % des praticiens interrogés) et a même augmenté dans la moitié des cas[32]. Dans l'enquête de 2022, des progrès sont cependant notés, avec une réduction de la charge de travail observée dans environ 22 % des cas, mais cela demeure tout de même relativement faible.

- A propos de l'intention d'acquérir un logiciel basé sur l'IA à l'avenir : Cette question a été posée à l'ensemble des 690 répondants, et non uniquement à ceux ayant déjà

une expérience avec l'IA, et la majorité d'entre eux ont répondu par la négative (avec 52,6 % de « non », 13,3 % de « oui », et 34,1 % sans réponse). Les raisons qui sont avancées pour les réponses négatives portent sur des doutes à propos de la valeur ajoutée, les performances annoncées et des inquiétudes concernant une charge de travail supplémentaire, comme on peut le voir dans le graphique suivant tiré de l'article :

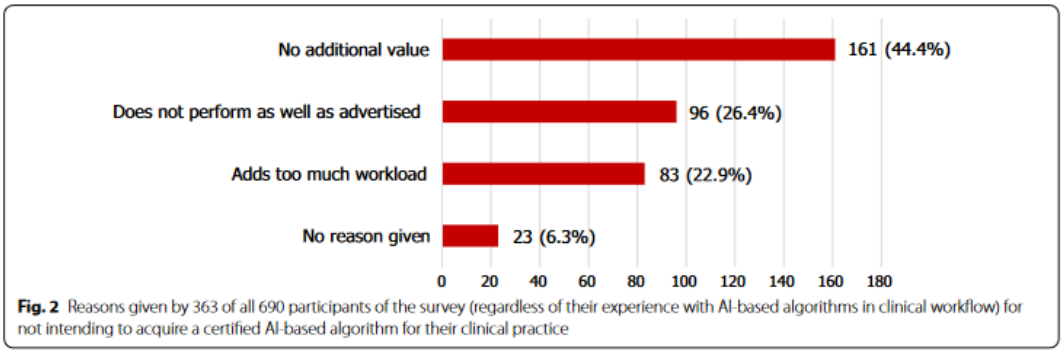


Fig. 2 Reasons given by 363 of all 690 participants of the survey (regardless of their experience with AI-based algorithms in clinical workflow) for not intending to acquire a certified AI-based algorithm for their clinical practice

Diagramme issu de l'article de Becker & al[26] portant sur les raisons des praticiens les poussant à ne pas souhaiter acquérir d'outils basés sur l'IA

Dans une autre enquête menée récemment auprès des membres de l'American College of Radiology (ACR)[33], il a été estimé que près de 30 % des radiologues états-uniens utilisaient déjà des systèmes à base d'IA. Ce qu'il en est ressorti, c'est que deux freins principaux s'opposent à la diffusion plus généralisée de ces outils : les radiologues américains se préoccupent du fait que des performances incohérentes puissent être observées et que la productivité puisse potentiellement se trouver réduite de par l'utilisation de ces systèmes.

Les résultats des deux enquêtes sont difficilement comparables. Néanmoins, elles convergent toutes deux pour conclure que, par rapport aux prévisions et aux attentes initiales, l'impact global des algorithmes basés sur l'IA sur la pratique radiologique actuelle demeure modeste.

Dans un autre article de Tonekaboni & al[27] on peut trouver ce tableau qui nous montre qu'en 2019 nombre de techniques d'explicabilité manquaient encore d'évaluations : cela souligne le fait que c'est un domaine encore très récent.

Table 1: Summary of Explainable ML methods Contextualized for Clinical Applicability

Explanation Class	Representative Existing Methods	Possible shortcomings for clinical settings
Feature Importance	Sensitivity Analysis (Saltelli *et al.*, 2008), LRP (Bach *et al.*, 2015) (Yang *et al.*, 2018) LIME, Anchors, Shapley Values (Ribeiro *et al.*, 2016, 2018; Lundberg *et al.*, 2018)	Complex correlation between features of clinical models can be a challenge Further evaluation for consistency required
Instance Level Explanation	Influence functions (Koh and Liang, 2017) Prototypes and Criticisms (Kim *et al.*, 2016)	Not evaluated on complex clinical models Limited applicability
Uncertainty	Distributional shift (Subbaswamy and Saria, 2018) Parameter uncertainty (Gal and Ghahramani, 2016; Schulam and Saria, 2019)	Not evaluated on complex clinical models
Temporal Explanations	RETAIN, RAIM (Xu *et al.*, 2018; Choi *et al.*, 2016)	Potential lack of consistency due to the attention mechanism (Jain and Wallace, 2019)
Transparent Design	Rule Based Methods (Lakkaraju *et al.*, 2016; Wang and Rudin, 2015)	Less powerful in modeling more complex applications; Generally assume a trade off of accuracy and explainability

Tableau issu de l'article de Tonekaboni & al[27] à propos de la présence et de la qualité des évaluations pour différentes méthodes d'explicabilité en 2019

Nous allons maintenant nous intéresser, non pas à des retours d'expérience, mais à ce que certains professionnels de santé pourraient attendre, pour le futur, de l'intégration à leur quotidien d'outils basés sur l'IA et la XAI. De fait, les auteurs ont entrepris une étude basée sur des entretiens qualitatifs dans le but de comprendre les attentes des praticiens à l'égard de la XAI. Ils ont estimé que les sondages via des questionnaires étaient insuffisants et trop fermés pour appréhender le sujet dans toute sa complexité, ce qui était une limite évoquée dans le premier article. Dans cette étude, il nous est dit que la saturation a été atteinte après environ 10 entretiens, conformément aux recommandations. Cette saturation correspond à un point où aucune nouvelle information pertinente n'est apportée au cours de ces entretiens [71, 72].

Ainsi, l'étude repose sur 10 entretiens avec 10 cliniciens provenant de services de soins intensifs et de services d'urgences, qui ont des niveaux d'expérience variés. Ces services

ont été choisis parce qu'ils sont considérés comme des domaines où l'IA peut être utilisée en soutien pour améliorer les soins, et parce qu'ils ont déjà une expérience pratique avec des systèmes d'alerte précoce ou des outils d'aide à la décision. Pour leur étude, les auteurs définissent également « l'explicabilité dans le ML pour les soins de santé » comme un ensemble d'attributs mesurables, quantifiables et transférables associés à un système de ML destiné aux cliniciens pour calibrer la confiance dans le modèle.

Voici ce qu'il en ressort :

- L'explicabilité est considérée comme un moyen de justifier les décisions cliniques aux yeux des patients et des collègues notamment. En effet, les cliniciens ressentent le besoin de comprendre les caractéristiques du modèle qui sont cliniquement pertinentes et en accord avec la médecine basée sur les preuves (Evidence-Based Medicine, EBM) actuelle. Comme l'a expliqué un clinicien senior en urgence : "Si un écart se manifeste entre le protocole clinique actuel et les prédictions du modèle, sans que nous en comprenions la raison, cela suscite de l'inquiétude." Une fois que le modèle a émis une prédiction, il est essentiel de pouvoir la comprendre et la rationaliser. Cela permet aux praticiens de comparer les résultats du modèle à leur propre jugement clinique, surtout lorsqu'ils ne concordent pas.

- Les praticiens veulent avoir à disposition des informations sur les domaines où le système est compétent (où il a de fortes probabilités de réussite) et ceux où il ne l'est pas. Avec cela, même un système dont la précision est considérée comme insuffisante peut être considéré comme acceptable par les praticiens interrogés si les raisons de ses sous-performances sont clairement expliquées.

- Les attributs utilisés par le système pour prendre des décisions doivent être clairement définis. Cela permet d'instaurer une certaine confiance de la part des praticiens et de garantir que le système est utilisé de manière adéquate avec les paramètres appropriés sur la population appropriée. Ce niveau de transparence est également considéré comme essentiel à communiquer par Mitchell & al [73].

- Ces deux aspects peuvent être résumés par la notion de **« savoir quand cela fonctionne et quand cela atteint ses limites. »** De plus, même si la fiabilité, la spécificité et la sensibilité d'un système sont satisfaisantes lors de sa conception, ce sont ses

performances dans le monde réel, y compris son apprentissage continu après sa mise en service, qui revêtent la plus grande importance.

- La présentation des prédictions doit être conçue sous une forme visuelle favorisant une compréhension rapide et claire. Ceci est particulièrement important lorsque les cliniciens doivent gérer plusieurs systèmes de manière concomitante. Il est préférable d'éviter les informations redondantes entre les différents flux d'informations accessibles aux praticiens, sauf si ces informations nécessitent d'être mises en évidence pour susciter une action potentielle. Dans des exercices où le temps est une ressource précieuse, comme dans un service d'urgences, il est essentiel que l'attention des praticiens soit attirée unique sur les informations pertinentes et utiles, sans introduire de surcharge cognitive inutile. Cela implique que les informations fournies par le système soient filtrées : les explications qui n'ont pas d'impact direct sur le flux de travail sont de moindre importance. Dans ce même esprit, l'explication doit être concise et donnée au moment opportun.

- Pour continuer sur ce concept de dispensation uniquement des informations utiles, toujours en raison de la nécessité de gérer simultanément plusieurs systèmes, le système utilisé doit pouvoir anticiper un changement de l'état du patient qui soit significatif et qui corresponde à la prédiction avec, surtout, **la possibilité d’exploiter cette prédiction** (en proposant par exemple des interventions potentielles ou une collecte de données supplémentaire). Un score de certitude peut être considéré par les praticiens interrogés comme une forme d'explication qui accompagne la prédiction et sert de seuil pour déclencher une alerte uniquement lorsque le système est vraiment certain de sa prédiction. D'autres études ont révélé que de nombreuses alertes ne sont pas suivies de changements cliniquement exploitables, ce qui compromet leur acceptation par les cliniciens [74, 75]. Dans le même registre, la "fatigue liée aux alarmes ou aux clics" est déjà une préoccupation de premier plan signalée par de nombreux cliniciens dans un article d’Embi et Leonard en 2012 [76].

- Les praticiens souhaitent également comprendre quels changements dans l’état d’un patient ont conduit à une prédiction spécifique de la part d’un système basé sur l’IA. De fait, les explications temporelles sur les trajectoires des patients sont jugées importantes pour pouvoir comprendre les prédictions suivant l’évolution de l’état des patients.

- Quand le temps est limité, les praticiens ne trouvent pas l'approche locale de la XAI très attrayante. Ils considèrent que cela correspond à de la recherche de patients similaires [77, 78, 79] et soulignent que, même pour des résultats similaires, les trajectoires cliniques des patients peuvent différer considérablement pour arriver à ces résultats, et que cela ne peut donc être utile que dans des cas spécifiques. Par exemple, parmi ces cas spécifiques, certains cliniciens souhaitent connaître des cas similaires à ceux de leurs patients afin de comprendre quelles mesures ont été prises dans ces situations et quels ont été les résultats de ces mesures.

Plus personnellement, j'ajouterais que j'accorderais tout de même de l'importance à avoir une explication locale pour un patient, car cela me permettrait de vérifier que l'algorithme a fonctionné correctement et que ce sont bien les variables attendues qui ont eu le plus d'impact sur la décision. Ceci afin d'éviter les situations où une variable aléatoire ou inappropriée aurait pu influencer de manière prédominante la décision.

L'article conclue sur le fait que les concepteurs de ces systèmes et leurs utilisateurs (les praticiens) peuvent avoir des perspectives différentes vis-à-vis de ces systèmes et qu'il est donc essentiel de tenir compte de leur point de vue lors de leur développement. Cela permet d'augmenter les chances que les produits issus de ce développement soient réellement utilisés par les praticiens et non mis au rebut par manque de confiance (surtout dans les cas où le risque associé à un mauvais diagnostic est haut, comme dans les soins de santé) ou de praticité, d'applicabilité dans un contexte de soins réel. Cela rejoint le concept d'« humain dans la boucle » que nous reverrons par la suite, et ce dès la conception des systèmes à base d'IA.

Ainsi, malgré une sorte d'explosion démographique de l'IA dans les produits de santé ces dernières années (un genre de « bAIby boom »), son impact réel dans la pratique est encore en dessous des espérances. Il en ressort qu'il existe encore de nombreux besoins des professionnels de santé à satisfaire pour que de plus en plus d'IA puissent se voir utiliser concrètement. Ainsi les praticiens recherchent, notamment et en plus d'une amélioration globale de la qualité des soins dispensés, de la transparence et de la confiance dans les outils qu'ils utilisent, la XAI étant un outil de choix pour y parvenir, une réduction de leur charge de travail ou au moins une absence d'augmentation de celle-ci et une ergonomie de ces outils pour pouvoir être utilisables en pratique réelle.

b. Les attentes éthiques et légales

Après avoir examiné les attentes des professionnels de santé envers l'IA, nous pouvons désormais aborder les considérations éthiques et législatives de notre société concernant l'utilisation de l'IA dans le domaine médical. D'un point de vue plus large et plus ancien, en 1942 Isaac Asimov énonce les trois lois de la robotique qui s'appliquent aux machines capables de penser et d'interagir avec les humains (à ce moment, le concept d'IA n'était pas encore né). Ces lois sont les suivantes :

I – Un robot ne doit pas porter atteinte à un être humain ou, par inaction, le laisser exposé à un danger.

II – Un robot doit obéir aux ordres d'un être humain à moins que cela ne contrevienne à la première règle.

III – Un robot doit protéger sa propre existence à moins que cela ne contrevienne avec les deux premières règles.

Ces règles peuvent donc s'appliquer à l'IA, qu'elle soit physique comme un robot ou virtuelle, à travers des auteurs qui déclarent notamment que « le développement de l'IA devrait garantir que ces technologies intelligentes ne nuisent pas aux humains ni au statut moral des machines elles-mêmes » [2, 68]. Bien que cela constitue déjà un socle de réflexion pour l'éthique de l'IA, d'autres considérations sont à prendre en compte comme celles portant sur la confidentialité des données avec la création, le partage et l'usage de jeux de données médicaux [2, 25, 69], la création et le renforcement d'inégalités pour l'accès aux soins médicaux[2, 70], le consentement éclairé du patient quand une boîte noire a participé à une prise de décision [23, 25], l'absence de biais discriminatoires dans le processus de décision d'une IA [25], la responsabilité légale d'un dommage causé par une IA [25], le degré d'autonomie alloué à l'IA [24, 25], etc. C'est un sujet que je trouve passionnant, mais nous nous contenterons dans cette sous-partie d'aborder brièvement quelques-unes de ces problématiques et de voir en quoi la XAI peut être une aide à la résolution de ces problématiques. Il faut également souligner le fait que l'application concrète de l'IA dans le domaine de la santé est relativement récente et en pleine explosion, débordant le cadre juridique actuel qui tente donc de s'adapter. Mais pour le moment, il est encore ambigu et

flou et est susceptible d'évoluer fortement dans un avenir proche pour encadrer ces problématiques.

