SIL-LAB experts

DOCUMENT CERTIFIÉ AdvanThink

INTELLIGENCE ARTIFICIELLE ET BIOLOGIE MÉDICALE

Le Machine Learning pour l'aide à la validation biologique

Expérimentation menée avec le laboratoire Bio86

A PROPOS DE SIL-LAB EXPERTS

SIL-LAB Experts est la première société de Conseil en Informatique pour les Laboratoires de Biologie Médicale (LBM). SIL-LAB développe également des solutions innovantes dans les domaines de l'analyse de données, de la gestion des prélèvements en mobilité, et de la cartographie des LBM. Le fondateur de SIL-LAB Experts, Serge Payeur, a également créé, avec SELIC, une nouvelle entité, SIL-LAB Innovations, dont l'objectif est de faciliter la dématérialisation des informations liées au prélèvement à domicile des patients.

A PROPOS DE BIO86

BIO86 est un laboratoire multi-sites, issu du regroupement de différents laboratoires de Biologie Médicale privés de la Vienne. Sa création est une réponse à l'évolution très importante de la Biologie dans un contexte réglementaire de plus en plus ouvert à l'Europe. Le regroupement est un gage d'efficacité dans la prise en compte de toutes les nouvelles contraintes de la profession, tout en restant très attaché à l'exercice d'une Biologie praticienne de proximité. Le site de la Polyclinique de Poitiers est dirigé par Bruno Gauthier qui est particulièrement investi sur les sujets relatifs à l'informatique et aux systèmes d'information dans le milieu de la Santé. Il est Vice-Président au Syndicat des Biologistes et membre du bureau de la Société Française d'Informatique de Laboratoire.

A PROPOS D'ADVANTHINK (ISOFT)

AdvanThink (ISoft) est un éditeur de logiciels hautes performances dédiés aux Métiers. Sa technologie est le fruit de 25 ans d'innovations et d'une avance historique en Machine Learning. Les solutions d'AdvanThink (ISoft) adressent les applications critiques, telles que l'analyse du risque en Temps Réel de 90% des transactions bancaires françaises. Impliqué depuis ses débuts dans des projets de recherche français et européens, le département de Bioinformatique d'AdvanThink (ISoft) a une expérience de 20 ans dans l'analyse des données biologiques (de type omique) et médicales. Pour l'expérimentation présentée dans ce document, AdvanThink (ISoft) a apporté sa technologie, mais aussi son expertise en Machine Learning et en analyse de données médicales.

Sommaire

Résumé du document

Aujourd'hui, l'Intelligence Artificielle (IA) est l'objet de nombreux discours alarmistes, notamment sous l'angle éthique et les risques que l'IA devienne plus puissante que l'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 ces dix dernières années, notamment liés à l'accréditation ISO-15189. Le champ d'utilisation des Systèmes d'Aide à la Validation Biologique (SAVB), est défini de la manière suivante :

> Les SAVB 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'IA, sous certaines conditions que nous allons détailler.

Le document aborde ensuite en détail une expérimentation, réalisée dans le Laboratoire de Biologie Médicale BIO86, à Poitiers, pour évaluer l'utilisation du Machine Learning dans un SAVB. Le document présente le retour d'expérience de cette évaluation, menée sur des données réelles de Biologie Médicale, en expliquant les étapes suivies, les 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 la 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.

Les approches Machine Learning vont permettre d'explorer les données plus en profondeur, mettre en lumière des relations inédites entre les données et ainsi acquérir des connaissances nouvelles.

Utilisation actuelle de l'IA en 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.

Rappel sur les domaines de l'IA

L'Intelligence Artificielle est un ensemble de techniques qui permettent à la machine de reproduire des raisonnements humains. Dans certains domaines de l'IA, comme les Systèmes Experts, la machine applique simplement des règles programmées au préalable. A l'inverse, le Machine Learning consiste en une approche statistique permettant à la machine, sans être programmée, d'apprendre progressivement en analysant les données qui lui sont fournies.

