IA et Biologie Médicale – Cas Réel

SIL-LAB experts

1. Résumé du document

En 2019, l'Intelligence Artificielle (IA) est l'objet de nombreux discours alarmistes, notamment sous l'angle éthique et sur les risques que l'IA devienne plus puissante que l'être humain. Ce document développe une approche différente, et propose, de manière pragmatique, de tracer les premiers contours de l'utilisation de l'IA en Biologie Médicale.

Contrairement à de nombreux autres domaines médicaux, l'utilisation des Systèmes Experts pour assister les médecins, voire pour réaliser les diagnostics, est une réalité depuis plus de 15 ans dans la Biologie Médicale. Cette évolution a entraîné des débats, en particulier liés à l'accréditation ISO-15189, ces 10 dernières années. Aujourd'hui, le champ d'utilisation des Systèmes d'Aide à la Validation Biologique (SAVB), est défini de la manière suivante : « Les Systèmes d'Aide à la Validation Biologique sont utilisés par le laboratoire de Biologie Médicale sous la responsabilité du Biologiste après une phase d'évaluation et de validation documentée et traçable et une analyse de risque associée ». Ce principe peut très bien inclure des techniques d'Intelligence Artificielle, sous certaines conditions que nous allons détailler.

Le document aborde ensuite en détail une expérimentation réalisée dans le laboratoire de biologie médicale BIO 86 à Poitiers pour évaluer l'utilisation du Machine Learning dans un Système d'Aide à la Validation Biologique. Le document présente le retour d'expérience de cette évaluation, menée sur des données réelles de Biologie Médicale, en expliquant les étapes suivies, les différents problèmes rencontrés et les bénéfices identifiés. Une comparaison est faite entre les Systèmes d'Aide à la Validation Biologique actuels et ce que pourrait être un système intégrant le Machine Learning. L'expérimentation n'étant pas allée jusqu'à une mise en place réelle en routine, un parallèle est réalisé avec une application de l'IA dans un autre domaine, comparable sur certains aspects.

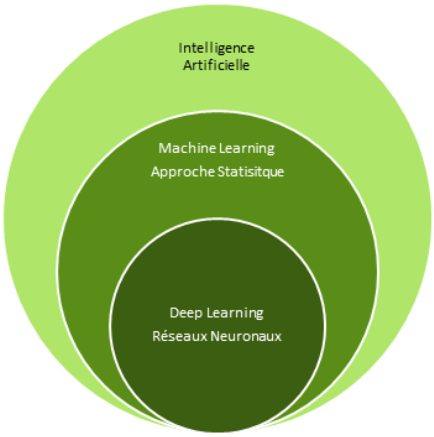
Dans une dernière partie de ce document, les premiers contours d'une méthodologie d'utilisation de l'IA en Biologie Médicale, combinant Machine Learning et Systèmes Experts, sont abordés.

2. Applicabilité de l'IA à la Biologie Médicale

Dans le rapport parlementaire porté par Cédric Villani « Donner du sens à l'Intelligence Artificielle : stratégie nationale et européenne »[1], la biologie est citée page 79 comme un domaine cible, et un chapitre complet est dédié à la santé.

2.1 Rappels sur les différents domaines de l'IA

L'Intelligence Artificielle est un ensemble de techniques qui permettent à la machine de mimer l'intelligence humaine. Un premier sous-ensemble de ces techniques est appelé « Machine Learning » et se définit par une approche statistique sur des ensembles de données qui permet à la machine de s'approcher progressivement du raisonnement humain avec l'expérience. Une technologie particulière du Machine Learning est appelée « Deep Learning » et utilise les réseaux neuronaux pour cet apprentissage.



Aujourd'hui, l'utilisation de réseaux neuronaux ne permet pas de comprendre le raisonnement effectué par la machine. Le Deep Learning ne permet pas d'avoir un système déterministe, reproductible et explicable. L'avènement du Deep Learning a permis de faire un bond remarquable dans l'analyse d'image et la reconnaissance vocale par les machines. Il nécessite une supervision préalable de l'être humain qui va lui soumettre des jeux de données avec les explications associées puis signaler les erreurs quand le raisonnement est erroné.