Dans leur article, Muller et al. [23] expriment dix règles qui, d'une certaine manière, sont une version évoluée des trois règles d'Asimov pour prendre en compte les problématiques actuelles et sont spécialisées sur une approche éthique de l'IA en médecine :

1. Il doit être reconnaissable quelle partie d'une décision ou action est prise et exécutée par l'IA.

2. Il doit être reconnaissable quelle partie de la communication est effectuée par un agent IA.

3. La responsabilité d'une décision, action ou processus de communication par IA doit incomber à une personne physique ou juridique compétente.

4. Les décisions, actions et processus de communication par IA doivent être transparents et explicables.

5. Une décision par IA doit être compréhensible et reproductible.

6. L'explication d'une décision par IA doit reposer sur les dernières données avérées de la science.

7. Une décision, action ou communication par IA ne doit pas être manipulatrice en prétendant à une précision qu'elle n'a pas.

8. Une décision, action ou communication par IA ne doit pas enfreindre la loi applicable et ne doit pas causer de préjudice humain.

9. Une décision, action ou communication par IA ne doit pas être discriminatoire, en particulier dans la formation des algorithmes.

10. La définition d'objectifs, le contrôle et la surveillance des décisions, actions et communications par IA ne doivent pas être effectués par des algorithmes.

Nous restreindrons donc ici notre discussion aux principes 1, 3, 4, 8, 9 et 10 en se basant essentiellement sur trois articles : celui de Muller et al. [23], celui de Mezrich et al. [24] et enfin celui de Schneeberger et al. [25].

Principe N°1 : Il doit être reconnaissable quelle partie d'une décision ou action est prise et exécutée par l'IA.

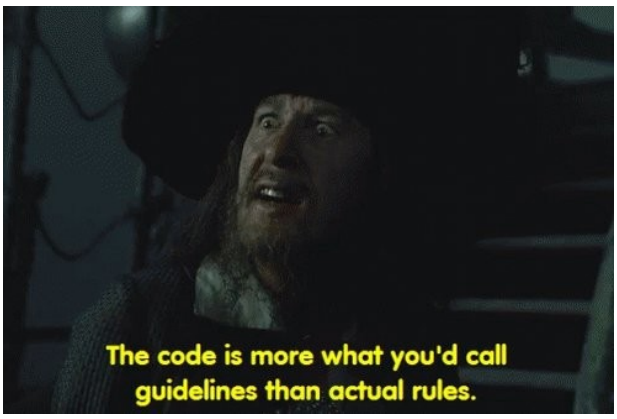
Globalement, trois niveaux d'autonomie sont actuellement plus ou moins reconnus par les différents auteurs :

1. L'IA en tant que simple outil d'aide au praticien, comme dans les logiciels de Diagnostic Assisté par Ordinateur (*Computer-Aided Diagnostic* – CAD).
2. L'IA avec une autonomie équivalente à celle d'un assistant médical.
3. L'IA avec une autonomie complète, à un niveau équivalent à celui du praticien.

Actuellement, le premier niveau d'autonomie tend à être favorisé par les différents systèmes judiciaires, notamment car c'est celui qui permet l'attribution de la responsabilité la plus simple. Le second niveau peut également être comparé à un externe en odontologie ou en médecine : il peut interagir avec les patients, participer au diagnostic et au traitement d'un patient, mais demeure sous la supervision et la responsabilité d'un praticien diplômé.

Ces trois niveaux revêtent également une importance dans le cadre du consentement éclairé du patient : il doit pouvoir être informé de l'utilisation d'une IA dans le processus décisionnel et de son degré d'autonomie afin de pouvoir prendre, avec le praticien, une décision thérapeutique en connaissance de cause. Cependant, le cadre légal européen est assez flou autour de la nature des informations qui doivent réellement être fournies au patient et sur le contexte de quand elles doivent être fournies. Le Règlement Général sur la Protection des Données (RGPD), lui, statue sur la nature de ces informations de la manière suivante : elles doivent englober l'existence d'une IA dans le processus, l'explicabilité

globale de cette IA et les implications du traitement proposé, le tout d'une manière à pouvoir être comprise par le patient. Dans le cas où une IA possède une autonomie complète, son usage est interdit sans le consentement explicite et documenté du patient. Cependant, dans les autres cas, le cadre juridique doit encore être précisé (notamment au niveau de certaines définitions floues et sujettes à des interprétations différentes) et il semble aujourd'hui plus s'agir de recommandations que d'obligations, bien que ces informations doivent pouvoir être fournies au patient s'il en fait la demande (selon l'article 15 du RGPD).



Ligne de dialogue issue du film *Pirates des Caraïbes* illustrant le cadre légal actuel autour de l'IA

Principe N°4 : Les décisions, actions et processus de communication par IA doivent être transparents et explicables.

Il devient compliqué de parler de consentement éclairé, d'une manière à ce que le patient soit à même de comprendre les informations qu'on lui fournit, quand le fonctionnement de l'IA devient une boîte noire et que l'on n'a aucune idée des processus décisionnels sous-jacents. La XAI peut ainsi entrer ici en jeu, mais à quel degré ? C'est là, notamment, où la Cour de justice Européenne est assez floue. Quand elle évoque le « droit à une explication » du patient, elle ne définit en réalité pas ce qu'est une explication. Cela peut être interprété, selon Shneeberger et al. [25], comme des « informations sur le fonctionnement de base du système », ce qui correspond à de l'explicabilité globale, ou des « explications des causes et des processus internes conduisant à une décision individuelle », ce qui correspond plutôt à de l'explicabilité individuelle ou même encore à de l'interprétabilité. Et encore, comment le présenter à un patient qui n'a pas nécessairement le bagage théorique pour comprendre le fonctionnement d'une IA et donc les explications qui y sont associées ? Cela mérite réflexion afin de trouver des solutions.

Aujourd'hui, l'explicabilité globale semble plutôt être privilégiée comme information à fournir au patient (et d'ailleurs également comme information à fournir au praticien, comme vu précédemment et bien que je ne sois pas entièrement d'accord).

Le RGPD établit notamment la transparence comme un principe clé, essentiel au respect des normes éthiques et juridiques de la part de l'IA. Cela complique notamment son application dans la pratique, notamment dans les secteurs où les conséquences peuvent être jugées comme étant critiques comme dans le secteur médical, et promeut la XAI au rang d'outil de choix pour rendre l'utilisation de l'IA la plus transparente et explicable possible.

Principe N°8 : Une décision, action ou communication par IA ne doit pas enfreindre la loi applicable et ne doit pas causer de préjudice humain.

D'un point de vue législatif autour de la confidentialité des données, le RGPD recommande l'intégration de cette gestion de la confidentialité dès la conception de l'IA et exige une ou des évaluations sur celle-ci avant la mise en application de l'IA. Notamment, lors de cette ou de ces évaluation(s), une description des mesures mises en œuvre pour respecter cette confidentialité doit être fournie et des approches telles que la XAI peuvent être mises en jeu à cette étape pour garantir notamment que des données sensibles soient protégées.

Tournons-nous maintenant vers un autre point de vue législatif, cette fois-ci tourné vers les dommages à un tiers. Même si on imagine une IA du futur très nettement supérieure au meilleur des spécialistes, il est inévitable qu'elle commette parfois des erreurs entraînant un préjudice au niveau des patients. Ce genre de cas doit alors être encadré par la loi, notamment au niveau de la responsabilité, comme on le ferait avec un être humain. Cela nous ramène au principe n°3.

Principe N°3 : La responsabilité d'une décision, action ou processus de communication par IA doit incomber à une personne physique ou juridique compétente.

Bien que le premier niveau d'autonomie de l'IA (en tant que simple outil d'aide au praticien) soit assez simple à gérer niveau responsabilité (dans ce cas toute la responsabilité demeure celle du praticien), plus l'IA gagne en autonomie et plus cela se complexifie. Qui sera légalement responsable en cas d'erreur ou de dysfonctionnement de l'IA ? Le développeur, le fabricant, le personnel de maintenance, l'hôpital, le praticien ou bien l'IA elle-même ? Et comment le prouver ? Ce sont des questions auxquelles les réponses sont encore floues dans le cadre juridique actuel. Face à cette situation, il serait possible de laisser les tribunaux statuer et créer une jurisprudence au fur et à mesure des cas qui se présenteront. Cependant la Commission Européenne vise plutôt à établir un cadre juridique uniforme sur la responsabilité civile de l'IA en amont afin de limiter la période d'incertitude le temps que ces jurisprudences voient le jour ainsi que les disparités qui en découleraient entre les différents membres de l'Union Européenne. Voici désormais un descriptif des différents types de responsabilités qui pourraient encadrer l'IA.

La **responsabilité civile** traditionnelle repose sur la preuve d'une faute, d'un responsable, d'un dommage, et d'un lien de causalité entre ces éléments. Cependant, dans certains cas spécifiques, la responsabilité peut être de nature stricte ou objective, ce qui signifie qu'elle est imputée à une personne déterminée par la loi, sans nécessité de prouver les éléments précédemment mentionnés. Par exemple, dans un accident de voiture, le propriétaire du véhicule peut être tenu responsable, même s'il n'a pas commis de faute, en vertu de la responsabilité objective.

Lorsqu'il s'agit d'IA, la complexité et l'opacité des modèles rendent difficile l'attribution des dommages à un comportement humain et l'établissement d'un lien de causalité. Cela soulève la question de savoir si la charge de la preuve devrait incomber à la victime, mais c'est également là où la XAI pourrait être d'une grande aide pour comprendre à quelle étape du processus l'erreur a potentiellement pu être commise. Cependant, au vu de la complexité de cette approche, certains suggèrent que l'IA devrait être soumise à une forme de responsabilité stricte ou objective. Cependant, les auteurs soulèvent que cette approche comporte également le risque de limiter l'innovation dans le domaine de l'IA.

La **responsabilité produit** vise à responsabiliser les producteurs et les vendeurs de produits tout en protégeant les consommateurs en cas de dommages. Cependant, lorsqu'il s'agit de l'IA, plusieurs défis se posent. En général, la responsabilité produit est appliquée à des produits physiques, or l'IA a tendance à être virtuelle quand elle n'est pas intégrée à un produit physique. De plus, en cas d'apprentissage continu, l'IA peut évoluer après sa mise en service, ce qui la distingue progressivement du produit initial. Cela soulève la question de la responsabilité du développeur en cas de dommage. Une proposition pour limiter cette évolution consiste à bloquer la capacité d'évolution d'un algorithme une fois qu'il est jugé suffisamment entraîné, bien que cela puisse limiter ses avantages et son développement. Aux États-Unis, il existe également l'**exception de l'intermédiaire informé**, qui rend responsable toute personne capable de contrôler les résultats de l'IA et de signaler les erreurs avant leur application, ce qui peut restreindre l'autonomie de l'IA et ne pas diminuer la charge de travail du praticien.

De plus, si une IA est destinée à être utilisée sur des êtres humains à des fins médicales, elle est considérée comme un **dispositif médical**. En conséquence, elle est classée comme présentant un risque potentiel « moyen » à minima et doit suivre un processus d'approbation pour accéder au marché, suivi d'une évaluation continue après sa mise sur le marché. Cependant, les lois en vigueur en Europe ne font pas actuellement de distinction entre les systèmes statiques et les systèmes d'apprentissage automatique continus, contrairement à la FDA aux États-Unis. Par conséquent, les risques spécifiques associés à l'IA ne sont pas pleinement pris en compte dans les critères de sécurité existants. La FDA a proposé que les IA non verrouillées, capables de s'adapter tout au long de leur utilisation, soient soumises à une surveillance continue tout au long de leur cycle de vie.

Lorsque l'IA agit de manière plus autonome en tant que subordonnée ou assistante d'un praticien, la **responsabilité** peut être imputée **par ricochet**, ce qui signifie que le superviseur, en l'occurrence le praticien, en assume la responsabilité, similaire à la responsabilité des praticiens hospitaliers vis-à-vis de leurs étudiants.

Une autre approche envisagée consiste à attribuer une **personnalité juridique distincte propre à l'IA**, qui pourrait être poursuivie en tant que telle en cas de dommage, bien que cela ne soit pas une solution simple et soulève d'autres questions juridiques complexes.

Aux États-Unis, l'exemple des voitures autonomes peut également être pris comme point de comparaison. Lorsque des accidents impliquant ces véhicules ont été jugés, la tendance générale a été de considérer le conducteur humain comme responsable s'il aurait raisonnablement pu éviter l'accident, par exemple, en restant attentif au volant au lieu de s'endormir pendant que la voiture traversait une zone de sortie d'école.

En revanche, lorsque l'IA est utilisée uniquement comme un outil par le praticien, la responsabilité est généralement attribuée au praticien lui-même comme nous le disions au début, ce qui simplifie la question de la responsabilité.

Principe N°9 : Une décision, action ou communication par IA ne doit pas être discriminatoire, en particulier dans la formation des algorithmes.

L'IA présente un potentiel considérable dans le domaine médical, mais son efficacité dépend en grande partie de la qualité des données d'entraînement. Si ces données sont biaisées, l'IA sera également biaisée, et si ce biais est discriminatoire, l'IA produira des résultats discriminatoires. Cela a été clairement illustré par des algorithmes d'attribution de prêts aux États-Unis, qui se sont révélés discriminatoires envers les Afro-Américains. Les risques de discrimination peuvent également se manifester en fonction de l'âge, du sexe, des revenus, et d'autres caractéristiques.

Pour minimiser ces biais et garantir des pratiques éthiques en matière d'IA médicale, les développeurs doivent être attentifs à ces questions. Des guides de bonnes pratiques doivent être élaborés pour normaliser les processus et réduire les biais, dans le cadre d'un effort collaboratif qui englobe l'ensemble de la communauté, y compris les praticiens, les entreprises, les institutions académiques et les patients. Il est également essentiel de surveiller en permanence le fonctionnement de l'IA pour détecter l'apparition de biais, même après son déploiement. Pour ce faire, des techniques d'explicabilité basée sur les données peuvent être mises en places pour analyser les jeux de données d'entraînement en amont et des techniques basées sur le modèle peuvent être mises en place tout au long de

l’utilisation de l’IA pour surveiller notamment quelles variables ont le plus d’impact sur la prédiction.

De plus, si une IA est appliquée à des groupes de population insuffisamment représentés dans le jeu de données d’entraînement, la pertinence de ses résultats doit être remise en question. Toutefois, il est tout aussi important de garantir que ces groupes puissent eux aussi bénéficier des progrès liés à l’IA. Pour résumer, elle doit être accessible à tous et la diversité des individus doit être prise en compte lors de son développement et de son utilisation afin d’éviter toute discrimination dans la qualité des soins prodigués.

Principe N°10 : La définition d’objectifs, le contrôle et la surveillance des décisions, actions et communications par IA ne doivent pas être effectués par des algorithmes.