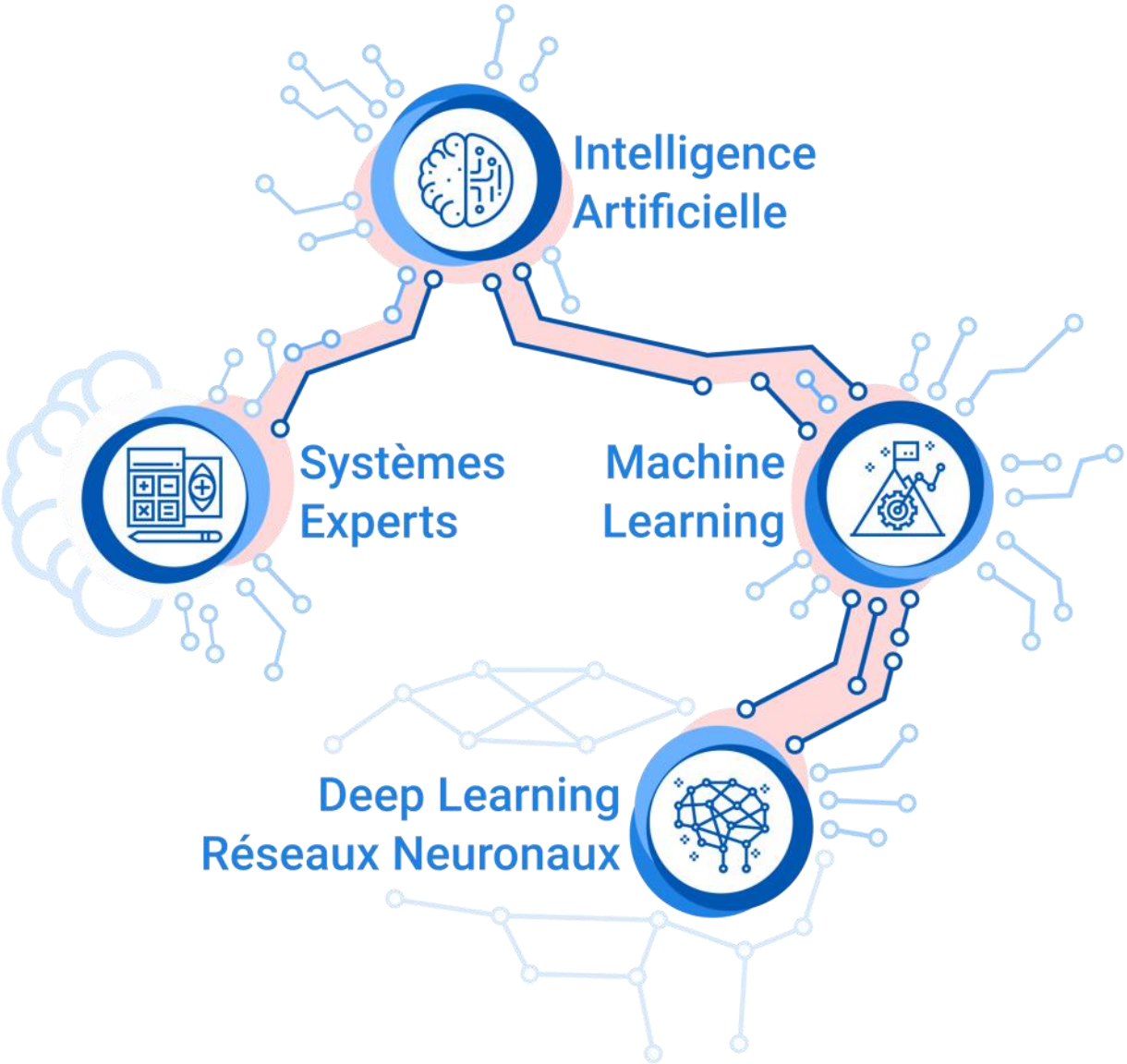
Un sous-ensemble du Machine Learning est appelé Deep Learning et utilise ce qu'on appelle les « Réseaux de Neurones » pour apprendre sur les données. L'avènement du Deep Learning a permis des progrès remarquables pour l'analyse des images et de sons. Dans l'analyse d'image, le Deep Learning s'avère très performant, en particulier quand les images qui lui sont fournies sont de très haute définition. La machine est capable trouver des facteurs explicatifs invisibles à l'œil humain et d'analyser ces facteurs en temps réel, dépassant ainsi l'analyse humaine aussi bien en performance qu'en rapidité.