Dans l'analyse d'image, le Deep Learning a démontré sa capacité à être plus performant que l'être humain, en particulier quand les images qui lui sont fournies sont de très haute définition. La machine peut analyser l'ensemble des données sur un même plan, là où l'humain va devoir zoomer / dézoomer sur différentes zones, limité par la capacité de restitution de son œil et des écrans qui affichent les données.

Si cette performance est très intéressante, elle est aussi la preuve que pour l'instant, l'Intelligence Artificielle en est encore à apprendre, à voir et à entendre. Certains experts reconnus affirment, pour cette raison, que l'Intelligence Artificielle en est encore à l'étape d'apprentissage d'un enfant de 4 ou 5 ans avec des capacités de vue ou d'écoute exceptionnelles.

Pour une présentation plus approfondie des domaines de l'IA, se référer au rapport parlementaire cité ci-dessus.

2.2 Système d'Aide à la Validation Biologique

Le Système d'Aide à la Validation Biologique (SAVB) est un terme générique défini dans le SH-GTA-02 qui désigne tout système basé sur des règles permettant d'aider le Biologiste à réaliser sa validation biologique. Dans certains laboratoires et sous certaines conditions contrôlées par le Biologiste, ces systèmes peuvent fournir eux-mêmes la validation biologique, le Biologiste gardant la responsabilité des résultats émis.

Deux types de SAVB existent :

- Ceux fournis par des industriels incluant des règles d'expertise existantes en « boîte noire ». Il s'agit de Systèmes Experts dont les règles, définies au préalable, ne sont pas accessibles. Dans certains cas, les règles peuvent être complétées ou réécrites.
- Ceux permettant aux Biologistes et aux techniciens de paramétrer eux-mêmes leurs algorithmes et où aucune règle n'est fournie par défaut.

Dans les deux cas, aujourd'hui, le Machine Learning n'a pas été utilisé pour paramétrer ces outils et les algorithmes sont complètement déterministes. Quand on prend deux fois de suite les mêmes données, on obtient exactement les mêmes résultats. Le système n'a pas d'apprentissage intégré.

Dans le cas des règles en « boîte noire », le Biologiste est obligé de faire une longue phase de validation de son SAVB avant de le passer en mode validation automatique / supervisée. Le déterminisme du système est rassurant car il garantit une reproductibilité dans le temps. Néanmoins, le fait que le Biologiste ne puisse pas comprendre le raisonnement de l'algorithme, le vérifier et l'adapter à la population étudiée, constitue régulièrement un frein à l'utilisation de ces systèmes en routine.

2.3 Limite à l'utilisation de l'Intelligence Artificielle en Biologie Médicale

Avec l'utilisation des Systèmes d'Aide à la Validation Biologique déjà en place, le marché de la Biologie Médicale est très en avance sur l'utilisation d'algorithmes pour assister le professionnel de santé dans le diagnostic. Les Systèmes Experts apportent une aide précieuse au diagnostic, mais l'exécution des algorithmes doit se faire sous la responsabilité des professionnels de santé. Quant aux algorithmes « boîte noire », les fournisseurs doivent être en mesure d'apporter des explications sur leur fonctionnement. A ce titre, l'utilisation de Systèmes Experts dont les règles ne sont pas accessibles pose question. Comment les Biologistes qui utilisent ce type de système peuvent-ils s'assurer que les règles du Système Expert sont cohérentes avec les particularités de leurs patients ? En effet, de nombreux critères (par exemple : la région, la catégorie socio-professionnelle du patient, les caractéristiques populationnelles) peuvent justifier le fait d'avoir des règles spécifiques.

Il est donc important de s'assurer, dans les contrats passés avec les fournisseurs du Laboratoire de Biologie, que les systèmes n'utilisent que des algorithmes déterministes, explicables et reproductibles.

Par ailleurs, les approches de Machine Learning tardent à être adoptées dans le secteur. Le Machine Learning ayant souvent été présenté, à tort, comme un domaine hermétique, sa mise en œuvre suscite des inquiétudes dans le domaine de la Biologie Médicale.

La manière dont les algorithmes de validation biologique sont créés importe peu, du moment que leurs décisions sont explicables par le professionnel de santé ou par le fournisseur, et que ces algorithmes ont un fonctionnement déterministe et reproductible. Il est donc tout à fait envisageable d'utiliser des moteurs de Machine Learning statistiques pour élaborer des règles compréhensibles et les mettre en application en routine de manière déterministe.

