UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'INTERPRÉTABILITÉ AVEC L'HYBRIDATION ET LA SOPHISTICATION DU SYMBOLISME : L'EXEMPLE DE LA LOGIQUE FLOUE PROBABILISTE POUR LE CHOC SEPTIQUE.

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

BOUMEDIENE EL MOUENIS MESSAOUDI

MAI 2026

REMERCIEMENTS

Je tiens à remercier mes directeurs de thèse, les professeurs, Serge Robert de l'Université du Québec à Montréal et Ismaïl Biskri de l'Université du Québec à Trois-Rivières ainsi que mes anciens directeurs : Valérie Plante, directrice de la direction de l'analyse et de l'intelligence d'affaires au ministère de l'économie, de l'innovation et de l'énergie, Stéphane Asselin, directeur de la direction de la gestion des données numériques gouvernementales au ministère de la cybersécurité et du numérique (MCN), les directeurs du centre d'expertise en intelligence artificielle, analytique et automatisation du MCN : Matthias Raison, Martin Soucy, Sarah Gagnon-Turcotte, et Mohamed Nabil Tir, ainsi qu'Elena Birsan, cheffe du service de la gestion des données numériques ministérielles à la direction de la sécurité du ministère de l'immigration, de la francisation et de l'intégration.

Pour l'éternité, avec tout mon amour et respect perpétuel, je suis très redevable et reconnaissant à mes parents. Je réalise aujourd'hui que ma vie et ma stabilité sont le fruit de leur affection inconditionnelle, bonté, sacrifices et dévouement inébranlables. Que votre vie soit comblée de santé, joie, paix, sérénité et bonheur, à l'image de ce que vous m'avez donné. Longue vie à vous dans ce qui est bien.

Également, j'exprime ma gratitude à mes sœurs, mon frère ainsi que mes amis.

Ultimement, je tiens, tout autant, à rendre grâce à ceux qui ont eu à endurer mes années d'études au Canada. Je pense à ceux qui ont été patients tout au long de ces années : ma conjointe et mes enfants.

DÉDICACE

Ce travail est *in memoriam* à ceux qui évoquent un brin de biographies prophétiques… À des lignées du passé qui ne se retrouvent que peu de nos jours… À leurs départs de nos vies; ornés par la beauté de vols de paons blancs… Des éclipses qui nous rendent étrangers et dont on ne se remet jamais … Que de beaux souvenirs avec eux… Que de soupirs quand nous pensons à eux… Merci pour vos visites dans nos rêves… Ce sont les seuls endroits où notre solitude peut vous regagner… Nous nous reverrons dans la vraie vie et la seule nation… À la mémoire de vos chers parents et grands-parents ainsi qu'aux miens : Zana Beghdadi (1929 - 8 mars 2022), Mohamed Saci (1923 – 10 mai 2021), El Maadene Mbarka (1923 – 2006), Messaoudi Ali (1918 – 1968)…

Le présent.

« Hier, c'est derrière toi, demain, c'est un mystère, mais aujourd'hui,
c'est un cadeau et c'est pour ça qu'on l'appelle le présent. »

« Ton travail, personne ne va le faire pour toi, fais le par toi-même,
et dis-toi que personne ne va venir pour te donner des présents. »

« Si le ciel était de cuivre et la terre de plomb, si toute la création était tes enfants,
ne crains pas la pauvreté, enrichis ton cœur à l'instant présent. »

« N'attends pas l'orage à chaque fois,
car cela t'empêchera de bénéficier de l'éclat du soleil présent. »

« Ne lis pas ton bonheur à l'avenir,
ne vis pas un bien-être différé. Ton bonheur se conjugue au présent. »

« Ta béatitude n'est pas en retard pour toi, elle est maintenant,
et ce que tu vois n'est pas du hasard,
c'est une précision délirante de calculs, alors, profite de ton souffle présent. »

« Sois comme une fleur qui ne pense pas à rivaliser avec la fleur d'à côté,
elle s'épanouit. Ainsi, ta beauté est la grâce du moment présent. »

Citations adaptées.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

A/IS	*Autonomous and Intelligent Systems*
ACM	*Association for Computing Machinery*
AIC	*Akaike information criterion*
ANCHOR	*high-precision model-agnostic explanations*
APACHE II	*Acute Physiology And Chronic Health Evaluation II*
APR	*Area under Precision Recall*
ARM	*Association rule learning*
ATE	*Average Treatment Effect*
AUROC	*Area Under the Receiver Operating characteristic Curve*
BGM	*Bayesian Gaussian Mixture*
BIC	*Bayesian information criterion*
BN	*Bayesian Network*
CAM	*Class Activation Mapping*
CBR	*Case-based reasoning*
CDC	*Centers for Disease Control and Prevention*
CFCMC	*Cumulative Fuzzy Class Membership Criterion*
CI	*Contextual Importance*
CIU	*Contextual importance and utility*
CNN	*Convolutional Neural Network*
CU	*Contextual Utility*
CPN	*Causal Probabilistic Network*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DAC	*Directed Acyclic Graph*
DARPA	*Defense Advanced Research Projects Agency*
DL	*Deep Learning*
D-N	*Deductive-Nomological*
DSE	*Dossiers de Santé Électroniques*
EAD	*Ethically Aligned Design*
EASE	*Early Assessment of Sepsis Engagement*
EBM	*Explainable Boosting Machine*
ECE	*Expected Calibration Error*
FATE	*Fuzzy Average Treatment Effect*
FBE	*Feature-based Explanations*
FCRLC	*Fuzzy classification Rules Learning via Clustering*
FDA	*Food and Drug Administration*
FPT	*Fuzzy Probability Tree*
FTC	*Federal Trade Commission*
GMM	*Gaussian Mixture Model*
GradCAM	*Gradient-weighted Class Activation Mapping*
GRU	*Gated Recurrent Unit*
HMM	*Hidden Markov Model*
IA	*Intelligence artificielle*
IAI	*Interpretability Artificial Intelligence*
IEEE	*Institute of Electrical and Electronics Engineers*

IQR	*Interquartile Range*
LightGBM	*Light Gradient Boosting Machine*
LIME	*Local Interpretable Model-agnostic Explanations*
LRP	*Layer-wise Relevance Propagation*
LSTM	*Long Shot-Term Memory*
mAP	*Mean Average Precision*
Matlab	*matrix laboratory*
MCC	*Matthews Correlation Coefficient*
MEDS	*Mortality score in Emergency Department*
MEWS	*Modified Early Warning Score*
MGP-Att	*Multitask Gaussian Process and attention*
MIMIC-III	*Medical Information Mart for Intensive Care III*
ML	*Machine Learning*
NeSy	*Neuro-Symbolic*
NEWS2	*National Early Warning Score 2*
NFATE	*Normalized Fuzzy Average Treatment Effect*
NLP	*Natural Language Processing*
NLR	*Negative Likelihood Ratio*
NPV	*Negative Predictive Value*
NSIN	*National Security Innovation Network*
OCDE	*Organisation de Coopération et de Développement Économiques*
ODNI	*Office of the Director of National Intelligence*
OMS	*Organisation mondiale de la santé*
PFI	*Permutation Feature Importance*
PFL	*Probabilistic Fuzzy Logic*
PGM	*Probabilistic Graphical Model*
PIRO	*Predisposition/Insult-infection/Response/Organ dysfunction*
PLR	*Positive Likelihood Ratio*
PP	*Predictive Processing*
PPV	*Positive Predictive Value*
qSOFA	*quick SOFA*
RGPD	*Règlement général sur la protection des données*
RNN	*Recurrent Neural Network*
AUC SOFA	*Area Under the Curve (AUC) of the Sequential Organ Failure Assessment (SOFA)*
SAPS II	*Simplified Acute Physiology Score II*
SHAP	*SHapley Additive exPlanation*
SIRS	*Systemic Inflammatory Response Syndrome*
SOFA	*Sequential Organ Failure Assessment*
SPEEDS	*Sepsis Patient Evaluation Emergency Department Score*
SQL	*Structured Query Language*
SR	*Statistical Relevance*
SSS	*Sepsis Severity Score*
SWAM	*Shallow and Wide Attention convolutional Mechanism*
TAN	*Tree Augmented Naive*
TCN	*Temporal Convolutional Network*
TraCE	*Training calibration-based explainers*
t-SNE	*t-distributed Stochastic Neighbor Embedding*
UMAP	*Uniform Manifold Approximation and Projection*

Voxel	*Volume & pixel*
XAI	*Explainable Artificial Intelligence*

RÉSUMÉ

Dans des contextes sensibles marqués par des degrés d'incertitude et d'imprécision, les experts ont la responsabilité de fournir des justifications pour les prédictions. Il est préférable que ce soient des analyses basées sur une interprétabilité intrinsèque enrichie par la causalité. De surcroît, la recherche d'un équilibre de valeurs stratégiques (ex. éthique…) avec l'intelligence artificielle (IA) centré sur ces analyses est un questionnement important. Le rapprochement de cette symbiose répond, également, à un besoin d'appui pour la prise de décision, d'une personnalisation, d'automatisation, ainsi que pour une avantageuse interaction homme-machine avec l'IA. Cependant, dans ces contextes et en utilisant des boîtes noires avec peu de données, il est difficile de la faire valoir en fournissant ces analyses. D'ailleurs, certaines de ces boîtes ne déduisent que les associations et les corrélations. Elles ne se prêtent qu'à l'explicabilité. En revanche, notre méthode d'enrichissement (sophistication) du symbolisme avec la logique floue probabiliste (*Probabilistic Fuzzy Logic* : *PFL*) et d'intégration de modèle arborescent *LightGBM* permet ces analyses. Notre objectif est d'avoir les avantages de différents paradigmes pour une solution valorisante. Ainsi, pour ces analyses, nous utilisons cette méthode, peu faite et récente, pour une étude de cas sur le choc septique. Cela à travers des probabilités, du clustering (ex. *Gaussian Mixture (GM), Bayesian Gaussian Mixture (BGM)*) ainsi que des règles « *IF-THEN* » par la fuzzification. Notre approche se base, dès lors, sur des degrés d'appartenance flous. De plus, notre *PFL* permet d'estimer de façon segmentée l'effet moyen du traitement (*Average Treatment Effect* : *ATE*), offrant ainsi une granularité, mettant en évidence l'influence différenciée des variables sur l'effet du traitement. Autant, ces « *IF-THEN* » permettent la transformation des prédictions en concepts linguistiques et intervalles en se rapprochant du raisonnement pour une compréhension globale et locale. Leur combinaison avec l'arborescence crée un autocontrôle. Aussi, les alertes permettent une validation et il y a un calcul du score de fidélité pour prouver que la règle reflète le comportement. D'autant plus, les experts sont dans la boucle pour avoir leur révision. *In fine*, avec une applicabilité potentielle dans d'autres domaines sensibles, tout en soulignant la nécessité d'optimiser le choix des paramètres en raison de la susceptibilité de la méthode, cette démarche leur ouvre des perspectives intéressantes.

Mots clés : interprétabilité, causalité, valeurs, *PFL*, *LightGBM, Gaussian Mixture (GM), Bayesian Gaussian Mixture (BGM)*.

ABSTRACT

In sensitive contexts characterized by uncertainty and imprecision, experts have a responsibility to provide justifications for their predictions. It is preferable for these justifications to be based on intrinsic interpretability enhanced by causality.Furthermore, the search for a balance between strategic values (e.g., ethics…) and artificial intelligence (AI) centered on these analyses is a critical issue. The development of this symbiosis also addresses a need for decision-making support, personalization, automation, and beneficial human-machine interaction with AI. However, in these contexts and when using black boxes with limited data, it is difficult to demonstrate this value through such analyses. Further, some of these models only infer associations and correlations. They are limited to explainability. In contrast, our method of enriching (enhancing) the symbolic representation with Probabilistic Fuzzy Logic (PFL) and integrating the LightGBM tree-based model enables these analyses. Our goal is to leverage the advantages of different paradigms to deliver a valuable solution. Thus, for these analyses, we use this novel and recently developed method for a case study on septic shock. We do so by employing probabilities, clustering (e.g., Gaussian Mixture (GM), Bayesian Gaussian Mixture (BGM)), and "IF-THEN" rules through fuzzification. Our approach is therefore based on fuzzy membership degrees. Also, our PFL allows for a segmented estimation of the Average Treatment Effect (ATE), thereby providing granularity and highlighting the differentiated influence of variables on the treatment effect. As well, these "IF-THEN" rules enable the transformation of predictions into linguistic concepts and intervals, approximating reasoning for both global and local understanding. Their combination with the decision tree creates a self-checking mechanism. Additionally, alerts facilitate validation, and a fidelity score is calculated to demonstrate that the rule reflects the actual behavior. Moreover, experts are involved in the process to provide their review. Ultimately, with potential applicability in other sensitive fields, while highlighting the need to optimize parameter selection due to the method's sensitivity, this approach opens up interesting prospects for them.

Keywords: interpretability, causality, values, PFL, LightGBM, Gaussian Mixture (GM), Bayesian Gaussian Mixture (BGM).

0.1 Introduction

L'usage que nous faisons avec l'IA concrétise son sens et lui confère des valeurs de principes. Elle doit être transparente, responsable, de confiance, éthique, conforme, réglementaire, sécuritaire, auditable, certifiable, frugale et souveraine. Dès lors, sa virtuosité et sa portée gratifiante doivent s'acquérir par ces valeurs. Cette valorisation est indispensable à son application concrète. De sorte que, pour les systèmes décisionnels dans des contextes critiques et à haut risque, nous aurons de bons résultats avec des données non biaisées. Au-delà de l'explicabilité, la fonction d'interprétabilité, est, ainsi, garante de cette mise en valeur. Semblablement, elle doit être enrichie par la déduction d'une fonction de subtilités causales, au-delà des corrélations[1]. Également, elle doit être précise dans sa prédiction, personnalisable, robuste[2], généralisable dans d'autres contextes, flexible et raisonnante. À la lumière de ceci, la recherche d'un équilibre avec ces valeurs stratégiques évite les angles morts de l'IA. Cette prospection est importante pour la prise de décisions et pour une avantageuse interaction homme-machine. Pareillement, l'un des objectifs d'une IA de qualité n'est pas de statuer de manière binaire sur ces deux fonctions, mais de structurer une réflexion évolutive autour de la recherche de cet équilibre. Enfin, cette recherche devient un levier nécessaire qui va bien au-delà d'une prédiction précise avec seulement de l'explicabilité et de l'associatif.

Cependant, dans des contextes critiques avec peu de données et en utilisant les boîtes noires connexionnistes, il est difficile d'atteindre ces fonctions. De même, certaines boîtes ne déduisent que les corrélations. Toutefois, il y a des exemples intéressants de combinaison de la logique floue et du probabilisme (*PFL)* pour ces fonctions d'interprétabilité ainsi que pour des subtilités causales. Successivement, la technique d'hybridation et d'enrichissement du symbolique, peu faite et

[1] La corrélation n'est pas la causalité et la première n'implique pas la deuxième.

[2] La robustesse (généralisation) assure une bonne prédiction précise.

récente, avec cette combinaison, reste un exemple permettant d'aboutir à cette valorisation. Si bien que cette technique se fait miroir de ces valeurs pour un outil utile.

Également, pour simplifier le discernement du flou et du probabilisme, racontant l'histoire imaginaire d'un médecin spécialiste en septicémie. Dans un objectif d'acceptation valorisante et social, il est important pour lui d'interpréter clairement à ses patients les décisions de son système. De ce fait, il interprète la causalité dans le cadre du sepsis. En procédant de la sorte, l'objectif de ses analyses est d'interpréter les causes des effets de la bactérie A sur la prédiction avec un ordre de grandeur de 0,7 pour trois patients. Sur une première analyse, il définit que chaque patient a un degré d'appartenance de 0,7 pour A et 0,3 pour la bactérie B. Ensuite, il analyse en deuxième temps, aussi, que l'un de ces trois patients avait A alors que le reste a uniquement B. Donc, dans le monde des logiciens, pour la première analyse de ce récit, ce médecin a une logique floue. Dans la deuxième, il est probabiliste.

Revenant aux réalités des processus cognitifs du cerveau et dans l'imprécis ainsi que l'incertain, avec plein de surprises en erreurs, il réussit extraordinairement à interpréter et à bien expliquer les causalités du risque à travers des minimisations de ces surprises. Par la suite, ses interprétations incluent au moins des traitements d'incertitudes.

La différence entre les probabilités et la logique floue se trouve dans le type d'incertitude qu'elles modélisent : la première concerne l'incertitude associée à la survenance d'un événement, alors que la seconde s'intéresse à l'incertitude relative au niveau d'appartenance à un concept vague. L'incertitude représente une condition incomplète d'une information. C'est un terme couvrant toutes les manifestations d'ignorance, de flou et d'erreur. Autant, l'incertitude floue n'est pas le résultat de notre manque de connaissance sur l'observable, mais plutôt de la limite de notre langage ou de nos concepts à segmenter le monde en catégories strictement booléennes. Elle se produit lorsque les limites d'une catégorie ne sont pas clairement définies. La proposition n'est pas « vraie » ou « fausse », elle est plutôt « partiellement vraie ». L'incertitude persiste même après l'observation. Par ailleurs, l'incertitude probabiliste est due à une méconnaissance de l'état précis de l'observable, souvent associée à l'aléatoire. Elle représente la fréquence à laquelle un événement se produit ou la probabilité qu'une affirmation soit vraie. L'événement peut appartenir ou non à la classe, on ne sait simplement pas lequel des deux est le cas en ce moment.

S'y ajoute que, une incertitude stochastique, c'est une branche de l'incertitude probabiliste incorporant la dimension temporelle. Elle décrit des processus aléatoires en mouvement. Elle est inhérente au système physique temporel et ne peut être complètement éliminée par l'incorporation de nouvelles données.

Ces incertitudes sont reliées au raisonnement des observables vagues (comme pour la logique floue) avec l'ignorance et l'indétermination quant à la survenance d'un observable (comme pour le probabilisme). Cette ignorance représente une condition cognitive de manque partiel ou total de données. Devant une ignorance, il est difficile d'attribuer des niveaux d'appartenance ou des probabilités aux événements. Aussi, l'indétermination est une condition ne pouvant pas être définie ou résolue, non en raison d'un manque de données, mais parce que le système ne fournit pas d'information claire. À la différence de l'incertitude qui s'atténue grâce à l'observation, l'indétermination demeure persistante, même en présence de données parfaites.

Ensuite, dans une zone grise d'un contexte bien spécifique, il discerne les dissociations de plusieurs causes pour un seul effet. Ultérieurement, il simplifie ses modèles cognitifs en priorisant les observables causant l'effet. Il s'ensuit, pour un observable, une intellection du portrait global de ce contexte avec des détails chirurgicaux. Par conséquent, il prédit les menaces.

Tout bien considéré, en utilisant des exemples *PFL*, nous pourrons tenter de reproduire une partie des processus cognitifs d'analyses des nuances causales et des interprétations du cerveau. De surcroît, nous mettons en proposition le besoin d'exploiter la pertinence statistique pour la causalité afin de proposer l'enrichissement d'une interprétabilité intrinsèque. À l'image de Salmon (1984, 1998) qui avait mis en exergue la nécessité de saisir cette pertinence statistique et la causalité afin de proposer des explications scientifiques. En conséquence, notre étude de cas sur les données de la maladie du choc septique acquiescera à voir le côté cognitif en cherchant cet essai de représentation du cerveau. Puis, celui de l'informatique utilisant des algorithmes *PFL*. Autant, ces deux côtés s'entremêlent dans notre projet pour une solution de recherche d'équilibre avec une perspective de valeurs.

0.2 Le contexte

L'IA a radicalement modifié divers secteurs, y compris les domaines critiques, comme la santé, en promettant des progrès notables en matière de diagnostic et de personnalisation. Toutefois, malgré leurs performances remarquables, beaucoup d'IA, notamment les modèles opaques, font défaut en termes d'interprétabilité et d'analyse causale indispensables pour une intégration totale dans ces secteurs. Cette opacité provoque une suspicion justifiée parmi les experts et les utilisateurs demeurant dans l'incapacité de saisir les processus sous-tendant les décisions prises par l'IA. Il devient donc primordial de créer des IA non seulement dotées d'une bonne précision prédictive, mais aussi capables de proposer des interprétations claires, des études de causalité et des analyses intelligibles de leurs logiques. Saisir pourquoi un modèle indique un risque élevé ou préconise une action spécifique peut avoir un impact direct sur les décisions et potentiellement optimiser les résultats pour ces experts et ces utilisateurs. Cette étude est motivée par cette exigence basique : concevoir une IA qui soit un interlocuteur fiable, capable d'interagir avec l'expertise humaine en fournissant des arguments transparents et fondés, plutôt qu'une entité opaque dont les décisions sont acceptées sans discernement.

En parallèle, nous avons choisi le choc septique comme étude de cas. C'est une affection médicale grave avec une mortalité élevée, nécessitant des actions immédiates et exactes. Le choix d'administrer un ou plusieurs antibiotiques, par exemple, peut considérablement influencer le pronostic pour le patient. Néanmoins, l'interprétabilité et l'identification des causalités sont complexes en raison de la complication des interactions physiologiques et de la variabilité des réponses individuelles. L'impulsion majeure derrière cette étude réside dans le besoin d'avoir cette interprétabilité et de transcender les méthodes corrélationnelles et associatives pour adopter de récents modèles dans le but de favoriser des interventions plus précises et performantes. L'hybridation et l'intégration des probabilités dans le symbolisme flou présentent une approche prometteuse pour saisir cette impulsion.

0.3 La problématique de recherche

L'évolution rapide des IA dans des contextes sensibles a révélé une question prépondérante : comment allier l'hybridation, le symbolique et l'efficacité algorithmique aux exigences pressantes

d'interprétabilité et de causalité, surtout dans des situations où les décisions entraînent des répercussions vitales ? Les méthodes actuelles ont du mal à répondre à ses exigences, se limitant souvent à des explications, des associations, puis des corrélations sans une interprétabilité et une différenciation des relations de cause à effet, une distinction impérative pour une perspective de valorisation.

En outre, les données sensibles se distinguent par leur nature intrinsèquement marquée par une incertitude et une imprécision qui se définit comme l'impossibilité de fournir une valeur précise. Cela peut être attribué à un déficit d'information détaillée. Les modèles conventionnels, habituellement fondés sur une logique dichotomique, peinent à traiter l'ambiguïté du monde réel, ce qui entrave leur personnalisation, flexibilité et robustesse. Il est donc pressant de créer des systèmes IA qui tentent de reproduire aussi le processus de pensée humain. Ils devraient fournir des interprétations limpides, mesurer l'incertitude et soutenir les experts dans la compréhension des éléments qui sous-tendent les prédictions du risque (en particulier en ce qui concerne notre cas de mortalité chez les patients atteints de choc septique).

De surcroît, avec les grandes performances informatiques, les solutions *Deep Learning (DL)* rencontrent un succès croissant dans divers domaines, avec des résultats qui sont très précis et de grandes capacités de classification. Dès lors, ces modèles connexionnistes de *DL* ou de *Machine Learning (ML)* utilisent des paramétrages compliqués avec de grands traitements de calculs sur de gros volumes de données. Ensuite, si une solution est développée sur la base de ces modèles opaques et sur des données biaisées, le pourquoi du résultat de cette solution devient difficile à saisir. Successivement, cette boîte noire connexionniste ne permet pas de comprendre comment les prédictions sont faites. De la sorte, l'expert du contexte (tel un médecin) ne pourra pas comprendre le processus de prédiction et vérifier les résultats donnés par cette boîte. Autant, selon Benno Schwikowski, responsable du groupe Biologie des systèmes à l'Institut Pasteur, devant la complexité biologique de la maladie, il y a une difficulté de prédiction. Après quoi, il souligne que « l'apprentissage profond est très efficace pour donner la bonne réponse à une question donnée, mais on ne sait pas comment, par quel processus logique : c'est une boîte noire. Nous cherchons à « ouvrir » la boîte noire pour que les experts puissent interpréter les résultats

obtenus par une IA et lui faire confiance. »[3]. Assurément, le manque d'interprétabilité des modèles *ML* empêche les experts de faire confiance et d'accepter ces modèles comme pour les cliniciens (Chen et coll., 2022).

Ainsi, le manque de clarté des solutions informatiques dans le processus décisionnel suscite de plus en plus d'inquiétudes (Yang et coll., 2023). Selon Masís (2023), l'interprétabilité des modèles d'apprentissage automatique se heurte à plusieurs défis majeurs. Premièrement, il mentionne que la complexité intrinsèque de nombreux algorithmes rend l'analyse de leurs mécanismes internes particulièrement ardue, ce qui rend difficile la compréhension de leur fonctionnement. En plus, il rappelle que certains modèles sont opaques, ce qui signifie qu'il est impossible de voir comment ils prennent des décisions. Subséquemment, il évoque qu'un compromis apparaît fréquemment entre la performance et l'interprétabilité, et souvent, les modèles les plus précis sont les moins intelligibles. Aussi, le manque de transparence et de responsabilité des modèles prédictifs peut avoir (et a déjà eu) de graves conséquences et de mauvaises décisions (Rudin, 2019). Dans la continuité, le manque de validation externe, la mauvaise interprétabilité et la non-transparence des problèmes de boîte noire ont limité les applications cliniques des *ML* (Li et coll., 2023). Ensuite, à l'exemple du *DL* qui a permis d'atteindre, une performance dans de nombreux domaines médicaux (Mehnatkesh et coll., 2023), il y a des problèmes de transparence. Dans ce contexte clinique, les boîtes noires sont inacceptables (Shortliffe et coll., 2018). Tant, dans des domaines sensibles comme les soins de santé, le nucléaire, l'armement ou le commerce financier de grande envergure, ou la cybersécurité où la solution informatique peut avoir un impact sur des vies humaines, l'interprétabilité et la responsabilité ne sont pas seulement certaines propriétés souhaitables, mais aussi des exigences légales (Kaminski, 2019). De même, « les modèles ne sont pas adaptés à la prédiction individualisée des risques et aux approches plus ciblées (personnalisées) de la recommandation de traitement » (Rajkomar et coll. (2018) et Martens et coll. (2023) cités par Woodman et Mangoni (2023)). Pareillement, dans leur questionnement à savoir si l'IA devrait-elle être utilisée pour appuyer la prise de décision éthique clinique, Benzinger et coll. (2023) présentent une revue systématique des raisons où les biais et l'interaction homme-machine aient été mentionnés comme des problématiques. Ensuite, pour d'autres ambiguïtés, ils en citent d'auteurs comme Gundersen et Bærøe (2022) pour la

[3] https://www.pasteur.fr/fr/file/39734/download

responsabilité dans les décisions prises grâce à l'IA. De surcroît, « des inquiétudes concernant la confiance accordée à une « boîte noire » ont été exprimées » (Biller-Andorno et Biller, 2019). De plus, « l'automatisation de la prise de décision éthique par l'IA n'est actuellement ni faisable ni éthique » (Barwise et Pickering, 2022). Également, Ferrario et coll. (2023) pour le manque de transparence et d'explicabilité. Tout autant, selon Musanga et coll. (2025), dans un contexte clinique où la transparence et le suivi des actions sont pressants, ce défaut d'interprétabilité inhérente pose un problème majeur. Dès lors, selon eux, bien que des techniques d'explicabilité post-hoc (comme *Grad-CAM* et *LIME*) fournissent des éclaircissements précieux sur les prédictions des modèles, elles ne dévoilent pas intégralement le mécanisme de raisonnement interne qui guide ces choix.

Semblablement et selon Pearl, les *ML* (y compris ceux qui utilisent des *DL*) opèrent presque exclusivement de manière associative (Bishop, 2021). « Pearl identifie certains obstacles majeurs qui nuisent encore à la capacité de raisonnement des systèmes d'IA, une voie proche de celle des humains, à surmonter en équipant les machines d'outils de modélisation causale (Pearl, 2019). Parmi ces obstacles, c'est le manque de robustesse des systèmes d'IA pour reconnaître ou répondre à de nouvelles situations sans étant spécifiquement programmés (c'est-à-dire l'adaptabilité) ainsi que leur incapacité à saisir les relations de cause à effet. Au lieu de cela, ces capacités sont des caractéristiques innées des êtres humains pouvant communiquer avec, apprendre et instruire et où tous leurs cerveaux raisonnent en termes de relations de cause à effet (Pearl, 2018). » (Carloni et coll., 2025). De même, l'élaboration d'une explication scientifique à travers la pertinence statistique et la causalité (Salmon, 1984, 1998, 2006, 2010) permet à l'IA une meilleure aide décisionnelle et gestion de l'implication de l'expert. Dans l'analyse d'un modèle intégrant une part de causalité, une prise de décision médicale améliorée et une capacité à travailler avec moins de données spécifiques, Llored & Bouchnita (2022) ont noté que « les IA alliées aux modèles conceptuels restent des grands comparateurs, des outils d'identification ou de regroupement de similarités ; outils qui ont un pouvoir heuristique puissant malgré leurs limites dans la mesure où le pouvoir explicatif et le lien causal sont, et resteront, pour l'essentiel, perdus (sauf pour les IA symboliques dont les utilisations sont très limitées). Cette alliance pourrait permettre de renforcer les dispositifs d'assistance qui aident un médecin dans l'étape qui consiste à établir un diagnostic. ». Aussi bien, selon Pearl (2018), l'analyse causale pourra être un outil pour résoudre certaines limites actuelles du *ML*. Sans une correspondance claire entre

les principes, les pratiques et les mécanismes de responsabilité, les discours autour de l'« IA éthique » pourraient se réduire à des engagements symboliques, incapables de relever les défis sociotechniques complexes que posent ces systèmes (Jobin et coll., 2019). De plus, dans leur article, traitant de la responsabilité causale en utilisant des comparaisons de probabilités, Qi et coll. (2024) ont écrit que : « alors que l'intégration de l'IA continue de se développer dans des secteurs tels que la santé et l'éducation, la nécessité d'une attribution précise des responsabilités devient de plus en plus importante, en particulier pour les résultats indésirables qui peuvent avoir un impact sur la confiance, la responsabilité et le déploiement éthique de l'IA. »

A fortiori, selon Salmon, la causalité et la pertinence statistique sont utiles pour fournir une explication scientifique (Woodward, 1989). Mais, « les algorithmes de l'IA actuels ne permettent pas d'élaborer une explication scientifique au sens de Salmon ni d'établir, par comparaison, l'existence d'une cause commune. Également, ne sont-ils pas de bons candidats pour ce type de modèle de l'explication scientifique. » (Llored & Bouchnita, 2022). Identiquement, les explications des résultats des solutions connexionnistes ne sont pas fidèles aux mécanismes causaux (Lipton, 2018). En plus, « traditionnellement, l'inférence causale est axée sur des causes uniques » (Mehta, 2020). De même, selon Broadbent et Grote (2022), « Lin et Ikra[m] sont clairs sur le fait que les « techniques » d'apprentissage automatique peuvent fournir un support informatique : elles ne peuvent pas contribuer à « la définition et à l'identification de la question causale », et « l'apprentissage automatique en soi est par définition insuffisant pour déduire la causalité » (Lin et Ikram, 2020, 184) ». Par ailleurs, « les méthodes modernes d'apprentissage automatique sont diamétralement opposées : ils trouvent des corrélations idiosyncrasiques [4] entre les caractéristiques des données, mais ne révèlent pas d'effets causaux. » (Mehta, 2020).

Viens ensuite l'exemple du choc septique où PhysioNet a organisé un défi informatique en 2019 pour sa détection (Reyna et coll., 2019), cependant, malgré leur capacité d'optimisation, la plupart des algorithmes *ML*, en particulier les *DL* de pointe n'étaient pas efficaces pour expliquer ce qu'ils ont appris et encore moins pour élucider les facteurs utilisés par le modèle et ainsi de prendre la décision d'une manière interprétable (Strickler et coll., 2023). Ainsi, l'exemple de l'application

[4] « Relatif aux caractéristiques propres à chaque individu, qui le distinguent des autres et qui déterminent sa façon particulière de réagir à son milieu et aux agents extérieurs. » https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8364359/idiosyncratique#:~:text=D%C3%A9finition,milieu%20et%20aux%20agents%20ext%C3%A9rieurs.
En anglais "peculiar" (mettant en avant un caractère étrange ou unique) ou "specific" (soulignant une singularité).

d'une IA sur les données du choc septique est une étude de cas intéressante. Sans compter que, la mortalité causée par la septicémie est très importante et cela nécessite une attention urgente (Rudd et coll., 2020). Récemment, de nombreux systèmes de notation ont été conçus pour prédire le pronostic des patients atteints de septicémie, notamment le *Systemic Inflammatory Response Syndrome (SIRS), Mortality score in Emergency Department (MEDS), Sepsis Severity Score (SSS), Predisposition/Insult-infection/Response/Organ dysfunction (PIRO), National Early Warning Score (NEWS), Acute Physiology And Chronic Health Evaluation score (APACHE II), Modified Early Warning Score (MEWS), Sequential Organ Failure Assessment (SOFA),* et *quick SOFA (qSOFA)* (Subbe et coll., 2001, Rubulotta et coll., 2009) cités par Yapici et coll., 2022)), mais « aucun des scores de gravité de la maladie existants et largement utilisés, spécifiquement développés pour les patients gravement malades, n'était cliniquement utile pour l'identification des patients présentant une infection en cours » (van Ruler et coll., 2011). Par exemple, le score *SOFA* peut être utilisé pour l'évaluation rapide des patients atteints de sepsis, mais il n'est pas assez sensible pour évaluer sa gravité (Zhang et coll., 2023). Ensuite, Maryani et Wisudarti (2023) ont effectué la comparaison des caractéristiques et des valeurs des systèmes *Sepsis Patient Evaluation Emergency Department Score (SPEEDS), MEDS, SOFA, APACHE II,* et *Simplified Acute Physiology Score (SAPS II)* dans les articles scientifiques publiés en anglais entre 2012 et 2022[5]. Cependant, ils estimés qu'ils sont inefficaces et non rentables, ce qui rendait le score *SOFA* défavorable pour améliorer la qualité des soins de santé.

In fine, notre problématique de recherche se résume au manque d'interprétabilité, de causalité et de perspective de valorisation de l'IA dans un contexte sensible avec de rares données. Les méthodes actuelles se limitent à des explications, des associations et des corrélations statistiques. C'est pourquoi, pour cette perspective, nous visons une contribution comblant cette lacune avec l'hybridation et l'enrichissement des approches symbolistes par le recours à la *PFL*.

[5] Ces auteurs ont découvert que chacun de ces cinq scores conduisait à la prédiction de mortalité chez les patients atteints de sepsis. Ils déterminent que *MEDS* a prédit la mortalité de ces patients mieux que *SAPS II* après 28 jours, mais le *SPEEDS* était plus précis que *MEDS*. Pour eux, la note *SOFA* prédisait mieux la mortalité que l'*APACHE II*. Et, *APACHE II* a une validité moindre que *SAPS II*. Ils estiment que les scores du *AUC SOFA* sont plus importants pour diagnostiquer cette mortalité que d'autres scores.

0.4 La justification

Notre recherche et sa structure reposent sur la conviction que l'IA ne sera véritablement valorisante et ne sera adoptée dans des contextes critiques, tels que la médecine que si elle peut faire plus que simplement prédire ; elle doit être interprétable, causale et valorisante. Il est vital d'interpréter les raisons de notre approche. Malgré leur efficacité, les IA en boîte noire engendrent un écart en matière d'intellection et de valeurs. Un expert ne peut se fier à une décision sans en saisir l'enchaînement sous-jacent, surtout quand des vies sont en jeu (comme dans l'exemple du choc septique). Notre mission trouve son sens dans le besoin de pallier cette lacune en développant une IA capable de tenter de représenter le raisonnement de manière transparente, identifiant autant que possible les corrélations par rapport aux causalités et en traitant l'incertitude inévitable des données sensibles grâce à la *PFL*.

Aussi, cette méthode nous donne la capacité de produire des règles « *IF-THEN* » intelligibles fournissant une augmentation au processus de pensée. Il s'agit d'un effort visant à concevoir une IA personnalisable et éthique assistant l'intelligence humaine tout en contribuant à identifier et à minimiser les risques de façon proactive reposant sur une compréhension réciproque entre l'homme et la machine. Cette motivation oriente tous les aspects de cette thèse du choix des méthodes à la présentation et l'interprétation des résultats avec toujours la visée de promouvoir une adoption sûre et performante de l'IA dans un contexte critique.

De plus, l'objectif de ce travail de recherche est d'exploiter la possibilité de développer des systèmes d'IA capables non seulement de faire des prédictions, mais aussi d'aider à réfléchir sur les causes et les conséquences. Dans le secteur sensible de la médecine où chaque choix a un impact direct sur la santé des patients, il est non seulement souhaitable, mais majeur d'avoir une IA qui peut détecter et minimiser les risques dans un contexte d'incertitude. L'utilisation de la logique floue, spécifiquement, offre la possibilité de symboliser les nuances et l'ambiguïté des notions médicales (comme « bas » ou « haut » taux d'une variable) de manière plus naturelle et limpide.

0.5 L’importance du sujet

Ce projet s'aligne sur l'initiative de rendre l'IA plus intelligible, en se concentrant précisément sur l'incorporation du *PFL* dans les modèles d'IA pour renforcer leur interprétabilité, leur robustesse et leur aptitude à établir des liens causaux. Le champ d'application sélectionné, le choc septique est particulièrement pertinent du fait de sa sévérité et de l'exigence d'une intervention rapide et bien informée. Avec sa capacité à représenter l'incertitude des données et des idées (comme un nombre « élevé » de globules blancs au lieu d'un seuil précis), la logique floue associée aux probabilités pour mesurer les incertitudes de façon rigoureuse constitue une approche efficace pour tenter de représenter le raisonnement humain.

Également, cette collaboration donne lieu à des « *IF-THEN* » qui sont non seulement intelligibles, mais qui assurent également un équilibre entre la précision de l'algorithme et l’intellection humaine. Ce sujet est significatif pour développer une IA adaptable, c'est-à-dire une technologie respectant un ensemble de valeurs principales pour une intégration réussie dans des contextes sensibles. Ces critères englobent la clarté, l'interprétabilité, la robustesse, l'aptitude à la généralisation, l'habileté a essayer de symboliser le raisonnement ainsi que la transparence pour davantage de : responsabilité, confiance, éthique, conformité aux normes, réglementation stricte, sécurité renforcée et frugalité avec souveraineté dans une solution vérifiable et certifiée.

Autant, cette IA ne se limiterait pas à perfectionner les diagnostics et prédictions, elle renforcerait aussi la confiance des médecins et des patients en proposant des justifications claires, traçables et argumentées de ses actions, transformant ainsi l'IA en un véritable partenaire clinique au lieu d'un simple outil opaque.

De plus, cette étude se situe à l'intersection de l'IA et de la médecine clinique. L'enjeu du sujet est de pouvoir révolutionner la prise de décisions médicales en proposant des renseignements causaux au lieu de simples corrélations. Une connaissance approfondie des processus sous-jacents aux maladies complexes, via des modèles causaux intelligibles, pourrait aboutir à des approches thérapeutiques plus individualisées et à une minimisation des dangers pour les patients. Dans le contexte actuel où l'interprétabilité des modèles d'IA est désormais une obligation éthique et réglementaire, cela revêt une importance accrue.

Par conséquent, nous envisageons de tenter de représenter la dimension cognitive du raisonnement par un ensemble d'hybridation et d'enrichissement (sophistication) du symbolisme. De ce fait, nous voulons déterminer comment pour une IA cet ensemble améliore la capacité d'interpréter les prédictions et de distinguer les causalités des corrélations dans une perspective de valorisation. Dès lors, ce questionnement va guider notre approche. Nous estimons que cet essai de simulation des processus cognitifs avec la *PFL* diminue les risques, découvre de nouvelles connaissances d'une manière valorisante et permet de prendre des décisions plus éclairées en utilisant l'analyse causale avec l'*IAI*. De la sorte, cela rend possible la démonstration, via une étude de cas sur le choc septique que l'intégration de l'analyse causale et de l'*IAI* permet de valoriser la solution comme un outil de collaboration Homme-Machine. Nous souhaitons prouver que cette approche facilite la prise de décision en contextualisant les prédictions et en révélant les mécanismes causaux sous-jacents renforçant ainsi la confiance des experts et une prise en charge personnalisée des patients. Subséquemment, cela se matérialise dans notre contribution à une nouvelle vision de l'IA transcendant la simple corrélation pour s'engager dans une analyse causale plus profonde et intelligible. Dès lors, en s'inspirant du raisonnement, notre travail enrichit l'*IAI* en offrant une analyse causale. Cela permet non seulement de valider théoriquement l'approche, mais aussi de poser les fondations d'une IA valorisante.

Parallèlement, l'informatique de la *PFL* est au service de la causalité et de l'interprétabilité. Ainsi, nous concentrons notre travail sur la conceptualisation des *PFL* intégrant du probabilisme et du flou pour traiter les nuances causales, générer des prédictions précises, et produire des règles interprétables. Pour y parvenir, nous comptons valider qu'une *PFL* de fuzzification basée sur un clustering et un modèle arborescent dont les entrées sont enrichies par les scores flous et les probabilités des clusters, permet une recherche d'équilibre robuste, flexible, et généralisable en obtenant un bon rapprochement d'une harmonie de précision, de causalité, et de règles interprétables. Comme cela, cette approbation est la clé pour implémenter un *PFL*-Arborescent capable de générer des « *IF-THEN* » interprétables. Notre but est de parvenir à un équilibre entre de bonnes précisions prédictives et une calibration des probabilités de risque (avec l'exemple du choc septique). Aussi, ce doit être un modèle personnalisé et ajusté pour s'adapter à des contextes sensibles pour une perspective valorisante. En procédant ainsi, l'implémentation réussie de cet objectif mène directement à notre présentation d'une hybridation utile combinant

un clustering pour la fuzzification, un modèle causal, un autre arborescent produisant des « *IF-THEN* » ainsi que des métriques pour évaluer la causalité et l'*IAI*.

Tout bien considéré, la symbiose et la progression des liens cognitifs et informatiques forment une chaîne logique cohérente où chaque élément dépend du précédent. Successivement, les réponses à nos questions dépassent la corrélation et l'explicabilité. Ultérieurement, la validation de nos hypothèses permet de considérer la *PFL* comme un miroir de la cognition. Par la suite, nos objectifs acquiescent à une IA valorisante. Au bout du compte, nos contributions enrichissent l'interprétabilité avec de la causalité.

0.6 Les questions de recherche : Dépasser la corrélation et l'explicabilité

Pour orienter notre recherche, nous nous sommes inspirés des processus cognitifs du cerveau en vue d'améliorer l'interprétabilité (*IAI*) des prédictions et de déchiffrer les causalités pour la prise de décisions dans un contexte marqué par l'incertitude et l'imprécision. Par conséquent, nous visons à répondre aux questions ci-après :

Question cognitive

La question cognitive (Qc) cherche à déterminer comment une IA d'hybridation et d'enrichissement (sophistication) du symbolisme teste la reproduction du raisonnement humain en améliorant la capacité d'interpréter les prédictions et distinguant les causalités des corrélations dans une perspective de valorisation ?

Question informatique

La question informatique (Qi) se concentre sur comment conceptualiser une *PFL* intégrant le probabilisme avec du flou pour traiter les nuances causales, générer des prédictions précises et produire des règles interprétables ?

Éclaircissements de la symbiose Qc-Qi

Concernant notre aspect cognitif, l'avancement de l'IA pourrait-il nous permettre d'élaborer une solution qui tente de représenter le raisonnement de façon plus raffinée se rapprochant davantage de l'expertise humaine ? Plus précisément, dans quelle mesure un modèle *PFL*-Arborescent enrichi de symbolisme et d'hybridation (basé sur une étude de cas en choc septique) peut-il renforcer l'interprétabilité de ses prédictions et la différenciation entre les relations causales et les simples corrélations ? Ce défi majeur dans l'analyse de données complexes et sensibles pourrait donc contribuer à une meilleure intellection, interaction et confiance des experts, leur offrant la possibilité d'appréhender et de minimiser les risques de façon plus proactive et éclairée en cas de situations incertaines et vagues. De plus, cet exemple de *PFL* représente-t-il une démarche de prospection d'équilibre et de maximisation de valeurs stratégiques ?

D'un point de vue informatique, comment élaborer et implémenter une IA symbolique et hybride qui fusionne les théories probabilistes et les concepts de logique floue avec une architecture arborescente ? Cette méthode devrait être apte à gérer efficacement l'incertitude et l'imprécision des données importantes (en étudiant l'exemple du choc septique), tout en produisant des prédictions robustes et des « *IF-THEN* » compréhensibles. De plus, quel effet l'intégration de ces fonctions d'interprétabilité intrinsèque (comme les scores flous et les probabilités des clusters) aurait-elle sur les performances prédictives du modèle, mais également sur la concordance entre les prédictions et les interprétations fournies par le modèle, ainsi que sur la crédibilité des probabilités (dans le cas du diagnostic et pronostic du choc septique) ?

Par la suite, dans le cadre de l'analyse causale, on se questionne sur la manière dont la combinaison de la logique floue et des méthodes probabilistes peut optimiser l'identification et la mesure des effets causaux de l'administration d'antibiotiques sur le taux de mortalité en situation de choc septique. Quels sont les éléments confondants les plus significatifs dans le lien entre l'administration d'antibiotiques et le résultat pour le patient et comment la logique floue peut-elle apporter des nuances à leur influence ? Finalement, à quel point l'approche suggérée facilite-t-elle une meilleure compréhension des résultats causaux comparativement aux techniques conventionnelles ?

Également, nous ne pouvant pas dissocier ces questions, qu'elles relèvent de la cognition ou de l'informatique. Une meilleure intellection du système (Qi) simplifiera la saisie de la façon dont une IA peut tenter de reproduire le processus de pensée en interagissant avec un utilisateur (Qc). L'élaboration des solutions technologiques est orientée par les contraintes cognitives. La question testant l'aptitude d'un modèle d'IA à cette tentative de reproduction pour soutenir cette vision encourageante reflète les inquiétudes grandissantes liées aux valeurs et à la société en ce qui concerne l'IA. Des travaux récents mettent en lumière l'importance de travailler avec des spécialistes d'autres domaines pour traiter ces problématiques.

Simultanément, l'aspect informatique, centré sur la création d'un système hybride, complexe et symbolique capable de gérer l'incertitude et de produire des « *IF-THEN* » intelligibles, est intrinsèquement lié à l'avancement des travaux dans le domaine de l'*IAI*. L'intégration des techniques *PFL* dans les modèles informatiques suggérés vise à créer des architectures robustes, aptes à traiter des données défectueuses, un enjeu pivot dans des secteurs clés, tels que la santé. Ce lien réciproque est fondamental : la Qi offre les instruments techniques pour traiter la question cognitive, alors que les besoins cognitifs stimulent l'élaboration de solutions informatiques plus mixtes et avancées.

0.7 Les hypothèses : La *PFL* est-elle un miroir de la cognition ?

Notre travail se fonde sur l'idée que l'on peut doter les IA de la capacité à interpréter et analyser les causalités en s'inspirant des processus cognitifs. Dès lors, nous formulons nos hypothèses comme suit :

Hypothèse cognitive (Hc)

L'hypothèse cognitive postule qu'en testant la simulation des processus cognitifs par l'hybridation et l'enrichissement du symbolisme, la *PFL* peut diminuer les risques, découvrir de nouvelles connaissances d'une manière valorisante et prendre des décisions plus éclairées en utilisant l'*IAI* avec l'analyse causale.

Hypothèse informatique (Hi)

On suppose qu'une *PFL* de fuzzification basée sur un clustering et un modèle arborescent dont les entrées sont enrichies par les scores flous et les probabilités des clusters permettra une recherche d'équilibre robuste, flexible et généralisable en obtenant un bon rapprochement d'harmonie de précision, de causalité et de règles interprétables.

Éclaircissements de la symbiose Hc-Hi

L'intellect humain a la capacité de généraliser dans divers contextes et il est simultanément robuste et adaptable. Il cherche à obtenir précision et équilibre dans son processus de réflexion, d'interprétation et d'analyse des nuances causales. L'analyse de causalité et l'interprétabilité sont des processus cognitifs. En suivant l'exemple de ces processus, notre objectif est de répondre à nos questions de recherche en soutenant l'hypothèse principale : grâce à l'essai de représenter ces processus, l'hybridation et la sophistication du symbolisme pourrait favoriser une prise de décision plus éclairée pour réduire les risques et acquérir de nouvelles connaissances valorisantes. Également, la modélisation causale pourrait constituer un début de réponse à l'insuffisance de raisonnement en IA, d'adaptabilité et de robustesse des IA. L'interprétabilité et l'inférence causale comme méthode scientifique de l'analyse causale sont étroitement connectées. Subséquemment, notre Hc postule que l'interprétabilité et la causalité favorisent une tentative de représenter le processus de pensée intelligible, transparent, responsable et fiable, en envisageant une IA qui soit sécurisée, réglementée, vérifiable, certifiable, éthique, économe et souveraine. Nous présumons que le progrès du symbolisme dans l'IA va significativement améliorer l'efficacité des modèles pour cette tentative de reproduction de la pensée, particulièrement en permettant une différenciation plus précise entre causalité et corrélation dans l'analyse des données (avec notre étude de cas sur le choc septique). Cette approche, qui gère intrinsèquement l'inexactitude et l'ambiguïté des données, donnera lieu à des interprétations de modèle plus intuitives et fiables pour des experts. Ainsi, nous anticipons que cette compréhension approfondie renforcera leur confiance dans les systèmes d'IA permettant une gestion plus efficace des cas (comme c'est l'exemple des patients atteints de choc septique).

Tout autant, notre Hi soutient que l'hybridation du symbolisme avec un exemple simple de *PFL* à travers une solution informatique permet de reproduire certains mécanismes cognitifs relatifs à l'analyse causale et le discernement des prédictions. Cette solution combinant la probabilité et la logique floue permet de rechercher un équilibre dans une IA causale à travers l'analyse de cas pratiques basés sur des données médicales réelles. Elle est à la fois adaptative et généralisable.

Aussi, nous avançons que l'élaboration et la mise en œuvre d'un système d'IA sophistiqué et hybride, combinant un module de fuzzification basé sur le clustering (ex. *GMM*) et un modèle arborescent (ex. *LightGBM*) intégrant les scores flous et les probabilités des clusters comme entrées permettra d'atteindre un équilibre satisfaisant entre les précisions prédictives, l'interprétabilité et la causalité. Plus précisément, nous escomptons que ce modèle conserve une précision notable tout en générant des « *IF-THEN* » dénormalisés et intelligibles décodant les prédictions de risque. En outre, nous anticipons que cette structure permettra d'ajuster les probabilités de risque, d'assurer une robustesse face aux données imprécises ou perturbées et de généraliser efficacement les résultats à de nouveaux environnements sensibles, renforçant ainsi les valeurs d'intelligibilité, de transparence et de confiance pour l'IA en milieu critique.

En dernière analyse, les hypothèses Hc et Hi se complètent mutuellement parce que leur implémentation informatique permet de vérifier et de confirmer que l'IA peut effectuer une tentative de représentation du processus cognitif de raisonnement. Les suppositions exposées s'insèrent dans un courant visant à réduire l'écart entre les modèles d'IA obscur et l'intellection. L'Hc suggérant que la reproduction des mécanismes cognitifs tels que la généralisation et l'analyse causale peut être bénéfique. Elle est soutenue par les études récentes dans le domaine des sciences cognitives et de l'informatique. L'Hi, qui offre une solution *PFL*-Arborescent pour simuler ces processus, s'appuie simultanément sur les recherches en inférence causale et sur les architectures hybrides récentes. Le croisement des deux suppositions crée un terrain propice d'avancées technologiques en hybridation (Hi) servant de test pour confirmer les hypothèses concernant la capacité de l'IA à tenter de représenter le raisonnement (Hc) tout en se dirigeant vers des systèmes plus robustes, adaptables et éthiques.

0.8 Les objectifs : Vers une IA valorisante

Objectif cognitif (Oc)

L'Oc est de démontrer, via une étude de cas sur le choc septique, que l'intégration de l'analyse causale et de l'*IAI* permet de valoriser la solution comme un outil de collaboration Homme-Machine. Nous souhaitons prouver que cette approche facilite la prise de décision en contextualisant les prédictions et en révélant les mécanismes causaux sous-jacents renforçant ainsi la confiance des experts et une prise en charge personnalisée des patients.

Objectif informatique (Oi)

L'Oi est une implémentation d'un *PFL*-Arborescent capable de générer des « *IF-THEN* » interprétables. Notre but est de rechercher un équilibre entre de bonnes précisions prédictives et une calibration des probabilités de risque (avec l'exemple du choc septique). Aussi, l'Oi devra être un modèle personnalisé et ajusté pour s'adapter à des contextes sensibles pour une perspective valorisante.

Éclaircissements de la symbiose Oc-Oi

L'interprétabilité se situe à l'interface des sciences cognitives et de l'informatique (Miller, 2019). En nous appuyant sur un exemple concret du choc septique, nous visons à mettre en lumière l'importance d'une prédiction précise du risque et l'entendement des subtilités causales. La solution s'appuyant sur des exemples de *PFL* favorise la recherche d'un équilibre et la maximisation de valeurs stratégiques. Cette solution permettra à l'expert de mieux prendre ses décisions en considérant un contexte d'incertitude et d'imprécision.

Également, l'Oc cherche à illustrer la puissance de l'hybridation et de l'enrichissement symbolique avec la *PFL* afin d'améliorer l'interprétation des prédictions et l'analyse des subtilités causales. Cette approche détermine les signaux de causes, forts, modérés et faibles, afin de simplifier la caractérisation des risques et d'éviter les alertes erronées. En conjuguant deux perspectives (floue et probabiliste), elle tente de reproduire certains mécanismes cognitifs du raisonnement. Ceci

offre une vision renouvelée de l'IA en tant qu'outil d'interprétation des décisions et des prédictions, tout en étant un dispositif collaboratif qui ne se substitue pas à l'expert. On tient également compte de la dimension éthique et sociale de l'IA. Un autre but est d'offrir un accompagnement personnalisé avec une boîte moins opaque.

Ainsi, notre but est de vérifier que l'enrichissement du symbolisme par la création des « *IF-THEN* » intelligibles et l'évaluation des degrés d'appartenance floue augmente la capacité de l'IA à distinguer les corrélations des causalités, un critère majeur pour une prise de décision éclairée. Cette méthode vise à synchroniser l'IA avec le processus de pensée des experts, où les choix sont fondés sur une évaluation graduelle des symptômes plutôt que sur des paramètres dichotomiques. La combinaison des approches probabilistes et de la logique floue aspire à élaborer une IA plus compréhensible, transparente et digne de confiance, capable d'offrir des interprétations reflétant l'expertise humaine, contribuant à une meilleure appréhension et atténuation des risques (dans des situations critiques telles que le choc septique).

De même, les Oi se définissent par l'élaboration de modèles probabilistes et flous basés sur les règles *« IF-THEN »*. L'emploi des modèles *PFL* permettra aux spécialistes de mieux saisir la causalité et l'interprétabilité des prédictions dans une perspective de valorisation. Cette simulation, qui fait appel à divers paramètres, générera de nouvelles informations avec différentes règles en fonction des regroupements (clustering), ce qui permettra un meilleur entendement des dynamiques du risque et d'amélioration de la gestion. Les prototypes seront conçus pour être évolutifs, modulables et adaptables à tout autre contexte critique.

Par conséquent, l'objectif est de créer, de mettre en œuvre et d'évaluer un modèle sophistiqué et hybride tirant parti des atouts complémentaires des méthodes probabilistes et de la logique floue. Le modèle *PFL*-Arborescent est élaboré afin d'accroître la précision et l'intellection dans un contexte sensible (tel que le choc septique). Cela nécessite l'implémentation d'un module de fuzzification utilisant le *BGMM* pour convertir des variables en niveaux d'appartenance flous. Aussi, des caractéristiques floues viendront compléter un modèle arborescent (*LightGBM*) afin de tirer parti de sa capacité de précision tout en lui offrant une tentative de raisonnement symbolique. L'élaboration d'algorithmes capables de produire des « *IF-THEN* » intelligibles directement à partir des centroïdes de clusters flous et des probabilités conditionnelles est un

enjeu intéressant. Nous avons également pour objectif de mettre en place des méthodes de calibration afin d'assurer la crédibilité des probabilités de risque prévues. De plus, le dispositif comprendra des seuils cliniques prédéfinis et un mécanisme d'alerte pour détecter les incohérences, ce qui garantira sa robustesse. L'intention est de prouver que le système peut conserver une bonne performance prédictive tout en offrant des analyses interprétables, vérifiables et certifiables, contribuant ainsi à une IA plus transparente et responsable.

Tout bien considéré, valoriser l'interprétabilité et les subtilités causales pour une prise de décision informée répond à l'exigence grandissante d'une IA bénéfique. L'idée de créer un modèle hybride *PFL*-Arborescent apte à produire des règles « *IF-THEN* » et à ajuster les probabilités répond au challenge technique imposé par l'exigence d'une IA éthique. L'objectif du projet est de concevoir une IA qui non seulement propose des réponses (Oi), mais qui offre également les justifications de ces réponses (Oc) en combinant ces finalités. Cela encourage une coopération et une communication plus fructueuses entre le spécialiste et la machine dans le processus de prise de décision.

0.9 Les contributions : Au-delà de l'enrichissement de l'interprétabilité avec de la causalité

Cette recherche propose des contributions informatiques qui tentent de reproduire le cognitif, repoussant les limites des analyses de corrélation et d'explicabilité.

Contribution cognitive (Cc)

La contribution cognitive contribue à une nouvelle vision de l'IA transcendant la simple corrélation pour s'engager dans une analyse causale plus profonde et intelligible. Ainsi, en s'inspirant du raisonnement, notre travail enrichit l'*IAI* en offrant une analyse causale. Cela permet non seulement de valider théoriquement l'approche, mais aussi de poser les fondations d'une IA valorisante.

Contribution informatique (Ci)

La contribution informatique est de présenter une hybridation utile combinant un clustering pour la fuzzification, un modèle causal, un autre arborescent produisant des « *IF-THEN* » ainsi que des métriques pour évaluer la causalité et l'*IAI*.

Éclaircissements de la symbiose Cc-Ci

Cette recherche offre une analyse causale susceptible de surmonter certaines limites existantes dans le domaine du *ML* (Pearl, 2018). Nous aspirons à participer à l'identification de mécanismes causaux à partir de données observées, une tâche qui selon Morandé et Tewari (2023) marque un tournant paradigmatique pour le *ML* en fournissant des capacités de généralisation plus robustes. Nous travaillons à doter les systèmes d'IA de compétences en analyse causales pour améliorer leur fiabilité, leur adaptabilité et leur crédibilité. Nous croyons que cette tâche contribuera à l'intégration de la découverte des relations causales en démontrant un potentiel pour augmenter la précision des modèles, faciliter une extrapolation précise et améliorer la résilience face aux changements. Il est primordial de faire preuve d'innovation dans le domaine de la découverte causale et d'évaluer avec rigueur les améliorations en matière de généralisation. Pour faire face au défi des boîtes noires en *ML* et *DL*, nous privilégions une approche basée sur les instruments d'explication des relations de cause à effet.

Également, il y a peu de prototypes qui allient le flou et l'approche probabiliste et ils sont tous très récents. Notre tâche s'insère dans cette approche de fusion, qui, d'après Faghihi et coll. (2020) procure davantage de flexibilité. La solution *PFL* offre la possibilité d'analyser les subtilités causales et l'essai de génération du mode de pensée face à l'incertitude. Ainsi, dans notre analyse de cas, le *PFL* facilite l'explication causale.

Autant, les utilisateurs exigent une valorisation de l'interprétation et de l'analyse causale pour donner leur approbation à l'IA. Dans nos modèles, les contributions se répartissent en deux éléments : cognitif et informatique. Cognitivement, l'intégration du symbolisme enrichi par la *PFL* permet de tenter de reproduire certains mécanismes du raisonnement dans ses prédictions exactes. Dans le domaine de l'informatique, nos objectifs sont d'optimiser l'interaction entre le

spécialiste et la machine pour une valorisation accrue, en mettant l'accent sur l'interprétabilité des prédictions et le discernement causal. Ces apports favorisent aussi la révélation de nouvelles connaissances entre les variables et l'acquisition de nouvelles connaissances.

De plus, d'un point de vue cognitif, cette recherche valorisera notre compréhension de la façon dont les systèmes d'IA peuvent transcender une simple corrélation pour aboutir à une analyse causale plus intelligible, aspect central pour leur adoption et leur crédibilité dans des contextes sensibles. En intégrant le *PFL*, nous démontrerons comment l'IA peut traiter l'incertitude et l'imprécision des données sensibles (comme celles du domaine médical) de façon plus sophistiquée et contextualisée, à l'image du fonctionnement du cerveau humain. Cette approche donne au modèle la faculté de penser en termes d'éléments indéfinis, tels que « faible valeur » pour des paramètres nécessaires et représente une base théorique pour le développement d'IA plus réfléchies, aptes non seulement à prévoir les événements (comme la mortalité liée à un état septique), mais aussi de fournir des interprétations sur les raisons et les modalités de ces anticipations.

Aussi, en fournissant des analyses « *IF-THEN* » interprétables et des évaluations de risque adaptées, nous aspirons à supprimer l'équivoque et l'incertitude inhérentes aux prédictions en boîte noire, ce qui simplifie le discernement et la réduction des risques pour les experts. Cela trace la route vers une IA plus valorisante dans laquelle les décisions capitales reposent sur des logiques transparentes et contrôlables augmentant de ce fait la confiance des experts. En outre, d'un point de vue cognitif, elle enrichit l'entendement des processus causaux (ex. associés au choc septique et à la réaction aux antibiotiques) offrant une méthode plus précise que les recherches corrélationnelles.

De même, d'un point de vue technique, cette étude met en avant une structure sophistiquée et hybride intégrant le clustering pour la fuzzification des données, ainsi qu'un modèle décisionnel dont l'efficacité et le discernement sont intrinsèquement optimisés par l'incorporation des scores d'affiliation flous et des probabilités liées aux clusters. Nous aspirons à démontrer que l'établissement d'un cadre capable de générer des « *IF-THEN* » lisibles et dénormalisés directement dérivés de la structure du modèle simplifie les complexités inhérentes aux modèles en un format compréhensible pour les experts. Cette contribution comprend aussi l'élaboration

de métriques pour évaluer la correspondance entre les prédictions du modèle et les interprétations générées, offrant ainsi des instruments innovants pour apprécier la valeur de l'interprétabilité.

De surcroît, le système intégrera des procédures d'étalonnage afin d'assurer la fiabilité des probabilités prévues ainsi que des limites et des alertes pour une utilisation pratique. En fin de compte, nos méthodes renforcent la robustesse de l'IA et sa faculté à être appliquée dans des situations sensibles. Dans la partie informatique, nous allons exposer une architecture d'IA qui associe la robustesse des modèles probabilistes à l'adaptabilité du flou, fournissant ainsi une méthode pour la modélisation causale en contexte d'incertitude. Les apports de cette étude participent à une évolution de l'IA, s'éloignant d'une simple corrélation pour privilégier la causalité.

Au final, l'introduction d'une nouvelle structure technique (Ci) qui génère des « *IF-THEN* » ainsi que des indicateurs pour quantifier la correspondance entre prédictions et analyses participe directement à l'avancement d'une IA plus éthique, vérifiable et claire. *In fine*, les Cc et Ci collaborent pour faire avancer l'IA vers des systèmes fiables, en fournissant des instruments techniques (Ci) qui favorisent une meilleure compréhension et une interaction avec les experts (Cc). Ces apports imbriqués cherchent à mettre en place des systèmes d'IA plus sûrs et éthiques, aptes à appuyer les décisions vitales des experts.

0.10 Le cadre théorique

L’angle mort traité dans notre cadre théorique (Figure 1) se résume à la recherche d’un équilibre d’enrichissement d’une interprétabilité intrinsèque; avec les subtilités causales, la flexibilité, la robustesse et la personnalisation pour une perspective de valorisation d’une IA transparente, « responsable », de confiance, éthique, conforme, sécuritaire, auditable, certifiable, frugale et souveraine. De même, cet angle se situe dans un contexte sensible, imprécis et incertain. Partant de là, notre solution se rapproche de cet équilibre en essayant d’imiter le raisonnement humain pour plus de collaboration homme-machine et de soutien à la prise de décisions. En sachant que dans ce même contexte de zone grise, le cerveau avec ses processus cognitifs et sa mémoire, équilibre mathématiquement sa précision, son raisonnement, ses causalités, ses corrélations, ses inductions de règles, sa flexibilité, sa généralisation ainsi que son interprétabilité globale et

chirurgicale. Il s'ensuit ainsi notre solution informatique d'hybridation par un modèle arborescent et une sophistication du symbolisme avec des exemples d'IA *PFL* permettant cet équilibre en modélisant des degrés d'appartenance avec le clustering. Dès lors, notre étude de cas traitant le choc septique utilise des règles « *IF-THEN* » avec une estimation de l'effet du traitement moyen (*Average Treatment Effect : ATE*) avec *Lightgbm*, le mélange gaussien bayésien (*BGM*), *GMM* et *Kmeans*. En dernier lieu, en combinant la *PFL* avec d'autres paradigmes, ce cadre se retrouve dans des voies de recherche interdisciplinaire où les sciences permettent de combiner l'informatique et le cognitif pour des interventions qui ne sont pas strictement binaires.

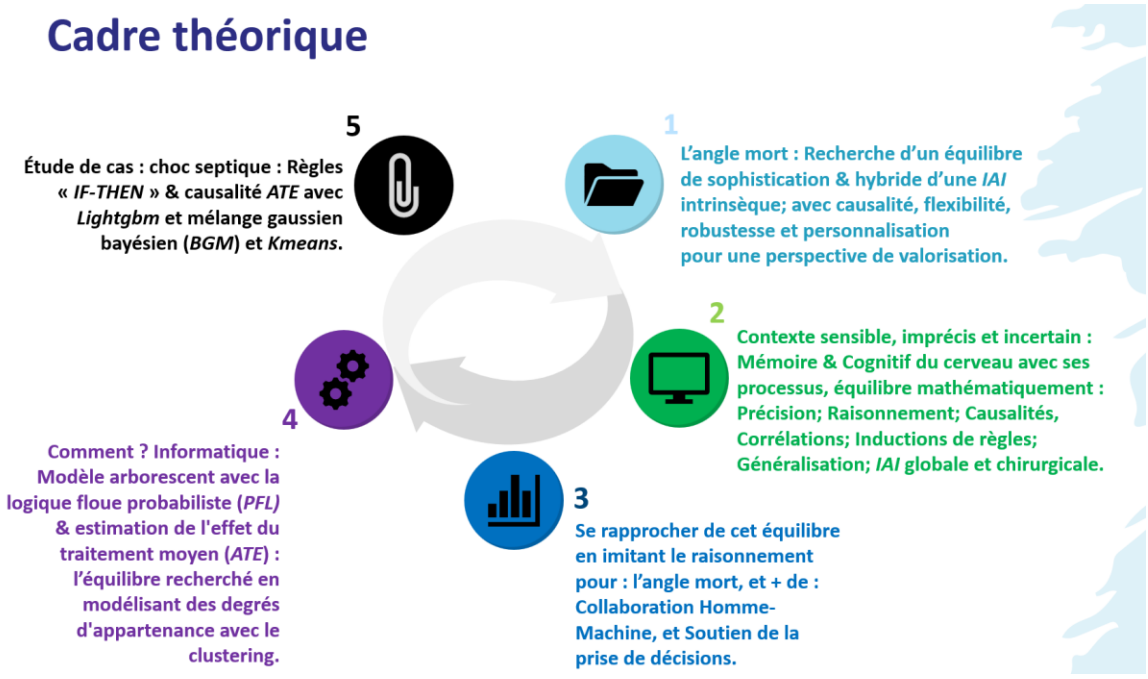

Figure 1 : Cadre théorique.

0.11 L'état de l'art

Une revue de la littérature a été effectuée afin de rassembler les concepts pertinents. De ce fait, l'objectif de la recherche documentaire était de repérer les publications majeures, les études académiques et les travaux de référence permettant la compréhension des différents éléments de notre recherche. Ainsi, la méthode de recherche consistait à consulter diverses bases de données (BD) universitaires, y compris *Google Scholar, arXiv, Institute of Electrical and Electronics Engineers Xplore*, ainsi que le site de l'*Association for Computing Machinery*. De même, des termes

liés principalement à l'interprétabilité, la causalité et la logique floue probabiliste ont servi à déterminer les articles appropriés sans restriction de date de publication pour garantir une perspective globale exhaustive.

De surcroît, les critères de choix pour la sélection des travaux littéraires comprenaient : des articles axés spécifiquement sur les méthodes et les techniques; des revues exhaustives; des recherches spécifiques dans différents contextes; actes de conférence avec références faisant autorité; et des publications abordant les perspectives à venir. Aussi, l'analyse a été conduite en recourant à une synthèse qualitative afin de déceler les thèmes majeurs dans le but de délivrer un compte rendu de la situation actuelle.

Aussi, notre recherche exploratoire du domaine est basée sur les articles de Boehm (1988) et de Basili et coll. (1986, 1994, 2013) traitant de développement de prototypes en génie logiciel. Ainsi, elle a été structurée en quatre phases : définition, planification, expérimentation et discussion des résultats obtenus. Dès lors, l'étape de définition a consisté à bien énoncer, préciser, investir et documenter notre recherche. De plus, elle a permis de bien comprendre ce qui est inclus dans notre travail. Ensuite, elle a facilité le travail sur l'analyse des besoins, l'état de lieux, la contribution, les validations techniques et les étapes à accomplir pour la finalisation de la thèse. Par la suite, la deuxième étape de planification a montré l'évolution de l'établissement des modèles *PFL* sur les données du choc septique. Ainsi, cette étape a permis la mise au point final des limites de notre état des lieux pour bien consolider les hypothèses, la méthode, les techniques, les défis et les contretemps. De la sorte, le tableau 1 suivant de l'étape de planification montre l'évolution de la recherche vers l'établissement des prototypes *PFL* sur les données de cette maladie.

Tableau 1 : L'étape de planification.

Étapes	Déclencheurs	Livrables
1-Définition.	Analyse.	-Clarification de la direction de recherche.
2-Recherche de travaux académiques	Analyse de la littérature avec ce qui suit :	-Comprendre les défis de l'état de l'art.

et techniques pour l'état de l'art de l'objectif de recherche.	"interprétabilité", "explicabilité", "flexibilité", "robustesse", "probabilisme", flou", "probabiliste flou", "prédiction", "choc septique", "septicémie", "causalité", "raisonnement causal", "clustering". Et, cela, aussi, avec les équivalents en anglais.	
3-Itérations algorithmiques.	Validations techniques.	-Conception finale des algorithmes. -Rédaction de la thèse.