Cependant, les Réseaux de Neurones fonctionnent en « boîte noire », et ne permettent pas de comprendre comment la machine a pris la décision. Ces algorithmes sont donc moins pertinents dans les domaines où la prise de décision doit s'accompagner d'explications. A noter que le Deep Learning souffre d'autres limites, notamment le risque de « sur-apprentissage ».

Si l'Intelligence Artificielle est source de progrès pour l'analyse de données, elle est encore loin d'être infaillible.

L'Intelligence Artificielle n'est pas magique : elle apprend, entend, voit et se trompe. Toute analyse s'appuyant sur l'IA implique des phases de test, d'itération et d'amélioration constante des algorithmes.

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



Les champs d'étude de l'Intelligence Artificielle

Le Système d'Aide à la Validation Biologique

Le Système d'Aide à la Validation Biologique (SAVB) est un terme générique[2] 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 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, le Machine Learning n'est pas utilisé pour paramétrer ces outils et les algorithmes sont complètement déterministes : quand on prend deux fois les mêmes données, on obtient exactement les mêmes résultats.

Dans le cas des règles en boîte noire, une longue phase de validation est nécessaire avant de passer le SAVB en mode validation automatique / supervisée. Le déterminisme de ce 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 et l'adapter à la population étudiée, constitue encore un frein à l'utilisation de ces systèmes en routine. Les Systèmes Experts les mieux adaptés aux Laboratoires de Biologie Médicale sont donc ceux permettant aux Biologistes de paramétrer leurs propres règles.

Il est 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.

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

Avec l'utilisation, répandue depuis plusieurs années, des Systèmes d'Aide à la Validation Biologique, 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 donc des gains de temps considérables ainsi qu'une aide précieuse au diagnostic. Cependant, l'exécution des algorithmes doit se faire sous la responsabilité des Professionnels de Santé. Il faut que les décisions prises restent traçables et explicables.

Quant aux Systèmes Experts en « boîte noire », les fournisseurs doivent être en mesure d'apporter des explications sur leur fonctionnement. A ce titre, l'utilisation de ce type de Systèmes Experts en Biologie Médicale pose question. Comment les Biologistes, qui utilisent ces Systèmes en boîte noire, 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 Experts 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 le Machine Learning 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 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.

Les apports du Machine Learning

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 de Biologie Médicale, avec des données réelles (anonymisées) de patients.
- Comparaison avec le domaine de la Sécurité Financière, qui partage avec le domaine Biomédical des problématiques similaires de mise en œuvre du Machine Learning.

Ces deux approches complémentaires nous ont permis d'appréhender, dans un contexte réel, l'utilisation de méthodes de Machine Learning pour l'analyse des données Biomédicales. Les obstacles rencontrés lors de l'expérimentation ont montré la nécessité d'établir un protocole pour faciliter la mise en œuvre de projets basés sur le Machine Learning. Dans cette perspective, nous proposons dans la dernière partie de ce document une démarche méthodologique pour la mise en œuvre du Machine Learning au sein des Laboratoires de Biologie Médicale.

Expérimentation en laboratoire sur des données réelles

L'objectif principal de cette expérimentation était de concevoir un prototype de Système d'Aide à la Validation Biologique, basé sur le Machine Learning. En termes de méthode d'analyse, cette approche diffère des Systèmes Experts classiques. En effet, un Système Expert repose sur des règles déterministes, explicitement programmées au préalable. Dans une approche basée sur le Machine Learning, l'algorithme apprend sur un jeu de données, et au terme de cette phase d'apprentissage, infère un ensemble de relations expliquant les résultats observés dans ce jeu de données. Ces relations sont exprimées sous la forme de probabilités.