Aujourd'hui le Deep Learning et l'utilisation de réseaux neuronaux créent des systèmes si complexes que le raisonnement de l'IA n'est pas explicable, ce que souligne le rapport parlementaire. Tant que des avancées sur l'explicabilité du raisonnement du Deep Learning ne seront pas réalisées, ce dernier ne pourra pas être utilisé dans le cadre de la Biologie Médicale.

3. Le Machine Learning appliqué à la Biologie Médicale

Afin de proposer une démarche méthodologique pour l'utilisation du Machine Learning en Biologie Médicale, nous avons opté pour une double approche :

- Réalisation d'une expérimentation au sein d'un groupe de laboratoires d'analyse médicale, avec des données réelles (anonymisées) de Biologie Médicale
- Comparaison avec le domaine de la sécurité financière, qui partage avec le domaine biomédical des problématiques similaires de mise en œuvre du Machine Learning

En combinant ces deux approches, nous avons élaboré une démarche méthodologique pour l'utilisation du Machine Learning dans le cadre de l'aide à la validation biologique. Nous avons conçu cette méthodologie dans le but de faciliter la mise en œuvre des futurs projets d'utilisation du Machine Learning au sein des laboratoires de Biologie Médicale.

3.1 Evaluation en réel dans le contexte de la Biologie Médicale

Dans ce paragraphe, nous allons présenter un retour d'expérience sur la réalisation du prototype d'utilisation du Machine Learning dans un environnement réel.

3.1.1 Expérimentation sur des données réelles

Pour concevoir une démarche méthodologique d'utilisation du Machine Learning en Biologie Médicale, l'équipe projet a décidé de mener une expérimentation dans un contexte réel.

L'objectif principal était de concevoir un prototype de système d'aide à la validation biologique basé sur le Machine Learning. Contrairement aux Systèmes Experts, le Machine Learning ne repose pas sur des règles déterminées au préalable. La spécificité du Machine Learning est de proposer des règles de décision, en menant une analyse statistique sur de grandes volumétries de données. Le Machine Learning est ainsi capable, sur les données, de mettre en évidence des relations inédites, que l'être humain n'avait jamais identifiées. Une innovation prometteuse dans le cadre de la Biologie Médicale, où le Machine Learning pourrait, à l'avenir, permettre de détecter en amont les premiers symptômes d'une pathologie.

Le second objectif était de comparer les résultats proposés par le prototype de Machine Learning à ceux obtenus avec un Système Expert classique. L'expérimentation a été menée avec des données anonymisées du Laboratoire BIO 86 à Poitiers. Ce laboratoire utilisait depuis quelques mois un Système Expert au quotidien pour le traitement des données biomédicales. Il a donc été possible de comparer les résultats respectifs du prototype et du Système Expert existant.

Le Laboratoire BIO 86 est un laboratoire composé de 12 sites, qui traite environ 3000 dossiers par jour, soit près de 25 millions de résultats annuels. Le Machine Learning a besoin d'un maximum de données biologiques pour pouvoir identifier puis proposer des règles. Le défi, dans le domaine de la Biologie Médicale, est de disposer d'un grand historique, notamment pour pouvoir étudier des cas de pathologie extrêmement rares. Pour cette raison, dans le cadre de l'expérimentation, 5 années d'historique d'activité du laboratoire ont été utilisées.

Les données biomédicales sont à la fois très volumineuses et très complexes. Ces deux éléments sont à prendre en compte dans le choix d'un outil de Machine Learning. Il est indispensable de s'assurer que les outils ont une technologie capable de gérer la volumétrie et la complexité des données. Préalablement à la phase de prototype, plusieurs solutions de Machine Learning ont été examinées. Les outils ont été évalués sur trois critères :

- Capacité à intégrer et stocker de grandes volumétries de données.
- Capacité à réaliser des traitements et des statistiques sur ces volumétries.
- Capacité à faire analyser des données complexes par un algorithme de Machine Learning sur de grandes volumétries.

Le logiciel Amadea, de la société ISoft, a été retenu pour l'expérimentation. Le moteur de traitement de données d'Amadea se distingue par sa capacité à intégrer et analyser les données en temps réel, quelles que soient leur complexité et leur volumétrie.

Pour alimenter Amadea, de nombreux types de connecteurs de données sont disponibles. Dans le cadre de l'expérimentation, un simple connecteur ODBC a été utilisé.