L’un des concepts que j’ai le plus vu revenir à travers tous les articles que j’ai pu lire sur l’IA (pas seulement ceux qui portaient spécifiquement sur l’éthique ou la législation) c’est le concept de « human in the loop », ou d’humain dans la boucle. Les meilleures performances sont obtenues lorsque humain et machine travaillent conjointement et non lorsque les machines sont seules. Dans une approche éthique de ce travail main dans la main, il est essentiel qu’il y ait toujours un humain qui supervise le travail d’un algorithme (et comment surveiller une boîte noire sans l’explicabilité qui va avec?) pour vérifier que, même si les performances peuvent être potentiellement impressionnantes, il n’y ait pas d’aspect humain de négligé durant le processus. Les auteurs vont jusqu’à dire qu’il est impératif que les patients ne soient pas considérés comme de simples objets et que, par conséquent, la décision finale doive toujours découler de l’humain. C’est pour cela, notamment, que le RGPD est bien plus strict dans son encadrement des IA autonomes et les restreint à certaines applications seulement. Mais même dans ces cas, le patient conserve le droit d’obtenir une intervention humaine, d’exprimer son point de vue et de contester la décision.

Attention cependant, dans certains cas même si l’humain est présent il peut ne pas avoir la possibilité d’interroger le résultat donné par l’IA ni la possibilité de modifier la décision

qui en découle. Un exemple donné est le cas où une infirmière serait obligée de suivre strictement le résultat et la thérapeutique indiqués par l'IA [25]. L'IA est alors considérée comme agissant de manière autonome dans ce cas et est soumise à la législation appropriée.

Même si l'humain est bien intégré dans la boucle, une autre problématique peut également se poser. Il est entendu que tout praticien doit pratiquer des soins selon les dernières données acquises de la science et dispose d'une obligation de moyen. Ceci est encadré par les états de l'Union Européenne qui s'assurent que les soins soient prodigués selon ces standards. De fait, une IA démontrant des performances supérieures à celles d'un humain peut-elle devenir la nouvelle norme de soin et entraîner une obligation d'utilisation ainsi que des sanctions pour le praticien en cas de non-utilisation ? Cela entraîne nombre de nouveaux questionnements, qui entrent parfois en conflit avec les dix principes de Muller et al. :

– Si cette IA est si performante qu'elle dépasse les capacités du praticien, y compris ses capacités de compréhension, on entre en conflit avec le principe n°5 de compréhensibilité. De nouvelles techniques, de XAI par exemple, doivent alors être développées afin de la rendre compréhensible à moins que l'on considère que les bénéfices apportés sont supérieurs à cette règle censée participer à un encadrement éthique de l'IA.

– Si le praticien décide de ne pas suivre les indications d'une IA « gold standard », devra-t-il justifier du fait de s'en être écarté ? On peut par exemple imaginer une utilisation de l'explicabilité locale dans ce but, afin de démontrer une différence de raisonnement à propos d'une instance ou de contrôler que le modèle ne produit pas un résultat incongru pour cette instance, rendant le résultat inadapté par exemple. Et si, au contraire, le praticien décide de se fier au résultat de cette IA, pourtant erroné, il reste difficile d'évaluer les éléments suivants en raison de l'opacité des modèles d'IA :

1. La rationalité de la décision d'utiliser l'IA.
2. La justification ou l'erreur du médecin (en ce qui concerne la conformité à la norme médicale requise) de s'écarter d'une recommandation de l'IA.
3. La détermination de la responsabilité du médecin dans ce contexte. [25]

– De plus, on peut se poser la question de l'éligibilité d'un tiers à juger de ces éléments et de la raison (ou du tort) du praticien de suivre ou non les indications de l'IA. En temps normal, on ferait appel à un expert dans le domaine pour juger de la faute du praticien, mais dans le cas ou une IA pourrait être plus performante qu'un spécialiste, cet expert aura-

t-il les compétences pour juger de l'erreur ou de la justesse dans le diagnostic de cette IA ? Encore une fois, la XAI peut ici jouer un rôle crucial. On peut également imaginer une sorte d'IA-expert à même de juger ce genre de cas, mais l'on vient alors, d'une certaine manière, à l'encontre du principe n°10 en laissant un algorithme juger d'un autre algorithme.

Malgré toutes ces problématiques, de grands organismes comme l'Organisation Mondiale de la Santé ou les Nations Unies considèrent tout de même l'IA comme un élément-clé pour la réalisation des trois objectifs suivants de développement durable en matière de santé de bien-être de la population : étendre la couverture de santé universelle à plus de personnes, protéger plus de personnes contre les urgences sanitaires et améliorer la santé et le bien-être de toujours plus de personnes [23]. Ce qu'il faut principalement retenir, c'est que :

– les complexités juridiques augmentent à mesure que l'autonomie de l'IA augmente ;

– l'utilisation de l'IA dans le secteur médical nécessite une surveillance humaine (« human in the loop ») et une explicabilité ;

– le cadre légal actuel autour de l'IA et de sa responsabilité est flou et lacunaire, mais il va être amené à changer prochainement (par exemple : par la mise en place de procédures d'autorisation de mise sur le marché, d'une potentielle responsabilité stricte associée à un régime d'assurance obligatoire, des modifications dans la charge de la preuve, etc.).

La transparence et l'incitation à l'usage de la XAI seront donc mis en avant de la scène pour intégrer ce futur cadre.

2. L'intégration de l'IA et de la XAI au système de soins

a. À propos des performances et du potentiel de l'IA

La XAI étant un outil servant l'IA, elle n'aurait que peu d'intérêt pour nous si l'IA se révélait dénuée de potentiel dans le domaine de la santé. Bien qu'étant loin d'avoir atteint son plein potentiel, notamment à cause du retard du domaine de la santé sur le nombre de jeux de données de qualité disponibles et accessibles, comparativement à d'autres domaines (comme l'imagerie spatiale, la finance, la F1, etc.), l'IA se révèle tout de même très prometteuse.

Les exemples vus durant la première partie [16 - 21] nous montrent ainsi que les performances de l'IA se révèlent comparables voir légèrement supérieures à celles des praticiens. Cette observation est accompagnée d'un potentiel certain pour alléger la charge de travail des professionnels de la santé bien qu'il ne soit pas encore atteint dans nombre d'applications. Bharati et al. [29] affirment également que les algorithmes d'IA peuvent être encore plus précis que les praticiens pour certaines tâches de diagnostic.

Ainsi, dans leur revue [36], Raman & al comparent les preuves actuelles à propos de différents modèles de deep learning pour le diagnostic de la rétinopathie diabétique. Ils observent ainsi que si l'on suit les critères d'évaluation de la FDA, la plupart des méthodes à base d'IA disponibles aujourd'hui ont la capacité d'apporter une plus-value pour le dépistage de cette rétinopathie avec de meilleures performances que les praticiens, qu'il s'agisse de sensibilité, de spécificité ou de rapidité.

De leur côté, Muddamsetty & al [22] comparent les performances d'ophtalmologues avec celles de deux modèles de CNN sur l'analyse d'images de fonds d'oeil rétiniens. Ces deux modèles atteignent 94 % et 85 % de précision, ce qui est équivalent à la précision d'un humain. Mais surtout, là où cet article devient intéressant, c'est qu'ils comparent, à l'aide d'un eye-tracker pour les médecins et de la XAI pour les CNN, les zones d'intérêt qui ont permis ce diagnostic pour les uns et pour les autres. Il en ressort que, pour ces algorithmes, les zones d'intérêt sont proches entre les praticiens et les CNN, ce qui rend les résultats vraiment comparables entre un humain et la machine. Cependant il est à préciser qu'il s'agit, dans cet article, uniquement de ces deux modèles. Ce n'est pas toujours le cas et bien que cela soit un exemple de ce qui peut se faire il n'est pas possible de l'extrapoler à tous les algorithmes. Car même si les algorithmes à base d'IA en général peuvent se révéler excellents en conditions expérimentales, il en va autrement dans les contextes cliniques réels où ils rencontrent plus de difficultés (on peut reprendre par exemple l'exemple d'E. J. Topol [66] qui parle d'un taux de faux positifs trop élevé) [29].

En effet, l'une des limitations de l'IA en santé est le manque de jeux de données de qualité et correctement annotés afin de pouvoir réaliser un apprentissage supervisé sur des données suffisamment proches des données réelles pour que les algorithmes puissent être également performants en conditions réelles. Pour cela il existe une autre possibilité : l'apprentissage non supervisé. Cependant, ce dernier pose des difficultés pour « identifier le modèle initial » [2]. De plus, bien que cela résolve le problème de l'annotation des jeux de données, ça ne change rien à celui de leur qualité.

Une autre solution pourrait être le développement d'IA basées sur la connaissance. À la différence des algorithmes d'apprentissage automatique qui tirent des règles à partir de jeux de données (construction ascendante), le concept d'IA basée sur la connaissance tend à imiter le fonctionnement du cerveau humain : l'IA possède de prime les connaissances puis s'emploie à les appliquer (construction descendante). Néanmoins, d'après Montani et Striani [67], il existe deux goulots d'étranglements qui ralentissent la recherche et demandent beaucoup d'efforts de recherche pour être passés : l'acquisition des connaissances et la formalisation de ces mêmes connaissances. Pour l'instant, les algorithmes d'IA en santé sont donc essentiellement des IA à construction ascendante [2].

b. Comment intégrer l'IA au flux de travail dans le domaine de la santé ?

En 2018, l'American Medical Association a introduit le concept d'intelligence augmentée, qui est une conceptualisation de l'IA dans les soins de santé comme une assistance pour les praticiens (Crigger et Khoury, 2019). On en revient au même point récurrent : l'IA doit être conçue pour coopérer avec l'humain, l'objectif est d'augmenter l'intelligence humaine et non de remplacer l'expertise humaine (pour reprendre les termes de l'article de Shan et al.[2]). Dans leur article[28], Leuween et al. déclarent même « AI is a means, a tool, not the goal in itself » (L'IA est un moyen, un outil, et non le but en soi).

Toujours dans ce même article[28], les secteurs d'action de l'IA sont répartis en 6 pôles : améliorer l'efficience du flux de travail, réduire la durée de lecture des examens, améliorer la précocité des détections, réduire la dose de rayonnement et de produit de contraste nécessaire à l'obtention d'images radiologiques lisibles, augmenter la précision des diagnostics et personnaliser les diagnostics. Le schéma suivant illustre ces six pôles :

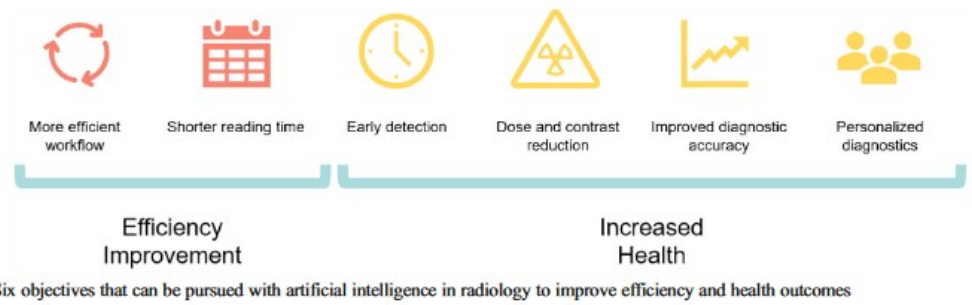


Fig. 1 Six objectives that can be pursued with artificial intelligence in radiology to improve efficiency and health outcomes

Schéma issu de l'article de Leuween et al[28] illustrant les secteurs d'action de l'IA

Ces 6 pôles sont eux-mêmes répartis en deux grandes catégories que sont les IA destinées à augmenter la qualité des soins de santé et celles destinées à augmenter leur efficience. L'efficience est un concept intéressant dans la problématique actuelle du vieillissement de la population et de l'évolution de la technologie, qui causent tous deux une augmentation continue de la demande de soins et de leur coût.

En 1991, Fryback et Thornbury ont formulé un modèle hiérarchique, ultérieurement adapté à l'évaluation de l'efficacité de l'IA. Ce modèle se décline en six niveaux, chacun englobant différents aspects dont les différents pôles que nous venons de voir :

- **Niveaux 1 et 2 :** On s'intéresse, à ces niveaux, au fonctionnement et aux performances de l'IA. La plupart des études les concernent et se concentrent sur les performances des IA.
- **Niveaux 3 à 5 :** On s'intéresse ici aux influences de l'IA sur le diagnostic, la thérapeutique et les résultats obtenus pour les patients. Mais en 2021, une étude prospective a démontré que les bénéfices de l'IA une fois appliquée à la clinique étaient limités et ne concernaient que 18 des 100 produits évalués lors de cette étude [37].
- **Niveau 6 :** Enfin, ce dernier niveau évalue les effets que peut avoir l'IA à grande échelle, comme sur les coûts et le niveau de santé global. Cependant, les preuves restent limitées à ce niveau aussi. A titre d'exemple les articles éligibles répertoriés dans une revue systématique sur l'impact économique de l'IA dans les soins de santé en 2020 ne sont qu'au nombre de 6 [46].

Voici le tableau issu de l'article[28] détaillant ces 6 niveaux :

Table 1 Hierarchical model of efficacy to assess the contribution of artificial intelligence (AI) software to the diagnostic imaging process, adapted from [5, 6]

Level	Explanation
Level 1t	Technical efficacy
	Study demonstrates the technical feasibility of the software
Level 1c	Potential clinical efficacy
	Study demonstrates the feasibility of the software to be clinically applied
Level 2	Diagnostic accuracy efficacy
	Study demonstrates the standalone performance of the software
Level 3	Diagnostic thinking efficacy
	Study demonstrates the added value to the diagnosis
Level 4	Therapeutic efficacy
	Study demonstrates the impact of the software on the patient management decisions
Level 5	Patient outcome efficacy
	Study demonstrates the impact of the software on patient outcomes
Level 6	Societal efficacy
	Study demonstrates the impact of the software on society by performing an economic analysis

Tableau issu de l'article de Leuween et al[28] portant sur la hiérarchie de l'évaluation de l'efficacité de l'IA

Les bénéfices attendus à propos de l'IA dans le domaine de la médecine sont donc les suivants :

- **Pour l'amélioration de la qualité des soins, l'IA peut contribuer à :**

- personnaliser les diagnostics et les décisions médicales ;

- adapter les pratiques de santé et les médicaments à des individus spécifiques ;

- proposer des options préventives ;

- réduire les doses de rayonnement et d'agents de contraste en radiologie ;

- détecter précocement les maladies ;

- améliorer la précision des diagnostics ;

le tout dans le but d'améliorer le bien-être et la santé de la population [28, 29].

- **Pour la réduction des coûts des soins, l'IA peut contribuer à améliorer l'efficacité du flux de travail en :**

- réduisant le temps de lecture des examens ;

- aidant à détecter précocement les pathologies ;

- proposant des mesures de prévention.