L'un des apports du Machine Learning est l'analyse approfondie de grandes volumétries de données, et la mise en évidence de relations inédites, qui ne peuvent pas être détectées par la seule analyse humaine. Dans le cadre de la Biologie Médicale, le Machine Learning constitue un domaine prometteur, car il pourrait permettre à l'avenir de détecter, en amont, les premiers symptômes d'une pathologie.

L'expérimentation a permis de comparer les résultats proposés par le prototype de Machine Learning à ceux d'un Système Expert classique. L'expérimentation a été menée avec des données anonymisées du Laboratoire BIO86, à Poitiers. Ce laboratoire utilisait, depuis quelques mois, un Système Expert au quotidien pour le traitement des données Biomédicales.

Le Laboratoire BIO86, composé de 12 sites, traite environ 3 000 dossiers par jour, soit près de 25 millions de résultats annuels. L'utilisation du Machine Learning, dans le domaine de la Biologie Médicale, est d'autant plus pertinente que l'on dispose d'un grand historique de données, notamment pour pouvoir étudier des cas de pathologies extrêmement rares. Pour cette raison, dans le cadre de l'expérimentation, 5 années d'historique d'activité du Laboratoire BIO86 ont été utilisées.

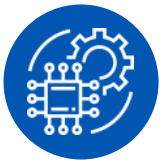
Les données Biomédicales sont très volumineuses et complexes, et ces caractéristiques doivent être prises en compte. Ce point est particulièrement critique pour la phase d'analyse et de préparation des données qui précède la phase d'apprentissage du Machine Learning.

Plusieurs solutions de Machine Learning ont donc été examinées avant la phase de prototype. Les outils ont été évalués sur les critères suivants :

Capacité à intégrer et stocker de grandes volumétries de données

Capacité à réaliser des traitements et des statistiques sur ces volumétries

Capacité à analyser, préparer et standardiser des données brutes et hétérogènes pour les rendre exploitables



Capacité à faire analyser des données complexes par un algorithme de Machine Learning sur de grandes volumétries

Le logiciel Amadea, de la société AdvanThink (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. Nativement, Amadea a la capacité de se connecter à de nombreux formats de données. Dans le cadre de cette expérimentation, un connecteur ODBC a été utilisé. Amadea a aussi été sélectionné car il dispose d'une bibliothèque d'algorithmes de Machine Learning, capables 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.

BIO86 utilise le Système de Gestion de Laboratoire Odancio de DL Santé (Groupe Dedalus), 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.

Standardisation des données

Après la phase d'intégration, nous avons vérifié la qualité des données et la cohérence de chaque source utilisée. Nous avons identifié les 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 5 ans, certains paramètres ont changé[3]. 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 des données.

Changements de terminologie. Par exemple, des changements dans la dénomination des bactéries ont été observés. Durant les 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 corriger les erreurs de terminologies. Amadea a également permis d'homogénéiser et de standardiser les données. Nous préconisons aux Laboratoires d'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, la standardisation des données permet 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.

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, ou en commentaire libre sur d'autres sites. Des données saisies de manière structurée seront plus facilement exploitables durant 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 avons détecté que, par exemple, pour certaines règles, les patients en réanimation, ou les patients insuffisants rénaux, présentaient des profils si atypiques qu'ils ne pouvaient pas être intégrés à la population générale, au risque de biaiser complètement les seuils de détection des différentes variables biologiques.

Ces profils particuliers illustrent la nécessité d'établir des catégories de patients devant faire l'objet d'une analyse séparée. Dans certains cas, la structuration des données rend possible cette classification : par exemple, un code spécifique est assigné au service des patients en réanimation. Dans d'autres cas, le manque d'informations structurées empêche d'établir directement une typologie (par exemple, « le patient est diabétique »), il est alors nécessaire d'établir un « profiling » de ces patients.