BIO 86 utilise le système de gestion de laboratoire Odancio, qui dispose lui-même d'un module de Business Intelligence (BI) assez ouvert, maîtrisé par SIL-LAB Experts. Le travail déjà réalisé dans le cadre de ce module a été réutilisé pour le prototype.

Amadea a aussi été sélectionné car il dispose d'un module de Machine Learning capable de proposer des règles sur les données analysées. Amadea permet donc de répondre à l'ensemble des besoins de l'expérimentation avec un seul et même outil intégré.

3.1.2 Standardisation des données

Après la phase d'intégration, nous avons vérifié la qualité des données en vérifiant la cohérence de chaque source utilisée. Nous avons rapidement identifié des ajustements nécessaires pour standardiser les données.

Deux cas de figure ont été identifiés :

Changements de paramétrage dans le temps : sur les 5 dernières années, le laboratoire a régulièrement changé de machines. Certains automates réalisaient peu de volume, d'autres en produisaient beaucoup plus. Sur une période de 5 ans, certains paramétrages ont changé[2]. Une rupture peut alors se produire dans l'interprétation historique des données. Il faut tenir compte de ces variations au moment d'analyser l'historique de données.

Changements de terminologie : des changements dans la dénomination des bactéries ont par exemple été observés. Durant la période de 5 ans, le laboratoire s'est équipé d'une chaîne automatisée de bactériologie et en a profité pour harmoniser la dénomination des bactéries avec des référentiels du leader Français dans ce domaine.

Grâce aux fonctionnalités d'Amadea, il a été possible d'identifier et de nettoyer rapidement les terminologies erronées. Amadea a également permis d'homogénéiser et de standardiser les données. Cependant, les laboratoires de Biologie Médicale devraient, le plus vite possible, utiliser les codifications recommandées pour la standardisation des résultats, LOINC, UCUM et SNOMED. Plus les données seront standardisées, plus les projets de Machine Learning seront rapides à mettre en œuvre.

Par ailleurs, une standardisation générale des données permettra aux différents laboratoires d'un même groupe de collaborer au moment de mettre en œuvre un système de Machine Learning. Cette collaboration entre laboratoires est souhaitable pour la réussite des projets. En effet, le Machine Learning ne fournira des résultats pertinents que si la phase d'apprentissage se fait sur de grands jeux de données. En collaborant, les laboratoires peuvent donc constituer une base de données commune plus large, et obtenir plus rapidement des résultats probants.

3.1.3. Structuration des données

La qualité des données (biologiques et cliniques) et leur structuration, sont cruciales pour la mise en œuvre du Machine Learning. La saisie des informations doit être aussi structurée que possible. Prenons l'exemple d'une information « Patient sous antibiotique », elle peut être saisie de manière structurée dans une liste déroulante sur certains sites et elle peut être saisie en commentaire libres sur d'autres sites. Des données saisies de manière structurée seront d'autant plus exploitables dans la phase d'apprentissage. Les Data Scientists parlent de « préparation des données » pour évoquer cette phase d'homogénéisation.

Dans l'analyse de données, quand une règle s'applique à 98% ou 99% des cas, les 1% manquants doivent aussi être investigués. Dans l'expérimentation, nous nous sommes assez vite aperçus par exemple que pour certaines règles, les patients en réanimation, ou, par exemple, les patients insuffisants rénaux présentent des profils si atypiques qu'ils ne peuvent pas être intégrés à la population générale, au risque de biaiser complètement les seuils de détection des différentes variables biologiques. Le défi est d'identifier ces cas particuliers qui doivent faire l'objet d'une analyse séparée. Dans le cas des patients en réanimation, il est possible de les identifier facilement, car un code bien spécifique est assigné au service de réanimation. Dans d'autres cas, il manque des données structurées

qui permettraient d'identifier certaines typologies (par exemple, « le patient est diabétique »). Le périmètre des données a donc été de plus en plus étendu, et, surtout, nous avons vu la nécessité d'établir un « profiling » des patients permettant de cibler le diagnostic en fonction des pathologies connues.

3.1.4. Phases d'apprentissage et de test du Machine Learning

L'expérimentation s'est faite sur une durée très courte, et l'objectif principal a été de mettre en œuvre un prototype de Machine Learning pour tester ce type de solution dans un contexte réel.