En fin de compte, l'étape de l'expérimentation des modèles de l'étude de cas sera faite à travers une exploration des parties suivantes : analyse, compréhension et préparation. Ainsi que modélisation, évaluation et développement. Nous utilisons Python pour finaliser les algorithmes de nos modèles *PFL-Lightgbm*. Durant l'étape d'explication des résultats, nous allons analyser et interpréter les résultats des flux de travail pour le choc septique.

0.12 La structure du travail

Cette étude est organisée en plusieurs sections liées entre elles, chaque partie apportant une contribution au discernement détaillé de notre démarche spécifique et de ses conséquences. L'introduction a établi les fondements en précisant le contexte, l'enjeu, le cadre théorique, ainsi que les Oc, Oi, Qc, Qi, Hc, Hi, Cc et Ci de notre recherche. La revue de littérature explore les recherches actuelles sur les techniques d'IA interprétables (*IAI*) et explicables (*XAI*), la causalité, la logique floue, les systèmes probabilistes ainsi que leur mise en application dans le cadre du choc septique. Le but étant de déceler l'absence de la valorisation de l'IA avec des modèles de sophistication et d'hybridation que notre étude aspire à combler. Les chapitres consacrés à la méthodologie, l'implémentation et les expérimentations exposeront les méthodes de collecte, de traitement des données, les considérations éthiques ainsi que la conception de ces modèles.

Également, la synthèse et l'analyse exposent les informations recueillies et les manipulations réalisées, fusionnant les conclusions initiales pour faire ressortir les premières observations importantes. Le chapitre consacré aux résultats présente de manière claire les performances du

modèle, les principales conclusions concernant la calibration et l'interprétabilité des règles ainsi que leur classement. La discussion s'appuie sur ces résultats pour les analyser à travers le prisme des théories en vigueur, envisager leurs conséquences pratiques et réitérer la pertinence de notre thèse. Pour finir, la conclusion récapitule l'ensemble des contributions scientifiques, aborde l'ampleur de l'étude, suggère des pistes de recherche futures et des conseils tout en mettant l'accent sur les conséquences pour une IA personnalisée et fiable dans un environnement sensible. Cette disposition est élaborée pour assurer une progression cohérente de notre raisonnement et explication.

0.13 Conclusion

Pour donner le ton à notre sujet d'envergure, nous avons présenté le contexte et les raisons fondamentales de notre étude. Nous avons souligné l'enjeu majeur qu'est l'interprétabilité et la causalité de l'IA dans des contextes sensibles (notamment pour le choc septique) et exposé nos buts, interrogations et hypothèses de recherche. Il a été évoqué les apports escomptés sur les plans cognitif et technologique, mettant en avant notre détermination à promouvoir une IA plus valorisante (transparente, éthique et fiable). L'esquisse du travail a établi les fondations pour une exploration méthodique. Pour maintenir cette dynamique et mettre notre démarche en contexte, le prochain segment se plongera dans l'examen de la littérature. Nous explorerons les bases théoriques et les recherches existantes qui nous ont conduits à identifier le vide de recherche que cette étude ambitionne de combler.

CHAPITRE 1 :
REVUE DE LA LITTÉRATURE : L'EXPLICABILITÉ ET L'INTERPRÉTABILITÉ

1.1 Introduction

L'examen de la littérature constitue une phase indispensable dans toute démarche de recherche universitaire. Elle offre la possibilité de contextualiser le travail au sein du panorama scientifique actuel, de repérer les théories pertinentes, les méthodes actuelles et les progrès notables. Ce chapitre se penchera sur les processus cognitifs, le symbolisme, le connexionnisme, les définitions, la classification des techniques et approches de l'interprétabilité et de l'explicabilité, la comparaison entre les boîtes : blanches, grises et noires, les bénéfices de l'interprétabilité ainsi que son importance. L'idée est non seulement de compiler les connaissances en cours, mais aussi d'établir un état des lieux de l'interprétabilité avant d'identifier les contraintes et le vide légitimant notre démarche, notamment l'intégration du symbolisme pour optimiser l'interprétabilité et la faculté à se rapprocher de la causalité. Cette identification pourra affiner notre contribution et garantir son originalité et sa pertinence.

1.2 Les processus cognitifs

La cognition désigne l'ensemble des processus dynamiques par lesquels le cerveau humain acquiert, traite, transforme et utilise l'information issue de l'environnement ou de l'organisme lui-même (ex. souvenirs). Tout autant, sa prédiction est surprenante. Il aime bien prédire où il se retrouvera. À un bon endroit et au bon moment. Ainsi, il construit un fort modèle prédictif. Ce modèle du cerveau ne modélise pas toujours ses observables d'une façon binaire. Sa « machine prédictive » élabore continuellement un modèle interne en synthétisant les données présentes dans son environnement. Ensuite, ce modèle lui permet de générer des inférences afin de minimiser ses erreurs (surprises) de prédictions associées aux phénomènes nouveaux qu'il expérimente. Ultérieurement, pour ne pas avoir des « bogues » dans cette machine, il utilisera, ces surprises pour de futurs succès. Subséquemment, la mise à jour des prédictions lui permet de

déjouer l'apprentissage des bruits non désirés. En conséquence, la réalisation d'actions lui acquiesce d'acquérir de nouvelles informations sur son contexte.

De plus, le cerveau est fréquemment lié à des attributs : la gestion souple de données vagues, imprécises ou indéterminées avec un niveau d'appartenance associé à un raisonnement basé sur des éléments non binaires ; aussi, il administre l'ignorance et l'incertitude relatives à la réalisation d'un observable. De même, l'équilibre robuste du cerveau offre une capacité prédictive efficace sur des situations qu'il n'a jamais rencontrées ainsi que sur celles qu'il connaît déjà. Donc, cela le rend plus proche d'une prédiction de plus en plus précise.

Le cerveau a également une grande capacité d'apprentissage; il mesure l'incertitude de sa connaissance et l'état imprécis des événements dans son environnement, d'autant plus, si les données de cet environnement sont inexactes ou qu'elles manquent de consistance. Il a une vigueur ainsi qu'une tolérance à l'ambiguïté et à l'indétermination. Aussi, il prend en compte les imperfections de l'observable. D'autre part, il interprète ses décisions. De plus, il a des portraits globaux et détaillés de son environnement. Il s'adapte aux changements dans son contexte spécifique. Subséquemment, il définit les observables, forts, moyens et faibles. De même, il induit des règles. Il génère fréquemment des anticipations. Il valide souvent ses connaissances implicites. Il minimise ainsi le désordre. Il a, en outre, une prédiction plus ou moins précise. Il utilise pareillement ses croyances pour générer des anticipations ou attentes pour prédire les données en entrée. Il met à jour autant ses prédictions précises pour aboutir à de bonnes décisions. Puis, il essaye de projeter son modèle prédictif sur son futur. Immanquablement, il utilise la mémoire du passé pour prédire son futur. Ensuite, il orchestre une recherche d'équilibre de ces processus cognitifs en prenant en considération un maximum de variables en parallèle. Néanmoins, pour le cerveau c'est difficile de prédire avec plusieurs variables interférentes en simultanée.

De surcroît, nous supposons que « la mémoire du passé n'est pas faite pour se souvenir du passé, elle est faite pour prévenir le futur. La mémoire est un instrument de prédiction », c'est-ce qu'écrivait le neurobiologiste Alain Berthoz. Lorsque le cerveau ne pratique plus, il égare sa mémoire. De plus, selon Lestienne, « la mémoire conditionne la cognition » (Pévet et coll., 2011). Aussi, il se limite sans dépenser beaucoup d'énergie et de temps. De surcroît, nous estimons que le cerveau construit un modèle prédictif avec des logiques et des corrections de calculs en moins

de temps possible. Ainsi, ces calculs pourraient être de la mathématique complexe. Ensuite, son modèle dépend de ces calculs mathématiques et physiques pour modéliser ses observables. Pareillement, il corrige ses erreurs par rapport à ce modèle.

Aussi, le cognitif du cerveau implique une hiérarchie, une mise à jour, une minimisation des surprises et une précision de la prédiction des événements. De plus, le cerveau « encode des croyances (états probabilistes) pour générer des prédictions à propos des entrées sensorielles, puis utilise les erreurs de prédictions pour mettre à jour ses croyances. » (Bottemanne et coll., 2022) et où la valeur de la conclusion peut changer à la suite de nouvelles données. Ainsi, si le cerveau calcule la vraisemblance d'un nouvel événement en fonction de croyances préalables, puis change ses croyances précédentes en fonction de nouveaux événements, nous estimons que pour ses traitements cognitifs, cela ressemble à des processus bayésiens.

Autant, le fonctionnement du cerveau est extraordinaire. Si les données de son environnement sont inexactes ou qu'elles manquent de consistance, il a une robustesse ainsi qu'une tolérance à l'ambiguïté et à l'hésitation. Aussi, il interprète ses décisions et il fait des mises à jour précises pour aboutir à de bons jugements. De plus, il définit plus que des subtilités fortes, moyennes et faibles. Il a des portraits chirurgicaux, globaux et nuancés de la causalité. Ensuite, il orchestre une recherche d'équilibre de ses processus cognitifs en prenant en considération un maximum de variables en simultané.

Également, pour les limitations de la mémoire, Miller (1956) donne un maximum de sept éléments à emmagasiner, à court terme. Aussi, ajouté à cette limitation, il y a celle, en terme comportemental donnant une rationalité limitée (Simon, 1990) au cerveau. De plus, « selon le *satisficing* ou le *good enough* de Simon, on recherche spontanément ce qui nous est accessible, à savoir des solutions plus satisfaisantes que moins satisfaisantes » (Serge Robert, 2024). Subséquemment, la mémoire serait le puits des traitements interprétables et causaux.

Pareillement, nous supposons que le bayésien[6] explique une partie du processus cognitif causal. Le bayésianisme est originellement une conception philosophique et mathématique issue des travaux de Bayes et théorisée par le mathématicien Frank Ramsey au XXe siècle. De plus, le cerveau fonctionnerait selon un mode bayésien (Jaynes, 2003). De même, « l'approche bayésienne est un outil mathématique sophistiqué qui peut aider le raisonneur spontané, qui a tendance à prendre des décisions basées sur une estimation qualitative et approximative de la probabilité. Dans cette perspective, l'approche bayésienne donne une estimation quantitative et précise de la probabilité. Cela justifie la pertinence de l'approche bayésienne pour aider le raisonnement humain. » (Serge Robert, 2024). Par ailleurs, ce bayésianisme implique une hiérarchie, une mise à jour, une minimisation des surprises et une précision de la prédiction des événements. Également, si le cerveau pouvait calculer la vraisemblance d'un nouvel événement en fonction de croyances préalables, puis de changer ses croyances précédentes en fonction de nouveaux événements, nous estimons qu'il est en partie bayésien. De surcroît, « la théorie du cerveau bayésien, une approche computationnelle issue des principes du traitement prédictif (PP, *Predictive Processing*[7]) propose une formulation mécanistique et mathématique des processus cognitifs. Cette théorie suppose que le cerveau encode des croyances (états probabilistes) pour générer des prédictions à propos des entrées sensorielles puis utilise les erreurs de prédictions pour mettre à jour ses croyances. » (Bottemanne et coll., 2022). De plus, « l'inférence bayésienne peut calculer le degré de confiance à accorder à une cause hypothétique. La connaissance *a*

[6] La formule de Bayes calcule la probabilité inconnue d'un événement (A) sachant une preuve (B) où la plausibilité ou le postériori P(A|B) est égale à la vraisemblance multipliée par la probabilité avant l'événement : P(B|A) P(A) / l'évidence P(B). Ce calcul se fait à partir de probabilités conditionnelles connues, généralement dans un sens causal.

[7] Selon le « codage prédictif » (« predictive coding », ou « predictive processing »), le cerveau fait des prédictions (en « top down ») : les états sensoriels à venir. Et, les compare à ses erreurs sur ces prédictions (en « bottom up ») : états sensoriels actuels en transmettant son « feedback » et il corrigera, ainsi, ses erreurs en minimisant la différence entre les prédictions et les erreurs (Ramstead et coll., 2018). Selon ce codage, le cerveau pourra avoir deux possibilités : soit de faire une inférence perceptuelle en mettant à jour son modèle pour le faire correspondre à la réalité de son environnement. Sinon, faire une inférence active en changeant et agissant sur son environnement pour qu'il corresponde davantage au modèle du cerveau s'il estime que c'est le bon (Friston, 2009). Dans la théorie du modèle du traitement prédictif du bayésien, les prédictions descendantes « top down » du cerveau génèrent des prédictions et engendrent inlassablement des modèles sur son environnement. Elles se basent sur les croyances postérieures du cerveau. Les erreurs de prédictions montantes « bottom up », en contrecoup, lui disent si la correspondance avec son univers est plutôt bonne ou mauvaise. Elles informent le cerveau des erreurs qu'il va faire en fonction des signes de ce qui se passe dans son univers et dont ses sens vont intercepter. Ainsi, il y a des erreurs de prédiction ascendantes et descendantes qui sont l'aboutissement des connaissances effectuées en nous (Clark, 2017). Les prédictions descendantes avec une plus grande précision formeront un pouvoir prédictif plus imposant. La première énonciation du paradigme prédictif se retrouve dans les travaux de Hermann von Helmholtz qui a notamment apporté d'importantes contributions à la thermodynamique (Bottemanne et coll., 2022).

priori ou antérieure contribue au calcul de probabilité »[8]. D'autre part, Hermann von Helmholtz « suggère que notre cerveau génère des inférences inconscientes prédisant l'état de son environnement immédiat et évoluant en fonction de la réalité perçue par les organes sensoriels. Pour Helmholtz, ces prédictions sont équivalentes à une simulation mathématique du monde générée par le cerveau témoignant de l'écart manifeste entre la perception subjective et la réalité objective. Sans faire d'hypothèses précises sur l'implémentation cérébrale des processus prédictifs, sa conception mécanistique constitue l'un des premiers fondements de la théorie du traitement prédictif. ».

1.3 Le symbolisme et le connexionnisme

Le symbolisme encode la connaissance et le raisonnement de manière explicite, via des symboles et des règles. Les critères précis et mesurables du symbolisme peuvent être le nombre de relations alimentant un modèle. Aussi, le taux de satisfaction des contraintes, ou le nombre d'étapes d'inférence logique (chaînage avant ou arrière) qu'un système peut exécuter sans erreur avant d'atteindre une conclusion. Selon Pallanca et Read (2021), le symbolisme, « c'est cette catégorie qui est plébiscitée en médecine, car pour que l'intelligence artificielle (IA) soit acceptée dans certains domaines à haut risque, son comportement doit être vérifiable et explicable. Ceci est souvent très difficile à réaliser par des algorithmes connexionnistes. De ce fait, les systèmes symbolistes impliquent souvent le raisonnement déductif, l'inférence logique, ou l'approche bayésienne. ». De plus, selon eux, le symbolisme se focalise sur la gestion des symboles en les associant entre eux. Ensuite, selon ces auteurs, cela consiste à coder un modèle au problème en traitant les entrées par son système selon ce modèle pour donner une solution. Les systèmes experts, de règles ou d'arbres de décision sont de cette catégorie symbolique. En parallèle, l'enrichissement du symbolisme avec la combinaison des deux théories mathématiques du flou et des probabilités permet de se rapprocher d'un équilibre d'interprétabilité, de robustesse, de flexibilité, en tentant de représenter le raisonnement avec différenciation entre les causalités et les corrélations pour saisir et réduire ainsi les risques dans un contexte d'incertitude et d'imprécision. De même que la *PFL* permet de traiter les données pour l'aide à la décision. Joint à

[8] L'inférence bayésienne : l'intelligence artificielle par la statistique Par Tomas Rojas Vazquez – auxiliaire de recherche au Laboratoire de cyberjustice. https://www.cyberjustice.ca/2020/12/16/les-techniques-algorithmiques-de-lia-linference-bayesienne/

cela, ce rapprochement est un ensemble de processus se retrouvant dans la mire des chercheurs dans les différentes écoles en IA. D'autre part, le flou se concentre sur l'incertitude que les humains utilisent dans le langage pour communiquer et raisonner et la probabilité se concentre sur l'incertitude naissant de l'ignorance quant à la survenance d'un événement (Madrid, 2023). De surcroît, les deux théories utilisent des degrés pour représenter l'incertitude, mais avec deux significations différentes.

Autant en IA, avec l'école des méthodes probabilistes (telles que les réseaux bayésiens), il y a d'autres méthodes comme le clustering (en regroupant des objets similaires sans les avoir prédéfinis), l'évolutionnisme (ex. programmation génétique) et le connexionnisme. Ce dernier utilise l'apprentissage du *ML* et les réseaux de neurones (RN). De plus, les algorithmes de ces réseaux prennent comme modèle le cerveau. Le *DL* est une utilisation de ces réseaux avec de nombreuses couches cachées. Également, dans les dernières années, les réseaux *Transformers* permettent une rapidité d'entraînement sur de très grands jeux de données. Aussi, « les algorithmes d'apprentissage automatique procèdent par itérations. Les premières itérations améliorent le modèle, mais souvent, passé un certain nombre d'itérations, le modèle commence à surajuster[9] sur les données d'entraînement. Il perd alors en généralisation. » (Coulombe, 2020).

De plus, la généralisation ou la robustesse est la performance du modèle sur des données qui n'ont pas été utilisées lors de l'entraînement. L'erreur de généralisation (ou de test) est l'erreur du modèle sur ces données. De même, cette erreur est mesurée avec les données de test. La réduction peut être faite par la minimisation de l'erreur d'entraînement. Par ailleurs, cette robustesse est le principal obstacle au déploiement du *ML* dans la pratique médicale et une taille de données suffisamment grande est déterminante pour que l'entraînement du *ML* obtienne de bonnes performances (Chen et coll., 2022). D'autre part, la flexibilité est « la capacité d'un système à s'ajuster en temps opportun lorsqu'une condition nouvelle ou à multiples facettes se produit de manière à maximiser la résolution de problèmes sans perturber les performances globales » (Zarrin, 2022). En IA, elle fait référence à la capacité de s'adapter.

[9] Si le modèle apprend par cœur, on a de l'apprentissage du bruit : surapprentissage (surajustement).

En définitive, le symbolisme et le connexionnisme représentent deux paradigmes clés dans l'étude de l'interprétabilité et la causalité. Ensuite, l'explicabilité à travers la causalité peut enrichir l'interprétabilité. En effet, comprendre consiste à relier des causes et des effets dans un contexte, ce que le symbolisme interprète par des règles et le connexionnisme découvre par des associations.

1.4 Les définitions de l'explicabilité et de l'interprétabilité

Le besoin d'interprétabilité est de plus en plus reconnu comme un thème central à aborder en recherche en IA, surtout lorsqu'elle opère dans un contexte sensible (ex. la santé) (Nguyen et coll., 2022). De même, « la nécessité de modèles entièrement compréhensibles par l'homme est de plus en plus reconnue comme un thème central dans la recherche en IA » Ambag et coll. (2023). De plus, selon eux, les méthodes d'*XAI* s'appuient sur une analyse a posteriori (*post hoc*) du processus de décision offrant ainsi des aperçus (*insights*) sur la façon dont le modèle a pris ses décisions; par conséquent, l'*XAI* ne passe en revue le modèle qu'une fois celui-ci créé. Ensuite, ils citent Başağaoğlu et coll. (2022), qui selon lui, en pratique, un modèle *XAI* nécessite d'abord l'élaboration d'un modèle en tant que boîte noire suivie de l'examen de ses processus internes pour saisir la valeur des différentes caractéristiques et les décisions qui en résultent; en revanche, l'*IAI* est considérée comme un terme plus large que l'*XAI*. Aussi, pour l'*IAI*, le processus décisionnel de ces systèmes est directement observable, ils citent aussi Fuchs et coll. (2020), qui selon lui, l'*IAI* génère des modèles qui, dès leur création (*a priori*), sont conçus pour être compréhensibles par les humains, c'est-à-dire interprétables dès leur conception; et selon Doran et coll. (2017), un système interprétable offre à l'utilisateur la possibilité non seulement d'observer, mais également de saisir comment les données sont mathématiquement converties en résultats.

Autant, l'*XAI* est une « mesure par laquelle un être humain peut comprendre la cause d'une décision » (Miller, 2019). De même, l'*XAI* permet de comprendre pourquoi un modèle produit une décision ou une classe de prédiction particulière sans ouvrir la boîte noire ou saisir sa mathématique. Ainsi, elle résume les raisons derrière le comportement. De plus, c'est une connaissance de ce qu'évoque une fonctionnalité et de son importance pour les performances de ce modèle. Aussi, après la construction de la boîte noire, l'*XAI* permettra d'expliquer ses rouages

en comprenant l'importance des différentes caractéristiques et décisions auxquelles elles peuvent conduire (Başağaoğlu et coll., 2022).

En revanche, on regroupe souvent sous le terme interprétabilité : le contrôle de l'utilisateur, la transparence, le « *right to explain* », ou l'explicabilité, qui sont parfois synonymes ou distincts (Falip, 2019). De même, l'*IAI* est la capacité à déterminer la cause et l'effet à partir d'un modèle en se référant à des notions purement mathématiques et elle traduit le fonctionnement mathématique du modèle. L'*IAI* permet de voir et de comprendre comment les entrées sont mathématiquement mappées aux sorties (Doran et coll., 2017). De plus, elle permet de saisir ce que le modèle a prédit. L'*IAI* permet de comprendre le pourquoi des sorties et de faire confiance aux résultats en comprenant comment de petits changements dans l'entrée entraînent ou pas d'importants changements dans la sortie. Ensuite, de savoir si cette sortie de prédictions ou de décisions est équitable (ex. sans faire de ségrégation contre un groupe bien précis). Par ailleurs, l'*IAI* du processus employé depuis l'entrée jusqu'aux sorties en classes prédites permettra de déchiffrer les propositions de la machine à travers une interaction et une collaboration interactive avec l'expert du contexte dont la boîte noire ne pourra offrir. La machine sera ainsi bien complémentaire à l'expert. Également, l'*IAI* pourra être analysée selon un niveau basé sur la complexité et un autre basé sur la sémantique (Gacto et coll., 2011).

À titre d'exemple, pour un contexte sensible, les résultats indiquent que l'*IAI* qui tente de représenter le raisonnement humain pourrait être d'une grande aide pour les cliniciens dans la prise de décisions; c'est une démarche prudente puis judicieuse pour élaborer des classes de produits médicaux; et à l'opposé de la notion plus stricte d'*XAI* qu'ils jugent insuffisante pour obtenir la confiance et l'adhésion dans le contexte clinique (Ambag et coll., 2023).

1.5 La taxonomie des méthodes d'explicabilité et d'interprétabilité

Selon Bourguin et coll. (2021), les méthodes explicatives s'appuient principalement sur les coefficients d'importance des caractéristiques d'entrée et se basent sur l'idée d'attribuer à chacune une valeur indiquant son importance dans le processus de prédiction. Selon eux, il est donc envisageable d'obtenir des éclaircissements sur une prédiction spécifique, ou des interprétations plus générales illustrées par divers graphiques liant caractéristiques, indice de

pertinence et prédictions. De même, les explications globales essayent de rendre le modèle transparent en expliquant son comportement complet alors qu'une explication locale ne vise que des prédictions individuelles et non pas le modèle entier. À côté, les méthodes ad hoc comprennent des approches intrinsèquement interprétables. Aussi, du fait de la simplicité de la structure des modèles intrinsèques, elles peuvent interpréter leurs prédictions par eux-mêmes. De plus, elles n'ont pas besoin de techniques supplémentaires pour l'intellection. Alternativement, les techniques post hoc sont celles pouvant être appliquées après la conception du modèle et l'entraînement. Joint à cela, ces post hoc construisent un second modèle explicable pour s'accoler au premier modèle sous-jacent et expliquer ses prédictions. Également, elles pourraient ne pas imiter le sous-jacent avec précision, ce qui engendre des erreurs d'interprétations. Pareillement, si le premier modèle est biaisé ou donne de mauvaises relations parmi les données d'entraînement, ces erreurs seront encore plus grandes. Identiquement, l'*XAI* haute résolution fournit une valeur d'attribution par voxel[10], tandis que celle de basse résolution fournit une seule valeur d'attribution pour plusieurs voxels (de Vries et coll., 2023). Ensuite, les méthodes d'explicabilité du modèle spécifique sont applicables à certaines classes spécifiques du modèle ou bien elles peuvent être pour toutes les classes du modèle et ainsi être indépendantes (agnostiques) du modèle (Arrieta et coll., 2020). De même, le besoin d'accès aux données et au modèle, les temps de calcul lents et des explications potentiellement mal interprétées représentent des inconvénients de l'agnostique comme *LIME* (*Local Interpretable Model-agnostic Explanations*) ou *SHAP (SHapley Additive exPlanation)*. Subséquemment, cette dernière fait partie de ce que nous avons recensé comme méthodes *IAI/XAI*. La Figure 2 présente la taxonomie des méthodes *XAI* & *IAI.*

[10] Le mot voxel est la contraction de « volume » et de « pixel » qui est une contraction de « *picture* » et « *element* ».

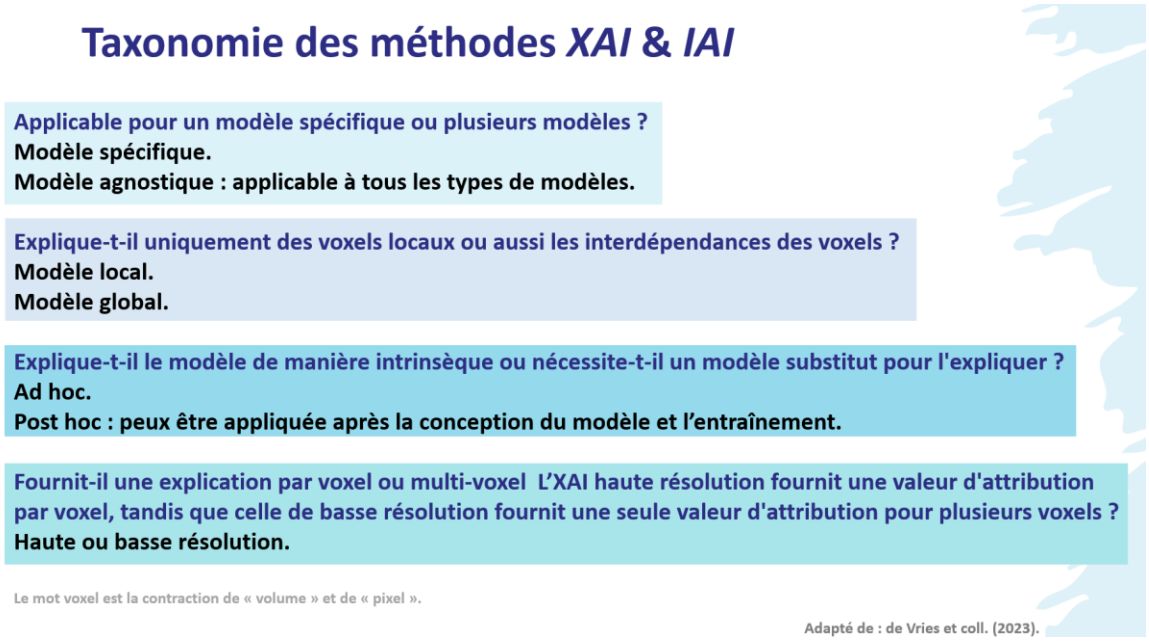

Figure 2 : Taxonomie des méthodes *XAI* & *IAI* (de Vries et coll., 2023).

1.6 Les méthodes d'explicabilité et d'interprétabilité

L'objectif de ce chapitre est d'explorer les méthodes d'explicabilité et d'interprétabilité de la Figure 3.

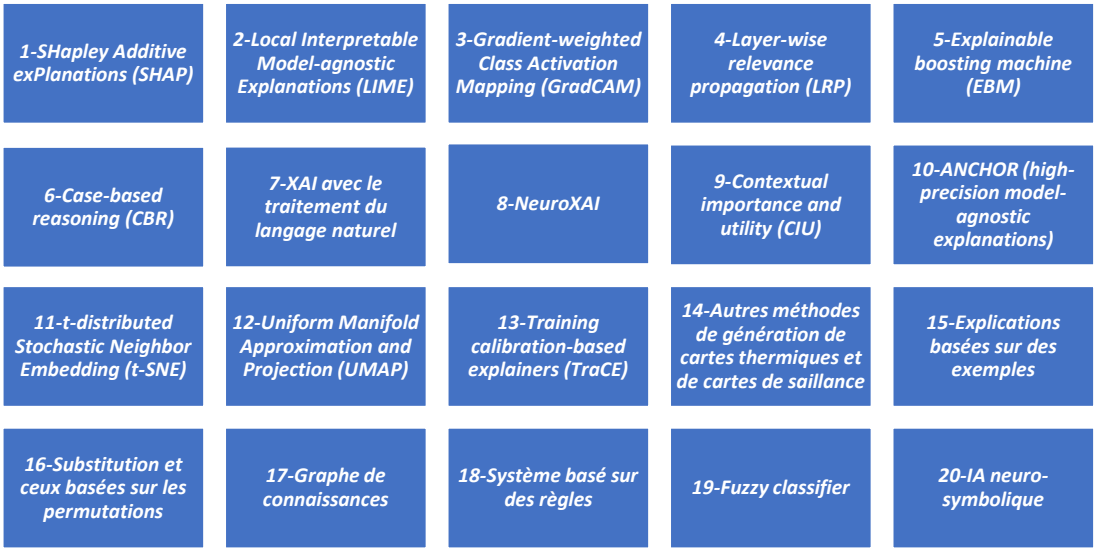

Figure 3 : Méthodes d'explicabilité et d'interprétabilité.

Préalablement, la méthode des valeurs *SHAP* du mathématicien Shapley (1953) qui fait partie des procédés d'explications basées sur les caractéristiques (*Feature-based Explanations*) se base sur la théorie des jeux collaboratifs qui permet de « déterminer dans quelle mesure chaque joueur dans un jeu collaboratif a contribué à son succès. De même, c'est une méthode d'attribution des paiements aux joueurs en fonction de leur contribution au paiement total. Aussi, les joueurs coopèrent au sein d'une coalition et tirent un certain profit de leurs contributions individuellement » (Meraihi, 2019). Le profit est représenté également par la prédiction et les joueurs par les variables d'entrée. De la sorte, un tracé *SHAP* (Figure 4) montre le résultat du calcul de l'action des variables importantes pour prédiction du choc septique. Par exemple : Prédiction (0,7) = bicarbonate (0,2) – calcium (0,2) + créatinine (0,7).

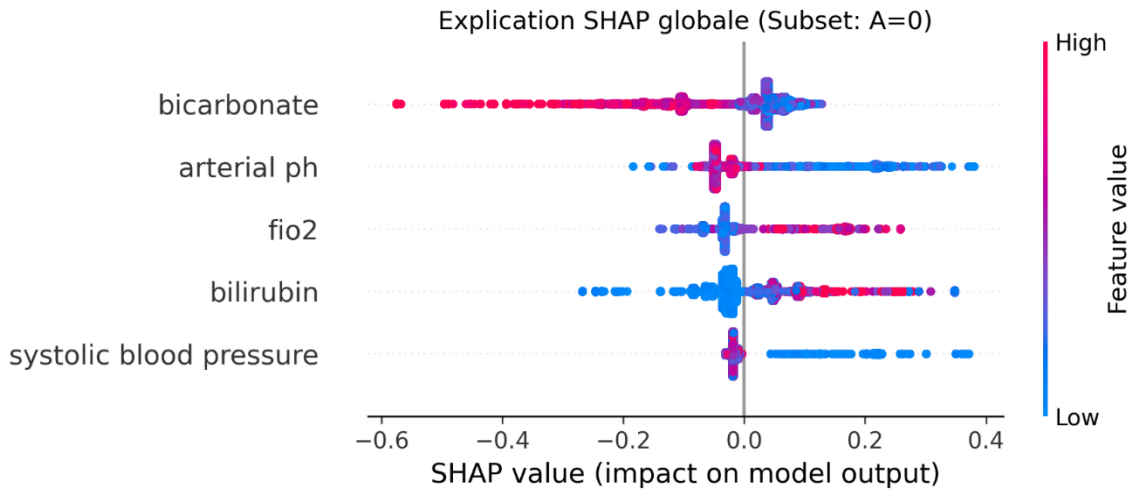

Figure 4 : Tracé *SHAP* pour les cinq variables importantes pour la prédiction du choc septique.

Deuxièmement, l'algorithme *LIME* (Ribeiro et coll., 2018) faisant partie aussi des procédés d'explications basées sur les caractéristiques engendre un modèle autour d'une prédiction pour l'approximer localement. Plus exactement, selon ces auteurs, la méthode *LIME* conçoit de nouvelles données à savoir des données proches de la prédiction à expliquer, puis les apprend à l'aide d'un modèle explicable (régression linéaire ou arbre). En revanche, l'inconvénient de *LIME* est qu'il ne fournit pas une généralisation de l'interprétabilité issue du modèle local à un niveau plus global (John-Mathews, 2019). La Figure 5 présente le résultat d'un modèle explicable avec *LIME* pour le choc septique.

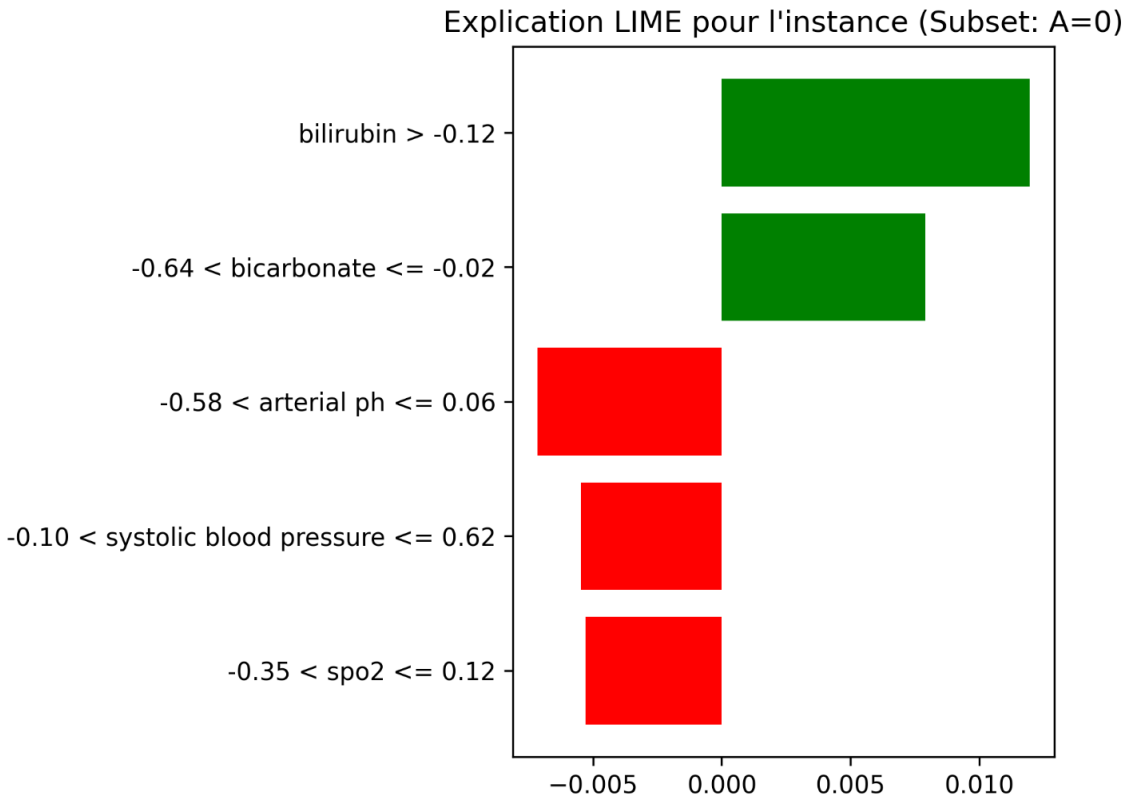

Figure 5 : Résultat du modèle *LIME* pour le choc septique.

Aussi, selon Chaudhury et coll. (2023), la carte d'activation de classe (*Class Activation Mapping*) est une stratégie de localisation approximative pour expliquer les classifications de certains *Convolutional Neural Network* (*CNN*) ou *DL*. Ainsi, en troisième, selon ces auteurs, la cartographie d'activation de classe pondérée par gradient *Grad CAM* (*Gradient-weighted CAM*) est une méthode de localisation utilisant la discrimination de classe et produit des explications visuelles en mettant en évidence les principales données responsables de la prédiction. De même, *Grad-CAM* est une visualisation qui « utilise le gradient d'un concept cible pour produire des heatmaps identifiant les régions de l'image qui ont le plus participé à sa reconnaissance. » (Bourguin et coll., 2021). Aussi, les dérivés de la *Grad-CAM* fournissent une explication des associations apprises par un *CNN*.

Quatrièmement, sous forme de cartes thermiques, la méthode de propagation de pertinence par couche *Layer-wise Relevance Propagation (LRP)* est un exemple de méthodes basées sur la rétropropagation (*Backpropagation*). De plus, elle permet de chiffrer les participations des

vecteurs d'entrées (images ou vidéos) des RN. Aussi, en appliquant l'algorithme *LRP*, on peut analyser les paramètres de poids dans le modèle et tenter de déterminer l'influence de chaque fonctionnalité d'entrée par rapport à la prédiction (Yang et coll., 2018). En plus, l'approche *LRP* offre une explicabilité et des échelles pour des réseaux neuronaux profonds qui peuvent être extrêmement complexes (Montavon et coll. 2019). Également, les *LRP* et les méthodes basées sur la rétropropagation et traitées par les auteurs Nielsen et coll. (2022) comme ceux intégrant le gradient ou ceux utilisant l'activation d'un neurone présentent des avantages cités par ces auteurs, comme la granularité locale fine et le calcul efficace. Ensuite, ils citent aussi des inconvénients comme la sensibilité au « bruit » du gradient et une interprétabilité globalement limitée.

Cinquièmement, avec une bonne performance, la méthode de l'algorithme *Explainable Boosting*[11] *Machine* (*EBM*) permet d'expliquer la contribution des variables individuelles à la prédiction à la fois d'un point de vue global et local. De même, elle permet la détection automatique des interactions (Nori et coll., 2019).

Sixièmement, la méthode du *Case-based reasoning* (*CBR*) qui est basée sur une mémoire où un système résout de nouveaux problèmes sur la base de cas mémorisés des situations antérieures similaires. De même, la *CBR* comporte les phases de caractérisation du problème, récupération, réutilisation, révision et conservation. La caractérisation est un portrait adéquat du problème de départ. De plus, la récupération découvre le cas précédent le plus similaire dans cette base. Également, la réutilisation essaye de reprendre la solution de l'un des éléments récupérés pour apposer la solution du cas au problème présent. Aussi, la révision pourra ajuster le cas de base distingué en raison d'un problème de dissemblances dans la caractérisation pour dénouer le problème en question. Finalement, la conservation dans le cas de la réussite de la solution, garde le problème du cas résolu pour une utilisation future[12].

[11] « L'approche *séquentielle* de la construction d'un ensemble cherche à créer des modèles faibles qui se concentreront sur les erreurs du modèle. On parle alors de *boosting* : par itérations successives, des apprenants faibles viennent exalter (« booster ») les performances du modèle final qui les combine » (Azencott, 2018).

[12] « Le raisonnement à base de cas (CBR) est analogue au raisonnement humain : 1. Un nouveau problème survient et on cherche dans nos expériences passées des problèmes similaires; 2. Parmi ces problèmes similaires, on retient celui qui semble le plus prometteur pour s'en servir comme modèle; 3. On fait un essai avec la solution la plus similaire trouvée et on l'applique au problème en cours. a) si cela fonctionne : nous avons notre solution ; b) sinon, il faut modifier la solution. » (Zhang, 2013).

Septièmement, l'*XAI* avec le traitement du langage naturel (*Natural Language Processing : NLP*) permet d'aider les utilisateurs avec des analyses, des quantifications et des extractions rapides des données importantes dans les textes. À l'égard de la *NLP*, Hu et coll. (2021) ont incorporé une approche explicable avec un *CNN* pour la prédiction des codes médicaux à partir des textes cliniques. De même, leur approche a utilisé un outil d'attention pour analyser et sélectionner ces textes importants pour chaque code.

Huitièmement, Zeineldin et coll. (2022) ont proposé la méthode *NeuroXAI* pour l'explicabilité des *DL*. *NeuroXAI* fournit des cartes convolutives de visualisation des clusters sur les couches internes. De même, elle est évolutive, facile à mettre en œuvre et elle est basée sur les gradients pour expliquer les *DL* en les clarifiant. De plus, elle peut fournir une haute résolution en implémentant des présentations localisées améliorées.

Neuvièmement, selon Främling (2023), la méthode *Contextual importance and utility (CIU)* comme explication post-hoc, est : indépendante du modèle, a été présentée par Kary Främling en 1992 pour expliquer les recommandations ou les résultats des systèmes d'aide à la décision, a une base mathématique différente des méthodes *SHAP* et *LIME*, et ne se limite pas qu'à l'influence des fonctionnalités. De plus, il a montré comment le *Contextual Importance (CI)* et le *Contextual Utility (CU)* peuvent fournir des explications contrefactuelles et donner un discernement plus approfondi du comportement du modèle.

Dixièmement, Ribeiro et coll. (2018) ont introduit la méthode *ANCHOR (high-precision model-agnostic explanations)* qui utilise l'apprentissage par renforcement. Selon eux, elle est indépendante du modèle et explique localement le comportement avec des règles précises et flexibles. Aussi, ils ont utilisé le terme « *anchors »* en représentant des conditions suffisantes pour les prédictions. Pour chaque instance expliquée, leur méthode analyse les voisins des données. De même, ils ont montré comment ces « *anchors »* permettaient aux utilisateurs de prédire comment un modèle se comportera sur des instances invisibles avec moins d'effort et une plus grande précision. Selon ces auteurs, les « *anchors »* génèrent des explications locales dans un format de règle simple *« IF-THEN »*, et elles sont basées sur un algorithme de recherche de graphiques.

Onzièmement, dans le travail de Hussain et coll. (2022), la méthode probabiliste non linéaire *t-distributed Stochastic Neighbor Embedding (t-SNE)* est un algorithme d'exploration, de visualisation et de réduction de dimension (deux ou trois) qui peut être appliquée sur des données de haute dimension. Joints à cela, avec la *t-SNE*, ils ont pu chercher des données dans un espace en présentant les structures locales du voisinage de chaque point. Par ailleurs, ce voisinage est représenté par une probabilité conditionnelle. Pour calculer la relation de ce voisinage entre les points, *t-SNE* convertit la distance euclidienne en probabilité de vraisemblance. De même que, seuls les voisins immédiats sont pris en compte localement. Pareillement, on parle ainsi de perplexité faible. De plus, si elle est globale, tous les points sont tous les voisinages de perplexité grande[13]. Également, la *t-SNE* est flexible en l'utilisant pour interpréter les *CNN*. Aussi, elle est très utilisée en clustering pour voir si les données ont une structure en clusters ou pour la validation des résultats (Chakraborty et coll., 2020). En parallèle, la *t-SNE* fonctionne de manière aléatoire. Donc, on n'obtiendra pas précisément la même sortie à chaque fois qu'on l'utilise.

Douzièmement, selon McInnes et coll. (2018), comme la *t-SNE*, la méthode non linéaire *Uniform Manifold Approximation and Projection (UMAP)*, plus récente, améliorée, efficace et rapide, est un algorithme de réduction de dimension qui cherche une structure plus globale avec une représentation des données en plus faible dimension. De même, l'*UMAP* définit une perplexité dépendante de la densité du voisinage. Joint à cela, la principale différence entre *t-SNE* et *UMAP* vient de la définition de la similarité. Ainsi, l'*UMAP* permet de construire une matrice représentant la similarité entre chaque paire de points et rechercher les points dans l'arrivée qui vérifie cette similarité. Également, elle minimise l'entropie croisée et elle peut déterminer la dimensionnalité des données. Ensuite, inversement au *t-SNE*, elle capture à la fois les dissimilitudes et l'écart entre les clusters. De plus, quand les points de données sont plus nombreux, de sorte que la dimension est plus élevée, la vitesse est beaucoup plus rapide que la *t-SNE* et il n'y a aucune limite au calcul. Par contre, une des limites de l'*UMAP* est sa mauvaise interprétabilité en raison de sa non-

[13] Selon Hussain et coll. (2022), il y a trois étapes. Dans une première, nous calculons les similarités des points dans l'espace initial en grande dimension. De plus, pour chaque point, une distribution gaussienne est centrée sur ce point. De même, ces points sont normalisés d'où est notée une liste de probabilités conditionnelles. Une deuxième étape consiste à construire un espace de plus petite dimension dans laquelle sont représentées les données. Également, comme à l'étape première, le calcul des similarités des points se fait dans un nouvel espace, mais en utilisant une distribution t qui n'est pas gaussienne. Ensuite, une liste de probabilités notées est obtenue. Dans la troisième étape, une mesure est utilisée pour comparer les similarités obtenues dans les deux espaces (faible et grande dimension). Enfin, dans l'espace de petite dimension, la descente de gradient est minimisée pour obtenir les meilleures possibilités.

linéarité (Band et coll., 2023). Subséquemment, selon ces auteurs, elle peut aboutir à des défis lors de la présentation des clusters dans certaines dispositions.

Treizièmement, selon Thiagarajan et coll. (2022), la méthode *Training calibration-based explainers (TraCE)* est conçue pour la classification et est composée de trois éléments principaux. Primo, un *CNN* pour les données d'entraînement. Secundo, un modèle prédictif qui prend des représentations en entrée et génère l'attribut cible. Tertio, une stratégie d'optimisation et de génération de contrefactuelles exploitant un objectif d'étalonnage basé sur cette incertitude pour découvrir les relations entre les entrées avec cet attribut. Enfin, selon eux, elle facilite un discernement global.

Quatorzièmement, Loh et coll. (2022) ont inclus des méthodes *XAI* supplémentaires qui, bien que moins connues que ceux décrits précédemment, pourraient également produire des cartes thermiques. Ils citent Liz et coll. (2021) qui ont utilisé le *package Keras Vis Python* comprenant diverses méthodes de génération de ces cartes. De même, selon eux, dans l'ensemble, ces générations de cartes thermiques fonctionnent bien avec les *DL*, en particulier les *CNN* et l'apprentissage par transfert. Aussi, « les cartes de saillance sont des méthodes spécifiques à l'imagerie ou l'analyse de textes permettant de visuellement mettre en valeur (masque de surlignage) les parties d'images ou du texte ayant significativement participé à la décision de l'algorithme black-box (souvent un réseau de neurones profonds). Joint à cela, le calcul de la carte de saillance étant basé sur l'algorithme d'apprentissage (représentation des gradients), la méthode n'est pas agnostique aux familles d'algorithmes black-box » (John-Mathews, 2019).

En quinzième, et selon Poché et coll. (2023), les méthodes d'explications basées sur des exemples (*Example-Based Explanations*) constituent une famille de techniques d'*XAI* post-hoc dont l'explication est fournie sous la forme d'exemples concrets (ou parties d'exemples) prélevés directement dans les données d'entraînement, sans génération artificielle : il peut s'agir de prototypes, de contre-exemples (*counterfactuals*), d'instances influentes ou de concepts illustrés par des exemples. De même, ils estiment que cette approche s'appuie sur le fait que l'humain raisonne naturellement par analogie et exemples, rendant l'explication plus intuitive et accessible pour l'utilisateur.

En seizième, Molnar (2025) définit les méthodes de substitution et celles basées sur les permutations : un modèle de substitution est un modèle interprétable (par exemple une régression linéaire ou un arbre de décision) entraîné sur les mêmes données en utilisant comme cible les prédictions d'un modèle boîte noire. D'autre part, il précise que l'idée est de remplacer (« *surrogate* ») partiellement la boîte noire pour fournir une explication de son comportement global. De même, il estime que l'un des avantages est un découplage modèle où le choix du « *surrogate* » est indépendant de cette boîte noire, et les inconvénients se résument à une fidélité limitée, une perte de granularité et une dépendance aux données. Ensuite, selon lui, l'importance d'une caractéristique de permutation (*Permutation Feature Importance : PFI*) évalue l'accroissement de l'erreur de prédiction du modèle à la suite de la permutation de ses valeurs, ce qui perturbe le lien entre la caractéristique et le résultat effectif. Aussi, il considère qu'une caractéristique est « importante » si sa permutation entraîne une augmentation de l'erreur du modèle parce que cela indique que le modèle s'est appuyé sur cette caractéristique pour effectuer la prédiction. De plus, il détermine qu'une caractéristique est considérée comme « sans importance » si sa permutation n'affecte pas l'erreur du modèle, car dans ce scénario, le modèle ne tient pas compte de la caractéristique pour effectuer la prédiction. Il précise également qu'une importante détérioration de la performance indique que le modèle a largement utilisé la caractéristique pour faire des prédictions. Pareillement, il énonce qu'une simple implémentation, une mesure de l'impact global sur la performance et une robustesse sont les avantages de la *PFI*. Il exprime subséquemment qu'une création d'instances irréalistes, une dépendance aux corrélations et une variabilité due au hasard sont les inconvénients de la *PFI*.

En dix-septième, et selon Tiddi et Schlobach (2022), un graphe de connaissances (*Knowledge Graphs*) comme BD structurées en triplets (entité-relation-entité) capturant des concepts d'un contexte et une ontologie formalisée décrivant les types d'entités, leurs propriétés et les relations valides entre elles sont des représentations symboliques générant des explications basées sur la sémantique des relations et des classes. Ensuite, selon eux, les avantages se résument à la richesse sémantique, l'interopérabilité ainsi que l'auditabilité et la traçabilité; avec en revanche des inconvénients se résumant à un coût de construction, une rigidité et une couverture inégale.

En dix-huitième et comme méthode d'explication globale, un système peut se baser sur un ensemble de règles. Il est flexible parce qu'il est modifiable et facile à expliquer en utilisant des

règles simples « si ceci, alors cela ». Cependant, à l'égard de l'accroissement de la complexité, ce système reposera sur plus de connaissances approfondies de l'expert et nécessitera beaucoup de son énergie en temps avec un travail manuel pour l'analyse, l'écriture et la génération de ces règles. De plus, il n'a pas de capacité d'autoapprentissage.

En dix-neuvième, les modèles intrinsèquement interprétables constituent une classe importante de méthodes interprétables (Huang et coll., 2022). De même, Sabol et coll. (2020) ont pu générer une carte thermique sur les images pour identifier le cancer colorectal en utilisant un critère d'appartenance à la classe floue cumulative (*Cumulative Fuzzy Class Membership Criterion*). Ensuite, pour diagnostiquer les patients atteints d'une maladie cardiaque, Bahani et coll. (2021) ont utilisé la méthode *Fuzzy classification Rules Learning via Clustering* pour extraire des règles linguistiques à partir des données cliniques. Cependant, les fonctionnalités de grande dimension et les règles excessives peuvent augmenter la complexité du modèle et détériorer l'interprétabilité des systèmes flous (Huang et coll., 2022).

En vingtième et selon Michel-Delétie & Sarker (2024), l'IA neuro-symbolique (*Neuro-Symbolic* (*NeSy*)) est une hybridation des méthodes symboliques et connexionnistes qui injecte des règles écrites en logique dans la structure du réseau de neurones; elle cherche ainsi à tirer parti du meilleur des deux méthodes en intégrant le symbolique; donc, la *NeSy* permet de renforcer l'interprétabilité grâce à cette injection dans le modèle avant l'entraînement sur les données et d'assigner la responsabilité des décisions prises (pour savoir « qui est responsable ? »).

1.7 Les avantages et la pertinence de l'interprétabilité

D'après Masís (2023), l'interprétabilité des modèles *ML* est un domaine en évolution constante cherchant à rendre les décisions prises par ces systèmes intelligibles pour les experts. De plus, il souligne que c'est fondamental pour les décisions à fort enjeu, comme des contextes sensibles : de la santé ou des finances, où les modèles peuvent considérablement influencer la vie des individus. De même, selon lui, trois motifs primordiaux justifient l'importance de l'interprétation de ces modèles :

• **Décisions optimales :** Quand nous saisissons le mécanisme décisionnel d'un modèle, cela nous permet de repérer et rectifier les fautes tout en maximisant son efficacité.

• **Renforcement de la confiance :** Si nous sommes capables de justifier le flux d'un modèle, il est plus probable que nous fassions confiance à ses prédictions, un aspect décisif pour sa mise en œuvre.

• **Aspects éthiques :** Saisir la manière dont un modèle prend ses décisions peut nous permettre de déceler et de réduire les biais, en nous assurons que le modèle soit équitable et responsable.

L'interprétabilité de l'IA est pertinente pour diverses raisons, y compris :

1-Confiance et transparence

La confiance se manifeste lorsque les utilisateurs considèrent l'IA comme étant fiable, équitable et en phase avec leurs intérêts. De même, la transparence signifie que les données, les algorithmes et les critères de prise de décision indispensables doivent être mis à disposition. Encore, elle offre la possibilité de surveiller les biais et d'assurer la conformité légale.

2-Conformité réglementaire

L'exigence légale est incontournable où des réglementations internationales comme le RGPD (Règlement général sur la protection des données) imposent des normes strictes relatives à la transparence et à la responsabilité des systèmes automatisés. Aussi, ces règles imposent l'interprétabilité comme une exigence et non pas comme une option. Également, la conformité réglementaire ne freine pas l'innovation, elle en est un moteur responsable. Pareillement, elle établit un cadre structuré où l'interprétabilité est une exigence technique et éthique, comme le souligne le Comité consultatif national d'éthique français[14].

3-Détection des erreurs et débogage

L'identification et la correction des erreurs sont des étapes capitales pour assurer la fiabilité et l'adéquation aux exigences des utilisateurs, en particulier dans des situations sensibles. De même, le processus de débogage nécessite que les concepteurs consignent les choix de l'IA et qu'ils rendent ces données accessibles. Joint à cela, dans de tels scénarios, le repérage des erreurs est

[14] https://www.ccne-ethique.fr/

une obligation éthique et réglementaire afin d'éviter des répercussions sévères. De plus, les erreurs non identifiées rendent les institutions vulnérables aux problèmes juridiques et mettent en question la validité de l'IA.

4-Aperçu des relations entre les données

L'interprétabilité offre la possibilité de comprendre les corrélations et causalités présentes dans les données.

5-Compréhension approfondie du contexte

L'emploi de méthodes intelligibles dans un secteur critique ne permet pas seulement de confirmer les choix effectués par ces modèles, mais aussi d'en tirer des informations inédites et d'encourager la créativité.

6-Responsabilité partagée

L'interprétabilité est un aspect déterminant pour établir une responsabilité collective dans une situation sensible. De même, elle favorise la clarté des décisions, facilitant ainsi la collaboration entre des systèmes d'IA et des experts.

7-Atténuation des risques

L'interprétabilité aide à repérer les erreurs possibles, à garantir une validation constante et à minimiser les dangers liés à des traitements inappropriés.

Hassanlou (2025) en citant ces 7 raisons, elle en ajoute la gouvernance efficace de l'IA, l'impact sur les entreprises ainsi que l'engagement des parties prenantes. Également, en novembre 2020, le conseil d'administration d'AstraZeneca a pris des mesures pour répondre aux risques en publiant un ensemble de principes éthiques (Tableau 2) pour les données et l'IA.

Tableau 2 : Principes d'AstraZeneca pour l'utilisation éthique des données et de l'*IA* (Mökander et Floridi, 2023).

Principe	Opérationnalisation
Confidentialité et sécurité	Nous respectons la vie privée et agissons de manière compatible avec l'utilisation prévue des données.
	Nous utilisons des systèmes de données et d'IA conçus pour être sécurisés.
Explicabilité et transparence	Nous sommes transparents sur l'utilisation, les forces et les limites de nos systèmes de données et d'IA.
	Nous veillons à ce que les hypothèses soient claires, les algorithmes correctement documentés, les décisions explicables, et les processus de gestion des conséquences imprévues soient en place.
Équité	Nous nous efforçons d'utiliser des ensembles de données robustes et inclusifs dans nos systèmes de données et d'IA.
	Nous traitons les individus et les communautés de manière juste et équitable dans la conception, le processus et la distribution des résultats de nos systèmes d'IA.
Responsabilité	Nous appliquons une gouvernance proportionnelle à l'impact et au risque des systèmes de données et d'IA.
	Nous anticipons et atténuons l'impact des conséquences défavorables potentielles de l'IA grâce à des tests, une gouvernance et des procédures adaptées.
Centré sur l'humain et bénéfique pour la société	Lorsque des données et de l'IA sont impliquées, des humains supervisent le système et sont responsables de générer des bénéfices clairs et attendus pour les individus et la société.
	Nous mettons en place une gouvernance dirigée par des humains sur nos systèmes d'IA. Nous respectons la dignité et l'autonomie humaines et nous efforçons de refléter cela dans nos systèmes d'IA.

1.8 La comparaison des boîtes : blanches, grises et noires

Dans le domaine de l'IA, la sélection du modèle est fréquemment confrontée à une double contrainte : obtenir des prédictions exactes tout en garantissant une certaine transparence dans

le processus de prise de décision. Subséquemment, dans le cadre de l'interprétabilité, on distingue trois grandes catégories de modèles, qu'on pourrait désigner sous les termes de boîtes blanches, noires et grises.

Également, les boîtes qualifiées de blanches se distinguent par leur grande transparence. Ils s'appuient sur des principes mathématiques clairs et faciles à décoder, ce qui donne aux experts la capacité de saisir exactement comment chaque facteur influe sur le choix final. Leur avantage majeur est l'interprétabilité, qui favorise la conformité aux régulations et la confiance des utilisateurs. Cependant, la précision n'est pas en rendez-vous, surtout lorsque le lien entre les variables est compliqué. On peut par exemple citer les arbres de décision comme l'un de ces modèles. En revanche, les boîtes dites noires, telles que les *DL*, mettent l'accent sur la précision. Ces méthodes ont la capacité de saisir des interactions complexes présentes dans les données, ce qui conduit généralement à des performances de prédiction supérieures. Toutefois, leur fonctionnement interne reste obscur, ce qui complique l'interprétation des choix effectués, d'où la problématique dans des secteurs sensibles dans lesquels la transparence est indispensable.

Aussi, les boîtes dites grises se situent entre ces deux pôles extrêmes. Donc, ils associent, d'un côté des aspects d'interprétabilité dérivés de modèles transparents et de l'autre, des méthodes visant à augmenter la précision généralement tirées d'approches plus compliquées. Par exemple, un modèle hybride pourrait combiner une structure de base interprétable (telle qu'un arbre décisionnel) avec des modifications ou des pondérations basées sur des méthodes d'optimisation sophistiquées. De même, cette solution offre un équilibre entre la performance fiable et l'intellection du processus décisionnel, répondant ainsi aux besoins pratiques de divers contextes sensibles.

1.9 Conclusion

Ce chapitre a approfondi les notions indispensables d'interprétabilité et d'explicabilité dans le domaine des systèmes d'IA, en particulier au regard de leur application dans des contextes complexes. Nous avons exploré les fondements de ces notions en les reliant aux processus cognitifs, à la distinction fondamentale entre le symbolisme ainsi que le connexionnisme et en fournissant des définitions pour ces concepts clés.

Autant, la taxonomie et la revue des méthodes d'explicabilité et d'interprétabilité ont mis en lumière la diversité des approches disponibles allant de techniques explicables aux modèles intrinsèquement interprétables. La comparaison des boîtes blanches, grises et noires a permis de nuancer les compromis entre performance et transparence, soulignant que le choix de l'architecture et de ces méthodes dépend intrinsèquement du cas d'usage, des exigences réglementaires et du niveau de confiance requis par les utilisateurs finaux. Ainsi, l'option entre les modèles se fonde fondamentalement sur le contexte d'utilisation et les restrictions spécifiques de ce contexte d'application. De ce fait, les boîtes blanches présentent une transparence élevée ; cependant ils peuvent rencontrer des difficultés en matière de complexité et de précision. Pour leur part, les boîtes noires tendent à obtenir des performances élevées, à l'égard d'une opacité croissante. Les boîtes grises apparaissent comme une option équilibrée, offrant une interprétabilité adéquate tout en renforçant la puissance prédictive. Cette concession rend ces outils particulièrement captivants pour les applications exigeant à la fois un strict respect des normes scientifiques et une conformité aux exigences de transparence.

Également, les avantages de l'interprétabilité sont multiples et indéniables. Au-delà du simple entendement du fonctionnement interne des modèles, elle renforce la valorisation par la confiance des utilisateurs, facilite la validation et la conformité réglementaire et permet une meilleure prise de décision. Sa pertinence s'accroît exponentiellement à mesure que les systèmes d'IA s'immiscent dans des domaines critiques tels que la santé, la défense ou la finance où les conséquences d'une décision non comprise ou erronée peuvent être désastreuses.

En somme, l'interprétabilité n'est plus une option, mais une nécessité impérieuse. Elle est le pont primordial entre la performance brute des algorithmes et leur acceptation sociétale. Le défi réside désormais dans le développement de méthodes toujours plus robustes, généralisables et flexibles, capables de concilier la complexité inhérente des modèles de pointe avec un niveau d'interprétabilité adapté aux besoins des différents acteurs. Les recherches futures devront continuer à explorer cette tension, en s'appuyant sur une compréhension approfondie des processus cognitifs pour concevoir des systèmes d'IA non seulement précis, mais aussi valorisants et responsables.

CHAPITRE 2 :
REVUE DE LA LITTÉRATURE : DES VALEURS IMPORTANTES ET DIFFICILES À RÉALISER EN IA

2.1 Introduction

Plusieurs défis importants sont mis en évidence par le contexte de ce chapitre. En premier lieu, les biais dans les données peuvent gravement affecter la qualité et l'équité des systèmes d'IA. Il est donc indispensable d'élaborer des techniques pour détecter, vérifier et corriger ces anomalies afin d'assurer des résultats justes. Par la suite, on discutera de valeurs fondamentales s'avérant être des exigences incontournables.

De plus, l'implémentation des IA dans des secteurs sensibles (tels que la médecine) présente une complexité accrue qui ajoute un niveau supplémentaire de défis. Aussi, dans de telles situations, les erreurs algorithmiques et les déficiences en matière de valeurs, l'absence des spécialistes du contexte et une démarche non multidisciplinaire peuvent entraîner des répercussions significatives.

Également, les systèmes d'IA transforment profondément nos sociétés, mais leur adoption avec des valeurs fondamentales se heurte à des défis complexes, allant des biais inhérents aux données jusqu'aux impératifs moraux. De même, les données biaisées constituent un point de départ critique : un algorithme, aussi sophistiqué soit-il, reproduira inévitablement les imperfections de ses données d'apprentissage, qu'il s'agisse de biais d'échantillonnage, de déséquilibres structurels ou de faux positifs. Ces écueils techniques ne sont pas neutres : ils soulèvent les questions des valeurs de recherche d'équilibre de notre projet de thèse.

Face à ces enjeux, l'interprétabilité et la valorisation par la transparence émergent comme des piliers incontournables. Par exemple, l'OCDE (Organisation de Coopération et de Développement Économiques), les chercheurs ou encore les lignes directrices des pays développés pour une « IA digne de confiance », insistent sur l'articulation entre licéité, éthique et robustesse technique.

Toutefois, une IA de morale exige une tentative de représentation du raisonnement. Elle doit intégrer une responsabilité intrinsèque, soutenue par des cadres juridiques stricts, comme le RGPD, l'*Algorithmic Accountability Act* ou ceux comme de l'Union européenne. Ces exigences se concrétisent dans des domaines sensibles où des erreurs algorithmiques (comme celles rapportées dans des hôpitaux) rappellent l'urgence d'un équilibre entre innovation et vigilance vertueuse.

Ce parcours des biais aux valeurs puis aux applications concrètes illustre une question de tension centrale : comment concilier progrès technologique et valeurs ? La réponse réside dans une symbiose entre la technique et les vertus, comme en témoignent certaines solutions IA médicales alliant performance et intelligibilité.

2.2 Les définitions des valeurs en IA

La valorisation en IA est un processus par lequel une architecture algorithmique est transformée en un actif générant une utilité concrète. Les critères précis et mesurables de la valorisation peuvent être le gain opérationnel généré par cette architecture. Aussi, le taux d'amélioration de la pertinence des résultats comparé à d'autres architectures.

Transparence

Le concept de transparence des technologies IA renvoie au niveau de compréhension qu'a l'utilisateur du mécanisme de ces technologies (Andrada et coll., 2023) ; et aussi, il est ardu de saisir comment ces IA assimilent (quand ils collectent, manipulent et conservent des données) durant les phases d'entraînement et il peut être complexe de transposer algorithmiquement des idées dérivées en expressions intelligibles pour l'être humain (Camilleri, 2024)[15]. Aussi, « on entend par « transparence » le fait que, au moment où le système d'IA est mis sur le marché, tous les moyens techniques disponibles conformément à l'état de la technique généralement reconnu

[15] https://ovsm.ch/wp-content/uploads/2024/10/OVSM_Livreblanc_IA-defis_01oct2024.pdf

sont utilisés pour garantir que les résultats du système d'IA sont interprétables par le fournisseur et l'utilisateur. »[16].

Responsabilité

Selon l'UNESCO[17] et l'OCDE[18], la responsabilité en matière d'IA implique de préciser quels acteurs (développeurs, ou autres) sont chargés de gérer les répercussions des IA. Cela implique la capacité de suivre un résultat d'IA jusqu'à son auteur responsable. De plus, selon l'UNESCO, les IA « doivent être vérifiables et traçables » afin que leurs impacts (qu'ils soient positifs ou négatifs) puissent être attribués à ceux qui les ont conçus ou exploités.

Confiance

Un système digne de confiance doit démontrer sa prévisibilité, sa capacité à interpréter ses actions, sa résistance aux erreurs et son alignement avec les valeurs humaines (Lee et See, 2004). Également, la confiance ne se limite pas à un aspect technique : c'est un engagement moral et social entre les concepteurs, les utilisateurs et la société (Cowls et coll., 2019). De même, dans les IA, elle fait référence à la volonté d'une personne ou d'une institution de se fier à un système intelligent, en présumant que ses actions sont dignes de foi et répondent aux anticipations. Ainsi, elle est multidimensionnelle : elle englobe des éléments techniques (robustesse), éthiques (honnêteté), sociaux (acceptabilité) et psychologiques (perception du risque). De surcroît, elle est précaire étant donné que les décisions manquent souvent de valorisation, ce qui intensifie la peur d'erreurs invisibles ou de biais cachés. Ces biais algorithmiques constituent l'un des principaux freins à la confiance, étant donné qu'ils mettent en péril l'équité des décisions. En outre, elle est influencée par le contexte d'utilisation : un algorithme médical devrait susciter une confiance plus grande qu'un système de traitement dans le cadre d'un contexte non sensible. Par conséquent, il est nécessaire d'ajuster les exigences de confiance en fonction du niveau de criticité du système. Sans compter qu'il faut établir la confiance dès la phase de conception en suivant le principe du «

[16] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_FR.pdf
[17] https://www.unesco.org/fr/artificial-intelligence/recommendation-ethics
[18] https://oecd.ai/en/dashboards/ai-principles/P9

trust by design ». En somme, dans les situations critiques (santé, armement, sécurité), il est impératif de fonder la confiance sur des preuves vérifiées.

Éthique

L'éthique de l'IA s'appuie sur quatre principes primordiaux tirés de la bioéthique : bienveillance, non-nuisance, autonomie et justice (Cowls et coll., 2019). De plus, elle se réfère à l'examen des principes et des standards moraux qui doivent orienter la conception, la création, l'exploitation et le déploiement des IA. Aussi, l'éthique algorithmique se concentre en particulier sur les biais, la discrimination, l'absence de transparence dans les décisions et la surveillance à grande échelle. Ainsi, on ne peut pas ajouter l'éthique ultérieurement : elle doit être incorporée dès le cycle de vie du système (*ethics by design*). Or, cela nécessite de questionner dès l'origine les objectifs du système, les parties prenantes impliquées et les valeurs en présence.

Conformité réglementaire

La conformité réglementaire en matière d'IA nécessite de se conformer à tous les standards, lois et normes pertinents lors de la création et de l'exploitation des IA. Également, cela inclut notamment la protection des données (RGPD), l'anti-discrimination, la sûreté des produits et les nouvelles réglementations spécifiques à l'IA (telles que l'*AI Act* de l'Union européenne). Successivement, la conformité aux règlements en matière d'IA est une obligation collective incombant aux concepteurs, fournisseurs, et à ceux qui la déploient et aux utilisateurs (Edwards & Veale, 2018).

Sécurité

La sécurité de l'IA englobe toutes les méthodes et principes assurant que les IA sont conçues et employés d'une manière qui protège les individus et la société. D'après IBM[19], l'objectif est de s'assurer que l'IA « bénéficie à l'humanité tout en réduisant au minimum tout dommage ou issue défavorable potentielle ». Ensuite, selon IBM, la sécurité de l'IA implique de prévoir et de réduire

[19] ibm.com/fr-fr/think/topics/ai-safety

des menaces, comme les biais algorithmiques, les divulgations de données, les attaques "*adversariales*" ou encore les pannes systémiques. De ce fait, elle se réfère à la capacité d'un système à résister aux interventions malintentionnées, aux défauts internes et aux perturbations extérieures. Pareillement, elle comporte la robustesse face aux données bruitées, aux dysfonctionnements logiciels et à l'exploitation de failles algorithmiques. Ainsi, l'UNESCO exhorte les intervenants de l'IA à prévenir ou gérer les risques associés à la sûreté (dommages non souhaités) et à la sécurité (offensives malintentionnées). Aussi, l'ISO/IEC 23894 (2023) met en place un cadre normatif pour gérer les risques associés à la sécurité des IA[20]. Il est donc nécessaire que la sécurité de l'IA s'inscrive dans une démarche de cybersécurité qui comprend un pare-feu, un contrôle d'accès et une supervision des flux de données. Irrémédiablement, dans le domaine de la santé, une IA non sécurisée pourrait fausser un diagnostic, proposer un traitement inapproprié ou être exposée à des cyberattaques ciblant les données des patients.

Auditabilité

Une IA auditée offre la possibilité de détecter les préjugés systématiques, les faiblesses et les problèmes d'un système, même post-déploiement (Raji et coll., 2020). Aussi, l'enregistrement systématique des décisions intermédiaires facilite une étude approfondie des raisons d'une sortie algorithmique anormale (Brundage et coll., 2020). De plus, des standards d'audit sont en voie d'émergence dans les systèmes critiques, tels que le standard ISO/IEC 42001[21]. Dans la continuité, l'audit d'une IA fait référence à sa capacité à être examiné, décortiquée et jugée rétroactivement, en fonction de critères techniques, éthiques ou organisationnels. Dès lors, elle est vitale pour assurer la traçabilité des décisions algorithmiques et la responsabilité en cas d'incident ou de conflit. En somme, les défis sont plus importants pour les IA, puisque leur fonctionnement interne n’est souvent pas accessible à l’interprétation directe.