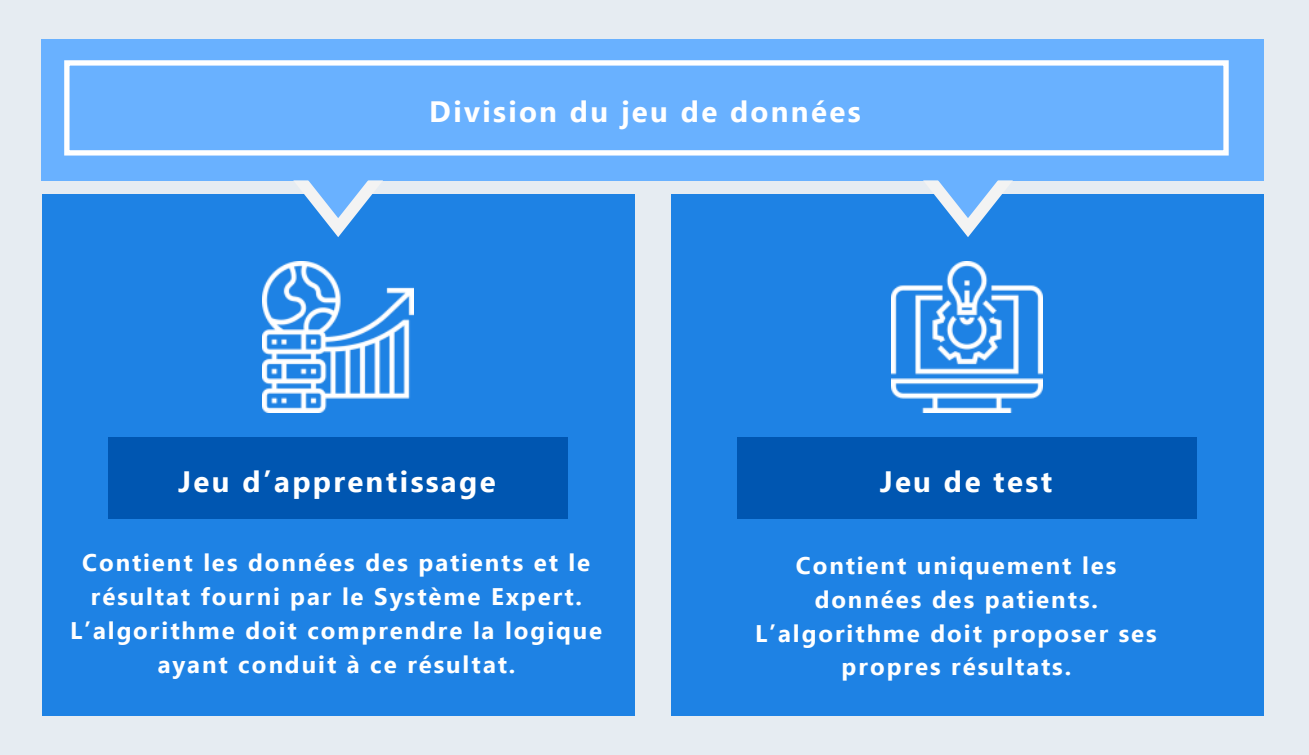
Phases d'apprentissage et de test du Machine Learning

L'objectif principal de l'expérimentation a été de mettre en œuvre, dans un temps très court, 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 distincts :

1. Un **jeu d'apprentissage**, sur lequel est entraîné l'algorithme de Machine Learning.

2. Un **jeu de test**, sur lequel l'algorithme propose ses propres résultats.

> **Une collaboration entre laboratoires d'un même groupe est souhaitable, car le Machine Learning ne fournira des résultats pertinents que si la phase d'apprentissage se fait sur de grands jeux de données.**



Phase d'apprentissage

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 ces résultats. 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.

Phase de test

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. L'algorithme exécute les règles définies lors de la phase d'apprentissage, 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 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 Biologistes. 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 processus d'apprentissage est rejoué, et la pertinence des règles proposées s'affine.

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.

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.
- **Les résultats obtenus par le Machine Learning reposent sur des probabilités, et non sur un raisonnement binaire (validation / invalidation).** Cela permet, contrairement aux résultats 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, et où l'on ne peut jamais être sûr à 100% d'un résultat.