Les phases d'apprentissage et de test du Machine Learning se déroulent de la manière suivante. L'historique de données est divisé en deux jeux de données distincts :

- Un jeu d'apprentissage, sur lequel est entraîné le Machine Learning.
- Un jeu de test, sur lequel le Machine Learning va proposer ses propres résultats.

Dans la phase d'apprentissage, les données des patients sont analysées (par exemple l'identifiant du dossier patient, l'identifiant de l'analyse, le résultat, l'unité, la date, l'âge, le sexe, les normes…). La machine analyse également les résultats fournis par le Système Expert, et l'objectif est de déduire la logique qui a conduit à ce résultat. Au terme de la phase d'apprentissage, le Machine Learning propose donc un ensemble de relations pour expliquer les résultats du Système Expert. Cet apprentissage se fait de manière itérative, en affinant progressivement les règles proposées par le Machine Learning.

Dans la phase de test, le jeu de données ne contient que les données biologiques des patients, sans le résultat fourni par le Système Expert. Le champ « résultat » du Système Expert devient ainsi le champ cible de l'analyse. L'algorithme exécute les règles définies lors de la phase de test, et propose cette fois-ci ses propres résultats.

Une règle statistique proposée par le Machine Learning peut, par exemple, se présenter ainsi : « 99% de validation automatique si LDL-Cholestérol a un résultat est inférieur à 1,7 g/l ».

Les règles peuvent être beaucoup plus complexes et inattendues. Les règles établies par le Machine Learning doivent être étudiées et validées par les experts Métiers. Certaines des relations identifiées par le Machine Learning ne seraient pas venues naturellement à l'esprit humain et méritent donc d'être explorées par les Biologistes. Dans certains cas, les Biologistes peuvent vérifier la pertinence des règles proposées par le système en collectant des informations de contexte clinique supplémentaires.

Le rôle du Biologiste est essentiel dans cette phase de tri des règles identifiées par le Machine Learning. Son expertise lui permet d'affiner certaines règles proposées. Par exemple, dans le cas du LDL-Cholestérol évoqué ci-dessus, le Biologiste peut décider de fixer la borne à 1,6g/l si cela s'explique scientifiquement.

A chaque fois que des notions supplémentaires sont ajoutées, le Machine Learning est à nouveau exécuté et la pertinence des règles proposées s'affine.

3.1.5. Résultats de l'expérimentation

Une comparaison a été faite entre les décisions du Système Expert utilisé par le laboratoire et les décisions du prototype de Machine Learning. A la suite de l'expérimentation, nous avons tiré plusieurs conclusions. Tout d'abord, cela peut sembler une évidence, la qualité et la structuration des données est primordiale pour la pertinence des règles proposées. C'est d'autant plus vrai dans le cadre du Machine Learning, qui implique une phase d'apprentissage déterminante pour la suite.

Sur les apports respectifs du Machine Learning et des Systèmes Experts, nos conclusions sont les suivantes :

- Le Machine Learning permet d'identifier des relations entre les données jamais envisagées par les Biologistes. Ces relations, après validation par les biologistes, peuvent compléter la connaissance autour d'une pathologie donnée, et enrichir les possibilités de diagnostic.

- La validation par les règles de Machine Learning permet d'avoir un taux de confiance de l'algorithme au moment de la validation. Cela permet, contrairement aux résultats binaires des SAVB, d'avoir une plus grande subtilité dans la validation. Cette approche est plus en adéquation avec la réalité biologique, où l'on sait que toute réponse n'est jamais vraiment binaire.

Cependant, le Machine Learning n'est pas « magique », et il implique une logique exploratoire qui demande un investissement humain plus important que pour un Système Expert. Mais cette combinaison de l'Intelligence Humaine et de l'Intelligence Artificielle est l'approche la plus pertinente pour caractériser finement une pathologie.

La méthode la plus efficace consiste donc à utiliser les forces de chacune de ces 2 approches :

- D'une part, l'utilisation de règles basées sur les études scientifiques alimentées par des Experts Métiers qui permettent de reproduire des conclusions scientifiquement explicables.
- D'autre part, les approches Machine Learning vont permettre d'explorer les données plus en profondeur, mettre en lumière de nouvelles relations entre les données et ainsi acquérir des connaissances nouvelles, qui viendront enrichir le système d'aide à la validation biologique. La capacité à formuler des relations complexes entre les données, et à associer un intervalle de confiance à chaque mesure, font du Machine Learning une approche prometteuse pour la détection future des états pré-pathologiques.