[20] ibm.com/fr-fr/think/topics/ai-safety
[21] https://assets.kpmg.com/content/dam/kpmgsites/ch/pdf/ISOIEC-42001-certification.pdf.coredownload.inline.pdf

« Certifiabilité »

Une IA qualifiée de « certifiable » est conçue pour être examinée et validée selon une norme officielle (standard ou référence). Autrement dit, elle a la possibilité d'obtenir une certification qui prouve qu'elle adhère à un cadre de qualité, de sécurité ou de gestion spécifié pour l'IA. Par exemple, la norme facultative ISO/IEC 42001 met en place un « système de gestion de l'IA » et autorise la vérification qu'une organisation suit des processus organisant la gouvernance, le management des risques et l'implication des parties prenantes ainsi que la capacité à certifier un système ou une organisation d'IA renforçant la confiance des parties prenantes. Également, la certification (telle que celle fondée sur la norme ISO 42001) atteste que des critères stricts (issus de normes de qualité reconnues) ont été respectés lors de l'élaboration et de l'implémentation de l'IA[22].

Frugalité

L'IA frugale vise à optimiser l'efficience d'une solution d'IA tout en réduisant au minimum son utilisation de ressources. Ce qui implique en pratique la création d'algorithmes exploitant le moins de données, de ressources informatiques et d'énergie possible tout en répondant aux performances exigées[23]. Aussi, le but est de satisfaire les exigences professionnelles sans dilapider des ressources informatiques ni accroître indûment l'impact carbone. De plus, cette méthode est guidée par les enjeux de durabilité et l'augmentation des coûts associés aux grands modèles d'IA. Par exemple, les modèles génératifs massifs nécessitent une consommation d'énergie considérable. Subséquemment, l'IA frugale préconise l'utilisation de modèles allégés lorsqu'ils sont adéquats ainsi que l'amélioration des procédures d'entraînement et d'inférence. Elle s'aligne donc sur une perspective d'IA responsable tenant compte des répercussions environnementales de la technologie.

[22] https://www.afnor.org/actualites/une-norme-et-une-certification-qualite-pour-lia

[23] https://www.intelligentcio.com/apac/2024/07/29/what-is-frugal-ai-marrying-ai-advancements-with-environmental-responsibility/

Souveraineté

La souveraineté en IA fait référence au contrôle par un État ou une entité nationale du « cycle de données » et des technologies qui y sont liées. En d'autres termes, il faut superviser le lieu et la manière dont les données sont conservées, manipulées et diffusées.

Ainsi, il est nécessaire[24] :

1- de créer des infrastructures internes (cloud, puces, algorithmes) pour assurer son indépendance technologique.

2- de sauvegarder son écosystème d'IA et ne pas être totalement tributaire des fournisseurs étrangers pour ses algorithmes et plateformes. Les questions de souveraineté en matière d'IA sont prépondérantes. Confrontées à l'essor mondial de l'IA, les nations tentent de protéger leurs innovations et leurs données. Par exemple, l'Europe finance des initiatives nationales en matière d'IA et des plateformes « de confiance » afin de réduire sa dépendance vis-à-vis des grands acteurs technologiques étrangers.

3- d'offrir également la possibilité de se conformer aux réglementations locales (protection des données privées, normes éthiques) sans être soumis à des standards étrangers. Plusieurs moyens complémentaires sont nécessaires pour atteindre la souveraineté en matière d'IA. Sur le plan technique, cela consiste à privilégier l'open source et l'échange de données entre les nations partenaires afin de stimuler l'innovation locale tout en préservant la sécurité (en évitant des silos trop restrictifs).

4- de soutenir la recherche et développement interne (ex. formation) et d'élaborer des régulations appropriées dans le cadre d'un plan politique. Le but est de parvenir à une autonomie qui permettra de maintenir sa compétitivité sur le long terme, tout en préservant ses intérêts technologiques.

[24] https://www.informatiquenews.fr/quest-ce-que-la-souverainete-en-matiere-dia-et-pourquoi-est-elle-si-importante-pour-leurope-dimitri-casvigny-aiven-101909

2.3 Les données biaisées

À travers ses premières observations, Saporta (2022) montre que les biais peuvent compromettre la qualité des BD utilisées pour entraîner les algorithmes décisionnels. Ensuite, un algorithme prédictif en lui-même n'est pas intrinsèquement inéquitable. Aussi, il est conçu pour maximiser l'exactitude des prédictions en se basant sur les données d'apprentissage en partant du principe que les données futures suivront la même distribution. Toutefois, si les données d'apprentissage contiennent des biais, l'algorithme les reproduira inévitablement. Également, il est courant de rencontrer des biais d'échantillonnage lors de la sélection préférentielle de certains cas par rapport à d'autres. En plus des erreurs liées à l'échantillonnage, des erreurs sérieuses et difficiles à rectifier peuvent survenir en raison des biais de sélection et des données absentes. Pareillement, lorsqu'il s'agit de prédire un comportement rare, les données d'apprentissage sont souvent déséquilibrées. Subséquemment, cela ne constitue pas un problème de biais de sélection (d'échantillonnage) au sens technique, par ce qu'elles peuvent refléter les proportions réelles. Cependant, le danger réside dans l'obtention d'un taux élevé de faux positifs lors de la tentative de prédiction avec une précision appréciable de la catégorie rare. De ce fait, d'autres observations montrent que pour le bruit et le biais, un algorithme peut également être considéré comme non robuste, étant donné que la variabilité des prédictions possibles peut entraîner des risques imprévus. Définitivement, on estime que ces observations de Saporta (2022) sont importantes pour comprendre les défis des données biaisées.

2.4 L'IA intègre

Une des lignes directrices déontologiques est qu'une « IA digne de confiance » doit être licite : en respectant toutes les lois et réglementations ainsi que morales : en respectant des principes et des valeurs déontologiques et robustes : d'un point de vue technique en tenant compte de son contexte (ex. social). De plus, l'IA doit faire fonctionner en harmonie trois caractéristiques : respecter l'éthique, se conformer aux législations et réglementations en vigueur (licéité) et être fiable sur les plans technique et contextuel afin d'éviter des préjudices involontaires (robustesse). De même, ces lignes directrices éthiques pour une « IA digne de confiance » ont été présentées sous l'égide de la Commission européenne, un groupe d'experts de haut niveau en IA (*High-Level Expert Group on Artificial Intelligence*). Aussi, ce document européen propose un cadre de

plusieurs rangs, allant des principes éthiques généraux à une liste d'évaluation concrète pour guider les acteurs dans le développement et le déploiement responsables de l'IA. Également, ils ont choisi ces trois caractéristiques de licéité, éthique et robustesse pour définir ces lignes visant à promouvoir cette confiance.

De plus, ils soulignent que pour garantir le développement, le déploiement et l'utilisation fiables des systèmes d'IA, il est approprié de se conformer à des principes moraux enracinés dans les droits fondamentaux. Encore, ces principes sont perçus comme des obligations éthiques que les spécialistes de l'IA devraient s'appliquer à suivre en tout temps. L'interprétabilité est l'un de ces principes. Identiquement, pour 2019-2024, le parlement européen a adopté, le 14 juin 2023, des textes d'amendements sur la législation de l'IA où « l'utilisateur est en mesure de comprendre et d'utiliser correctement le système d'IA en connaissant généralement le fonctionnement du système d'IA et les données qu'il traite, ce qui lui permet d'expliquer les décisions prises par le système d'IA à la personne concernée … ». Ainsi, afin d'assurer la transparence, les processus doivent être clairs, les compétences et l'objectif des systèmes d'IA doivent être divulgués sans réserve et les résolutions devraient pouvoir être justifiées aux individus concernés. En outre, nous devons porter une attention spéciale aux modèles dits boîte noire.

Aussi, ils soutiennent que la robustesse technique est indispensable pour développer des systèmes d'IA fiables. De même, elle implique l'élaboration d'IA basées sur une stratégie de prévention des risques afin que ces systèmes se comportent de façon crédible conformément aux attentes, tout en minimisant au maximum les dommages non intentionnels. De plus, d'après leur perspective, un des éléments de cette robustesse technique inclut : la fiabilité, la reproductibilité et l'exactitude. Également, cela dépend de la capacité de l'IA à porter un résultat, par exemple en classifiant correctement les données dans les catégories appropriées, ou de sa capacité à effectuer des prédictions, des conseils ou des décisions justes. Donc, il ne suffit pas de tenter de représenter le raisonnement humain pour penser de manière éthique, et pour qu'une machine soit intrinsèquement morale, il faudrait qu'elle soit responsable (Peterson et Hamrouni, 2022).

2.5 L’intelligibilité rime avec la confiance et la transparence

Les premiers éléments pour garantir un déploiement et une exploitation interprétable et responsable des systèmes d'IA sont de mettre l'accent sur la transparence et la confiance. Ainsi, dans le but de bâtir cette confiance, ceux qui œuvrent dans le domaine de l'IA sont tenus de s'engager à communiquer de manière imputable les données concernant ces systèmes. En pratique, cela signifie fournir des informations explicites et contextualisées qui permettent :

- Si possible, de fournir des interprétations claires et intelligibles sur les sources de données, les variables, les processus et les logiques qui sous-tendent les prédictions, contenus, suggestions ou décisions émit par l'IA ;
- De promouvoir une intellection approfondie de ces systèmes, en soulignant leurs potentialités et leurs limites ;
- D'éclairer les utilisateurs sur la nature de leurs interactions avec ces dispositifs, y compris dans un contexte professionnel ;
- De donner aux individus potentiellement affectés la possibilité de contester les résultats fournis par ces systèmes.

Il est également prépondérant d'assurer la traçabilité tout au long du cycle de vie des systèmes d'IA. Cela implique de documenter soigneusement les ensembles de données, les processus et les décisions à chaque étape de l'utilisation des données à la décision finale. Ainsi, ce contrôle minutieux est indispensable pour permettre une analyse détaillée des résultats de ces systèmes et, si nécessaire, pour répondre aux questions et contestations concernant ces décisions. Définitivement, les éléments d'interprétabilité, de confiance et de transparence, soutenue par des systèmes de traçabilité, sont des fondements majeurs pour garantir un déploiement responsable de l'IA. Elles permettent pareillement aux utilisateurs et intervenants de comprendre et si besoin de remettre en question les décisions issues de ces technologies. Selon l'OCDE, tous ces éléments sont impératifs pour garantir un déploiement responsable de l'IA.

Aussi, dans leur exploration des diverses familles d'explications, telles que les arbres de décision, Ribeiro et coll. (2016) soutiennent que « la confiance est cruciale pour une interaction humaine efficace avec les systèmes d'apprentissage automatique et que l'explication des prédictions

individuelles est importante pour évaluer la confiance ». De plus, travailler sur l'éthique du projet IA constitue un levier pour penser et enrichir ce projet. De même, cette approche invite à dépasser la simple performance technique pour intégrer la dimension fondamentale de la confiance. Identiquement, les recherches de Adamyk et coll. (2023) sur l'influence des principes de l'IA digne de confiance des auteurs Fjeld et coll. (2020) sont des thèmes clés comme :

- Transparence et explicabilité ;
- Responsabilité ;
- Confidentialité ;
- Sécurité et sûreté ;
- Équité et non-discrimination ;
- Contrôle humain de la technologie ;
- Responsabilité professionnelle ;
- Promotion des valeurs humaines.

Pareillement, bien que les algorithmes IA aient fait leurs preuves dans de nombreuses disciplines, leur adoption pour des domaines extrêmement sensibles, comme les contextes cliniques sont encore restreints en raison de problèmes liés à la confiance et à la transparence (Ahmad et coll., 2018). Aussi, la nécessité de transparence et d'interprétabilité est de plus en plus reconnue comme un enjeu majeur à traiter par les travaux de recherche en IA, spécialement dans le secteur de la santé (Nguyen et coll., 2022).

Autant, Floridi & Cowls (2022) ont identifié un cadre composé de cinq principes fondamentaux pour une IA éthique. Quatre d'entre eux sont des principes couramment utilisés en bioéthique : la bienfaisance, la non-malfaisance, l'autonomie et la justice. Aussi, sur la base de leur analyse comparative, ils ont soutenu qu'un nouveau principe est nécessaire en complément : l'explicabilité, comprise comme intégrant à la fois le sens d'intelligibilité (en réponse à la question « comment cela fonctionne-t-il ? ») et le sens éthique de responsabilité (en réponse à la question : « qui est responsable de la manière dont cela fonctionne ? »).

Également, la recommandation révisée du Conseil de l'OCDE sur l'IA établit un cadre international pour une IA digne de confiance et centrée sur l'humain. De plus, elle propose un principe clé parmi

d’autres : transparence et interprétabilité en fournissant des informations claires sur le fonctionnement et les décisions des IA. Aussi, leur texte affirme que l’interprétabilité est un élément clé de l’IA éthique. Ainsi, leur discours comporte deux dimensions:

(i) Intelligibilité technique (« Comment cela fonctionne-t-il »?) où les acteurs de l’IA devraient fournir des informations intelligibles sur les données utilisées, les processus algorithmiques et les logiques décisionnelles. De même, cela inclut la transparence sur les limitations des systèmes et les sources des données, notamment pour les applications critiques (ex. pour la santé).

(ii) Responsabilité éthique (« Qui est responsable ? ») où la traçabilité des décisions et des données est vitale pour identifier les responsables en cas de préjudice. Subséquemment, les mécanismes de contestation permettent aux utilisateurs ou personnes affectées de contester les résultats de l'IA renforçant ainsi la justice procédurale.

Ensuite, leur élément connexe à l’interprétabilité, parmi d’autres, est la robustesse où un système robuste doit être prévisible et sécurisé, ce qui nécessite une conception transparente pour éviter les erreurs ou les manipulations. Selon eux, pour un exemple de limites et défis, l’interprétabilité peut être techniquement difficile pour les systèmes d’IA boîte noire (ex. réseau de neurones profonds) nécessitant des outils d’interprétation. En parallèle, pour parfaire et renforcer la confiance dans l’IA, l’OCDE institue l’interprétabilité en l’associant à des mécanismes de gouvernance dynamique et de coopération internationale. D’autre part, cela reflète une approche où transparence technique et responsabilité éthique sont indissociables.

D'après Casey et coll. (2019), le RGPD confère un droit à l’interprétabilité solide qui a des conséquences juridiques sur la conception, le prototypage, les tests pratiques et la mise en œuvre des systèmes automatisés de gestion des données. Aussi, ce droit ne requiert pas forcément une transparence sous la forme d'une interprétation personnalisée exhaustive pour les protections qu'il offre. Toutefois, une analyse approfondie de l'élucidation des autorités de protection des données montre que la véritable force de ce droit réside dans ses impacts synergiques lorsqu'il est associé aux approches d'audit algorithmique et de « protection des données dès la conception » stipulées dans les autres sections du RGPD.

Autant, selon eux, certaines clauses du RGPD requièrent que les gestionnaires de traitement offrent "des informations significatives sur la logique impliquée" dans le processus de décision automatisée, ainsi que sur son importance et ses impacts prévus. De même, ils proposent que le droit à l'interprétabilité doive être compris de façon fonctionnelle et flexible pour permettre aux individus concernés d'exercer leurs droits selon le RGPD. De plus, ils soulignent que, bien qu'il y ait des divergences sur l'éclaircissement des dispositions du RGPD, un consensus de plus en plus large s'établit quant à l'idée que le RGPD introduit de nouvelles obligations de transparence et de responsabilité algorithmique, et ce en dépit du fait que la nature précise des interprétations demeure sujette à débat.

Également, Cecilia Panigutti et coll. (2023) ont examiné les clauses de la Loi sur les systèmes d'IA qui se rapportent aux problématiques d'opacité de certains de ces IA. Subséquemment, ils ont précisé le rôle des méthodologies d'interprétabilité dans le cadre de la stratégie réglementaire de l'Union européenne. De même, ils ont particulièrement mis l'accent sur la transparence. Ils ont précisé que l'objectif politique de la réglementation est d'assurer que les utilisateurs des IA à risque élevé saisissent comment exploiter correctement les résultats issus du système et sont en mesure de le contrôler efficacement. D'autre part, ils concluent principalement que cet objectif peut être réalisé en fournissant des renseignements et des dossiers fonctionnels appropriés aux utilisateurs ainsi qu'en instaurant des mesures de contrôle adaptées, sans qu'il soit nécessaire d'exiger l'emploi de méthodes d'interprétabilité ou de modèles intrinsèquement transparents. Joints à cela, pour soutenir leur argument, ils ont fourni des exemples d'initiatives visant à instaurer la transparence et la supervision humaine dans un système de surveillance piloté par l'IA sans se fonder sur ces instruments d'interprétabilité ou ces modèles. Aussi, ils ont aussi abordé les contraintes techniques auxquelles se heurtent les méthodes d'interprétabilité actuelles. De plus, ils ont noté certaines contraintes, comme : l'absence de cadre normatif pour juger les interprétations, la robustesse et la fiabilité.

Selon leur étude, la Loi sur l'IA ne requiert pas spécifiquement l'emploi de l'interprétabilité ni n'interdit les systèmes d'IA à boîte noire. Néanmoins, l'interprétabilité peut servir d'instrument pour répondre aux attentes. Aussi bien que d'autres mesures centrées sur la transparence et le contrôle humain pour assurer l'usage adéquat des systèmes d'IA, en particulier ceux classés comme à haut risque, pourraient être mis en œuvre. De plus, la législation impose que ces IA à

risque élevé soient élaborées de manière à garantir une transparence suffisante permettant aux utilisateurs de comprendre correctement leurs résultats. De même, cela implique une documentation exhaustive et des directives d'utilisation fournissant des informations approfondies sur les caractéristiques, les capacités et les restrictions du système. En parallèle, l'étude propose une étude de la Loi sur l'IA, en se concentrant sur sa stratégie visant à gérer le manque de transparence des IA. D'autre part, cette méthode englobe non seulement la transparence et le contrôle humain, mais aussi la gestion des risques, la documentation technique, l'exactitude, la robustesse, la sécurité informatique ainsi que des données gouvernées et enregistrées. De surcroît, les auteurs affirment que la confiance dans les IA peut être établie en combinant ces attributs interdépendants plutôt qu'en s'appuyant sur des instruments techniques précis, tels que l'interprétabilité.

Ensuite, il est capital de concevoir de nouvelles méthodes d'apprentissage automatique qui génèrent des modèles plus intelligibles. Un autre objectif, pareillement, est de développer des méthodes d'interaction personne-machine performantes pour produire des interprétations. De plus, selon Gunning (2019), ces objectifs sont ceux de l'initiative de la *Defense Advanced Research Projects Agency (DARPA)* qui est à la pointe du développement de l'IA interprétable grâce à son programme spécialisé. Aussi, il note que l'objectif du programme intelligible de la *DARPA* est de développer des IA non seulement efficaces, mais aussi aptes à justifier leurs choix et comportements de manière claire et fiable pour les utilisateurs.

D'après Gunning & Aha (2019), ce programme *XAI* conçoit et teste un large éventail de nouvelles méthodes *ML* :

• Des *DL* modifiés qui apprennent des caractéristiques interprétables.

• Ceux qui acquièrent des schémas plus organisés, compréhensibles et basés sur la causalité.

• Ceux d'induction de modèles qui déduisent un modèle intelligible à partir de tous ceux considérés comme une boîte noire.

Ils notent qu'un an après le démarrage de ce programme des procédés d'interprétabilité, les premières démonstrations technologiques et les résultats suggèrent que ces trois principales méthodes nécessitent une exploration plus détaillée et proposeront à terme aux futurs développeurs des alternatives de conception prenant en compte l'équilibre entre performance et

intelligibilité. De même, ils estiment que pour les équipes de développement, ces procédés sont examinés afin d'évaluer la valeur des interprétations qu'ils offrent, en identifiant les apports spécifiques de chacune de trois méthodes dans ce compromis.

Une décision déterminante influence de manière légale ou matérielle des secteurs requis, exigeant une plus grande transparence. Aussi, afin d'assurer cette clarté, les organisations doivent déterminer dans quelle mesure elles informent les utilisateurs sur l'emploi d'un système automatisé et leur proposent des moyens pour : premièrement, saisir la décision en accédant aux éléments qui l'influencent, y compris une élucidation des composants qui, s'ils étaient altérés, modifieraient l'issue. Ensuite, deuxièmement contester ou rectifier en documentant les réclamations, les rectifications et les appels ainsi que les actions entreprises pour traiter les préoccupations. En termes d'interprétabilité, il est élémentaire que les utilisateurs aient la possibilité de contester une décision et de se désengager du système automatisé. Aussi, l'objectif central est de garantir la transparence (et la possibilité d'interpréter les décisions) ainsi que la responsabilité dans l'application des algorithmes en imposant des exigences rigoureuses d'intelligibilité. Subséquemment, la confiance ne sera que renforcée où les décisions automatisées doivent être :

• Compréhensibles : où les éléments déterminants d'une décision doivent être à la portée de tous.
• Contestables : où les utilisateurs devraient avoir la possibilité de solliciter une réévaluation par un humain.
• Justifiées : où les entités sont tenues de consigner et rectifier les biais ou anomalies détectées.

Cette définition de la décision déterminante ainsi que ces liens entre interprétabilité, responsabilité, confiance se retrouve dans le texte de Loi connu sous le nom « *Algorithmic Accountability Act of 2022* ». De même, il a pour objectif de réguler l'usage des systèmes décisionnels automatiques et les procédures critiques soutenues par l'IA aux États-Unis. Pareillement, il charge la *Federal Trade Commission (FTC)* d'exiger des évaluations d'impact pour ces systèmes, surtout lorsqu'ils influencent des décisions pressantes. Aussi, cette Loi envisage l'établissement d'un bureau technologique à la *FTC* et donne aux États le pouvoir de faire respecter ses dispositions. De plus, cette législation représente un progrès vers une régulation éthique de l'IA, en harmonisant innovation et sauvegarde des droits individuels.

Ambag et coll. (2023) soulignent que l'interaction entre l'humain et l'IA, conforme aux lignes directrices du RGPD sur la prise de décision automatisée et aux exigences de la Loi européenne, garantit une régulation éthique, responsable et transparente des modèles dans les processus décisionnels. Également, ils rappellent que la Commission européenne, depuis 2016, encadre les systèmes opaques, notamment en renforçant les normes pour les outils d'aide à la décision clinique afin d'aligner innovation et protection des droits en imposant des exigences renforcées à ces outils dans les contextes médicaux critiques.

2.6 Le cadre réglementaire dans le contexte médical

Un cas récent rapporté par un article[25] citant le *Wall Street Journal* met en lumière les lacunes en matière de transparence et de responsabilité où une infirmière du *Kaiser Permanente Medical Group* dans un centre d'appel en Californie a suivi les recommandations d'un algorithme de triage, retardant le diagnostic d'une pneumonie et le patient est décédé quelques jours après l'appel. De même, l'infirmière a été tenue pour responsable. Le jugement a souligné que les professionnels doivent conserver leur jugement clinique, malgré les directives algorithmiques. *Kaiser* a été condamné à verser environ trois millions de dollars à la famille.

Dans un autre cas, cité dans cet article, un autre professionnel de la santé œuvrant au *UC Davis Medical Center* en Californie a été prévenu par le programme informatique d'un potentiel état septique chez un patient. Aussi, doté de quinze ans d'expertise, il était conscient qu'il s'agissait d'une erreur et que le patient était en réalité atteint de leucémie. De plus, dans ce contexte, les procédures de l'hôpital le contraignent à respecter les protocoles lorsqu'il y a un signalement du choc septique. Ainsi, il a décidé de suivre les instructions de l'algorithme en procédant à une analyse sanguine du patient, en dépit des dangers d'infection et des coûts médicaux additionnels. En fin de compte, la décision de l'algorithme s'est révélée erronée, intensifiant le dilemme éthique avec lequel il était confronté.

[25] https://www.lespecialiste.be/fr/actualites/e-health/etats-unis-des-soignants-obliges-de-suivre-les-consignes-de-l-algorithme.html

Autant, l'article note que ce sont de bons exemples où l'IA devrait servir d'outil d'aide à la prise de décision. Aussi, selon l'auteur, le manque de flexibilité dans son usage pose des problèmes éthiques et souligne l'importance de prendre conscience des contraintes d'un algorithme dans un cadre clinique. De plus, il mentionne qu'il ne faut pas percevoir l'IA comme une réponse indépendante, mais plutôt comme un instrument qui complète les compétences et le savoir-faire des professionnels de la santé. Également, il indique que les responsables d'hôpitaux devraient saisir les contraintes de l'IA en médecine et garantir son application en partenariat étroit avec les professionnels médicaux. De même, il souligne que ceci assurera une gestion améliorée des patients et préviendra les désordres dus à des erreurs algorithmiques.

Pour le défi de manque de transparence dans l'application de l'IA, un autre article[26] cite le médecin José Rodríguez Sendín, membre du groupe de bioéthique de la Société espagnole de médecine générale et familiale, où il invoque que « souvent, les objectifs et les méthodes des développeurs ne sont pas clairs, ce qui crée de la méfiance. La complexité de ces systèmes et le manque de compréhension des utilisateurs finaux peuvent conduire à un manque de responsabilité et de reddition de comptes. ». Aussi, « sans une méthodologie claire et éprouvée, les promesses de l'IA risquent de ne pas se matérialiser, ou pire, de causer des dommages importants », ajoute-t-il.

2.7 La symbiose entre l'éthique et le progrès technologique

Premièrement, il est urgent de garantir la transparence des processus pour comprendre et accorder la confiance aux décisions prises par les systèmes d'IA. Pareillement, il est impératif que les fondements des décisions prises par des systèmes spécifiques soient toujours limpides. En outre, ces systèmes doivent pouvoir justifier leur logique d'une façon claire pour les êtres humains. En second lieu, il est nécessaire d'établir une responsabilité claire. Ainsi, il est primordial de préciser la manière dont les responsabilités sont réparties entre les divers intervenants engagés dans la création, l'élaboration et l'exploitation de ces systèmes. De ce fait, il est basique que ces systèmes d'IA soient construits et utilisés de sorte à fournir une interprétation claire pour chaque décision prise. En troisième lieu, il est vital d'adopter une approche pluridisciplinaire. Aussi,

[26] Comment l'IA pourrait contribuer à désengorger les centres de santé – et pourquoi elle n'y parvient pas encore : https://elpais.com/sociedad/2024-06-15/como-la-ia-podria-ayudar-a-desatascar-los-centros-de-salud-y-por-que-todavia-no-lo-hace.html

l'incorporation de l'éthique dans les programmes de recherche sur l'IA nécessite une collaboration pluridisciplinaire faisant appel aux sciences humaines et sociales, à l'ingénierie et à d'autres domaines. Il est donc essentiel d'incorporer les enjeux éthiques lors de la conception, du développement et de l'implémentation des systèmes. L'objectif de la conception éthique est d'incorporer des aspects morales lors de la création, du développement et de l'implémentation de ces systèmes.

Pareillement, ces trois points et cette conception éthique aident à prévenir des répercussions défavorables. De plus, ces éléments sont le résultat du travail fondateur de l'*IEEE (Institute of Electrical and Electronics Engineers)*. Aussi, ce travail, nommé *Ethically Aligned Design (EAD)*, a été publié par la *Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS)* de l'*IEEE*. De plus, en y examine ces aspects, y compris l'incorporation des normes éthiques dans les systèmes.

Également, il est indispensable d'établir un cadre juridique pour assurer l'imputabilité. Pareillement, l'un des premiers éléments de la problématique juridique entourant ces systèmes est l'obligation d'établir la responsabilité et de déterminer qui est en tort lorsqu'ils provoquent des préjudices. De plus, il est nécessaire d'ajouter des commentaires concernant la façon dont le droit devrait s'adapter aux défis éthiques et juridiques particuliers posés par l'élaboration et l'implémentation de ces systèmes. Autant, il est fondamental de mettre en avant l'importance de la responsabilité et de la transparence afin d'encourager la confiance dans l'adoption du système. Il est également primaire d'anticiper les dangers associés à la mise en place du système. Ensuite, il est primordial d'établir des normes d'adoption qui réduisent ces risques. Il est aussi nécessaire de préciser la distribution des responsabilités. Subséquemment, il est primordial d'établir des responsabilités bien définies et de s'assurer que les personnes impliquées aient un discernement clair de leurs rôles, obligations et responsabilités possibles. De surcroît, l'élément décisif reste de garantir la transparence avec une disponibilité d'information. De même, la divulgation des informations sur la manière dont le système prend certaines décisions renforce la confiance dans leur mise en œuvre. Définitivement, un élément concluant est de déterminer quelles informations doivent être divulguées et à qui elles doivent être communiquées, et ce, dans le respect des enjeux liés à la sauvegarde de la propriété intellectuelle.

En définitive, un équilibre est recherché qui renforce l'innovation tout en sauvegardant les valeurs éthiques. De plus, cela nécessite de définir la légalité du système, d'instaurer des critères d'adoption sûrs et d'assurer une responsabilité et une transparence accrues. De même, ces éléments et l'équilibre souhaité ont été mentionnés par la structure relative aux *A/IS*. Ensuite, l'*EAD* a examiné le cadre légal des *A/IS* en éclairant la relation complexe entre le droit, l'éthique et les avancées technologiques. *In fine*, il ne se limite pas à étudier la manière dont le droit répond aux avancées technologiques, mais aussi comment il les régule et les façonne.

2.8 Des exemples d'innovations technologiques validées

En 2024, la *Food and Drug Administration (FDA)* des États-Unis a donné son aval au *Sepsis ImmunoScore* de *Prenosis*, un instrument d'IA examinant 22 biomarqueurs pour anticiper le risque du sepsis dans les prochaines 24 heures. Aussi, ce *Sepsis ImmunoScore*, en fonction de sa capacité, attribue un score de risque et catégorise les patients en quatre niveaux de danger distincts et assiste les professionnels de santé dans la prise rapide de décisions thérapeutiques informées, un aspect substantiel pour les patients présentant un haut risque. De plus, la nécessité de disposer d'instruments de diagnostic sophistiqués tels que celui-ci est due à l'ampleur dévastatrice du sepsis, affectant plus de 1,7 million d'adultes aux États-Unis chaque année et causant environ 350 000 décès, d'après les *Centers for Disease Control and Prevention (CDCP)*. Aussi, un diagnostic et une prise en charge rapides sont décisifs, étant donné que la mortalité s'accroît de manière significative lorsque la détection du sepsis est retardée. Ensuite, le diagnostic du sepsis a toujours été un défi en raison des symptômes initiaux non spécifiques, tels que la fièvre et l'augmentation de la fréquence cardiaque pouvant être similaires à ceux d'autres affections. Subséquemment, l'*ImmunoScore* répond à ces enjeux en offrant une estimation mesurable du risque de sepsis chez un patient, ce qui facilite des actions plus précises et plus immédiates [27].

Ensuite, dans cette partie, nous présentons XTRACTIS®. Cette technologie symbolique d'IA cognitive floue augmentée a été proposée dès 2003. Elle a à peu près 1700 robots en compétition qui résolvent des problèmes de modélisation complexes. Également, avec son système multiagent, ces robots chargent le même problème en partant de la même BD avec en leur disposition une

[27] https://dig.watch/updates/fda-approves-new-ai-tool-for-sepsis-detection

famille de stratégies de raisonnement inductif et où au départ, chacun est pour soi. Ainsi, pour résoudre ce problème, un robot pourrait trouver un modèle à deux règles et d'autres vont donner d'autres modèles avec d'autres nombres de règles et selon des temps différents. De plus, pour chaque nouveau cas, il y a un déclenchement instantané de ces règles impliquées pour un modèle validé. Il y a aussi une validation croisée. Les prédictions se font en temps réel pour plus de 70 000 décisions par seconde, sur *CPU* standard, hors ligne ou en ligne. Subséquemment, c'est une IA de raisonnement où des « *Exobrain* » augmentent les 3 modes de raisonnement humain. Tout d'abord, il y a l'induction impliquant l'extraction automatique de modèles basés sur la connaissance à partir des données, à l'instar des scientifiques appliquant la méthode expérimentale scientifique. En second lieu, la déduction engageant une prédiction instantanée des résultats pour un nouveau cas. En troisième lieu, l'abduction supposant la découverte des solutions les plus optimales répondant à une demande multiobjectif. Du coup, les rapports de prédiction interprètent comment les décisions sont prises à partir des règles préétablies par le modèle. Puis, chaque décision prédite est accompagnée d'une interprétabilité et d'une traçabilité. Par conséquent, cette IA raisonnante générale délivre des décisions de confiance, par sa conception même. Pareillement, pour assurer cette confiance, elle met au point des modèles robustes et intelligibles dotés d'une bonne capacité prédictive. De la sorte, cette IA de confiance produit de bons résultats dans des applications critiques. Elle maintient cette confiance dans l'utilisation de ces systèmes et permet d'assurer leur utilisation éthique. Il s'agit donc d'une approche collective, compétitive, réflexive, coopérative, évolutive et adaptative pour les trois types de raisonnements. Par la suite, une aide est fournie par des spécialistes certifiés par le régulateur avant la mise en service pour les utilisateurs finaux. Avec cela, les décisions sont déterminées en temps réel. C'est une IA générale polyvalente adaptée à des domaines commerciaux ou scientifiques. Particulièrement en ce qui concerne les décisions entraînant des conséquences sur la vie humaine, l'environnement ou provoquant des pertes économiques. En même temps, elle esquive les pièges inacceptables de manque de transparence des autres modèles IA, tout en conservant le plus haut degré possible de performance prédictive. Conjointement, cette technologie offre une intelligibilité des systèmes de décision dans un contexte critique avec des risques élevés. En harmonie, elle autorise une exploration automatique de connaissances à partir de données structurées ou non, de qualité inférieure ou en quantité limitée ; une vérification de ces systèmes et leur déploiement pratique dans ce contexte pour tout domaine d'activité ; et une reconnaissance légale et judiciaire à l'aube de l'AI Act. Parallèlement,

elle a été jaugée par la *DARPA*, la *National Security Innovation Network (NSIN)*, l'*Office of the Director of National Intelligence (ODNI)* et l'*US ARLAB*. En conséquence, la compagnie française Intellitech de cette technologie souveraine (du fait de son indépendance à l'open source) se déploie rapidement et stratégiquement dans les contextes précités. Cette présentation d'*XTRACTIS* a été faite par le professeur Zyed ZALILA, inventeur et fondateur de cette technologie [28].

2.9 Conclusion

Les enjeux de l'IA, allant des données biaisées aux problèmes de valeurs, mettent en lumière une réalité impérative : la technologie n'est pas suffisante. Les recherches des experts sur les biais, les valeurs et les cadres réglementaires, tels que le RGPD ouvrent une voie rigoureuse, mais indispensable. Pour être digne de confiance, l'IA doit non seulement se montrer robuste sur le plan technique, mais également faire preuve de transparence dans son fonctionnement, de responsabilité dans ses conséquences et d'équité dans son utilisation.

Autant, les cas sensibles (ex. médicaux) qu'ils soient marqués par des réussites ou des échecs algorithmiques illustrent ce problème multidimensionnel. De plus, ils démontrent que l'innovation responsable s'appuie sur un échange constant et pluridisciplinaire : ingénieurs, éthiciens, experts (ex. médecins)... Également, ils sont tenus de travailler ensemble pour s'assurer que les systèmes d'IA sont au service de l'homme et non en train de le remplacer. Pareillement, les programmes d'interprétabilité de la *DARPA* ou les recommandations de l'*IEEE* soulignent que l'intelligibilité n'est pas un obstacle, mais plutôt un outil pour une IA qui est à la fois efficace et de valeur ajoutée.

De plus, l'évolution de l'IA dépend de sa capacité à représenter des valeurs morales. Comme le mettent en évidence les directives internationales et les avancées récentes en prédiction pour les contextes sensibles (ex. médecine), cette recherche d'équilibre est bénéfique. Cependant, il requiert une surveillance constante, des ajustements adaptatifs et surtout l'admission qu'il y a des vies humaines, des droits et une imputation collective à protéger derrière chaque algorithme.

[28] https://www.youtube.com/watch?v=sKBhtPRtP4Q & https://www.youtube.com/watch?v=UFi_8cccMwo

Définitivement, pour les systèmes décisionnels critiques et à haut risque, l'IA symbolique est un outil d'aide à la prise de décision. Aussi, elle donne de bons résultats avec des données de qualité inférieure ou en quantité limitée en honorant ce qui suit comme valeurs : intelligible, interprétable, robuste, généralisable, raisonnante, transparente, responsable, de confiance, éthique, réglementaire, sécuritaire, auditable, certifiable, frugale, et souveraine.

CHAPITRE 3 :

REVUE DE LA LITTÉRATURE : LOGIQUE FLOUE PROBABILISTE

3.1 Introduction

Dans plusieurs secteurs critiques, la prise en charge de l'incertitude représente un enjeu central. Historiquement, la théorie des probabilités a été le cadre de référence privilégié pour représenter cette incertitude, en mesurant la probabilité d'événements aléatoires. Toutefois, nous reconnaissons de plus en plus que l'incertitude ne découle pas uniquement du hasard et peut aussi être le fruit d'une imprécision. C'est dans ce cadre que la logique floue, mise en avant par Zadeh en 1965, s'est révélée être un modèle puissant pour comprendre le flou, à savoir l'aptitude à décrire et à raisonner avec des notions qui n'ont pas de limites nettement établies.

Tandis que la théorie des probabilités se concentre sur la fréquence ou la probabilité d'événements binaires (vrai/faux), la logique floue autorise l'expression de niveaux d'appartenance à des ensembles, ce qui reflète l'incertitude associée à la subjectivité ou au caractère graduel. L'association et la fusion de ces deux méthodes qui sont à la fois distinctes et complémentaires, ont conduit à l'élaboration de cadres sophistiqués, dont la *PFL* occupe une position dominante. Cette branche de recherche cherche à combiner les atouts de la modélisation probabiliste et de la logique floue afin de saisir à la fois l'incertitude probabiliste et l'incertitude floue, fournissant un instrument plus souple et résistant pour l'étude et le processus décisionnel dans des contextes sensibles.

Dans cette revue de littérature, nous examinerons les principes fondamentaux de ces trois notions, leurs relations mutuelles et les progrès récents en matière de logique floue probabiliste. Nous mettrons également l'accent sur leur importance grandissante pour surmonter les problématiques actuelles liées à la gestion de la donnée incertaine et vague.

3.2 La logique floue

Nous nous intéressons à la logique floue[29] par ce qu'elle permet l'interprétabilité et l'analyse des nuances de causalités. Cela en se caractérisant par sa capacité d'explication scientifique, son intelligibilité, et sa tentative de représentation du raisonnement. Également, « elle permet de mesurer l'état imprécis des événements dans le monde (comme lorsque la personne est plus ou moins malade) » (Serge Robert, 2024). Ainsi, comme pour l'imprécision, elle a aussi, une tolérance à l'incertitude. De même, nous considérons que cette logique constitue un moyen important de flexibilité, une représentation des connaissances incomplètes à partir des données ambiguës associées à des contextes complexes, un bon temps de réponse, et une bonne maintenabilité. Pareillement, elle est basée sur la connaissance qui utilise l'ambiguïté pour attribuer le degré de compatibilité d'une opinion d'expert (Meghdadi et Akbarzadeh-T, 2001). Joint à cela, selon ces auteurs, l'incertitude peut être représentée par cette logique en décrivant une vérité partielle et un raisonnement approximatif.

Autant, la logique floue tente de reproduire les fonctionnalités du cerveau. Aussi, pour comprendre cette logique, nous prenons l'exemple de l'affirmation suivante : « mon niveau de risque cardiaque est au-dessus de la normale ». Ainsi, il existe plusieurs mesures entre 0 et 1 qui rendent cette affirmation vraie : Moyennes [0.27], Élevées [0.40] et Très élevées [0.85] où le risque cardiaque est une variable linguistique et les valeurs linguistiques sont Moyennes, Élevées, Très élevées. De plus, pour notre exemple de l'introduction, dans le flou, la bactérie A est « Élevées » avec une valeur d'appartenance de 0,7. De même, la fonction d'appartenance (*Membership function*), caractérisant le flou d'un ensemble A dans X associe chaque point dans X avec un nombre réel dans l'intervalle [0, 1]. Joint à cela, c'est un degré de vérité des valeurs précises d'une variable d'entrée qui peut être de forme triangulaire, trapézoïdale, gaussienne ou sigmoïde. Également, le choix de cette fonction dépend habituellement du problème et est fréquemment décidé subjectivement. Pareillement, le flou définit la probabilité en utilisant cette fonction. L'inférence floue est une construction de règles basées sur les variables linguistiques, puis une agrégation des règles pour obtenir une seule sortie. À la lumière de cela, la théorie de la logique floue n'est surtout pas floue.

[29] La logique floue résulte des travaux de Jan Lukasiewicz (1930), Max Black (1937) et surtout Lotfi A. Zadeh (1965) qui avait publié la théorie des *Fuzzy Sets*.

En 1974, Mamdani a construit un contrôleur flou basé sur cette théorie et l'a appliqué au contrôle des chaudières et des machines à vapeur. Aussi, selon lui, les quatre étapes de l'inférence floue sont la fuzzification où il y a définition des fonctions d'appartenance de toutes les valeurs d'entrée. Ensuite, un passage aux grandeurs physiques avec ces variables linguistiques et ainsi une transformation des entrées en grandeurs floues. Puis, il y a une évaluation des règles floues. Par la suite, une agrégation des conclusions des règles. Et, enfin, la défuzzification par une inférence de transformation des grandeurs floues à des sorties binaires non floue. Selon AYAD (2023), dans le modèle flou Mamdani, l'antécédent de la règle prémisse et de la conclusion sont des propositions floues utilisant des variables linguistiques. Après, selon lui, en se basant sur la structure singulière de la conclusion de la règle conséquente, on peut différencier un deuxième type de modèles flous : Takagi-Sugeno, dans lesquels le conséquent utilise des variables numériques plutôt que des variables linguistiques, sous la forme d'une constante, une fonction, un polynôme ou une équation différentielle et dépends des variables associées à la proposition antécédente.

Assurément, pour évaluer l'interprétabilité, nous pouvons utiliser au moins le nombre de fonctions d'appartenance et le nombre de règles en essayant d'analyser la complexité et la sémantique (Gacto et coll., 2011, Ouifak et Idri, 2023). Aussi, l'interprétabilité des systèmes flous est une caractéristique représentant les connaissances d'une manière reflétant la cognition humaine et fournissant une compréhension détaillée des systèmes complexes et de leur dynamique sous-jacente (Chen et Chen, 2001). Le modèle Mamdani est plus interprétable, mais moins précis et c'est le contraire pour Takagi-Sugeno (Tung et Quek, 2009).

3.3 Le probabilisme

Nous nous intéressons au probabilisme, parce que c'est une solution tolérante à l'imprécision, adaptable et possède une capacité d'apprentissage. Ce probabilisme se caractérise par une facilité de développement. Mais avant, nous définissons le déterminisme qui joint une cause à un résultat certain qui s'oppose au stochastique et dépend de l'aléatoire. Cette stochastique est subséquemment impossible à prédire avec exactitude. Une même cause ne va pas toujours provoquer le même effet. De plus, selon Meghdadi et Akbarzadeh-T (2001), l'incertitude statistique pourrait être considérée comme une incertitude concernant la survenance d'un

événement dans le futur. Selon eux, la probabilité représente mieux l'incertitude statistique en donnant la probabilité d'un événement qui peut ou non se produire. Pareillement, « la probabilité mesure l'incertitude de notre connaissance (comme lorsque nous croyons qu'il est plus ou moins probable que la personne soit un peu malade) » (Serge Robert, 2024). En parallèle, les probabilités permettent de quantifier l'incertitude contextuelle. De même, le probabilisme avait une combinaison de la logique avec la théorie des probabilités de telle manière que l'implication logique probabiliste se réduise à la logique ordinaire où des applications de l'IA nécessitaient le raisonnement avec des connaissances incertaines en recherchant des généralisations logiques (Nilsson, 1986). Ainsi, cette logique a combiné l'algèbre booléenne avec les probabilités pour les modèles relationnels.

Aussi, les algorithmes probabilistes se composent de ceux explorant des règles d'association, d'appariement probabiliste et ceux des modèles graphiques selon (Woodman et Mangoni, 2023). Selon eux, les algorithmes d'exploration de règles d'association (*Association Rule Mining : ARM algorithms*) mirent à trouver des combinaisons de différents items qui surviennent plus fréquemment que d'autres comme trouver des modèles cachés dans un texte médical. De plus, les algorithmes *ARM* comme *Apriori* (Agrawal et Srikant, 1994)*, AIS* (Agrawal et coll., 1993) ou *SETM* (Houtsma et Swami, 1993) permettent la génération de règles d'association[30]. De même, des mesures de probabilité conditionnelle peuvent être utilisées pour définir le degré d'incertitude dans les règles.

Autant, les algorithmes d'appariement probabilistes (*probabilistic matching algorithms*) estiment la probabilité d'une correspondance entre les éléments d'une collection. Également, ils sont nécessaires dans les cas où les données sont dépendantes et où les taux d'erreur sont faibles. Comme *Las vegas* ou *Monte carlo*, ils utilisent un degré aléatoire comme partie de leur procédure (Woodman et Mangoni, 2023). De même, ces algorithmes se comportent différemment lorsque nous les faisons exécuter plus d'une fois sur les mêmes données. Ils ont différents temps d'exécution et le résultat peut changer amplement à chaque accomplissement[31].

[30] Un exemple de règle d'association est que les personnes qui ont mal au dos (le conséquent) ont une tendance à faire du télétravail (l'antécédent de la règle).

[31] *Las Vegas* retourne, dans la plupart du temps, une réponse exacte, mais parfois ne trouve pas de réponse du tout, et son temps d'exécution est de nature inconnue en étant dépendant aux données aléatoires. Le *Monte Carlo* utilise

Aussi, ces auteurs estiment que les algorithmes probabilistes peuvent être plus adaptés aux dossiers de santé électroniques (DSE) que les algorithmes déterministes, puisque ces DSE contiennent des données avec des erreurs de saisie et peuvent donc obtenir une proportion plus élevée de liens.

De plus, les modèles graphiques probabilistes (*Probabilistic Graphical Model : PGM*) sont des cadres utilisés pour modéliser des associations entre les variables dans des domaines complexes. Aussi, les *PGM* permettent la modélisation de la causalité, l'amélioration de l'explicabilité, l'intégration des connaissances des experts, l'incorporation des données corrélées pour une représentation des relations probabilistes entre ces variables et des données hétérogènes dans un environnement d'incertitude avec la spécification de la dépendance des variables aléatoires (nœuds) d'un problème. Les *Bayesian Network (BN)* et *Hidden Markov Model (HMM)* sont des exemples de *PGM* pour représenter des modèles probabilistes avec une structure graphique (Woodman et Mangoni, 2023).

Premièrement, le *BN* est un modèle probabiliste représentant les relations causales entre les nœuds à l'aide d'un graphe orienté acyclique (*Directed Acyclic Graph*) (Bai, 2023). Aussi, chaque nœud a une attribution conditionnelle en fonction de ses parents. Un *BN* se base sur la théorie des graphes et les probabilités permettant de classer les entrées en fonction des connexions dirigées entre les variables et du degré de corrélation exprimé par des probabilités (Biedermann et Taroni, 2022). De même, les *BN* fournissent une analyse décisionnelle cohérente dans des conditions d'incertitude. Pareillement, ils peuvent être générés en utilisant soit des connaissances d'experts, soit en étant appris à partir des données et ensuite utilisés pour estimer les probabilités d'événements ultérieurs (Woodman et Mangoni, 2023). Deuxièmement, « le modèle de Markov, aussi appelé chaîne de Markov, est un modèle statistique composé d'états et de transitions. »; et « un modèle de Markov caché est basé sur un modèle de Markov, sauf qu'on ne peut pas observer directement la séquence d'états : les états sont cachés »[32]. Contrairement à une chaîne de Markov classique où les transitions sont inconnues, mais où les états sont connus, dans un *HMM*, les états sont inconnus. Identiquement, selon Woodman et Mangoni (2023), les nœuds des *PGM* et les

l'aléatoire pour retourner la meilleure réponse possible avec une grande probabilité, mais peut se tromper avec une certaine incertitude; et son temps d'exécution est toujours borné.

[32] https://igm.univ-mlv.fr/~dr/XPOSE2012/HiddenMarkovModel/description.html

arêtes se reliant représentent leurs relations. Selon eux, leurs distributions de probabilité et relations entre eux en capacités (bords) sont spécifiées subjectivement, avec le modèle capturant la « croyance » d'un domaine complexe. Identiquement, ces auteurs évaluent que les dépendances desdits nœuds sont capturées via les arêtes dirigées avec les connexions manquantes définissant des indépendances conditionnelles.

3.4 La logique floue probabiliste

Nous nous étions résolus à la *PFL* comme solution pour l'interprétabilité et l'explication causale. De la sorte, nous avons misé sur cette solution qui tente de représenter le raisonnement du cerveau dans ses prises de décision. De même, la *PFL* est un outil récent. Elle est peu exploitée. Il est communément admis que la probabilité et le flou sont complémentaires plutôt que compétitifs (Zadeh, 1996). Aussi, il est démontré que beaucoup des concepts de la théorie des probabilités standard sont transposés dans le contexte flou (Gudder, 1998). Pareillement, la *PFL* intègre le flou et le probabiliste en comblant le fossé entre les deux. En parallèle, pour la *PFL* et dans l'exemple de l'introduction, il y a 71 % de probabilités que la bactérie A ait une valeur d'appartenance « élevée » de 0,7. D'autre part, un exemple de *PFL* implique des règles dont les antécédents sont des conditions floues typiques dont les conséquences sont des distributions de probabilité (Razab-Sekh, 2021).

Aussi, « en général, les classificateurs flous probabilistes auront la forme suivante :

$$\begin{aligned} \textit{Rule } R_j \textit{ : IF x is } A_j \textit{ THEN} &= \omega_1 \textit{ with probability } Pr(\omega_1|A_j) \\ &= \omega_2 \textit{ with probability } Pr(\omega_2|A_j) \\ &\ldots \\ &= \omega_c \textit{ with probability } Pr(\omega_c|A_j) \end{aligned}$$

où j = 1, ..., *J* correspond au numéro de règle, *Aj* est l'ensemble flou antécédent de la *j* ième règle et $Pr(\omega_c|A_j)$ est la probabilité de la classe respective ω_c pour la règle *j*. Les ensembles flous antécédents A_j sont définis dans l'espace d'entrée *X* à *N* dimensions. Enfin, les paramètres de probabilité $Pr(\omega_c|A_j)$ doivent satisfaire $Pr(\omega_c|A_j) \geq 0$, pour $c = 1, \ldots C$ et $\sum_{c=1}^{C} Pr(\omega_c|A_j) = 1$. Soit

maintenant, μ_{Aj} désignant la fonction d'appartenance d'un ensemble flou A_j.» (Razab-Sekh, 2021).

Autant, selon Faghihi et coll. (2021), un exemple de *PFL* excelle dans des problèmes réels et dans le raisonnement avec des degrés de certitude (Yager et Zadeh, 2012) traitant trois incertitudes : l'aléatoire, l'incertitude probabiliste et le flou (Robert et coll., 2020); elle peut à la fois gérer l'incertitude de la connaissance par l'utilisation des probabilités; avec le flou inhérent à la complexité du monde par la fuzzification (Yager et Zadeh, 2012) et elle a été utilisée pour la recherche des causes des événements (Robert et coll., 2020). Ainsi, il est pertinent de combiner la probabilité et le flou.

Aussi, dans le domaine de l'IA, le flou est souvent associé à des caractéristiques : de gestion flexible des données ambiguës, imprécises ou incertaines avec un degré d'appartenance relié à un raisonnement d'observables vagues. En ce qui concerne les probabilités, il s'agit d'une gestion de l'ignorance et l'indétermination quant à la survenance d'un observable. De plus, ce sont souvent des caractéristiques intrinsèques à la cognition. La *PFL* permet une bonne capacité de prédiction sur des cas que nous ne connaissons pas et sur ce que nous connaissons déjà. De ce fait, on se rapproche plus d'une prédiction précise.

Pareillement, « un nouveau concept de logique floue probabiliste est introduit comme un moyen de représenter et/ou de modéliser le caractère aléatoire existant dans de nombreux systèmes du monde réel » (Meghdadi et Akbarzadeh-T, 2001). Subséquemment, Ishibuchi et Nakashima (2001) ont utilisé des règles de poids correspondant aux probabilités dans certaines conditions. Ensuite, une autre approche a été développée dans le travail (Waltman et coll., 2005) où les paramètres sont évalués simultanément en utilisant l'estimation du maximum de vraisemblance. Aussi, l'article de Fialho et coll. (2016) (Figure 6) détaille une technique pour anticiper le décès des patients à la suite d'un choc septique.

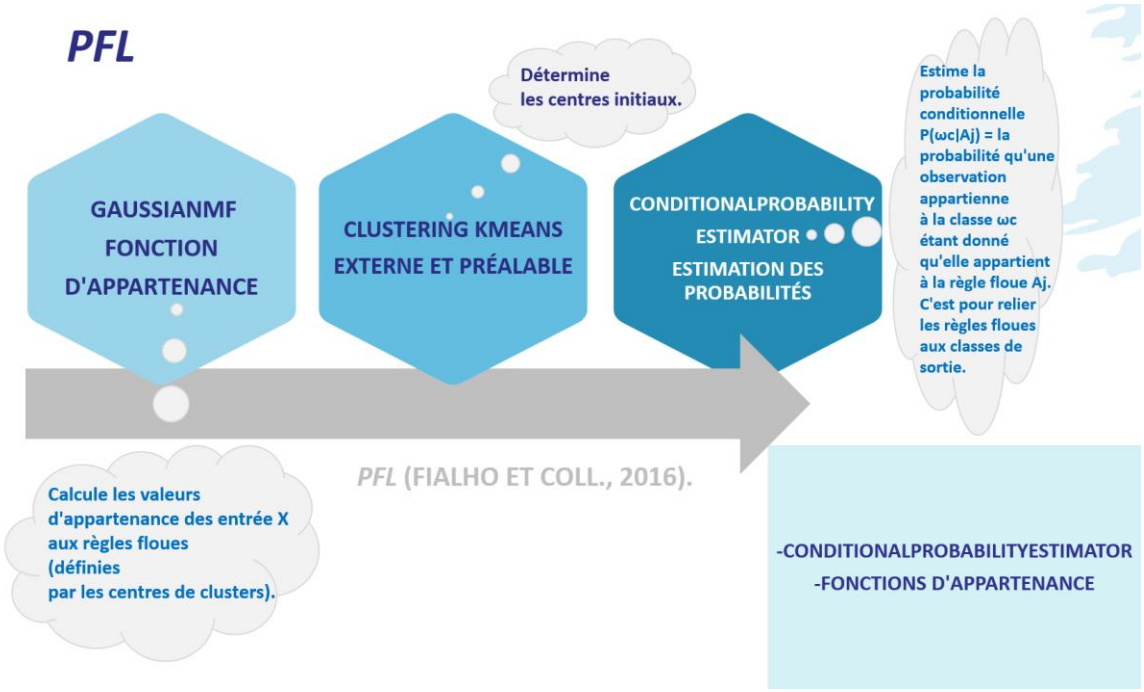

Figure 6 : *PFL* (Fialho et coll., 2016).

Ensuite, Ambag et coll. (2023) ont proposé une nouvelle méthode de prise de décision interprétable basée sur les arbres probabilistes et la théorie des ensembles flous. De même, ils ont testé et validé le *Fuzzy Probability Tree (FPT)* sur deux études de cas cliniques. Dans l'ensemble, ce *FPT* a obtenu un bon score en détectant les patients souffrant de la maladie. Pareillement, l'article de Jamshidnejad et De Schutter (2024) a présenté une méthode qui a associé les probabilités et la logique floue pour une modélisation dynamique, en considérant les incertitudes des données. En parallèle, leur technique se servait d'ensembles flous pour incorporer simultanément les incertitudes de nature probabiliste et floue. Leur méthode a optimisé la fiabilité des prédictions et fournissait des modèles en mesure de mieux administrer les interactions complexes. L'article de Cardiel-Ortega et Baeza-Serrato (2024) a suggéré un système probabiliste flou. L'incertitude était intégrée dans leur modèle qui a employé la théorie des probabilités et le flou. L'interprétation de leur système était simple, fiable et facilite la prise de décisions. Le travail de Li et coll. (2024) a présenté un algorithme reposant sur une méthode floue de regroupement probabiliste incertaine, en incorporant les impacts des données locales pour optimiser la gestion du bruit et des répartitions aléatoires des données. Également, leur algorithme a renforcé la précision et la robustesse face au bruit en exploitant la distribution

gaussienne pour déterminer la "dissimilarité". Autant, leur article a suggéré une technique associant un ensemble probabiliste flou à une optimisation du regroupement par clustering.

Par contre, selon Meng et coll. (2022), si les règles floues sont directement apprises à partir d'un entraînement de peu de données, un problème pourra être dans des règles floues acquises ne correspondant pas au mieux aux données fournies. De ce fait, selon eux, l'objectif est d'avoir des résultats de modélisation pouvant être interprétés comme des règles activées en probabilité. Subséquemment, selon ces auteurs, un exemple de solution s'illustre par un *PFL* en deux étapes avec une première étape d'entraînement, en se basant sur l'apprentissage bayésien pour maximiser les probabilités conjointes des données et de ces règles floues en s'appuyant sur le gaussien pour classer les données et apprendre leurs probabilités, ce qui permettra d'apprendre les règles et de s'adapter à la fois au flou et aux incertitudes stochastiques. Ultérieurement, selon eux, dans une deuxième étape, une inférence *PFL* basée sur l'apprentissage bayésien pourra être développée pour que les règles floues activées correspondant au mieux aux paires de données d'entrée et de sortie et interprétant les résultats de la modélisation avec ces règles sous un maximum de vraisemblance. *In fine*, la *PFL* permet de se rapprocher de l'équilibre et de la perspective de valorisation explorés dans ce projet pour une meilleure collaboration en interaction entre la machine et l'expert pour de nouvelles connaissances.

3.5 Conclusion

L'étude conjointe des notions de flou, de probabilités et de la *PFL* met en évidence une convergence indispensable pour comprendre intégralement la complexité des incertitudes propres aux systèmes réels. Tandis que les probabilités sont particulièrement efficaces pour quantifier l'aléatoire et la logique floue pour modéliser l'imprécision, la *PFL* apparaît comme un cadre évolué apte à traiter ces deux aspects de l'incertitude de manière simultanée. La faculté de la *PFL* à assimiler des données stochastiques et incertaines confère une flexibilité notable à cette méthode dans un environnement sensible. Les progrès récents dans les bases théoriques, les algorithmes d'inférence puis l'utilisation pratique de la *PFL* illustrent son développement croissant et sa capacité à interpréter la prédiction.

En fin de compte, certains enjeux subsistent, particulièrement en ce qui touche à la précision des *PFL* et au développement de méthodes robustes pour l'obtention ainsi que la validation des connaissances floues et probabilistes. Ensuite, l'interaction entre le flou et les probabilités représentée par la *PFL*, offre des opportunités encourageantes pour l'évolution de systèmes plus interprétables, adaptatifs, résilients et permettant l’analyse causale.

CHAPITRE 4 :

REVUE DE LA LITTÉRATURE : EXPLICATION CAUSALE

4.1 Introduction

Au centre de l'approche scientifique se trouve la recherche du discernement et de l'explication des phénomènes. Une explication scientifique dépasse la simple observation de corrélations afin de fournir une connaissance approfondie des processus et des motifs qui sous-tendent un phénomène. L'émergence et la multiplication des modèles IA sophistiqués ont rendu nécessaire le fait de pouvoir offrir des explications, ce qui a conduit à l'apparition du principe d'explicabilité. L'objectif de l'explicabilité est de rendre les décisions et les prédictions des modèles non transparents intelligibles pour l'homme en dévoilant les mécanismes internes qui les gouvernent.

Ainsi, les techniques de regroupement destinées à structurer des données non labellisées en ensembles homogènes basés sur leurs ressemblances sont recommandées pour la révélation de structures invisibles. Dans ces techniques, le clustering bayésien se démarque par son orientation probabiliste, intégrant l'incertitude dans le processus de classification. Cette méthode, basée sur le théorème de Bayes offre non seulement une robustesse face aux perturbations et aux données absentes, mais elle donne aussi une évaluation du degré de certitude associé à chaque attribution d'un point de donnée à un groupe. L'importance primordiale réside dans l'interaction entre ces notions, l'explication scientifique pour une intellection approfondie, l'explicabilité pour la clarté des modèles, et le regroupement (par exemple, bayésien) pour la découverte structurée probabiliste dans les recherches contemporaines.

Assurément, la revue de littérature suivante examine la manière dont ces secteurs se superposent et s'appuie mutuellement pour surmonter les obstacles complexes de l'analyse des données et de l'élaboration de connaissances robustes et intelligibles. Elle met également en évidence les liens de causalité qui sous-tendent les dynamiques du monde réel.

4.2 La causalité

Le cerveau utilise quotidiennement la causalité pour désigner une relation entre deux ou plusieurs observables. Ainsi, la causalité fait partie de la vie de l'être humain en jouant un rôle important dans la cognition et l'aide à la décision. Aristote avait écrit : « nous pensons avoir connaissance d'une chose seulement lorsque nous en avons compris la cause » (Physics, 194 b17–20) » (Acharki, 2022). De plus, les vues philosophiques des causes sont des vues de régularité où elles sont fréquemment suivies par leurs effets probabilistes où elles les changent, mécanistes, manipulationnistes et contrefactuelles où aucun des deux n'est concrétisé (Maziarz, 2020). De même, la vue des résultats contrefactuels est une vue représentant deux variables aléatoires ne pouvant pas véritablement se produire simultanément (Hernán et Robins, 2010). Également, les contrefactuels, l'association et l'intervention sont trois niveaux distincts de capacités cognitives où le cerveau organise ses connaissances du monde en les incarnant dans des échelons distincts de l'échelle de la causalité (Pearl et Mackenzie, 2018). Ensuite, entre la cause et l'effet, il y a trois niveaux de causalité : absolue où il y a nécessité et suffisance entre les deux; conditionnelle : s'il y a nécessité sans suffisance et contributive : sans nécessité et sans suffisance[33]. La découverte de la relation entre la cause et ses effets est appelée raisonnement causal. Dans ce raisonnement, nous analysons cette relation et essayons de comprendre comment et pourquoi quelque chose s'est produit. Pareillement, le raisonnement est une composante importante de l'explication scientifique.

Si bien que, « dans la vision de Hume, la causalité est déterminée par trois facteurs empiriques principaux : (i) la cause précède l'effet, (ii) la cause et l'effet sont liés temporellement et spatialement (iii) la cause et l'effet covarient » (Béghin, 2022). Hume soutient que la cooccurrence des événements observée par expérience ne donne pas d'informations causales utiles (Miller, 2019). Aussi, selon Hume, la formalisation la plus traditionnelle de la causalité est celle où nous pouvons affirmer qu'un événement E_1 est responsable de E_2 si les deux conditions suivantes sont remplies : à chaque fois qu'on examine E_1, on voit ensuite E_2 (E_1 est une condition suffisante pour E_2); et à chaque fois que l'on observe E_2, on a déjà vu E_1 (E_2 est une condition nécessaire de E_1). Également, pour les causes multiples nécessaires et ceux suffisants, un ensemble de E sont tous

[33] http://media.acc.qcc.cuny.edu/faculty/volchok/causalMR/CausalMR3.html

nécessaires pour provoquer un seul E et dans un autre schéma, il existe plusieurs manières possibles de provoquer cet événement E et une seule d'entre elles est requise (Kelley, 1987).

Aussi, la théorie de la causalité de l'interventionnisme (Halpern & Pearl, 2005) présente un outil d'intervention ou de manipulation stipulant que si la cause C comme levier d'un événement E, alors on peut contrôler et agir sur E en modifiant C. Ainsi, la causalité est liée à notre capacité cognitive à intervenir sur le cours des E. Ensuite, un certain facteur F est statistiquement important pour un événement E si sa présence entraîne une modification de la probabilité de sa survenue. Un ensemble de E explicatifs et statistiquement pertinents augmente la probabilité de la justification de l'expert. Les théories probabilistes s'appuyant sur l'interventionnisme dans le contexte causal sont reliées à l'idée que modifier un facteur causal devrait modifier la probabilité de son effet. Reichenbach (1956), un défenseur de cette idée, a observé que lorsqu'une cause commune est présente, la corrélation entre ses effets disparaît. Par la suite, l'exploration des liens entre causalité et probabilité a mené au développement des outils bayésiens causaux des outils intéressants pour représenter les relations complexes.

Autant, le raisonnement causal est généralement considéré comme une forme inductive. « La prédiction est une estimation de ce qui va se passer dans le futur; elle se fait généralement en trouvant des modèles inductifs. » (Solomonoff, 2009). Aussi, l'induction où la répétition d'un observable accroît la probabilité de le revoir dans l'environnement, la déduction où on va de la cause aux effets et l'abduction permettant une recherche des causes avec une interprétation de l'observable sont trois formes de la pensée. Ces trois formes font partie du raisonnement et de la cognition du cerveau.

Également l'abduction, initialement formulée par Aristote, était réinventée par Peirce en 1965. Selon ledit, c'est une inférence de la meilleure hypothèse explicative à partir d'événements (Gaudet & Robert, 2018). Il l'avait décrit comme « l'art de deviner ». Il a perçu sa contribution opérationnelle à la mémoire. Ensuite, c'est un mécanisme cognitif spontané s'appuyant sur des connaissances et en génère d'autres. C'est la première croyance en recherche de cause. De même, selon Serge Robert, le raisonnement logique permet de découvrir que l'abduction, bien que pertinente pour connaître, est non logiquement valide. Il mentionne que pour ce raisonnement, si les prémisses sont vraies, la conclusion l'est nécessairement et elle est possiblement vraie pour

l’abductif. De surcroît, selon lui, il y a deux types d’abductions : rétrospective d’explication où la cause est recherchée à partir de l’effet et prospective de prédiction de l’effet à partir de la cause. Il souligne que dans le raisonnement spontané en contexte de découverte, le respect de la logique est un accident abductif.

Pareillement, l'abduction est d'une importance scientifique majeure parce qu’elle touche les problématiques de causalité et d'*XAI* (Ghidalia, 2024). Selon Karam (2010), le diagnostic médical ou la représentation de l’incertitude sont des exemples de contextes couvrant les applications abductives. Il évoque que l’approche probabiliste de Pearl avec des probabilités calculées par le théorème de Bayes (Arbib, 2003) était un exemple d’approche d’abduction où la déduction est inversée. De même, il indique que disposant d'un champ de connaissance et d'une observation à élucider, l'inférence abductive génère une ou plusieurs hypothèses explicatives basées sur ces contextes. Également, il mentionne que cette approche a exposé la connaissance causale relative à une situation spécifique et son côté bayésien reposé sur la découverte de la causalité comme fondement structurant de la connaissance du contexte examiné.

De même, l'abduction est une inférence plausible jusqu'à sa confirmation, surtout pour les contextes nuancés. De plus, sa sélection d'une interprétation jugée meilleure peut conduire à une focalisation sur une hypothèse spéculative empirique et sans fondement dès les premières étapes de l'analyse. Elle apporte une information inédite susceptible d'être vraie ou fausse, se limitant à suggérer une possibilité unique sans la prouver. Aussi, elle s'appuie fréquemment sur des évaluations subjectives et difficilement mesurables. Dans les systèmes complexes, on peut rencontrer plusieurs causes interconnectées. Il y a des risques de biais cognitifs.

De la sorte, en raison de ses contraintes et pour éviter des risques dogmatiques et rigides, l’expert peut utiliser une attitude scientifique de remise en question et d’analyse critique en utilisant un processus de calcul de probabilités et de logique floue. Également, à travers la richesse de ce processus, il peut corriger les croyances spontanées et ainsi ses connaissances. De ce fait, chaque élément est examiné de façon nuancée, évitant les dérives abductives qui pourraient mener à des intolérances. Aussi, ce processus permet d’évaluer systématiquement les hypothèses et leur robustesse. Pareillement, cela implique une diminution de biais et une confrontation des suppositions aux données accessibles. De même, cela met à l'épreuve les hypothèses en les

comparant à des preuves empiriques et à des normes de cohérence organisée. En parallèle, c'est l'assujettissement à un examen approfondi afin d'éviter d'accepter des hypothèses non justifiées. Joint à cela, il endosse une modification des modèles assurant ainsi la fiabilité des déductions. Ensuite, il met en évidence les degrés d'incertitude et encourage l'expert à maintenir diverses pistes d'exploration. En optant pour ce processus, il y a aussi une assurance méthodologique, un questionnement critique et une actualisation des convictions en se basant sur les preuves en évitant de ce fait les conclusions rigides. Par conséquent, il confrontera les prédictions aux données et corrigera ainsi l'abduction.

En conséquence, si un expert (ex. médecin) observe des effets et des signaux (ex. symptômes) d'une observation (ex. éruption cutanée chez un patient), il pourrait "abduire" fortement une seule problématique (ex. maladie virale, bien que d'autres causes comme une allergie soient également envisageables). En revanche, avec la *PFL*, il interprétera les vraisemblances et plus au moins les nuances des effets. Ensuite, il pourra décomposer les causalités, adapter et personnaliser un traitement robuste (ex. polythérapie) en ajustant les quotités (ex. doses de médicaments).

4.3 L'explication scientifique

Hempel et Oppenheim ont fondé une théorie de l'explication scientifique en lien avec le positivisme (empirisme) logique (1948), puis Hempel l'a reprise en 1965 en proposant le modèle Deductive-Nomological (D-N) (Salmon, 2006). Dès lors, pour l'aspect nomologique, Hempel remarque que les conditions initiales d'une expérience sont liées à des lois générales. De là, certains faits spécifiques produits par cette expérience sont déduits dans un système logique. Depuis ce modèle, cette explication est devenue l'un des sujets importants des sciences cognitives. Aussi, une explication est définie par la plupart des philosophes comme une réponse à la question « pourquoi », qui vise à reconnaître et à décrire la logique derrière l'apparition d'événements spécifiques (Salmon, 1984). De plus, l'explanandum désigne ce qui décrit l'événement à expliquer (et non l'événement en tant que tel), tandis que l'explanans (ce qui explique) regroupe les éléments invoqués pour interpréter cet événement (Hempel & Oppenheim, 1948). Globalement, une explication peut être peinte comme un modèle composé

d’explanandum et d’explanans. Ainsi, ce modèle inclut des variables, des causalités, des probabilités et des relations les reliant.

Autant, une fois que les relations statistiques entre les événements ont été identifiées, Salmon chercha à expliquer les événements en analysant la notion de pertinence statistique. Ainsi, selon lui, l'explication scientifique commence par l’analyse des relations stochastiques dans les données. Encore, son modèle de *Statistical Relevance* (SR) introduit cette notion pour vérifier les processus influençant ce qui est pertinent en termes d'explanandum (Salmon, 2010). Aussi, dans ce SR, l’impact de l’explanans comme regroupement de faits statistiquement pertinents afin que l’expert fournisse son explication et augmente la probabilité de l'explanandum. Subséquemment, cet impact en détermine la pertinence. Ensuite, l'interprétation de l’expert est appropriée et justifiée.