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. Cette combinaison de l'intelligence humaine et de l'IA 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 :

L'utilisation de règles basées sur les études scientifiques et alimentées par les Biologistes permettront de reproduire des conclusions scientifiquement explicables.

Le Machine Learning permettra d'explorer les données en profondeur, de mettre en lumière des relations inédites entre les données et d'acquérir de nouvelles connaissances, qui enrichiront 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 également 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 des règles expertes** basées 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 pour améliorer les modèles.

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

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.

Les données Biomédicales ont leurs caractéristiques propres, et une solution basée sur le Machine Learning devra tenir compte de ces spécificités. Le domaine Biomédical se distingue notamment par la complexité des données (rareté de certaines pathologies…) mais aussi par le besoin de transparence : la décision prise par le SAVB doit être explicable. Certains algorithmes fonctionnent en « boîte noire », ce qui rend impossible l'interprétation des résultats obtenus. Or, le médecin doit toujours disposer des informations nécessaires pour comprendre le résultat proposé par l'algorithme.

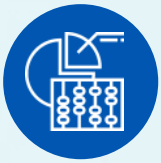
Il faut donc tenir compte des spécificités du domaine Biomédical, et de ses besoins propres. Comment mettre en œuvre une solution de Machine Learning maîtrisable de bout en bout par les Biologistes ? Comment entraîner un modèle pour détecter des pathologies rares ? Comment s'assurer que le médecin dispose constamment d'une information transparente ? Bien que peu communes, ces problématiques ont déjà été rencontrées dans d'autres domaines d'application.

La technologie d'AdvanThink (ISoft), fruit de 25 ans de R&D continue en Machine Learning, est utilisée pour des applications critiques dans de nombreux secteurs d'activité. Au cours de l'expérimentation menée avec SIL-LAB, AdvanThink (ISoft) a retrouvé des problématiques similaires à celles de la Sécurité Financière, un domaine où sa technologie est utilisée en Temps Réel par 90% des banques françaises pour la lutte contre la fraude. Même si les deux secteurs demeurent très éloignés, la Biologie Médicale peut s'inspirer des innovations développées au sein des banques pour nourrir sa réflexion autour de l'Intelligence Artificielle.

Des problématiques communes

Explicabilité des décisions. La décision prise par un algorithme de valider ou non une transaction doit être explicable et documentée. A l'instar du monde bancaire, un algorithme conçu comme une « boîte noire » est incompatible avec les exigences de transparence et d'explicabilité du domaine Biomédical.

Analyse comportementale. Pour être pertinents, les modèles doivent prendre en compte 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 AdvanThink (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 le client. Ainsi, pour un client souvent en déplacement, une transaction effectuée à l'étranger sera considérée plus facilement « dans la normalité » que pour d'autres clients.



Détection d'événements rares. La modélisation d'événements rares est un défi dans plusieurs domaines, et implique des techniques d'Intelligence Artificielle spécifiques. C'est le cas pour la détection des fraudes, qui ne représentent qu'une infime proportion de l'ensemble des transactions financières. Il en est de même pour le milieu médical. Même en disposant d'un large historique, les Biologistes manqueront probablement de données d'apprentissage sur certaines pathologies. Utiliser l'IA dans le milieu médical implique donc une expertise dans la détection d'événements rares.

Le Biologiste au centre du processus. Il nous paraît essentiel qu'un SAVB, même s'il intègre des méthodes de Machine Learning, reste au service des Biologistes. En effet, si le Machine Learning fournit un apport précieux à l'analyse, la connaissance humaine reste primordiale pour les diagnostics. Une solution d'IA déployée dans le domaine médical devra donc être hybride. Elle devra comporter des fonctions de Machine Learning, mais aussi permettre aux Biologistes d'ajouter, en temps réel, des règles expertes basées sur la connaissance scientifique. Dans la finance, le besoin est équivalent : l'expertise humaine reste cruciale, et les Responsables Risque doivent pouvoir comprendre et affiner, en temps réel, leur scénario de détection, notamment lorsqu'ils sont confrontés à un type de fraude inédit.