Alors que le coût de la santé augmente continuellement en raison de l'évolution des technologies et du vieillissement de la population, le concept d'efficience prend de plus en plus d'intérêt afin de faire toujours mieux avec les ressources limitées dont nous disposons. L'IA peut également aider à améliorer cette efficience, en agissant, en plus du domaine clinique, sur les domaines périphériques à la clinique comme en aidant à la gestion des patients ou de la logistique par exemple. Comme ce sont des applications moins risquées pour l'IA qu'une prédiction de diagnostic par exemple, la réglementation est simplifiée et il en va de même pour leur application concrète dans la vie réelle.

Leuween et al. [28] posent l'équation valeur(des soins de santé)=résultat/coût, où la valeur des soins de santé est donc égale à leurs résultats divisés par leur coût. Suivant cette équation, l'IA peut donc apporter de la valeur en améliorant la qualité des soins (le premier point évoqué ci-dessus) ou en diminuant les coûts (le second point), ce qui fait que les

auteurs considèrent ces deux aspects comme les objectifs ultimes de l'IA en radiologie, mais l'on peut les étendre au domaine médical dans sa globalité.

D'un point de vue concret, un graphique issu de l'article de Becker et al. [26] nous montre, en pratique, comment sont actuellement employées les IA par les radiologues de l'ESR. On y retrouve les différents pôles sus-mentionnés ainsi que ces deux aspects : en vert les points améliorant la qualité des soins et en bleu le point diminuant le coût des soins.

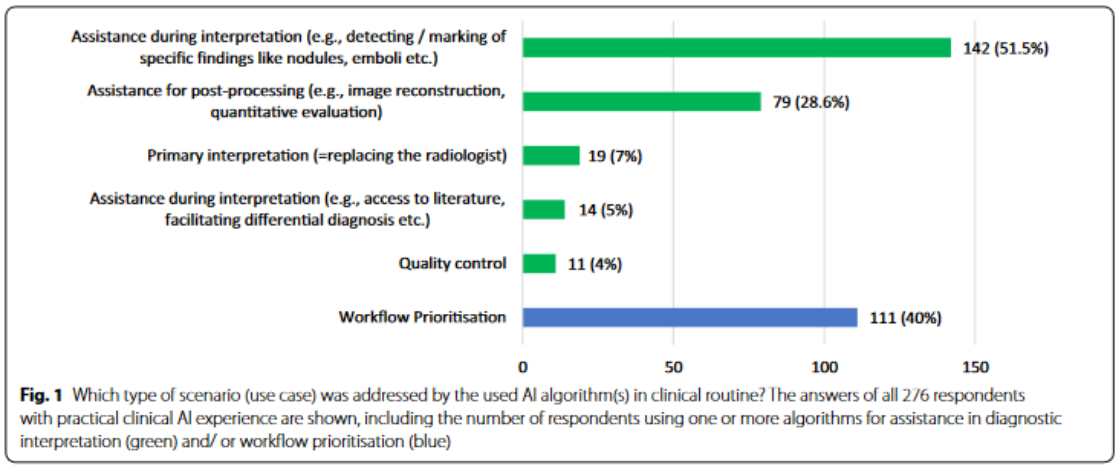


Diagramme issu de l'article de Becker et al [26] à propos des rôles de l'IA en pratique clinique

En médecine, on peut donc retrouver les exemples suivants d'utilisation de l'IA suivant ces pôles :

- **Amélioration de l'efficience du flux de travail et diminution du temps de lecture des examens :**

L'IA peut par exemple aider à planifier les rendez-vous d'un service hospitalier afin de limiter la perte de temps due aux patients qui ne se présentent pas à leurs rendez-vous : une étude de Chong et al. [43] a développé un modèle d'IA prédictif à même d'identifier les patients les plus susceptibles de manquer leur rendez-vous afin de pouvoir leur envoyer un rappel téléphonique supplémentaire. Ils ont ainsi réussi à réduire le taux d'absence de 19,3 % à 15,9 %, ce qui, cumulé sur un service entier (d'imagerie par résonance magnétique dans cet exemple) et sur la durée, permet de soigner un nombre non-négligeable de patients supplémentaires.

Un autre exemple de problématique au Royaume-Uni montre qu'entre 2013 et 2018 la quantité d'examens d'imagerie réalisée a augmentée de 50 % environ tandis que le nombre

de professionnels de santé à même de les traiter n'a augmenté que de 19 % [38]. Même s'il n'a pas d'exemple déjà existant à nous présenter pour ce problème, l'article de Leuween et al. [28] le cible comme un exemple type de situation qui pourrait être améliorée grâce à l'IA. En effet, celle-ci pourrait permettre de réduire le temps de lecture des radiographies afin de lutter contre l'augmentation de la charge de travail et de la fatigue professionnelle que génère cette croissance différentielle.

En revanche ils nous fournissent une panoplie d'autres exemples concrets issus des articles [39-42] :

- une réduction du délai entre le diagnostic et l'intervention dans les cas d'AVC (accidents vasculaires cérébraux) de 281 minutes en moyenne à 243 minutes grâce à l'IA ;
- toujours dans ces mêmes cas d'AVC, une réduction de la durée de séjour des patients (ce qui correspond à une évaluation, pour cette étude, aux niveaux 4 et 5 de la hiérarchisation précédemment mentionnée : *therapeutic efficacy* et *patient outcome efficacy*) ;
- une réduction du délai de compte-rendu pour les radiographies réalisées dans des cas critiques de 80 minutes en moyenne à 35-50 minutes ;
- une réduction du temps d'attente pour les cas d'hémorragie intracrânienne de 16 minutes en moyenne à 12 minutes grâce à une IA de priorisation.

- **Diminution de la dose d'irradiation et de la quantité de produit de contraste :**

Cela revient à améliorer la qualité des soins mais également leur efficience car une réduction de l'irradiation reçue par les patients et des doses de produits de contraste qui leur sont injectées conduit à une réduction du risque de développer une pathologie iatrogène et donc une réduction du nombre total de pathologies à devoir soigner. C'est particulièrement important en pédiatrie car le risque de cancer radio-induit est accru. Pour ce faire, l'IA est notamment utilisée pour améliorer et accélérer la reconstruction des images et leur post-traitement, ce qui permet une meilleure qualité d'image avec une dose réduite [28].

- **Augmentation de la précision diagnostic :**

Cela améliore la qualité des soins mais également l'efficience globale du parcours de santé, car une erreur de diagnostic fait perdre du temps avec des examens, des thérapeutiques, etc, inutiles. Le coût est multiple : il est ponctionné sur la santé du patient, sur son temps, mais aussi sur celui des professionnels de santé et sur le coût de ces examens, traitements, etc, inutiles.

Actuellement, dans les produits à base d'IA commercialisés pour la radiologie, environ la moitié de ces produits se concentrent sur l'amélioration de la précision diagnostic. On y retrouve un grand nombre de logiciels de CAD (computer-aided detection) qui aident à la lecture des radiographies par le biais de boîtes de délimitation, de marqueurs et de scores de probabilité. Des études ont été réalisées pour évaluer l'apport de l'IA sur la précision du diagnostic (niveau 2 de la hiérarchisation : *diagnostic accuracy efficacy*), concluant à une amélioration significative de celle-ci dans le cas de la détermination de l'âge osseux comparée à une détermination à partir de l'atlas de Greulich-Pyle. En réalité, dans ce cas-là l'IA est déjà largement adoptée en pédiatrie, cependant ce type d'IA n'est pas autonome et son efficacité repose sur une synergie entre elle et le praticien, ce qui classe son évaluation dans le niveau 3 de la hiérarchisation : *diagnostic thinking efficacy* [28, 44, 47].

Dans une de leurs publications [22], Muddamsetty et al. comparent les performances d'ophtalmologues avec deux modèles de réseaux de neurones pour l'analyse de fonds d'œil rétiniens. Outre des performances globalement similaires entre les IA et les spécialistes, le point fort de cette étude est l'utilisation d'un eye-tracker sur les ophtalmologues pour comparer les zones d'intérêts les plus consultées sur les images avec celles les plus consultées par les réseaux neuronaux et révélées par la XAI, révélant de très fortes similitudes entre les deux et donc une logique et des règles sous-jacentes dans l'établissement d'un diagnostic a priori proches.

- **Personnalisation des diagnostics et détection plus précoce :**

Cela améliore la qualité des soins mais aussi leur efficience, car une pathologie traitée plus précocement requière moins de soins. De plus, l'IA peut permettre d'anticiper les risques propres à chaque patient et d'allouer ainsi plus précisément les ressources en dirigeant les

ressources adéquates (comme des traitements ou des tests supplémentaires) vers les patients les plus susceptibles d'en bénéficier.

Par exemple dans le cas de cancers du sein, où la densité mammaire est reconnue comme étant un facteur de risque, une étude menée aux Pays-Bas a permis d'établir un groupe de 40 000 femmes avec une densité mammaire extrêmement élevée. La densité a été identifiée de manière automatisée à l'aide d'une IA commercialisée, afin de mettre en place un dépistage personnalisé avec un suivi plus fréquent ou modifié (avec par exemple la réalisation d'IRM supplémentaire). Cela a permis d'améliorer significativement le dépistage du cancer du sein (niveau 5 : *patient outcome efficacy*) [28, 45].

Dans le domaine de l'odontologie, les applications de l'IA sont pour la plupart virtuelles[1] et servent par exemple à distinguer les lésions des structures physiologiques, à hiérarchiser les facteurs de risque et à simuler et évaluer les résultats attendus d'un traitement. Cependant, contrairement à la médecine d'un point de vue plus globale, elle y est en retard de plusieurs années [49]. Pour expliquer ce retard, on retrouve des limites communes à différents secteurs médicaux comme un manque de data sets qualitatifs dû, d'après Hosny et al. [51], à un manque de conservation et de partage des données, mais aussi plus spécifiquement à un « manque d'informations sur le traitement, la mesure et la validation des données » qui est un défaut de la recherche sur l'IA en dentaire [50]. Cela correspond effectivement à mon expérience et mon ressenti lorsque j'ai souhaité chercher un jeu de données pour l'expérimentation sur l'intérêt de la XAI. De fait, Shan et al. [2] déclarent donc qu'à l'avenir l'amélioration de la quantité, de la qualité et de la lisibilité des données est cruciale pour le développement de l'IA en odontologie. Pour ce faire, ils proposent de standardiser la méthodologie de conservation et de report des données. Mais en attendant, ils proposent déjà un petit aperçu de ce qui existe déjà aujourd'hui. Néanmoins, j'ai choisi de classer ces exemples d'application de l'IA un peu différemment de ceux abordés précédemment, en fonction des catégories d'application que je distingue pour répondre aux problématiques actuelles rencontrées dans l'exercice dentaire en France :

1 Les IA virtuelles sont des algorithmes intégrés dans des logiciels pour soutenir la prise de décision clinique par exemple, contrairement aux IA physiques comme les robots ou les bras robotiques automatisés par exemple [48].

- **Applications capables de suppléer l’omnipraticien en cas de difficulté d’accès à un spécialiste :**

En France, la démographie actuelle des praticiens en santé bucco-dentaire est telle que les spécialistes sont majoritairement regroupés autour des villes et les campagnes (comme les déserts médicaux) sont bien en peine d’accéder à ces spécialistes. Certains patients se retrouvent ainsi à parfois faire plusieurs heures de route pour se rendre à un rendez-vous avec un spécialiste. De plus, les rendez-vous avec ces spécialistes sont souvent difficiles à obtenir (on observe notamment une errance médicale prolongée, notamment lorsqu’un patient doit consulter plusieurs spécialistes) avec des délais plus que conséquents. Ainsi, je considère tout outil capable d’assister l’omnipraticien de manière à diminuer le besoin de renvoyer le patient vers un spécialiste comme une plus-value énorme pour les patients. Et ceci sans compter les patients qui abandonnent les soins faute d’un accès facile à un spécialiste, le stress engendré par une prise en charge complexifiée et qui traîne en longueur, etc, problèmes qui pourraient être réduits pas une prise en charge améliorée chez l’omnipraticien. De plus, avec la complexification et la diversification des actes liées aux avancées de la science, un omnipraticien ne peut être excellent dans tous les domaines. L’accès à des outils capables de supporter un omnipraticien dans les domaines où il se trouve moins à l’aise permettrait également d’améliorer la qualité globale des soins apportés aux patients.

Voici donc quelques applications à base d’IA à même de remplir ce rôle :

- Yilmaz et al. ont développé un modèle à même de différencier les kystes périapicaux des tumeurs odontogènes dans 94 % des cas à partir de radiographies 3D [52].
- Sunny et al. ont développé un modèle capable de différencier les lésions buccales malignes avec une sensibilité de 93 % (et les lésions potentiellement malignes de haut grade avec une sensibilité de 73%) à partir d’images cytologiques [53].
- Jeon et al. [54] ont développé des modèles capables de distinguer le lichen plan buccal d’autres lésions blanches à partir de l’expression de gènes de cytokines inflammatoires.
- Une étude de Kise et al. [55] a montré la supériorité d’un modèle par rapport à des radiologues inexpérimentés pour différencier un véritable syndrome de Sjögren d’une xérostomie à partir d’échographies.
- Feres et al. [56] ont montré de bons résultats dans le diagnostic différentiel de grades de parodontites avec de l’IA, en se basant sur les antécédents médicaux, les informations

cliniques et les radiographies des patients, ainsi que des pistes non-intégralement élucidées sur l'étiologie de ces parodontites.

- Thanathornwong [57] a montré un bon niveau de concordance entre un modèle et des orthodontistes à propos du diagnostic de besoins en traitement ODF de patients.

Ces exemples sont donc porteurs de promesses à propos d'applications futures de l'IA. On peut aisément imaginer un omnipraticien usant d'un logiciel capable de déterminer les probabilités diagnostiques d'une lésion en bouche et de proposer une intervention au cabinet en cas de fort indice de confiance associé à un risque faible pour le patient ou au contraire de proposer une réorientation vers un spécialiste en cas de faible indice de confiance et/ou risque élevé dans la pathologie suspectée. Cela permettrait par exemple de mieux trier le flux de patients adressés aux chirurgiens oraux, spécialisés dans ce type de diagnostic. On peut également imaginer un logiciel capable de prédire la nécessité ou non d'envoyer un patient chez un ODF afin de ne pas passer à côté d'un traitement bénéfique pour le patient.
Cependant il reste encore un peu de chemin à parcourir avant d'en arriver là, afin d'avoir des logiciels capables de performer en conditions réelles tout en fournissant de bonnes explications au chirurgien-dentiste pour répondre aux obligations et besoins évoqués dans les parties consacrées à la juridiction autour de l'IA et aux attentes des professionnels de santé. De plus, beaucoup de modèles se concentrent encore sur un seul type de données, comme les radiographies, alors qu'un cumul d'informations médicales supplémentaires à propos du patient permettrait un diagnostic encore plus précis [2].