De plus, le SR n'est pas convenablement développé pour donner une explication suffisante (Salmon, 1998). Ainsi, selon Salmon et Pearl, la science doit se plier à ce qui semble être un ordre causal du contexte. Certaines structures causales sont largement sous-déterminées par la pertinence statistique (Pearl, 2019). Encore, dans un temps subséquent, afin d'apporter une explication exhaustive, il s'agissait de faire un compte rendu de façon causale, c'est-à-dire en termes d'interaction de processus de la notion de pertinence statistique et de ses relations (Woodward, 1989). Les mêmes relations de cette pertinence entre leurs éléments peuvent servir à décrire diverses structures causales. Pareillement, cela rend facile la différenciation des liens causaux directs avec une analyse de cette pertinence. De cette manière, si l’analyse de la pertinence statistique est importante, elle doit néanmoins être mise en perspective avec une interrogation sur les fondements mêmes de la causalité.

Également, le rapport de nécessité entre les énoncés explicatifs et l'événement à expliquer est logique. Mais, cette notion logique ne suffit pas. L’objectif est d'obtenir une satisfaction pour l’explication scientifique (Salmon, 1984). De plus, selon Weber et coll. (2013), en 1984, Salmon a proposé un modèle pour cette explication qui s'appuie sur la causalité. Aussi, selon eux, son ouvrage a eu une ambition centrée autour de l’entendement de cette explication; il commence par identifier la structuration d'un événement qui a besoin d'explication au sein d'un réseau causal avec d'autres événements qui l'auraient précédé dans le temps et il a promulgué ensuite une idée

qui, de prime abord, semble assez simple, mais qui dans son développement fournit une vision causale de ce que pourrait être une explication scientifique.

Identiquement, selon González del Solar et coll. (2019), Salmon précise qu'il ne faut pas trop se focaliser sur des explications déductives et négliger d'autres facettes vitales de l'explication causale. Selon ces auteurs, pour Wesley, une explication scientifique authentique ne se limite pas à présenter une déduction, elle doit également inclure des facteurs causaux pour offrir un aperçu plus exhaustif de la raison d'un événement; il adopte ensuite une démarche plus mécaniste dans le but de démontrer que certaines explications scientifiques requièrent la détection de processus sous-jacents et il estime que ces mécanismes également appelés descriptions de processus causaux sont déterminants pour fournir une explication exhaustive et compréhensible, surtout dans le domaine des sciences où les phénomènes ne peuvent pas se limiter à des lois universelles.

Aussi, une explication est une attribution de responsabilité causale ((Miller, 2019) citant les travaux de (Josephson et Josephson, 1996)). Selon Salmon, l'explication en termes causaux: expliquer un événement c'est citer ces causes nécessaires et suffisantes. Ainsi, tandis que cette vision met en avant des conditions empiriques fondamentales pour déterminer la causalité et s'appuie sur les notions de conditions nécessaires et suffisantes, l'interventionnisme enrichit cette perspective en introduisant la capacité d'agir sur les causes pour modifier les effets reliant ainsi causalité et probabilité à travers des outils bayésiens pour affiner notre explication causale.

En plus, Druzdzel & Simon (1993) ont abordé le problème de l'interprétation causale dans une structure bayésienne. Selon eux, la connaissance causale dans un système est nécessaire pour prédire les effets des changements dans sa structure et, en raison du rôle de la causalité dans le raisonnement, elle est vitale dans les interfaces homme-machine des systèmes d'aide à la décision. Pearl a développé des modèles bayésiens pour spécifier les relations causales, en soulignant que la causalité ne peut pas être complètement capturée par le calcul des probabilités. De la sorte, on peut incorporer l'explanandum et l'explanans en utilisant des modèles probabilistes pour formaliser les relations causales. Pearl critique l'approche purement bayésienne affirmant que certaines connaissances sont organisées autour de relations causales plutôt que probabilistes (Pearl, 2001). Il avait écrit que « les considérations de fiabilité (du jugement) appellent à enrichir

le langage des probabilités d'un vocabulaire causal et à admettre les jugements causaux dans le répertoire bayésien. ».

Tout bien considéré, il y a en réalité deux processus dans l'explication et un produit : 1. Produit : l'explication découlant du processus cognitif est un résultat de l'explication cognitive, 2. Processus social de transfert de connaissances entre l'explicateur et la personne à qui l'on explique un processus social qui se déroule généralement entre un groupe de personnes et 3. Processus cognitif : l'inférence abductive pour trouver une explication pour un E spécifique d'explanandum, où les causes de l'événement E sont identifiées et un sous-ensemble de ces causes est choisi comme explication ou explanans (Miller, 2019).

4.4 L'explicabilité et la causalité

Des chercheurs ont analysé trois perspectives où la causalité et l'*XAI* peuvent être liées : critiques de l'*XAI* sous l'angle de la causalité, l'*XAI* pour la causalité et la causalité pour l'*XAI* (Carloni et coll., 2025). L'analyse de l'article de ces auteurs démontre que la causalité et l'*XAI* sont étroitement liées. Ainsi, cette analyse met en évidence trois points de vue majeurs sur ce lien : le premier soulève l'absence de causalité comme un frein majeur des méthodes actuelles de l'*XAI*, le second perçoit l'*XAI* comme un instrument facilitant l'investigation causale et la dernière envisagerait la causalité comme une préparation préliminaire à l'étude de l'*XAI*. De plus, ils considèrent que ce point de vue est le plus prometteur pour la conception de systèmes véritablement fiables et bénéfiques pour l'homme.

Autant, en ce qui concerne leur première perspective, la causalité est une critique à l'*XAI*. Aussi, elle a été repérée dans la littérature en tenant compte de l'absence de causalité comme l'un des principaux défauts des méthodes actuelles d'*IA* et d'*XAI*. Également, cette représentation critique interroge les aspects fondamentaux de la robustesse des *XAI* actuels. Pareillement, un nombre considérable de publications identifie la causalité comme un élément absent dans les recherches d'*XAI* actuelles pour parvenir à des systèmes robustes et explicables.

Aussi, leur deuxième perspective envisage l'*XAI* comme un outil d'exploration causale permettant d'accéder à la causalité. De même, les recherches estiment que l'*XAI* pourrait stimuler

l'exploration scientifique dans le cadre de l'investigation causale. En parallèle, cette démarche considère l'*XAI* comme un instrument apte à repérer des manipulations expérimentales méritant d'être poursuivies, assistée ainsi par les chercheurs dans la formulation d'hypothèses concernant de potentielles relations causales à examiner. Cette représentation renverse le lien conventionnel entre cause et explication, établissant l'*XAI* comme un catalyseur de la découverte causale plutôt qu'en tant que résultat final.

Encore, pour leur troisième perspective, la causalité est vue comme un fondement pour l'*XAI*. Elle est jugée la plus prometteuse en défendant l'idée que la causalité est préliminaire à l'*XAI* de trois façons possibles. D'abord et en premier lieu, elle utilise les notions tirées de la causalité pour appuyer ou perfectionner l'*XAI*. En deuxième lieu, elle fait appel à des contrefactuels pour assurer l'explicabilité. En troisième lieu, elle voit dans l'accès à un modèle causal une explication en soi. Cette méthode admet que l'on peut utiliser des outils et des indicateurs causaux afin de mettre en œuvre l'*XAI* et que des stratégies précises d'*XAI* peuvent être référées à leur définition causale formelle pour optimiser leurs aptitudes à la généralisation. Le concept unique de cette approche est que l'accès au modèle causal d'un système représente un moyen intrinsèque d'en expliquer le fonctionnement.

Au final, selon ces auteurs, le cadre conceptuel de la causalité de Pearl est central pour saisir les limites des méthodes *XAI* actuelles. De même, cette échelle identifie trois niveaux de cognition : l'association (observation passive des données), l'intervention (modification du système) et les contrefactuels (imagination de scénarios alternatifs). En parallèle, ils citent la vision de Pearl estimant que les méthodes traditionnelles d'IA pour la classification ou la régression se trouvent actuellement au premier niveau d'association, expliquant ainsi leur incapacité à identifier la cause de l'effet. Aussi, selon eux, cette contrainte cardinale met en lumière des interrogations capitales concernant la qualité des explications délivrées par les *XAI* actuels, fonctionnant majoritairement sur une base associative sans toucher aux échelons supérieurs du raisonnement causal.

4.5 Le clustering bayésien

Selon Pevneva et Pivovarova (2022), le clustering bayésien probabiliste a des performances aussi bonnes que les méthodes classiques sur des données de test. Aussi, selon ces auteurs, il a une

simplicité, une flexibilité et des capacités de paramétrage. Ensuite, l'article de Zhu, et coll. (2022) a exposé un *BGM* dans une perspective probabiliste. De plus, leur modèle a optimisé la précision dans des contextes de données volumineuses. Aussi, dans l'objectif d'une composante robuste, Wang et coll. (2023) ont proposé un *BGM* qui a détecté les irrégularités potentielles en analysant la répartition des données normales. De la sorte, leur modèle a facilité la gestion des données de grande envergure et a géré les conséquences des données perturbées. Ultérieurement, dans le travail de Mohammed et coll. (2024), le *BGM* a augmenté la précision en incorporant divers scores de catégorisation. Également, leur modèle convenait spécifiquement aux variantes de signification incertaine, offrant ainsi des interprétations plus crédibles dans un contexte probabiliste. Subséquemment, l'article de Guan et coll. (2024) a exposé un *BGM* dans des contextes de bruit complexes. De même, leur modèle a permis un ajustement dynamique du nombre de mélanges gaussiens et a optimisé la précision.

Premièrement, selon ces articles, dans un modèle de mélange gaussien (*Gaussian Mixture Model : GMM*), les données proviennent d'un assortiment de différentes distributions gaussiennes, chacune se distinguant par sa moyenne et sa covariance. De la sorte, le *GMM* détermine et pondère cette moyenne et cette covariance pour chaque composant gaussien. Ce modèle illustre la part de chaque composant qui contribue dans la distribution générale. De même, chaque gaussien est chargé de positionner un unique point de données là où il se trouve. Aussi, dans le *GMM*, on utilise l'algorithme d'espérance-maximisation pour estimer les paramètres, tandis que dans le modèle *BGM*, on se sert de la méthode bayésienne pour les calculer.

Deuxièmement, le *BGM* est un modèle statistique basé sur des mélanges gaussiens qui intègre des principes bayésiens. Également, comparé à d'autres outils ayant des objectifs similaires, tels que les *GMM* ou les *k-means*, le *BGM* présente plusieurs avantages distinctifs découlant de son approche probabiliste. Ainsi, le premier avantage majeur du *BGM* réside dans sa gestion de l'incertitude grâce au bayésien. Aussi, contrairement au *GMM* reposant sur des estimations de maximum de vraisemblance pour les paramètres (en trouvant les valeurs des paramètres qui rendent les données observées les plus probables), le *BGM* utilise des distributions probabilistes pour capturer l'incertitude. De plus, il permet d'obtenir une interprétation probabiliste de ces estimations avec des intervalles de confiance rendant les résultats plus robustes. Ensuite, le *BGM* excelle dans la sélection automatique du nombre de clusters ou de composantes gaussiennes. De

même, il utilise une pénalisation bayésienne pour identifier automatiquement le nombre effectif de clusters nécessaires. Aussi, les composantes inutiles sont éliminées en leur attribuant des poids proches de zéro évitant ainsi des ajustements manuels fastidieux ou le recours à des critères comme les critères : d'information d'*Akaike (Akaike information criterion : AIC)* et d'information bayésienne *(Bayesian information criterion : BIC)*.

Troisièmement, le *BGM* est également particulièrement bon pour gérer des données rares ou bruitées. Les distributions a priori jouent un rôle de régularisation, diminuant les surapprentissages que d'autres outils peuvent rencontrer avec ces données. De la sorte, cela confère au *BGM* une bonne stabilité avec son approche probabiliste. Également, contrairement à d'autres outils qui attribuent chaque point à un cluster de manière déterministe, le *BGM* fournit des probabilités d'appartenance pour chaque point à chaque cluster. Aussi, cela offre une interprétation plus riche. De plus, le *BGM* évite les problèmes de singularité souvent rencontrés dans les *GMM* où certaines composantes gaussiennes peuvent coller excessivement aux données. Subséquemment, le *BGM* peut gérer des distributions complexes et chevauchantes, tout en restant flexible sur le nombre de clusters. De ce fait, il est performant pour des situations dynamiques où le nombre de clusters n'est pas fixe. De même, le *BGM* est un outil puissant pour modéliser des clusters dans des situations où une interprétation probabiliste est nécessaire. Ensuite, il se distingue par sa flexibilité et sa précision. Par ailleurs, les statistiques *BGM* contribuent à optimiser la modélisation des données. De surcroît, les modèles utilisent l'inférence bayésienne pour calculer d'une façon adaptable le nombre d'éléments d'un mélange offrant une bonne visualisation des distributions complexes de données. En plus, pour les *BGM*, l'approche d'inférence variationnelle est employée pour l'approximation des modèles probabilistes complexes. En parallèle, le *BGM* est plus adapté si les données présentent des configurations de groupes irrégulières ou une diversité entre les groupes. Donc, l'adaptation du *BGM* aux données permet de détecter les groupes sous-jacents. En outre, Rubin (1978) avait examiné comment utiliser l'inférence bayésienne pour déterminer les effets de causalité. Pareillement, il avait évoqué comment la méthode bayésienne pouvait incorporer des données préexistantes et traiter l'incertitude statistique dans le calcul des effets causaux.

4.6 Conclusion

L'expert est responsable de la justification de ses explications scientifiques. Ensuite, comme on avait mentionné Miller (2019), citant les travaux de Josephson et Josephson (1996), où une explication est une attribution de responsabilité causale. Ainsi, la prise de décision de l'expert est enrichie par l'analyse de la causalité de ces explications. La causalité permet de comprendre les relations entre les variables dans les modèles IA et ainsi une meilleure imputabilité des résultats. Cette intégration de la causalité consent que les décisions prises par ces modèles soient justifiées. Dans leur article, traitant de cette responsabilité causale en utilisant des comparaisons de probabilités, Qi et coll. (2024) ont écrit que : « alors que l'intégration de l'IA continue de se développer dans des secteurs tels que la santé et l'éducation, la nécessité d'une attribution précise des responsabilités devient de plus en plus importante, en particulier pour les résultats indésirables qui peuvent avoir un impact sur la confiance, la responsabilité et le déploiement éthique de l'IA. »

De surcroît, l'analyse de l'article de Carloni et coll. (2025) montre que la troisième approche est celle qui offre le plus de potentiel pour fusionner efficacement ces deux disciplines. Ainsi, selon eux, cette vision facilite la progression de la recherche en vue d'instaurer des systèmes fiables qui présentent une réelle utilité pour les êtres humains. Aussi, l'innovation de cette méthode se trouve dans sa compétence à réduire minutieusement la distance entre l'*XAI* et la causalité, en soulignant les secteurs à progresser et en révélant les contraintes présentes. De plus, selon eux, l'inclusion de la causalité dans l'*XAI* n'est pas seulement un perfectionnement méthodologique, mais elle est centrale pour la création de systèmes d'IA réellement explicables et dignes de confiance. De même, selon eux, cette récapitulation trace la route vers des orientations de recherche inédites qui ont le potentiel de transformer notre manière d'aborder l'*XAI*.

En outre, dans l'article de Faghihi et coll. (2020), l'inférence causale est utilisée par les informaticiens pour le raisonnement dans le domaine de cette inférence, les scientifiques se penchent sur la détermination du lien entre deux faits observables. Avec cet article, ils ont examiné l'identification de la causalité en se basant sur la *PFL*. Ensuite, ils ont démontré aussi que la *PFL* est plus exacte que le modèle de causalité proposé par Pearl.

Autant, Simon prend en compte les contraintes cognitives et Salmon ancre la causalité dans une structure physique. De ce fait, ces visions se rejoignent dans les travaux de Pearl et Zadeh, enrichissant la conception de la causalité en intégrant le probabilisme et le flou pour mieux interpréter les événements.

Tout bien considéré, l'explication scientifique évolue autour de la relation entre causalité, pertinence statistique et mécanismes explicatifs de l'expert. Initialement basée sur un modèle logique (D-N), l'explication scientifique s'est enrichie pour inclure avec les notions de processus causaux, les probabilités et une reconnaissance croissante de l'importance des outils bayésiens. Une explication justifiée et responsable combine des processus cognitifs pour répondre aux questions du "pourquoi" en reliant l'explanandum à l'explanans. Ainsi, les approches mécanistes et réalistes proposées par Salmon soulignent que cette explication doit s'appuyer sur des relations causales objectives et observables.

CHAPITRE 5 :
ÉTUDE DE CAS POUR LE CHOC SEPTIQUE

5.1 Introduction

Initialement, l'exploration de la causalité et de valeurs, telles que l'éthique dans l'entendement des prédictions relatives à une maladie aussi complexe que le choc septique ainsi qu'à ses données imparfaites et incohérentes a suscité notre intérêt. Aussi, ces observations ont été faites par des chercheurs.

En outre, le choc septique constitue une situation d'urgence médicale majeure marquée par une défaillance aiguë due à une réaction inflammatoire non maîtrisée face à une infection. Cette infection, couplée à un taux de mortalité important, représente un défi considérable en matière de diagnostic précoce, d'estimation précise et de prise en charge thérapeutique sur mesure. La modélisation et la prédiction de son évolution par des méthodes conventionnelles sont compliquées en raison de sa complexité impliquant des interactions non linéaires entre différents systèmes physiologiques.

Également, confrontées à cette complexité, les approches connexionnistes, en particulier les *DL* se sont révélées être des instruments prometteurs pour la précision et l'association. Leur aptitude à déchiffrer des motifs compliqués et à gérer d'importants volumes de données variées les positionne comme des candidats parfaits pour la détection précoce des indicateurs de cette maladie ou pour anticiper la réponse aux soins. Toutefois, ces modèles manquent fréquemment d'interprétabilité et d'analyse causale. Cela restreint leur acceptation dans le domaine clinique où l'intellection des processus sous-jacents aux décisions est nécessaire.

Autant, afin de surmonter cette contrainte et d'incorporer la multitude de données incertaines et imprécises propres au contexte sensible, la *PFL* propose une méthode utile, innovante, puis originale en associant la robustesse de la modélisation probabiliste à l'aptitude de la logique floue à tenter de représenter le raisonnement ainsi que les expertises. La *PFL* offre la possibilité

d'élaborer des modèles plus interprétables et adaptables. L'intégration de la *PFL* dans le domaine du choc septique pourrait potentiellement faciliter la combinaison de diverses données cliniques avec des notions médicales imprécises (telles que « fréquence cardiaque élevée avec une certaine probabilité ») dans le but d'élaborer des indices de risque ou des conseils thérapeutiques plus personnalisés, détaillés et intelligibles.

En définitive, toute initiative de recherche et de développement de traitements pour des affections médicales aussi graves que le choc septique doit se conformer à un cadre éthique strict. L'acquisition, la conservation et l'exploitation de données de santé sensibles suscitent des interrogations déterminantes liées à la protection de la vie privée des patients, au consentement pleinement informé, à la protection des données ainsi qu'à l'évitement des préjugés algorithmiques. Il est également indispensable de garantir la validité et la généralisation des modèles élaborés en assurant la fiabilité et la représentativité des sources de données. L'étude de cas suivante analysera divers aspects de l'implémentation des méthodes connexionnistes et de la *PFL* à la prise en charge du choc septique, tout en mettant en avant l'importance primordiale d'une démarche éthique et scrupuleuse dans l'utilisation des données médicales.

5.2 Le choc septique

La septicémie est une réaction sévère du corps susceptible de provoquer des lésions aux tissus, des amputations, ou le décès (Liu et coll., 2023). « « Le sepsis est considéré comme un dysfonctionnement d'organes potentiellement mortel résultant d'une réponse dérégulée de l'hôte à l'infection, et dont la forme la plus grave est le choc septique » et « le sepsis touche surtout les âges extrêmes de la vie, les nouveau-nés (sepsis néonatal) et les seniors. L'état septique est responsable d'un décès sur 5 dans le monde, d'après l'Organisation mondiale de la santé (OMS). On dénombre 49 millions d'états septiques chaque année, dont près de la moitié concernent des enfants. L'infection est contractée à l'hôpital pour la moitié des patients environ. Dans les pays industrialisés, le sepsis représente autant de décès que l'infarctus du myocarde. Dans les pays en développement, le sepsis puerpéral demeure une cause de mortalité importante des femmes après leur accouchement. » ; subséquemment « les projections dans l'avenir suggèrent un doublement du nombre de cas d'ici cinquante ans, en particulier en raison du

vieillissement de la population » »[34]. Il a été constaté qu'annuellement, il y a plusieurs millions de décès liés à la septicémie dans le monde. Ensuite, le coût moyen de l'hospitalisation d'un choc septique était de 25 327 € [+/− 23 396] en France (Mageau et coll., 2018). De ce fait, le choc septique a des mortalités avec des coûts élevés, ce qui pourrait être réduit par une IA aidant à la décision.

5.3 Le connexionnisme et le choc septique

La prédiction précoce du choc septique est nécessaire pour un diagnostic et des traitements rapides. De ce fait, les algorithmes *ML* et *DL* ont été utilisés pour développer des modèles pour cette prédiction. Ainsi, ces modèles ont montré des résultats encourageants en termes de précision. Subséquemment, Nemati et coll. (2018) ont combiné les *CNN* et les *Long Shot-Term Memory (LSTM)* pour construire un *DL* permettant de prédire la septicémie à l'aide des données des DSE[35]. De même, le modèle le plus performant avait atteint un *AUROC (Area Under the Receiver Operator characteristics Curve)* de 0,97 parmi de nombreuses IA pour ce diagnostic en prédisant la mortalité dans le temps (Schinkel et coll., 2019). Aussi, Wang et coll. (2021) ont développé un algorithme IA pour la prédiction de la septicémie, créant avec succès un modèle de forêt aléatoire utilisant 55 variables cliniques à partir des données des patients en soins intensifs. De plus, les *ML* et *DL* sur des « mégadonnées » (*big data*) se sont révélés prometteurs dans ce diagnostic et cette prédiction (Ramlakhan et coll., 2022). Ultérieurement, ces *ML* et *DL* peuvent être utilisés avec les DSE et d'autres données cliniques pour identifier des indicateurs de sepsis pouvant ne pas être immédiatement apparents aux cliniciens (Coggins et Glaser, 2022). De la sorte, selon Islam et coll. (2023), les *ML* et *DL* utilisent des signes vitaux (ex. la fréquence cardiaque et la tension artérielle), des résultats de laboratoire (ex. le nombre de globules blancs) et des facteurs cliniques (ex. l'âge et les comorbidités) pour prédire la probabilité de sepsis. Donc, selon eux, en utilisant les DSE provenant des visites aux urgences, ces *ML*, y compris les modèles de régression logistique, *Gradient Boosting*, forêt aléatoire et réseau neuronal, ont présenté de bonnes performances prédictives (ex. *AUROC* de 0,93) par rapport aux modèles de référence[36]

[34] https://www.pasteur.fr/fr/centre-medical/fiches-maladies/sepsis-septicemie

[35] Leur modèle a eu une mesure d'*AUROC* de 0,83.

[36] *AUROC* 0,635 pour *qSOFA*, 0,688 pour *MEWS* et 0,814 pour *SIRS*.

soulignant leur potentiel à améliorer le diagnostic de sepsis chez les patients d'urgence. De surcroît, la capacité du *DL* a permis de développer automatiquement des représentations hiérarchiques à partir de données complexes en générant un grand intérêt pour la prédiction du sepsis (Islam et coll., 2023). Par conséquent, selon ces auteurs, les *DL*, les *CNN* et les *Recurrent Neural Network (RNN)* ont montré des résultats encourageants dans la détection du sepsis. Ils résument les modèles *ML* et *DL* ainsi que les mesures de performance couramment utilisées pour la prédiction du sepsis. Ils les ont répertoriés ci-dessous avec d'autres *ML* : arbres de décision, machine à vecteurs de support, *Naïve Bayes*, et k-voisin le plus proche; d'autres modèles *DL* : *LSTM, CNN, Gated Recurrent Unit (GRU), Multitask Gaussian Process and Attention-based Deep Learning Model (MGP-AttTCN), Temporal Convolutional Network (TCN), CNN-LSTM, CNN-GRU*. De la sorte, selon eux, les indicateurs de performance étaient : *AUROC, Sensitivity (Recall), Specificity, Accuracy, Precision, F1 score, Matthews Correlation Coefficient (MCC), Mean Average Precision (mAP), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Positive Likelihood Ratio (PLR),* et *Negative Likelihood Ratio (NLR)*. Pareillement, selon eux, l'application du *ML* et du *DL* pour prédire la septicémie à l'aide des DSE a attiré une attention considérable pour une intervention rapide. En outre, le modèle *Extremely Randomized Trees* de Zhang et coll. (2023) a montré des performances supérieures à celles des autres modèles. Également, le modèle *Early Assessment of Sepsis Engagement (EASE)* de Guo et coll. (2023) a démontré d'excellentes performances prédictives et une capacité de diagnostic.

5.4 Le flou probabiliste pour le choc septique

Nous estimons que la *PFL* permet l'interprétabilité avec flexibilité, robustesse, explication causale, tentative de représentation du raisonnement et face aux données, une tolérance à l'incertitude et à l'imprécision. C'est une avenue intéressante dans un contexte sensible, telle une maladie aussi complexe que le choc septique.

Primo, la logique floue a été utilisée pour le diagnostic, la détection, l'interprétabilité d'une solution éthique pour le choc septique. De la sorte, elle permet de mesurer l'état imprécis des événements de cette maladie. De ce fait, une étude avait proposé une méthode de regroupement basée sur des K moyens flous pour la prédiction et la classification de la septicémie (Efosa et Akwukwuma, 2013). Ainsi, leur méthode a réussi à agréger des caractéristiques de variables

temporelles et invariantes en utilisant différentes mesures de similarité. De même, ils ont démontré de bonnes performances en tant que classificateurs binaires dans une application médicale visant à prédire l'issue des patients en choc septique. Ensuite, une autre étude a développé un système d'inférence floue utilisant la boîte à outils de logique floue de *Matlab* pour créer un système basé sur les connaissances en diagnostic et en détection de la septicémie (Porouhan, 2023). De plus, ce dernier système comprenait des fonctions d'appartenance en entrée spécifiées par des experts, une fonction d'appartenance en sortie et des règles *« IF-THEN »*. Ensuite, il se comportait bien dans des scénarios hypothétiques et a généré des règles pour l'aide à la décision. Ainsi, on estime que les approches floues seraient prometteuses pour améliorer l'interprétabilité, l'explication causale, les connaissances et la compréhension d'un contexte sensible, tel que le choc septique.

Secundo, la probabilité peut mesurer l'incertitude de notre connaissance qu'une personne est plus ou moins un peu malade de chocs septiques. De la sorte, pour cette maladie, les outils probabilistes sont tolérants à l'imprécision, adaptables et ont une bonne capacité d'apprentissage. De même, ils ont été développés pour aider à la décision clinique dans les cas liés à la septicémie conduisant à des résultats favorables pour les patients avec des prédictions précises de la mortalité (Tsoukalas et coll., 2015). De plus, Ward et coll. (2017) ont utilisé le *ML* pour affiner un *Causal Probabilistic Network (CPN)* et ainsi à évaluer la gravité du syndrome de réponse inflammatoire systémique (*SIRS*). Pour le diagnostic du sepsis, ils ont eu une bonne capacité discriminatoire pour la prédiction de la mortalité à 30 jours. Leurs résultats se comparaient bien aux autres systèmes. Ensuite, ils ont eu une modeste capacité à faire la distinction entre le sepsis et le *SIRS* non infectieux. Aussi, les modèles de *HMM* ont été utilisés pour faire des inférences sur l'état de septicémie des patients sur la base de prédicteurs cliniques fournissant ainsi un diagnostic basé sur les probabilités (Parente et coll., 2018). Ultérieurement, la recherche sur l'analyse des données sur la septicémie a également exploré les études de causalités à l'aide d'approches probabilistes, telles que les *BN* et la prédiction de survie à l'aide du *ML* (Sheetrit et coll., 2019). Par la suite, une nouvelle approche probabiliste avec un objectif interprétable a été utilisée pour établir des classificateurs pour le diagnostic de la septicémie surpassant les classificateurs alternatifs (Yao et coll., 2021). Également, des modèles graphiques probabilistes ont été explorés en tant qu'outil de diagnostic de la septicémie pédiatrique démontrant une précision et une spécificité élevées (Nguyen et coll., 2023). Puis, Valik et coll. (2023) ont démontré

qu'un *ML* pouvait prédire l'apparition d'une septicémie dans les 48 heures en utilisant les rares données des DSE. De plus, leur score était basé sur un réseau causal probabiliste – *SepsisFinder*[37] – qui avait des similitudes avec le raisonnement clinique. Subséquemment, selon Valik et coll. (2023) identifier un risque élevé avec cette méthode pourrait être utilisée pour adapter les interventions cliniques et améliorer les soins aux patients.

Tertio, les modèles *PFL*, comme solutions valorisantes, permettent l'interprétabilité pour la prédiction de la mortalité chez les patients en choc septique et l'explication scientifique à travers la causalité. De la sorte, les *PFL* combinant des descriptions avec les propriétés statistiques des données ont montré une précision comparable à celle d'autres méthodes, telles que les modèles flous de *Takagi-Sugeno* et les modèles de régression logistique (Fialho et coll., 2016). De même, ces derniers modèles détaillent une technique pour anticiper le décès des patients à la suite d'un choc septique. Aussi, ce genre de dispositif allie les bénéfices de la logique floue et des modèles probabilistes pour traiter l'incertitude tant sur le plan du flou que celui probabiliste, ce qui en fait un instrument particulièrement approprié dans les systèmes critiques (ex. de santé) où ces incertitudes se manifestent fréquemment. Ensuite, selon Balentien et coll. (2016), ces modèles de régression ont fourni des estimations interprétables du risque de mortalité, augmentant ainsi la transparence du système apprise à l'aide de règles floues. En utilisant les *PFL,* ces auteurs avaient fait un projet sur les prédictions personnalisées du risque de mortalité chez les patients en choc septique et en réanimation. De plus, les *PFL* donnent des mesures de probabilité fournissant des informations (ex. cliniques) supplémentaires sur lesquelles les experts (ex. médecins) peuvent agir (Morales, 2016). À l'aide de la *PFL*, Mustafayev (2017) avait travaillé ultérieurement sur la prédiction du risque de mortalité chez les patients en choc septique en réanimation. Subséquemment, Razab-Sekh (2021) avait élaboré la modélisation prédictive du sepsis à l'aide de la *PFL* et de la programmation génétique.

[37] Le *SepsisFinder* se déclenché plus tôt et avait un *AUROC* (0,950 contre 0,872), ainsi qu'une *Area under Precision Recall (APR)* (0,189 contre 0,149). Ce *SepsisFinder* a signalé une médiane de 5,5 heures avant l'administration d'antibiotiques. Comparé au *NEWS2 (National Early Warning Score 2)*, qui est une méthode établie pour identifier le sepsis.

5.5 Conclusion

Ce chapitre a exploré l'interconnexion entre les paradigmes IA et la gestion du choc septique, soulignant l'importance significative de la valorisation dans l'exploitation des sources de données sensibles. Nous avons démontré comment les avancées en IA et plus particulièrement les approches connexionnistes offrent des perspectives prometteuses pour améliorer les associations entre les variables et la précision de prédiction de cette pathologie dévastatrice.

En outre, l'étude de cas dédiée au choc septique illustre l'application concrète d'une solution basée sur la *PFL*. Cette approche, par sa capacité à gérer l'incertitude et la variabilité inhérentes aux contextes critiques, révèle son potentiel pour affiner les constats précoces, optimiser les stratégies et, ultimement, améliorer les prédictions. La *PFL* se distingue par sa flexibilité à intégrer des connaissances expertes et des données empiriques permettant ainsi de construire des modèles robustes et interprétables, impératifs dans un domaine aussi critique que l'exemple de la médecine.

Cependant, l'intégration des technologies soulève des questions éthiques fondamentales. L'accès, l'utilisation et la protection des données de contexte sensible exigent une vigilance constante pour garantir la confidentialité, prévenir les biais algorithmiques et assurer la valorisation des décisions prises par les systèmes. L'éthique ne doit pas être une considération secondaire, mais un pilier central guidant le développement et le déploiement de toute solution basée sur l'IA pour les domaines critiques.

En conclusion, la synergie entre les paradigmes de l'IA, une gestion valorisante des données et des outils comme la logique floue probabiliste représente une voie d'avenir pour révolutionner la prise en charge du choc septique. Les défis demeurent, notamment en ce qui concerne la validation à grande échelle et l'intégration clinique de ces solutions. Néanmoins, les bénéfices potentiels en termes de réduction de la mortalité et d'amélioration de la qualité des soins justifient amplement la poursuite de ces recherches interdisciplinaires avec une attention rigoureuse portée aux impératifs valorisants.

CHAPITRE 6 :
MÉTHODOLOGIE

6.1 Introduction

Premièrement, nous illustrons l'exemple d'un jeu de données non réelles lié à des patients souffrant d'une maladie chronique. Les données comportent l'âge, les antécédents médicaux et les résultats d'examens sanguins. Sur le plan informatique, nous organisons ces données sous forme de variables normalisées. Cognitivement, nous évaluons l'importance de ces variables pour le spécialiste médical. La nécessité de cette double approche découle du fait que la machine requiert un format numérique pour effectuer des calculs alors que l'expert a besoin d'une valorisation clinique pour interpréter les résultats. Ce processus garantit que chaque donnée n'est pas simplement traitée, mais également comprise dans son contexte pratique.

Dès lors, notre méthodologie constitue une démarche scientifique spécifique à l'atteinte de nos objectifs cognitifs et informatiques. Nous soulignons un flux de travail distinctif pour notre recherche. Ce chapitre expose, de la sorte, les éléments les plus importants de ce flux, avec les choix théoriques, les méthodes utilisées pour traiter les données, les exemples et les justifications. Notre but est, de ce fait, de développer une solution IA qui tente de représenter le raisonnement et qui se rapproche d'un équilibre en prédiction précise en intégrant le triptyque : en premier par l'interprétabilité, en deuxième par l'enrichissement en analyse causale, pour un troisième cairn en perspective de valorisation : intelligible, robuste, généralisable, transparente, responsable, de confiance, éthique, réglementaire, sécuritaire, auditable, certifiable, frugale et souveraine.

Exemple de parcours de données non réelles

Pour les besoins de compréhension, nous utilisons des données non réelles pour l'illustration de notre démarche. Notre approche est organisée en différentes étapes clés, chacune intégrant un aspect technique et un autre cognitif démontrant une coopération ininterrompue entre la

technologie et l'intellection humaine. Pour saisir cette démarche méthodologique, nous envisageons un ensemble fictif de données d'essai portant sur des patients (Figure 7).

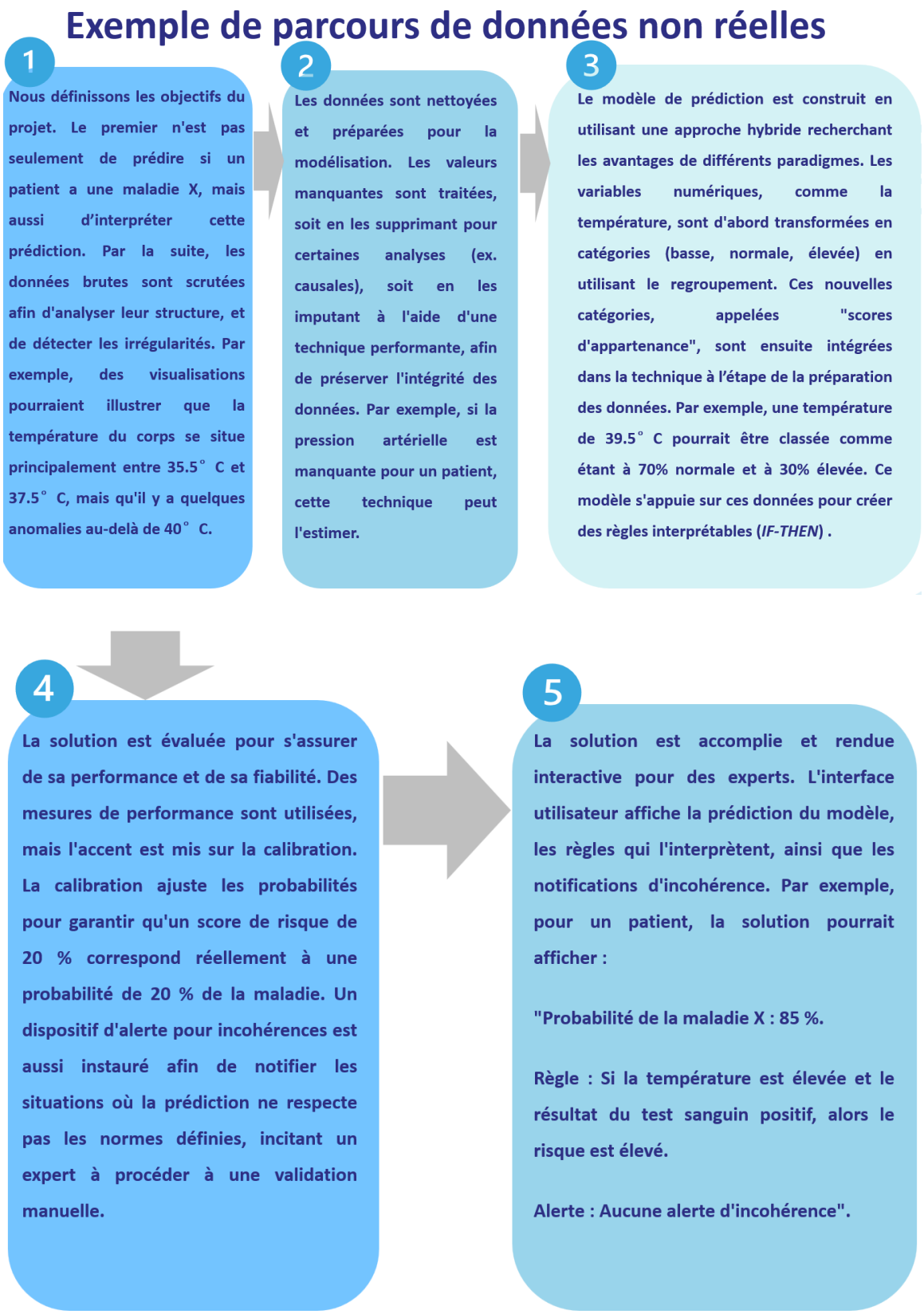

Figure 7 : Exemple de parcours de données non réelles.

Autant, pour atteindre nos objectifs, nous envisageons de développer une stratégie méthodologique, sophistiquée et hybride combinant les avantages de divers paradigmes d'IA. Notre choix méthodologique intègre, de la sorte, la logique floue, les probabilités, les modèles arborescents et la régression logistique. Donc, notre méthodologie est organisée en phases distinctes, de la préparation des données à l'évaluation pour garantir la fiabilité de nos architectures.

En outre, notre flux méthodologique (Figure 8) repose sur une articulation progressive entre les dimensions informatiques et cognitives. Le volet informatique fournit les fondations techniques avec notre stratégie pour notre implémentation, nos algorithmes et nos protocoles d'expérimentation. Le volet cognitif s'intéresse à la manière dont notre flux reflète, dans un contexte sensible, notre triade. Ces deux volets interagissent en permanence parce que l'un ravitaille l'autre. Les choix algorithmiques sont guidés par les attentes et analyses du volet cognitif validant la pertinence du volet informatique. L'association et l'avancement de ces deux volets se manifestent par notre stratégie et par une intégration plus fine des mécanismes de notre triade. Le flux de travail permet ainsi de passer d'une interprétation brute des données à une méthodologie distinctive et customisée où chaque étape et objectif informatique est relié à une fin cognitive, garantissant une cohérence globale.

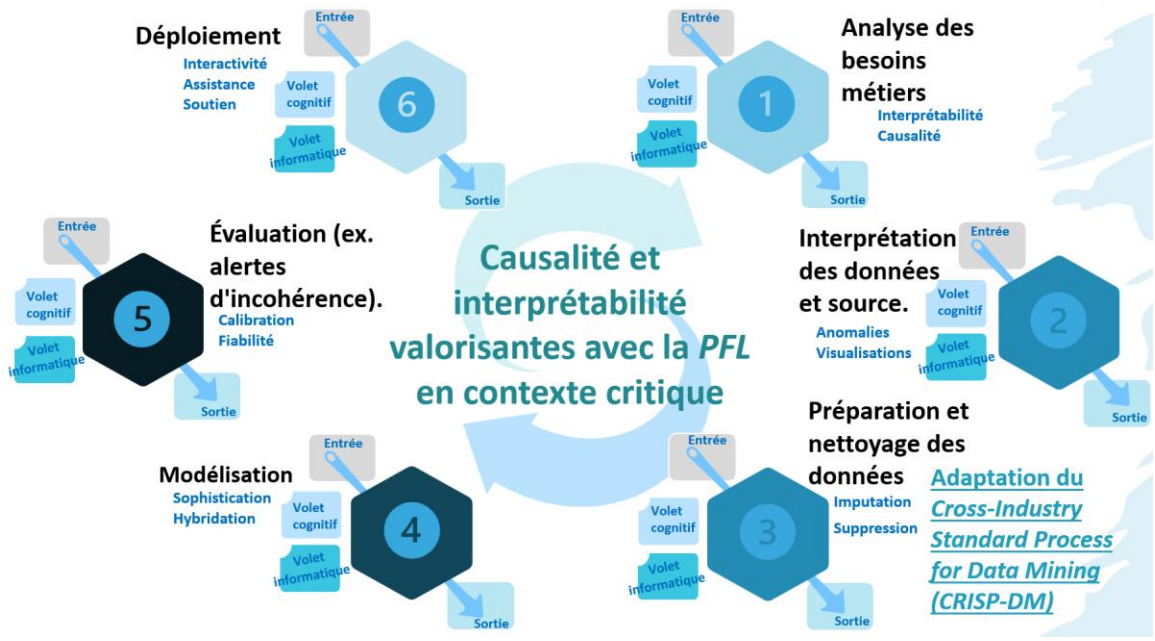

Figure 8 : Flux méthodologique.

6.2 Notre analyse des besoins métiers

Nous illustrons l'analyse des besoins métiers en prenant l'exemple d'un service hospitalier où les médecins doivent anticiper le risque de récidive d'un patient. Informatiquement, nous mettons en place des algorithmes capables de traiter plusieurs variables cliniques. En sachant que l'expert est limité dans le nombre de ces variables à traiter, nous devons fournir une interprétation claire détaillant pourquoi tel patient est considéré comme présentant un risque, et ce dans le cadre cognitif. La raison est claire : les experts ne se basent pas sur des données brutes pour prendre des décisions nécessitant un processus de réflexion organisé. Cela se manifeste par des règles compréhensibles et catégorisables comme « si le taux de sucre dans le sang excède X et l'âge dépasse Y, alors le risque de récidive s'accroît ». Ce genre de représentation offre aux médecins une assurance directe dans le modèle.

Notre méthodologie ad hoc répond aux besoins de notre triade. Dès lors, l'objectif est de développer cette méthodologie en construisant des modèles pour interpréter la prédiction et la causalité depuis l'entrée des données vers la sortie des résultats. Ces modèles sont évolutifs, dynamiques et interactifs. Nous nous inspirons par conséquent du travail de Razab-Sekh (2021) en adaptant le *Cross-Industry Standard Process for Data Mining (CRISP-DM)*.

A fortiori, notre méthodologie est animée par la nécessité de relever le défi des systèmes opaques en IA. C'est pourquoi le but est de concevoir des architectures recherchant l'équilibre de notre triade. De cette façon, prenons l'exemple d'un contexte critique, ce qui implique de donner des résultats concernant le risque lié à une problématique tout en proposant une interprétation facile à comprendre. En procédant ainsi, l'illustration de la valorisation par un élément comme la transparence n'est pas une simple possibilité, mais un devoir d'amélioration pour permettre aux experts de faire confiance à l'IA et d'en assumer la responsabilité.

Tout bien considéré, notre étape d'analyse mobilise une double perspective. Du point de vue informatique, nous définissons les spécifications techniques à respecter, comme la capacité à gérer un certain nombre de variables (caractéristiques) et volume de données sensibles, ou l'exemple de la robustesse face aux valeurs manquantes. Du point de vue cognitif, nous identifions les attentes des experts (spécialistes) à savoir notre triade et pourquoi une prédiction précise s'est-elle produite. Le lien entre ces deux points de vue se traduit par l'intégration de modèles capables de livrer à la fois un résultat chiffré et une tentative de représentation du raisonnement. L'avancement de nos deux volets repose sur la formalisation progressive des besoins en règle techniques et en contraintes interprétatives. Le flux de travail démarre donc par cette analyse, orientant toutes les étapes suivantes. Exemple concret, dans la condition critique, l'exigence est de générer non seulement une probabilité de risque, mais aussi une justification claire à l'intention du spécialiste.

6.3 Notre interprétation des données

Nous utilisons l'exemple de l'étude d'une BD concernant le suivi des patients atteints d'une maladie spécifique pour illustrer comment interpréter des données. Sur le plan informatique, nous produisons des visualisations et des graphiques pour vérifier la répartition d'âge et de pression artérielle. D'un point de vue cognitif, nous vérifions si ces distributions correspondent à des réalités médicales possibles. Pourquoi est-il impératif de franchir cette étape ? En effet, un déséquilibre, tel qu'un manque de jeunes patients, pourrait entraîner l'anormalité du modèle. Comment peut-on prévenir cela ? En utilisant des techniques comme l'*Isolation Forest* détectant

les anomalies, ce qui garantit que l'analyse médicale repose sur un ensemble de données cohérent, sans irrégularités dissimulant de réelles vérités.

De la sorte, cette phase préliminaire a pour objectif d'explorer et de comprendre les données à disposition. Nous faisons appel à des illustrations graphiques, tels que les histogrammes pour examiner la répartition des variables et leur association avec la prédiction. Ainsi, le but est de détecter les attributs susceptibles d'affecter le résultat et de localiser les possibles éléments perturbateurs. Nous étudions aussi la présence de données absentes et d'observations atypiques (*outliers*) en recourant à des techniques telles que l'écart interquartile (*IQR*) et l'*Isolation Forest*. Pour ces éléments, cette étape est majeure pour guider les choix relatifs à la préparation des données et à la modélisation.

Pour finir, nous abordons cette étape sous deux angles. Informatique, impliquant la caractérisation statistique, l'identification des anomalies et la validation des distributions. Cognitive en évaluant les relations entre les variables et leur signification pour l'expert. La corrélation entre ces aspects se manifeste lors du choix des caractéristiques appropriées tant pour le modèle que pour l'interprétation. Les progrès dans les deux aspects impliquent l'amélioration de la description de nos variables et la création d'un ensemble de données intelligible. Ainsi, le processus de travail évolue d'une base initiale vers une base structurée et améliorée. Par exemple, un diagramme représentant la distribution d'une variable est initialement considéré comme une donnée statistique puis réinterprété pour être pertinent aux yeux du spécialiste.

6.3.1 Notre source de données

Nous illustrons l'usage de la source de données par la supposition que nous accédions à une BD hospitalière. En matière d'informatique, nous veillons à ce que les données soient anonymisées et conformes aux standards de sécurité. D'un point de vue cognitif, nous vérifions que le spécialiste médical est d'accord pour gérer ces données, parce qu'elles garantissent la protection des patients. Notre interprétation est double : préserver l'intégrité des patients tout en renforçant la reconnaissance et la confiance des professionnels. Cela signifie qu'une approbation institutionnelle doit être obtenue avant de procéder à l'exploitation des données. Par exemple,

l'accès est accordé par un comité d'éthique assurant que les prédictions générées seront considérées comme dignes de confiance et valides.

Dès lors, la composante informatique se concentre sur l'anonymisation, la standardisation et la conformité des données sensibles. La composante cognitive insiste sur la valorisation que le spécialiste doit avoir dans ces données. L'association des deux est pressante parce qu'un algorithme performant perd sa valeur si les données ne sont pas acceptables d'un point de vue valorisant. Par exemple, l'avancement des deux composantes se traduit par un processus de validation qui va du respect des règles techniques à l'acceptation institutionnelle par les comités d'éthique. Le flux de travail s'organise ici autour de la sécurisation des données et de leur mise à disposition pour les étapes suivantes d'implémentation et d'expérimentations. Exemple, l'autorisation d'accès à une BD sensible garantit que les données utilisées sont fiables, ce qui renforce la valorisation (ex. en confiance) cognitive des experts dans les résultats prédictifs.

6.4 Notre préparation des données

Nous illustrons la préparation des données en prenant pour exemple un fichier où 33 % des valeurs de la tension sont absentes. Techniquement, nous effectuons une imputation en utilisant un modèle arborescent. Sur le plan cognitif, nous nous assurons que cette approche n'engendre pas de valeurs contradictoires, comme des tensions médicalement irréalisables. Pourquoi opter pour un modèle de ce type ? Il préserve les liens entre les variables assurant de cette façon la constance du jeu de données. Comment évaluer l'impact ? En examinant les prédictions générées avant et après la substitution afin de garantir que l'analyse médicale demeure intacte.

C'est pourquoi nous réalisons notre préparation de données (avec du nettoyage, collection et structuration) pour des tests de construction de modèles (sélection des techniques et entraînement) et une évaluation des modèles (performance prédictive, sélections d'algorithmes et exploration des variables) (Okafor et coll., 2023). Implicitement, cette phase de préparation des données est déterminante pour assurer la qualité et l'adéquation du jeu de données destiné à la modélisation. Nous gérons, par exemple, les données manquantes en supprimant les lignes si besoin. Autre exemple, en ce qui concerne nos variables, nous employons un modèle arborescent (ex. *LightGBM*) pour l'imputation des valeurs absentes, garantissant ainsi une bonne précision.

Durant cette étape, la composante informatique implique le nettoyage, l'imputation et le formatage des variables. L'aspect cognitif se rapporte à la certitude que ces modifications ne distordent pas le sens subtil des données. L'équilibre entre la qualité statistique et la fidélité de notre triade en constitue le lien. Le progrès des deux parties se reflète dans l'utilisation de jeux de données de plus en plus perfectionnés validés tant sur le plan technique que cognitif. Le processus de travail progresse en convertissant graduellement des données non traitées en un ensemble organisé et compréhensible. Par exemple, le recours à un logiciel d'estimation de valeurs manquantes (ex. le modèle arborescent) n'est pas choisi uniquement pour sa précision, mais également parce qu'il permet des résultats explicatifs.

6.4.1 Notre nettoyage des données

Nous illustrons le processus de nettoyage des données en prenant l'exemple d'une variable présentant 77 % de valeurs manquantes. Techniquement, nous choisissons de l'écarter, car cela nuirait à la fiabilité des statistiques. Sur le plan cognitif, nous établissons que cette variable ne présente plus d'intérêt pour l'expert, étant donné qu'elle est trop fragmentée. La raison de ce filtrage est la quête de clarté. L'idée est de consigner ce choix dans un protocole afin que l'expert puisse comprendre la raison pour laquelle une donnée a été éliminée. Par exemple, l'exclusion d'une variable rare et peu quantifiée. Cette suppression rend l'ensemble plus robuste et interprétable.

De la sorte, afin de préserver la fiabilité des données, nous vérifions initialement l’insuffisance de données. De plus, pour gérer les valeurs absentes, nous choisissons d'adopter des stratégies combinées. Par exemple, pour la validation croisée, l'attribution d’une variable est réalisée au moyen du modèle arborescent, une décision soutenue par sa grande précision. Également, pour l’analyse des variables confondantes, les lignes avec des valeurs absentes sont éliminées, ce qui diminue la taille de l'ensemble des données tout en garantissant la qualité des observations.

En dernier lieu, l'aspect informatique repose sur des techniques systématiques pour gérer les valeurs manquantes et rectifier les incohérences. L'aspect cognitif examine l'effet de ces décisions sur l'interprétation de l'expert. Par exemple, la combinaison de ces deux aspects est manifeste dans la décision d'éliminer certaines lignes, parce que cela permet d'éviter des interprétations

erronées. Nous pouvons évaluer la progression des deux aspects, puisqu'à chaque phase, l'ensemble de données gagne en cohérence et en clarté. Le processus de travail respecte une logique où chaque choix technique est associé à une explication cognitive. Un autre exemple, l'exclusion d'une variable est justifiée non seulement sur le plan technique par une insuffisance de données, mais aussi sur le plan cognitif en raison du fait qu'un expert ne serait pas en mesure d'interpréter une variable trop morcelée.

6.5 Notre modélisation

Nous illustrons la modélisation en prenant l'exemple de la construction d'un modèle prédictif pour évaluer le risque cardiaque. Sur le plan informatique, nous utilisons un regroupement flou pour convertir des valeurs numériques, telles que la pression artérielle en trois classes distinctes. Sur le plan cognitif, nous démontrons que ce langage reflète une tentative de représentation du processus de raisonnement d'un médecin. Pourquoi avoir fait cette sélection ? Un expert médical ne voit pas une tension à 151 mmHg (millimètres de mercure) comme une valeur unique, mais plutôt comme un indicateur de risque. Le processus est guidé par les règles « *IF-THEN* », telles que « si tension artérielle haute et cholestérol élevé, alors risque cardiaque important ». Cette affiliation assure à la fois la précision et la clarté du modèle.

Dès lors, nos modèles sont construits sur des architectures alliant la puissance prédictive des modèles basés sur l'arborescent à l'utilisation des règles et symboles de la *PFL*. L'étape de modélisation est celle où notre stratégie méthodologique commence à se concrétiser. Nous employons le regroupement (clustering) pour la fuzzification des variables, ce qui facilite la transformation de valeurs numériques en niveaux d'appartenance à différentes catégories. Ces indices d'appartenance sont par la suite incorporés comme attributs additionnels dans le modèle arborescent. Cette approche combine aussi le potentiel prédictif du modèle arborescent avec la compétence d'interprétation inhérente de la logique floue fournissant une forme de tentative de représentation du raisonnement plus alignée à la cognition.

Autant, les modèles de la première architecture se concentrent sur l'inférence causale en utilisant une pondération basée sur l'appartenance floue. Également, nous y associons un calcul manuel de l'*ATE* à une segmentation probabiliste par le biais d'un regroupement bayésien. En procédant

ainsi, les flux débutent par l'extraction des variables sensibles, la détermination de l'issue observée et de l'impact de chaque variable. Donc, les variables confondantes sont évaluées grâce à des régressions pondérées par des niveaux d'appartenance floue dérivés de ce modèle de regroupement. Chaque variable est rendue floue sur trois segments ordonnés. Aussi, cet *ATE* est déterminé pour chaque segment, puis ajusté par un coefficient de correction basé sur la valeur manuelle. Dès lors, les résultats comprennent les effets corrigés moyens et la validation interne de ces effets.

En outre, les modèles de la deuxième architecture utilisent une structure modulaire pour la prédiction incluant l'interprétabilité et la calibration. Commençant, de ce fait, avec un module de configuration centralisée, puis passant à un gestionnaire de données élaborant les ensembles d'entraînement en normalisant et indexant les variables. Ensuite, le modèle *PFL*, se chargeant de manière dynamique, réalise un regroupement flou et produit des résultats sur différents niveaux. Les processus comprennent, également, la pondération des caractéristiques interprétables, l'entraînement du modèle arborescent, la validation croisée, la calibration isotonique et l'élaboration de métriques approfondies. De la sorte, l'analyse est enrichie par une catégorisation du risque par regroupement incluant des alertes en cas de discordance entre les prédictions et les règles établies.

In fine, cette étape constitue le cœur de l'intégration informatique et cognitive. Informatiquement, avec l'assemblage des architectures de notre stratégie associant modèles arborescents, logique floue et régression. Cognitivement, avec l'élaboration de règles et d'indicateurs interprétables. Le lien se traduit dans la création de règles « *IF-THEN* » reprenant le langage du spécialiste. L'évolution des deux aspects implique une progression de la simple prédiction précise et chiffrée vers un processus tentant de reproduire le raisonnement. Le flux de travail structure le passage des données transformées vers un modèle opérationnel respectant notre triade. Exemple, la fuzzification transforme des valeurs numériques en catégories proches des jugements comme « faible », « modéré » ou « élevé », ce qui rapproche la machine du langage de l'expert.

6.6 Notre évaluation

Nous illustrons l'évaluation en utilisant l'exemple d'un modèle prédictif testé sur trois échantillons : apprentissage, validation et test. Techniquement, nous évaluons la performance en utilisant les mesures de précision, de rappel et de score F1. Sur le plan cognitif, nous vérifions si le modèle offre une probabilité qui concorde avec l'appréciation d'un médecin. Pourquoi est-il nécessaire de procéder à une calibration ? Car un score brut de 0.7 devrait en réalité représenter 70 % de risque. Comment l'acquérir ? Par régression isotonique, on ajuste les probabilités. Par exemple, un modèle incorrectement ajusté indique 0.7 pour un patient alors que le risque réel observé ne s'élève qu'à 30 %. Ce retard diminue la valorisation et la confiance. Une calibration rectifie ce biais.

De la sorte, afin d'examiner le rapprochement vers notre équilibre lors des phases clés de conception, nous effectuons une exploration de données et choisie, puis combinée des variables interprétatives. De plus, nous élaborons des modèles avec divers paramétrages et nous procédons à l'évaluation de ces derniers (Denis et Varenne, 2019). Également, pour la stratégie d'évaluation, et afin de minimiser les biais et d'accroître la capacité de généralisation des modèles, des approches d'estimation sont mises en œuvre où dans certains modèles, nous utilisons une validation croisée en plusieurs phases (Musanga et coll., 2025). Ainsi, l'ensemble de données sera divisé en plusieurs sous-ensembles, dont certains sont utilisés pour l'apprentissage, la validation et le test à chaque itération. De même, nous effectuons des évaluations avec des mesures de performance (ex. exactitude, précision, rappel, score F1 et *ROC- AUC*), ainsi que d'autres mesures, comme la moyenne ± écart-type sur plusieurs plis. Autant, l'accent est mis sur l'ajustement des probabilités afin d'assurer que les scores du modèle correspondent aux risques véritables. Pareillement, nous utilisons l'erreur de calibration attendue (*Expected Calibration Error : ECE*) pour évaluer la fiabilité avant et après la calibration. Nous évaluons aussi la capacité d'interprétation grâce à des diagnostics et un système d'alerte d'incohérence. Subséquemment, cette étape assure la robustesse, la fiabilité et la cohérence de nos modèles.

Autant, dans notre vision informatique, l'évaluation sert à mesurer la performance, la calibration et la robustesse. Du point de vue cognitif, elle évalue la lisibilité et la valorisation (par exemple, via la confiance). Le lien est visible dans certains indicateurs (comme l'*ECE*) reflétant la relation

entre le score technique et l'appréciation humaine du risque. Le progrès des deux aspects se manifeste par une capacité de plus en plus grande à générer des prédictions sûres et respectables. Chaque phase de validation technique est associée à une approbation cognitive dans le déroulement du processus de travail.

Calibration des probabilités

Il est primordial d'être fiable. De ce fait, les scores bruts du modèle arborescent sont ajustés par la régression isotonique, afin de garantir qu'un score de 0.2 correspond véritablement à une probabilité de 20 %. De la sorte, cette approche accroît la crédibilité du modèle.

En outre, la calibration fait le lien entre l'informatique et la cognition. Elle garantit que les résultats numériques correspondent à la perception humaine du risque. Le progrès des deux aspects repose sur l'amélioration de la précision des probabilités, augmentant ainsi la fiabilité du modèle. Le processus de travail assure la conversion d'un score brut en une mesure précise pour l'élaboration des décisions.

Alertes d'incohérence

Un système d'alerte est inclus pour indiquer les divergences significatives (ex. dépassant un seuil de 15%) entre les prédictions et l'interprétation des règles. Dès lors, ce mécanisme de confiance encourage l'expert à effectuer une validation manuelle.

Également, ces alertes sont une interface directe entre les deux dimensions. Informatiquement, elles sont détectées par ces seuils de divergence. Cognitivement, elles renforcent la vigilance de l’expert. L’avancement des deux volets consiste à faire les liens entre des prédictions précises et leurs signaux interprétables. Le flux de travail rend ainsi possible une interaction humaine avec le modèle.

Évaluation de la causalité

Notre méthodologie tente de s'aligner sur une représentation du raisonnement causal. De ce fait, nous estimons l'*ATE* en appliquant un poids basé sur le degré d'appartenance flou pour chaque segment d'une variable. Autant, l'évaluation standardisée de l'effet causal est fournie par le calcul de l'effet de traitement moyen flou normalisé (*Normalized Fuzzy Average Treatment Effect : NFATE*). De la sorte, cette méthode cherche à différencier la causalité des simples corrélations.

Tout bien considéré, le volet informatique se rapporte aux méthodes de calcul statistiques et floues utilisées dans cet *ATE*. L'aspect cognitif se rapporte à la certitude que ces effets sont interprétables et utilisables. La distinction entre causalité et simple corrélation est au cœur de l'objectif. On évalue le progrès des deux aspects en fonction de notre capacité grandissante à délivrer non seulement un résultat, mais aussi son analyse causale. Le processus de travail assure que chaque estimation causale est associée à une explication précise.

6.7 Notre déploiement

Nous illustrons le déploiement en prenant pour exemple l'application de la solution dans un département hospitalier. Au niveau informatique, un outil présente les prédictions ainsi que les règles qui y sont liées. Sur le plan cognitif, un médecin met en parallèle ces résultats avec son expérience. Pourquoi cela a-t-il de l'importance ? Parce qu'un expert ne délègue pas sa prise de décision à une machine sans préalablement examiner le processus de réflexion qui la sous-tend. Comment simplifier ce processus ? En intégrant un système d'alerte qui indique les divergences entre les prédictions et les règles. Par exemple, lorsque le modèle signale un risque minime, mais que plusieurs règles activées indiquent une forte probabilité de risque, une alerte incite le médecin à prendre des mesures.

Ainsi, l'objectif de cette illustration est de saisir la mise en place de notre stratégie de solution *PFL*. C'est de la rendre interactive et collaborative pour les spécialistes. L'idée est de concevoir une solution ne se limitant pas à faire des prédictions précises, mais qui sert également d'outil d'interprétation et de causalité pour aider à la prise de décisions. Par conséquent, nous envisageons l'élaboration d'algorithmes exposant les prédictions, les règles « *IF-THEN* » associées et les notifications d'incohérence de façon claire. Donc, cela offre aux praticiens la possibilité de corroborer les résultats du modèle avec leur propre savoir-faire, ce qui est primordial pour une intégration sûre et valorisante (ex. éthique) dans des situations critiques.

En définitive, notre angle informatique construit une plateforme avec des algorithmes commentés et des règles automatiques. La cognition évalue la pertinence de ces règles pour la décision réelle. L'association des deux angles se manifeste dans la conception d'un outil renforçant les capacités du spécialiste plutôt que de le remplacer. Leur avancement repose sur la

transformation future d'implémentation et de prototypes expérimentaux en un outil validé pour un contexte critique. Le flux de travail s'achève par une intégration pratique. Exemple, le déploiement de résultats affichant simultanément les prédictions chiffrées et les règles « *IF-THEN* » offrant à un expert la possibilité de comparer les résultats de l'outil informatique à son propre processus cognitif de raisonnement.

6.8 Un exemple d'application pour un expert

Nous fournissons une illustration d'application pour un expert en présentant le cas d'un spécialiste en oncologie. Le modèle prévoit le risque de récidive chez un patient soigné pour un cancer. Sur le plan informatique, il produit une directive : « si âge > 60 et présence de mutation génétique, alors risque accru ». Sur le plan cognitif, l'expert confirme cette règle en la confrontant à son expérience en clinique. Pourquoi est-ce déterminant ? Un médecin doit s'assurer que la prédiction se conforme à un raisonnement médical reconnu. Comment peut-on rendre cette validation faisable ? Par la création d'un ensemble restreint de règles élémentaires reflétant la logique du modèle. Par exemple, quatre règles couvrant 79 % des situations, au lieu d'un ensemble trop étendu.

De plus, l'un des buts primordiaux est de proposer une IA servant de véritable collaborateur dans le processus de raisonnement de l'expert. Dès lors, elle tente de reproduire ce processus cognitif d'abord. Nos modèles produisent également des règles interprétables à partir des centroïdes des groupes flous et des probabilités conditionnelles. Par exemple, le système pourrait générer soit quatre ou sept règles en fonction du nombre de regroupements ou une règle globale constituée de sept caractéristiques clés, voire l'ensemble complet de ces caractéristiques. De plus, ces règles sont dénormalisées afin d'être plus pertinentes pour un expert. En conséquence, cette méthode donne à ce spécialiste la possibilité de comprendre le flux derrière une prédiction, ce qui est concluant pour assurer la valorisation (en confiance et en élaboration de décisions éthiques).

Également, l'IA ne se limite pas à fournir une prédiction exacte, elle la soutient aussi avec une analyse causale et une interprétation que l'expert peut confirmer avec son expertise personnelle. Dès lors, l'évaluation de l'IA en prenant en compte la valorisation, à l'exemple de la transparence, n'est pas une question de choix, mais une nécessité impérative.

Pour clore, cette étape illustre l'interférence évidente entre les deux envergures. Informatiquement, les modèles produisent des règles et des probabilités. Cognitivement, le spécialiste évalue la pertinence de ces règles pour la prise de décision. L'association de ces deux envergures se manifeste lorsque l'expert comprend immédiatement la logique derrière une prédiction. Leur avancement se traduit par une meilleure acceptation des modèles par des experts. Le flux de travail est ici validé par l'expérience utilisateur. Exemple, un spécialiste recevant une prédiction accompagnée de règles simples pour un contexte critique est en mesure de juger si ces règles correspondent à sa propre pratique.

6.9 Nos justifications des méthodes retenues

Nous mettons en lumière les raisons des approches choisies en prenant l'illustration de l'utilisation de la logique floue afin de gérer l'incertitude. Techniquement, cette approche transforme des données en catégories successives. Sur le plan cognitif, elle tente de reproduire la façon dont les spécialistes énoncent leurs évaluations. Pourquoi ne pas conserver uniquement les données brutes ? Parce que le jugement médical ne repose pas sur une seule nuance de données, mais sur une évaluation comparative. Comment cela se manifeste-t-il ? En utilisant les termes médicaux pour catégoriser la pression artérielle comme « normale », « pré-hypertension » ou « hypertension ». Ainsi, l'ajustement est nécessaire pour faire correspondre les probabilités techniques avec les perceptions humaines. Par exemple, une probabilité ajustée assure que la machine et le médecin communiquent dans le même langage du risque.

Donc, à travers cette illustration, les décisions méthodologiques relatives aux composants clés de notre flux de travail sont motivées par l'objectif de développer des modèles crédibles en mettant en avant la valorisation (ex. éthique).

Pourquoi et comment choisir les données ?

Le choix de la BD est motivé par son accessibilité et la richesse de ses données. Cela permet de reproduire des résultats intéressants dans un contexte critique. De plus, son utilisation à des fins éducatives et de développement d'algorithmes pour l'IA doit être déjà établie. Ce qui valide son emploi pour nos expérimentations. Autant, sa nature anonymisée, ce qui respecte les considérations éthiques.

Pourquoi nos choix de gestion de données ?

Notre stratégie combinée pour la gestion des données manquantes est une démarche rigoureuse. Aussi, la suppression des lignes pour les variables confondantes permet une certaine qualité des analyses interprétables et causales. Pareillement, l'imputation avec un modèle performant pour nos variables préserve l'intégrité de l'ensemble des données.

Pourquoi le regroupement pour la fuzzification ?

La logique floue offre la possibilité de représenter l'incertitude et le caractère subjectif des données. Dès lors, en employant le regroupement pour la fuzzification, nous essayons de reproduire la logique humaine s'appuyant sur des catégorisations non binaires dans le processus de décision. En conséquence, nous optons pour ce regroupement, parce que celui-ci déduit les groupes directement à partir des données sans avoir besoin de règles prédéterminées manuellement.

Pourquoi un modèle arborescent ?

L'incorporation des scores d'appartenance flous dans le modèle arborescent constitue une approche pour pallier le dilemme entre précisions prédictives et l'analyse de notre triade. De ce fait, ceci confère au modèle la puissance prédictive des modèles arborescents tout en préservant une clarté inhérente au symbolisme à travers ces règles et ces symboles.

Pourquoi la calibration et les alertes d'incohérence ?

La précision est décisive pour la confiance. Subséquemment, des probabilités peu fiables peuvent tromper un expert. Les alertes d'incohérence constituent un dispositif de sécurité destiné à signaler lorsque le modèle et ses règles ne sont pas en accord incitant ainsi à une intervention humaine.

En somme, les choix techniques ne sont pas isolés ; ils répondent directement à des contraintes cognitives. Le choix d'une BD dans notre contexte illustre l'importance de la valorisation (ex. confiance) dans la source de données. La gestion des données manquantes traduit la nécessité d'avoir des résultats interprétables. Le recours à la fuzzification relie les valeurs numériques au langage naturel des experts. Exemple, l'utilisation d'un modèle arborescent équilibre, performance et lisibilité. Enfin, la calibration et les alertes d'incohérence montrent que chaque

décision technique est motivée par la recherche de cette valorisation cognitive. L'avancement des deux volets réside dans la consolidation progressive d'une méthodologie qui ne sépare jamais ces deux dimensions. Le flux de travail s'en trouve renforcé parce qu'il relie chaque décision technique à un objectif d'interprétation.

6.10 Notre synthèse de la méthodologie

6.10.1 Notre analyse des besoins métiers

Cette phase initiale ancre notre méthodologie dans une approche centrée sur l'assistance d'un expert.

- **Volet cognitif :** Nous identifions un besoin fondamental de l'utilisateur : la valorisation. Aussi, l'expert ne cherche pas une simple prédiction précise, il veut comprendre le processus derrière celle-ci pour une prise de décision éclairée. Ce volet cognitif établit l'exigence d'une analyse causale et d'une interprétabilité intrinsèque et non d'une explication après les résultats prédictifs.
- **Volet informatique :** Nous transformons ce besoin cognitif en spécifications techniques. Les architectures informatiques doivent non seulement offrir une meilleure précision, mais également produire des résultats pouvant être interprétés en langage humain, ce qui nous oriente vers des modèles alliant la force prédictive à l'interprétabilité, tels que les modèles basés sur les modèles arborescents associés à la logique floue.
- **Liens des deux volets :** Les exigences éthiques et la recherche d'une dimension cognitive de valorisation orientent l'élaboration des modèles selon notre stratégie pour le volet informatique. Cette collaboration assure que le produit final respecte des normes appréciables tout en démontrant une bonne performance technique.
- **Avancement des deux volets :** La prise de conscience que l'IA doit agir en tant que partenaire plutôt qu'une simple machine de prédiction, favorise l'évolution du design architectural. C'est une augmentation de la pensée influant sur l'avancement technique.
- **Flux du travail :** Cette phase sert de guide synchronisant toutes les étapes suivantes avec l'objectif central.
- **Modules :** Aucun module spécifique n'est en cours d'utilisation pour l'instant, mais cette phase détermine la structure des modules à venir.