Dans ce cadre, l'humain a toute sa place, et apportera son expertise sur les points suivants :

- Etudier plus en profondeur les relations identifiées par les algorithmes de Machine Learning.
- Proposer leurs règles Métier basée sur leur expérience clinique et la connaissance fine des spécificités de leur population de patients.
- Dans les cas de pathologies complexes, ou à la limite entre 2 états physiologiques, le rôle du médecin est crucial. Par l'observation clinique et l'échange avec le patient, le médecin peut récupérer des informations déterminantes, qui pourront être injectées dans les algorithmes, afin d'améliorer les modèles.

Ainsi, l'Intelligence Artificielle, loin de déposséder les médecins de leurs attributions, va au contraire leur fournir une aide décisionnelle quotidienne, mais constitue aussi une piste supplémentaire dans leurs efforts pour détecter, de plus en plus tôt, les symptômes de potentielles futures pathologies.

3.2. Parallèle avec la détection de fraude dans le secteur financier

Dans le but d'élaborer une approche méthodologique de l'utilisation du Machine Learning en Biologie Médicale, nous avons dans un premier temps mené une expérimentation sur des données réelles de laboratoires. Dans un second temps, nous allons étudier des cas d'utilisation du Machine Learning dans un autre domaine d'application. En effet, les enjeux et les besoins du secteur biomédical sont très similaires à ceux d'autres secteurs d'activité, qui utilisent depuis longtemps le Machine Learning pour répondre à leurs problématiques.

Les données biomédicales ont leurs caractéristiques propres, et une solution de Machine Learning dédiée à l'analyse de données biologiques devra tenir compte de ces spécificités. Le domaine biomédical se distingue notamment par la complexité des données (rareté de certaines pathologies…) mais aussi par la sensibilité des informations médicales. En effet, la gestion des données de patients doit se faire en garantissant un traitement sécurisé, confidentiel, et des résultats facilement traçables. Une exigence à prendre en compte au moment de déployer une solution de Machine Learning pour l'aide au diagnostic. Certains algorithmes fonctionnent en « boîte noire », ce qui rend impossible

l'interprétation des résultats obtenus. Même dans le cadre du Machine Learning, le médecin doit toujours disposer des informations nécessaires pour comprendre le résultat proposé par l'algorithme.

Il faut donc tenir compte des spécificités du domaine biomédical, et de ses besoins propres. Comment mettre en œuvre une solution de Machine Learning dans un cadre confidentiel ? Comment entraîner un modèle pour détecter des pathologies rares ? Comment s'assurer que le médecin disposera, sur tout le processus, d'une information transparente et explicable ?

Bien que peu communes, ces problématiques ont déjà été rencontrées dans d'autres secteurs. C'est avec le domaine de la sécurité financière que nous avons pu établir les parallèles les plus riches en enseignements. Par de nombreux aspects, l'utilisation du Machine Learning par les établissements financiers correspond aux besoins que nous avons identifiés pour les laboratoires d'analyse médicale. Même si les deux domaines demeurent très éloignés, la Biologie Médicale peut s'inspirer des innovations développées au sein des banques, ces dix dernières années, pour nourrir leur réflexion autour des applications possibles de l'Intelligence Artificielle.

Le partenaire ISoft, dont la technologie est utilisée depuis plus de 10 ans par plus de 90% des banques françaises dans la Lutte contre la Fraude, a retrouvé dans l'expérimentation menée avec SIL-LAB de nombreuses similitudes avec les problématiques sur lesquelles ils ont travaillé pour l'analyse des transactions bancaires.

Voici quelques-unes des caractéristiques que les solutions d'Intelligence Artificielle, appliquées au milieu médical, doivent partager, selon nous, avec les solutions d'IA dédiées à la sécurité financière :

Traçabilité et explicabilité des données : les transactions bancaires sont des données hautement confidentielles, et l'outil d'analyse doit garantir un traitement sécurisé de ces données. De plus, la décision, prise par un algorithme, de valider ou de signaler une transaction, doit être explicable et documentée. Dans le monde médical comme dans le monde bancaire, une décision prise par un algorithme doit être explicable. Un algorithme conçu comme une « boîte noire » est incompatible avec les exigences de transparence et d'explicabilité du secteur financier, et la problématique est la même pour les diagnostics médicaux.