- **Applications capables d'améliorer l'efficience du flux de travail :**

Aujourd'hui, le ratio nombre de praticiens/nombre de patients est bien trop faible en France alors que la demande de soins dentaires ne cesse d'augmenter (notamment avec le vieillissement de la génération du baby boom de l'après-guerre). Résultat : l'accès à un omnipraticien et globalement aux soins dentaires devient de plus en plus compliqué avec des délais de plus en plus longs. Les applications de l'IA capables d'améliorer l'efficience de travail sont donc un moyen plus qu'intéressant pour pouvoir, à temps égal, générer une plus grande quantité de soins dentaires de qualité. C'est aussi un moyen de réduire potentiellement la charge de travail et la charge mentale des praticiens, ce qui, dans un

contexte d'arrêts maladies ou de changements de métier pour cause de surmenage croissants, est un atout non-négligeable.

Voici quelques exemples d'applications prometteuses pour l'amélioration du flux de travail :

- Setzer et al. [58] évoquent un modèle capable de segmenter les différentes structures d'une radiographie 3D (dent/os/restauration/lésion/arrière-plan) avec une précision comparable à celle d'un praticien. Cet exemple aurait tout aussi bien pu être placé dans le point précédent mais j'ai choisi de le placer ici, imaginant une application en pratique à même d'assister la lecture de radiographies pour en diminuer le temps et pour éviter de passer à côté d'éléments-clés.
- Alarifi et AlZubi [59] ont développé un modèle qui analyse les données des patients, le système d'implant et les précédents actes du chirurgien pour estimer le taux de réussite de la pose de l'implant.
- Yamaguchi et al. [60] ont développé un modèle capable de prédire la probabilité de décollement de couronnes avec une précision de 98,5 % à partir d'images des piliers.
- Zhang et al. [61] ont développé un modèle à même de prédire le gonflement post-opératoire du visage des patients avulsés de leurs 3^{e} molaires mandibulaires incluses à partir de facteurs personnels, anatomiques et chirurgicaux, avec une précision de 98 %.

On peut imaginer une utilité à ces trois dernières applications de pronostic de réussite ou de complications pour prévoir les suites de soins à plus ou moins long terme par exemple et organiser ces suites de soins à l'avance. On peut également imaginer des propositions de techniques opératoires ou thérapeutiques alternatives en cas de mauvais pronostic sur un soin donné afin d'en diminuer les risques d'échec. Cela améliore la qualité des soins, mais également l'optimisation du temps de travail en ayant moins besoin de revoir les patients pour des échecs de traitement.

Pour reprendre les mots de Shan et al. [2] « Les modèles de classificateur et d'IA prédictive aident à prioriser les facteurs de risque et à prédire les résultats à long terme des maladies dentaires en explorant les associations entre les maladies et les données des patients ».

Alors que nous parlions plus tôt du manque de modèles qui se basaient sur plusieurs types de données, nous pouvons voir ici que ceux qui le font, en se basant notamment sur les symptômes cliniques, les antécédents du patient, les données démographiques, le mode de

vie et les facteurs cliniques et génétiques, obtiennent de très bons résultats. Cela conforte l'intuition que cette piste est à explorer de plus en plus.

Je n'ai pas trouvé d'exemples spécifiques à la dentisterie mais l'on peut également imaginer les applications vues précédemment en médecine pour le rappel des patients les plus susceptibles de manquer un rendez-vous par exemple et qui soient également appliquées en odontologie. Globalement, la plupart de ce qui est applicable en médecine peut être transposé au domaine dentaire du fait de la variété de compétences requises et utilisées pour les soins dentaires (radiologie, chirurgie, pathologie…).

- **Applications capables d'améliorer la qualité absolue des soins dentaires :**

Il est à différencier deux choses : ce qui se fait de mieux à l'heure actuelle, et sa démocratisation au sein des cabinets. Si dans les deux points précédents nous avons vu des exemples d'IA à même d'aider à démocratiser les soins de qualité pour les rendre accessibles à tous, l'IA peut également être un moyen de faire avancer la science. Voici quelques exemples de domaines où l'IA permet de faire progresser le champ des possibles :

- Fu et al. [62] ont développé une technique de conversion vocale pour transformer les murmures non-audibles et de patients en phase de rééducation post-opératoire après une chirurgie de la face en discours compréhensible pour l'entourage.
- Bindal et al. [63] ont développé un modèle capable de prédire la viabilité de cellules souches pulpaires en fonction des concentrations en lipopolysaccharides bactériens et des durées de traitement reçu. Shan et al. [2] déclarent à propos « Cela pourrait constituer un futur outil pour évaluer les protocoles de dentisterie régénérative dans un microenvironnement inflammatoire simulé. »
- Patcas et al. [64] ont développé un modèle à même de définir un score d'attractivité objectif et reproductible de l'apparence d'un visage, pouvant être utilisé en orthodontie pour évaluer l'objectif à atteindre. Cette application est tout de même à surveiller avec beaucoup de précautions du fait des dérives qui peuvent en découler et nuire à l'éthique, de plus, un résultat objectif à atteindre peut ne pas correspondre à la subjectivité du patient et ses attentes. Enfin, si le modèle a été entraîné sur un jeu de données annoté par des humains ou suit des règles définies par des humains, peut-on parler de réelle objectivité ?
- Wang et al. [48] on développé un système de préparation des dents en vue de la réalisation de prothèses fixes qui allie l'IA et un système robotique. Ils ont obtenu une bonne précision (de l'ordre de 50 à 60 microns) in vitro mais cette technique n'est pour le

moment pas applicable in vivo du fait du retard de la robotique sur l'IA et du manque de recherches dans ce domaine : la durée de la préparation tourne autour des trois heures et demi.

Pour plus de références à propos de l'IA appliquée à l'odontologie, voir l'annexe 1.

Revenons maintenant à l'IA en santé d'un point de vue plus global. Le graphique suivant issu de l'article de Leeuwen et al. [28] illustre l'évolution du nombre de produits basés sur l'IA disponibles sur le marché pour le domaine de la radiologie et présente une augmentation soudaine à partir de 2019.

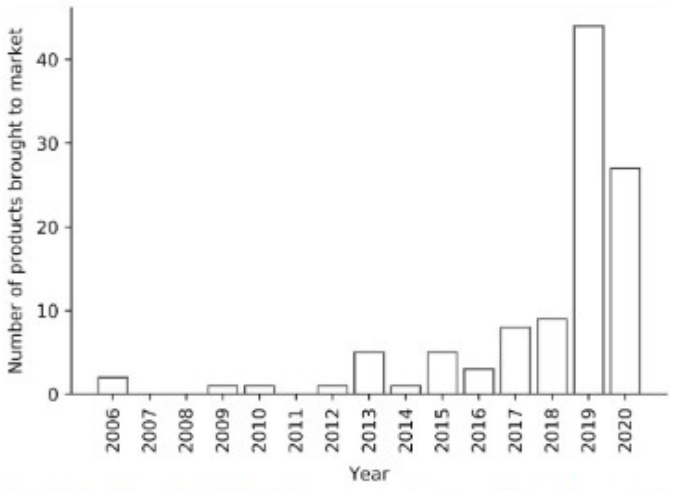


Fig. 2 Number of artificial intelligence products in radiology brought to market based on data from [3]

Graphique issu de l'article de Leuween et al [28] sur l'évolution du nombre de produits basés sur l'IA, en radiologie, disponibles sur le marché

D'après les auteurs, l'IA est déjà largement déployée en santé publique, en recherche dans le domaine de la biologie et de la pharmacologie. Cependant, d'après Bharati et al. [29] et malgré son emploi dans divers secteurs de la santé, l'IA n'est pas encore déployée à grande échelle. Des facteurs tels que le manque de robustesse des modèles, la complexité des tâches de modélisation clinique et les enjeux élevés contribuent à expliquer ce retard [27].
Malgré son potentiel, l'impact réel de l'IA dans des domaines de santé tels que la radiologie n'a encore été prouvé que de manière limitée jusqu'à présent et est souvent établi sur des simulations ou des études rétrospectives, ce qui n'est donc pas du plus haut niveau de preuve. Outre la phase encore précoce de l'application de l'IA en médecine dans

laquelle nous nous trouvons, les vues des entreprises sur certaines applications plus que d'autres peuvent également expliquer ce défaut de preuves. En effet, celles-ci se concentrent davantage sur les applications commercialisables rentables ce qui concentrent les IA développées sur les mêmes applications et en laissent de nombreuses autres inexploitées.

De plus, il est intéressant de noter qu'une certification de la part de la FDA ou du marquage CE ne sont pas synonymes de valeur clinique ajoutée pour les produits commercialisés. Cette valeur clinique ajoutée est très variable suivant les conditions et la pertinence de l'application clinique concrètes de ces produits, notamment du fait de la variabilité des méthodes de travail dans les différents hôpitaux, des différences de population, etc.

Cela a mené à l'apport de propositions pour instaurer une réglementation systémique sur ces produits afin de mieux en évaluer l'impact réel. Il ressort ainsi des études préliminaires que ces applications pourraient être bénéfiques aux patients notamment de deux manières sur le long terme, en réduisant leur incapacité et leurs besoins en soins futurs, ce qui fait que, bien que les coûts initiaux de la mise en place de ces applications seraient supportés à court terme par les départements hospitaliers par exemple [65], à long terme il pourrait en découler des avantages économiques et médicaux d'un point de vue systémique [28].

L'intégration croissante de l'IA dans le domaine médical soulève également d'autres problématiques. Par exemple, comme en témoigne l'article de Becker et al. [26], la divulgation de l'utilisation d'un algorithme de diagnostic aux patients (dans 17,3% des cas) et son enregistrement dans le rapport (dans 34,6% des cas) sont encore particulièrement rares. Cette situation soulève des interrogations éthiques et légales, notamment en ce qui concerne le consentement éclairé comme nous avons pu le voir précédemment.

Par ailleurs, une autre étude [30] met en lumière divers biais, tels que les biais d'ancrage, de confirmation, d'automatisation et de complaisance, susceptibles de survenir avec l'utilisation de l'IA en tant qu'outil d'aide au diagnostic. Dans cette étude, des étudiants en dentaire doivent poser un diagnostic de lésion inter-radiculaire à partir de radiographies retro-alvéolaires, certains ayant en plus à disposition une prédiction réalisée par une IA et d'autres non. Il en ressort qu'il n'y a pas réellement de différence significative entre les deux groupes dans ce cas mais que, cependant, le groupe assisté par IA tend à accorder une confiance excessive dans cet outil au point de se tromper lorsque l'IA se trompe. Ces

résultats soulignent l'importance d'une utilisation prudente de l'IA, évitant une dépendance excessive qui pourrait compromettre la précision des diagnostics médicaux.

Face à certaines de ces problématiques, des mécanismes d'assurance qualité sont déployés par les professionnels de la santé [26]. Ces mécanismes incluent notamment l'enregistrement des divergences de diagnostic entre les radiologues et les algorithmes, ainsi que l'établissement de courbes ROC basées sur les diagnostics des radiologues. Cependant, il est crucial de noter que ces processus peuvent être chronophages, nécessitant des ressources importantes et une régularité dans leur mise en œuvre.

Par ailleurs, une enquête menée en 2020 [37] souligne des lacunes dans la documentation scientifique des produits à base d'IA (destinés à un usage radiologique dans le cadre de cette étude). Seulement 18 sur 100 produits provenant de 54 fournisseurs ont des preuves d'efficacité potentielle de niveau 3 ou supérieur, et la plupart manquent de preuves de leur impact clinique réel. De plus, seulement 36 des 100 produits analysés disposent de preuves examinées par des pairs attestant de leur efficacité. Ces résultats mettent en évidence la nécessité d’établir à l’avenir une évaluation plus approfondie et rigoureuse de l'efficacité réelle des produits à base d'IA dans le domaine de la santé.

Nous venons donc de voir tout un panel d’exemples d’applications potentielles de l’IA pour la pratique clinique. Cependant, l’une des notion-clés qui revient au fil des différents angles d’approches vis-à-vis de l’IA est la coopération entre l’IA et l’humain pour obtenir de meilleurs résultats qu’avec un humain seul ou une IA seule.

Aujourd’hui, c’est en partie dû au fait que les performances de l’IA sont encore restreintes à des tâches spécifiques (par exemple, un algorithme capable de repérer les lésions périapicales sur des radiographies sera incapable de réaliser l’avulsion de ces mêmes dents, ou de diagnostiquer une pathologie de la muqueuse buccale, ou encore de réaliser un enseignement à l’hygiène bucco-dentaire auprès du patient, il est compétent pour la tâche pour laquelle il a été développé) et sujettes à des erreurs que le praticien doit surveiller, notamment à l’aide de la XAI. Il reste encore de nombreux progrès à réaliser avant qu’un système composite à même de posséder un champ de compétences aussi vaste qu’un professionnel de santé voit le jour, mais le jour où cela sera possible, peut-on imaginer un système capable de remplacer un professionnel de santé ?

Selon l’article de Shan et al. [2], la réponse est non : une machine serait incapable de remplacer « l’intuition clinique, la perception ineffable et l’empathie, qui sont essentielles

à la prestation de soins de santé personnalisés et au professionnalisme » et que la gestion de situations complexes et ambiguës, qui nécessite une approche humaine pour la relation et la discussion avec le patient, évaluer les objectifs et résultats esthétiques, résoudre une situation éthique,etc, nécessitent la compréhension et l'expertise d'un chirurgien-dentiste (ou d'un professionnel de santé à plus large échelle). Ainsi d'après Raman et al. [36] l'IA servirait d'élément auxiliaire, d'assistant, sans pouvoir remplacer le rôle du praticien dans le diagnostic et la gestion clinique des patients.

Cependant, par rapport à l'argument de l'humain, il ne faut pas oublier qu'il existe d'ors et déjà des algorithmes conversationnels capables d'avoir des rapports humains avec leurs utilisateurs et même de nouer des relations comme des amitiés avec eux, comme Xiaoice ou Replika. Cela questionne sur ce qui appartient réellement à l'humain ou non, on peut par exemple imaginer un algorithme qui se soucie plus du patient que certains professionnels de santé (l'approche de médecine centrée sur le patient a d'ailleurs été développée pour lutter contre cette déshumanisation et réduction du patient à un système liée à la sur-spécialisation des spécialités médicales). Le patient pourrait alors développer une meilleure relation de confiance avec ce modèle qu'avec son praticien. Également, une IA éthique ne serait pas soumise aux aléas inter et intra-personnelles qui pourraient mener un praticien à réaliser des actes contraires à l'éthique (allant d'un simple manque de professionnalisme à cause de fatigue personnelle à un acte répréhensible susceptible d'entraîner une sanction de la part de l'ordre). Mais qui dicterait cette éthique à l'IA ? Et n'y a-t-il qu'une seule éthique possible que toutes les IA devraient suivre ?
De plus, avec une standardisation des soins où l'humain est remplacé à 100 % vers l'IA, les prises en charges seraient uniformisées. Mon maître de stage m'a dit un jour qu'il était tout à fait souhaitable que chaque praticien soit différent dans son relationnel avec les patients car cela permet à chaque patient de trouver un praticien qui lui correspond, et cette idée que j'aime particulièrement serait perdue avec la répartition du marché sur un nombre limité de systèmes d'IA autonomes.
Et quid des cas uniques, ou presque uniques et non-référencés dans les jeux de données où le praticien doit s'adapter ou innover pour gérer une problématique qui sort des cadres connus, comme j'ai par exemple pu en voir en gériatrie durant mes années d'externat où l'on ne pouvait pas appliquer les mêmes protocoles qu'avec des patients lambdas.
Avec toutes ces questions que cela soulève, le remplacement complet du chirurgien-dentiste semble aujourd'hui et dans un futur proche une solution insatisfaisante. L'approche la plus prometteuse est, encore une fois, une approche coopérative entre

l’humain et la machine. De plus, en tant qu’humains nous faisons le choix d’utiliser les outils qui nous intéressent. À moins que les différences de performances ne soient réellement conséquentes, pourquoi choisir d’utiliser un système automatisé pour réaliser une avulsion au cabinet quand on apprécie de la réaliser soi-même ?