- **Entrée :** Les exigences non organisées des spécialistes et les enjeux de l'IA opaque.
- **Sortie :** Une énonciation précise des buts de notre démarche spécifique : exactitude, lisibilité et causalité.

6.10.2 Notre interprétation des données

Cette phase primordiale prépare le terrain pour l'ensemble du processus.

- **Volet cognitif :** Nous élaborons une interprétation intuitive des données. Nous analysons les distributions, examinons les liens entre les variables et identifions les irrégularités. Cette étape est primordiale pour la construction d'une tentative d'interprétation des données nécessaires pour les analyses à venir.
- **Volet informatique :** Nous utilisons des instruments d'analyse statistique et de représentation visuelle (ex. des histogrammes) facilitant l'identification des attributs primordiaux et des problématiques éventuelles dans les données. C'est une démarche technique d'exploration visant à préparer les données pour qu'elles puissent être traitées par des algorithmes.
- **Liens des deux volets :** L'analyse informatique est capable d'identifier des motifs que l'esprit humain peut percevoir de manière cognitive. L'application de ces algorithmes sur un plan informatique peut être influencée par les intuitions cognitives concernant les données, comme le choix préalable de variables appropriées.
- **Avancement des deux volets :** Une connaissance plus approfondie des données mène à de meilleures décisions en matière de prétraitement informatique, facilitant ainsi la production d'interprétations plus précises et plus fiables pour l'aspect cognitif.
- **Flux du travail :** Cette phase constitue une préparation analytique. Elle assure la qualité des intrants lors de la phase de modélisation, élément primordial pour la crédibilité de nos architectures.
- **Modules :** Nous faisons appel à des bibliothèques de programmation et des modules d'analyse statistique.
- **Entrée :** Données brutes réelles du contexte sensible.
- **Sortie :** Un rapport d'analyse exploratoire et une identification initiale des variables clés et des anomalies.

6.10.3 Notre préparation des données

Cette étape garantit la fiabilité et l'excellence des données en vue des phases subséquentes.

- **Aspect cognitif :** Nous préparons les données de telle manière qu'elles soient exemptes de biais et puissent être analysées de façon cohérente. Par exemple, la gestion des valeurs manquantes doit assurer que les inférences finales soient correctes et non altérées par des données partielles.
- **Volet informatique :** Nous appliquons des techniques de nettoyage, de normalisation et d'imputation. Par exemple, pour une variable importante, nous utilisons un modèle performant comme un modèle arborescent pour l'imputation. Cela se justifie par la recherche de la précision et la préservation des relations complexes dans les données. Nous privilégions l'élimination de certaines données pour assurer la rigueur de l'analyse causale.
- **Liens des deux volets :** La rigueur dans le nettoyage informatique des données se reflète par une plus grande fiabilité dans les analyses cognitives. Un modèle qui est entraîné sur des données de haute qualité a plus de chances de produire des règles d'interprétation valables et exactes.
- **Avancement des deux volets :** L'emploi de méthodes sophistiquées, telles que l'imputation par un modèle arborescent témoigne d'une progression consolidant la crédibilité des futures analyses cognitives.
- **Flux du travail :** Cette étape consiste à passer de l'interprétation à la construction. Elle permet une utilisation directe des données par les algorithmes de modélisation.
- **Modules :** Des fonctions de nettoyage, des scripts pour l'imputation et des bibliothèques destinées à la manipulation des données.
- **Entrée :** Données traitées à partir de l'étape précédente.
- **Sortie :** Un jeu de données prêt à l'emploi, bien structuré et complet pour les variables clés.

6.10.4 Notre modélisation

Les éléments suivants sont au cœur de notre méthodologie spécifique.

- **Aspect cognitif :** Le regroupement flou permet de convertir une valeur numérique d'une variable en une classification dans des catégories cognitives (inférieure, normale, supérieure). Ceci simule un processus de tentative de représentation du raisonnement, où nous n'effectuons pas de catégorisation binaire des éléments.
- **Aspect informatique :** Nous employons des techniques telles que le flou et un modèle arborescent pour effectuer des prédictions. Les indices d'appartenance floue sont intégrés en tant que variables additionnelles. Cela autorise le modèle informatique à intégrer l'essai de reproduction du raisonnement cognitif dans ses opérations de calcul.
- **Liens des deux volets :** L'assemblage informatique produit les centroïdes des groupes flous en guise de résultat cognitif qui sont par la suite exploités par le modèle arborescent pour établir les règles « *IF-THEN* » en tant que seconde sortie cognitive. Ces règles constituent le point d'intersection des deux volets.
- **Avancement des deux volets :** L'aspect informatique progresse en incorporant une dimension sémantique. Le côté cognitif progresse en se basant sur des calculs exacts pour produire des règles précises et objectives.
- **Flux de travail :** Cette phase représente le processus de création. C'est ici que les architectures singulières sont élaborées, fusionnant les diverses méthodes pour constituer les solutions retenues.
- **Modules :** Regroupement de modules du flou, des probabilités, du modèle arborescent et de la régression logistique.
- **Entrée :** Données préparées et épurées.
- **Sortie :** Modèles entraînés, directives d'interprétation et estimations de l'*ATE*.

6.10.5 Notre évaluation

Nous vérifions que les solutions sont à la fois crédibles et praticables.

- **Volet cognitif :** Nous évaluons l'utilisabilité et la fiabilité de la solution pour l'expert. Les alertes d'incohérence sont nécessaires. Elles permettent à l'expert de valider

manuellement un résultat lorsqu'il y a un désaccord notable entre la prédiction et la règle. C'est la validation finale par un expert.

- **Volet informatique :** Nous utilisons des métriques de performance standards et surtout la calibration isotonique et l'*ECE*. Cet étalonnage garantit que les probabilités modélisées reflètent fidèlement les risques réels assurant la fiabilité technique.
- **Liens des deux volets :** La robustesse des probabilités du volet technologique renforce la crédibilité des règles du volet cognitif. Les alertes d'incohérence, de leur côté, instaurent un échange entre les deux, offrant à un expert la possibilité de reprendre la main.
- **Avancement des deux volets :** L'emploi de l'*ECE* et la calibration démontrent une progression en termes de rigueur informatique. L'élaboration d'alertes d'incohérence et de l'indice *ATE* pour établir la causalité constitue un progrès dans notre approche visant à rendre la solution compréhensible du point de vue cognitif.
- **Flux du travail :** Cette phase constitue une boucle de validation. Elle garantit que les modèles sont adaptés à une utilisation concrète réduisant ainsi les risques de biais et d'erreurs.
- **Modules :** Calibration (régression isotonique), mesures d'évaluation et système d'alerte.
- **Entrée :** Modèles entraînés et données de test.
- **Sortie :** Rapports de performance, modèle calibré et un système d'alerte fonctionnel.

6.10.6 Notre déploiement

C'est l'aboutissement de toute la méthodologie où le travail devient un outil pratique.

- **Volet cognitif :** Nous créons une interface utilisateur qui est interactive et intuitive. Un expert peut interagir avec les résultats. Il voit les prédictions, les règles qui les soutiennent et les alertes d'incohérence de manière claire.
- **Volet informatique :** Les algorithmes sont mis pour générer les prédictions en temps réel. La solution est prête à être intégrée dans un flux de travail critique.
- **Liens des deux volets :** Le déploiement est la matérialisation de l'intégration. Le calcul informatique complexe est présenté sous une forme cognitive simple et utilisable, ce qui permet à un expert de s'appuyer sur la solution sans avoir besoin d'une expertise très poussée en IA.

- **Avancement des deux volets :** La solution est rendue opérante. Elle passe du stade théorique à un outil pratique assistant la prise de décision.
- **Flux du travail :** C'est la dernière étape. Elle met la solution à la disposition des utilisateurs finaux et marque le succès du projet.
- **Modules :** Algorithmes de génération d'analyses causales, de règles et d'alertes, ainsi que des interfaces de visualisation.
- **Entrée :** Modèles examinés et prêts à l'utilisation.
- **Sortie :** Des solutions IA déployées, interactives et collaboratives, assistant un expert dans sa prise de décision.

6.10.7 Le tableau de synthèse de notre méthodologie

Étape	Volet cognitif	Volet informatique	Liens des deux volets	Avancement des deux volets	Flux du travail	Modules	Entrée	Sortie
Analyse des besoins métiers	Analyse du besoin de la triade d'interprétabilité, de causalité pour une valorisation, alignement sur la prise de décision des experts.	Définition des exigences techniques pour la précision prédictive et l'analyse causale.	Les besoins cognitifs de valorisation orientent la conception informatique des modèles interprétables et causaux.	L'interprétation des besoins cognitifs fait avancer le design des architectures informatiques.	La première étape : Guider l'ensemble du projet en définissant les objectifs.	s.o.	Besoins métiers de la triade.	Définition des objectifs de la stratégie sophistiquée et hybride.
Interprétation des données	Appréhension des relations entre variables et de leur impact potentiel sur le résultat, et	Utilisation de techniques d'analyse exploratoire pour résumer et visualiser les données.	Les observations cognitives chiffrées guident les choix informatiques de prétraitement.	L'exploration informatique visuelle et statistique enrichit l'interprétation cognitive des données.	Seconde étape : Base de l'interprétation de la problématique.	Histogrammes , Isolation Forest, IQR.	Données brutes.	Statistiques descriptives, visualisations, et identification des données problématiques.

	détection des anomalies.							
Préparation des données	Mise en forme des données.	Application de techniques de nettoyage, de gestion des données manquantes et de structuration.	Un jeu de données propre pour le volet informatique garantit la qualité cognitive des interprétations causales et des règles.	La préparation informatique des données assure la validité des futures interprétations cognitive.	Troisième étape : Rends les données utilisables pour la modélisation.	Modèle arborescent pour l'imputation, filtres pour l'élimination de variables et de lignes.	Données brutes et analysées.	Données propres, structurées et prêtes pour la modélisation.
Modélisation	Tentative de reproduction du raisonnement via la logique floue et des règles.	Utilisation de modèles arborescents pour la prédiction et de la régression pour la causalité.	Le regroupement flou transforme les variables informatiques en catégories cognitives. Les scores d'appartenance floue sont ensuite intégrés au modèle arborescent.	La fusion de l'approche informatique et cognitive permet d'avancer vers un équilibre entre précision (informatique) et interprétabilité (cognitive).	Quatrième étape : Cœur du projet par la création de nos architectures.	Regroupement flou, modèle arborescent, régression logistique.	Données préparées et structurées .	Architectures de modèles pour la triade.

Évaluation	Validation de la fiabilité et de la clarté des règles et des interprétations .	Utilisation de métriques de performance et de calibration pour évaluer la précision et la robustesse des prédictions.	Les alertes d'incohérence signalent une divergence entre la prédiction informatique et la règle d'interprétation cognitive. La calibration assure que les probabilités du modèle informatique sont fiables pour la prise de décision cognitive de l'expert.	Les diagnostics informatiques valident et certifient la fiabilité de la solution cognitive pour l'expert.	Cinquième étape : Valider les choix méthodologiques et la performance.	Métriques, validation croisée, calibration isotonique.	Résultats des modèles.	Modèles calibrés et validés, rapports de performance et d'interprétabilité .
Déploiement	Création d'une interface interactive et collaborative.	Mise en place des algorithmes pour générer les prédictions,	Le déploiement rend les résultats informatiques (prédiction, causalité) accessibles et interprétables via	Les solutions sont rendues cognitivement opérationnelles et utiles pour un expert grâce à la	Sixième étape : Livrer les solutions collaboratives.	Algorithmes de génération de règles et d'alertes, avec interface	Modèles finaux et calibrés.	Solutions interactives, outil d'aide à la décision.

		les règles et les alertes.	un format cognitif (règles, alertes).	robustesse informatique.		utilisateur conviviale.		

Tableau 3 : Synthèse de la méthodologie

6.11 Conclusion

Globalement, nous illustrons la méthodologie à travers l'exemple d'une chaîne unifiée qui fait le lien entre l'informatique et la cognition. Par exemple, dans le contexte d'une application hospitalière, chaque donnée recueillie est scrutée, épurée, modélisée, évaluée et ensuite mise en œuvre au sein d'un environnement maîtrisé. Sur le plan informatique, cela garantit précision, robustesse et performance. Ensuite, cela assure un discernement clair, une valorisation et une adoption par les experts du point de vue cognitif. Pourquoi ce flux présente-t-il une certaine cohérence ? Parce que chaque décision technique est liée à une explication humaine. Comment cela se présente-t-il ? Grâce à des normes, des ajustements et des validations institutionnelles, cela garantit que le modèle ne se limite pas à être un instrument de calcul, mais qu'il constitue aussi un authentique allié décisionnel.

Autant, nous expliciterons notre conclusion à travers d'une autre illustration d'une situation clinique où dans l'exemple d'une unité hospitalière cherchant à évaluer le risque de complications après une opération. En termes de traitement de données, le modèle arborescent génère une prédiction précise, améliorée par des scores flous dérivés du regroupement probabiliste. Sur le plan cognitif, l'expert reçoit une directive explicite en retour, tel que « si l'âge est supérieur à 65 ans et que la pression artérielle est haute alors risque élevé ». Pourquoi cette combinaison est-elle déterminante ? Parce qu'elle facilite la conversion d'une sortie numérique en une directive directement applicable. Comment cette interprétation est-elle garantie ? En intégrant directement les règles floues au sein du modèle prédictif, chaque résultat devient intelligible.

De plus, nous intégrons une illustration en utilisant le processus de calibration. En termes informatiques, l'exemple de la régression isotonique modifie la probabilité de risque générée par le modèle. Sur le plan cognitif, le médecin obtient une valeur correspondant réellement à la fréquence observée. Pourquoi ce réglage est-il indispensable ? Car un expert base ses choix sur des probabilités fiables. Comment ce contrôle est-il intensifié ? Par un système d'alerte détectant une disparité entre la prédiction et la règle. Par exemple, si une directive signale un risque élevé, mais que la probabilité affichée est faible, une alerte attire l'attention du professionnel de santé.

Nous illustrons ensuite la différenciation de corrélation et causalité. Dans le domaine informatique, l'exemple du calcul de l'*ATE* permet d'évaluer l'effet concret d'une intervention, telle que la délivrance

d'un médicament. Cela permet au professionnel de juger, sur le plan cognitif, si le progrès observé est dû au traitement et pas simplement à un facteur associé. Quelle est l'importance décisive de cette différence ? Parce qu'elle se garde de prendre une décision fondée sur une illusion statistique. Comment cela est-il incorporé ? À l'aide d'une modélisation floue déterminant les niveaux d'affiliation des patients aux divers profils de risque. Par exemple, un traitement est considéré efficace pour un profil « moyen », mais inefficace pour un profil « élevé ».

Tout bien considéré , nous illustrons la méthodologie comme un processus cohérent. Sur le plan informatique, elle lie la préparation, la modélisation, l'étalonnage et l'évaluation dans un processus intégré. Sur le plan cognitif, elle transforme des opérations abstraites en instructions concrètes et applicables. Aussi, pourquoi est-il indispensable de maintenir cette consistance ? Parce qu'elle apporte au système une fiabilité et une capacité d'adaptation. Qu'est-ce que cela signifie concrètement ? Lorsqu'un expert reçoit une prédiction, il a la possibilité de consulter la probabilité révisée, la règle interprétable et de simuler une intervention. Par exemple, diminuer le taux de sucre dans le sang d'un patient dans le simulateur et constater la réduction correspondante du risque.

Dès lors, nous mettons au point une méthodologie dédiée pour la création de solutions IA intégrant des composants de logique floue, de probabilités, de modèle arborescent et de régression logistique. Elles visent à pallier les insuffisances des systèmes actuels en termes d'interprétabilité et d'analyse causale dans un but de valorisation. De la sorte, notre méthode fusionne la précision prédictive d'un modèle arborescent avec l'interprétabilité des systèmes *PFL*. Autant, au centre de notre approche se trouve le processus de fuzzification des données par le biais du regroupement et l'insertion de scores flous dans le modèle prédictif. De ce fait, grâce à cette approche, nous pouvons produire des « *IF-THEN* » intelligibles fournissant une interprétation des prédictions.

En outre, la fiabilité est un élément central de notre démarche méthodologique. Nous employons la calibration isotonique ainsi qu'un mécanisme d'alerte en cas d'incohérences pour assurer que l'IA demeure un collaborateur de confiance pour le spécialiste. Subséquemment, notre méthode d'explication de la causalité, prenant en compte les éléments d'incertitude dans l'*ATE*, a pour but de différencier les corrélations des causalités, un point décisif pour les interventions dans un contexte critique.

Également, cette méthodologie adopte un point de vue valorisant, utile et pratique. Pareillement, notre but est de créer une IA transparente et vérifiable ; nous consolidons la confiance des spécialistes et encourageons une prise de décision informée. L'ambition est de développer un système qui soit non seulement efficace, mais qui aide également à la responsabilisation et l'interprétation, en mesure de satisfaire les besoins des situations critiques. Cette méthodologie nous permet donc de construire un système concret.

En somme, notre méthodologie met en relation plusieurs éléments cardinaux. Informatiquement et selon notre stratégie, ce sont des modèles robustes. Cognitivement, ce sont des mécanismes d'interprétation et de causalités adaptés. L'association de ces deux portées garantit que les résultats sont fiables et intelligibles. Leur avancement est marqué par des progrès à chaque étape, de la préparation à l'évaluation. Le flux de travail tisse ces étapes en une chaîne cohérente aboutissant à un système intégrable. Exemple, l'ajout de règles « *IF-THEN* » transforme un calcul statistique en une information exploitable par un spécialiste. *In fine*, notre méthodologie prépare l'implémentation concrétisant ces choix en produisant une solution interactive en soutenant la prise de décision.

CHAPITRE 7 :
IMPLÉMENTATION

7.1 Introduction

Tout d'abord, nous utilisons l'illustration d'un établissement hospitalier de base qui intègre des données relatives aux patients en cours de traitement. Techniquement, l'exemple des algorithmes de la catégorie *LightGBM* estime les probabilités de récidive. Sur le plan cognitif, des règles interprétables, décodant ces prédictions, assistent les spécialistes. Pourquoi est-il indispensable d'avoir cette dualité ? Parce qu'un système efficace, mais non transparent, ne serait pas adopté par les professionnels. Comment relever ce défi ? En associant un modèle prédictif performant à une couche d'interprétabilité. Par exemple, un patient présente un risque de 67 %, mais ce chiffre est associé à une règle interprétable qui fait le lien entre la glycémie, l'âge et la tension. Cela convertit une sortie codée en un résultat intelligible.

Aussi, notre chapitre d'implémentation illustre la mise en œuvre de notre méthodologie. Nous exposons, dès lors, les algorithmes élaborés pour répondre à nos questions de recherche. Le but est d'établir, également, que les solutions suggérées sont à la fois pratiques et reproductibles. Nous détaillons, en conséquence, les architectures de ces algorithmes, les choix technologiques et les principaux modules.

Par conséquent, comme mentionné dans notre méthodologie, nous mettons en œuvre une stratégie hybride et sophistiquée. Elle répond à un double objectif. Nous visons à construire des architectures informatiques robustes et performantes. Cela implique l'utilisation d'algorithmes de pointe en précision pour notre triangulation d'interprétabilité, d'enrichissement en analyse causale, pour une perspective de valorisation. De la sorte, nous répondons à un besoin cognitif : celui d'offrir à l'expert un outil de décision transparent, fiable et cohérent. Ce dernier volet se traduit par la génération de règles intelligibles et une quantification de la confiance dans ces règles. Notre approche conjugue ainsi interprétation, analyse causale et précision.

7.2 L’architectures de nos algorithmes

Nous illustrons les architectures des algorithmes par l’imagination d’un service d’urgences recevant un flot constant de données cliniques. Dans le cadre de l'informatique, un pipeline Python gère la normalisation, la prédiction et l'étalonnage. Sur le plan cognitif, les médecins ne se contentent pas de recevoir des probabilités ; ils obtiennent également des interprétations accompagnées de règles. Pourquoi est-il judicieux d'opter pour ces outils ? Parce qu'ils associent les choix internes du modèle aux variables cliniques effectivement employées. Comment ça marche ? Les règles indiquent l'impact proportionnel de chaque variable sur une prédiction spécifique. Par exemple, l'outil suggère que pour un patient donné, 77 % du risque estimé est attribuable à un taux de glucose anormal. Ce lien tangible facilite la vérification par un expert.

Par la suite, nous avons élaboré plusieurs algorithmes représentant notre stratégie. Certains algorithmes se concentrent sur l'inférence causale et d’autres sur l'interprétabilité de la prédiction précise. L'architecture informatique repose sur une chaîne de modules Python orchestrés dans différents pipelines. Nous nous appuyons sur des librairies reconnues pour la science des données, comme *numpy* et *pandas* pour la manipulation des données. Pour le volet prédictif, le modèle arborescent *LightGBM* est le composant utilisé. Ce choix est justifié par ses performances élevées, sa rapidité d'exécution et ses capacités de traitement des données de grande dimension. Ensuite, nous utilisons *scikit-learn* pour le clustering probabiliste (*Gaussian Mixture Model : GMM*) et la régression logistique pour la fusion des modèles, ce qui crée le cœur du système d'interprétabilité. Aussi, nous utilisons les librairies *SHAP* et *LIME* pour visualiser les explications locales et globales. Cette combinaison de briques logicielles démontre un avancement des deux dimensions informatiques et cognitives, étant donné qu’elle intègre à la fois des méthodes performantes et des outils de valorisation.

7.2.1 Notre architecture d’analyse causale floue

Nous illustrons l’architecture d’analyse causale floue par l’exemple d’un cas de traitement antibiotique administré dans un service hospitalier. Sur le plan informatique, le *BGMM* classifie les patients en trois profils de risque flou. Sur le plan cognitif, les cliniciens perçoivent ces profils comme des catégories bien connues. Pourquoi est-il indispensable d'adopter cette logique floue ? Étant donné que la réaction au traitement n'est pas absolue et fluctue en fonction des caractéristiques. Comment cela est-il intégré ? En

assignant à chaque patient un niveau d'affiliation à chaque profil. Par exemple, on évalue un patient à 77 % « moyen » et 23 % « élevé ». Cette issue illustre la situation dans le domaine médical où un cas peut se trouver à cheval entre deux conditions.

Dans l'exemple d'une architecture utilisant le clustering, les subtilités causales représentent un ensemble de profils (ex. faible, moyen, élevé) pour une seule variable de confusion. Cet ensemble permet de comparer les corrélations à ces relations de cause à effet. Les critères précis et mesurables de ces subtilités peuvent être la mesure de l'erreur d'estimation de l'effet moyen du traitement (*ATE*) entre un modèle causal découvert par un algorithme et un autre modèle. Aussi, la dégradation de la précision du modèle lorsqu'il est testé sur un jeu de données dont la distribution des variables de confusion a été modifiée.

Le flux de travail causal flou est élaboré pour guider l'expert dans l'évaluation de l'impact d'un traitement, en tenant compte de la variation des données. Le processus débute de la sorte par la collecte des variables (caractéristiques) cliniques, la cible et la caractéristique de ce traitement. Un mélange gaussien bayésien (*BGM*) est de plus utilisé pour segmenter chaque variable confondante en trois profils flous (subtilités causales). La figure 9 explique comment nous affichons les résultats pour l'architecture d'analyse causale floue.

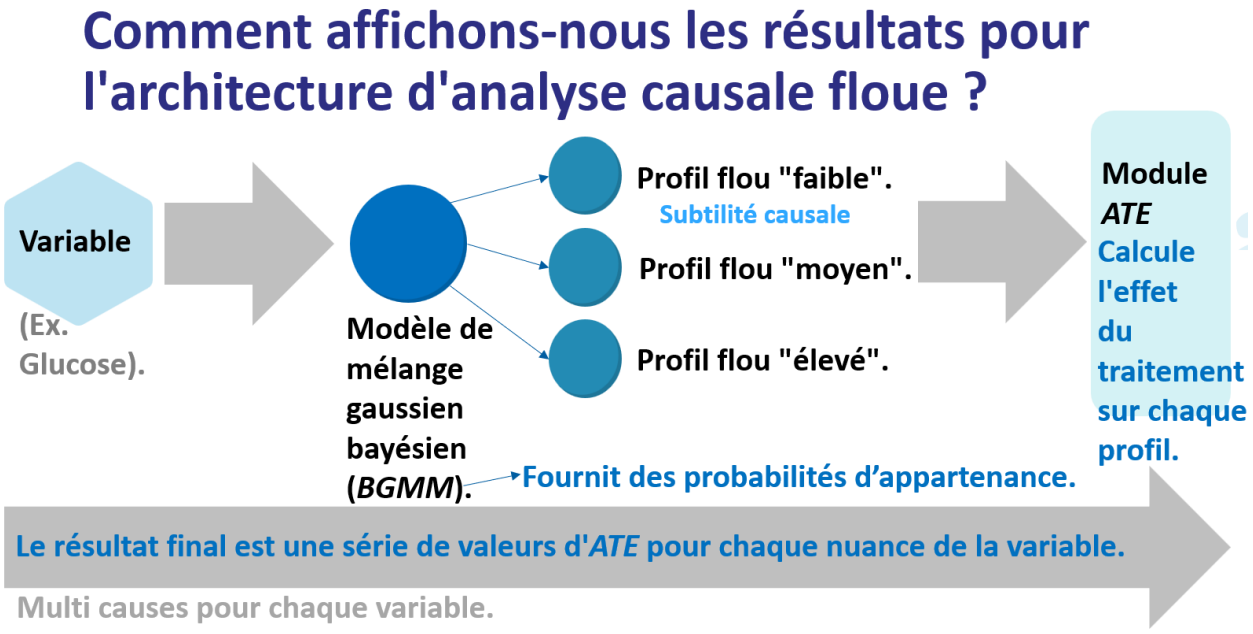

Figure 9 : Comment affichons-nous les résultats pour l'architecture d'analyse causale floue?

En définitive, la partie cognitive de cette architecture est la capacité à traduire des valeurs numériques précises, comme un taux de glucose, en concepts nuancés qu'un clinicien peut interpréter et utiliser dans son raisonnement. Nous choisissons le *Bayesian Gaussian Mixture (BGMM)* pour sa robustesse face aux données complexes, de plus, sa capacité à estimer de manière probabiliste les appartenances à des groupes. C'est une amélioration, puisque cela rend le modèle capable de saisir les transitions entre les états (par exemple, d'un risque « faible » à « moyen ») d'une manière plus naturelle. Par ailleurs, cette approche nous permet d'aller au-delà de la simple association en élaborant un raisonnement causal. C'est une avancée intéressante, étant donné que le système peut ainsi s'ajuster au processus décisionnel humain.

7.2.1.1 Notre conception des modules du flux causal flou

Nous illustrons la conception des modules du flux par l'exemple d'un module *ATE* grâce à une cohorte dans laquelle un médicament est examiné. Dans le domaine de l'informatique, une régression linéaire pondérée évalue l'impact moyen du traitement. Sur le plan cognitif, les cliniciens interprètent ces résultats comme une variation du risque. Pourquoi est-il déterminant d'incorporer des poids flous ? Parce qu'il distingue les patients en fonction de leur profil plutôt que de fournir un effet global. Comment cela se traduit-il ? Un patient se rapprochant du profil « élevé » a une plus grande influence sur le calcul. Par exemple, l'*ATE* ajusté indique que le médicament diminue le risque de 17 % dans le profil dit « moyen », mais de seulement 7 % dans le profil qualifié d'« élevé ».

De cette manière, nous avons les modules suivants :

- **Module *ATE***

  Nous avons créé une classe *ATE* pour analyser l'impact causal d'une intervention, comme par exemple, l'administration d'un antibiotique. Le module emploie une régression linéaire pour évaluer l'effet causal moyen du traitement. L'originalité de la mise en œuvre découle de l'emploi de poids (*memberships*) dérivés de notre modèle flou. Pourquoi ce choix ? Cette méthode offre une perspective nuancée en considérant l'intensité de l'appartenance de chaque patient à un profil de risque donné. C'est un avantage, du fait que cela nous permet d'étudier comment l'effet d'un traitement varie non seulement selon sa présence ou son absence, mais aussi selon les caractéristiques du patient. Le calcul est un exemple de l'entrelacement du volet informatique et

cognitif où la modélisation mathématique de l'*ATE* est affinée par l'interprétation floue des données.

- **Module *FuzzyBGMM***

  La classe *FuzzyBGMM* gère la fuzzification des variables. Elle utilise, de cette manière, le *BGMM* de *Scikit-learn* pour attribuer à chaque observation des degrés d'appartenance aux clusters. Ainsi, nous employons un *BGMM* pour la segmentation des variables cliniques. Ce module a pour entrée une variable et la divise en plusieurs profils ou clusters flous (par exemple : « faible », « moyen », « élevé »). L'avancement des deux dimensions réside dans la capacité du *BGMM* à déterminer automatiquement le nombre de ces profils sans que nous ayons besoin d'une intervention manuelle. Le but est de créer des concepts clairs et cliniquement pertinents à partir de données brutes, ce qui est un élément central de notre approche cognitive. Le module génère une matrice d'appartenance floue quantifiant dans quelle mesure chaque patient correspond à chaque profil. Ces scores d'appartenance sont ensuite utilisés dans le calcul des effets causaux.

7.2.2 Notre architecture interactive *PFL-LightGBM*

Nous illustrons l'architecture interactive *PFL-LightGBM* en prenant l'exemple du suivi des patients souffrant de maladies cardiaques. Techniquement, le pipeline *PFL-LightGBM* comprend la préparation des données, l'apprentissage, l'étalonnage et les alertes d'incohérence. Sur le plan cognitif, les médecins se réfèrent aux règles floues qui traduisent les prédictions. Pourquoi cette modularité est-elle déterminante ? Parce qu'elle offre la possibilité de modifier chaque phase et de localiser l'endroit où une discordance se produit. Comment cela est-il bénéfique pour l'expert ? En présentant en même temps une chance et une règle. Par exemple, une estimation de 80 % de risque est liée à la condition « hypertension et taux élevé de cholestérol ». Le spécialiste contrôle tout de suite la pertinence clinique.

De cette façon, l'architecture *PFL-LightGBM* est modulaire. Elle débute avec un module de configuration, puis passe à un gestionnaire de données élaborant les jeux d'apprentissage. Le module *PFL*, qui est chargé de manière dynamique, effectue le clustering flou et génère, en plus, les scores d'interprétabilité. Ces scores sont par la suite intégrés au modèle prédictif *LightGBM* lors de l'exécution de la validation croisée. Le processus comprend également la calibration (l'étalonnage) isotonique après l'entraînement et la

création de métriques. L'analyse est, en conséquence, améliorée grâce à une classification du risque par ensemble et accompagnée d'alertes en cas d'incohérence. La Figure 10 résume cette architecture.

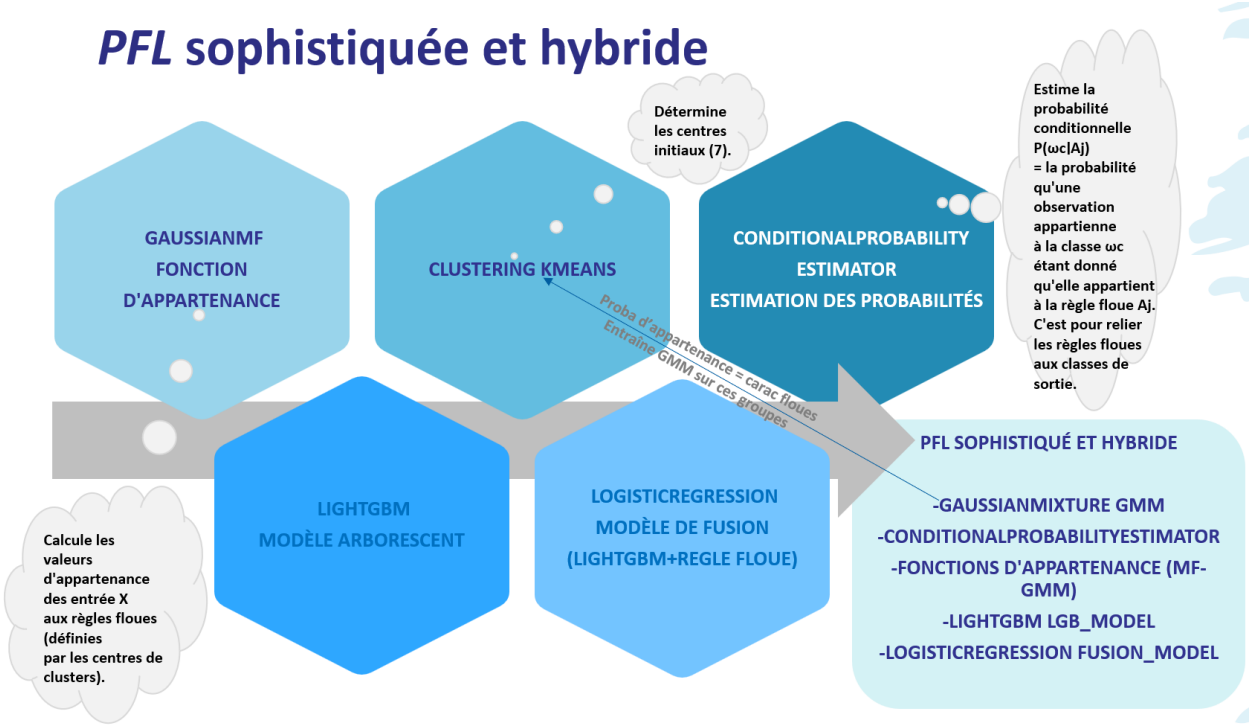

Figure 10 : *PFL-LightGBM*.

Ainsi, l'un des résultats souhaités de l'architecture interactive *PFL-LightGBM* est à la Figure 11.

```
✨ Cluster 1 - RISQUE MOYEN
    • Risque calibré (modèle arborescent): 10.8%
    • P(mortalité|cluster) (Règle Floue): 12.30%
    • Fidélité du Cluster                                    : 0.9851
    • Recommandation clinique: Surveillance renforcée
    • Nombre de patients: 44,880
    • Pourcentage du subset: 9.3%
    • Règle clinique: arterial base excess ∈ [-5.131, 7.862] ET arterial pco2 ∈ [-4.307, 8.
687] ET arterial ph ∈ [-7.049, 5.945] ET arterial po2 ∈ [-7.011, 5.982] ET bicarbonate ∈ [-
4.791, 8.202] ET bilirubin ∈ [-6.930, 6.064] ET calcium ∈ [-5.540, 7.454] ET creatinine ∈
[-7.108, 5.885] ET crp ∈ [-4.409, 8.584] ET cvp ∈ [-6.497, 6.497] ET diastolic blood pressu
re ∈ [-6.378, 6.615] ET fio2 ∈ [-5.574, 7.419] ET glucose ∈ [-6.645, 6.348] ET got(asat) ∈
[-6.669, 6.325] ET gpt(alat) ∈ [-6.718, 6.276] ET heart rate ∈ [-5.890, 7.104] ET hematocri
t ∈ [-6.530, 6.464] ET platelets ∈ [-5.131, 7.863] ET potassium ∈ [-5.650, 7.343] ET ptt ∈
[-6.864, 6.130] ET sodium ∈ [-6.372, 6.622] ET spo2 ∈ [-7.813, 5.181] ET systolic blood pre
ssure ∈ [-6.768, 6.225] ET temperature ∈ [-6.287, 6.707] ET urea ∈ [-6.819, 6.174] ET white
blood cells ∈ [-6.975, 6.019]
    • Cohérence Règle-modèle arborescent: ✅ Excellente (écart: 1.5%)
```

Figure 11 : L'un des résultats souhaités de l'architecture interactive *PFL-LightGBM*.

Interprétation d'une règle « *IF-THEN* »

Par exemple, pour les résultats quantitatifs, le modèle identifie un Cluster 1 comme un groupe à risque moyen. De même, il est établi que ce groupe inclut 44 880 patients, ce qui représente 9.3 % de l'ensemble des données. De plus, le modèle arborescent calibré attribue un risque de 10.8 % à ce groupe. Ensuite, la règle floue liée à ce cluster indique un taux de mortalité de 12,3 %. Également, nous observons une bonne concordance entre ces deux prédictions, avec un écart de seulement 1,5 %. Un coefficient de fidélité du cluster de 0.9851 valide la pertinence du regroupement des patients. En somme, cela indique que les individus de ce regroupement partagent des traits communs.

Autant, nous élaborons une directive clinique pour ce groupe. Cette règle est énoncée en suivant la logique « *IF-THEN* ». Cette approche est un fondement de la logique floue et de l'interprétabilité des modèles. Elle offre une traduction des relations du modèle en un langage clair. Si un patient montre des caractéristiques spécifiques (par exemple, si son *arterial base excess* est dans l'intervalle [-5.131, 7.862] ET si son *arterial pco2* est dans l'intervalle [-4.307, 8.687] et ainsi de suite pour toutes les variables).

De plus, le patient appartient au groupe à risque moyen. Cela se traduit par une probabilité de mortalité associée de 12.30 % et une recommandation clinique de surveillance renforcée. L'utilisation de la règle « *IF-THEN* » permet de justifier la classification du patient. Il s'agit d'une démarche de raisonnement qui se fonde sur l'observation pour arriver à une conclusion. Nous définissons des règles comportant des plages pour chaque variable. Ces intervalles sont établis à partir des caractéristiques moyennes des patients de ce cluster. Un patient est classé dans ce groupe s'il respecte toutes les conditions de la règle.

En fin de compte, le modèle ne se base pas sur une seule variable. Il considère une combinaison de valeurs de variables. Cela fournit une interprétation plus complète. L'interprétation nous permet de comprendre pourquoi un patient est classé à risque moyen. Cela rend les prédictions pratiques pour les professionnels de la santé. Dans certains cas le système recommande de renforcer la surveillance pour les patients de ce groupe parce que leur risque de mortalité est notable bien qu'il ne soit pas élevé.

7.2.2.1 Notre conception des modules du flux *PFL-LightGBM*

Nous illustrons la conception des modules du flux en prenant l'exemple d'un patient en oncologie. Sur le plan informatique, le pipeline importe ses données, les normalise puis forme le modèle *LightGBM*. Sur le

plan cognitif, l'interprète produit également une règle simple pour interpréter la prédiction. Quelle est la pertinence de cette intégration multimodale ? Elle assure que chaque phase, de l'insertion des données à la production des résultats, génère des conclusions uniformes et intelligibles. Comment cela se manifeste-t-il de manière tangible ? Avec le simulateur causal, un oncologue a la possibilité d'examiner l'impact d'une modification, telle que la diminution d'un marqueur inflammatoire et d'observer comment le risque se modifie. Par exemple, à la suite de l'intervention, le risque diminue de 71 % à 29 %, confirmant l'impact escompté d'un traitement.

De la sorte, nous avons les modules suivants :

- **Pipeline principal**

  L'ensemble du processus est orchestré par le pipeline principal. Il gère dès lors le montage du *Drive*, le chargement des données, l'arrimage dynamique du module *PFL* et la coordination des différentes étapes d'entraînement, d'évaluation et d'analyse.

- **Prise en charge des données**

  Ce module fondamental intègre les données cliniques, les épure, et effectue les calculs statistiques requis pour la normalisation. Le jeu de données brut sert d'entrée. Le résultat est un lot de données préparées pour les phases suivantes.

- **Modèle d'entraînement**

  Il combine les prédictions de *LightGBM* et les évaluations des règles floues pour produire des probabilités de risque. Les données préparées constituent l'entrée, tandis que le modèle entraîné et ajusté représente la sortie.

- **Interpréteur**

  Ce composant produit des indices d'interprétabilité, des directives cliniques pour la normalisation et des diagnostics pour chaque groupe. Le modèle et les données brutes constituent les éléments d'entrée. Le résultat est une collection d'interprétations et de notifications.

- **Module de calibration**

Nous utilisons la régression isotonique pour modifier les probabilités de sortie du modèle, garantissant ainsi un alignement entre la confiance estimée et l'exactitude réelle.

- **Simulateur causal**

  Ce module permet de tester l'impact d'une intervention donnée par un expert sur le risque prédit. L'entrée est un profil patient et des modifications de variables. La sortie est la nouvelle prédiction de risque et l'interprétation associée.

Aussi, la partie d'interprétabilité du modèle combine la puissance prédictive du *LightGBM* avec l'intelligibilité d'un système de règles floues. Cette architecture se base sur un modèle hybride où une régression logistique fusionne les prédictions du *LightGBM* avec les probabilités des règles floues.

Partie informatique : Le cœur de l'implémentation est la classe *PFL*. Elle utilise un *GMM* pour regrouper les patients en clusters, qui correspondent aux règles floues. De surcroît, elle utilise la régression logistique pour combiner les prédictions des deux modèles. Le choix de ces modèles nous permet d'avoir un modèle à la fois performant et transparent. Nous utilisons également un outil pour la normalisation des valeurs afin que les règles soient exprimées avec les unités cliniques d'origine, rendant ainsi leur interprétation plus aisée pour un clinicien.

Partie cognitive : L'élaboration de règles floues, telles que « [*bilirubin*] ∈ [-0.48, 13.72] ET [*crp*] ∈ [-5.22, 4.29] », est primordiale pour l'approbation clinique de l'instrument. Ces directives expriment les processus complexes du modèle dans une forme facile à comprendre et à mettre en œuvre. Ce système offre un double niveau d'interprétation.

Interprétation locale : pour un patient donné, le modèle interprète les facteurs clés influençant sa prédiction. Par exemple, pour un patient, « un niveau de *bilirubin* élevé augmente le risque de mortalité ».

Interprétation globale : des règles généralisées sont produites pour des groupes de patients aux caractéristiques similaires. Ces règles fournissent des interprétations de la prédiction du modèle ainsi que

la probabilité de l'issue (par exemple, 25% de mortalité dans ce cluster). Cette approche répond au besoin de contextualisation du clinicien et lui permet de mieux comprendre les recommandations du système.

7.3 La synthèse de notre flux de travail et notre assistance à un expert

Nous illustrons la synthèse du flux de travail et assistance à l'expert en citant le cas d'un service où un clinicien gère plusieurs patients à risque élevé. Sur le plan informatique, le flux offre une série de probabilités ajustées. Sur le plan cognitif, il fournit des règles limpides et des diagnostics d'incohérence. Pourquoi cette combinaison apporte-t-elle un meilleur soutien au spécialiste ? Parce qu'elle fait la différence entre les cas sûrs et les cas incertains. Comment cela prend-il forme ? En utilisant des indicateurs de fidélité indiquant lorsque la prédiction et la règle diffèrent. Par exemple, pour un patient donné, une probabilité de 37 % va à l'encontre d'une règle signalant un risque élevé. L'alerte capte l'attention du médecin qui décidera de réexaminer ce dossier.

C'est pourquoi nous avons organisé notre système en un processus de travail modulaire et cohérent pour aider le spécialiste dans ses décisions. Chaque composant joue un rôle précis et collabore avec les autres pour réaliser des performances, des analyses causales et une interprétabilité. Aussi, ce flux de travail assiste le clinicien de plusieurs manières. La partie informatique génère une prédiction de risque précise et calibrée, mais également un ensemble de règles interprétables. Autant, la partie cognitive est servie par la traduction des nombres en concepts cliniquement pertinents et par l'analyse des divergences entre le modèle et ses règles. Cela permet d'identifier des cas où le modèle est le plus fiable.

En somme, les relations entre les deux parties se révèlent à travers les indicateurs de fidélité et de cohérence. Elles précisent si les interprétations des règles sont conformes aux prédictions du modèle, assurant ainsi une certaine fiabilité pour le clinicien. Le simulateur illustre de manière tangible cette connexion offrant à l'expert la possibilité d'ajuster les entrées pour saisir les prédictions.

7.4 Nos justifications

Nous illustrons les justifications du choix *BGMM* à travers l'exemple d'une analyse causale dans le contexte d'un suivi de traitement complexe. Techniquement, le modèle segmente avec précision les variables. Sur

le plan cognitif, il génère des profils pertinents. Pourquoi cela dépasse-t-il une simple corrélation ? Parce que la segmentation probabiliste saisit la subtilité des transitions. Comment cela perfectionne-t-il la pratique ? En démontrant qu'un traitement efficace pour un profil moyen n'a pratiquement aucune incidence sur un profil élevé. Par exemple, un facteur de rectification indique qu'un effet réel diverge de la moyenne brute, renforçant ainsi la validité de l'interprétation médicale.

Dès lors, pour la partie causale, l'emploi du *BGMM* permet une segmentation probabiliste saisissant les subtilités de la variable. De ce fait, l'analyse détaillée pour chaque segment, suivie de son ajustement grâce à un facteur de correction, présente une perspective nuancée. Outre cela, cette approche, tirée des recherches de Faghihi et coll., offre la possibilité d'évoluer d'une simple corrélation vers une tentative d'élaboration du raisonnement causal. Le coefficient de correction garantit, conséquemment, l'accord entre les évaluations floues et une valeur manuelle standard.

Tout autant, l’architecture modulaire de la partie d’interprétabilité prédictive est élaborée pour assurer robustesse et reproductibilité. La gestion des données maintient, pareillement, les échelles réelles des variables afin de produire des règles intelligibles. L'incorporation de scores flous dans le *LightGBM* dote le modèle d'une faculté prédictive tout en préservant une connexion directe avec l'interprétation. Les modules de calibration et d'alerte fonctionnent, du coup, comme des garde-fous, renforçant la fiabilité et la crédibilité des architectures.

7.5 Conclusion

Globalement, nous illustrons la représentation de l’implémentation par un système intégré transformant des données brutes en prédictions claires, calibrées et interprétées. Elle se base sur des pipelines robustes, des algorithmes efficaces et des modules spécialisés d'un point de vue technologique. Sur le plan cognitif, elle convertit les résultats en règles, diagnostics et simulations faciles à comprendre. Pourquoi est-il logique d'avoir cet ancrage ? Parce qu'un système n'est véritablement fiable que s'il est adopté par les spécialistes qui l'utilisent. Comment peut-on le confirmer ? Grâce à des directives compréhensibles, des ajustements et des alertes d'incohérence ; les praticiens sont en mesure d'évaluer eux-mêmes la pertinence. Par exemple, dans un contexte clinique, un médecin confronte directement la prédiction de l'IA avec ses propres standards, facilitant ainsi son adoption.

En procédant ainsi, l'implémentation de nos architectures, l'une centrée sur la causalité et l'autre sur la capacité d'interprétation de la prédiction, représente l'application pratique de notre approche méthodologique. Les divers composants logiciels sont en outre liés entre eux et constituent un système cohérent. De là, ce système est à présent préparé pour être testé.

Également, ce chapitre assemble plusieurs éléments importants. Nous avons mis en place une partie informatique performante avec des modèles hybrides et une infrastructure de pipeline. La partie cognitive est prise en compte par la création de règles floues et de concepts interprétables ainsi que par un simulateur causal. Les liens entre ces parties sont formalisés par les métriques de cohérence et de fidélité. Notre avancement des deux dimensions réside dans l'intégration réussie de ces deux volets, ce qui donne un outil d'analyse causale de prédiction et d'interprétation fiable. Le flux de travail modulaire et ses sorties dédiées assistent le clinicien en lui offrant un discernement des prédictions, au-delà d'un simple chiffre.

En définitive, nos futures expérimentations vont explorer davantage le volet causal en intégrant les méthodes d'inférence causale. Ces expérimentations vont chercher à personnaliser les règles floues en fonction du patient et de l'expertise de chaque clinicien, en intégrant un système de boucles de rétroaction. Cela nous permettrait d'adapter le système aux besoins individuels de chaque utilisateur. Dans le prochain chapitre, nous détaillerons, tout compte fait, les protocoles expérimentaux mis en œuvre pour juger les performances et valider nos hypothèses.

CHAPITRE 8 :
EXPÉRIMENTATIONS

8.1 Introduction

Nous illustrons notre introduction du chapitre d'expérimentations par le cas d'un service de soins intensifs où un modèle anticipe le risque de décès dans les 72 heures. Sur le plan informatique, il utilise *LightGBM* pour les prédictions et un clustering flou pour l'étude de causalité. Sur le plan cognitif, il produit des règles compréhensibles accompagnant les prédictions. Pourquoi est-il décisoire de passer par cette évaluation ? Par ce qu'un modèle de valeurs et digne de confiance doit être validé à travers des protocoles expérimentaux stricts. Quelle est l'organisation de ce processus ? À travers une évolution assurant la liaison entre la préparation, la modélisation et l'interprétation des données. Par exemple, un simulateur offre à un professionnel de la santé la possibilité d'expérimenter l'impact d'un antibiotique sur un patient et de suivre l'évolution du risque.

Donc, ce chapitre évalue la performance de nos implémentations en exposant les expérimentations[38] réalisées pour confirmer notre stratégie IA sophistiquée et hybride. Il s'agit, dès lors, de mettre à l'épreuve les solutions développées pour vérifier nos hypothèses cognitives et informatiques. Simultanément, nous utilisons une méthode scientifique pour valider ces hypothèses, tester nos solutions dans un environnement contrôlé, utiliser des métriques, et analyser les résultats. Nos expérimentations, telles que : l'analyse causale, la sélection de variables importantes via la permutation bayésienne et l'interprétabilité des prédictions visent de la sorte à confirmer la capacité de nos modèles à prospecter un équilibre de prédictions précises avec la triade d'interprétabilité, d'enrichissement en analyse causale pour une perspective de valorisation. Nous élaborons, de ce fait, des procédures pour ces expérimentations. Notre but est de prouver par conséquent que nos modèles sont capables de tenter de reproduire les processus cognitifs, en distinguant, par exemple, causalité et corrélations pour leur intégration dans des contextes

[38] https://github.com/Boumediene-ELMouenis

critiques. D'où, nous tentons de maintenir une certaine rigueur pour chaque protocole afin d'assurer la fiabilité des résultats.

Tout bien considéré, notre approche est une démonstration de notre stratégie intégrant l'informatique et la cognition. La partie informatique s'appuie sur des architectures, fusionnant l'analyse causale, la puissance prédictive de *LightGBM* et la transparence de règles floues. Notre avancement sur les deux parties se concrétise dans cette capacité à combiner différents paradigmes IA. Le volet cognitif est au centre de notre démarche. Nous traduisons les données brutes en concepts cliniquement pertinents avec des règles lisibles pour augmenter le raisonnement d'un clinicien. Le flux de travail est un pipeline structuré. Il va de la préparation des données à la génération des interprétations causales et prédictives. Ce processus est conçu pour assister ce clinicien. Les liens entre les deux volets sont mesurés par des métriques de cohérence. Elles garantissent que les interprétations des règles floues reflètent fidèlement les prédictions du modèle. Un exemple de cette intégration est notre simulateur. Il permet au clinicien de tester des scénarios et de visualiser l'impact causal d'un traitement sur un patient. En conséquence, la validation de ces hypothèses par des expérimentations rigoureuses est notre objectif.

8.2 L'éthique et notre source de données

Nous illustrons l'éthique et source de données par la considération des données *MIMIC-III*. Informatiquement, elles offrent des milliers d'observations structurées et anonymisées. Cognitivement, elles garantissent la confiance des cliniciens grâce au respect des règles éthiques. Pourquoi ce choix est-il pertinent ? Parce qu'il associe richesse statistique et sécurité. Comment est-il appliqué ? Par une autorisation éthique, une anonymisation stricte et une sélection de 26 variables cliniques pertinentes. Par exemple, des paramètres comme la glycémie, la tension et la bilirubine sont retenus, parce qu'ils traduisent directement l'état physiologique d'un patient atteint de sepsis.

De ce fait, nous étions conscients que notre travail va impliquer des contraintes éthiques. Régler ces contraintes était la première balise pour notre flux de travail, de méthodologie, d'implémentation et d'expérimentations. Dès lors, nous avons eu l'autorisation d'utiliser les données du site du *Medical Information Mart for Intensive Care III (MIMIC-III)* : https://physionet.org/content/synthetic-mimic-iii-health-gym/1.0.0/. De plus, selon ce site, le risque de divulgation d'identité associé à ces données anonymisées a été estimé comme très faible (0,045 %).

Pour ces raisons éthiques, nous avons réalisé une demande d'autorisation pour accéder aux données *MIMIC*. Nous avons rempli, donc, un formulaire dont le but principal était d'informer des conditions d'utilisation de la BD. Aussi, il leur a été fourni une description avec les motifs d'accès. De surcroît, nous avons suivi une formation éthique sur les sujets humains en signant un accord d'utilisation des données. Ensuite, notre demande a été approuvée par un comité d'examen. Ainsi, après l'accord de consentement, les données de cette BD ont été accessibles. Autant, il faut savoir qu'elles sont anonymisées et standardisées.

En outre, *MIMIC* regroupe les signes vitaux, les valeurs de laboratoire et les données démographiques. De même, elle compte plusieurs milliers de patients avec une quarantaine de variables cliniques. Ces variables peuvent être divisées en signes vitaux (fréquence cardiaque, tension artérielle, etc.), en valeurs de laboratoire (pH, numération plaquettaire, hémoglobine, etc.) et en données démographiques (âge, sexe) (Strickler et coll., 2023). En parallèle, ce sont les données de 2164 patients atteints de sepsis dans l'unité de soins intensifs. Les ensembles de données ont été générés et publiés pour développer des algorithmes d'apprentissage automatique, et ce à des fins éducatives. De plus, notre idée était d'avoir une variable de traitement antibiotique (A) qui est binaire, soit le patient a pris un antibiotique unique (A=1), soit multiple (A=0). Aussi, la variable de résultat (y) est la mortalité à 72 heures. Enfin, pour l'interprétabilité et l'analyse causale, 26 variables cliniques sont sélectionnées comme facteurs de confusion.

Également, dans le secteur de la santé, chaque choix entraîne des conséquences immédiates sur la vie des patients. Une IA mal élaborée ou sans validation valorisante (ex. en éthique) peut engendrer des préjugés et exacerber les disparités en matière de soins. Le manque de transparence des algorithmes suscite la méfiance des praticiens et également du grand public. Nous cherchons la transparence et la traçabilité de notre travail. Notre approche, intégrant le volet cognitif et l'interprétabilité, répond directement à ce besoin valorisant en rendant le processus de prise de décision de l'IA intelligible. Elle se focalise sur l'assurabilité des résultats.

Précisément, nous avons opté pour les données *MIMIC-III*, considérant qu'elles représentent un cas d'étude pertinent et réel concernant le choc septique. Cette BD jouit d'une grande réputation au sein de la communauté scientifique. Nous nous focalisons sur les variables cliniques les plus significatives pour plusieurs raisons. Du côté informatique, l'inclusion d'un trop grand nombre de variables, un phénomène

connu sous le nom de malédiction de la dimension, augmente la complexité du modèle. Le risque de surapprentissage est réel. La sélection des variables les plus pertinentes garantit un modèle robuste et généralisable. Par ailleurs, du côté cognitif, la restriction du nombre de variables à un sous-ensemble pertinent rend les règles d'interprétation plus digestes et plus faciles à valider pour un clinicien. La simplicité des règles favorise leur adoption et leur application. Notre choix se porte sur les variables suivantes : *1. arterial base excess, 2. arterial pco2, 3. arterial ph, 4. arterial po2, 5. bicarbonate, 6. bilirubin, 7. calcium, 8. creatinine, 9. crp, 10. cvp, 11. diastolic blood pressure, 12. fio2, 13. glucose, 14. got(asat), 15. gpt(alat), 16. heart rate, 17. hematocrit, 18. platelets, 19. potassium, 20. ptt, 21. sodium, 22. spo2, 23. systolic blood pressure, 24. temperature, 25. urea, et 26. white blood cells*. Ces caractéristiques ont un lien direct avec les processus inflammatoires et hépatiques (qui a rapport au foie) du choc septique. Ces variables sont familières aux cliniciens. Leur inclusion facilite l'intégration de notre système dans la pratique courante.

Fréquence des antibiotiques

Nous avons préparé les données médicales pour la modélisation du choc septique en nous basant sur la mise en œuvre de Reijnders (2020) avec notre requête SQL. La première partie d'analyse de notre travail sur la causalité permet de déterminer la fréquence de prescription de certains antibiotiques pour tous les patients en choc septique. Les résultats montrent que la Vancomycine est de loin l'antibiotique le plus couramment prescrit, avec 11319 occurrences, suivies de loin par la Levofloxacine (3398) et le Métronidazole (1620). À l'inverse, la Moxifloxacine (3), la Pipéracilline (5) et la Tobramycine (7) sont les moins prescrites. Cette hiérarchie de prescription suggère que la Vancomycine pourrait être un traitement de première ligne ou un antibiotique à large spectre souvent utilisé pour cette condition. Le nombre d'occurrences de chaque antibiotique par ordre décroissant dans *MIMIC III* est résumé dans le tableau 4.

Nom de l'antibiotique	Nombre d'occurrences de chaque antibiotique par ordre décroissant dans *MIMIC III*.
Vancomycin	11319
Levofloxacin	3398
Metronidazole	1620
Gentamicin	662
Azithromycin	586
Aztreonam	435

Clindamycin	374
Ciprofloxacin	313
Cefepime	225
Ceftazidime	213
Erythromycin	180
Amikacin	165
Cefazolin	158
Rifampin	114
Ampicillin	22
Tobramycin	7
Piperacillin	5

Tableau 4 : Tableau des nombres d'occurrences de chaque antibiotique par ordre décroissant dans *MIMIC III*.

Cas spécifiques de patients

Pour notre partie de l'analyse sur notre travail sur la causalité, nous avons examiné les prescriptions d'antibiotiques pour les exemples de trois patients (112, 157 et 166), tous en choc septique.

- Le patient 112 a reçu de la Lévofloxacine et du Métronidazole et ces prescriptions ont été répétées.
- Le patient 157 a reçu du Céfazolin, du Métronidazole et de la Lévofloxacine. La prescription de Céfazolin a été répétée.
- Le patient 166 a reçu uniquement de la Lévofloxacine.

Ces résultats illustrent que, même pour des patients atteints de la même condition, les schémas de traitement peuvent varier, reproduisant la dissemblance des infections et des pratiques médicales.

Analyse statistique des prescriptions

Toujours pour notre partie de l'analyse sur notre travail sur la causalité, nous avons testé la classification des patients en choc septique en deux groupes basés sur le nombre d'antibiotiques prescrits.

- A = 0 : Patients ayant reçu plus d'un antibiotique. Nous observons 1 431 cas dans cette catégorie.
- A = 1 : Patients ayant reçu un seul antibiotique. Nous comptons 843 cas dans ce groupe.

Ces chiffres révèlent que la majorité des patients en choc septique (1431 sur un total de 2274, soit environ 63 %) ont été traités avec une combinaison de plusieurs antibiotiques (polythérapie). Seuls 37 %

des patients ont reçu un seul antibiotique. La polythérapie est une stratégie courante en cas de choc septique pour cibler un éventail plus large de bactéries potentielles. Les exemples de patients 112 et 157 confirment cette tendance à la polythérapie.

Définitivement, pour assister un utilisateur, le flux de travail a un protocole (mode d'emploi). Cet utilisateur commence par l'étape de gestion des données. Il télécharge un fichier de données patient. Un module de nettoyage entre en action. Ce module substitue les valeurs absentes. Le système dénote le niveau de complétude pour les patients, mettant en évidence les lignes où un grand nombre de données est manquant. Il comprend les choix d'imputation avant de continuer. Cette interaction constitue le premier module de notre flux de travail. L'entrée est le fichier brut, la sortie est le jeu de données prétraité et un rapport de qualité de ces données.

8.3 Nos stratégies expérimentales

Nous illustrons nos stratégies expérimentales par l'exemple d'association de *LightGBM* et du clustering flou. Informatiquement, *LightGBM* fournit une prédiction rapide et précise. Cognitivement, les règles issues du clustering traduisent cette prédiction en concepts cliniques. Pourquoi fusionner ces approches ? Parce que des chiffres isolés ne suffisent pas à un expert médical. Comment assurer la cohérence ? Par une régression logistique combinant les deux sorties et valide leur concordance. Exemple, une prédiction de risque de 80 % est accompagnée d'une règle « si calcium élevé et température élevée alors risque fort » renforçant la valorisation et la confiance clinique.

En conséquence, notre stratégie expérimentale est basée sur l'élaboration de modèles d'*ATE* et *PFL-LightGBM* incorporant la possibilité d'établir la causalité et intrinsèquement l'interprétabilité. Nous basons nos expérimentations de causalité sur les recherches de Faghihi et coll. (2020), nous permettant ainsi de passer d'une simple corrélation à une démarche causale. En outre, nous nous appuyant sur les recherches de Fialho et coll. (2016) dans le but de l'interprétabilité de la prédiction, et aussi nous concevons un processus intégrant des modèles de clustering flou et *LightGBM*. Cette stratégie nous permet d'établir des règles interprétables tout en conservant une bonne précision prédictive.

Aussi, notre stratégie expérimentale repose sur la fusion de deux approches. Une première approche informatique utilise le *LightGBM* pour sa performance prédictive. Une autre approche inspirée du volet cognitif génère des règles transparentes basées sur un clustering flou. Le flux de travail est

structuré pour que les prédictions du modèle et les probabilités des règles floues soient fusionnées par un modèle de régression logistique. Cette fusion est un avancement au niveau des deux approches, eu égard qu'elle combine l'analyse causale, la puissance de prédiction et l'intelligibilité des règles.
Assurément, le lien entre ces deux parties reste notre priorité. Le système ne fournit pas une prédiction abstraite. Il la soutient par une règle concrète. Par exemple, si le système prédit un risque élevé, il l'associe à la règle « Si le taux de [calcium] est élevé et la [température] est élevée, alors le risque de mortalité est de X % ». Nous assurons que cette règle est conforme à la prédiction du modèle. Nous évaluons la fidélité et la cohérence. Il est primordial d'utiliser ces indicateurs pour confirmer la justesse de l'interprétation. Elles aident à l'acceptation clinique du système.

8.4 Nos protocoles expérimentaux

Nous illustrons les protocoles expérimentaux par l'exemple d'organisation de trois ensembles de données, en fonction des prescriptions d'antibiotiques. Sur le plan informatique, les protocoles intègrent des calculs d'*ATE*, la sélection de variables et la calibration. Sur le plan cognitif, ils génèrent des directives précises accompagnant les prédictions. Pourquoi est-il utile d'avoir cette structuration ? Parce qu'elle établit un lien entre la rigueur statistique et la clarté clinique. Comment cela se manifeste-t-il ? Par une analyse détaillée des conséquences de cause à effet. Par exemple, l'*ATE* ajusté indique qu'un antibiotique diminue la mortalité chez les patients ayant un faible taux de glucose, mais n'a aucun impact sur les patients avec un taux élevé.

Subséquemment, nos expérimentations se concentrent sur la prédiction de la mortalité à 72 heures pour les patients en choc septique. Autant, les modèles sont entraînés et testés sur trois jeux de données :

1. A=1 : Patients ayant reçu un antibiotique unique.
2. A=0 : Ceux qui ont reçu plusieurs antibiotiques.
3. A=Tous : L'ensemble complet.

En plus, nos protocoles sont établis pour divers objectifs spécifiques. Dans le cadre de la première expérience, portant sur l'analyse causale, nous déterminons, de ce fait, l'*ATE* pour chaque variable confondante. Par la suite, nous employons un algorithme afin de décomposer ces variables en nuances, calculons un *ATE* ajusté et pondéré pour chacune d'elles. Également, un second protocole se rapportant à

la sélection des variables. Nous utilisons, par ailleurs, une technique bayésienne de permutation pour identifier les sept variables les plus significatives pour la prédiction. Subséquemment, un autre protocole examine la capacité d'interprétation du *LightGBM*. Il s'appuie autant sur l'incorporation des scores d'appartenance floue que ceux de fiabilité dans le processus d'entraînement, d'où il est évalué la concordance entre les prédictions et les règles créées.

Aussi, notre approche protocolaire est conçue pour l'intégration de la partie informatique et de celle cognitive. Pour le volet causal, le protocole débute par l'application de l'algorithme *ATE*. Cet algorithme évalue l'effet causal d'un traitement sur un groupe de patients. Mais le protocole ne s'arrête pas là. Il utilise ensuite le modèle *BGM* pour segmenter chaque variable, comme le taux de glucose, en trois profils flous. Pourquoi cette étape est-elle un avancement au niveau des deux parties ? Elle permet de calculer un *ATE* nuancé pour chaque profil. Par exemple, nous pouvons déterminer si le traitement a un effet sur les patients avec un taux de glucose faible. La partie cognitive est assistée par cette transformation. Elle convertit un chiffre en une notion clinique intelligible.

Également, le protocole de sélection des variables est une autre brique de notre flux de travail. Nous employons une technique de permutation bayésienne. Elle évalue l'importance de chaque variable pour la prédiction. L'objectif est de trouver les variables qui ont le plus grand impact sur la mortalité. Nous gardons les sept plus significatives. Pourquoi pas la totalité des variables ? C'est un compromis informatique réduisant la dimensionnalité et prévient le surapprentissage. C'est aussi une décision cognitive rendant l'interprétation des règles plus simples pour un clinicien. Les règles générées sont plus courtes, plus faciles à retenir et à valider.

En définitive, le protocole d'interprétation relie les deux volets. Le flux de travail intègre les scores d'appartenance floue et ceux de fiabilité dans le modèle de régression logistique. Cette intégration crée le lien entre les règles et le modèle de prédiction. Nous évaluons la concordance entre la prédiction et l'interprétation de la règle. C'est une mesure de la fidélité. Cette mesure garantit que le modèle ne donne pas de prédiction contradictoire avec sa propre règle. Si un désaccord est présent, le système déclenche une alerte. Il y a un avancement pour les deux parties parce que cela signale un besoin d'analyse par un clinicien.

8.5 Notre présentation des outils de travail

Nous illustrons la présentation des outils de travail par l'exemple d'utilisation d'une série d'outils Python intégrés pour enrichir la présentation des outils de travail. En matière de programmation, *Scikit-learn*, *LightGBM* et *BGMM* garantissent la cohérence et la reproductibilité. Sur le plan cognitif, les résultats sont interprétés en directives nettes et ajustées. Pourquoi a-t-on besoin de cette association ? Par ce qu'elle assure la robustesse technique et l'utilité pratique. Comment cela se met-il en pratique ? En utilisant la régression isotonique pour ajuster les probabilités. Par exemple, une prédiction ajustée à 80% signifie en réalité qu'elle correspond à 8 cas sur 10 observés, ce qui renforce la confiance d'un clinicien.

De la sorte, nous faisons appel à divers outils Python pour mener nos expérimentations. Pour la voie causale, une classe, sur mesure intégrant des poids d'appartenance floue est utilisée pour effectuer les calculs d'*ATE*. Nous utilisons, de surcroît, le *BGM* pour effectuer le clustering flou. L'implémentation de la permutation bayésienne se fait, également, grâce à l'utilisation des fonctions de la bibliothèque *Scikit-learn* pour le calcul des scores. Nous utilisons, ainsi, la bibliothèque *LightGBM* comme modèle de prédiction. Nous employons, du coup, la régression isotonique de *Scikit-learn* pour calibrer les probabilités.

Aussi, la mise en œuvre des algorithmes repose sur un ensemble d'outils informatiques robustes et bien ancrés. L'*ATE* est une illustration de la personnalisation. Elle est spécifiquement élaborée pour notre projet. Elle prend en charge les poids d'appartenance floue, un perfectionnement par rapport à l'*ATE* standard. Elle permet ainsi d'ajuster l'estimation de l'effet causal en fonction de l'appartenance des patients à des groupes. Par la suite, le *BGM* est la brique logicielle pour notre volet cognitif. Il sert à la segmentation des variables et crée des profils de patients.

En fin de compte, un lien entre les différentes parties du flux de travail est l'utilisation des bibliothèques standards, comme *Scikit-learn*. Par exemple, le *BGM* et la régression isotonique proviennent de cette librairie, ce qui garantit la cohérence et la compatibilité entre les différents modules. Pour la partie interprétabilité, ce flux inclut l'intégration de la régression isotonique à la fin du pipeline. Cette intégration est décisive pour ajuster les probabilités du modèle. Pourquoi cela a-t-il de l'importance ? Dans une perspective clinique, une prédiction de 80% devrait indiquer que le patient a effectivement 80% de probabilité d'avoir un résultat spécifique. L'étalonnage confère au système sa fiabilité. Cet instrument est capital pour la valorisation (par exemple, la confiance d'un clinicien).

8.6 Nos plateformes de développement

Nous illustrons les plateformes de développement par l'exemple de réalisation des expérimentations dans un cadre sécurisé. En termes informatiques, les pipelines garantissent la gestion des modules ainsi que la reproductibilité. Sur le plan cognitif, les plateformes génèrent des directives intelligibles et émettent des alertes lorsqu'elles détectent une incohérence. Pour quelle raison cette infrastructure est-elle incontournable ? Parce qu'elle sauvegarde les données et encourage l'adoption en milieu clinique. Quelle est la signification de cela ? Grâce à un simulateur offrant aux cliniciens l'opportunité d'éprouver leurs hypothèses sur des situations concrètes. Par exemple, changer la prescription d'un antibiotique et observer son effet sur la mortalité prédite.

Subséquemment, nous réalisons toutes nos expériences dans un cadre de développement sécurisé. Nous avons eu, de même, la possibilité d'exploiter les capacités de calcul indispensables tout en assurant la protection des données sensibles provenant de *MIMIC-III*. Un pipeline gère, en conséquence, le chargement des modules sur mesure et la gestion des données. Nos plateformes nous permettent, en définitive, d'utiliser des bibliothèques standards et de reproduire nos résultats dans un cadre cohérent.

Au bout du compte, les plateformes assurent la protection des données sensibles tirées de *MIMIC-III*. Elles sont indispensables pour notre processus de travail. Elles supervisent le flux de la gestion des données jusqu'à l'analyse des résultats. Cet arrangement garantit la reproductibilité de nos expérimentations. De plus, l'avancement cognitif est également rendu possible par ces plateformes. Le flux de travail génère des sorties concrètes pour un clinicien. Par exemple, le système produit des règles cliniques. De même, il envoie des alertes en cas d'incohérence entre ces règles et les prédictions. Le lien entre les deux volets se matérialise sur ces plateformes. Le système rend visible le comportement du modèle et justifie ses prédictions. Ce clinicien n'est pas un utilisateur passif. Il peut interagir avec le simulateur causal pour tester ses hypothèses et explorer les scénarios de traitement.