Analyse comportementale : Pour être plus pertinents, les modèles doivent prendre en compte ce que l'on appelle des données de contexte. En effet, le cas de certains patients est si particulier (par exemple, les patients diagnostiqués « insuffisants rénaux ») qu'il faut tenir compte de leur profil atypique dans l'analyse. Ainsi, si les analyses de ces patients dépassent certains seuils, les résultats a priori « hors-norme » sont parfois explicables au regard de l'historique du patient. Cette problématique se rapproche de l'analyse comportementale effectuée par ISoft dans le cas de la sécurité financière. Dans certains cas, une transaction financière a priori suspecte est en fait explicable lorsque l'on dispose de données de contexte sur la personne. Ainsi, une transaction effectuée en Asie alors que le porteur de la carte est français paraîtra frauduleuse. Mais si l'on sait que la personne est actuellement en voyage d'affaires, l'analyse en sera changée. La solution d'Intelligence Artificielle devra donc être capable de collecter et d'analyser ces données de contexte.

Détection d'événements rares : il est très complexe d'établir un modèle statistique pour détecter des événements qui ne se produisent que rarement. Or, l'Intelligence Artificielle est confrontée à ce cas de figure dans plusieurs domaines. C'est l'un des défis de l'utilisation de l'IA pour détecter les cas de fraude, qui ne représentent qu'une proportion minuscule de l'ensemble des transactions financières. De même, utiliser l'IA dans le milieu médical implique une expertise dans la détection d'événements rares. Comme expliqué précédemment, l'expérimentation a nécessité un large historique, afin de disposer de suffisamment de cas pour chaque pathologie.

Combiner le Machine Learning et l'expertise humaine : ce point découle du précédent. En effet, le Machine Learning utilisé seul ne peut pas suffire à détecter des événements rares. Les algorithmes actuels de Machine Learning ont besoin, pour apprendre, d'un large historique de données. Il faut donc disposer de nombreuses données pour l'ensemble des cas de patients, y compris les cas de maladies rares. Or, dans certains cas, les laboratoires ne disposeront pas de l'historique nécessaire pour entraîner efficacement les algorithmes. Il faut donc un système hybride, avec une part de règles basées sur la connaissance humaine, et une part de Machine Learning pour explorer de nouvelles pistes et mettre en évidence des phénomènes biologiques non détectables, à l'heure actuelle, par l'analyse humaine. Les solutions d'Intelligence Artificielle, dans le domaine médical, devront donc combiner les techniques de Machine Learning, et la possibilité, pour les médecins, de créer des règles expertes.

4. Principe d'une méthodologie appliquée à la Biologie Médicale

L'expérimentation réalisée avec le laboratoire BIO86, et le parallèle avec l'utilisation de l'Intelligence Artificielle pour la sécurité financière, a permis d'identifier plusieurs étapes à prendre en compte pour une utilisation judicieuse de l'Intelligence Artificielle dans le domaine biomédical. Les points présentés dans cette section ont pour objectif de tracer une première méthodologie pour l'application de l'IA dans la Biologie Médicale.

4.1. Une solution hybride alliant Machine Learning et expertise humaine

Nous avons vu que les systèmes d'aide au diagnostic seront particulièrement efficaces s'ils combinent le Machine Learning et les règles basées sur la connaissance scientifique. Les solutions mises en œuvre dans les laboratoires doivent permettre aux utilisateurs de faire évoluer les modèles de façon simple et réactive. Cela implique certaines fonctionnalités, comme la possibilité de créer ou d'ajouter, en temps réel, de nouvelles règles expertes dans l'outil.

La place de l'expertise humaine dans les solutions d'Intelligence Artificielle est aussi cruciale pour l'explicabilité. Les solutions d'IA mises en œuvre doivent fonctionner de manière transparente, et livrer des résultats traçables et explicables. Les analystes doivent être en mesure de comprendre les décisions de l'outil, pour les valider ou au contraire enrichir les modèles et les rendre plus pertinents.