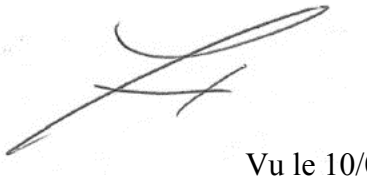
Synthèse et conclusion

Nous avons donc vu brièvement comment fonctionne l'IA et quelles sont les limitations de la boîte noire qu'elle génère. Malgré les performances prometteuses de l'IA, de par sa capacité de computation de données bien supérieure à celle d'un cerveau humain, l'incertitude que génère la boîte noire est une problématique qui se doit d'être solutionnée avant de pouvoir observer un usage pratique réel de l'IA en santé. Ceci est notamment dû au haut risque associé pour les patients en cas d'erreur, que l'on ne retrouve pas forcément dans d'autres domaines, et à toutes les questions éthiques qui découlent de l'usage d'une IA. Pour ce faire, la XAI est un outil au service de l'IA afin d'essayer de redonner de la transparence à cette boîte opaque. Nous avons donc vu une proposition de protocole élaborée afin d'essayer d'évaluer l'impact de la XAI sur les chirurgiens dentistes suite à une observation quant au manque d'études qualitatives à ce sujet. Nous avons également vu ce que les praticiens en santé attendent de la XAI, à savoir un outil de confiance et de contrôle des IA cliniques qui demeure ergonomique et ne génère pas une mobilisation supplémentaire trop conséquente d'attention et de ressources cognitives, sources de surcharge mentale chez les praticiens pouvant entraîner l'abandon de ces outils. D'un point de vue juridique, de nombreuses questions restent en suspens suite à l'apparition relativement récente de l'IA dans une pratique clinique concrète, en dehors du cadre expérimental. Se posent notamment les questions de degré d'autonomie de l'IA, de consentement du patient et de responsabilité en cas d'erreur. Cependant les autorités compétentes en Europe et aux États-Unis semblent alertes à ce sujet et des évolutions législatives adéquates pour encadrer l'utilisation de l'IA paraissent imminentes, quand les organismes se seront mis d'accord sur des définitions précises et claires de certains concepts. En attendant, du côté clinique, l'IA propose déjà de nombreuses promesses d'application à même d'apporter des changements dans l'exercice odontologique, que ce soit dans l'amélioration dans la qualité des soins mais également dans leur dispensation d'une manière plus efficiente et favorable à un meilleur accès aux soins dans la population. Cependant, de nombreux progrès restent encore à accomplir qui passent, par exemple, par l'enregistrement de jeux de données qualitatifs en santé pour gagner en fiabilité, performances et perdre en biais dans la formation des algorithmes. De plus, à travers son évolution rapide et nombre d'œuvres de science-fiction qui pointent, à raison, du doigt les failles éthiques qu'elle peut générer, l'arrivée de l'IA au cabinet peut s'accompagner de septicisme et de peurs, auxquelles la XAI peut apporter une solution au moins partielle. Par

exemple, nombre de chercheurs et de praticiens semblent s'accorder pour dire que l'explication des performances d'une IA est plus importante que ses performances en elles-mêmes car cela permet une meilleure synergie entre l'homme et la machine par la compréhension des limites et des domaines de compétence de l'algorithme et par une amélioration de la confiance entre les deux. Il est cependant à noter que tout dépendra de ce que l'on en fait et les observations sur l'évolution de l'IA dans notre quotidien de soignants dans les années et décennies à venir promettent d'être intéressantes. Qui sait, peut-être que si Marvin, personnage robotique animé par une IA, de Douglas Adams dans *Le Guide du voyageur galactique*, avait eu à sa disposition la XAI pour se psychanalyser, son existence en aurait été adoucie ?

Vu, le Président du jury et Directeur de thèse

Pr Paul MONSARRAT :



Vu le 10/01/2024

Annexe 1 :

Applications de l'IA dans le domaine de la dentisterie en 2021 présentées dans l'article de Shan & al[2]. Ces tableaux sont directement issus de l'article.

Table 1. Dental Applications of AI in Diagnosis.

Field								
Diagnosed Disease	Clinical or Experimental Data	Sample Size, Training: Testing	Validation Method	AI Methods	Accuracy	Sensitivity	Specificity	AUC
Oral and maxillofacial surgery								
Oral cancer	Fourier-transform infrared spectra of salivary exosomes (Zlotogorski-Hurvitz et al. 2019*)	34: —	k-fold cross-validation	PCA and LDA	95%	100%	89%	
				SVM	89%			
	Hyperspectral images (Jeyaraj and Samuel Nadar 2019*)	100: —	7-fold cross-validation	CNN	91.4%	94%	91%	
	X-ray images (Al-Ma'aitah and AlZubi 2018*)		Mean square error rate	GSOESNN	99.2%	>95%	>95%	
Oral squamous cell carcinoma	Laser endomicroscopy images (Aubreville et al. 2017*)	7,894 augmented 2 times: —	Leave-1-patient-out cross-validation	CNN	88.3%	86.6%	90%	0.955
				LBP-based RF	81.4%	84.7%	78.2%	0.895
				GLCM-based RF	73.1%	77.5%	69.5%	0.807
	Oral tissue histopathologic slides (Das et al. 2018*)	80: 20		CNN and RF	96.9% (keratin pearl)			
	Brush cytology specimens (McRae et al. 2020*)			k-nearest neighbors	96.9%			
	Cytology images (Sunny et al. 2019)	11,981: —		Pretrained ANN (inception V3)		73% (high potential malignant), 93% (malignant)		
Cervical lymph node metastasis	CT images (Ariji et al. 2019*)	441 augmented to 21,362: —	5-fold cross-validation	Pretrained CNN (Alex Net)	78.20%	75.4%	81.0%	0.8
Ameloblastoma, keratocystic odontogenic tumor	Panoramic images (Poedjiastoeti and Suebnukarn 2018*)	400 augmented to 800: 100 augmented to 200		Pretrained CNN (VGG-16)	83%	81.8%	83.3%	0.88
Dentigerous cysts, keratocystic odontogenic tumor and periapical cysts	Panoramic radiography (Lee et al. 2020*)	648 augmented 100 times: 228		Pretrained CNN (GoogLeNet)		96.1%	77.1%	0.914
	CBCT images (Lee et al. 2020*)	592 augmented 100 times: 197		Pretrained CNN (GoogLeNet)		88.2%	77.0%	0.847
Periapical cyst and keratocystic odontogenic tumor	CBCT images (Yilmaz et al. 2017)		10-fold cross-validation	SVM	94%			
				ANN	92%			
				Naïve Baves	Inferior to ANN			
Cariology and endodontics								
Dental caries	Periapical radiographic images of premolar and molar (Lee et al. 2018*)	2,400: 600		Pretrained CNN (GoogLeNet)	82%	81%	83%	0.845
	Near-infrared transillumination images (Casalegno et al. 2019)	185: 32	Monte Carlo cross-validation	CNN				0.836 (occlusal), 0.856 (proximal)
	Near-infrared-light transillumination images (Schwendicke et al. 2020*)	226 with augmentation: —	10-fold cross-validation	Pretrained CNN (Resnet18)	69%	85%	46%	0.73
				Pretrained CNN (Resnext50)	68%	76%	59%	0.74
Root fracture	Panoramic images (Fukuda et al. 2019*)	240: 60		Pretrained CNN (DetectNet)	93%			
	Periapical radiographs (Johari et al. 2017*)	180: 60		PNN	70%	93.3%	63.6%	
		120: 120		PNN	65%	81.7%	61.3%	
		60: 180		PNN	63.9%	88.9%	59.3%	
	CBCT images (Johari et al. 2017*)	180: 60		PNN	96.7%	93.3%	100%	
		120: 120		PNN	95.0%	90%	100%	
		60: 180		PNN	93.3%	86.7%	100%	
Periapical pathosis	Panoramic radiographs (Ekert et al. 2019*)	2,001 teeth from 85 panoramic images: —	10-fold repetition	CNN		65%	87%	0.84
	CBCT images (Setzer et al. 2020)	20: —	5-fold cross-validation	Pretrained CNN (U-net)	93%	93%	88%	
	CBCT images (Orhan et al. 2020)	3,900: 153		CNN		0.89 estimated recall		
Cracked dental root, caries, hypodontia and bone resorption	X-ray images (Ngan et al. 2016*)			Fuzzy aggregation operators	93.0%			
Periodontics								
Gingivitis and periodontitis	Risk factors, clinical periodontal parameters (Ozden et al. 2015*)	100: 50		SVM	98% precision			
				DT	98% precision			
				BPNN	46% precision			
Periodontally compromised teeth	Periapical radiographs (Lee et al. 2018)	1,044 augmented to 104,400: 348		CNN	82.8% (premolar)			0.826 (premolar)
					73.4% (molar)			0.734 (molar)

Table 1. (continued)

Diagnosed Disease	Field: Clinical or Experimental Data	Sample Size, Training: Testing	Validation Method	AI Methods	Accuracy	Sensitivity	Specificity	AUC
Aggressive and chronic periodontitis	Clinical and immunologic data sets (Papantonopoulos et al. 2014*)		10-fold cross-validation	MLP neural network	90% to 98%			
	Microbial profiles (Feres et al. 2018)	2,740: 1,175		SVM		86%	79%	0.83
	TMDs							
TMD-mimicking conditions	Mouth opening size and the frequency of word usage in chief complaint (Nam et al. 2018)	29 (TMD-mimicking cases), 290 (TMD): —	10-fold cross-validation	Text-mining method	96.6%	69.0%	99.3%	
Bone changes and disc displacement	Magnetic resonance images (Iwasaki 2015*)	590: —	Resubstitution validation	BBN with path condition	99.8%			
			10-fold cross-validation	BBN with path condition	99.5%			
Deformation of condyles with osteoarthritis	CBCT images (Shoukri et al. 2019)	259: 34		ANN	97.1%			
	Orthodontics							
Cervical vertebral maturation	Measurement data from lateral cephalometric radiographs (Kok et al. 2019)	300: —	5-fold cross-validation	ANN	93.0% (CVS 1)			
				k-nearest neighbors	78.7% (CVS 1)			
				Naïve Bayes	92.1% (CVS 1)			
				DT	97.1% (CVS 1)			
				SVM	84.8% (CVS 1)			
				RF	91.8% (CVS 1)			
	Measurement data from lateral cephalometric radiographs (Amasya et al. 2020)	498: 149		ANN	0.926 (κ value)			
				SVM	0.874 (κ value)			
				RF	0.908 (κ value)			
				DT	0.921 (κ value)			
Automated cephalometric analysis	Lateral cephalometric radiographs (Kunz et al. 2020*)	96.6% of 1,792 samples after augmentation: —		CNN	High correlation with humans ($r > 0.864$)			
	CBCT images (Gupta et al. 2015)	—: 30		A knowledge-based algorithm	2.01 ± 1.23 mm (mean error)			
	Others							
Oral lichen planus	Clinical data, normalized expression of interleukin-12 receptor β2 and tumor necrosis factor receptor superfamily member 8 (Jeon et al. 2015)	81: —		SVM	0.76	0.77	0.75	0.87
				RF	0.86	0.90	0.80	0.92
				ANN	0.78	0.79	0.77	0.83
				LDA	0.82	0.87	0.75	0.88
				Naïve Bayes	0.80	0.77	0.85	0.87
Sjögren's syndrome	Ultrasonography images of parotid gland (Kise et al. 2020)	160 augmented to 8,000: 40	5-fold cross-validation	Pretrained CNN (VGG16)	89.5%	90.0%	89.0%	0.948
	Ultrasonography images of submandibular gland (Kise et al. 2020)	160 augmented to 8,000: 40	5-fold cross-validation	Pretrained CNN (VGG16)	84.0%	81.0%	87.0%	0.894

Table 2. Dental Applications of AI in Treatment of Diseases.