Méthodologie

L'expérimentation réalisée avec le laboratoire BIO86, et le parallèle avec l'utilisation de l'Intelligence Artificielle pour la Sécurité Financière, a permis de dégager plusieurs critères déterminants pour une intégration probante du Machine Learning dans le domaine Biomédical. Cette section présente les principaux critères identifiés, et propose une première méthodologie pour l'application de l'IA dans la Biologie Médicale.

- **Une solution hybride alliant Machine Learning et Expertise Humaine.** Nous avons vu que les systèmes d'aide au diagnostic sont particulièrement efficaces quand 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.

- **Une double expertise en Biologie et en Data Science.** La Biologie Médicale relève de l'analyse de cas pathologies rares et/ou complexes. En Data Science, on parle alors d'événements rares. Ce champ d'étude nécessite des techniques de Machine Learning spécifiques, et donc une expertise en la matière. De plus, même avec les techniques adéquates, l'historique de données disponibles et la phase de préparation des données sont déterminants pour la réussite de l'analyse. Cette phase de préparation consiste par exemple à traiter différemment les patients ayant un taux élevé de cholestérol, et nécessite l'expertise d'un Biologiste. La réussite d'un projet de Machine Learning dans un SAVB passe donc par une double expertise, à la fois en Biologie et en Data Science.

- **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 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.

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 classiques. 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. Détection d'événements rares, explicabilité et transparence, technologies accessibles à des utilisateurs non-experts : les deux domaines partagent de nombreuses problématiques.

L'analyse médicale permet d'obtenir un très grand nombre d'informations invisibles par simple constatation des signes cliniques par un médecin. La connaissance en biologie et les techniques d'analyses évoluent en permanence, ouvrant la voie à une analyse biologique de plus en plus personnalisée, mais aussi à une complexité d'interprétation exponentielle, due à la combinatoire des interactions possibles entre les paramètres biologiques.

A l'ère de la production massive de données, quel que soit le secteur, la volumétrie ne permet plus une analyse uniquement humaine, et même les méthodes statistiques doivent être adaptées à la haute dimensionnalité : c'est dans ce contexte que le Machine Learning devient pertinent. Loin d'être une finalité, il est avant tout un outil d'aide décisionnelle précieux. A travers le prototype réalisé dans cette étude, nous avons montré qu'il permet d'établir des corrélations non évidentes par l'analyse humaine et 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.

L'Intelligence Artificielle apparaît alors comme une solution prometteuse. Il s'agit avant tout d'un outil qui va permettre au Biologiste d'appliquer des raisonnements de plus en plus poussés dans un temps acceptable cliniquement. Le laboratoire Biomédical et le Biologiste, en saisissant les opportunités offertes dès aujourd'hui par le Machine Learning, peuvent devenir la pierre angulaire de la médecine préventive et personnalisée de demain.

Notes

1 : Rapport de Cédric Villani, Donner un sens à l'Intelligence Artificielle : pour une stratégie nationale et européenne : https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