4.2. Un historique de données exhaustif

Le Machine Learning nécessite une première phase d'apprentissage, pendant laquelle les algorithmes vont être entraînés sur un historique de données. Les algorithmes actuels ont besoin d'un vaste historique pour apprendre efficacement. Les données doivent aussi être équilibrées, c'est-à-dire que les différents cas de pathologie doivent tous être représentés de manière significative. Un algorithme ne pourra pas apprendre de manière probante sur un phénomène biologique, si ce cas n'apparaît que très rarement dans les données d'apprentissage. Pour des cas rarissimes, il sera bien sûr impossible d'établir un historique suffisant – dans ce genre de cas extrême et complexe, l'expertise humaine est plus que jamais nécessaire. Mais il faut, dans la mesure du possible, constituer au préalable un historique de données exhaustif pour réussir la phase d'apprentissage.

4.3. Tenir compte de l'historique du patient

Les données biologiques des patients sont analysées par les Systèmes Experts en fonction de certains seuils. Par exemple, si le taux de cholestérol du patient ne dépasse pas le seuil limite fixé, le résultat sera validé par l'outil. Mais pour être pertinent, les seuils ne doivent pas être tout à fait rigides. Les données biologiques des patients peuvent sensiblement varier si, par exemple, le patient suit un traitement au long cours, ou présente un terrain familial. Un Système d'Aide à la Validation Biologique doit prendre en compte ces données de contexte. Si ce type d'information n'est pas disponible, l'algorithme ne validera pas certains résultats, qui sont pourtant « normaux », au regard de l'historique du patient.

En s'assurant que la solution dispose de toutes ces fonctionnalités, les laboratoires pourront mettre en œuvre plus simplement le Machine Learning, mais aussi les règles expertes identifiées au préalable. L'Intelligence Artificielle prendra alors tout son sens, en combinant à la fois des règles déterminées par la connaissance scientifique, et des techniques de Machine Learning pour explorer de nouvelles relations entre les données, et détecter des phénomènes biologiques qui échappent à l'analyse humaine.

Conclusion

Pour appréhender la mise en œuvre du Machine Learning dans la Biologie Médicale, nous avons procédé en deux temps. Nous avons d'abord mené une expérimentation en laboratoire sur des données réelles, pour évaluer les apports du Machine Learning par rapport aux Systèmes Experts habituellement utilisés. Nous avons ensuite étudié les similitudes avec le domaine de la fraude bancaire, où les apports de l'IA sont aujourd'hui reconnus et incontestables. Ces deux domaines sont très différents, mais les besoins en termes d'analyse sur des données volumineuses, complexes et hautement hétérogènes, avec l'objectif de trouver des cas rares sont finalement très proches. Sécurité, traçabilité et explicabilité des données sont également des prérequis dans les deux domaines.

A l'ère de la production massive de données, quel que soit le domaine, la volumétrie ne permet plus une analyse uniquement humaine, et même les méthodes statistiques doivent être adaptées à la haute dimensionnalité : c'est la raison d'être du Machine Learning. Loin d'être une finalité, il est avant tout un outil d'aide décisionnelle précieux. Il permet d'établir des corrélations non évidentes par l'analyse humaine et il ouvre de nouvelles perspectives. L'analyse humaine permettra alors d'explorer ces dernières, de faire évoluer des concepts scientifiques, et, in fine, de mieux comprendre la causalité de certains phénomènes physio-pathologiques.

Le laboratoire de biologie médicale et le biologiste, s'ils se saisissent des opportunités offertes dès aujourd'hui par le Machine Learning, devraient devenir la pierre angulaire de la médecine préventive et personnalisée de demain. L'analyse médicale permet d'obtenir un très grand nombre d'informations invisibles par simple constatation des signes cliniques par un médecin. La connaissance du fonctionnement biologique du corps humain et les techniques d'analyses évoluent en permanence permettant en même temps d'avoir une biologie de plus en plus personnalisée et une complexité d'interprétation exponentielle par la combinatoire de l'interaction possible entre tous les paramètres biologiques. L'intelligence artificielle n'est qu'un outil qui va permettre au biologiste d'appliquer des raisonnements de plus en plus poussés dans un temps acceptable cliniquement.

Notes

[2] : Dans le meilleur des cas, il ne s'agit que d'un changement d'automate, donc de code d'analyse, avec des résultats a priori comparables. Dans d'autres cas, il s'agit d'un changement de technique, et il est alors possible d'établir une règle de comparabilité entre les algorithmes. Dans le dernier cas, il s'agit d'un changement complet de démarche, soit dû à une nouvelle stratégie de diagnostic interne, soit dicté par les politiques de santé publiques.