Fields						
Aim	Function	Data	Sample Size, Training: Testing	Assessment	AI Method	Results
Oral and maxillofacial surgery						
Provide anatomic guidance	Distinguish interdigitated tongue muscles (Ye et al. 2015*)	Limited diffusion MRI			Bayesian approach	
	Segmentate the mandibular canal (Gerlach et al. 2014)	CBCT images	13: —	Compared with histologic sections of the same region	ASM	The difference ranged from −3.45 to 3.27 mm
					AAM	The difference ranged from −4.44 to 4.44 mm
	Segmentate parotid gland (Yang et al. 2014*)	MRI		Compared with the physicians' manual contours	SVM combined with Atlas registration	The average volume differences were 7.98% (left) and 8.12% (right)
	Classifying different tissue types (Engelhardt et al. 2014*)	Diffuse reflected spectra of laser scalpel	1,000: 1,000	Calculated misclassification rate	*k*-nearest neighbors	0.26 misclassification rate
					ANN	0.98 misclassification rate
					LDA	0.34 misclassification rate
					QDA	0.27 misclassification rate
					PDA	0.02 misclassification rate
					Random forests	0.34 misclassification rate
					Classification and regression trees	0.50 misclassification rate
Improve speech intelligibility of patients	Voice conversion (Fu et al. 2017)	Sentences uttered by oral surgical patients and target speakers	70: 40	Compared with conventional exemplar-based NMF method	Joint dictionary learning based non-NMF algorithm	Higher short-time objective intelligibility scores in the proposed algorithm
Prosthodontics						
Classify tooth	Classify dental cusps (Raith et al. 2017)	3D surface scans of dental casts	119: 10		ANN	93.3% accuracy (cusp distance method), 93.5% accuracy (range image method)
	Classify and number tooth (Tuzoff et al. 2019*)	Panoramic radiographs	1,352: 222	Compared with experts	Pretrained CNN (VGG-16)	0.9941 sensitivity and 0.9945 precision
	Detect and classify tooth (Zhang et al. 2018*)	Periapical radiographs	700: 200		Pretrained CNN (VGG-16)	95.8% accuracy and 96.1% recall in total
	Classify enamel, dentin, and pulp layer (Wang et al. 2017*)	Micro–computed tomography data sets			*k*-means++	0.83, 0.85, and 0.77 accuracy in classifying enamel, dentin, and pulp
Optimize conditions	Correlate resin composite hardness with light conditions (Deniz Arısu et al. 2018*)	Processing parameters and corresponding results			ANN	The mean square error value of the model was 0.0373
	Predict color outcomes of porcelain powder (Li et al. 2015*)	Parameters from spectrophotometer colorimetric instrument	75% of 119 sets of data: 25% of 119 sets of data		Back propagation neural network with genetic algorithm	Smaller mean square error than back propagation neural network
Resin restoration removal and tooth preparation	Discriminate tooth and restorative materials (Zakeri et al. 2015*)	Cutting sounds of an air-turbine handpiece			SVM	89% accuracy for resin composite, 92% accuracy for amalgam
	Achieve tooth crown preparation (Wang et al. 2014)	Shape of the target tooth acquired by a laser scanner		Matched with expected tooth preparation	Automatic laser ablation system	Wax resin: 0.05-mm linear error, 4.33° angel error, and 0.1-mm depth error. Dentin: 0.06-mm linear error, 0.5° angel error, and 0.1-mm depth error
	Achieve tooth preparation for porcelain laminate veneers (Otani et al. 2015*)	Image of tooth models scanned with 3D laser scanner	10: —	Compared with freehand tooth preparation	A robotic arm	The mean absolute deviation was 0.112 mm in the control group and 0.133 mm in the experimental group
Assist designing RPDs	Produce RPD designs according to the most similar cases (Chen et al. 2016)	Data about of oral conditions	104: —	Compared with professionals	An ontology-driven, case-based clinical decision support model	The mean average of precision was 0.61; AUC = 0.96; normalized discounted cumulative gain = 0.74
Measure masticatory efficiency	Identify patterns of masticated chewing gums (Vaccaro et al. 2018*)	Images of chewing gums obtained by flatbed scanner	400: —		Expert system	Mathews correlation coefficient score = 0.97

Table 2. (continued)

Fields						
Aim	Function	Data	Sample Size, Training: Testing	Assessment	AI Method	Results
Orthodontics						
Assist therapeutic decisions	Evaluate the orthodontic treatment needs (Thanathornwong 2018)	Patients' oral examination data sets	800: 200	Compared with 2 orthodontists on the newly recruited patients	Bayesian network	A high degree of agreement between the system and orthodontists (kappa value was 1.00 and 0.894)
	Identify pretreatment patients from posttreatment and normal individuals (Wang et al. 2016)	The eye-tracking data of observers	440: —	Leave-1-out cross-validation	SVM	97.2% and 93.4% accuracy for classifying pretreatment patient from posttreatment patient and normal people, respectively
Evaluate aesthetic outcome	Assess facial attractiveness and apparent age (Patcas et al. 2019)	Pre- and posttreatment photographs	469: —		Pretrained CNN (VGG-16)	Facial attractiveness was scaled from 0 to 100; apparent age was compared with real age
	Assess facial attractiveness (Patcas et al. 2019*)	Frontal and left-side profile images		Compared with lay people, orthodontists, and oral surgeons	Pretrained CNN (VGG-16)	No significant differences with regard to the evaluation of attractiveness
Endodontics						
Produce anatomic guidance	Identify root morphology (Hiraiwa et al. 2019)	Panoramic images	22,476: 64	Examined on CBCT images	Pretrained CNN (AlexNet)	87.4% accuracy, 77.3% sensitivity, 97.1% specificity
					Pretrained CNN (GoogleNet)	85.3% accuracy, 74.2% sensitivity, 95.9% specificity
	Measure root canal curvature (Christodoulou et al. 2018*)	CBCT images	30: —		An algorithm for the 3D measurement	
Optimize conditions	Predict the viability of stem cells (Bindal et al. 2017)	Processing variables and cell viability			Fuzzy inference system with ANN	The root mean square error was 0.028855. The coefficient of determination was 0.8111
Dental implantology						
Classify implant types	Identify 4 types of implant fixtures (Kim et al. 2020*)	Periapical radiographs	60% of 801 images: 20% of 801 images		Pretrained CNN (MobileNet-v2)	97% accuracy, 96% precision
					Pretrained CNN (ResNet-50)	98% accuracy, 98% precision
					Pretrained CNN (ResNet-18)	98% accuracy, 98% precision
					Pretrained CNN (GoogleNet)	93% accuracy, 92% precision
					Pretrained CNN (SqueezeNet)	96% accuracy, 96% precision

Table 3. Dental Applications of AI in Disease Prediction.

Field		Sample Size, Training: Testing	Validation Method	AI Methods	Accuracy	Sensitivity	Specificity	AUC
Prediction	Data							
Oral and maxillofacial surgery								
Survivability of patients with oral cavity squamous carcinoma	5 factors related to quantitative nuclear histomorphometric features of tissue sections (Lu et al. 2017*)	50: 65		Wilcoxon rank sum test and quadratic discriminant analysis	70.8%	61.5%	73.1%	0.72
Nodal metastasis in early oral cavity squamous cell carcinoma	5 factors pertaining to clinical and pathologic features of patients (Bur et al. 2019)	1,570: 391	5-fold cross-validation	Decision forest		0.753	0.492	0.712
				Kernel SVM		0.649	0.636	0.698
				Gradient boosting		0.773	0.492	0.704
Stage, type, and survivability of oral cancer	12 attributes regarding clinical details, personal history, and habits of patients (Sharma and Om 2015*)	1,025	Leave-1-out method	PNN and general regression neural networks		Benign and malignant		
					70.0%	91.0%	55.7%	0.7491
						Survivability		
					67.0%	91.0%	55.7%	0.7491
Locoregional recurrence of oral tongue squamous cell carcinoma	11 features pertaining to clinical, pathologic, and treatment features of patients (Alabi et al. 2020)	50% of 254 cases: 59	5-fold cross-validation	SVM	68%	0.84	0.60	
				Boosted DT	81%	0.79	0.83	
				Decision forest	78%	0.79	0.78	
				Naïve Bayes	70%	0.84	0.63	
3-y survival rate of patients with oral cancer	10 features selected from 17 features pertaining to patients (Tan et al. 2016*)	31: —	5-fold cross-validation	Genetic programming	83.9%			0.8341
				SVM	64.8%			0.5000
				Logistic regression	64.5%			0.5000
Occurrence of bisphosphonate-related osteonecrosis after dental extraction	9 factors related to medical history, tooth conditions (Kim et al. 2018*)	70%: 30% of the 125 patients		RF		100%	83.3%	0.973
				ANN		100%	76.7%	0.915
				SVM		81.8%	86.7%	0.882
				DT		90.9%	79.0%	0.821
				Logistic regression		90.9%	70.0%	0.844
Facial swelling after impacted mandibular third molars extraction	15 factors related to patients, the third molars, bone, and surgical conditions (Zhang et al. 2018)	300: 100	5-fold cross-validation	ANN based on improved conjugate grads BP algorithm	98.00%			
Cariology and endodontics								
Root caries	15 factors related to personal, nutrition, lifestyle, and clinical factors (Hung et al. 2019)	7,272: 1,818		SVM	97.1%	99.6%	94.3%	0.997
				XGBoost	94.7%	100.0%	88.9%	0.987
				RF	94.1%	100.0%	87.5%	0.999
				k-nearest neighbors	83.2%	97.1%	67.9%	0.881
				Logistic regression	74.2%	77.1%	71.1%	0.818
Tooth surface loss index	11 factors related to patients and teeth (Al Haidan et al. 2014*)	81: 15		ANN	73.3%, ±5 scores			
Gene expression of radicular cyst and periapical granuloma	Times of the gene associated to the diseases in the PubMed database (Poswar et al. 2015*)			MLP neural network				
Difficulty of endodontic cases	83 features from AAE case difficulty assessment form (Mallishery et al. 2020)		10-fold cross-validation	SVM	94.8%	95.0%	94.6%	
				Deep neural network	93.4%	93.0%	93.8%	
Dental implantology								
Failure of dental implants	20 attributes pertaining to patients and surgeon (Liu et al. 2018*)	747: —	10-fold cross-validation	DT	0.679	0.590	0.768	0.670
				SVM	0.628	0.581	0.675	0.628
				Logistic regressions	0.624	0.607	0.641	0.644
Individual implant mean bone levels	6 factors related to demographic and clinical features (Papantonopoulos et al. 2017)	237: —	10-fold resampling cross-validation	Ensemble selection		55%	91%	
	Age, tooth number, and implant surface (Papantonopoulos et al. 2017)	237	10-fold resampling cross-validation	SVM with Particle Swarm Optimization		62%	85%	
Successive rate of implant treatment	Dental data, patient self-behavior, health condition, and attitude (Alarifi and AlZubi 2018)			Memetic search optimization with genetic scale RNN	99.2%	97.6%	98.3%	
Periodontics								
Presence of oral malodor	16s ribosomal RNA sequence from microbiota in saliva (Nakano et al. 2018*)	90: —	Leave-1-out cross-validation	DL	96.7%	100%	93.3%	
				SVM	78.9%	77.8%	80.0%	
Tooth mobility	9 factors related to patients (Yoon et al. 2018)		10-fold cross-validation	MLP and deep neural networks	88.4%			0.72
Orthodontics								
Perioperative blood loss in orthognathic surgery	Demographic, clinical and surgical data (Stehrer et al. 2019*)	760: 190		RF	7.4-mL mean difference			

Table 3. (continued)

Field								
Prediction	Data	Sample Size, Training: Testing	Validation Method	AI Methods	Accuracy	Sensitivity	Specificity	AUC
Prosthodontics								
Longevity of dental restorations	Attributes related to patients and teeth (Aliaga et al. 2015)	4,336: 1,714	Leave-1-out validation	Bayesian network and MLP neural network	Error: 0.42 y (composites), 0.21 (amalgam)			
Debonding probability of CAD/CAM composite resin crowns	2D images captured from 3D stereolithography (Yamaguchi et al. 2019)	6,480: 2,160		CNN	98.5%			0.998
Facial deformation after complete denture prosthesis	3D pre- and postoperative models (Cheng et al. 2015*)	43: 5		BP neural network				

BIBLIOGRAHIE

[1] Salih, Ahmed, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Gloria Menegaz, et Karim Lekadir. « Commentary on explainable artificial intelligence methods: SHAP and LIME ». arXiv, 8 mai 2023. https://doi.org/10.48550/arXiv.2305.02012.

[2] Shan, T., F. R. Tay, et L. Gu. « Application of Artificial Intelligence in Dentistry ». *Journal of Dental Research* 100, n° 3 (mars 2021): 232-44. https://doi.org/10.1177/0022034520969115.

[3] Liu, Peng-Ran, Lin Lu, Jia-Yao Zhang, Tong-Tong Huo, Song-Xiang Liu, et Zhe-Wei Ye. « Application of Artificial Intelligence in Medicine: An Overview ». *Current Medical Science* 41, n° 6 (décembre 2021): 1105-15. https://doi.org/10.1007/s11596-021-2474-3.

[4] Kaul, Vivek, Sarah Enslin, et Seth A. Gross. « History of Artificial Intelligence in Medicine ». *Gastrointestinal Endoscopy* 92, n° 4 (octobre 2020): 807-12. https://doi.org/10.1016/j.gie.2020.06.040.

[5] Choi, Rene Y., Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang, et J. Peter Campbell. « Introduction to Machine Learning, Neural Networks, and Deep Learning ». *Translational Vision Science & Technology* 9, n° 2 (s. d.): 14. https://doi.org/10.1167/tvst.9.2.14.

[6] Shah, Tarang. « About Train, Validation and Test Sets in Machine Learning ». Medium, 10 juillet 2020. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7.

[7] Margot, Vincent. « A Brief Overview of Methods to Explain AI (XAI) ». Medium, 15 mars 2022. https://towardsdatascience.com/a-brief-overview-of-methods-to-explain-ai-xai-fe0d2a7b05d6.

[8] Bhattacharya, Aditya. *Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More*. Packt Publishing Ltd, 2022.

[9] « What Is Global, Cohort and Local Explainability? | Censius AI Observability Blog ». Consulté le 18 juillet 2023. https://censius.ai/blogs/global-local-cohort-explainability.

[10] Sheu, Ruey-Kai, et Mayuresh Sunil Pardeshi. « A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System ». *Sensors* 22, n° 20 (janvier 2022): 8068. https://doi.org/10.3390/s22208068.

[11] Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, et Himabindu Lakkaraju. « Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods ». In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180-86. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3375627.3375830.

[12] Duell, Jamie, Xiuyi Fan, Bruce Burnett, Gert Aarts, et Shang-Ming Zhou. « A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records ». In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 1-4, 2021. https://doi.org/10.1109/BHI50953.2021.9508618.

[13] Roberts, Claudia V., Ehtsham Elahi, et Ashok Chandrashekar. « On the Bias-Variance Characteristics of LIME and SHAP in High Sparsity Movie Recommendation Explanation Tasks ». arXiv, 9 juin 2022. https://doi.org/10.48550/arXiv.2206.04784.

[14] « Applied Sciences | Free Full-Text | Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP ». Consulté le 18 septembre 2023. https://www.mdpi.com/2076-3417/13/5/3125.

[15] « Diagnostics | Free Full-Text | Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP ». Consulté le 18 septembre 2023. https://www.mdpi.com/2075-4418/13/11/1932.

[16] Blaivas, Michael, Srikar Adhikari, Eric A. Savitsky, Laura N. Blaivas, et Yiju T. Liu. « Artificial Intelligence versus Expert: A Comparison of Rapid Visual Inferior Vena Cava Collapsibility Assessment between POCUS Experts and a Deep Learning Algorithm ». *Journal of the American College of Emergency Physicians Open* 1, n° 5 (2020): 857-64. https://doi.org/10.1002/emp2.12206.

[17] Fourcade, A., et R. H. Khonsari. « Deep learning in medical image analysis: A third eye for doctors ». *Journal of Stomatology, Oral and Maxillofacial Surgery*, 55th SFSCMFCO Congress, 120, n° 4 (1 septembre 2019): 279-88. https://doi.org/10.1016/j.jormas.2019.06.002.

[18] Kolbinger, Fiona R., Franziska M. Rinner, Alexander C. Jenke, Matthias Carstens, Stefanie Krell, Stefan Leger, Marius Distler, Jürgen Weitz, Stefanie Speidel, et Sebastian Bodenstedt. « Anatomy Segmentation in Laparoscopic Surgery: Comparison of Machine Learning and Human Expertise – An Experimental Study ». medRxiv, 26 juin 2023. https://doi.org/10.1101/2022.11.11.22282215.

[19] McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. « International Evaluation of an AI System for Breast Cancer Screening ». *Nature* 577, n° 7788 (janvier 2020): 89-94. https://doi.org/10.1038/s41586-019-1799-6.

[20] Rajpurkar, Pranav, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, et al. « Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists ». *PLOS Medicine* 15, n° 11 (20 novembre 2018): e1002686. https://doi.org/10.1371/journal.pmed.1002686.