8.7 Le déroulement de nos expérimentations

Nous dépeignons l'illustration du processus des expérimentations par l'exemple de mise en œuvre d'un protocole spécifique. Sur le plan informatique, chaque jeu de données est analysé et des indicateurs de performance sont établis. Sur le plan cognitif, des règles normalisées sont générées et mises en

comparaison avec les prédictions. Pourquoi cette méthode est-elle rigoureuse ? Par ce qu'elle évalue à la fois la précision et la cohérence. Comment cela fonctionne ? Via un module activant une notification lorsque l'écart entre la règle et la prédiction excède 10 %. Par exemple, une prédiction de 85% de taux de mortalité est liée à une règle indiquant 60%. Un avertissement indique cette divergence au professionnel de santé.

De la sorte, les expérimentations sont menées en suivant un protocole détaillé. Chaque sous-ensemble de données est, dès lors, traité par notre pipeline, qui a géré l'entraînement, la validation et l'analyse.

- **Analyse causale :** Nous calculons un *ATE* ajusté pour chacune des trois nuances pour chaque variable confondante.
- **Interprétation :** Le système génère des règles intelligibles. De même, nous utilisons la normalisation des règles pour plus de clarté avec comparaison aux prédictions du modèle.
- **Détection des incohérences :** Nous utilisons un mécanisme d'alerte pour identifier les clusters de patients où l'écart de cohérence entre le modèle et les règles dépassait un seuil préétabli (10%). Cette alerte incite ainsi à une analyse approfondie.

Aussi, le processus expérimental suit un flux de travail organisé, établissant la connexion entre l'aspect informatique et celui du cognitif. La première phase consiste à organiser les trois ensembles de données : les groupes A=1, A=0 et celui de « tous ». Chaque jeu de données est ensuite traité par notre pipeline d'entraînement. Pour la partie informatique, cela signifie que les modèles, tels que *LightGBM*, sont entraînés et les métriques de performance sont calculées.

En parallèle, le processus d'interprétation commence. Le système génère des règles intelligibles pour un clinicien. Ces règles sont une représentation qualitative du comportement du modèle. Elles sont un avancement cognitif, dans la mesure où elles traduisent un processus complexe en un format simple et actionnable. Le flux de travail inclut un module de détection des incohérences. Ce module compare la prédiction du modèle à la règle correspondante. L'écart est mesuré. Si l'écart dépasse un seuil de 10 %, une alerte est déclenchée. Pourquoi cet outil ? Il assiste ce clinicien. Cette alerte l'incite à un examen plus poussé du cas. Il ne s'agit pas de rejeter le modèle, mais d'en saisir les nuances.

Tout bien considéré, le processus d'analyse causale suit un protocole précis. Le système ne fournit pas seulement un *ATE* global. Il calcule un *ATE* ajusté pour chaque nuance de variable. Par exemple, pour

chaque variable, il y a plusieurs nuances. Le système fournit un *ATE* pour chaque groupe de patients. C'est un avancement cognitif favorisant une prise de décision plus fine. Autre exemple, un traitement peut avoir un effet bénéfique pour les patients avec un faible taux de glucose, mais sans effet pour les autres. Cette capacité à fournir une analyse détaillée est nécessaire pour un clinicien.

8.8 La validation de nos expérimentations

Nous illustrons la validation des expérimentations à travers l'exemple des indicateurs de cohérence, de fidélité et de fiabilité. Sur le plan informatique, l'*ECE* évalue la précision des probabilités. Sur le plan cognitif, l'adéquation entre la règle et la prédiction assure une intelligibilité. Pourquoi cette vérification est-elle centrale ? Parce qu'elle indique une valorisation et une confiance. Comment cela se met-il en pratique ? Par des seuils spécifiques provoquant une alerte lorsqu'un écart se produit. Par exemple, une fidélité de 92 % indique que les règles sont représentatives du comportement du modèle et que le praticien peut s'y fier.

De ce fait, nous confirmons nos expérimentations en évaluant, par exemple, la concordance entre les prédictions du modèle et les interprétations des règles. Une incohérence est mesurée si cette différence excède un seuil préétabli (10%), une alerte est activée. Cette notification incite, en conséquence, un clinicien à une évaluation plus poussée des écarts. En outre, nous évaluons la fidélité et la fiabilité en déterminant l'*ECE* avant et après calibration.

Aussi, la validation de nos expérimentations est centrale dans notre flux de travail et représente un pont déterminant entre les aspects informatiques et cognitifs. Nous ne nous limitons pas à évaluer la précision des modèles. Nous évaluons également la qualité de ses interprétations. Cette approche est illustrée par l'évaluation de la cohérence. Elle mesure la différence entre l'estimation du *LightGBM* et la probabilité donnée par la règle floue. Le progrès cognitif consiste à rendre ce lien perceptible pour un clinicien. Par exemple, si le modèle anticipe un risque de mortalité de 85% tandis que la règle associée fixe le chiffre à 60%, nous constatons une divergence de 25%. Une alerte est déclenchée. Pourquoi est-ce important ? Ce clinicien ne se fie pas aveuglément à la prédiction. Il peut examiner ce cas et comprendre l'incohérence.

En plus, la fidélité est une autre mesure de validation. Elle vérifie si les règles générées sont un reflet exact des prédictions du modèle sur un large ensemble de données. C'est une mesure de la partie

informatique qui a un impact direct sur celle cognitive. Si la fidélité est faible, les règles ne sont pas fiables. Le clinicien ne peut pas faire confiance aux interprétations du système. Une fidélité élevée, par exemple 92 %, valide l'efficacité de notre stratégie.

Assurément, la fiabilité est une autre validation élémentaire. Elle est mesurée par l'*ECE*. Cette mesure nous dit si les probabilités prédites sont fiables. Un modèle qui prédit 80% de chances de mortalité devrait avoir une issue de mortalité pour 80% des patients dans son groupe de prédiction. Une *ECE* faible, après notre calibration isotonique est un avancement informatique se traduisant par une confiance accrue du clinicien. C'est un élément clé pour l'intégration d'une IA dans un environnement clinique.

8.9 Les exemples d'application de nos expérimentations réalisées

Nous illustrons les exemples d'application des expérimentations réalisées par des cas de patients atteints de sepsis. Sur le plan informatique, le modèle génère un *ROC-AUC* de 0.88 et une précision moyenne de 92 %. Sur le plan cognitif, il interprète ces résultats en directives applicables. Pourquoi ces exemples sont-ils significatifs ? Par ce qu'ils démontrent l'harmonie entre performance et capacité d'interprétation. Comment cela se manifeste-t-il ? Grâce à un paramètre alpha modifiant l'amplitude des intervalles des règles. Par exemple, une directive « calcium et glucose » liée à un risque spécifique en pourcentage et assistant un praticien dans l'évaluation claire de la condition d'un patient.

Par conséquent, nous orientons nos expérimentations sur la prédiction de la mortalité à trois jours pour les patients souffrant de choc septique. Nous entraînons et testons, subséquemment, nos modèles sur trois jeux de données distincts : A=1, 0, et tous. Un clinicien peut, de même, interpréter les résultats des expérimentations. Également, dans le cadre de l'analyse causale, le système ne se limite pas à fournir un *ATE* global. Il propose, aussi, des *ATE* selon les nuances de la variable. Par exemple, le système pourra démontrer qu'un traitement a un impact considérable pour un niveau « bas » en glucose, alors qu'il est négligeable pour un niveau « haut ». Cela fournit, par conséquent, un degré de détail favorisant une décision plus précise. De ce fait, en ce qui concerne l'interprétabilité, le système a la capacité de produire une règle clinique pour : calcium $\in$ [-74.50, 58.02] ET glucose $\in$ [-62.23, 49.41] ALORS la règle floue est de 4.05%. Ces règles sont lisibles et peuvent être utilisées, en fin de compte, pour saisir le flux du modèle.
Pour la partie qualitative, nous cherchons à obtenir une cohérence et une fidélité rassurant le clinicien. Nous avons des résultats numériques illustrant cela. Par exemple, nous mesurons un *ROC-AUC* de 0.88 sur

l'ensemble des données, un chiffre démontrant une bonne performance prédictive du modèle *LightGBM*. Par ailleurs, la fidélité de nos règles floues est en moyenne de 92 %, ce qui indique que les règles reflètent de manière significative la prédiction du modèle sous-jacent.

Aussi, nous employons divers paramètres influençant les résultats. Dans notre modèle de règles, l'alpha est un paramètre déterminant la largeur des intervalles des règles. Un alpha supérieur produit des règles plus étendues, englobant un plus vaste éventail de patients ; toutefois, elles peuvent être moins exactes. Au contraire, un alpha réduit produit des règles plus spécifiques et précises, bien qu'elles soient moins applicables de manière générale. Nous établissons un alpha de 1.5 afin de rechercher un équilibre optimal entre généralisation et précision.

En plus, une autre décision importante concerne le nombre de clusters. Un nombre trop faible de clusters produit des règles trop générales. Un nombre trop élevé conduit à des règles spécifiques. Nous déterminons le nombre de clusters de manière automatique en nous basant sur la variabilité des données. Ces choix de paramètres constituent l'avancement informatique de notre approche. Ils optimisent la performance prédictive et la clarté des règles.

En somme, ces exemples démontrent un lien fort entre notre partie informatique et celle cognitive. Le système informatique génère une prédiction de risque et des calculs d'*ATE* nuancés. La partie cognitive du système transforme ces sorties en informations actionnables pour un clinicien. Par exemple, la capacité de notre modèle à fournir un *ATE* pour un niveau « bas » de « glucose » est une connaissance. Elle s'oppose au seul apport d'un *ATE* global. Ce détail est utile pour un clinicien cherchant à personnaliser ses décisions. Le flux de travail est conçu pour produire ces résultats. Il commence par la préparation des données, continue avec l'entraînement du modèle et se termine par la production des règles et des analyses causales. Par exemple, une règle sur le calcium et le glucose est le résultat direct de ce flux de travail. Pourquoi est-elle importante ? Elle est une traduction directe du comportement du modèle en termes cliniques clairs. Elle assiste un clinicien en lui fournissant des justifications pour son raisonnement.

8.10 La qualité de nos règles

Nous illustrons la qualité des règles par l’exemple des dérivées des centroïdes flous. Sur le plan informatique, elles sont normalisées et étalonnées. Sur le plan cognitif, elles sont compréhensibles pour

un clinicien. Pourquoi la qualité est-elle décisoire ? Parce qu'elle détermine la confiance dans le système. Comment garantir cela ? Avec un coefficient alpha fixé à 1.5, assurant un équilibre entre précision et généralisation. Par exemple, une règle pour une caractéristique médicale fournie dans son unité d'origine permet à un médecin de confirmer instantanément sa pertinence.

De cette manière, nos expérimentations se concentrent spécialement sur la qualité des règles générées. Ces règles sont établies, conséquemment, sur la base des centroïdes du clustering flou et des probabilités conditionnelles de décès. De surcroît, elles sont normalisées afin d'améliorer leur clarté clinique. Nous appliquons, accessoirement, un coefficient alpha (déterminé à 1,5) pour moduler la largeur des intervalles des règles, et cela dans le but de parvenir à une robustesse du système.

Aussi, la qualité des règles est un aspect central de notre démarche expérimentale. Elle représente un lien basique entre notre partie informatique et celle cognitive. Notre flux de travail génère des règles qui sont à la fois précises et intelligibles. Nous créons les règles en nous basant sur les centroïdes du clustering flou. Ces centroïdes représentent les profils types de patients. La normalisation des règles est un avancement informatique se traduisant par un gain cognitif. Pourquoi la normalisation est-elle importante ? Elle permet de présenter les règles avec les unités cliniques d'origine, par exemple les valeurs de *bilirubin* en unités intelligibles pour un clinicien, ce qui facilite leur interprétation.

En définitive, ce coefficient alpha est un paramètre important de notre partie informatique. Ce coefficient détermine la largeur des intervalles de nos règles floues. Ce choix est justifié. Un alpha trop grand rendrait les règles trop générales. Elles couvriraient une large partie des patients, mais les règles seraient moins précises. Un alpha trop petit produirait des règles trop spécifiques à certains cas, rendant le système moins robuste. Notre alpha assure un compromis. Il garantit la robustesse du système. Il montre aussi un avancement dans l'intégration des paramètres informatiques pour une meilleure interprétabilité cognitive. La qualité de ces règles assiste le clinicien. Elle lui fournit un outil fiable pour valider la logique du modèle.

8.11 Nos mesures et nos métriques

Nous illustrons les évaluations des modèles par l'exemple des lectures de plusieurs indicateurs en même temps. Par exemple, dans le plan informatique, le *ROC-AUC* et la validation croisée assurent la robustesse.

Sur le plan cognitif, la cohérence et l'*ATE* flou procurent une clarté clinique. Pourquoi est-il nécessaire de fusionner plusieurs mesures ? Par ce qu'elles allient exactitude technique et pertinence pratique. Comment cela fonctionne ? Par exemple, en ajoutant un calcul d'*ECE* validant la fiabilité des probabilités. Ainsi, un risque indiqué à 75% correspond réellement à 75% des cas observés, ce qui confirme son utilisation clinique.

De cette façon, nous formulons des hypothèses sur la capacité de nos modèles à produire des prédictions précises, à offrir une interprétation pertinente et à identifier des nuances causales. Pour les évaluer, nous utilisons, au bout du compte, un large éventail de métriques.

- **Métriques de performance :** Nous mesurons, par exemple, l'exactitude (*Accuracy*), l'équilibre de l'exactitude (*Balanced Accuracy*), le rappel (*Recall*) et le *ROC-AUC*. Nous utilisons, autant, une validation croisée à cinq plis pour minimiser les biais et accroître la généralisation des modèles.

- **Métriques de fiabilité :** La fiabilité des probabilités est évaluée par l'*ECE*, mesurée avant et après la calibration isotonique. Autant, un *ECE* faible indique que les probabilités prédites sont fiables.

- **Métriques d'interprétabilité et de causalité :** Nous mesurons la cohérence, qui est l'écart entre les prédictions du modèle et les règles floues associées. Une cohérence élevée est, ainsi, synonyme d'une bonne interprétabilité. Pour la causalité, nous calculons, pareillement, l'*ATE* pour chaque segment flou de la variable confondante, ce qui nous permet de mettre en évidence des nuances.

Aussi, les mesures et métriques sont des outils substantiels pour valider le flux de travail et le lien entre nos deux volets, informatique et cognitif. Du côté informatique, nous utilisons les métriques de performance standards, comme le *ROC-AUC*. L'utilisation de la validation croisée à cinq plis est un avancement informatique. Elle minimise le biais de sélection de l'ensemble de données et garantit la généralisation des modèles. Le *ROC-AUC* fournit une indication précise de la compétence du modèle à séparer efficacement les patients survivants de ceux qui ne le sont pas.
En plus, une autre mesure impérative est la fiabilité des probabilités. Elle est évaluée par l'*ECE*. Pourquoi le mesurons-nous ? Un faible *ECE* est une preuve que les probabilités prédites par le modèle sont fiables. Dans le contexte médical, un modèle qui prédit 75% de risque de mortalité doit avoir une mortalité observée de 75% pour les patients de ce groupe. C'est pressant pour la valorisation (ex. confiance d'un

clinicien). Cette fiabilité est un avancement cognitif permettant l'utilisation du système dans des contextes réels.

Assurément, le lien entre l'informatique et la cognition est mesuré par la cohérence. Cette dernière est l'écart entre les prédictions du modèle et la probabilité des règles floues. Une cohérence élevée est synonyme d'une bonne interprétabilité. La cohérence est le pont entre la précision de l'IA et le discernement humain. Pour la causalité, le calcul de l'*ATE* pour chaque segment flou de la variable confondante est une amélioration. Cela nous permet de mettre en évidence des nuances que l'*ATE* global ne montre pas. Ces métriques nous donnent une vision complète du système. Elles ne se concentrent pas seulement sur la performance, mais aussi sur la qualité et la fiabilité.

8.12 Nos justifications

Nous illustrons nos motivations pour chaque décision par l'exemple du besoin précis. Sur le plan informatique, la segmentation floue permet de diminuer la dimensionnalité et d'ajuster les probabilités. Sur le plan cognitif, elle illustre l'incertitude médicale de façon claire. Pourquoi est-il indispensable de fournir cette interprétation ? Parce qu'elle crée un lien entre la computation et la tentative de représentation du raisonnement. Comment cela se manifeste-t-il ? Grâce à la validation croisée améliorant la robustesse et à l'ajustement des probabilités qui accroît la crédibilité clinique. Par exemple, l'examen de cohérence indique que le modèle ne génère aucune contradiction avec ses propres règles.

C'est pourquoi la segmentation des données permet d'évaluer les performances et la capacité d'interprétation dans des contextes cliniques spécifiques. Cela nous donne, en fin de compte, la possibilité d'analyser l'impact potentiel du traitement sur différentes populations de patients. Également, chaque décision expérimentale est motivée par un but spécifique. L'intégration de modèles, par exemple, répond à la quête d'une recherche en IA qui soit à la fois précise dans ses prédictions et transparente dans son interprétation. L'utilisation du flou pour la segmentation des variables s'explique, de la sorte, par le besoin de représenter l'incertitude intrinsèque aux données médicales. L'ajustement des probabilités est décisif pour la crédibilité du clinicien. L'évaluation de la cohérence sert, ainsi, à mesurer la performance des résultats produits par le modèle.

De même, l'utilisation de multiples métriques nous offre une vision complète de la performance. La validation croisée garantit la robustesse des résultats. Tout bien considéré, les mesures de l'*ECE* et de la cohérence sont déterminantes pour notre objectif d'IA valorisante et transparente. En plus, chaque décision dans notre approche est justifiée par un besoin informatique ou cognitif. Leur avancement est un lien entre ces deux volets. Du côté informatique, la segmentation des données permet de concentrer les modèles sur des groupes de patients spécifiques, améliorant ainsi la précision des prédictions. Nous analysons l'impact d'un traitement sur différentes populations de patients. L'utilisation de multiples métriques, comme le *ROC-AUC* et l'*ECE* est aussi une justification. Ces métriques nous offrent une vue complète de la performance. La validation croisée garantit la robustesse des résultats.

Du côté cognitif, l'utilisation du flou pour la segmentation est une justification primordiale. Les données médicales sont incertaines. Le flou représente cette incertitude et permet de créer des profils de patients qui ont un sens clinique. La crédibilité d'un clinicien dépend de l'ajustement des probabilités. Une prédiction bien calibrée est plus fiable pour la prise de décision. Le flux de travail est justifié par sa capacité à générer des résultats fiables. Il fournit des règles et des interprétations pour ce clinicien. De même, l'évaluation de la cohérence est une justification majeure. Elle mesure la performance des résultats produits par le modèle. Elle sert à vérifier si le modèle se comporte de manière intelligible. C'est le lien entre les prédictions numériques et le flux qualitatif. Le système assiste ce praticien en lui fournissant une vision complète de la performance. Cela inclut la précision du modèle et la qualité de son interprétation.

8.13 Nos limites

Nous mettons en évidence deux aspects pour la reconnaissance des limites. Sur le plan informatique, la suppression des données manquantes crée un biais de sélection. Des facteurs tels que le nombre de regroupements ou l'alpha ont un impact sur la qualité des règles d'un point de vue cognitif. Pour ces deux aspects, quelle est l'importance de cette reconnaissance ? Elle définit la portée véritable des résultats. Comment peut-on les diminuer ? En employant le flou pour illustrer l'incertitude et en évaluant la cohérence et la fidélité. Par exemple, bien qu'une sélection de paramètres réduise la précision, le dispositif délivre une indication avertissant le praticien de cette contrainte.

Également, la reconnaissance des limites est une démarche scientifique rigoureuse. Irrémédiablement, cela permet de situer nos résultats dans notre contexte sensible et de fournir une base pour des recherches futures. Le principal défi, du point de vue de la partie informatique, est l'enrichissement de l'interprétabilité avec de la causalité. Même avec nos algorithmes, il est difficile d'établir l'interprétabilité et des liens de cause à effet sans l'aide d'un expert en choc septique. L'élimination des lignes avec des valeurs manquantes dans le jeu de données *MIMIC-III* peut aussi créer un biais de sélection. C'est une contrainte technologique influençant la validité de nos résultats.

En fin de compte, le progrès cognitif rencontre plusieurs types de restrictions. Le nombre de clusters ou le paramètre alpha exercent une influence sur la qualité des règles générées. Un mauvais choix de ces paramètres peut rendre les règles moins précises ou trop complexes, ce qui diminue leur utilité pour un clinicien. L'incertitude est une limite que nous reconnaissons. C'est pourquoi nous utilisons le flou et nous fournissons les mesures de cohérence et de fidélité. Ces mesures sont un lien entre les limites informatiques du modèle et le besoin cognitif de l'expert de comprendre les risques. Notre flux de travail est conçu pour minimiser ces limites, mais elles ne peuvent pas être complètement éliminées.

8.14 Le tableau de synthèse de nos expérimentations

Catégorie	Description
Objectif cognitif et informatique	Développer des IA symboliques, sophistiquées et hybrides capables de prédire avec précision la mortalité à 72 heures en offrant une interprétation et une analyse causale destinées à des cliniciens. L'objectif est de concilier la triade de précision prédictive, causalité, et valorisation.
Approche	Stratégie intégrant une approche informatique (analyse causale *ATE*, techniques de permutation bayésienne, *LightGBM*…) et celle cognitive (génération de règles transparentes via le clustering flou, interprétation des prédictions).
Source de données	Données anonymisées de *MIMIC-III*. Elles proviennent des patients atteints de sepsis et incluent des signes vitaux, des valeurs de laboratoire et des données

	démographiques. Une demande d'autorisation et une formation éthique sont nécessaires pour y accéder.
Choix et éthique de la BD	Pourquoi *MIMIC-III* ? Cette BD est une ressource reconnue pour le choc septique. Elle est réaliste et anonymisée. C'est un cas d'étude pertinent pour le développement de nos algorithmes. Justifications valorisantes (ex. éthiques) dans le monde réel : Une IA mal validée peut engendrer des préjugés et exacerber les disparités de soins. L'absence de valorisation (ex. transparence) des algorithmes dans le monde réel suscite la méfiance des cliniciens et du grand public. Nos expérimentations visent à obtenir cette valorisation et la traçabilité de notre flux de travail. Procédure d'accès : l'obtention des données a requis une démarche éthique. Nous avons rempli un formulaire, fourni les motifs d'accès, et signé un accord d'utilisation des données. La demande a été approuvée par un comité d'examen. Sélection des variables : nous nous focalisons sur 26 variables cliniques pertinentes. C'est une décision informatique pour éviter la malédiction de la dimension et le surapprentissage. C'est aussi une décision cognitive pour simplifier les règles d'interprétation pour un clinicien, ce qui facilite l'adoption du système.
Variable cible	Mortalité à 72 heures (variable binaire : y=1 pour la mortalité, y=0 pour la survie).
Variables d'entrée	26 variables cliniques choisies pour leur pertinence et leur familiarité pour les cliniciens.
Flux de travail (Pipeline)	Flux structuré en différentes phases : 1. Gestion des données (téléchargement, nettoyage, imputation), 2. Modélisation (entraînement des modèles), 3. Analyse et interprétation (génération de règles, calcul de l'*ATE*).
Protocole pour l'utilisateur	Le flux de travail est structuré pour assister un utilisateur voulant reproduire les modèles. Début d'utilisation : il commence par télécharger un fichier de données patient. Nettoyage des données : un module de nettoyage substitue les valeurs manquantes. Le système dénote le niveau de complétude des données. Il doit

	comprendre les choix d'imputation avant de continuer. L'entrée est le fichier brut, la sortie est le jeu de données prétraité et un rapport de qualité des données.
Protocoles expérimentaux	1. Analyse de causalité : Estimation de l'*ATE* concernant le traitement antibiotique (A) et son impact sur la mortalité (y). L'*ATE* est ajusté et pondéré pour chaque subtilité (segment flou) des variables. 2. Choix des variables : recours à une permutation bayésienne pour déterminer les 7 variables les plus pertinentes. 3. Interprétation : génération de règles floues lisibles. Les scores d'appartenance floue et ceux de fiabilité sont intégrés au modèle de prédiction.
Outils et technologies	Python avec ses bibliothèques telles que *Scikit-learn* (permutation bayésienne, régression isotonique), *LightGBM* pour la prédiction, et des classes personnalisées pour l'*ATE* et le clustering flou (*BGM*).
Mesures et métriques	Métriques de fiabilité : *ECE* pour mesurer la fiabilité des probabilités anticipées. Indicateurs d'interprétabilité : cohérence (différence entre la prédiction et la règle associée) et fidélité (représentation des prédictions par les règles). La validation va au-delà de la simple précision prédictive. Elle intègre l'évaluation de la cohérence et de la fidélité afin de garantir que les interprétations sont en adéquation avec les prédictions. Métriques de performance : *ROC-AUC* (mesure clé), exactitude, rappel, etc., avec une validation croisée à cinq plis pour garantir la robustesse. Métriques de fiabilité : *ECE* pour évaluer la fiabilité des probabilités prédites. Métriques d'interprétabilité : cohérence et fidélité.
Validation	La validation ne se limite pas à la précision prédictive. Elle inclut la mesure de la cohérence et de la fidélité pour s'assurer que les interprétations correspondent bien aux prédictions. Un mécanisme d'alerte est déclenché si l'écart de cohérence dépasse 10%.

Côté qualitatif (résultats et paramètres)	Ce qu'on doit obtenir : Nous cherchons la cohérence et la fidélité qui rassurent un clinicien. Paramètres et leur influence : • Coefficient alpha : ce paramètre, qui est fixé à 1.5, modifie la largeur des intervalles des règles. Un alpha élevé favorise des règles générales, tandis qu'un alpha faible tend à favoriser des règles spécifiques. • Quantité de regroupements : ce réglage est défini automatiquement en se basant sur la variabilité des données. Exemples de résultats quantitatifs : • *ROC-AUC* : Un score de 0,88 sur l'ensemble des données indique une bonne précision prédictive. • Fidélité des règles floues : Leur fidélité est en moyenne de 92 %, ce qui indique que les règles reflètent de manière significative la prédiction du modèle sous-jacent.
Limites reconnues	La complexité de la causalité, la difficulté d'établir des liens de cause à effet sans l'aide d'un clinicien, le biais de sélection lié à l'élimination des données manquantes, et l'influence des hyperparamètres (comme l'alpha et le nombre de clusters) sur la qualité des règles.
Conclusion	L'expérimentation valide une approche sophistiquée et hybride produisant un système d'IA précis, causal, intelligible, et valorisant pour des cliniciens. L'intégration des volets informatiques et cognitifs et leur avancement permettent d'assister à la prise de décision dans un contexte médical sensible.

Tableau 5 : Synthèse des expérimentations.

8.15 Conclusion

Intégralement, nous illustrons les expérimentations par l'exemple d'une présentation de boucle validant la méthodologie et l'implémentation dans un contexte appliqué. Informatiquement, elles présentent l'interprétabilité et l'analyse causale. Cognitivement, elles évaluent l'intelligibilité, la cohérence, la valorisation et l'acceptation par des experts. Pourquoi cette boucle est-elle homogène ? Parce qu'elle garantit que le système n'est pas seulement exact en théorie, mais aussi pertinent dans la pratique. Comment cela s'exprime-t-il concrètement ? Par la convergence entre mesures quantitatives

informatiques et appréciations qualitatives cognitives. Par exemple, un modèle performant, mais opaque serait rejeté alors qu'un modèle performant et intelligible est intégré dans le processus décisionnel.

En procédant ainsi, les expérimentations réunissent plusieurs éléments. La partie informatique a produit des architectures robustes avec des modèles performants et un pipeline de traitement des données. La partie cognitive est satisfaite par la création de règles intelligibles et par la traduction des concepts techniques en termes cliniquement pertinents. Le flux de travail que nous concevons du nettoyage de données à l'analyse des résultats assiste un clinicien à chaque étape. Le lien entre ces parties est assuré par la mesure de la cohérence entre les règles et les prédictions ainsi que par le simulateur causal permettant au clinicien de visualiser l'impact d'une intervention. Notre avancement sur ces deux parties réside dans la capacité de nos systèmes à fournir des prédictions précises accompagnées d'une interprétation fidèle, ce qui est requis dans un domaine sensible. De futures recherches pourraient s'attacher à automatiser l'optimisation des hyperparamètres comme le coefficient alpha. L'intégration de données dynamiques (en temps réel), telle que la série chronologique des signes vitaux, pourrait également optimiser la prédiction et l'analyse.

Tout bien considéré, nous exposons ces expérimentations visant à confirmer notre méthodologie ainsi que notre stratégie d'implémentation. Les protocoles que nous instaurons cherchent, de surcroît, à mettre en évidence les aptitudes de nos modèles en matière d'analyse causale, d'importances de variables et d'interprétabilité. L'analyse détaillée des résultats nous a dès lors donné la possibilité de valider nos hypothèses et de démontrer la pertinence de nos expérimentations dans le contexte clinique du choc septique. Ces résultats confirment, par conséquent, que nous pouvons construire un système efficace, transparent et intelligible. Nos expérimentations sont élaborées, en aboutissement, pour fournir une IA qui sert comme instrument valorisant pour des cliniciens.

CHAPITRE 9 :

COMBINAISON DE LA PROBABILITÉ FLOUE AVEC DES MODÈLES ARBORESCENT ET CAUSAL

9.1 Introduction

Ce chapitre examine l'usage d'un système hybride qui combine la *PFL* et un modèle arborescent afin d'accroître l'intellection des prédictions dans le cadre du choc septique. La visée est de surpasser les modèles opaques en offrant des interprétations précises et exploitables, tout en préservant une capacité de prédiction robuste. Le but est d'évaluer la capacité du système à essayer de simuler le processus de pensée lorsqu'il est confronté à l'incertitude et à l'imprécision, facteurs déterminants dans un contexte sensible.

9.2 La présentation des données collectées

Les données utilisées proviennent de *MIMIC-III*, spécifiquement des enregistrements de patients atteints de choc septique. Elles proviennent d'un ensemble d'événements cliniques transformés en un format adapté à la modélisation. Les caractéristiques incluent des paramètres physiologiques variés. Un élément clé est la variable 'A', qui représente le traitement antibiotique (1 pour antibiotique unique, 0 pour antibiotiques multiples). La variable cible 'y' indique la mortalité à 72 heures (1 pour la mortalité, 0 pour la survie). L'ensemble de données d'entraînement est composé de milliers d'observations. Le nombre total de caractéristiques sélectionnées, excluant la variable 'A' pour l'entraînement des modèles est de 26. Pour l'analyse causale, l'ensemble de données initial contenait 924 843 lignes et 28 colonnes, incluant des identifiants patients, des mesures de laboratoire et des observations physiologiques.

9.3 Les traitements de données

Pour l'interprétabilité, les données subissent plusieurs étapes de prétraitement pour assurer leur qualité et leur pertinence pour la modélisation. Initialement, les valeurs manquantes dans la variable 'A' sont

imputées à l'aide du modèle *LightGBM* entraîné sur les autres caractéristiques disponibles. Sur le total des valeurs manquantes, toutes sont imputées, entraînant en une complétude de cette variable. Les données sont ensuite divisées en trois ensembles d'entraînement, de validation et de tests, pour maintenir la distribution de la variable cible. Cette division aboutit à des proportions d'observations pour ces trois ensembles avec des distributions de mortalité cohérentes entre eux. Après la division, les caractéristiques sont imputées avec la moyenne et normalisées pour assurer une échelle cohérente entre les variables. Une étape significative est le calcul et le stockage des valeurs réelles min/max pour chaque caractéristique numérique (sauf 'A'), ce qui est standard pour la dénormalisation des règles interprétables. De plus, des poids d'instance sont calculés pour chaque observation en fonction du nombre d'instances par patient afin de gérer les déséquilibres potentiels liés aux patients ayant des enregistrements plus fréquents.

Pour l'analyse causale, les données ont été soumises à plusieurs phases de prétraitement minutieux afin d'assurer leur fiabilité et leur pertinence pour l'analyse :

- **Nettoyage des valeurs manquantes :** La première phase a impliqué l'identification et la gestion des données manquantes. Une première analyse a mis en évidence des taux importants de données manquantes pour certaines variables, notamment '*crp*' (85.04%) et '*cvp*' (66.97%). Pour assurer la précision des analyses causales et prévenir des imputations complexes susceptibles d'introduire des biais, les lignes comportant au moins une donnée manquante ont été éliminées. Cette action a diminué l'ensemble de données de 924 843 à 12 780 entrées. Il faut souligner que cette méthode est réaliste, mais pourrait fausser l'échantillon final si les données absentes ne sont pas aléatoires.

- **Conversion des types de données :** La colonne 'deathtime' a été transformée en format numérique afin de faciliter les calculs du 'time to death'. La variable 'A' a aussi été transformée en un entier afin d'être utilisée dans les modèles.

- **Détection des atypiques :** Nous avons utilisé deux techniques pour identifier les valeurs aberrantes :

Méthode IQR (*Interquartile Range*) : Cette technique a repéré un nombre variable d'anomalies pour chaque paramètre, par exemple, 301 pour 'excès de base artériel' (2.36%) et 1824 pour 'got (asat)' (14.27%).

Isolation Forest : Cette méthode qui repose sur le *ML* a repéré un taux d'anomalies autour de 4,5% à 5% pour la majorité des variables. Par exemple, pour 'excès de base artériel', 565 valeurs aberrantes ont été détectées (soit 4.42%). Cette distinction dans la détection met en évidence le caractère complémentaire des deux techniques.

- **Préparation pour l'analyse causale :** La variable '*time to death*' a été calculée en soustrayant '*charttime*' de '*deathtime*'. La variable cible 'y' a été créée, indiquant la mortalité dans les 72 heures. Le jeu de données nettoyé a été sauvegardé sous NoMissingData_cleaned.csv pour les analyses ultérieures.

9.4 Les résultats bruts

Pour l'interprétabilité, le modèle *LightGBM* a été entraîné et examiné sur plusieurs sous-groupes de données (A=0, A=1 et A=Tous). Les résultats bruts de ces modèles, avant toute étape de calibration, sont consignés. Par exemple, pour le sous-ensemble 'A=Tous', le modèle arborescent obtient un *ROC-AUC* de 80.01% accompagné d'une *ECE* brute de 0.0808. Pour 'A=0', le score *ROC-AUC* est de 82,85% et l'*ECE* brut se monte à 0,0935. Finalement, pour 'A=1', le *ROC-AUC* s'élève à 83,82% et l'*ECE* brut est de 0,0838. Ces mesures non traitées constituent une base de référence pour juger l'efficacité des phases suivantes de calibration dont l'objectif est de synchroniser la confiance du modèle avec ses résultats réels.

Les analyses initiales en causalité ont fourni plusieurs résultats bruts :

Comptage des enregistrements : À la suite du nettoyage des données, le jeu de données définitif (SansDonnéesManquantes) comprend 12 780 entrées. Sur l'ensemble de ces données, 9 658 patients ont survécu (y=0) tandis que 3 122 sont décédés (y=1). En termes de traitement, 11 468 patients ont été administrés par plusieurs antibiotiques (A=0), tandis que 1 312 ont bénéficié d'un seul antibiotique (A=1). Il est central de prendre en compte ces déséquilibres de classe et de traitement pour les analyses causales.

• **Corrélation entre traitement et issue :** Le lien brut (*Pearson* et *Spearman*) entre l'intervention (A) et le résultat (y) est de 0,0903. Cette corrélation positive faible indique une association modeste entre l'administration d'un antibiotique unique et le taux de mortalité, sans toutefois établir un lien de causalité sans prise en compte des facteurs confondants.

• **Corrélations des confondants :** Nous avons noté diverses corrélations entre les variables confondantes et l'intervention/l'issue. Par exemple, le '*bilirubin*' a une corrélation de *Pearson* de 0.3172 avec l'intervention et de 0.1631 avec le résultat, suggérant son éventuel rôle en tant que facteur de confusion. '*crp*' présente une corrélation de *Pearson* de -0.2111 avec le traitement et de 0.0484 avec l'issue. Ces fluctuations mettent en évidence la complexité des interactions.

• ***ATE* manuels :** Les *ATE* ont fluctué selon chaque variable confondante calculée manuellement. À titre d'exemple, l'*ATE* manuel pour 'excès de base artériel' s'élève à 0.1007, pour '*pCO2 arterial*', il est de 0.1173 et enfin pour '*crp*', la valeur atteint 0.1489. Ces chiffres illustrent l'effet causal moyen non corrigé du traitement sur le résultat, en prenant en compte la variable de confusion. Les effets moyens pour les groupes qui ont reçu un seul antibiotique et ceux qui n'en ont pas reçu (mais plusieurs autres) sont, respectivement, de 0.3590 et 0.2312.

• ***NFATE* :** Les scores du *NFATE* ont été déterminés pour chaque variable de confusion. Selon le classement des variables effectué par *NFATE*, 'crp' arrive en première position avec 0.15, suivi de près par 'arterial ph' et 'temperature', tous deux à 0.14. L'objectif du *NFATE* est de fournir une évaluation standardisée de l'effet causal en intégrant la logique floue.

9.5 La combinaison des résultats partiels

Pour l'interprétabilité, les probabilités brutes produites par les modèles *LightGBM* subissent une calibration après l'entraînement, en particulier par le biais de la régression isotonique. Cette phase est majeure puisqu'elle permet de réajuster les probabilités de sortie du modèle pour assurer leur fiabilité, c'est-à-dire que la probabilité estimée doit être en adéquation avec la fréquence véritable de l'événement. On évalue l'efficacité de cette calibration en se basant sur la diminution de l'*ECE*. Pour chaque sous-groupe, l'*ECE* calibré affiche une valeur de 0, témoignant d'une nette amélioration par rapport aux *ECE* non calibrés.

Par exemple, pour 'A=Tous', on observe une amélioration de l'*ECE* atteignant 0.0808, pour 'A=0', elle est de 0.0935, et pour 'A=1', elle s'établit à 0.0838.

De plus, les algorithmes intègrent *LIME* et *SHAP* afin d'expliquer les prédictions produites par le modèle arborescent. Par exemple, *LIME* offre la possibilité de comprendre comment une prédiction particulière est réalisée pour un patient spécifique en élaborant un modèle linéaire basique et explicable autour de cette instance. Ce qui apporte de la transparence en délivrant des renseignements sur les attributs les plus déterminants dans une prise de décision locale.

Pour l'effet causal, les résultats non filtrés sont fusionnés pour créer une représentation plus unifiée de cet effet.

- **Lien entre corrélation et causalité :** Les corrélations brutes entre la thérapie et le résultat, comparées aux *ATE* manuels et modifiés, mettent en évidence l'importance de l'inférence causale. Les variables dont la corrélation avec le traitement et l'issue est significative sont confirmées comme des confondants majeurs, nécessitant un ajustement pour produire des effets causaux valides.

- **Importance des éléments atypiques dans l'analyse :** Les divergences dans l'identification des valeurs anormales entre la méthode *IQR* et *l'Isolation Forest* indiquent que le caractère des valeurs aberrantes fluctue en fonction des variables. Bien que ces valeurs aient été éliminées de la version épurée des données, elles soulignent la variabilité et la disparité des données, ce qui justifie le recours à des méthodes robustes et floues pour traiter l'incertitude lors de l'analyse causale.

- **Effet des facteurs de confusion sur l'*ATE* :** L'examen comparatif des *ATE* manuels avec les *NFATE* ou les *ATE-PFL* ajustés (détaillés dans la section résultats) met en lumière l'influence de chaque facteur de confusion sur le lien causal. Les variables présentant un *NFATE* élevé sont celles qui influencent le plus l'impact du traitement sur le résultat, lorsqu'on prend en compte leur répartition floue.

9.6 La justification

Le principal moteur de notre démarche est le besoin d'élaborer des IA fiables dans des domaines critiques, tels que la santé. Les systèmes précis en prédiction ne suffisent plus ; il est immanent d'avoir une connaissance approfondie des processus de décision pour garantir leur acceptation critique et leur valorisation. Le choc septique, en raison de sa complexité multifactorielle et de ses implications vitales, constitue un bon exemple d'étude pour un tel système. L'implémentation de la *PFL* permet de saisir l'incertitude et l'inexactitude inhérentes aux données sensibles, tout en produisant des règles de risque intelligibles, semblables à la logique clinique humaine. L'association avec *LightGBM*, qui est intéressante pour ses performances, offre une recherche d'équilibre entre la puissance prédictive et l'interprétabilité.

Aussi, des seuils cliniques fixes (Faible : <10%, Moyen : 10-25%, Élevé : >25%) ont été établis pour ancrer les interprétations du modèle dans des catégories ayant une pertinence clinique, ce qui facilite la prise de décisions. L'ajustement post-entraînement est important pour assurer que les probabilités de décès anticipées par le modèle s'alignent précisément avec les fréquences observées, ce qui est indispensable pour la crédibilité clinique. Les vérifications de cohérence sont des mesures de sécurité, fournissant des perspectives additionnelles sur le comportement du modèle et signalant toute éventuelle déviation entre les prédictions et les interprétations des règles. La solution envisagée permettra une IA intelligible, interprétable, robuste, généralisable, raisonnante, transparente, digne de confiance, pour une conformité à la réglementation, à la sécurité, à une soumission à un audit et à une certification. Elle pourra répondre aux exigences grandissantes d'une IA éthique et responsable.

Concernant la causalité, l'explication de l'analyse préliminaire découle du besoin de poser les bases pour les phases d'inférence causale plus sophistiquées. Le nettoyage des données est une phase pour assurer la fiabilité des analyses qui suivront. L'identification des valeurs aberrantes aide à évaluer la qualité des données et à repérer les points extrêmes susceptibles d'affecter les modèles. L'analyse des corrélations simples et des *ATE* manuels offre un premier point de référence avant la mise en œuvre de la logique floue et de l'ajustement des facteurs confondants. Pour finir, c'est un processus clé pour garantir que les modèles reposent sur des fondations robustes et que les résultats finaux demeurent intelligibles.

9.7 Les synthèses préliminaires

Pour l'interprétabilité, les premières évaluations indiquent que le modèle arborescent présente de bonnes performances prédictives, en particulier concernant le critère *ROC-AUC* sur l'ensemble des sous-groupes de données. La calibration après l'entraînement s'est révélée utile, réduisant l'*ECE* à une valeur proche de zéro et renforçant ainsi la crédibilité des probabilités estimées. L'intégration des modules de la *PFL* a rendu possible la création de règles de risque interprétables liées aux groupes de patients. Ces règles, associées à des scores d'interprétabilité et des diagnostics de cohérence, établissent les fondations d'un système capable non seulement de fournir des prédictions, mais aussi de les saisir et de les valider. Les alertes de concordance entre le modèle et les règles floues indiquent les zones requérant une enquête plus détaillée, mettant en avant la nécessité de cette surveillance proactive. La capacité du système à dénormaliser les valeurs des caractéristiques contenues dans ces règles constitue une bonne avancée pour leur interprétation directe par les spécialistes.

Concernant la causalité, les résumés initiaux suggèrent que, malgré un rétrécissement suite au nettoyage, l'ensemble des données demeure représentatif des populations de patients en état de choc septique. La présence de valeurs extrêmes sur plusieurs variables souligne la nécessité d'adopter des méthodes robustes. Les corrélations brutes mettent en évidence la complexité des liens entre le traitement, l'issue et les facteurs de confusion d'où la nécessité d'une inférence causale ajustée. Les premiers *ATE* manuels fournissent une vision générale de l'effet moyen, mais c'est l'intégration de la logique floue et des *NFATE* qui conduira à un entendement plus détaillé de la causalité. Les données préliminaires du *NFATE* indiquent que des variables telles que la crp, le pH artériel et la température sont des facteurs de confusion inhérente dont les caractéristiques et l'impact causal requièrent une étude plus approfondie.

9.8 Conclusion

La phase initiale de ce projet a prouvé la viabilité et les avantages d'une méthode sophistiquée et hybride pour prédire le risque du choc septique. Les résultats initiaux confirment l'efficacité du modèle et la pertinence de sa calibration, tout en mettant en évidence l'opportunité offerte par les règles floues pour une meilleure interprétation. Ces bases posent les jalons pour une étude plus détaillée des liens entre les prédictions du modèle et les interprétations symboliques, pavant la voie vers une IA plus valorisante et plus digne de confiance dans des contextes sensibles.

Manifestement, l'examen et la synthèse initiale des données ont confirmé leur qualité, quantifié la présence d'éléments aberrants et posé les fondations pour des interprétabilités et des analyses causales plus détaillées. Ces phases sont élémentaires pour assurer la robustesse des résultats finaux et la pertinence des déductions. Cela nous conduit à une inspection plus approfondie des résultats obtenus, en nous concentrant sur l'analyse détaillée, les trouvailles majeures et le classement des données. La section suivante mettra l'accent sur une exposition approfondie des résultats, fournissant un aperçu exhaustif des principales conclusions de notre recherche.

CHAPITRE 10 :
RÉSULTATS

10.1 Introduction

Cette partie expose les résultats obtenus grâce au système optimisé avec le *PFL*-Arborescent, en fournissant une étude détaillée des performances de prédiction et de la capacité d'interprétation. Elle expose également de façon détaillée les résultats de notre étude, en soulignant l'inférence causale enrichie grâce à la *PFL*. Nous exposerons les résultats principaux, en les structurant de façon hiérarchique, dans un souci de faciliter la compréhension de leur incidence sur le lien entre l'administration d'antibiotiques et le taux de mortalité chez les patients souffrant de choc septique.

10.2 L'analyse complète : présentation et objectivité

Le *PFL-LightGBM* a été entraîné et évalué sur trois sous-ensembles de données : "A=Tous" (l'ensemble complet), "A=0" (patients sous antibiotiques multiples) et "A=1" (patients sous antibiotique unique). Les performances sont évaluées à l'aide d'un large éventail de métriques, couvrant à la fois la discrimination, la calibration et la robustesse.

Pour le sous-ensemble "A=Tous":

- *ROC-AUC* : 81.82%. Cette métrique mesurant la capacité du modèle à distinguer les classes positives de ceux négatifs, indique une bonne performance de discrimination globale.
- *ECE* brut: 0.0893, *ECE* calibré: 0. L'amélioration de l'*ECE* de 0.0893 est bonne et confirme l'excellente calibration post-entraînement du modèle, assurant que les probabilités prédites sont fiables et correspondent aux fréquences réelles.
- Rappel (*Recall*): 47.45%. Le rappel est déterminant dans le contexte médical, car il mesure la proportion de vrais positifs correctement identifiés, minimisant ainsi les faux négatifs.
- Cohérence globale: 0.8806. La cohérence, calculée comme 1 moins l'écart absolu moyen entre les prédictions du modèle arborescent et les scores d'interprétabilité des règles floues est un

indicateur clé de la correspondance entre les prédictions et leurs interprétations. Un écart de 0.1194% est observé, déclenchant une alerte globale.

Pour le sous-ensemble "A=0" (Antibiotiques multiples):

- *ROC-AUC*: 83.07%. La performance discriminatoire reste bonne.
- *ECE* Brut: 0.0917, *ECE* Calibré: 0. L'amélioration de 0.0917 de l'*ECE* confirme la robustesse de la calibration.
- Rappel (*Recall*): 49.68%. Le rappel est légèrement supérieur à celui du dataset "A=Tous".
- Cohérence globale: 0.7797. L'écart absolu moyen global est de 0.2203, ce qui est considéré comme significatif et déclenche une alerte de cohérence.
- Analyse par cluster: Tous les clusters présentent une "Excellente" ou "Bonne" cohérence règle-modèle arborescente (écart < 10%), avec des écarts allant de 0% à 7.9%. Notamment, le cluster 6, classifié comme "RISQUE ÉLEVÉ" affiche un risque calibré de 33.1% et une probabilité de mortalité inhérente à la règle floue de 25.19% avec un écart de 7.9%.

Pour le sous-ensemble "A=1" (Antibiotique unique):

- *ROC-AUC*: 85.68%. Ce sous-ensemble présente la meilleure performance discriminatoire.
- *ECE* Brut: 0.0817, *ECE* Calibré: 0. La calibration est également bonne, avec une amélioration de 0.0817 de l'*ECE*.
- Rappel (*Recall*): 82.73%. Ce rappel est très élevé, indiquant une excellente capacité à identifier les cas de mortalité dans ce groupe.
- Cohérence globale: 0.8993. L'écart absolu moyen global est de 0.1007, déclenchant également une alerte de cohérence.
- Analyse par cluster: Le cluster 3 présente une "Faible" cohérence (écart: 15.9%). Ce cluster est classé "RISQUE MOYEN" avec un risque calibré de 23.7% et une probabilité de mortalité inhérente à la règle floue de 7.78%. Cette divergence souligne la nécessité d'une analyse plus approfondie des interactions dans ce sous-groupe.

Visualisations : Les diagrammes représentent la distribution des scores et des prédictions, en plus des courbes de calibration, démontrent visuellement l'efficience de la calibration et la correspondance entre les prédictions du modèle et les scores d'interprétabilité des règles.

Concernant la causalité, les conclusions sont issues de l'application de notre méthode sur le jeu de données épuré comprenant 12 780 observations. L'étude a examiné la quantification de l'*ATE*, de l'*ATE-PFL*, du *NFATE* ainsi que des indicateurs de vagueness (entropie) et d'incertitude.

- *ATE* manuel et effets moyens du traitement :

  Pour l'ensemble des confondants analysés, l'*ATE* calculé manuellement varie, par exemple, de 0.0607 pour '*bilirubin*' à 0.1489 pour '*crp*'. L'effet moyen pour les patients traités avec un seul antibiotique (groupe traité) est de 0.3590, tandis que pour ceux traités avec plusieurs antibiotiques (groupe non traité), il est de 0.231274. La différence entre ces deux valeurs (0.3590−0.2312=0.1278) indique une mortalité plus élevée chez les patients recevant un seul antibiotique, ce qui souligne la nécessité d'une collaboration avec un spécialiste clinicien en choc septique et d'une analyse causale ajustée pour les facteurs de confusion.

- Segmentation floue (fuzzification) et *ATE-PFL* :

  L'application du *BGMM* a permis de segmenter chaque variable confondante en trois catégories floues : *"low", "medium" et "high"*, avec des degrés d'appartenance pour chaque observation. Ces segments capturent les nuances dans les valeurs des variables, par exemple, pour les *"white blood cells"* :

  Low : Moyenne = 7.6744, Min = 0.1000, Max = 11.6600 (5508 observations),

  Medium : Moyenne = 19.9998, Min = 15.0364, Max = 27.6000 (4398 observations),

  High : Moyenne = 43.2105, Min = 31.9500, Max = 94.3000 (725 observations).

  Les *ATE-PFL* (*ATE* pondérés par les degrés d'appartenance floue) pour chaque segment ont ensuite été calculés. Par exemple, pour les '*white blood cells*' :

  ATE pour le segment *low* : 0.1865,

  ATE pour le segment *medium* : 0.1232,

  ATE pour le segment *high* : -0.0564.

  Ces résultats bruts pour les *ATE* par segment sont ajustés de manière que leur moyenne corresponde à l'*ATE* manuel global. Pour les '*white blood cells*' :

  Adjusted ATE-PFL low: 0.2143,

  Adjusted ATE-PFL medium: 0.1415,

  Adjusted ATE-PFL high: -0.0648.

La validation a montré que la moyenne des *ATE-PFL* ajustés est bien égale à l'*ATE* manuel (par exemple, 0.0970 pour '*white blood cells*'), ce qui assure la cohérence du modèle.

- *Vagueness* et Incertitude :
  1. Les scores de *vagueness* et les écarts-types pondérés ont été calculés pour évaluer l'ambiguïté et l'incertitude des segments. Pour les '*white blood cells*' :
  2. Le segment "*medium*" est le plus ambigu (*vagueness score* : 9.0364).
  3. Le segment "*high*" est le plus incertain (*weighted standard deviation* : 15.9853).
  4. Le score de *vagueness* global est de 9.8963 et l'écart-type global pondéré est de 9.9220. Ces métriques fournissent des informations sur la qualité de la segmentation et la dispersion des données au sein de chaque catégorie floue.

- *NFATE* :
  1. Le *NFATE*, calculé pour chaque confondant offre une mesure normalisée de l'effet causal ajusté. Les résultats triés par *NFATE* décroissant mettent en évidence les confondants les plus influents.
  2. Ces résultats suggèrent que '*crp*', '*arterial ph*' et '*temperature*' sont les confondants qui modulent le plus l'effet causal de l'antibiotique sur la mortalité, une fois les effets flous pris en compte.

10.3 Les découvertes principales

Pour l'interprétabilité :

1. **Bonne calibration des probabilités :** Obtenir un *ECE* de 0 sur l'ensemble des sous-groupes après la calibration est une surprise inattendue. Cela indique que les probabilités estimées par le modèle sont fortement fiables, un critère requis pour les applications critiques où la confiance dans les indices de risque est primordiale.

2. **Rappel élevé pour le groupe A=1 :** Le modèle fait preuve d'une performance notable dans la détection des cas de mortalité parmi les patients traités par un seul antibiotique (Rappel de 82,73%), indiquant une plus grande précision dans ce sous-groupe.

3. **Génération de règles claires et dénormalisées :** Le système *PFL* produit des directives « *IF-THEN* » pour chaque regroupement, comportant des plages de caractéristiques dénormalisées et des adjectifs linguistiques (« plutôt basse », « moyenne », « plutôt élevée », « large éventail », « assez précise », « très précise »). Les experts peuvent facilement interpréter ces règles.

4. **Identification des zones d'incohérence :** Le système d'alerte de *cohérence* par regroupement est efficace pour repérer les sous-ensembles de patients où les prédictions du modèle arborescent diffèrent considérablement des probabilités inhérentes aux règles floues. Un exemple notable serait le cluster 3 du sous-ensemble A=1, qui présente un écart de 15.9%. Cette aptitude à identifier les zones d'incertitude où le modèle pourrait se baser sur des corrélations complexes plutôt que sur des relations de cause à effet explicites représente des éléments intéressants pour la perspective d'auditabilité.

5. **Règle globale synthétique :** La possibilité de créer une règle globale consolidée, pondérée selon le nombre de patients dans chaque groupe fournit une perspective macro et compréhensible du comportement général du modèle. Par exemple, la directive générale pour toutes les données signale un risque réduit de 9.52% P(mortalité|Règle floue globale) et fourni des plages dénormalisées pour chaque caractéristique, incluant leur emplacement et leur exactitude.

Pour l'analyse causale :

1. **Distinguer la causalité de la corrélation :** L'*ATE* manuel indique une légère liaison positive entre un antibiotique unique et le taux de mortalité tandis que les *NFATE* et *ATE-PFL* ajustés facilitent le discernement de l'influence des variables confondantes sur cette relation. Il est évident que la simple corrélation brute ne permet pas de comprendre la complexité causale.

2. **Impact de la fuzzification sur l'*ATE* :** La méthode *ATE-PFL* fournit une analyse de l'impact du traitement. Par exemple, concernant les *« white blood cells »*, un niveau « *low* » est lié à un *ATE* positif (0.2143) tandis qu'un niveau « *high* » est relié à un *ATE* négatif (-0.0648). Cela suggère que l'effet de l'antibiotique sur le taux de mortalité peut fluctuer considérablement selon les niveaux précis de la variable confondante, ce qui ne peut pas être perçu avec une seule *ATE* globale.

3. **Identification des confondants clés (via *NFATE*) :** Le *NFATE* facilite la classification des confondants selon leur influence causale ajustée. La crp, le pH artériel et la température se distinguent comme les variables les plus fondamentales, une observation cliniquement plausible qui indique qu'elles pourraient jouer un rôle prépondérant dans la réaction au traitement et l'évolution des patients en état de choc septique.

4. **Quantification de l'incertitude :** Les mesures quantitatives d'incertitude dans la segmentation sont fournies par les scores de vagueness et les écarts-types pondérés. L'ambiguïté associée à la catégorie *« medium »* des *« white blood cells »* ainsi que l'incertitude relative à la catégorie *« high »* peut influencer l'analyse clinique, en indiquant des régions où les données sont moins précises ou davantage dispersées.

10.4 La hiérarchisation des informations

Les informations produites par le système sont hiérarchisées en plusieurs niveaux pour différents utilisateurs:

- Niveau 1 : Global (pour une vue d'ensemble):

  ROC-AUC et rappel global: Mesures de performance synthétiques.

  ECE (brut et calibré): Indicateur de fiabilité des probabilités, avec l'amélioration.

  Cohérence globale et alerte: Indique la concordance générale entre le modèle et ses interprétations.

  Règle globale unique: Une seule règle synthétique dénormalisée pour un discernement rapide de la logique moyenne du modèle.

  Recommandations cliniques globales: Basées sur les seuils cliniques fixes.

- Niveau 2 : Par sous-groupe:

  Métriques de performance détaillées par A=0, A=1, A=Tous: Précision, *F1-Score*, Spécificité, etc.

  Alertes de cohérence par cluster: Identification des clusters problématiques nécessitant une investigation avec le clinicien. Pour le sous-ensemble A=1, le Cluster 3 est identifié avec une faible cohérence.

Diagnostics d'interprétabilité: Statistiques sur les scores bruts et normalisés ainsi que sur le coverage adaptatif.

- Niveau 3 : Par patient:

Cluster de risque dominant et sa probabilité inhérente: La classification du patient dans un cluster spécifique (ex.: Cluster 6 pour le patient hypothétique avec une mortalité de 25.19%).

Règle clinique détaillée du cluster dominant: Interprétation « *IF-THEN* » dénormalisée avec intervalles, position et précision des caractéristiques pour ce patient.

Simulation causale: Permet de visualiser l'impact d'une intervention sur la probabilité de mortalité et le profil de risque du patient, changeant potentiellement le cluster dominant et sa règle associée (ex.: changement du Cluster 4 au Cluster 6 pour le patient simulé).

Explication *LIME* locale: Fournit les caractéristiques les plus influentes pour la prédiction individuelle du modèle arborescent.

En ce qui concerne la causalité, l'ordre de priorité dans les informations favorise les *NFATE*. Ils combinent l'effet causal et l'approfondissement par la logique floue, offrant donc l'évaluation la plus appropriée de l'influence des variables confondantes. Il est également indispensable d'examiner les *ATE-PFL* par segment, car ils mettent en lumière la variation de l'effet causal en fonction des niveaux de la variable. Ces analyses sont enrichies par les scores de vagueness et d'incertitude, fournissant un aperçu de la fiabilité que l'on peut attribuer à ces résultats. Finalement, les *ATE* manuels et les corrélations brutes fonctionnent comme des références et des validations, soulignant la valeur de l'approche floue.

10.5 La vue d'ensemble des mesures

Le système est entraîné et évalué sur trois ensembles de patients :

- A=0 (Antibiotiques multiples) : Ce groupe représente les patients ayant reçu plusieurs antibiotiques.
- A=1 (Antibiotique unique) : Ce groupe concerne ceux ayant reçu un seul type d'antibiotique.
- A=Tous (Dataset combiné) : L'ensemble de la population de l'étude.

Les seuils cliniques fixes sont des points de référence pour notre interprétation. Ils définissent ce qui est considéré comme un risque faible, moyen ou élevé de mortalité, ce qui ancre les résultats du modèle dans une réalité clinique.

Analyse détaillée par groupe de patients

1. Groupe A=0 (Antibiotiques multiples)

Ce groupe se caractérise par une mortalité moyenne plus faible que le groupe A=1. Le modèle arborescent obtient de bonnes performances générales.

- Métriques de performance :
    - L'*Accuracy* (89,04%) et la spécificité (92,97%) sont élevées, ce qui montre que le modèle identifie les patients qui vont survivre (vrais négatifs). C'est décisif pour ne pas alarmer inutilement les cliniciens.
    - Le rappel (49,68%) est modéré. Cela signifie que le modèle a correctement identifié un peu moins de la moitié des patients qui sont réellement décédés. C'est un point d'amélioration, car les faux négatifs (patients décédés que le modèle a prédits comme survivants) sont un risque clinique important.
    - Le *F1-Score* (45,18%) et la précision (41,83%) sont autant modérés. La faible précision signale que parmi toutes les prédictions de décès, une proportion expressive est inexacte (faux positifs).
    - Le *ROC-AUC* (0,8307) est très bon, désignant une bonne disposition du modèle à particulariser les classes (survie vs décès).
- Calibration :
    - Le score *ECE* Brut (0,0917) montre que les probabilités initiales du modèle sont raisonnablement bien calibrées.
    - Le score *ECE* Calibré (0) est un bon résultat, montrant que la régression isotonique a parfaitement aligné les probabilités prédites avec les fréquences observées. Cela signifie que si le modèle prédit un risque de 10%, il y a 10% de patients dans ce groupe qui décèdent, ce qui affermit et de beaucoup la confiance clinique dans les résultats informatiques du modèle.
- Analyse des clusters :
    - Le modèle conçoit des clusters s'alignant bien avec les règles floues. Les valeurs de fidélité sont toutes élevées (proches de 1).
    - La cohérence est jugée « Excellente » pour approximativement tous les clusters et « Bonne » pour le Cluster 6 (écart de 7,9%). Un écart de cohérence faible signifie que la prédiction du modèle arborescent et la règle floue sont en accord pour ces profils de patients.

- Le risque calibré (9,7%) et la probabilité de mortalité du cluster (6,64%) pour le Cluster 0 sont très proches, ce qui valide l'interprétabilité de ce profil. Cependant, nous notons l'énorme différence entre le risque brut (13,6%) et le risque calibré (9,7%), soulignant l'importance du processus de calibration pour ce groupe de patients.

2. Groupe A=1 (Antibiotique unique)

Ce groupe expose un défi singulier, parce que sa mortalité moyenne globale est inconsidérément plus élevée que celle du groupe A=0.

- Métriques de performance :
  - Le rappel (82,73%) est élevé. Le modèle est bon pour identifier les patients à risque de décès dans ce groupe. C'est une force majeure d'un point de vue clinique, car cela réduit considérablement les faux négatifs.
  - Cependant, il y a un compromis : la précision (26,66%) est beaucoup plus faible, ce qui signifie qu'un grand nombre de patients prédits à risque de décès ne décèdent pas réellement (faux positifs). Cela peut conduire à des alarmes cliniques excessives.
  - Le *Balanced Accuracy* (77,81%) et le *ROC-AUC* (0,8568) sont très bons, ce qui confirme que le modèle a une forte capacité de prédiction, même si le compromis entre précision et rappel a été ajusté pour prioriser le rappel.
- Calibration :
  - De même, la calibration est bonne, avec un *ECE* Calibré (0), ce qui rend les probabilités prédites dignes de confiance.
- Analyse des clusters :
  - Une alerte de cohérence est déclenchée pour le Cluster 3 (écart de 15,9%). C'est le seul cas où le modèle arborescent et la règle floue divergent de façon significative. Le modèle prédit un risque calibré de 23,7% pour ce cluster, ce qui le classe comme un risque élevé, tandis que la règle floue lui attribue un risque de seulement 7,78% (faible). Cela indique que le profil de ce cluster est plus complexe que ce que la simple règle floue peut capturer. Il s'agit d'un cas où l'expertise clinique doit être sollicitée pour saisir les facteurs subtils que le modèle a découverts.

3. Groupe A=Tous (*Dataset* combiné)

Ce groupe sert de point de référence global, dévoilant la performance moyenne du modèle sur l'ensemble de la population.

- Métriques de performance :
    - Les résultats sont un mélange des deux sous-groupes, avec une performance générale robuste. Le rappel (47,45%) est modéré, ce qui est logique étant donné l'influence du groupe A=0.
    - Le *ROC-AUC* (0,8182) est légèrement inférieur à celui des sous-groupes, ce qui montre que la segmentation des patients en fonction du traitement (A) est une variable importante pour la prédiction.
- Calibration :
    - La calibration a également été très efficace pour ce groupe.
- Analyse des clusters :
    - La cohérence globale (0,8806) est qualifiée de "Faible" par l'alerte, car l'écart de 11,9% dépasse le seuil défini. Cela signifie que la règle globale moyenne, qui synthétise tous les clusters n'est pas un bon représentant de la complexité de l'ensemble du modèle. Il est bien plus pertinent d'utiliser les règles spécifiques à chaque cluster que la règle globale unique.
    - L'analyse par cluster révèle des divergences notables, notamment pour les Clusters 3 et 5, avec des écarts de cohérence de 8,1% et 5,4% respectivement, ce qui les classe dans la catégorie "Bonne", mais justifie une surveillance.

Nous résumons les vrais/faux positifs et négatifs dans le Tableau 6 et le comparatif des principales métriques dans Tableau 7.

Jeu de données	Vrais positifs (TP)	Faux positifs (FP)	Vrais négatifs (TN)	Faux négatifs (FN)
A=0	21,762	30,809	407,656	22,042
A=1	16,691	45,948	123,615	3,485
A=Tous	30,356	47,947	560,081	33,624

Tableau 6 : Tableau récapitulatif des vrais/faux positifs et négatifs.

	Accuracy	Balanced Accuracy	ROC-AUC	Rappel	Précision	F1-Score		ECE Calibré
A=0	0,8904	0,7133	0,8307	0,4968	0,4183	0,4518		0
A=1	0,7395	0,7781	0,8568	0,8273	0,2666	0,4032		0
A=Tous	0,8786	0,6978	0,8182	0,4745	0,3882	0,4268		0

Tableau 7 : Tableau comparatif des principales métriques.

Le Tableau 6 met en évidence le compromis de performance. Le modèle pour le groupe A=1 a beaucoup plus de vrais positifs (TP), mais aussi beaucoup plus de faux positifs (FP) que le groupe A=0, en raison de l'ajustement du seuil de classification pour maximiser le rappel. En revanche, le nombre de faux négatifs (FN) est beaucoup plus bas pour A=1, ce qui confirme l'efficacité du modèle pour détecter les cas de mortalité, même au prix de fausses alarmes.

Le Tableau 7 résume les observations précédentes. Le modèle A=1 a une *Balanced Accuracy* et *un ROC-AUC* plus élevés, montrant une meilleure performance globale, mais son *Accuracy* est plus faible en raison de la présence de plus de faux positifs. Le rappel de 82,73% pour le groupe A=1 est la métrique clé justifiant l'approche segmentée. L'*ECE* Calibré nul pour tous les groupes est un bon résultat, validant la fiabilité des probabilités du modèle.

10.6 La justification

L'objectif de ce classement est d'ajuster le niveau de l'information en fonction des besoins de l'utilisateur. Un administrateur pourrait se limiter aux indicateurs globaux et à la règle synthétique alors qu'un spécialiste exigera des précisions par regroupement et des analyses propres à un cas (par exemple, un patient). Cette modularité facilite le discernement et l'interprétation, rendant l'IA à la fois efficace, exploitable et digne de foi.

Le traitement des données manquantes, la stratification, l'ajustement et les seuils cliniques fixes sont motivés par le désir de garantir la robustesse et la pertinence clinique du système. L'identification des incohérences est vitale pour la santé des patients, parce qu'elle aide à repérer les situations où le modèle fonctionne comme une entité opaque difficile à analyser par les règles floues, demandant alors une intervention humaine plus marquée. Des interprétations précises des indicateurs sont fournies afin d'assurer un éclaircissement clair et sans ambiguïté des termes techniques.

Concernant la causalité, la raison pour laquelle ces résultats sont présentés de manière détaillée et structurée est d'assurer un discernement profond de l'impact de la *PFL* sur l'inférence causale. En distinguant l'*ATE* manuel des *ATE-PFL* ajustés et du *NFATE*, nous mettons en évidence la manière dont l'intégration du symbolisme flou permet de saisir des subtilités vitales qui seraient par ailleurs imperceptibles. Les représentations graphiques permettent de rendre ces idées abstraites plus tangibles pour un auditoire élargies, tout en préservant l'objectivité grâce à la mise en avant de mesures quantitatives, telles que *ATE*, *NFATE*, *Vagueness*, etc. C'est primordial pour appuyer nos suppositions et illustrer l'apport de notre méthode.

10.7 Le résumé des résultats

Les résultats révèlent un équilibre prometteur entre précision dans les prédictions et capacité d'interprétation. Le modèle arborescent calibré brille par la fiabilité de ses probabilités. La *PFL* produit des règles cliniquement intelligibles et pertinentes. Le système détecte les incohérences, facilitant ainsi une intervention précise. La règle générale fournit un résumé pertinent, tandis que l'exemple d'explications spécifiques à chaque instance, grâce à *LIME* ou au cluster prédominant, répond aux exigences particulières des patients. Cette méthode s'avère particulièrement performante pour le groupe « A=1 » avec un taux de rappel très haut, mais elle met également en lumière des problèmes d'interprétabilité pour certaines catégories.

De plus, nos conclusions indiquent que l'intégration de la *PFL* offre une évaluation plus précise et accessible des effets causaux de l'administration d'antibiotiques sur la mortalité chez les patients en choc septique. Selon le *NFATE*, nous avons repéré la crp, le pH artériel et la température comme les facteurs de confusion les plus significatifs. Les *ATE-PFL* révisés démontrent que l'impact du traitement fluctue largement selon les niveaux indéfinis des variables confondantes. Pour conclure, les indicateurs de *vagueness* et d'incertitude évaluent la fiabilité de ces segments, fournissant ainsi une assise pour une IA plus intelligible. Par conséquent, la section suivante sera dédiée à l'analyse de ces résultats, en les comparant aux travaux existants, en examinant leurs répercussions et en mettant l'accent sur les apports de cette recherche.

10.8 Conclusion

Les résultats démontrent que la segmentation des patients en fonction du traitement (A=0 vs A=1) est une approche intéressante. Nos modèles obtiennent de bonnes performances à la fois en termes de prédictions et d'interprétabilité.

- Pour les patients du groupe A=0, le modèle est précis pour identifier les survivants.
- Pour les patients du groupe A=1, le modèle est performant pour identifier les cas de mortalité, ce qui est cliniquement plus important.
- La calibration post-entraînement a été une étape critique, transformant des probabilités initiales acceptables en prédictions d'une bonne fiabilité, ce qui est fondamental pour l'adoption clinique du modèle.
- Les alertes de cohérence par cluster sont des outils précieux, comme le montre le Cluster 3 du groupe A=1. Elles indiquent les cas où les règles floues ne suffisent pas à expliquer la décision du modèle arborescent, signalant la nécessité d'une investigation plus poussée par les cliniciens.

En bref, ce système offre des outils à la fois robustes et interprétables pour la prise de décision clinique, en adaptant sa stratégie de prédiction à la spécificité des patients.

In fine, les résultats valident l'efficacité du modèle hybride pour prédire le risque de mortalité lié au choc septique, en apportant une contribution à la fois causale et interprétative. De ce fait, les découvertes soulignent l'importance de la calibration, la pertinence des règles floues dénormalisées et l'intérêt du système d'alerte de cohérence. Ces facteurs représentent des progrès vers une IA plus transparente et responsable dans des contextes sensibles. La prochaine section se concentrera donc sur l'analyse de ces résultats, leur mise en rapport avec les théories existantes et leur portée clinique, pour enrichir la discussion.