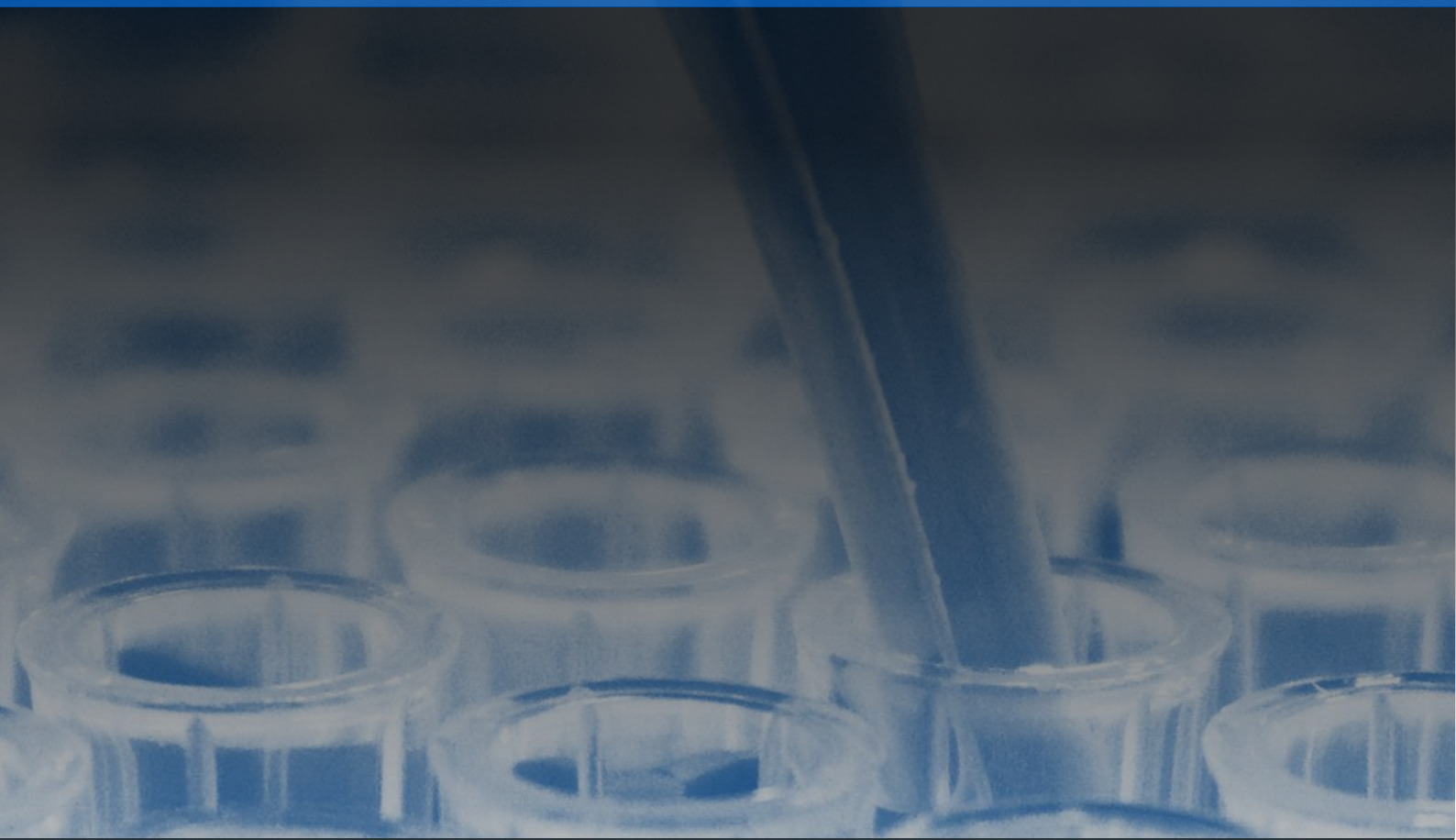
2 : Le SAVB est défini dans le guide technique SH-GTA-02.

3 : Dans le meilleur des cas, il s'agit 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 (on peut alors établir une règle de comparabilité entre les algorithmes) ou d'un changement complet de démarche, dû à une nouvelle stratégie de diagnostic interne, ou dicté par les politiques de santé publiques.

INTELLIGENCE ARTIFICIELLE ET BIOLOGIE MÉDICALE

Résumé

Pour appréhender la mise en œuvre du Machine Learning dans la Biologie Médicale, nous avons d'abord mené une expérimentation basée sur les données fournies par le Laboratoire BIO86, pour évaluer les apports du Machine Learning par rapport aux Systèmes Experts classiques. Nous avons ensuite étudié les similitudes avec le domaine de la fraude bancaire, où les apports de l'IA sont aujourd'hui reconnus. Ce document présente les résultats de cette expérimentation et propose les premiers contours d'une méthodologie d'utilisation de l'Intelligence Artificielle en Biologie Médicale.



LIVRE BLANC

www.sil-lab-experts.com | www.Advanthink.com