[21] Rueckel, Johannes, Wolfgang G. Kunz, Boj F. Hoppe, Maximilian Patzig, Mike Notohamiprodjo, Felix G. Meinel, Clemens C. Cyran, Michael Ingrisch, Jens Ricke, et Bastian O. Sabel. « Artificial Intelligence Algorithm Detecting Lung Infection in Supine Chest Radiographs of Critically Ill Patients With a Diagnostic Accuracy Similar to Board-Certified Radiologists ». *Critical Care Medicine* 48, n° 7 (juillet 2020): e574. https://doi.org/10.1097/CCM.0000000000004397.

[22] Muddamsetty, Satya M., Mohammad N. S. Jahromi, et Thomas B. Moeslund. « Expert Level Evaluations for Explainable AI (XAI) Methods in the Medical Domain ». In *Pattern Recognition. ICPR International Workshops and Challenges*, édité par Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, et Roberto Vezzani, 35-46. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-68796-0_3.

[23] Muller, Heimo, Michaela Theresia Mayrhofer, Evert-Ben Van Veen, et Andreas Holzinger. « The Ten Commandments of Ethical Medical AI ». *Computer* 54, n° 7 (juillet 2021): 119-23. https://doi.org/10.1109/MC.2021.3074263.

[24] Mezrich, Jonathan L. « Is Artificial Intelligence (AI) a Pipe Dream? Why Legal Issues Present Significant Hurdles to AI Autonomy ». *American Journal of Roentgenology* 219, n° 1 (juillet 2022): 152-56. https://doi.org/10.2214/AJR.21.27224.

[25] Schneeberger, David, Karl Stöger, et Andreas Holzinger. « The European Legal Framework for Medical AI ». In *Machine Learning and Knowledge Extraction*, édité par Andreas Holzinger, Peter Kieseberg, A Min Tjoa, et Edgar Weippl, 209-26. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-57321-8_12.

[26] Becker, Christoph D., Elmar Kotter, Laure Fournier, Luis Martí-Bonmatí, et European Society of Radiology (ESR). « Current Practical Experience with Artificial Intelligence in Clinical Radiology: A Survey of the European Society of Radiology ». *Insights into Imaging* 13, n° 1 (21 juin 2022): 107. https://doi.org/10.1186/s13244-022-01247-y.

[27] Tonekaboni, Sana, Shalmali Joshi, Melissa D. McCradden, et Anna Goldenberg. « What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use ». In *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359-80. PMLR, 2019. https://proceedings.mlr.press/v106/tonekaboni19a.html.

[28] Leeuwen, Kicky G. van, Maarten de Rooij, Steven Schalekamp, Bram van Ginneken, et Matthieu J. C. M. Rutten. « How Does Artificial Intelligence in Radiology Improve Efficiency and Health Outcomes? » *Pediatric Radiology* 52, n° 11 (1 octobre 2022): 2087-93. https://doi.org/10.1007/s00247-021-05114-8.

[29] Bharati, Subrato, M. Rubaiyat Hossain Mondal, et Prajoy Podder. « A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? » *IEEE Transactions on Artificial Intelligence*, 2023, 1-15. https://doi.org/10.1109/TAI.2023.3266418.

[30] Glick, Aaron, Mackenzie Clayton, Nikola Angelov, et Jennifer Chang. « Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians ». *JAMIA Open* 5, n° 2 (1 juillet 2022): ooac031. https://doi.org/10.1093/jamiaopen/ooac031.

[31] European Society of Radiology (ESR) (2019) Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. Insights Imaging 10:105

[32] Kwee TC, Kwee RM (2021) Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence. Insights Imaging 12:88

[33] Allen B, Agarwal S, Coombs L, Wald C, Dreyer K (2021) 2020 ACR data science institute artificial intelligence survey. J Am Coll Radiol 18:1153–1159

[34] Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M.E., Linkohr, B., Peters, A., Heid, I.M., Palm, C., Weber, B.H.: A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology 125(9), 1410–1420 (2018)

[35] Nayak, J., Acharya, R., Bhat, P.S., S., N., Lim, T.: Automated diagnosis of glaucoma using digital fundus images. Journal of medical systems 33(5), 337 (2009)

[36] Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C., Rajalakshmi, R.: Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. Eye (11 2018). https://doi.org/10.1038/s41433-018-0269-y

[37] van Leeuwen KG, Schalekamp S, Rutten MJCM et al (2021) Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 31:3797–3804. https://doi.org/10.1007/s00330-021-07892-z

[38] The Royal College of Radiologists (2018) Clinical radiology UK workforce census report 2018. RCR website. https://www.rcr.ac. uk/publication/clinical-radiology-uk-workforce-census-report2018. Accessed 4 May 2021

[39] Hassan AE, Ringheanu VM, Rabah RR et al (2020) Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. Interv Neuroradiol 26:615–622

[40] Grunwald IQ, Ragoschke-Schumm A, Kettner M et al (2016) First automated stroke imaging evaluation via electronic Alberta stroke program early CT score in a mobile stroke unit. Cerebrovasc Dis 42:332–338

[41] Baltruschat I, Steinmeister L, Nickisch H et al (2021) Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. Eur Radiol 31:3837–3845. https://doi.org/10. 1007/s00330-020-07480-7 28

[42] O'Neill TJ, Xi Y, Stehel E et al (2021) Active reprioritization of the reading worklist using artificial intelligence has a beneficial effect on the turnaround time for interpretation of head CT with intracranial hemorrhage. Radiol Artif Intell 3:e200024

[43] Chong LR, Tsai KT, Lee LL et al (2020) Artificial intelligence predictive analytics in the management of outpatient MRI appointment no-shows. AJR Am J Roentgenol 215:1155–1162

[44] Kim JR, Shim WH, Yoon HM et al (2017) Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. AJR Am J of Roentgenol 209:1374–1380

[45] Bakker MF, de Lange SV, Pijnappel RM et al (2019) Supplemental MRI screening for women with extremely dense breast tissue. N Engl J Med 381:2091–2102

[46] Wolff J, Pauling J, Keck A, Baumbach J (2020) The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res 22:e16866

[47] Martin DD, Deusch D, Schweizer R et al (2009) Clinical application of automated Greulich-Pyle bone age determination in children with short stature. Pediatr Radiol 39:598–607

[48] Wang L, Wang D, Zhang Y, Ma L, Sun Y, Lv P. 2014. An automatic robotic system for three-dimensional tooth crown preparation using a picosecond laser. Lasers Surg Med. 46(7):573–581.

[49] Schwendicke F, Golla T, Dreher M, Krois J. 2019. Convolutional neural networks for dental image diagnostics: a scoping review. J Dent. 91:103226.

[50] Schwendicke F, Samek W, Krois J. 2020. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 99(7):769–774.

[51] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. 2018. Artificial intelligence in radiology. Nat Rev Cancer. 18(8):500–510.

[52] Yilmaz E, Kayikcioglu T, Kayipmaz S. 2017. Computer-aided diagnosis of periapical cyst and keratocystic odontogenic tumor on cone beam computed tomography. Computer Methods Programs Biomed. 146:91–100.

[53] Sunny S, Baby A, James BL, Balaji D, NV A, Rana MH, Gurpur P, Skandarajah A, D'Ambrosio M, Ramanjinappa RD, et al. 2019. A smart tele-cytology point-of-care platform for oral cancer screening. PLoS One. 14(11):e0224885.

[54] Jeon SH, Jeon EH, Lee JY, Kim YS, Yoon HJ, Hong SP, Lee JH. 2015. The potential of interleukin 12 receptor beta 2 (IL12RB2) and tumor necrosis factor receptor superfamily member 8 (TNFRSF8) gene as diagnostic biomarkers of oral lichen planus (OLP). Acta Odontol Scand. 73(8):588–594.

[55] Kise Y, Shimizu M, Ikeda H, Fujii T, Kuwada C, Nishiyama M, Funakoshi T, Ariji Y, Fujita H, Katsumata A, et al. 2020. Usefulness of a deep learning system for diagnosing Sjogren's syndrome using ultrasonography images. Dentomaxillofac Radiol. 49(3):20190348.

[56] Feres M, Louzoun Y, Haber S, Faveri M, Figueiredo LC, Levin L. 2018. Support vector machine-based differentiation between aggressive and chronic periodontitis using microbial profiles. Int Dent J. 68(1):39–46.

[57] Thanathornwong B. 2018. Bayesian-based decision support system for assessing the needs for orthodontic treatment. Healthc Inform Res. 24(1):22–28.

[58] Setzer FC, Shi KJ, Zhang Z, Yan H, Yoon H, Mupparapu M, Li J. 2020. Artificial intelligence for the computer-aided detection of periapical lesions in cone-beam computed tomographic images. J Endod. 46(7):987–993.

[59] Alarifi A, AlZubi AA. 2018. Memetic search optimization along with genetic scale recurrent neural network for predictive rate of implant treatment. J Med Syst. 42(11):202.

[60] Yamaguchi S, Lee C, Karaer O, Ban S, Mine A, Imazato S. 2019. Predicting the debonding of CAD/CAM composite resin crowns with AI. J Dent Res. 98(11):1234–1238.

[61] Zhang W, Li J, Li ZB, Li Z. 2018. Predicting postoperative facial swelling following impacted mandibular third molars extraction by using artificial neural networks evaluation. Sci Rep. 8(1):12281.

[62] Fu SW, Li PC, Lai YH, Yang CC, Hsieh LC, Tsao Y. 2017. Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery. IEEE Trans Biomed Eng. 64(11):2584–2594.

[63] Bindal P, Bindal U, Lin CW, Kasim NHA, Ramasamy T, Dabbagh A, Salwana E, Shamshirband S. 2017. Neuro-fuzzy method for predicting the viability of stem cells treated at different time-concentration conditions. Technol Health Care. 25(6):1041–1051.

[64] Patcas R, Bernini DAJ, Volokitin A, Agustsson E, Rothe R, Timofte R. 2019. Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age. Int J Oral Maxillofac Surg. 48(1):77–83.

[65] Hassan AE, Ringheanu VM, Rabah RR et al (2020) Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. Interv Neuroradiol 26:615–622

[66] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature medicine, vol. 25, no. 1, pp. 44-56, 2019.

[67] Montani S, Striani M. 2019. Artificial intelligence in clinical decision support: a focused literature survey. Yearb Med Inform. 28(1):120–127.

[68] Keskinbora KH. 2019. Medical ethics considerations on artificial intelligence. J Clin Neurosci. 64:277–282.

[69] Char DS, Shah NH, Magnus D. 2018. Implementing machine learning in health care —addressing ethical challenges. N Engl J Med. 378(11):981–983.

[70] Fiske A, Henningsen P, Buyx A. 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res. 21(5):e13216.

[71] Oreilly, M. and Parker, N. (2013). unsatisfactory saturation: a critical exploration of the notion of saturated sample sizes in qualitative research. Qualitative research, 13(2), 190–197.

[72] Hennink, M. M., Kaiser, B. N., and Marconi, V. C. (2017). Code saturation versus meaning saturation: how many interviews are enough? Qualitative health research, 27(4), 591608.

[73] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2018). Model cards for model reporting. arXiv preprint arXiv:1810.03993 .

[74] Umscheid, C. A., Betesh, J., VanZandbergen, C., Hanish, A., Tait, G., Mikkelsen, M. E., French, B., and Fuchs, B. D. (2015). Development, implementation, and impact of an automated early warning and response system for sepsis. Journal of hospital medicine.

[75] Guidi, J. L., Clark, K., Upton, M. T., Faust, H., Umscheid, C. A., Lane-Fall, M. B., Mikkelsen, M. E., Schweickert, W. D., Vanzandbergen, C. A., Betesh, J., et al. (2015). Clinician perception of the effectiveness of an automated early warning and response system for sepsis in an academic medical center. Annals of the American Thoracic Society, 12.

[76] Embi, P. J. and Leonard, A. C. (2012). Evaluating alert fatigue over time to ehr-based clinical trial alerts: findings from a randomized controlled study. Journal of the American Medical Informatics Association, 19(e1), e145–e148.

[77] Sun, J., Wang, F., Hu, J., and Edabollahi, S. (2012). Supervised patient similarity measure of heterogeneous patient records. ACM SIGKDD Explorations Newsletter , 14.

[78] Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., and Wang, F. (2016). Measuring patient similarities via a deep architecture with medical concept embedding. In 2016 IEEE 16th International Conference on Data Mining (ICDM).

[79] Sharafoddini, A., Dubin, J. A., and Lee, J. (2017). Patient similarity in prediction models based on health data: a scoping review. JMIR medical informatics, 5.

[80] Kendall, Maurice George. "Rank correlation methods." (1948).

[81] Bernard, David, Emmanuel Doumard, Isabelle Ader, Philippe Kemoun, Jean-Christophe Pagès, Anne Galinier, Sylvain Cussat-Blanc, et al. « Explainable Machine Learning Framework to Predict Personalized Physiological Aging ». *Aging Cell* 22, n^{o} 8 (2023): e13872. https://doi.org/10.1111/acel.13872.

[82] Emmanuel Doumard, Julien Aligon, Elodie Escriva, Jean-Baptiste Excoffier, Paul Monsarrat, et al.. A quantitative approach for the comparison of additive local explanation methods. *Information Systems*, 2023, 114 (Special issue on DOLAP 2022: Design, Optimization, Languages and Analytical Processing of Big Data), pp.102162.

MAY Julien 2024 TOU3 3003

PROPOSITION D'UNE MÉTHODOLOGIE D'ÉVALUATION DE L'INTÉRÊT DE LA XAI POUR LE CHIRURGIEN-DENTISTE

RÉSUMÉ EN FRANÇAIS :

Nous voyons ici brièvement qu'est-ce que sont l'intelligence artificielle et l'explicabilité avant de nous intéresser à la comparaison entre deux techniques d'explicabilité : SHAP et LIME. Nous proposons ensuite un protocole d'évaluation de l'intérêt de l'explicabilité pour le chirurgien-dentiste. Enfin, nous nous intéressons aux attentes des professionnels de santé ainsi qu'aux attentes éthiques et légales vis-à-vis de l'intelligence artificielle et de l'explicabilité avant de nous intéresser aux usages que nous pouvons en avoir dans le domaine de la santé et plus précisément en odontologie, ainsi qu'aux promesses sur leur avenir.

TITRE EN ANGLAIS : Proposal for a methodology for evaluating the interest of the XAI for the dentist

DISCIPLINE ADMINISTRATIVE : Chirurgie dentaire

MOTS-CLÉS : intelligence artificielle, IA, explicabilité, XAI, soin augmenté, SHAP, LIME, CAD, éthique, législation

INTITULE ET ADRESSE DE L'UFR :

Université Toulouse III – Paul Sabatier

Faculté de santé – Département d'Odontologie 3 chemin des Maraîchers 31062 Toulouse Cedex 9

Directeur de thèse : Pr Paul MONSARRAT