CHAPITRE 11 :
DISCUSSION

11.1 Introduction

Cette partie vise à analyser les résultats générés par le système *PFL*-Arborescent, ceux de la *PFL* pour l'inférence causale et les comparer à notre niveau actuel de connaissances. Nous examinerons les relations et les conséquences de ces découvertes, composerons des arguments élaborés, défendrons la pertinence de la thèse sous-tendant notre méthode et mettrons en évidence nos apports.

11.2 Les résultats qualitatifs et quantitatifs

Résultats qualitatifs et quantitatifs pour l'analyse causale

Nous analysons l'effet causal de l'administration d'un seul antibiotique par rapport à plusieurs antibiotiques sur les taux de survie ou de décès des patients, en utilisant différentes variables comme facteurs de confusion. Nous employons une méthode combinant des moyennes pondérées avec comme modèle le *BGMM* pour évaluer l'*ATE*. Les résultats quantitatifs nous permettent de juger des effets du traitement dans des segments spécifiques des données, définis par les trois niveaux bas, moyens et élevés de chaque variable de confusion.

Résultats qualitatifs

Nous constatons que l'*ATE* pour l'administration d'un seul antibiotique est en moyenne de 0.1189 sur l'ensemble de ces variables. Par rapport au traitement multiple, l'administration d'un seul antibiotique a un effet moyen positif sur la survie. L'effet moyen chez les patients traités avec un seul antibiotique est de 0.3590, tandis que l'effet moyen chez ceux traités avec plusieurs antibiotiques est de 0.2312. Cela suggère que le traitement par antibiotique unique est lié à un meilleur taux de survie dans l'échantillon analysé.

Résultats quantitatifs

Les résultats quantitatifs montrent que l'*ATE* non ajusté (manuel) est positif pour toutes les variables de confusion. Cependant, lorsque nous affinons l'analyse en segmentant les patients sur les trois segments à l'aide de la *BGMM*, nous observons une variation importante des *ATE* ajustés. Par exemple, pour le taux de créatinine, l'*ATE* est de 0.1616 pour le segment « faible », de 0.8543 pour le « moyen » et de -0.7520 pour celui « élevé ». De même, pour la pression artérielle systolique, les *ATE* ajustés sont de 0.2118 pour le segment « faible », de 0.0752 pour le « moyen » et de 0.0431 pour celui « élevé ». Ces résultats démontrent que l'impact du traitement fluctue énormément selon les niveaux des variables de confusion. Pour chaque variable, la moyenne des *ATE* ajustés pour les trois segments correspond à l'*ATE* manuel initial.

Ces analyses indiquent que l'efficacité du traitement avec un seul antibiotique dépend fortement des caractéristiques des patients, représentées ici par les valeurs des variables de confusion. La classification en catégories floues nous offre une meilleure intellection de la manière dont l'impact du traitement se manifeste dans des sous-groupes spécifiques. Par exemple, l'*ATE* corrigé négatif de -0.7520 pour les patients présentant un taux de créatinine élevé, comparé à l'*ATE* positif de 0.1616 pour ceux avec un taux bas, nous pousse à formuler une supposition sur l'efficacité du traitement en lien avec la condition de la fonction rénale. Ces conclusions laissent à penser qu'un traitement antibiotique isolé pourrait ne pas être efficace, voire nuisible, pour les patients présentant des taux de créatinine élevés. Cette constatation nécessite une étude plus détaillée des effets du traitement et l'examen des traits cliniques individuels.

Pourquoi utilisons-nous cette approche ?

La méthode que nous utilisons, qui adapte le *BGMM*, permet d'analyser l'effet causal de manière plus nuancée. Au lieu d'assumer que le traitement a un effet uniforme sur l'ensemble de la population, le *BGMM* crée des segments de patients avec des caractéristiques similaires, basées sur une variable de confusion. Chaque segment représente un groupe flou, indiquant qu'un patient peut faire partie de plusieurs groupes avec différents niveaux d'appartenance. En déterminant l'*ATE* pour chaque segment, nous discernons si l'impact du traitement fluctue en fonction du contexte clinique du patient. La rectification mathématique mise en œuvre par un coefficient d'ajustement assure que la moyenne des effets segmentés correspond à l'*ATE* global. Cela permet une interprétation des résultats à la fois pour la population générale et pour des groupes spécifiques.

Ce que nous avons fait

Nous avons d'abord calculé l'effet causal *ATE* global en utilisant une régression linéaire simple avec une variable de confusion. Par la suite, nous avons segmenté la population pour chaque variable de confusion en trois groupes flous en utilisant un modèle de mélange gaussien bayésien. Nous avons ensuite calculé un *ATE* pour chaque segment. Enfin, nous avons ajusté les *ATE* de chaque segment pour que leur moyenne arithmétique soit égale à l'*ATE* global initial, assurant la cohérence des résultats et permettant une interprétation plus fine.

Résultats qualitatifs et quantitatifs pour l'interprétabilité

Pour l'interprétabilité, nous exposons les mesures quantitatives. Par la suite, nous procédons à l'interprétation des résultats qualitatifs.

Résultats quantitatifs

Nous effectuons des calculs de mesures de performance et de calibration. Ces calculs mesurent l'aptitude du modèle à prévoir avec précision le risque. Nous employons une validation croisée en cinq plis. Cette approche assure une évaluation robuste.

Le modèle basé sur l'arborescence obtient des performances quantifiables. Pour le sous-ensemble A=1, nous constatons une Précision de 26.66 % et un Rappel de 82.73 %. Le F1-Score est de 40.32 %. Cela démontre que nous gérons le compromis entre la détection des cas à risque et la minimisation des faux positifs. Le score de *ROC-AUC* est de 0.8568. Ce score indique que la capacité de discrimination du modèle est bonne.

Nous évaluons la fiabilité des probabilités. L'*ECE* brute est de 0.0817. Après la calibration, l'*ECE* est de 0. Cela représente une amélioration de 0.0817. Cela indique que nous améliorons de manière significative la fiabilité des prédictions.

Nous générons des règles cliniques pour chaque groupe de patients. Ces règles sont basées sur les centroïdes des groupes. Chaque règle propose une interprétation du risque.

- Groupe à risque faible : Le modèle calibré assigne un risque de 6.2 %. La probabilité de mortalité associée à la règle floue est de 4.83 %. La fidélité du groupe est de 0.9861. La prédiction est une

survie probable. Cette fraction constitue 31,8 % du total des données. Les patients dans ce groupe présentent des valeurs de variables dans des intervalles spécifiques.

- Groupe à risque moyen : Le modèle assigne un risque de 13.0 % à ce groupe. La probabilité de mortalité associée à la règle floue est de 11.45 %. La fidélité du groupe est de 0.9844. Nous recommandons une surveillance renforcée. Ce groupe représente 5.5 % de l'ensemble des données.
- Groupe à risque élevé : Le modèle assigne un risque de 20.4 % à un groupe et de 43.3 % à un autre. La probabilité de mortalité pour le premier groupe est de 12.26 % et pour le second, de 42.13 %. Nous recommandons une intervention urgente pour ces patients. Ces groupes représentent respectivement 15.7 % et 0.7 % de l'ensemble des données.
- Cohérence des règles : Nous constatons une divergence générale de 10.07 % entre les prédictions du modèle et les interprétations des règles. Cela témoigne d'une divergence notable. Il est conseillé de procéder à une analyse plus approfondie des écarts pour chaque groupe.

11.3 L'interprétation des résultats

L'obtention d'un *ECE* presque nul suite à une calibration isotonique témoigne de la crédibilité des probabilités anticipées par le modèle. Dans un contexte sensible, cela signifie que si le modèle anticipe un risque de 20%, ce risque sera réellement de 20% parmi les patients dans des circonstances comparables. Cette caractéristique est primordiale pour l'adhésion clinique et la décision médicale, car les professionnels de santé se basent non seulement sur la catégorisation (décès/survie), mais également sur le degré de confiance lié à chaque prédiction.

Aussi, cette méthode a la capacité de produire des « *IF-THEN* » pour chaque groupe, ce qui constitue l'un de ses points forts. Ces règles sont dénormalisées en se basant sur les réels intervalles de valeurs détectés dans les données d'apprentissage, ce qui facilite leur discernement par les cliniciens. L'utilisation de termes descriptifs comme « plutôt bas », « moyen » ou « plutôt haut » ainsi que la définition d'une « plage » (par exemple, « très précise », « assez précise », « large »), contribue à imiter le processus de réflexion clinique basé sur des seuils et des intervalles plutôt que sur des valeurs exactes. Ces règles symboliques offrent non seulement une intellection des prédictions du modèle pour un ensemble de patients, mais également les raisons sous-jacentes, en déterminant les combinaisons de caractéristiques appropriées à un profil de risque spécifique.

En plus, l'outil principal est le système d'alerte de cohérence. Il fonctionne comme un système d'autocertification, indiquant les situations où la prédiction du modèle arborescent (potentiellement plus sophistiqué et capable de saisir des interactions sensibles) s'écarte notablement de la probabilité liée à la règle floue prédominante. L'exemple du Cluster 3 pour le sous-groupe A=1, avec une différence de 15,9%, démontre un cas où le modèle pourrait se baser sur des liens plus sophistiqués ou non déchiffrables par la règle floue. Cela ne remet pas forcément en question la prédiction faite par le modèle arborescent, mais souligne l'importance d'une vigilance renforcée et d'une évaluation clinique manuelle par un spécialiste.
Également, bien que distincte du processus principal d'entraînement, la simulation causale améliore l'interprétabilité en offrant aux cliniciens la possibilité d'expérimenter avec des scénarios hypothétiques. En ajustant certaines caractéristiques précises et en analysant les variations dans le taux de mortalité, le dispositif propose une forme d'interprétabilité contrefactuelle. Ceci dépasse la simple prédiction pour favoriser l'entendement de l'impact possible des interventions thérapeutiques, distinguant ainsi les corrélations des inférences causales.

Autant, dans le cadre de l'analyse causale, les résultats obtenus viennent corroborer plusieurs éléments élémentaires de nos hypothèses. D'abord, l'écart entre les *ATE* globaux manuels et les *ATE-PFL* segmentés par logique floue est saisissant. La valeur manuelle pour l'*arterial base excess* est de 0.1007, cependant les *ATE-PFL* réajustés fluctuent grandement : 'faible' (0.0151), 'moyenne' (0.0310), et 'élevée' (0.2561). Cela illustre que l'effet causal de l'administration d'antibiotiques sur la mortalité varie, étant fortement lié aux subtilités des facteurs confondants. Une analyse causale classique, dépourvue de cette granularité indéterminée, pourrait omettre des effets hétérogènes centraux pour la décision clinique.

En outre, la validation de la moyenne des *ATE-PFL* ajustés comme étant équivalente à l'*ATE* manuel (par exemple, 0.1007 pour '*arterial base excess*' et 0.0970 pour '*white blood cells*') renforce la consistance interne de notre modèle. Cela implique que notre méthode floue décompose l'effet causal total en effets particuliers à des segments, tout en préservant l'ampleur globale de l'effet.

De même, la reconnaissance de la *crp*, du *pH arterial* et de la *temperature* comme les facteurs de confusion ayant le *NFATE* le plus élevé (0.15, 0.14, 0.14 respectivement) présente un intérêt clinique. Ces paramètres pourraient servir d'indicateurs décisifs pour évaluer la sévérité de l'inflammation et du déséquilibre dans le contexte du choc septique. Un *NFATE* élevé indique que ces variables ne sont pas seulement liées au

traitement et au résultat, mais qu'elles influencent également de façon significative l'effet causal de l'antibiotique, ce qui justifie une considération spéciale dans les approches de traitement.

En définitive, les résultats que nous avons obtenus sont enrichis d'une dimension cognitive grâce aux scores de vagueness et d'incertitude (écart-type pondéré). Le segment *« medium »* des *'white blood cells'*, étant le plus flou, et le segment *« high »* qui est le plus incertain indique que les informations dans ces plages tendent à être plus dispersées ou variables. Ceci pourrait être attribué à une diversité biologique accrue des patients dans ces catégories ou à des inexactitudes de mesure, mettant en évidence l'importance de la nuance pour représenter ce constat. Une IA capable de reconnaître et d'évaluer son incertitude est davantage fiable et transparente.

11.4 La confrontation aux théories existantes

Généralement, les systèmes experts fondés sur des logiques « *IF-THEN* » sont très transparents, mais ils peuvent faire défaut en termes de précision lorsqu'ils sont confrontés à la complexité et à l'instabilité des données sensibles. Inversement, les modèles *DL* brillent par leur précision, mais tendent fréquemment à être peu transparents. L'introduction de la logique floue vise à gérer l'incertitude et l'imprécision caractéristiques des systèmes réels. Cette recherche associant la logique floue à des approches probabilistes et des modèles arborescents se présente comme un lien entre ces deux paradigmes.

Aussi, l'approche *PFL* est en accord avec l'idée hybride, qui cherche à fusionner les atouts des modèles symboliques (capacité d'interprétation) et subsymboliques (efficacité). La fusion des probabilités *LightGBM* et des ceux de règles floues est un exemple de cette hybridation, permettant au système de bénéficier à la fois de la puissance prédictive du *LightGBM* et de l'interprétabilité de ces règles.

Également, le concept de fidélité des interprétations, bien que parfois implicite, est nouveau et central dans la littérature sur l'interprétabilité de l'IA. La fidélité mesure la conformité d'une interprétation par rapport au comportement du modèle. Dans notre cas, la cohérence entre le risque prédit par le modèle arborescent et la probabilité inhérente aux règles floues est une forme de mesure de la fidélité. Un écart minime suggère que ces règles représentent fidèlement la logique du modèle sous-jacent pour un cluster spécifique. Les situations de « faible cohérence » mettent en évidence des cas où le modèle se comporte comme une boîte noire, même face à des directives floues, mettant en relief les contraintes de

l'interprétabilité symbolique pour certains profils complexes. Autant, nos résultats s'alignent avec les principes de l'inférence causale postulant que les corrélations ne sont pas suffisantes pour établir la causalité et que l'ajustement pour les facteurs de confusion est fondamental. Le cadre de Pearl est la pierre angulaire de notre approche *ATE*. Cependant, nous étendons ces cadres en intégrant la flexibilité du flou.

Au final, les travaux de Serge, de Faghihi et coll. ont été une source d'inspiration majeure pour l'application du *BGMM* afin d'obtenir des degrés d'appartenance flous. Notre recherche valide leur approche en démontrant son utilité concrète dans un contexte sensible. La notion de *NFATE*, inspirée des travaux de Saki et Faghihi (2024) est une contribution directe normalisant et rendant l'interprétation des effets causaux flous plus aisée.

11.5 La comparaison avec la littérature

Le domaine de l'interprétabilité et de l'explicabilité de l'IA connaît une croissance rapide, avec des techniques telles que *SHAP* et *LIME* qui se sont imposées comme des références majeures en matière d'explicabilité. L'illustration fournie par *LIME*, incorporée dans ce processus, offre des explications spécifiques à chaque instance. Tandis que *LIME* génère des modèles simples, mais fidèles localement, le *PFL* a pour objectif d'intégrer les comportements de groupes de patients dans des règles organisées.

Aussi, les systèmes fondés sur la théorie du flou ont une longue tradition en médecine, fréquemment louée pour leur aptitude à saisir l'expertise pointue. Toutefois, leur dépendance à l'ingénierie des connaissances et les défis liés à la fourniture d'une précision ont parfois fait l'objet de critiques. Notre méthode *PFL* adresse ces contraintes en employant des stratégies d'arborescence pour générer automatiquement les règles floues (par le biais de *GMM* et *K-Means*) et leurs probabilités, diminuant de ce fait l'exigence d'ingénierie manuelle et augmentant la capacité d'adaptation aux données complexes. De surcroît, l'incorporation de *LightGBM* représente aussi un pont vers la documentation relative aux modèles intrinsèquement interprétables.

En plus, contrairement à la plupart des recherches sur la prédiction de la mortalité en cas de choc septique, notre étude ne se limite pas à anticiper le résultat. Elle cherche également à élucider les raisons pour lesquelles un traitement a un impact spécifique et comment cet impact est influencé par d'autres facteurs. Les modèles prédictifs classiques, bien qu'ils atteignent une grande précision, peinent à répondre à des

questions de causalité, telles que « quel serait l'impact si nous administrions un seul antibiotique à un patient ayant un taux de glucose 'très bas' ? » Notre méthode consistant à décomposer l'*ATE* en segments flous nous permet d'aborder ces problématiques.

En somme, les recherches utilisant le concept de flou en médecine se sont généralement focalisées sur le diagnostic ou la prédiction. En intégrant clairement notre travail dans une approche d'inférence causale et d'interprétabilité, nous ouvrons de nouvelles avenues pour l'application de la logique floue dans l'évaluation de l'efficacité des traitements et la formulation de directives cliniques plus subtiles.

11.6 Les liens et les implications

Les relations entre les divers composants du système sont à la fois interdépendantes et complémentaires :

- Les interprétations sont basées sur des données prétraitées et calibrées.
- Le module *PFL* organise l'espace des caractéristiques en regroupements interprétables et propose les probabilités conditionnelles orientant les règles.
- Le *LightGBM*, amélioré par les attributs flous du *PFL*, conserve une performance prédictive robuste. Le modèle fusionné tire parti des avantages du *LightGBM* et des règles floues pour fournir cette performance.
- Les outils d'interprétation (règles des regroupements) et de diagnostic (cohérence, calibration) apportent la transparence requise pour l'acceptation en milieu clinique.

Les conséquences cliniques revêtent une grande importance. Un tel dispositif pourrait :

- Aider à la prise de décision : Les praticiens pourraient rapidement repérer les patients à haut risque et saisir les éléments primordiaux derrière la prédiction.
- Améliorer l'éducation et la formation : Le système peut agir en tant qu'instrument éducatif pour les jeunes médecins, en démontrant les profils de risque et l'influence des caractéristiques.
- Faciliter une médecine sur mesure : En détectant des groupes de risque particuliers, le traitement pourrait être personnalisé et ajusté en fonction des traits distinctifs d'un patient.
- Solidifier la confiance en l'IA médicale : Le système, par sa clarté et son interprétabilité, facilite l'audit et favorise l'adhésion des cliniciens et du grand public.
- Améliorer l'attribution des ressources : En ciblant plus précisément les patients à risque élevé, il serait possible de répartir les ressources médicales de manière plus performante.

Pour l'analyse causale, les relations entre les divers résultats sont inhérentes. L'ajustement causal est justifié par l'identification des variables confondantes. Le *BGMM*, grâce à la segmentation floue des confondants, met en évidence le caractère hétérogène des effets de causalité. En normalisant ces effets, le *NFATE* fournit une mesure globale pour évaluer l'impact des différents facteurs confondants. Les mesures de flou et d'incertitude rendent compte de la robustesse des segmentations et des inférences.

Les conséquences cliniques ont une grande importance. Il serait bénéfique pour les cliniciens d'avoir des recommandations plus détaillées, comme : « pour les patients présentant un pH artériel dans la fourchette 'faible', l'administration d'un unique antibiotique pourrait avoir des effets distincts comparativement à ceux situés dans la fourchette 'élevée' ». Cela encourage une approche médicale plus personnalisée et exacte. De plus, une IA qui est interprétable par ses degrés d'appartenance flous et ses *ATE* ajustés est plus susceptible d'être adoptée par les professionnels de la santé.

11.7 La synthèse des arguments développés

L'approche d'interprétabilité du système proposé va au-delà des simples explications (ex. *LIME, SHAP*), ce qui le distingue. Il fusionne la force prédictive des modèles arborescents avec la clarté des règles floues et la fiabilité des probabilités ajustées. L'aptitude à transformer les règles en langage dénormalisé et à les qualifier sur le plan linguistique permet de réduire l'écart entre le résultat du modèle et le langage clinique. Finalement, le dispositif d'alerte de cohérence est prépondérant pour détecter les situations complexes où l'interprétation symbolique est plus ardue, favorisant ainsi une démarche décisionnelle plus nuancée et plus sécurisée.

In fine, les points soulevés tendent à confirmer notre proposition : l'augmentation du symbolisme grâce à la logique floue, combinée avec les méthodes probabilistes, semble être une approche prometteuse pour l'inférence causale en médecine. Cette approche hybride offre non seulement la possibilité de distinguer les causes des simples corrélations, mais également de modéliser et quantifier l'incertitude inhérente aux données sensibles. Les conclusions tirées à partir des données du jeu *MIMIC-III* concernant le choc septique en témoignent clairement.

11.8 L’affirmation de la validité de la thèse

La théorie selon laquelle des approches combinant probabilités, logique floue et modèles arborescents peuvent parvenir à une harmonie d'interprétabilité, de robustesse, de flexibilité et de causalité en imitant le processus cognitif est fermement soutenue par les données. L'ajustement précis des probabilités témoigne de la robustesse et de la crédibilité. L'élaboration de règles floues dénormalisées et qualifiées témoigne de l'interprétabilité et de la capacité d'adaptation. La faculté de repérer des sous-groupes complexes indique un progrès vers l’intellection causale et l'aptitude à la tentative de représentation du raisonnement. Cette IA intelligible, interprétable, robuste, généralisable, flexible et raisonnante représente l'optique de transparence, de responsabilité, de confiance, d'éthique, de conformité réglementaire, de sécurité, d'auditabilité, de certifiabilité, d'économie et de souveraineté.

De plus, la solidité de notre argument est prouvée par l'aptitude de notre modèle à :

1. Élaborer des *ATE* ajustés exposant des impacts causaux variés selon les niveaux indéfinis des confondants, une compétence absente dans les méthodes *ATE* non indéterminées.
2. Reconnaître et établir un ordre d'importance parmi les confondants les plus influents à l'aide du *NFATE*, fournissant des renseignements capitaux pour la pratique clinique.
3. Évaluer ce qui est incertain dans les clusters contribue à accroître la transparence et la confiance dans les conclusions.
4. Préserver la cohérence globale de l'*ATE* manuel avec la moyenne des *ATE-PFL* ajustés, garantissant ainsi la validité des ajustements flous.

11.9 La justification

Cette partie vise à situer les résultats obtenus dans un contexte plus vaste, à défendre les décisions techniques prises et à souligner l'importance de cette méthode pour le secteur de l'IA en environnement critique. En montrant la compétence du système à fournir des interprétations appropriées à divers niveaux d'abstraction et à détecter ses propres restrictions (alertes de cohérence), notre objectif est d'établir une confiance avec les utilisateurs finaux. L'IA pourra être totalement intégrée et avoir un effet bénéfique et éthique dans des contextes décisifs, tels que le choc septique, uniquement en dépassant les simples indicateurs de performance.

11.10 Les apports

La valeur scientifique significative de ce projet découle de l'élaboration et de l'approbation d'une architecture de modèle avancé et hybride qui s'attaquent à la fois aux enjeux de la performance prédictive et de l'intellection approfondie.

1. ***PFL-LightGBM* calibré en tant que modèle hybride :** La démonstration d'une calibration *ECE* presque idéale associée à de robustes performances au niveau du *ROC-AUC* représente un progrès notable en matière de fiabilité des probabilités dans un contexte sensible.
2. **Règles floues dénormalisées et qualifiées sur le plan linguistique :** L'approche de création de règles combinant des intervalles de valeurs concrètes et des descripteurs linguistiques constitue une contribution directe à la facilité d'analyse pour les spécialistes du domaine.
3. **Plateforme de simulation de causalité :** L'élaboration d'un simulateur offrant la possibilité d'examiner l'effet d'interventions théoriques sur le profil de risque d'un patient. Cela participe au progrès des techniques d'interprétation causale.

Cette méthode fournit une structure pour l'élaboration de l'IA qui non seulement anticipe avec exactitude, mais est également capable d'interpréter ses choix de manière à tenter de raisonner comme la logique humaine, encourageant ainsi l'utilisation éthique et responsable des systèmes d'IA dans des situations à haut risque.

Manifestement, notre apport se manifeste à travers l'élaboration et la validation empirique d'un cadre d'inférence causale qui intègrent la *PFL* de manière explicite afin de traiter l'incertitude des données médicales. Ce modèle, illustré par le calcul des *ATE-PFL* et du *NFATE*, fournit une perspective inédite sur l'hétérogénéité des effets de traitement, facilitant une intellection approfondie des mécanismes causaux sous-jacents. Ceci constitue un progrès vers une IA plus intelligible, analysable, robuste et réfléchissante, apte à appuyer des choix décisifs dans des contextes complexes, tels que les unités de soins intensifs.

11.11 Conclusion

La discussion a souligné la robustesse, la capacité d'interprétation et l'importance clinique du modèle *PFL*-Arborescent. En associant les résultats aux théories préexistantes et en mettant en évidence les

conséquences pratiques, nous avons corroboré la pertinence de cette méthode du point de vue éthique. Ces progrès facilitent une meilleure interprétation des décisions de l'IA et favorisent une intégration plus aisée dans les processus de travail clinique. De plus, la discussion met en évidence l'importance du *PFL* pour perfectionner l'inférence causale. Les résultats confirment l'aptitude de notre modèle à apporter des conclusions plus raffinées et à satisfaire les critères d'une IA fiable.

Finalement, il est maintenant nécessaire de résumer tous les apports et de tracer les voies futures pour cette étude, dans une perspective d'extension et d'individualisation constante. En dernière partie, la conclusion synthétisera l'impact scientifique global de ce travail, en examinant son étendue et en suggérant des directions pour les recherches à venir.

CHAPITRE 12 :
CONCLUSION GÉNÉRALE : RÉSUMÉ, PORTÉE, PERSPECTIVES ET LIMITES

12.1 Introduction

Cette dernière section synthétise les apports majeurs du travail exposé, récapitule son impact scientifique et examine l'étendue de la recherche. Elle résume les contributions majeures de ce travail de recherche, en établit un compte rendu par section, mesure la portée des découvertes et suggère des orientations pour des recherches à venir. Elle souligne à nouveau l'importance de notre démarche visant à promouvoir une IA plus intelligible, interprétable et valorisante dans un cadre sensible. Elle présente également des suggestions pour des extensions, aborde les conséquences pratiques et met l'accent sur les recommandations, les perspectives et les contraintes du système. L'idée est de renforcer la notion d'une IA adaptable, centrée sur l'utilisateur et conforme aux valeurs vitales.

12.2 Le bilan par partie

Le document a exposé une approche d'interprétabilité enrichie par la causalité en IA, structurée autour de plusieurs éléments clés :

Introduction : Nous avons souligné l'importance incontournable d'une IA interprétable et apte à différencier la causalité des corrélations dans le contexte complexe du choc septique. Ceci justifie l'hybridation et la sophistication du symbolisme de la *PFL* pour traiter l'incertitude et l'imprécision des données sensibles.

Revue de la littérature : Cette partie a souligné le manque dans les études actuelles portant sur nos intégrations du *PFL*. Nous avons illustré comment les recherches des scientifiques dans divers paradigmes constituent les bases théoriques et pratiques sur lesquelles repose et évolue notre démarche.

Méthodologie, implémentation et expérimentations : Nous avons élaboré un protocole minutieux, comprenant la préparation des données, l'identification des valeurs aberrantes, l'évaluation manuelle de

l'*ATE*, la fuzzification en utilisant *GMM* et *Kmeans BGMM*, l'association du *PFL* avec *LightGBM*, le calcul des *ATE-PFL* ajustés et la mise en œuvre du *NFATE*. L'accent a été mis sur la transparence et la reproductibilité de chaque étape, tout en reconnaissant les limites inhérentes à la manipulation de données dans un domaine critique.

Analyse et synthèse : Nous avons posé les fondements de la modélisation, en passant par le traitement minutieux des données, y compris l'estimation des valeurs absentes et l'élaboration des ensembles de données stratifiés. Il a été souligné que la collecte et l'entreposage des données statistiques réelles sont centraux pour une analyse causale et une future dénormalisation des règles interprétables.

Résultats : Cette partie a souligné les bonnes performances du modèle *LightGBM* ajusté, affichant un *ECE* nul sur l'ensemble des sous-groupes, assurant ainsi des probabilités dignes de confiance. Il a été remarquable de constater un taux de rappel élevé pour le sous-groupe A=1 (82.73%). Les règles floues, dénormalisées et enrichies de qualificatifs linguistiques, ont démontré leur utilité pour une interprétation clinique directe. L'outil d'alerte de cohérence a décelé des écarts notables dans certains regroupements (tel que le Cluster 3 pour A=1 avec un différentiel de 15,9%), indiquant des groupes requérant une surveillance accrue. Finalement, la règle globale intégrée a présenté une perspective condensée et intelligible du fonctionnement global du modèle. De plus, les résultats de l'analyse causale ont été partagés, mettant en lumière la compétence de l'*ATE-PFL* à ajuster l'effet causal selon les valeurs indéterminées des variables confondantes, le repérage des variables confondantes nécessaires grâce au *NFATE* (*crp, pH arterial, temperature*) et la mesure du flou et de l'incertitude dans les segmentations. Ces résultats sont objectifs et soutenus par des mesures quantitatives.

Discussion : Nous avons confronté, mis en évidence la manière dont le *PFL* comble l'écart entre les systèmes experts interprétables, bien que stricts et les modèles connexionnistes (par exemple, les *DL* efficaces, mais difficiles à comprendre). On a souligné l'importance de la calibration pour la crédibilité clinique et l'utilité du système d'alerte de cohérence du point de vue d'un audit. Les conséquences cliniques, comme l'assistance à la prise de décision sur mesure et la transparence, ont été exposées en détail. L'approche de simulation causale a été mise en avant comme un outil précieux pour évaluer l'effet des interventions, déplaçant ainsi l'attention de la simple corrélation vers une intellection causale. De plus, nous avons analysé d'autres résultats, mettant en évidence la manière dont notre méthode dépasse les contraintes des modèles conventionnels en fournissant une interprétation plus détaillée des effets

causaux hétérogènes. Les implications cliniques pour une médecine personnalisée et une IA plus fiable ont été explorées.

12.3 Le résumé de l'apport scientifique

Le principal apport scientifique de cette étude est l'élaboration et la démonstration d'un cadre hybride et sophistiqué d'IA destiné aux contextes critiques, se concentrant sur l'interprétabilité symbolique et probabiliste avec l'utilisation d'un modèle arborescent. Ce cadre propose :

- **Une modélisation robuste de l'incertitude :** Grâce à l'utilisation des *GMM* et de la *PFL*, le système aborde intrinsèquement les imprécisions des données prépondérantes, un enjeu caractéristique dans ce domaine.
- **Des probabilités ajustées et crédibles :** L'utilisation de la calibration isotonique a abouti à des *ECE* nulles, garantissant que les probabilités de risque générées par le modèle sont directement exploitables et dignes de confiance pour les cliniciens.
- **Des interprétations intelligibles par l'humain :** Les *« IF-THEN »* dénormalisées et étiquetées sur le plan linguistique (« plutôt basse », « large », etc.) offrent aux experts en santé la possibilité de saisir intuitivement les éléments influents sur le risque.
- **Un système d'alerte pour les zones ambiguës :** L'établissement d'un mécanisme de détection des discordances entre les prédictions du modèle arborescent et les règles floues constitue une avancée vers l'amélioration des perspectives d'audit et de sécurité. Il indique les situations où le modèle pourrait agir comme une boîte noire, appelant à une prudence dans la pratique clinique.
- **Une approche axée sur la causalité :** Le simulateur de causalité donne la possibilité d'examiner l'effet supposé des actions, fournissant une vision des relations causales plutôt que de simples corrélations.

Ces apports visent à construire une vision d'IA éthique et responsable, en harmonie avec les principes d'intelligibilité, de transparence et de confiance.

Un autre apport scientifique de cette étude réside dans l'élaboration et la validation d'un cadre d'inférence causale : la *PFL* pour la causalité. Ce dispositif permet de :

- **Modéliser l'incertitude et l'imprécision :** Par l'utilisation de la logique floue et du *BGMM*, des idées médicales indéterminées sont converties en niveaux d'appartenance mesurables, représentant de manière plus précise la réalité clinique.

- **Mesurer les effets causaux hétérogènes :** Les *ATE-PFL* ajustés montrent comment l'impact du traitement fluctue selon les niveaux (bas, moyen, haut) des facteurs confondants, fournissant une précision inaccessible avec les approches standards d'*ATE*.
- **Déterminer les confondants majeurs en utilisant le *NFATE* :** Le *NFATE* offre une mesure standardisée et intelligible pour classer l'impact causal des variables, simplifiant ainsi l'identification des interventions pressantes.
- **Améliorer l'interprétabilité et la transparence de l'IA :** En présentant les résultats sous forme de « segments flous » et en mesurant la vague et incertain, notre IA devient plus intelligible et interprétable, conformément aux critères d'une IA fiable et digne de confiance.

12.4 La portée de la recherche

Tenter de représenter la dimension cognitive du raisonnement par l'hybridation et la sophistication du symbolisme

Nous avons déterminé comment l'ensemble de l'hybridation et de l'enrichissement (sophistication) du symbolisme pour une IA tente de simuler le raisonnement et améliore la capacité d'interpréter les prédictions et de distinguer les causalités des corrélations dans une perspective de valorisation. Dès lors, nous avons répondu à notre question cognitive (Qc). Elle a aussitôt guidé notre approche. Cet essai de simulation des processus cognitifs avec la *PFL* diminue les risques, découvre de nouvelles connaissances d'une manière valorisante et permet de prendre des décisions plus éclairées en utilisant l'analyse causale avec l'*IAI*. Donc, notre hypothèse cognitive (Hc) est validée. Cette corroboration a rendu possible la démonstration, via l'étude de cas sur le choc septique, que l'intégration du raisonnement causal et de l'*IAI* permet de valoriser la solution comme un outil de collaboration Homme-Machine. Nous avons prouvé que cette approche facilite la prise de décision en contextualisant les prédictions et en révélant les mécanismes causaux sous-jacents, renforçant ainsi la confiance des experts et une prise en charge personnalisée des patients, d'où l'aboutissement à notre objectif cognitif (Oc). La validation de l'hypothèse (Hc) par l'atteinte de l'objectif (Oc) se matérialise dans notre contribution à une nouvelle vision de l'IA, transcendant la simple corrélation pour s'engager dans un raisonnement causal plus profond et compréhensible. De la sorte, en s'inspirant du raisonnement, notre travail enrichit l'*IAI* en offrant une analyse causale. Cela permet non seulement de valider théoriquement l'approche, mais aussi de poser les fondations d'une IA valorisante. Ainsi, ceci était notre contribution cognitive (Cc).

Enrichir l'interprétabilité avec la causalité via l'informatique de la *PFL*

Nous avons concentré notre travail sur la conceptualisation des *PFL* intégrant du probabilisme et du flou pour traiter les nuances causales, générer des prédictions précises et produire des règles interprétables. De ce fait, nous avons répondu à notre question informatique (Qi). Pour y parvenir, nous avons validé qu'une *PFL* de fuzzification basée sur un clustering et un modèle arborescent dont les entrées sont enrichies par les scores flous et les probabilités des clusters permet une recherche d'équilibre robuste, flexible et généralisable en obtenant un bon rapprochement d'une harmonie de précision, de causalité, et de règles interprétables. De la sorte, nous avons validé notre hypothèse informatique (Hi). Cette approbation a été la clé pour implémenter un *PFL*-Arborescent capable de générer des « *IF-THEN* » interprétables. Notre but était de parvenir à un équilibre entre de bonnes précisions prédictives, la calibration des probabilités de risque (avec l'exemple du choc septique). Aussi, ce devait être un modèle personnalisé et ajusté pour s'adapter à des contextes sensibles pour une perspective valorisante. En procédant ainsi, notre objectif informatique (Oi) était atteint. L'implémentation réussie de cet objectif (Oi) a mené directement à notre présentation d'une hybridation utile combinant un clustering pour la fuzzification, un modèle causal, un autre arborescent produisant des « *IF-THEN* » ainsi que des métriques pour évaluer la causalité et l'*IAI*. Par conséquent, ceci était notre contribution informatique (Ci).

Tout bien considéré, la symbiose et la progression des liens cognitifs et informatiques : Qc,i → Hc,i → Oc,i → Cc,i forment une chaîne logique cohérente où chaque élément dépend du précédent. Successivement, les réponses à nos questions de recherche dépassent la corrélation et l'explicabilité. Ultérieurement, la validation de nos hypothèses permet de considérer la *PFL* comme un miroir de la cognition. Par la suite, nos objectifs acquiescent à une IA valorisante. Au bout du compte, nos contributions enrichissent l'interprétabilité avec de la causalité.

En définitive, cette étude va au-delà de la simple question du choc septique. Les méthodes et les principes élaborés peuvent être utilisés dans d'autres secteurs indispensables où l'évaluation du risque est vitale et où la capacité d'interprétation revêt une importance capitale. De plus, la démarche de cette étude peut être appliquée à d'autres secteurs complexes où les données sont vagues et où la détermination des liens de cause à effet est primordiale. Elle propose un cadre pour développer des IA qui tentent de reproduire le raisonnement causal, diminuant ainsi les dangers associés à des décisions fondées sur de simples corrélations. Outre le domaine médical, cette méthode est applicable à toute utilisation de l'IA où les

répercussions d'une erreur sont significatives et où des analyses sont indispensables pour obtenir l'adhésion des spécialistes concernés (ex. en finance, ingénierie, défense).

12.5 La justification

Cette conclusion vise à renforcer l'intégralité du travail, à renforcer l'argumentation pour ce genre d'IA, à mettre en lumière les efforts déployés et à tracer des orientations pour le futur. Également, souligner l'aspect unique et l'influence de nos apports, tout en établissant un plan clair pour le futur. Il est temps de mettre en lien tous les éléments du casse-tête et de réitérer l'importance requise de notre démarche vis-à-vis de la science et de la société. En mettant l'accent sur la personnalisation et les principes éthiques, nous confirmons notre dévouement à une IA qui est utile, simple et novatrice au profit de l'homme.

12.6 Les implications

Cette recherche a de nombreuses implications :

Technologiques : Prouve la viabilité d'architectures d'IA complexes et hybrides alliant intelligibilité et performance.

Cliniques : Trace le chemin pour des outils d'aide à la décision plus fiables, transparents et personnalisés pour les problèmes critiques.

Éthiques et réglementaires : Propose un modèle intrinsèquement conçu pour aborder les questions grandissantes de responsabilité et d'audit des systèmes d'IA dans des contextes sensibles.

Scientifiques : Ouvre de nouvelles perspectives de recherche à la croisée de l'interprétabilité, la causalité, la sophistication du symbolisme, la *PFL*, l'hybridation, et la complexité.

12.7 La proposition de prolongements

Nous pouvons envisager diverses voies d'amélioration pour accroître la robustesse et l'applicabilité du système :

Inclusion de modules explicables : Même si *LIME* propose des explications à l'échelle locale, *SHAP* fournit une perspective d'ensemble sur les contributions des caractéristiques à l'ensemble des

données ou pour des sous-ensembles significatifs, apportant ainsi un complément à l'analyse par regroupement du *PFL*.

Affinement des hyperparamètres du *PFL* : Une étude plus poussée sur le nombre idéal de clusters (n_clusters), les paramètres du *GMM*, ainsi que les poids des caractéristiques explicables pourraient renforcer la cohérence et l'efficacité.

Gestion des relations causales plus interprétables : Étudier des techniques causales plus élaborées pour établir des connexions causales directes à partir des données, au-delà du modèle contrefactuel afin d'ajouter une dimension plus robuste et causale aux règles floues.

Prise en compte de la causalité temporelle : Développer le modèle afin d'inclure l'aspect temporel, ce qui facilite l'évaluation de l'impact causal des traitements à divers stades de la maladie.

Interactions homme-machine : Concevoir une interface utilisateur interactive offrant aux cliniciens la possibilité de parcourir les règles floues, d'observer les incohérences et d'effectuer des simulations causales de façon plus intuitive.

Homologation clinique prospective : La mise en œuvre et l'examen du système dans un cadre clinique authentique seraient des étapes élémentaires pour confirmer son efficacité et son acceptation.

Analyse des contraintes de la généralisation : Même si la stratification basée sur le patient diminue les dangers de surajustement, une vérification sur des groupes externes et diverses catégories de patients est indispensable pour valider l'universalité du modèle.

Adaptation à d'autres données : Le modèle pourrait être élargi afin d'inclure des données complexes, en se servant de modèles qui saisissent les dynamiques à long terme des paramètres dans un environnement critique.

Validation externe : Évaluer la méthode de travail sur d'autres ensembles de données cliniques ou différentes maladies afin de vérifier cette universalité.

Étude de diverses fonctions d'appartenance floues : Tester d'autres formes de fonctions d'appartenance (trapézoïdales, gaussiennes, etc.) ou un nombre de groupes différent pour le *BGMM* dans le but d'examiner leur influence sur les *ATE-PFL* et le *NFATE*.

Participation à la conception d'outils informatiques : Développer une librairie open source s'appuyant sur les classes *ATE* et *FuzzyBGMM* pour favoriser l'acceptation de la méthode de travail au sein de la communauté scientifique.

Analyse de sensibilité : Effectuer des études de sensibilité pour juger de la solidité des résultats en fonction des changements dans les hyperparamètres du modèle et les décisions concernant le prétraitement des données.

12.8 Les limites, les recommandations et les perspectives

Limites

Selon Musanga et coll. (2025), même si le jeu de données offre une multitude de cas, il ne concerne que les patients d'une seule zone géographique et cela pourrait engendrer des biais éventuels influençant la généralisation du modèle lorsqu'il est mis en œuvre sur des populations de diverses origines ethniques, tranches d'âge ou ensembles de données distincts. Selon lui, il faut considérer les biais en fonction de l'âge et du sexe où l'ensemble des données ne présente pas explicitement des distributions basées sur l'âge et le genre, ce qui pourrait influencer les performances des modèles si certaines données démographiques sont sous-représentées. En premier, pour l'exemple du sepsis, les modèles actuels de prédiction de l'IA pour diagnostiquer cette maladie présentent des risques majeurs de biais (Schinkel et coll., 2019). Subséquemment, les performances restent dépendantes de la connaissance du domaine et des biais des ensembles de données (Morandé et Tewari, 2023). De même, sans compter qu'on pourrait faire affaire aux faux survivants. Autant, les faux survivants seraient ceux qui auraient choisi d'arrêter un traitement et sont décédés peu de temps après leur sortie (Xu et coll., 2022). De plus, il y a des défis de la qualité du peu de données, les anomalies, ceux manquants (ex. dans le profil du patient) et leurs déséquilibres par rapport au nombre de cas de classes positives ou négatives.

Donc, la nécessité d'avoir des données d'entrée de haute qualité demeure une contrainte basique pour tous les systèmes d'IA. Il est indéniable qu'une collaboration avec des spécialistes du domaine est nécessaire. Également, il est décisif de collaborer afin de perfectionner les règles et confirmer les interprétations. Ainsi, une validation clinique en temps réel s'avère primaire pour attester de l'efficacité du système dans un contexte opérationnel. Par ailleurs, même si les règles floues essayent de générer la pensée, elles représentent une simplification des interactions complexes.

Successivement, dans le cadre de la logique floue, la conception des fonctions d'appartenance (par exemple, « malade » ou « très malade ») est limitée par une certaine subjectivité. Elles sont laissées à l'évaluation d'un expert. Deux spécialistes peuvent les interpréter de manière différente, conduisant à des résultats variés sans objectivité pour les définir. De surcroît, le flou peut être basé sur des *IF-THEN* établies par un expert et par conséquent, en raison de l'explosion combinatoire de ces règles avec un grand nombre de variables d'entrée, leur nombre requis pour englober toutes les interactions possibles augmente de manière exponentielle (« malédiction de la dimensionnalité »), rendant le système complexe à administrer. Tout bien considéré, en l'absence de mécanismes d'apprentissage intégrés, les systèmes flous ne tirent pas d'apprentissage des données. Ils transcrivent l'expertise en code. Ensuite, ils nécessitent un ajustement manuel si le contexte change.

En matière de probabilités, il existe une sensibilité aux présupposés. Dans le cadre de l'exemple du bayésien, la probabilité a posteriori (le résultat final) est étroitement liée à celle a priori (la croyance initiale). L'option de cette présomption est fréquemment subjective et peut déformer l'analyse, surtout en cas de manque de données. Tout comme, les suppositions d'indépendance ne sont pas toujours réalistes. Les modèles probabilistes (ex. bayésiens) présument l'indépendance entre les variables afin de simplifier les calculs. Dans la réalité, les symptômes sont généralement interconnectés, contredisant l'hypothèse fondamentale du modèle. Au bout du compte, il y a une incapacité à gérer l'imprécision. Les probabilités traitent de l'incertitude liée à un fait spécifique (ex. « il y a 70% de chances qu'il soit malade »), mais pas de l'indétermination d'un concept (ex. « il est plutôt malade »). Elles ne sont pas capables de bien représenter la progression d'un concept vague.

En dernier lieu, il y aura toujours une difficulté à la compréhension. Un modèle est bien interprétable s'il est petit (ex.: cinq règles). Cependant, un modèle avec des milliers de branches, bien que techniquement transparent (on peut lire toutes les règles), devient moins intelligible. Et encore, la compréhension causale reste difficile. Plus précisément, prenant l'exemple « si Mark avait eu une pneumonie, il n'aurait pas eu une maladie neurodégénérative », on modifie utopiquement l'issue d'un événement en modifiant l'une de ses causes. La *PFL* comme les autres IA sont incapables de faire cet exemple de raisonnement contrefactuel. Tout autant, concernant cette *PFL*, il existe un risque de confusion sémantique entre le niveau de véracité et la probabilité d'apparition. Par exemple, l'expression « Mark est probablement malade » n'a pas le même sens que « Mark soit un peu malade ». L'association crée une instabilité dans la sémantique du modèle. Aussi, la *PFL* présente une complexité en matière de modélisation. Il est

conceptuellement complexe de définir des modèles gérant à la fois des appartenances floues et des probabilités.

Recommandations:

Nous recommandons :

- **Surveillance continue :** Il est impératif de vérifier fréquemment la calibration du modèle arborescent avec de nouveaux patients. Si une détérioration de l'*ECE* est notée pour ce modèle, il est conseillé d'effectuer une recalibration régulière.
- **Utilisation stratégique de la cohérence :**
  Excellente cohérence (écart inférieur à 5% par regroupement) : En se basant sur les règles floues comme référence principale, le modèle arborescent valide les prédictions.
  Bonne cohérence (écart par cluster de 5 à 10%) : Il est nécessaire d'associer judicieusement les règles et le modèle arborescent, tout en maintenant une supervision rigoureuse.
  Mauvaise cohérence (> 10% de différence par groupe) : Favoriser les prédictions du modèle arborescent et explorer les interactions complexes ou ajuster les règles avec la compétence clinique. Il est donc conseillé de procéder à une étude détaillée des écarts par regroupement.
- **Partenariat clinique :** Il est indispensable de travailler en étroite coopération avec les spécialistes cliniques pour perfectionner les règles floues et confirmer les intellections.
- **Incorporation de visualisations explicatives :** Cette incorporation pourrait offrir une perspective supplémentaire sur l'importance des caractéristiques tant à un niveau global qu'à un niveau local.
- **L'exploration de la frugalité et de l'autonomie de l'IA :** Ceci est un enjeu majeur, cherchant à concevoir des modèles moins énergivores et des infrastructures de données décentralisées afin d'assurer la confidentialité et la conformité.

Par conséquent, nous suggérons que les recherches futures étudient l'effet de l'élimination des données manquantes et évaluent des techniques d'imputation plus évoluées. Une possibilité serait d'incorporer des réseaux bayésiens afin de représenter des relations plus complexes entre diverses variables confondantes et le traitement/résultat. Une des limites de notre recherche est le caractère rétrospectif de l'échantillon, qui, malgré sa taille considérable, provient d'une seule BD (*MIMIC-III*).

Perspectives

L'évaluation de nos prototypes sur d'autres contextes sensibles pourra confirmer leur généralisation, flexibilité et robustesse. Également, une collaboration avec des experts pour affiner les analyses d'interprétations des prédictions et des causalités permettra une adoption plus large de nos approches dans ces contextes avec peu de données, plus rares ou complexes. De plus, une coopération interdisciplinaire entre experts en IA, statistique et domaines applicatifs pourra accélérer le développement et l'adoption de méthodes hybrides et symboliques enrichies. Somme toute, l'intégration de ces prototypes dans des systèmes interactifs d'aide à la décision pourra grandement améliorer ces analyses en temps réel pour les utilisateurs (ex. les patients sur leurs mobiles) ainsi que les experts et la formation de ces derniers.

12.9 Conclusion

Pour la prise de décision éthique dans un contexte sensible, il y a deux éléments par rapport à l'utilisation de l'IA. Premièrement, l'IA peut avoir un impact positif sur les utilisateurs, les experts et l'environnement dans son ensemble. Deuxièmement, le principe d'interprétabilité, lié à l'IA, qui n'a pas non plus été suffisamment abordé et devrait être intégré plus fréquemment aux délibérations scientifiques. Ces deux éléments ont été débattus par Benzinger et coll. (2023) dans le domaine clinique pour les patients, les médecins ainsi que le personnel soignant et le système de santé.

Dès lors, si les données et les outils d'IA sont importants pour la créativité technologique, l'expertise humaine en connaissances, expériences, pratiques et ainsi que les valeurs morales restent pressantes pour expliquer les causalités, prendre des décisions garantes et diminuer les risques. Alors que ces risques font partie de l'environnement de l'humain, donc, pour la gestion dynamique de ces risques, l'hybridation et l'enrichissement du symbolisme en s'appuyant sur des exemples récents de la *PFL* avec un modèle arborescent est un projet recherchant un équilibre de ces valeurs d'analyse en interprétabilité, subtilités causales, robustesse, flexibilité, et précision. En conséquence, la *PFL* confectionne un cadre général favorisant l'aide à la décision avec une systématisation de la mise en œuvre des pièces de cet équilibre valorisant pour une éventuelle solution intelligible, éthique et transparente.

Aussi, le cerveau n'est notoirement pas seulement associatif. Il perçoit, de même, une succession d'événements. Pour ainsi dire, il ne discerne pas la causalité. Il la cherche. Sur le plan cognitif, il a besoin de la trouver. D'ailleurs, si selon Serge Robert, la causalité se relie à la catégorisation dans sa définition (ex. « bas », « normal » ou « élevé »), s'en suit, une solution IA essayant d'établir et de dépister cette causalité. Surtout, dans un contexte sensible avec plusieurs variables. C'est une solution d'hybridation et de sophistication du symbolisme, à l'instar de la *PFL*. Ce *PFL* est pourvu d'une logique floue émancipant l'analyse causale, en interprétant ce qui s'est passé, mieux qu'une logique classique. Dans la même veine *PFL*, ses probabilités, à l'exemple du bayésien, favorisent la prédiction. Les rôles de cette solution ne se résument pas seulement à l'apprentissage, à faire découvrir de nouvelles connaissances, au potentiel d'ajustement, de correction, et d'assistance à la prise de décision d'un expert. C'est un bon outil pour lui trouver des analyses d'interprétations des prédictions et des causalités. C'est, aussi, une solution de collaboration interactive entre cet expert et sa machine. Elle permet, de plus, dans son flux de travail, différents paliers de résultats évolutifs. En revanche, les processus cognitifs du cerveau n'ont accès à l'outil *PFL* qu'à travers la force informatique. Après quoi, cet expert peut raisonner en termes de logique *PFL*. Et, c'est finalement à lui seul qu'appartient la décision finale.

Également, il est important d'encourager l'utilisation des systèmes d'IA dans des domaines d'application sensibles (comme ceux de la santé). Donc, il est impératif que les gouvernements et les agences réglementaires prennent la responsabilité de protéger leurs citoyens et de minimiser les dommages potentiels pour les utilisateurs (Adamyk et coll., 2023). Ensuite, il est majeur de se rapprocher d'un équilibre entre différentes perspectives, comme celles des administrateurs et des techniciens, afin de construire un avenir d'IA qui soit à la fois novateur et moral. D'un côté, les critiques soulignent la nécessité de faire preuve de vigilance face aux dangers systémiques (biais algorithmiques, manque de transparence des décisions automatisées), nécessitant des mesures réglementaires pour sauvegarder les droits fondamentaux et éviter des préjudices à la société. D'un autre côté, les optimistes mettent en avant le potentiel révolutionnaire de l'IA (ex. pour des avancées médicales) et mettent en garde contre une régulation trop stricte pouvant freiner l'adoption de solutions vitales ou nuire à la compétitivité économique. En conséquence, une prépondérance de l'une des méthodes entraînerait des répercussions significatives. De surcroît, un risque de surrégulation en entravant l'innovation, en rendant l'accès aux technologies plus complexe ou en transférant les développements vers des juridictions moins restrictives, compromettant ainsi les normes éthiques mondiales. Sinon, une faible régulation avec un risque accru

d'abus (ex. dépendance à des systèmes non vérifiés) et érosion de la confiance des utilisateurs, qui pourrait à long terme compromettre l'acceptation sociale de l'IA.

De plus, l'incorporation de la déontologie et du sentiment de sécurité pose des exigences pratiques tout au long du cycle de vie du projet. Dès la phase initiale de conception, on doit procéder à l'examen des ensembles de données et des modèles du point de vue des biais et de l'équité, dans le but d'éviter toute forme de discrimination. Par ailleurs, les exigences réglementaires obligent à rendre les décisions prises par l'IA interprétables pour les utilisateurs, ce qui peut favoriser l'utilisation de modèles intelligibles et la création d'interfaces dédiées à l'affichage des prédictions. En outre, la sauvegarde des données est aussi un défi majeur, requérant l'implémentation de méthodes comme l'anonymisation afin de réduire les dangers de divulgation. Ensuite, pour garantir un projet fiable, des systèmes de suivi et de responsabilisation des parties prenantes devront être mis en place.

Par conséquent, l'approche médiane s'appuie sur des modèles enrichissants structurés sur les trois axes suivants. Premièrement, des régulations ciblées, fondées sur une évaluation des risques, imposent des exigences rigoureuses aux systèmes à fort enjeu (notamment en contexte sensible) tout en ménageant une certaine flexibilité pour les usages à faible risque. Deuxièmement, des mécanismes d'autorégulation (sous forme de certifications ou de labels de transparence) incitent les acteurs à adopter des pratiques garantes. Troisièmement, la co-construction de normes adaptatives réunit chercheurs, secteur privé et pouvoirs publics afin de garantir que ces standards évoluent de concert avec les avancées technologiques. Subséquemment, plutôt que d'opposer innovation et imputabilité, notre démarche de recherche d'équilibre vise à articuler les deux : la régulation devient alors un cadre favorisant le déploiement de l'IA au service de l'humain, sans sacrifier ni l'audace technologique ni les valeurs fondamentales.

Pareillement, l'incorporation de considérations valorisantes dans un projet d'IA constitue une approche à la fois déterminante et transformatrice. Cette méthode abordant les défis réglementaires et sociaux associés à l'IA, ne se restreint pas à une simple adhésion normative, mais transforme la manière dont un projet est envisagé, élaboré et mis en œuvre. Elle encourage à remettre en question les bases mêmes du projet, en s'interrogeant particulièrement sur les objectifs réels que l'IA doit atteindre. Il est donc décisif d'aller au-delà de l'accent mis sur la performance des algorithmes (précision, rapidité) pour prendre en compte des éléments tels que le partage et la protection de la vie privée, ce qui conduit à une perspective englobante aussi bien des paramètres techniques que non technique dans l'évaluation du succès.

Autant, à l'opposé d'une conception courante, les normes légales ne freinent pas la créativité, au contraire, elles encouragent l'innovation. Tout autant, un projet d'IA élaboré en suivant ces principes produit une valeur immatérielle envers les utilisateurs. Aussi, l'affichage des mesures (ex. du taux d'erreur) et l'obtention d'un consentement éclairé contribuant à renforcer l'engagement des utilisateurs, sont des moyens efficaces pour éviter d'éventuels risques.

En outre, notre approche entraîne une transformation évolutive et itérative des méthodes internes des équipes de développement. De surcroît, elle promeut l'interdisciplinarité, en rassemblant des aptitudes diversifiées pour traiter les problématiques techniques et de gestion de façon complémentaire et encourage une agilité clé dans un environnement réglementaire en perpétuel changement. Les développeurs doivent donc être formés à élaborer des systèmes modulables, et aptes à répondre aux nouvelles demandes des législations nationales ainsi qu'aux normes internationales.

Par ailleurs, il est donc décisif d'incorporer les valeurs stratégiques dans un projet d'IA, parce que cela permet de créer des solutions robustes, inclusives et durables, tout en évitant les pièges des technologies développées à la hâte. De ce fait, cette stratégie fait du projet un participant existentiel dans l'environnement de l'IA. En somme, l'approche ne transforme pas uniquement la méthode de développement technique, mais elle redéfinit également le projet lui-même en l'inscrivant dans une perspective à long terme où technologie et valeurs sont indissociables. En conséquence, elle sert d'instrument pour une innovation centrée sur l'humain et ses droits fondamentaux.

Identiquement, la valorisation de l'IA doit garantir que les systèmes décisionnels critiques restent à la fois intelligibles, c'est-à-dire capables de produire des décisions interprétables, permettant de mettre en évidence l'influence de chaque facteur sur le résultat. Pareillement, ils doivent être clairs, offrant des justifications robustes, c'est-à-dire suffisamment généralisables, pour résister aux variations des données d'entrée. Aussi, cette adaptabilité doit s'étendre dans d'autres contextes afin que les modèles conservent leurs performances hors de l'environnement d'entraînement. Par ailleurs, les algorithmes doivent tenter de représenter le raisonnement en appliquant une logique cohérente et transparente, pour que l'ensemble du processus soit accessible à l'examen. De plus, la chaîne de développement et de déploiement exige une pleine responsabilité et doit inspirer confiance en démontrant la fiabilité des résultats. Subséquemment, l'approche se veut pleinement éthique, respectant les principes de justice et d'équité et réglementaire, en conformité avec les normes en vigueur. Les mécanismes de défense intégrés

assurent la sécurité opérationnelle, tandis que des journaux de bord garantissent l'audit de chaque étape. De même, la mise en place de procédures de contrôle permet la certification formelle des systèmes, attestant de leur qualité et de leur conformité. Définitivement, pour limiter l'empreinte écologique, ces technologies doivent être frugales dans leur consommation de ressources et souveraines, en offrant la maîtrise pleine et entière des données et des modèles aux organisations qui les utilisent.

Ainsi, ce travail met en lumière des modèles d'IA pour la prédiction, axés sur l'amélioration de l'interprétabilité avec de l'analyse causale. En alliant la précision des modèles basés sur des arbres de décision à la clarté du *PFL* et à la fiabilité de l'étalonnage, nous pavons la route vers une IA plus intelligible et responsable. Le système propose des instruments de personnalisation, allant de la sélection du groupe de traitement à l'étude d'un scénario hypothétique, y compris l'interprétation des règles de risque prédominantes. Cette adaptation avec une IA interprétable, adaptable, robuste, généralisable respecte l'idée d'avoir une IA rationnelle, claire, fiable, éthique, réglementée, sécurisée, vérifiable, certifiable et économe. Nous contribuons à un futur où l'IA ne se contente pas de traiter des données, mais de faciliter la compréhension du monde, en essayant de symboliser la richesse du raisonnement pour améliorer la qualité de vie.

In fine, la valorisation de l'IA n'est pas un ensemble de règles figées, mais un processus dynamique devant évoluer selon des contextes différents et sensibles. Elle exige un dialogue collaboratif entre ingénieurs, philosophes, législateurs et citoyens afin de transformer l'IA en un levier d'émancipation plutôt qu'une simple recherche de précision. Il convient d'un gouvernail de valeurs pour les technologies IA afin de naviguer dans ces contextes avec sagesse et intelligence collective.

ANNEXE A
[Tableaux synthétiques]

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Introduction	Nous souhaitons situer cette étude dans le contexte de l'IA symbolique et hybride, en soulignant l'importance d'associer l'innovation technologique aux valeurs éthiques. Notre motivation est de développer une IA fiable capable de fournir des interprétations et une analyse causale, en particulier pour des systèmes à haut risque, comme dans le secteur de la santé. La contribution principale est de proposer une approche basée sur une logique floue	**Introduction:** L'IA ne possède pas de valeurs. Celles-ci sont conférées par son usage, ce qui rend un périmètre moral indispensable à son application. Pour les systèmes critiques, il est nécessaire d'avoir une IA avec des données non biaisées et qui adhère à des valeurs de principe. Ces valeurs incluent la transparence, la responsabilisation, la confiance, l'éthique, la conformité réglementaire, la sécurité, l'audit, la certification, la frugalité et la souveraineté. Une valorisation par ces valeurs est recherchée. **Contexte:** L'IA a transformé plusieurs domaines, y compris la santé. Cependant, les modèles opaques manquent d'*IAI* et d'analyse causale, entraînant un manque de confiance. Il est requis de créer des IA capables de proposer des interprétations claires et de saisir la causalité. L'étude de cas porte sur le choc septique, une condition médicale grave où les décisions ont un impact direct sur la vie des patients. **Problématique:** Le défi est d'allier efficacité algorithmique et exigences d'*IAI* et de causalité. Les méthodes actuelles se limitent souvent à des corrélations, sans distinguer les relations de cause à effet, ce qui peut conduire à des conclusions incorrectes. Les données sensibles sont caractérisées par l'incertitude et l'imprécision, que les modèles conventionnels peinent à traiter. Le but est de créer des IA qui tentent de représenter le raisonnement pour fournir des interprétations et gérer les incertitudes. Les modèles d'apprentissage profond et d'apprentissage automatique sont souvent des boîtes noires qui ne permettent pas de comprendre le processus de prédiction. Ce manque d'*IAI* est un obstacle majeur à leur adoption dans les domaines sensibles. **Justification:** La valeur de l'IA dans les contextes critiques dépend de sa capacité à interpréter (ex. avec des règles claires) et à être causale. Un expert ne peut se fier à une décision sans en saisir le processus sous-jacent, surtout lorsque des vies sont en jeu. Notre mission est de développer une IA capable de tenter de représenter le raisonnement de manière transparente et de traiter l'incertitude avec la *PFL*. Cette approche vise à concevoir une IA qui assiste l'humain en identifiant et minimisant les risques de façon proactive. **Importance du sujet:** Le projet a pour but de rendre l'IA plus intelligible en intégrant la *PFL* dans les modèles pour renforcer l'*IAI* et la capacité causale. L'association de la logique floue et des probabilités de cette représentation, en traduisant les concepts médicaux imprécis en règles *« IF-THEN »*. Cette étude	Les entrées de l'étude de cas sont des données médicales concernant le choc septique, qui se caractérisent par leur nature incertaine et imprécise. Ces données sont transformées en degrés d'appartenance flous, comme « faible », « moyen » ou « élevé ».	Les sorties attendues incluent des prédictions précises du risque de mortalité chez les patients atteints de choc septique. Nous visons aussi à produire des règles *« IF-THEN »* d'*IAI* qui interprètent la prédiction. Aussi, les sorties doivent permettre de différencier la causalité des corrélations et de fournir une analyse causale des effets du traitement.	La qualité de notre approche sera évaluée selon plusieurs critères. Le modèle doit conserver une bonne précision prédictive. Il doit être interprétable. La robustesse face aux données imprécises et la capacité de généralisation à d'autres contextes sont aussi des critères importants. Enfin, l'approche doit être conforme aux normes d'éthique et de valorisation.

	probabiliste (*PFL*) pour dépasser la simple corrélation et offrir une meilleure interprétation, une flexibilité, une plus grande robustesse et une capacité de généralisation.	vise à développer une IA adaptable qui vise le respect de valeurs fondamentales pour une intégration réussie. **Questions de recherche:** Qc : Comment l'hybridation du symbolisme, en essayant de représenter le raisonnement, peut-elle améliorer l'*IAI* des prédictions et la distinction entre causalité et corrélation? Qi : Comment concevoir une *PFL* pour gérer les nuances causales, générer des prédictions précises et produire des règles interprétables? **Hypothèses:** Hc : La tentative de générer des processus cognitifs par la *PFL* peut réduire les risques, découvrir de nouvelles connaissances et prendre des décisions éclairées grâce à l'analyse causale et à l'*IAI*. Hi : Une *PFL* basée sur un clustering et un modèle arborescent permettra de trouver un équilibre entre précisions, causalité et *IAI*. **Objectifs:** Oc : Démontrer que l'intégration du raisonnement causal et de l'*IAI* permet de valoriser l'IA comme un outil de collaboration Homme-Machine, en renforçant la confiance des experts. Oi : Implémenter un *PFL*-Arborescent capable de générer des règles interprétables, en cherchant l'équilibre optimal. **Contributions:** Cc : Une nouvelle vision de l'IA qui transcende la simple corrélation pour s'engager dans un raisonnement causal et intelligible, en s'inspirant du raisonnement. Ci : Présenter une sophistication et une hybridation utiles combinant un clustering pour la fuzzification, un modèle causal et un modèle arborescent produisant des règles, avec des métriques pour évaluer la causalité et l'*IAI*. **Cadre théorique:** Le cadre intègre l'hybridation et l'enrichissement du symbolisme (avec la *PFL*) pour analyser les données de contextes sensibles. Il combine des concepts de logique floue, de probabilités, de clustering, d'*IAI* et de causalité. Il vise à étendre les méthodes d'explicabilité et d'estimation de l'*ATE* à des contextes où les variables sont analysées graduellement. **État de l'art:** Une revue de la littérature a été effectuée pour comprendre l'état actuel de l'IA interprétable et explicable, de la causalité, de la logique floue et de leur application au choc septique. Elle a permis d'identifier le besoin de valoriser l'IA avec la sophistication et l'hybridation. **Conclusion:** Nous avons présenté le contexte, les motivations, les questions, les hypothèses, les objectifs et les contributions. Notre démarche vise à promouvoir une IA valorisante avec l'*IAI* et la causalité.			

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
INTERPRÉTABILITÉ	Nous souhaitons explorer les processus cognitifs, le symbolisme, le connexionnisme, les définitions, la classification, et les approches de l'*IAI* et de l'*XAI*. L'objectif est de motiver notre démarche en identifiant le manque de recherche en matière d'intégration du symbolisme pour optimiser l'*IAI* et l'analyse causale. Notre contribution sera de proposer une approche utile pour combler ce vide.	Cet examen de la littérature se concentre principalement sur l'*IAI*. • Le cerveau est un modèle prédictif efficace qui gère les données vagues et l'incertitude. Il minimise les erreurs de prédiction. Il interprète, s'adapte aux changements, définit des observables sur différents niveaux et utilise ses croyances pour anticiper son environnement. Il orchestre un équilibre de ces processus cognitifs. Nous estimons que le fonctionnement du cerveau s'apparente aux processus bayésiens. • Le symbolisme et le connexionnisme sont deux paradigmes clés pour comprendre l'*IAI* et la causalité. Le symbolisme est plébiscité en médecine, car ses systèmes, qui utilisent le raisonnement déductif, sont vérifiables et interprétables. Le connexionnisme, comme le *ML*, est efficace pour des performances de classification, mais il peut surajuster les données d'entraînement et perd en généralisation. La *PFL* enrichit le symbolisme en combinant les théories du flou et des probabilités pour atteindre un équilibre de causalité et d'interprétabilité. • L'*IAI* et l'*XA*I sont deux concepts distincts, mais liés. L'*XAI* est la mesure par laquelle un humain peut comprendre la cause d'une décision. Elle s'applique après la création du modèle (post hoc) pour expliquer ses rouages. L'*IAI* est intrinsèque et ses modèles sont conçus pour être intelligibles dès leur création (a priori). • La taxonomie des méthodes *IAI/XAI* se divise en plusieurs catégories, comme les explications globales contre locales, méthodes intrinsèquement interprétables contre post hoc, et méthodes spécifiques à un modèle contre indépendantes. • Les méthodes *IAI/XAI* sont variées. Elles comprennent les valeurs *SHAP,* l'algorithme *LIME, Grad-CAM, LRP, EBM, CBR, NLP, NeuroXAI, CIU, ANCHOR, t-SNE, UMAP, TraCE*, les cartes de saillance, les méthodes basées sur des exemples, les méthodes de substitution et de permutation, les graphes de connaissances, et les systèmes de règles. Les modèles intrinsèquement interprétables, comme la logique floue, permettent de générer des règles linguistiques. L'IA neuro-symbolique combine le meilleur des deux méthodes pour renforcer l'IAI et l'assignation de la responsabilité.	Les entrées des systèmes d'IA sont les données fournies pour l'*IAI*. Le cerveau, lui, prend en entrée des données sensorielles de son environnement. Les modèles d'IA interprétables ou explicables peuvent prendre en entrée des images, des vecteurs d'entrée, ou des textes cliniques.	Les sorties des modèles d'IA incluent des prédictions, des classifications, des décisions, des explications et des interprétations. Elles peuvent aussi être des cartes thermiques ou des règles. Le cerveau, quant à lui, génère des inférences et des prédictions à propos de son environnement.	Les critères de qualité de l'*IAI* sont la clarté et la fidélité. Les modèles doivent être transparents et leur fonctionnement interne doit être intelligible. Ils doivent aussi être robustes et avoir une bonne capacité de généralisation. La qualité est liée à l'aptitude à déceler et à atténuer les biais, ainsi qu'à la capacité de fournir des justifications qui renforcent la confiance des utilisateurs. La conformité réglementaire est également un critère primordial.

		• L'importance de l'*IAI* est primordiale pour les décisions à fort enjeu. Elle renforce la confiance et la transparence, assure la conformité réglementaire, facilite la détection des erreurs, offre une meilleure intellection des données et du contexte, établit une responsabilité partagée, et atténue les risques. • La comparaison des modèles se fait entre les boîtes blanches, noires et grises. Les boîtes blanches sont transparentes, mais manquent de précision. Les boîtes noires, comme les modèles *DL*, ont une grande précision, mais leur fonctionnement est opaque. Les boîtes grises offrent un équilibre entre la prédiction précise et l'interprétabilité décisionnelle.			

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
DES VALEURS EN MOUVEMENT	Nous souhaitons situer cette étude dans un contexte de défis liés aux biais des données, aux valeurs fondamentales et à l'implémentation de l'IA dans des secteurs sensibles. L'objectif est de montrer la tension entre le progrès technologique et les valeurs éthiques. Notre contribution est de démontrer que l'IA ne peut être fiable que si elle intègre la transparence, l'imputabilité, la confiance et d'autres valeurs capitales. Nous cherchons à concilier la technique et l'éthique pour une IA responsable.	Ce chapitre aborde les défis de l'IA en s'attardant sur les valeurs fondamentales. • **Définitions de valeurs en IA :** La transparence est l'intellection qu'a l'utilisateur des mécanismes de l'IA. La responsabilité est la capacité de suivre un résultat d'IA jusqu'à son auteur pour attribuer ses impacts. La confiance est la volonté de se fier à un système, ce qui exige des preuves vérifiées dans les contextes critiques. L'éthique s'appuie sur la bienveillance, la non-nuisance, l'autonomie et la justice. La conformité réglementaire est l'adhésion aux lois et standards, comme le RGPD ou l'AI Act de l'Union européenne. La sécurité garantit que les IA sont utilisées sans causer de dommages et résistent aux attaques. L'audit permet d'examiner les systèmes pour détecter les faiblesses. La certification est la validation selon une norme, comme l'ISO/IEC 42001. La frugalité optimise l'efficience des solutions d'IA en réduisant l'utilisation de ressources. La souveraineté fait référence au contrôle par un État des technologies et des données qui y sont liées. • **Données biaisées :** Les biais dans les données d'apprentissage peuvent compromettre la qualité et l'équité des systèmes d'IA. Un algorithme reproduira les biais présents dans les données d'apprentissage. Des erreurs peuvent survenir en raison des biais et des données manquantes. • IA intègre : Une IA digne de confiance doit être licite, et morale. Ces principes ont été présentés par la Commission européenne. L'interprétabilité est, aussi, un de ces principes. Ensuite, pour qu'une machine soit morale, elle doit être responsable. • Intelligibilité rime avec confiance et transparence : La transparence et la confiance sont vitales pour un déploiement responsable de l'IA. La transparence implique la communication sur les sources de données, les variables et les processus. L'OCDE et d'autres instances insistent sur la nécessité d'informations claires sur le fonctionnement des IA. Le RGPD et l'AI Act imposent des normes strictes de transparence et de responsabilité. Le programme DARPA cherche à développer des IA capables de justifier	Les entrées sont les données, souvent biaisées, utilisées pour l'entraînement des algorithmes décisionnels. En revanche, pour XTRACTIS®, les entrées sont des données, structurées ou non, de qualité inférieure ou en quantité limitée.	Les sorties attendues sont des résultats justes et des décisions fiables. Les systèmes d'IA doivent aussi pouvoir fournir des informations claires sur leurs prédictions et leurs limites. Les sorties doivent être intelligibles, contestables et justifiées. Pour XTRACTIS®, les sorties sont des rapports de prédiction avec une interprétation et une traçabilité.	Les critères de qualité incluent la licéité, l'éthique et la robustesse. Une IA de qualité est fiable, reproductible et exacte. Elle doit aussi être transparente, responsable, équitable et sécurisée. L'interprétabilité et la capacité à être audité sont des critères importants pour la traçabilité des décisions. Enfin, l'IA doit être frugale, souveraine et certifiable pour répondre aux normes et aux besoins spécifiques.

		leurs choix de manière claire et fiable. • Cadre réglementaire dans le contexte médical : Des cas concrets illustrent les défis éthiques et les lacunes en matière de responsabilité dans l'usage de l'IA en médecine. Un professionnel de santé doit conserver son jugement clinique et ne pas percevoir l'IA comme une réponse indépendante, mais plutôt comme un outil d'aide à la décision. • Symbiose entre éthique et progrès technologique : Il est urgent de garantir la transparence pour accorder la confiance aux décisions des IA. Une responsabilité claire doit être établie entre les intervenants. Une approche pluridisciplinaire est nécessaire pour incorporer les enjeux éthiques dans la conception des IA. • Exemples d'innovations technologiques validées : Le Sepsis ImmunoScore de Prenosis, validé par la *FDA*, est un exemple d'IA qui assiste les professionnels de santé dans la prise de décision pour le diagnostic du sepsis. XTRACTIS®, une IA symbolique et augmentée, est une autre innovation qui produit des décisions de confiance grâce à des modèles robustes, intelligibles et avec une bonne capacité prédictive.			

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
PFL	Nous cherchons à situer cette étude dans le domaine de la gestion de l'incertitude. L'objectif est de motiver l'usage de la *PFL* en tant que solution capable de gérer les incertitudes stochastiques et floues simultanément. Notre contribution est d'examiner les principes fondamentaux de la logique floue et du probabilisme, leurs relations et les progrès de la *PFL*, en soulignant son importance pour surmonter les défis de la donnée incertaine.	Ce chapitre présente la logique floue et le probabilisme ainsi que la *PFL*. • La logique floue a été introduite par Zadeh en 1965. Elle modélise l'imprécision en attribuant des degrés d'appartenance à des ensembles. Elle est caractérisée par sa capacité d'interprétation et d'analyse des nuances de causalité. Elle est un moyen de flexibilité, de représentation des connaissances incomplètes et de maintenance. Les modèles flous de Mamdani et de Takagi-Sugeno sont des exemples de cette logique. • Le probabilisme est une solution qui se veut tolérante à l'imprécision, adaptable et dotée d'une capacité d'apprentissage. La probabilité mesure l'incertitude de notre connaissance. Les algorithmes probabilistes se composent de ceux explorant des règles d'association, d'appariement probabiliste et de modèles graphiques. Les modèles graphiques probabilistes (*PGM*) comme le réseau bayésien (*BN*) sont des outils pour modéliser la causalité et améliorer l'interprétabilité. • La *PFL* est une solution pour l'interprétabilité et l'analyse causale. Elle combine les probabilités et le flou pour combler le fossé entre les deux. La *PFL* gère trois incertitudes : l'aléatoire, l'incertitude probabiliste et le flou. Elle est utilisée pour la recherche des causes des événements. Les travaux de plusieurs chercheurs présentent des exemples d'utilisation de la *PFL*.	Les entrées des modèles *PFL* sont des données contenant de l'incertitude probabiliste ou des imprécisions. Les données peuvent être transformées en grandeurs floues par fuzzification.	Les sorties des modèles *PFL* sont des prédictions, des règles, des distributions de probabilité et des interprétations. Ces sorties permettent de mieux gérer les interactions complexes et d'optimiser la fiabilité des prédictions.	La qualité d'un modèle flou s'évalue par sa capacité d'interprétation, sa précision et sa maintenabilité. L'interprétabilité des systèmes flous se mesure par le nombre de fonctions d'appartenance et de règles. Les modèles probabilistes sont évalués selon leur capacité à gérer l'incertitude statistique. La *PFL* est évaluée sur sa capacité à traiter les incertitudes, à être robuste et à fournir des interprétations pour la prédiction.

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
EXPLICATION CAUSALE	Nous voulons situer cette recherche dans la recherche d'une intellection et d'une *XAI* des phénomènes qui dépassent la simple corrélation. Nous sommes motivés par la nécessité de rendre les modèles d'IA non transparents intelligibles pour l'homme en révélant leurs mécanismes internes. Notre contribution est d'analyser la superposition de l'*XAI* scientifique, et du clustering bayésien pour surmonter les obstacles de l'analyse des données, en mettant en évidence les liens de causalité sous-jacents aux	Ce chapitre traite de la causalité, de l'*XAI* scientifique, de la causalité et du regroupement bayésien. • La causalité est un élément central de la cognition. Aristote a écrit qu'on a connaissance d'une chose seulement quand on en a compris la cause. Le raisonnement causal est la découverte de la relation entre une cause et ses effets. Il est une composante majeure de cette *XAI*. • Cette *XAI* est une réponse à la question « pourquoi », qui vise à reconnaître et à décrire la logique derrière l'apparition d'événements spécifiques. Le modèle SR de Salmon introduit la notion de pertinence statistique pour vérifier les processus influençant l'événement à expliquer. • L'explicabilité et la causalité sont étroitement liées. Carloni et coll. (2025) ont identifié trois perspectives sur ce lien : une critique de l'*XAI* due à l'absence de causalité, l'*XAI* comme outil d'exploration causale, et la causalité comme fondement pour l'*XAI*. La troisième perspective est jugée la plus prometteuse et considère que l'accès à un modèle causal est une interprétation en soi. • Le clustering bayésien (*BGM*) est un modèle qui gère l'incertitude grâce aux principes bayésiens. Il excelle dans la sélection automatique du nombre de clusters, gère les données rares ou bruitées, et fournit des probabilités d'appartenance pour chaque point de données.	Les entrées sont des données non labellisées pour les techniques de regroupement.	Les sorties attendues incluent des explications. Le regroupement bayésien fournit une évaluation du degré de certitude associé à chaque attribution d'un point de donnée à un groupe. Il fournit aussi des probabilités d'appartenance pour chaque point à chaque cluster. Les modèles bayésiens causaux fournissent une analyse décisionnelle cohérente en cas d'incertitude.	Les critères de qualité sont la capacité du modèle à être fiable et à fournir des interprétations pour la prédiction. L'*XAI* scientifique doit s'appuyer sur des relations causales objectives et des mécanismes explicatifs de l'expert. Les *BGM* se distinguent par leurs simplicité, flexibilité, capacité de paramétrage, gestion de l'incertitude, robustesse et précision. L'approche doit être capable de réduire la distance entre l'*XAI* et la causalité.

	dynamiques du monde réel.				

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
ÉTUDE DE CAS POUR LE CHOC SEPTIQUE	Nous souhaitons situer cette étude dans la problématique de l'*IAI*, l'*XAI* de la causalité et des valeurs (ex. éthiques) pour l'intellection des prédictions d'une maladie comme le choc septique. Nous sommes motivés par la complexité de cette maladie et la nécessité d'une prise en charge rapide. Notre contribution est d'analyser l'implémentation de l'hybridation et de la sophistication de la *PFL* pour la gestion du choc septique, en mettant l'accent sur l'importance d'une démarche valorisante.	Ce chapitre présente une étude de cas sur le choc septique, en se concentrant sur les approches connexionnistes et la *PFL* symboliques. • le choc septique est une réaction sévère du corps qui peut causer des lésions ou le décès. C'est une urgence médicale avec un taux de mortalité important, un défi en matière de diagnostic et de prise en charge. La modélisation de son évolution est compliquée en raison des interactions non linéaires entre différents systèmes physiologiques. • Les approches connexionnistes (*ML & DL*) ont montré des résultats prometteurs pour la prédiction et le diagnostic précoce du choc septique. Ces modèles utilisent des signes vitaux, des résultats de laboratoire et d'autres facteurs cliniques pour prédire la probabilité du sepsis. Cependant, ils manquent d'interprétabilité et de causalité, ce qui limite leur acceptation en milieu clinique. • La *PFL* est une approche qui offre l'interprétabilité, la causalité, la flexibilité et la robustesse face aux données incertaines et imprécises. Les modèles flous ont été utilisés pour le diagnostic et la détection du choc septique. Les outils probabilistes sont aussi efficaces pour la prédiction de la mortalité. La *PFL* combine les bénéfices des deux approches pour traiter ces incertitudes, ce qui la rend appropriée pour les systèmes critiques de santé. • Considérations éthiques : L'acquisition et l'exploitation de données de santé sensibles posent des questions éthiques importantes, notamment la protection de la vie privée, le consentement et la prévention des biais algorithmiques. L'éthique doit être un pilier central pour le développement de toute solution IA dans les domaines critiques.	Les entrées sont les données cliniques, les DSE (Dossiers de Santé Électroniques), les signes vitaux (fréquence cardiaque, tension artérielle), et les résultats de laboratoire (nombre de globules blancs). Ces données peuvent être imparfaites et incohérentes.	Les sorties sont des prédictions, des diagnostics et des traitements pour le choc septique. Les modèles *PFL* produisent des estimations de risque de mortalité qui sont interprétables. Ils peuvent aussi donner des mesures de probabilité qui fournissent des informations supplémentaires sur lesquelles les experts peuvent agir.	Les critères de qualité incluent la validité et la généralisation des modèles. Les modèles doivent être robustes et interprétables, et ils doivent gérer l'incertitude et la variabilité. L'approche doit respecter un cadre éthique strict, garantir la confidentialité, prévenir les biais et assurer la transparence des décisions.

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Méthodologie	Nous visons à élaborer une méthodologie qui concilie précision prédictive, interprétabilité et analyse causale pour des systèmes d'IA. Notre motivation est de relever le défi des systèmes opaques en concevant des architectures hybrides. Nous contribuons en proposant un flux de travail distinctif pour développer une solution IA qui se rapproche d'un équilibre entre ces éléments, en nous inspirant du travail de Razab-Sekh (2021) et en adaptant les processus *CRISP-DM*. Notre but est de créer des modèles interactifs et évolutifs pour	• Analyse des besoins métiers : Notre recherche vise à résoudre le problème des systèmes d'IA opaques en équilibrant la précision prédictive avec l'interprétation et l'analyse causale. Cela est déterminant dans le domaine médical, où les résultats liés à une maladie comme le choc septique doivent être compréhensibles pour que les experts puissent faire confiance à l'IA et en assumer la responsabilité. • Interprétation des données : Cette phase explore et comprend les données. Nous utilisons des graphiques comme les histogrammes pour examiner la répartition des variables et leur relation avec la prédiction. Nous détectons les données absentes et les observations atypiques avec des techniques comme l'écart interquartile et l'Isolation Forest. o Éthique et source de données : Pour des raisons éthiques, nous avons obtenu l'autorisation d'utiliser les données anonymisées de *MIMIC-III*. Cette BD sensible de patients en unité de soins intensifs atteints de sepsis. Le risque de divulgation d'identité est très faible. Nous avons sélectionné 26 variables comme facteurs de confusion pour l'analyse causale. • Préparation des données : Cette phase inclut le nettoyage, la collection et la structuration des données. Pour les données manquantes, nous avons combiné des stratégies en utilisant un modèle arborescent *LightGBM* pour l'imputation de la variable de traitement et en supprimant les lignes avec des valeurs manquantes pour l'analyse des variables de confusion. • Modélisation : Nous construisons des modèles hybrides combinant la puissance prédictive des modèles arborescents avec la *PFL*. o La première architecture se concentre sur l'inférence causale en utilisant une pondération basée sur l'appartenance floue et un calcul manuel de l'*ATE* avec un mélange gaussien bayésien (*BGM*). Chaque variable est rendue floue sur trois segments ordonnés. L'*ATE* est calculé pour chaque segment et ajusté. o La deuxième architecture utilise une structure modulaire pour la prédiction, intégrant l'interprétabilité et la calibration. Elle utilise le clustering pour la fuzzification des variables, et les scores d'appartenance sont incorporés comme attributs dans un *LightGBM*. • Évaluation : Nous évaluons nos modèles en utilisant une exploration des données, des paramétrages variés, et des approches d'estimation comme la validation croisée en plusieurs phases. Nous utilisons des	Les entrées de notre méthodologie sont les données anonymisées de *MIMIC-III*. Pour l'analyse causale, seulement 26 variables cliniques ont été sélectionnées comme facteurs de confusion.	Les sorties sont des informations de modèles IA prédictifs, interprétables et causaux. Ces modèles produisent des règles *« IF-THEN »* intelligibles et dénormalisées, des prédictions de mortalité ajustées par calibration, et des alertes d'incohérence. Elles fournissent également une estimation de l'*ATE* pour différencier la causalité des corrélations.	Les critères de qualité sont la fiabilité, la robustesse, la généralisation et la cohérence. La fiabilité est assurée par la calibration des probabilités. La robustesse et la généralisation sont évaluées par la validation croisée. Un système d'alerte d'incohérence est également un critère de qualité. L'évaluation de la causalité est un critère central pour différencier les corrélations des causalités.

	aider l'expert à prendre des décisions éthiques.	mesures de performance (exactitude, précision, rappel, score F1, *ROC-AUC*) et l'*ECE* pour évaluer la fiabilité. o Calibration des probabilités : Les scores du modèle arborescent sont ajustés par régression isotonique pour assurer leur fiabilité. o Alertes d'incohérence : Un système d'alerte signale les divergences entre les prédictions et les règles, incitant à une intervention humaine. o Évaluation de la causalité : Nous estimons *l'ATE* avec un poids basé sur le degré d'appartenance flou et nous utilisons le *NFATE* pour différencier la causalité des corrélations. • Déploiement : L'objectif est de rendre la solution interactive et collaborative pour les spécialistes en exposant les prédictions, les règles *« IF-THEN »* et les notifications d'incohérence. • Justifications des méthodes retenues : o *MIMIC-III* est choisi pour sa nature anonymisée, son accessibilité et sa richesse de données. o La gestion des données manquantes est rigoureuse, en supprimant des lignes pour les variables de confusion et en utilisant un modèle performant pour l'imputation. o Le clustering pour la fuzzification est utilisé pour reproduire la logique humaine qui s'appuie sur des classifications non binaires. o Le modèle arborescent est choisi pour sa puissance prédictive, qui est renforcée par l'interprétabilité des scores d'appartenance flous. o La calibration et les alertes d'incohérence sont des dispositifs de sécurité qui renforcent la confiance des experts.			

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Implémentation	Nous avons mis en œuvre notre méthodologie pour prouver que les solutions proposées sont pratiques et reproductibles. Notre contribution est de fournir des algorithmes pour répondre à nos questions de recherche sur la causalité et l'interprétabilité. Nous cherchons à guider l'expert dans l'évaluation de l'impact d'un traitement tout en tenant compte de la variation des données.	Notre travail est basé sur le développement de plusieurs algorithmes. Certains se concentrent sur l'inférence causale, tandis que d'autres sur l'interprétabilité des prédictions. Le flux de travail causal flou est élaboré pour évaluer l'effet d'un traitement en considérant la variabilité des données. Nous avons conçu une classe *ATE* pour déterminer l'effet causal à l'aide d'une régression linéaire pondérée. Une classe *FuzzyBGMM* gère la « fuzzification » des variables en utilisant *BayesianGaussianMixture* de *Scikit-learn* pour attribuer des degrés d'appartenance à des groupes. L'architecture *PFL-LightGBM* est modulaire. Elle inclut des modules de configuration, de gestion des données, et un module *PFL* qui effectue le clustering flou. Les scores d'interprétabilité sont intégrés au modèle prédictif *LightGBM*. Une calibration isotonique est appliquée après l' entraînement. Le système de données est géré pour maintenir les échelles réelles des variables.	Les entrées des algorithmes sont les variables cliniques, la cible et la caractéristique de traitement. La classe *FuzzyBGMM* utilise *BayesianGaussianMixture* de *Scikit-learn* pour segmenter chaque variable. Le module de calibration utilise la régression isotonique pour ajuster les probabilités de sortie du modèle.	Les sorties des algorithmes sont les scores d'interprétabilité et les règles cliniques pour les experts. La classe *ATE* détermine l'effet causal moyen. L'architecture *PFL-LightGBM* produit une classification du risque par ensemble et des alertes en cas d'incohérence. La calibration isotonique ajuste les probabilités de sortie du modèle pour assurer leur fiabilité.	La qualité des solutions est assurée par leur caractère pratique et reproductible. L'architecture modulaire de la partie d'interprétabilité prédictive a été élaborée pour garantir la robustesse et la reproductibilité du système. L'utilisation du *BGMM* permet une segmentation probabiliste saisissant les subtilités de la variable. L'intégration des scores flous dans *LightGBM* dote le modèle d'une faculté prédictive tout en préservant une connexion directe avec l'interprétation. Les modules de calibration et d'alerte agissent comme des garde-fous, renforçant la fiabilité et la

					crédibilité des architectures.

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Expérimentations	Nous visons à valider notre stratégie d'IA hybride pour la prédiction, l'interprétation et l'analyse causale. Nous situons notre travail dans le contexte clinique du choc septique, en nous concentrant sur la prédiction de la mortalité à 72 heures. Nous contribuons à l'élaboration d'un système capable d'imiter les processus cognitifs, distinguant la causalité de la corrélation pour une utilisation dans des dispositions sensibles.	Nous élaborons des modèles intégrant l'analyse causale et l'interprétabilité. Nous nous basons sur les recherches de Faghihi et coll. (2020) pour l'analyse de la causalité, et sur celles de Fialho et coll. (2016) pour la prédiction et l'interprétabilité. Notre approche hybride combine un modèle de clustering flou et un modèle *LightGBM* pour établir des règles interprétables tout en maintenant une bonne précision prédictive. Nos expérimentations se concentrent sur la prédiction de la mortalité à 72 heures chez les patients en choc septique. Nous utilisons une permutation bayésienne pour identifier les sept variables les plus significatives pour la prédiction.	Nous utilisons trois jeux de données distincts : A=1 (patients ayant reçu un antibiotique unique), A=0 (ayant reçu plusieurs antibiotiques), et A=Tous (l'ensemble complet). Le jeu de données *MIMIC-III* sert de base, et nous gérons le chargement des données via un pipeline. Pour le protocole d'interprétation, nous incorporons des scores d'appartenance floue et des scores de fiabilité dans le processus d'entraînement.	Le système fournit des *ATE* pour chaque variable confondante, ainsi que des *ATE* ajustés et pondérés pour leurs nuances. Il génère des règles interprétables basées sur les centroïdes du clustering flou et les probabilités conditionnelles de décès. Il génère également des alertes lorsque l'écart de cohérence entre les prédictions et les règles dépasse un seuil de 10%, ce qui incite à une analyse approfondie.	Nous évaluons la performance par l'exactitude, l'équilibre de l'exactitude, le rappel, et le *ROC-AUC*, en utilisant une validation croisée à 5 plis pour minimiser les biais. La fiabilité des probabilités est évaluée par l'*ECE*, mesurée avant et après la calibration. L'interprétabilité est mesurée par la cohérence, l'écart entre les prédictions du modèle et les règles floues associées. Pour la causalité, nous calculons l'*ATE* pour chaque segment flou de la variable confondante.

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Analyse et synthèse	Nous situons la recherche dans le domaine de l'IA pour la santé, un secteur critique. Nous la motivons par le besoin de concevoir des IA non seulement précises, mais aussi causales et interprétables. Les modèles opaques ne sont plus suffisants dans des contextes vitaux comme le choc septique, car l'acceptation clinique et la confiance dépendent d'une bonne intellection des décisions. Nous contribuons en proposant une approche hybride et sophistiquée qui associe un modèle prédictif puissant avec des techniques d'interprétabilité pour fournir des	La première idée est l'utilisation d'un système hybride combinant la *PFL* et un modèle arborescent : *LightGBM*. L'objectif de cette association est de lier la performance prédictive à l'interprétabilité des résultats. Une autre idée centrale est que les modèles précis ne suffisent plus. Il est impératif d'avoir un entendement approfondi des processus de décision pour garantir la valorisation et l'acceptation de ces systèmes en milieu clinique. Nous soulignons également l'importance d'une analyse causale pour aller au-delà de la simple corrélation. Le choc septique est un cas d'étude pertinent en raison de sa complexité et de ses implications. L'approche floue est justifiée, par ce qu'elle permet de capturer l'incertitude et l'imprécision des données biologiques.	Les entrées sont les données cliniques de patients atteints de choc septique. Elles proviennent de *MIMIC-III*. Elles sont composées d'un ensemble d'événements cliniques transformés en un format adapté. Les variables incluent des paramètres, une variable de traitement 'A' (antibiotique unique vs multiples), et une variable cible 'y' (mortalité à 72 heures). Le nombre de caractéristiques utilisées est de 26, plus la variable 'A'.	Les sorties du système hybride sont des prédictions précises de mortalité. Elles sont accompagnées de règles de risque interprétables produites par la *PFL*. Ces règles, qui utilisent des scores d'interprétabilité et des diagnostics de cohérence, interprétant les prédictions du modèle. D'autres sorties incluent des scores bruts, comme le *ROC-AUC* et l'*ECE*, ainsi que des résultats d'analyse causale tels que les *ATE* et les *NFATE*.	Le premier critère est la performance prédictive du modèle, mesurée par le *ROC-AUC*. Le deuxième est la fiabilité des probabilités, évaluée par la réduction de l'*ECE* à une valeur proche de zéro après calibration. Nous estimons aussi la pertinence clinique des interprétations du modèle en les ancrant à des seuils cliniques fixes. L'aspect le plus important est l'interprétabilité des règles générées. Ces règles doivent être intelligibles, fiables et se rapprocher de la logique clinique humaine. La robustesse et la fiabilité des analyses sont également jugées par la méthode de prétraitement des données, comme l'identification des

	interprétations claires. Nous visons à évaluer la capacité du système à essayer de reproduire le raisonnement face à l'incertitude.				valeurs aberrantes et la gestion des données manquantes.

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Résultats	Nous présentons une étude détaillée des performances de prédiction et de la capacité d'interprétation d'un système optimisé par le *PFL*-Arborescent. Nous exposons les résultats de l'inférence causale, soulignant l'apport de la *PFL*. Nous cherchons à quantifier le lien entre l'administration d'antibiotiques et la mortalité chez les patients souffrant de choc septique.	Nous avons évalué notre système hybride sur trois sous-ensembles de données. La bonne calibration post-entraînement est une découverte, garantissant la fiabilité des probabilités. La capacité du modèle à générer des règles claires et interprétables est un point fort. L'identification des incohérences entre les prédictions et les règles floues facilite l'audit. Nous avons hiérarchisé les informations en trois niveaux (global, par sous-groupe, et par patient) pour nous adapter à différents utilisateurs. L'analyse causale a montré que la simple corrélation ne suffit pas. L'intégration de la logique floue a révélé l'impact variable du traitement selon les niveaux des variables. Les scores *NFATE* ont aidé à identifier les facteurs de confusion clé, tels que la température.	Les données sont les enregistrements de patients du jeu de données *MIMIC-III* comprenant 12 780 observations épurées. Les entrées sont des mesures physiologiques et des variables cliniques. La variable de traitement 'A' indique l'administration d'antibiotiques (unique ou multiples).	Les sorties principales sont les suivantes. Le modèle présente de bonnes performances de discrimination (*ROC-AUC* élevé) et une bonne calibration (*ECE* calibré à 0). Le rappel pour le groupe sous antibiotique unique est très élevé. Le système produit des règles "*IF-THEN*" dénormalisées et facilement interprétables. Les analyses par cluster révèlent des zones d'incohérence, nécessitant une investigation clinique. En ce qui concerne la causalité, nous avons quantifié l'*ATE* manuel et les *ATE-PFL* ajustés. Les scores *NFATE* ont permis de classer les facteurs de confusion par influence. Les mesures de vagueness et d'incertitude quantifient	La fiabilité des probabilités est un critère central, validé par l'*ECE* calibré à 0. L'interprétabilité des règles générées par la *PFL*, avec des adjectifs linguistiques et des plages de valeurs dénormalisées, est une mesure de qualité. La capacité du système à détecter les incohérences et à déclencher des alertes est un critère de confiance. L'identification des facteurs de confusion clés par les *NFATE* et la quantification de l'incertitude par les scores de vagueness et les écarts-types pondérés sont des critères qui garantissent la robustesse de l'analyse causale.

				l'ambiguïté des segments flous.	

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Discussion	Nous interprétons les résultats du *PFL*-Arborescent en les confrontant aux connaissances actuelles. Nous examinons les conséquences de ces découvertes, défendons notre approche, et mettons en évidence nos contributions. Notre objectif est de positionner cette méthode dans le contexte de l'IA valorisante, en montrant comment la *PFL* enrichit l'*IAI* l'inférence causale.	Notre approche combine la force prédictive de modèles comme *LightGBM* avec la clarté de la logique floue. La calibration isotonique assure une crédibilité cardinale aux probabilités prédites. La capacité à générer des règles "*IF-THEN*" dénormalisées et descriptives tente de représenter le raisonnement clinique et offre un entendement des prédictions. Le système d'alerte de cohérence sert d'autocertification, signalant les cas complexes nécessitant une expertise humaine. L'analyse causale floue a révélé que l'effet du traitement varie selon les niveaux des variables. Par exemple, l'*ATE-PFL* pour l'excès de base artériel varie de manière significative entre les 3 segments. L'identification des facteurs de confusion clé par les *NFATE*, comme la température, a une pertinence clinique. La quantification de l'incertitude par les scores de vagueness et d'écart-type pondéré rend l'IA plus transparente.	Les entrées pour cette discussion sont les résultats du modèle *PFL*-Arborescent. Cela inclut les métriques de performance (*ECE, ROC-AUC*), les règles floues générées, les alertes de cohérence, et les résultats de l'analyse causale (*ATE, ATE-PFL, NFATE*, scores de vagueness et d'incertitude). Ces informations sont tirées de l'entraînement sur le jeu de données *MIMIC-III*.	Les principales sorties sont les interprétations des résultats. Nous confirmons la crédibilité des probabilités, la pertinence des règles "*IF-THEN*", et l'utilité du système d'alerte. Nous démontrons que l'approche causale floue surpasse les analyses causales classiques en capturant des effets hétérogènes. Nous prouvons que notre modèle peut s'aligner sur les principes de l'IA éthique, responsable et transparente. Nous validons également nos contributions spécifiques, notamment le modèle *PFL-LightGBM* calibré, les règles floues dénormalisées et un simulateur de causalité.	La validité de notre thèse repose sur plusieurs points : la crédibilité des probabilités ajustées, l'interprétabilité des règles floues, l'aptitude à détecter des incohérences, et la capacité à modéliser la causalité avec plus de finesse que les méthodes conventionnelles. La cohérence entre la moyenne des *ATE-PFL* ajustés et l'*ATE* manuel global renforce la consistance interne du modèle. L'identification des facteurs de confusion pertinents et la quantification de l'incertitude sont des critères qui justifient la robustesse de l'approche. Ces éléments soutiennent notre vision d'une IA intelligible,

					interprétable, fiable et éthique.

Chapitre	Objectifs	Principales idées	Entrées	Sorties	Critères de qualité
Conclusion générale : Synthèse, portée et perspectives.	Notre approche propose une modélisation robuste de l'incertitude. Elle fournit des probabilités ajustées. Les interprétations sont rendues accessibles a des experts par des règles. Un système d'alerte identifie les zones ambiguës. Le cadre se concentre sur une vision causale. L'ensemble de ces principes permet de construire un système d'IA éthique, fiable et axé sur l'enrichissement de l'interprétabilité par la causalité.	Notre approche propose une modélisation robuste de l'incertitude. Elle fournit des probabilités ajustées. Les interprétations sont rendues accessibles aux humains par des règles. Le cadre se concentre sur une vision causale. L'ensemble de ces principes permet de construire un système d'IA éthique, fiable et axé sur la causalité.	Le système utilise des données statistiques réelles. Ces données proviennent du contexte clinique et peuvent inclure des valeurs indéterminées. L'approche requiert la connaissance du domaine pour la construction des règles.	Les sorties incluent des probabilités de risque crédibles. Le modèle produit des règles floues, étiquetées, directement intelligibles par un clinicien. Un outil signale les discordances. L'analyse causale mesure l'effet des traitements et identifie les facteurs d'influence.	Nous évaluons la qualité par la transparence et la fiabilité des résultats. Le modèle présente un taux d'erreur de calibration presque nul. Il est aussi conçu pour être intelligible, adaptable et responsable. Ces critères sont nécessaires afin de garantir la confiance et la sécurité des utilisateurs dans des domaines sensibles, en respectant les normes réglementaires et éthiques.

ANNEXE B

[Mesures de performance]

Nous décrivons dans le tableau suivant les mesures utilisées pour la performance prédictive des modèles.

Score	Définition	Formule
VP (Vrai Positif)	Nombre de cas où le modèle a correctement prédit une observation positive.	–
VN (Vrai Négatif)	Nombre de cas où le modèle a correctement prédit une observation négative.	–
FP (Faux Positif)	Nombre de cas où le modèle a incorrectement prédit une observation positive.	–
FN (Faux Négatif)	Nombre de cas où le modèle a incorrectement prédit une observation négative.	–
Précision	Proportion de prédictions positives correctes.	Précision = VP / (VP + FP)
Rappel	Proportion d'observations positives correctement identifiées.	Rappel = VP / (VP + FN)
ROC *(Receiver Operating Characteristic :* fonction d'efficacité du récepteur*)*	Courbe représentant la relation entre le taux de faux positifs (FP) et le taux de vrais positifs (VP) pour différents seuils de classification.	–
AUC *(Area Under the ROC :* Aire sous la courbe *)*	Représente la proportion de la surface totale sous la courbe *ROC* d'un modèle. Plus précisément, l'*AUC* mesure la capacité d'un modèle à distinguer entre les classes positives et négatives.	Se calcule en intégrant la courbe *ROC*. Plus la surface totale située en dessous de la courbe est grande, plus le modèle est performant.
F1 Score	Indicateur de la performance d'un modèle de classification binaire mesurant le compromis entre la précision et le rappel.	F1 = 2 × (Précision × Rappel) / (Précision + Rappel)

Balanced Accuracy	Moyenne pondérée des taux de vrais positifs et de vrais négatifs, où les poids sont égaux aux proportions d'échantillons positifs et négatifs dans les données d'entraînement.	Balanced Accuracy = (TPR + TNR) / 2 * TPR (taux de vrais positifs) est le nombre de positifs correctement classés divisés par le nombre total de vrais positifs. * TNR (taux de vrais négatifs) est le nombre de négatifs correctement classé divisé par le nombre total de vrais négatifs.

Description des scores utilisés pour mesurer les performances des modèles de prédiction (Ed-Driouch, 2023).

Deuxièmement, la régression isotonique est une méthode efficace qui sert à représenter le lien entre deux variables tout en respectant une contrainte de relation monotone (c'est-à-dire soit non décroissante, soit non croissante) spécifique, aussi nommée isotonicité :

- Cette restriction assure que la liaison entre les variables sera monotone.
- La régression isotonique vise à déterminer une fonction par morceaux (monotone) fournissant les ajustements les plus précis aux points de données générés tout en respectant la contrainte de monotonie.

De plus, en ce qui concerne l'erreur de calibration attendue (*Expected Calibration Error : ECE*). C'est une évaluation qui calcule la différence moyenne entre les probabilités prévues et les résultats réels pour toutes les occurrences :

- ❖ Nous la déterminons en calculant la moyenne pondérée des écarts absolus entre les probabilités prévues et les fréquences observées, toutes classes de probabilité regroupées.
- ❖ L'*ECE* fournit une analyse complète de l'étalonnage, ce qui est particulièrement bénéfique pour la mise en comparaison de différents modèles.
- ❖ Un *ECE* bas suggère une calibration de meilleure qualité, indiquant que les probabilités estimées se rapprochent davantage des taux d'événements authentiques[39].

Aussi, pour le coefficient Kappa de Cohen, la spécificité, le coefficient de corrélation de Matthew, l'aire sous la courbe de rappel, Akhouayri (2023) les définit comme suit :

Coefficient Kappa de Cohen : Il s'agit d'un indicateur statistique qui varie entre 0 et 1, principalement employé pour évaluer la concordance entre deux modèles prédictifs quant à la manière de classifier des observations. Nous pouvons l'interpréter comme le rapport de concordances : la proportion d'éléments identifiés de manière similaire par les deux modèles. Cet indice reflète une forte concordance, d'autant plus remarquable qu'il se situe près de 1, ce qui signifie également l'absence d'erreur. Plus ce kappa est faible, plus il y a une divergence entre la réalité et les prédictions.

Spécificité : Elle examine la capacité d'un modèle prédictif à identifier, dans une population ciblée, les individus qui ne possèdent pas une caractéristique spécifique donnée (« non-cas »). La spécificité d'un

[39] https://cas.uqam.ca/pub/web/vignettes/calibration_metrics_vignette.html

modèle représente donc la probabilité qu'il réussisse à détecter correctement une absence de cas, ou la probabilité qu'un cas ne soit pas repéré par le modèle. Par conséquent, la spécificité se rapporte à sa capacité à faire une distinction entre les éléments. La spécificité est déterminée par le taux d'individus que le modèle a identifiés comme n'ayant pas une caractéristique spécifique (résultat négatif) parmi ceux qui ne possèdent effectivement pas cette caractéristique (« non-cas »). Nous nous basons donc sur de véritables négatifs. Un modèle de prédiction présentant un problème de spécificité identifie des sujets qui semblent avoir une caractéristique particulière, alors qu'en vérité, ils ne l'ont pas. Nous appelons ces erreurs discriminatoires des faux positifs.

Coefficient de corrélation de Matthew (*Matthews' Correlation Coefficient : MCC*) : Le *MCC* est une mesure statistique utilisée pour juger la performance des modèles. Cet outil sert à évaluer ou mesurer la différence entre les valeurs prévues et les valeurs réelles et se réfère à la statistique du chi carré pour un tableau de contingence 2 x 2 (Brian Matthews, 1975).

Pour le score Brier, cet indicateur évalue la précision des prédictions statistiques en mesurant la différence quadratique moyenne entre les probabilités prévues et les événements réels ; un score Brier faible indique une meilleure performance du modèle[40].

Finalement, pour la perte logarithmique (*Log-loss*), Hadiat (2022) note que :

- Les modèles de régression logistique sont évalués en utilisant les fonctions *Log-loss*, aussi appelées fonction de perte par entropie croisée.
- La *Log-loss* mesure l'écart entre les probabilités estimées et les valeurs réelles.
- Un score *Log-loss* inférieur témoigne de meilleures performances pour le classificateur.
- Nous pouvons évaluer l'importance d'une caractéristique précise en recourant à la *Log-loss*.
- Cette indication reflète la contribution précise d'une caractéristique particulière au modèle.
- Nous utilisons cette méthode pour juger de la pertinence d'une caractéristique particulière dans un classificateur de régression logistique.

[40] https://cas.uqam.ca/pub/web/vignettes/calibration_metrics_vignette.html

ANNEXE C

[Tableau des groupes de risque, des résultats quantitatifs clés, et des règles « *IF-THEN* » et interprétation.]

Selon le groupe de risque, nous récapitulons les résultats quantitatifs clés, les règles ainsi que leur interprétation dans le tableau suivant avec un suivi d'explication.

Groupe de risque	Résultats quantitatifs clés	Règle « *IF-THEN* » et interprétation
Faible	Risque modèle : 6.2 % Probabilité floue : 4.83 % Fidélité : 0.9861 Proportion : 31.8 %	Règle : Si un patient montre des valeurs de variables dans des intervalles spécifiques, alors son risque de décès est minime. Interprétation : Nous catégorisons ce groupe comme ayant une probabilité de survie. Il y a une cohérence entre le modèle et la règle.
Moyen	Risque modèle : 10.8 % Probabilité floue : 12.30 % Fidélité : 0.9851 Proportion : 9.3 %	Règle : Si un patient affiche une série de traits particuliers, alors le risque de décès qui lui est associé s'élève à 12.30 %. Interprétation : Nous recommandons une surveillance accrue pour ces patients. L'interprétation de cette dernière est justifiée par l'insignifiante différence entre la prédiction du modèle et la probabilité de la règle.
Élevé	Groupe A : Risque modèle : 20.4 % Probabilité floue : 12.26 %	Règle A : Si un patient satisfait à un certain nombre de critères, son niveau de risque est élevé.

	Proportion : 15.7 % Groupe B : Risque modèle : 43.3 % Probabilité floue : 42.13 % Proportion : 0.7 %	Règle B : Si un patient répond à un autre groupe de critères, son risque se trouve être extrêmement élevé. Interprétation : Nous préconisons une action immédiate pour les patients appartenant à ces catégories. Les règles reflètent la sévérité du danger.

Tableau des groupes de risque, les résultats quantitatifs clés et les règles « *IF-THEN* » et interprétation.

Interprétation des règles

Nous employons le principe « *IF-THEN* » pour l'élaboration de règles cliniques. Ces règles nous permettent d'expliquer comment le modèle arrive à ses conclusions. Chaque règle est associée à un groupe de patients. Les variables et les intervalles de valeurs de ces variables justifient l'appartenance d'un patient à un groupe.

Comment construisons-nous les règles

Nous déterminons les caractéristiques moyennes des patients appartenant à chaque groupe. Ces caractéristiques constituent les conditions de la règle « *IF-THEN* ». Par exemple, nous prenons en compte l'association de plusieurs variables, telles que le taux de potassium et la pression artérielle. Nous créons un intervalle de valeurs pour chaque variable. Nous disons que si un patient a ces valeurs, alors il fait partie du groupe. Cette approche fournit une interprétation transparente. Elle nous permet de comprendre pourquoi un patient est classé à risque faible, moyen ou élevé.

Justifications

Chaque règle a une signification clinique.

- Règle du risque faible : Nous justifions cette règle par des mesures quantitatives qui indiquent une probabilité de mortalité minimale. L'appartenance à ce groupe suggère une survie probable. Nous ne recommandons pas d'intervention spéciale pour ces patients, car leur risque est jugé très bas. La fidélité élevée du groupe confirme cette classification.
- Règle du risque moyen : Le modèle attribue à ce groupe un risque notable, mais pas critique. Nous justifions cette classification par les probabilités associées. La règle nous permet de comprendre

les critères cliniques qui mènent à ce risque. Nous suggérons une surveillance accrue. L'identification de ce groupe permet de prioriser les soins sans saturer les services (d'urgence).

- Règles du risque élevé : Le document distingue deux groupes à haut risque. Le premier groupe a une probabilité de mortalité élevée. Le second groupe a une probabilité de mortalité très élevée. Ces règles se justifient par les valeurs de risque. Nous suggérons une intervention urgente. Comprendre les critères de ces règles nous permet d'agir rapidement pour les patients les plus critiques.

Pour résumer, les règles « *IF-THEN* » nous fournissent une structure pour comprendre les prédictions du modèle. Elles convertissent des algorithmes sophistiqués et hybrides en collaborateurs pour les professionnels de la santé.

BIBLIOGRAPHIE

1. Acharki, N. (2022). Statistical learning and causal inference for energy production (Doctoral dissertation, Institut polytechnique de Paris).
2. Adamyk, O., Chereshnyuk, O., Adamyk, B., & Rylieiev, S. (2023, September). Trustworthy AI: a fuzzy-multiple method for evaluating ethical principles in AI regulations. In *2023 13th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 608-613).
3. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).
4. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).
5. Ahmad, M. A., Eckert, C., & Teredesai, A. (2018, August). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559-560).
6. Akhouayri, L. (2023). Genomic, bioinformatic and statistical approaches to refine invasive breast carcinoma classification (Doctoral dissertation).
7. Ambag, E. L., Capitoli, G., Imperio, V. L., Provenzano, M., Nobile, M. S., & Liò, P. (2023). Assisting clinical practice with fuzzy probabilistic decision trees. arXiv preprint arXiv:2304.07788.
8. Andrada, G., Clowes, R. W., & Smart, P. R. (2023). Varieties of transparency: Exploring agency within AI systems. *AI & society*, *38*(4), 1321-1331.
9. Arbib, M. A. (Ed.). (2003). The handbook of brain theory and neural networks. MIT press.
10. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.
11. AYAD, M. (2023). Commande en temps réel d'un pendule inversé rotatif par la logique floue (Doctoral dissertation, Directeur: Mr MERAD Lotfi/Co-Directeur: Mr ARICHI Fayssal).
12. Azencott, C.-A. (2018). Introduction au machine learning. Dunod.
13. Bahani, K., Moujabbir, M., & Ramdani, M. (2021). An accurate fuzzy rule-based classification systems for heart disease diagnosis. *Scientific African*, *14*, e01019.
14. Bai, H. (2023, April). Network Equipment Fault Maintenance Decision System Based on Bayesian Decision Algorithm. In 2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT) (pp. 1197-1202). IEEE.
15. Balentien, R. M., Kaymak, U., Walther van Mook, M. D., & Bergmans, D. (2016). Personalized mortality risk predictions for ICU Septic Shock patients: Making use of Probabilistic Fuzzy Systems Thesis.
16. Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 101286.
17. Barwise, A., & Pickering, B. (2022). The AI needed for ethical decision making does not exist. *The American Journal of Bioethics*, *22*(7), 46-49.
18. Başağaoğlu, H., Chakraborty, D., Lago, C. D., Gutierrez, L., Şahinli, M. A., Giacomoni, M., ... & Şengör, S. S. (2022). A review on interpretable and explainable artificial intelligence in hydroclimatic applications. Water, 14(8), 1230.
19. Basili, V. R. (2013). A Personal Perspective on the Evolution of Empirical Software Engineering. Dans: Münch J, Schmid K (éd.). Perspectives on the Future of Software Engineering, Springer, Berlin/Heidelberg. p. 255-273.

20. Basili, V. R., Caldiera G., Rombach H. D. (1994). Goal question metric paradigm. Encyclopedia of Software Engineering.
21. Basili, V. R., Selby R. W., Hutchens D. H. (1986). Experimentation in software engineering. IEEE Transactions on Software Engineering, vol. SE-12, n° 7, p. 733-743.
22. Béghin, G. (2022). Causalité et modèle des deux stratégies de raisonnement: étude des différences individuelles dans l'induction causale à partir d'informations de covariation.
23. Benzinger, L., Ursin, F., Balke, W. T., Kacprowski, T., & Salloch, S. (2023). Should Artificial Intelligence be used to support clinical ethical decision-making? A systematic review of reasons. *BMC medical ethics*, *24*(1), 48.
24. Bezdek, J. C., & Bezdek, J. C. (1981). Objective function clustering. Pattern recognition with fuzzy objective function algorithms, 43-93.
25. Biedermann, A., & Taroni, F. (2022). Bayesian Networks and Influence Diagrams. Biedermann A., Taroni F., Bayesian networks and influence diagrams, Encyclopedia of Forensic Sciences (Third Edition), Max M. Houck (Ed.), Oxford: Elsevier, 271-280.
26. Biller-Andorno, N., & Biller, A. (2019). Algorithm-aided prediction of patient preferences-an ethics sneak peek. *New England Journal of Medicine*, *381*(15), 1480-1485.
27. Bishop, J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. Frontiers in Psychology, 11, 513474.
28. Boehm, B. (1988). A Spiral Model of Software Development and Enhancement. IEEE Computer, 21 (5).
29. Bottemanne, H., Longuet, Y., & Gauld, C. (2022). L'esprit prédictif: introduction à la théorie du cerveau bayésien. L'Encéphale, 48(4), 436-444.
30. Bourguin, G., Lewandowski, A., Bouneffa, M., & Ahmad, A. (2021, June). Vers des classifieurs ontologiquement explicables. In *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence artificielle (PFIA'21)* (pp. pp-89).
31. Broadbent, A., & Grote, T. (2022). Can robots do epidemiology? Machine learning, causal inference, and predicting the outcomes of public health interventions. Philosophy & Technology, 35(1), 14.
32. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.
33. Camilleri, M. A. (2024). Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert systems*, *41*(7), e13406.
34. Cardiel-Ortega, J. J., & Baeza-Serrato, R. (2024). Probabilistic Fuzzy System for Evaluation and Classification in Failure Mode and Effect Analysis. *Processes*, *12*(6), 1197.
35. Carloni, G., Berti, A., & Colantonio, S. (2025). The role of causality in explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *15*(2), e70015.
36. Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines. *Berkeley Technology Law Journal*, *34*(1), 143-188.
37. Chakraborty, S., Paul, D., Das, S., & Xu, J. (2020, June). Entropy weighted power k-means clustering. In International conference on artificial intelligence and statistics (pp. 691-701). PMLR.
38. Chaudhury, S., Sau, K., Khan, M. A., & Shabaz, M. (2023). Deep transfer learning for IDC breast cancer detection using fast AI technique and Sqeezenet architecture. Math Biosci Eng, 20, 10404-27.
39. Chen, Q., Li, R., Lin, C., Lai, C., Chen, D., Qu, H., ... & Li, L. (2022). Transferability and interpretability of the sepsis prediction models in the intensive care unit. BMC Medical Informatics and Decision Making, 22(1), 1-10.
40. Chen, S. J., & Chen, S. M. (2001, December). A new method to measure the similarity between fuzzy numbers. In 10th IEEE International Conference on Fuzzy Systems. (Cat. No. 01CH37297) (Vol. 3, pp. 1123-1126). IEEE.

41. Clark, A. (2017). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. Noûs, 51(4), 727-753.
42. Claverie, B. (2021). What is cognition and how to make it one of the means of war?
43. Coggins, S. A., & Glaser, K. (2022). Updates in Late-Onset Sepsis: Risk Assessment, Therapy, and Outcomes. Neoreviews, 23(11), 738-755.
44. Coulombe, C. (2020). Techniques d'amplification des données textuelles pour l'apprentissage profond (Doctoral dissertation, Télé-université).
45. Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). Designing AI for social good: Seven essential factors. *Available at SSRN 3388669*.
46. de Vries, B. M., Zwezerijnen, G. J., Burchell, G. L., van Velden, F. H., & Boellaard, R. (2023). Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Frontiers in medicine, 10, 1180773.
47. Denis, C., & Varenne, F. (2019, July). Interprétabilité et explicabilité pour l'apprentissage machine: entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique. In National (French) Conference on Artificial Intelligence (CNIA)-Artificial Intelligence Platform (PFIA) (pp. 60-68).
48. Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.
49. Druzdzel, M. J., & Simon, H. A. (1993, January). Causality in Bayesian belief networks. In *Uncertainty in Artificial Intelligence* (pp. 3-11). Morgan Kaufmann.
50. Ed-Driouch, C. (2023). Dialogue humain machine pour l'aide à la décision médicale (Doctoral dissertation, Nantes Université).
51. Edwards, L., & Veale, M. (2018). Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"?. *IEEE Security & Privacy*, *16*(3), 46-54.
52. Efosa, I. C., & Akwukwuma, V. (2013). Knowledge-based fuzzy inference system for sepsis diagnosis. *International Journal of Computational Science and Information Technology*, *1*(3), 1-7.
53. Faghihi, U., Bouchard, S. M., & Biskri, I. (2021). Science of Data: A New Ladder for Causation. Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications, 33-45.
54. Faghihi, U., Kalantarpour, C., Baldé, I., & Saki, A. (2023). Causal Fuzzy Deep Learning Algorithm for the Differentiation of Gliosarcoma From Glioblastoma.
55. Faghihi, U., Robert, S., Poirier, P., & Barkaoui, Y. (2020, May). From association to reasoning, an alternative to pearls' causal reasoning. In The Thirty-Third International Flairs Conference.
56. Falip, J. (2019). Structuration de données multidimensionnelles: une approche basée instance pour l'exploration de données médicales (Doctoral dissertation, Université de Reims Champagne-Ardenne).
57. Ferrario, A., Gloeckler, S., & Biller-Andorno, N. (2023). Ethics of the algorithmic prediction of goal of care preferences: from theory to practice. *Journal of medical ethics*, *49*(3), 165-174.
58. Fialho, A. S., Vieira, S. M., Kaymak, U., Almeida, R. J., Cismondi, F., Reti, S. R., ... & Sousa, J. M. (2016). Mortality prediction of septic shock patients using probabilistic fuzzy systems. *Applied Soft Computing*, *42*, 194-203.
59. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1).
60. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.
61. Främling, K. (2023, May). Counterfactual, Contrastive, and Hierarchical Explanations with Contextual Importance and Utility. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 180-184). Cham: Springer Nature Switzerland.

62. Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? PloS one, 4(7), e6421.
63. Fuchs, C., Spolaor, S., Nobile, M. S., & Kaymak, U. (2020, July). pyFUME: a Python package for fuzzy model estimation. In *2020 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1-8). IEEE.
64. Gacto, M. J., Alcalá, R., & Herrera, F. (2011). Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, *181*(20), 4340-4360.
65. Gaudet, S., & Robert, D. (2018). L'aventure de la recherche qualitative: Du questionnement à la rédaction scientifique. University of Ottawa Press.
66. Ghidalia, S. (2024). Etude sur les mesures d'évaluation de la cohérence entre connaissance et compréhension dans le domaine de l'intelligence artificielle (Doctoral dissertation, Université Bourgogne Franche-Comté).
67. González del Solar, R., Marone, L., & Lopez de Casenave, J. (2019). Mechanismic approaches to explanation in ecology. *Mario Bunge: A Centenary Festschrift*, 555-573.
68. Guan, S., Li, J., & Luo, X. (2024). Variational Bayesian Gaussian mixture model for off-grid DOA estimation. *Electronics Letters*, *60*(3), e13114.
69. Gudder, S. (1998). Fuzzy probability theory. Demonstratio Mathematica, 31(1), 235-254.
70. Gundersen, T., & Bærøe, K. (2022). Ethical algorithmic advice: Some reasons to pause and think twice. *The American Journal of Bioethics*, *22*(7), 26-28.
71. Gunning, D. (2019). DARPA's explainable artificial intelligence (XAI) program. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19) (pp. ii–ii). Association for Computing Machinery. https://doi.org/10.1145/3301275.3308446.
72. Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58.
73. Guo, S., Guo, Z., Ren, Q., Wang, X., Wang, Z., Chai, Y., ... & Union, R. P. (2023). A PREDICTION MODEL FOR SEPSIS IN INFECTED PATIENTS: EARLY ASSESSMENT OF SEPSIS ENGAGEMENT. *Shock*, *60*(2), 214-220.
74. Hadiat, A. R. (2022). Topic modeling evaluations: The relationship between coherency and accuracy (Doctoral dissertation).
75. Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science*.
76. Hassanlou, M. (2025, February 25). *Follow 5 best practices to enhance the explainability of AI models (ID G00823629)*. Gartner. Retrieved from [Follow 5 Best Practices to Enhance the Explainability of AI Models].
77. Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of science*, *15*(2), 135-175.
78. Hernán, M. A., & Robins, J. M. (2010). Causal inference.
79. High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission, Brussels, Belgium. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
80. Houtsma, M., & Swami, A. (1993). Set-oriented mining for association rules. IBM Almaden research center. research report RJ 9567, San Jose.
81. Hu, S., Teng, F., Huang, L., Yan, J., & Zhang, H. (2021). An explainable CNN approach for medical codes prediction from clinical text. BMC Medical Informatics and Decision Making, 21, 1-12.
82. Huang, Y., Chen, D., Zhao, W., & Lv, Y. (2022). Fuzzy c-means clustering based deep patch learning with improved interpretability for classification problems. IEEE Access, 10, 49873-49891.
83. Hussain, S. M., Buongiorno, D., Altini, N., Berloco, F., Prencipe, B., Moschetta, M., ... & Brunetti, A. (2022). Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images: The Role of Explainable Artificial Intelligence. Applied Sciences, 12(12), 6230.

84. Ishibuchi, H., & Nakashima, T. (2001). Effect of rule weights in fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems, 9(4), 506-515.
85. Islam, K. R., Prithula, J., Kumar, J., Tan, T. L., Reaz, M. B. I., Sumon, M. S. I., & Chowdhury, M. E. (2023). Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review. Journal of clinical medicine, 12(17), 5658.
86. Jamshidnejad, A., & De Schutter, B. (2024). A combined probabilistic-fuzzy approach for dynamic modeling of traffic in smart cities.
87. Jaynes, E.T. (2003). Probability theory: the logic of science, Cambridge, Cambridge University Press.
88. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.
89. John-Mathews, J. M. (2019). Interprétabilité en Machine Learning, revue de littérature et perspectives. *Séminaire Good In Tech." Développement de technologies responsables*".
90. Josephson, J. R., & Josephson, S. G. (Eds.). (1996). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.
91. Kaminski, M. E. (2019). The right to explanation, explained. Berkeley Technology Law Journal, 34(1), 189-218.
92. Karam, H. (2010). L'abduction en conception architecturale: Une sémiose hypostatique.
93. Kelley, H. H. (1987). Causal schemata and the attribution process. In *Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969.* Lawrence Erlbaum Associates, Inc.
94. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.
95. Li, S., Dou, R., Song, X., Lui, K. Y., Xu, J., Guo, Z., ... & Cai, C. (2023). Developing an Interpretable Machine Learning Model to Predict in-Hospital Mortality in Sepsis Patients: A Retrospective Temporal Validation Study. Journal of Clinical Medicine, 12(3), 915.
96. Li, X., Zhao, L., & Chen, J. (2024). Remote Sensing Image Segmentation Based on Probabilistic Fuzzy Local Information Clustering. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *10*, 211-216.
97. Lin, S. H., & Ikram, M. A. (2020). On the relationship of machine learning with causal inference. European Journal of Epidemiology, 35, 183-185.
98. Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.
99. Liu, R., Hunold, K. M., Caterino, J. M., & Zhang, P. (2023). Estimating treatment effects for time-to-treatment antibiotic stewardship in sepsis. *Nature machine intelligence*, *5*(4), 421-431.
100. Liz, H., Sánchez-Montañés, M., Tagarro, A., Domínguez-Rodríguez, S., Dagan, R., & Camacho, D. (2021). Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis. *Future Generation Computer Systems*, *122*, 220-233.
101. Llored, J. P., & Bouchnita, A. (2022). Comment instaurer la médecine de précision? Pour une alliance entre intelligence artificielle et modélisation conceptuelle. *Ethique et Numérique*, *1*(1), 154-180.
102. Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 107161.
103. Madrid, N. (2023). Significance measures for rules in probabilistic-fuzzy inference systems based on fuzzy transforms. Fuzzy Sets and Systems, 467, 108575.
104. Mageau, A., Timsit, J. F., Perrozziello, A., Papo, T., & Sacré, K. (2018). Choc septique au cours du lupus érythémateux systémique: pronostic à court et à long terme. Résultats de l'analyse d'une base médico-administrative nationale. La Revue de Médecine Interne, 39, A58-A59.

105. Martens, C. R., Wahl, D., & LaRocca, T. J. (2023). Personalized medicine: will it work for decreasing age-related morbidities?. In *Aging* (pp. 683-700). Academic Press.
106. Maryani, N., & Wisudarti, C. F. R. (2023). The comparison score of SPEEDS, MEDS, SOFA, APACHE II, and SAPS II as predictor of sepsis mortality: A Systematic Review. Asian Journal of Social and Humanities, 2(1), 1493-1503.
107. Masís, S. (2023). *Interpretable Machine Learning with Python: Build explainable, fair, and robust high-performance models with hands-on, real-world examples*. Packt Publishing Ltd.
108. Maziarz, M. (2020). The philosophy of causality in economics: Causal inferences and policy proposals. Routledge.
109. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
110. Meghdadi, A. H., & Akbarzadeh-T, M. R. (2001, December). Probabilistic fuzzy logic and probabilistic fuzzy systems. In *10th IEEE international conference on fuzzy systems.(Cat. No. 01CH37297)* (Vol. 3, pp. 1127-1130). IEEE.
111. Mehnatkesh, H., Jalali, S. M. J., Khosravi, A., & Nahavandi, S. (2023). An intelligent driven deep residual learning framework for brain tumor classification using MRI images. Expert Systems with Applications, 213, 119087.
112. Mehta, P. H. (2020). A Principled Approach to Multiple Causal Inference (Doctoral dissertation).
113. Meller, G. (2023/2024). Explainable Artificial Intelligence: A Study of Methods, Applications, and Future Directions.
114. Meng, X. B., Li, H. X., & Chen, C. P. (2022). A two-stage Bayesian learning-based probabilistic fuzzy interpreter for uncertainty modeling. Applied Soft Computing, 131, 109786.
115. Meraihi, N. (2019). Modélisation de la fréquence des sinistres avec données télématiques par le modèle GBMP.
116. Michel-Delétie, C., & Sarker, M. K. (2024). Neuro-Symbolic methods for Trustworthy AI: a systematic review. *Neurosymbolic Artificial Intelligence*.
117. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.
118. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.
119. Mohammed, E. E., Fayez, A. G., Abdelfattah, N. M., & Fateen, E. (2024). Novel gene-specific Bayesian Gaussian mixture model to predict the missense variants pathogenicity of Sanfilippo syndrome. *Scientific Reports*, *14*(1), 12148.
120. Mökander, J., & Floridi, L. (2023). Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, *3*(2), 451-468.
121. Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3rd ed.). christophm.github.io/interpretable-ml-book/.
122. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193-209).
123. Morales Boada, C. (2016). *Investigating prognostic factors in Sepsis using Computational Intelligence methods* (Master's thesis, Universitat Politècnica de Catalunya).
124. Morandé, S., & Tewari, V. (2023). Causality in Machine Learning: Innovating Model Generalization through Inference of Causal Relationships from Observational Data. Qeios.
125. Musanga, V., Viriri, S., & Chibaya, C. (2025). A Framework for Integrating Deep Learning and Symbolic AI Towards an Explainable Hybrid Model for the Detection of COVID-19 Using Computerized Tomography Scans. *Information*, *16*(3), 208.
126. Mustafayev, S. (2017). Mortality Risk Prediction for ICU Septic Shock Patients Using Probabilistic Fuzzy Systems: Comparison of Sepsis-2 and Sepsis-3.

127. Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical care medicine*, *46*(4), 547-553.
128. Nguyen, T. L., Kavuri, S., Park, S. Y., & Lee, M. (2022). Attentive Hierarchical ANFIS with interpretability for cancer diagnostic. Expert Systems with Applications, 201, 117099.
129. Nguyen, T. M., Poh, K. L., Chong, S. L., & Lee, J. H. (2023). Effective diagnosis of sepsis in critically ill children using probabilistic graphical model. *Translational Pediatrics*, *12*(4), 538.
130. Nielsen, I. E., Dera, D., Rasool, G., Ramachandran, R. P., & Bouaynaya, N. C. (2022). Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, *39*(4), 73-84.
131. Nilsson, N. J. (1986). Probabilistic logic. Artificial intelligence, 28(1), 71-87.
132. Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.
133. Okafor, C. E., Iweriolor, S., Ani, O. I., Ahmad, S., Mehfuz, S., Ekwueme, G. O., ... & Chikelu, O. P. (2023). Advances in machine learning-aided design of reinforced polymer composite and hybrid material systems. *Hybrid Advances*, *2*, 100026.
134. Ouifak, H. et Idri, A. (2023). Sur les performances et l'interprétabilité des systèmes neuro-flou basés sur Mamdani et Takagi-Sugeno-Kang pour le diagnostic médical. Africain scientifique, 20, e01610.
135. Pallanca, O., & Read, J. (2021). Principes généraux et définitions en intelligence artificielle. Archives des Maladies du Coeur et des Vaisseaux-Pratique, 2021(294), 3-10.
136. Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., ... & Gomez, E. (2023, June). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1139-1150).
137. Parente, J. D., Möller, K., Shaw, G. M., & Chase, J. G. (2018). Hidden Markov models for sepsis classification. *IFAC-PapersOnLine*, *51*(27), 110-115.
138. Pearl, J. (2001). Bayesianism and causality, or, why I am only a half-Bayesian. In *Foundations of bayesianism* (pp. 19-36). Dordrecht: Springer Netherlands.
139. Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016.
140. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54-60.
141. Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.
142. Peterson, C., & Hamrouni, N. (2022, May). Preliminary thoughts on defining f (x) for ethical machines. In *The International FLAIRS Conference Proceedings* (Vol. 35).
143. Pévet, P., Sauvayre, R., & Tiberghien, G. (2011). Les sciences cognitives. Dépasser les frontières disciplinaires (p. 138). Presses Universitaires de Grenoble.
144. Pevneva, A., & Pivovarova, I. (2022, December). Comparison of the probabilistic model of a fuzzy Bayesian classifier and the Gaussian mixture problem. In Journal of Physics: Conference Series (Vol. 2373, No. 6, p. 062022). IOP Publishing. ou c-means flou (FCM) : Analyser et choisir une des optimisations.
145. Poché, A., Hervier, L., & Bakkay, M. C. (2023, July). Natural example-based explainability: a survey. In *World Conference on eXplainable Artificial Intelligence* (pp. 24-47). Cham: Springer Nature Switzerland.
146. Porouhan, P. (2023). Treatment process conformance checking of patients (with sepsis and septic shock) in compliance with SSC and WMA using fuzzy miner algorithm in Fluxicon Disco. *International Journal of Applied Decision Sciences*, *16*(4), 385-421.
147. Qi, Y., Schölkopf, B., & Jin, Z. (2024). Causal Responsibility Attribution for Human-AI Collaboration. *arXiv preprint arXiv:2411.03275*.

148. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).
149. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, *1*(1), 18.
150. Ramlakhan, S., Saatchi, R., Sabir, L., Singh, Y., Hughes, R., Shobayo, O., & Ventour, D. (2022). Understanding and interpreting artificial intelligence, machine learning and deep learning in Emergency Medicine. Emergency Medicine Journal, 39(5), 380-385.
151. Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. Physics of life reviews, 24, 1-16.
152. Razab-Sekh, N. N. (2021). Predictive Modelling of Sepsis using Probabilistic Fuzzy Systems and Genetic Programming.
153. Reichenbach, Hans (1956). The Direction of Time. Berkeley and Los Angeles, University of California Press.
154. Reijnders, K. (2020). Septic shock mortality prediction using deep learning and probabilistic fuzzy systems (Master's thesis, Eindhoven University of Technology).
155. Reyna, M. A., Josef, C., Seyedi, S., Jeter, R., Shashikumar, S. P., Westover, M. B., ... & Clifford, G. D. (2019). Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In 2019 Computing in Cardiology (CinC), pages Page–1. IEEE, Page–1.
156. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
157. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
158. Robert, S., Faghihi, U., Barkaoui, Y., & Ghazzali, N. (2020). Causality in probabilistic fuzzy logic and alternative causes as fuzzy duals. In *Advances in Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12* (pp. 767-776). Springer International Publishing.
159. Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34-58.
160. Rubulotta, F., Marshall, J. C., Ramsay, G., Nelson, D., Levy, M., & Williams, M. (2009). Predisposition, insult/infection, response, and organ dysfunction: A new model for staging severe sepsis. Critical care medicine, 37(4), 1329-1335.
161. Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., ... & Naghavi, M. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet, 395(10219), 200-211.
162. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.
163. Sabol, P., Sinčák, P., Hartono, P., Kočan, P., Benetinová, Z., Blichárová, A., ... & Jašková, A. (2020). Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. Journal of biomedical informatics, 109, 103523.
164. Saki, A., & Faghihi, U. (2024). Integrating Fuzzy Logic with Causal Inference: Enhancing the Pearl and Neyman-Rubin Methodologies. *arXiv preprint arXiv:2406.13731*.
165. Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
166. Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press.
167. Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh press.

168. Salmon, W. C. (2010). *Statistical explanation and statistical relevance* (Vol. 69). University of Pittsburgh Pre.
169. Saporta, G. (2022). Équité, explicabilité, paradoxes et biais. *Statistique et société*, (10| 3), 13-23. https://journals.openedition.org/statsoc/539.
170. Schinkel, M., Paranjape, K., Panday, R. N., Skyttberg, N., & Nanayakkara, P. W. (2019). Clinical applications of artificial intelligence in sepsis: a narrative review. *Computers in biology and medicine*, *115*, 103488.
171. Sheetrit, E., Nissim, N., Klimov, D., & Shahar, Y. (2019, July). Temporal probabilistic profiles for sepsis prediction in the ICU. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2961-2969).
172. Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. Jama, 320(21), 2199-2200.
173. Simon, H. A. (1990). Bounded rationality. In Utility and probability, pages 15–18. Springer.
174. Solomonoff, R. J. (2009). Algorithmic probability: Theory and applications. Information theory and statistical learning, 1-23.
175. Strickler, E. A., Thomas, J., Thomas, J. P., Benjamin, B., & Shamsuddin, R. (2023). Exploring a global interpretation mechanism for deep learning networks when predicting sepsis. Scientific Reports, 13(1), 3067.
176. Subbe, C. P., Kruger, M., Rutherford, P., & Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. Qjm, 94(10), 521-526.
177. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html
178. Thiagarajan, J. J., Thopalli, K., Rajan, D., & Turaga, P. (2022). Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific reports*, *12*(1), 597.
179. Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence, 302, 103627.
180. Tsoukalas, A., Albertson, T., & Tagkopoulos, I. (2015). From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR medical informatics*, *3*(1), e3445.
181. Tung, W. L., & Quek, C. (2009, August). A mamdani-takagi-sugeno based linguistic neural-fuzzy inference system for improved interpretability-accuracy representation. In 2009 ieee international conference on fuzzy systems (pp. 367-372). IEEE.
182. Valik, J. K., Ward, L., Tanushi, H., Johansson, A. F., Färnert, A., Mogensen, M. L., ... & Nauclér, P. (2023). Predicting sepsis onset using a machine learned causal probabilistic network algorithm based on electronic health records data. *Scientific reports*, *13*(1), 11760.
183. van Ruler, O., Kiewiet, J. J., Boer, K. R., Lamme, B., Gouma, D. J., Boermeester, M. A., & Reitsma, J. B. (2011). Failure of available scoring systems to predict ongoing infection in patients with abdominal sepsis after their initial emergency laparotomy. BMC surgery, 11, 1-9.
184. Waltman, L., Kaymak, U., & van den Berg, J. (2005, May). Maximum likelihood parameter estimation in probabilistic fuzzy classifiers. In The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05. (pp. 1098-1103). IEEE.
185. Wang, C., Liu, H., Li, C., Sun, Y., Wang, W., & Wang, B. (2023). Robust intrusion detection for industrial control systems using improved autoencoder and bayesian Gaussian mixture model. *Mathematics*, *11*(9), 2048.
186. Wang, D., Li, J., Sun, Y., Ding, X., Zhang, X., Liu, S., ... & Sun, T. (2021). A machine learning model for accurate prediction of sepsis in ICU patients. Frontiers in public health, 9, 754348.

187. Ward, L., Paul, M., & Andreassen, S. (2017). Automatic learning of mortality in a CPN model of the systemic inflammatory response syndrome. *Mathematical Biosciences*, *284*, 12-20.
188. Woodman, R. J., & Mangoni, A. A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. Aging Clinical and Experimental Research, 35(11), 2363-2397.
189. Woodward, J. (1989). The causal mechanical model of explanation.
190. Xu, P., Chen, L., Zhu, Y., Yu, S., Chen, R., Huang, W., ... & Zhang, Z. (2022). Critical care database comprising patients with infection.
191. Yager, R. R., & Zadeh, L. A. (Eds.). (2012). *An introduction to fuzzy logic applications in intelligent systems* (Vol. 165). Springer Science & Business Media.
192. Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., ... & Kang, B. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Human-Centric Intelligent Systems, 1-28.
193. Yang, Y., Tresp, V., Wunderle, M., & Fasching, P. A. (2018, June). Explaining therapy predictions with layer-wise relevance propagation in neural networks. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 152-162). IEEE.
194. Yao, S., Yang, J. B., Xu, D. L., & Dark, P. (2021). Probabilistic modeling approach for interpretable inference and prediction with data for sepsis diagnosis. *Expert Systems with Applications*, *183*, 115333.
195. Yapici, S. B., Üzücek, D. M., Urfalioglu, A. B., Yildiz, D., Sener, K., Kaya, A., ... & Yolcu, S. (2022). Quick Sequential Organ Failure Assessment-Mortality (qSOFAm): A New Scoring System to Predict the Mortality of Sepsis Patients. Southern Clinics of Istanbul Eurasia, 33(4).
196. Zadeh, L. A. (1996). Discussion: Probability theory and fuzzy logic are complementary rather than competitive. In *Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh* (pp. 805-810).
197. Zarrin, M. (2022). Inferring causal networks of health care resilience and safety performance indicators: A two-stage fuzzy cognitive map approach. Socio-Economic Planning Sciences, 84, 101389.
198. Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2022). Explainability of deep neural networks for MRI analysis of brain tumors. *International journal of computer assisted radiology and surgery*, *17*(9), 1673-1683.
199. Zhang, T. Y., Zhong, M., Cheng, Y. Z., & Zhang, M. W. (2023). An interpretable machine learning model for real-time sepsis prediction based on basic physiological indicators. European Review for Medical & Pharmacological Sciences, 27(10).
200. Zhang, X. (2013). *Analyse de la similarité du code source pour la réutilisation automatique de tests unitaires à l'aide du CBR* (Doctoral dissertation, Université Laval).
201. Zhu, W., McBrearty, I. W., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2022). Earthquake phase association using a Bayesian Gaussian mixture model. *Journal of Geophysical Research: Solid Earth*, *127*(5), e2021JB023249.