Attribution par argumentation de caractéristiques sensibles au contexte

Jinfeng Zhong*, Elsa Negre*

*Paris-Dauphine University, PSL Research University,
CNRS UMR 7243, LAMSADE, 75016 Paris France
jinfeng.zhong@dauphine.eu
elsa.negre@lamsade.dauphine.fr

Résumé. L'attribution des caractéristiques est une tâche fondamentale à la fois en apprentissage automatique et en analyse de données, qui consiste à déterminer la contribution des caractéristiques ou variables individuelles à la sortie d'un modèle. Ce processus aide à identifier les caractéristiques les plus importantes pour prédire un résultat. L'histoire des méthodes d'attribution de caractéristiques remonte aux Modèles Additifs Généraux (GAMs). Ces dernières années, des méthodes basées sur les gradients et des modèles de substitution ont été appliqués pour démêler les systèmes complexes d'Intelligence Artificielle (IA), mais ces méthodes ont leurs limites. Pour pallier les limitations des méthodes existantes et faire avancer l'état de l'art, nous définissons un nouveau cadre d'attribution de caractéristiques appelé ***C**ontext-**A**ware **F**eature **A**ttribution **T**hrough **A**rgumentation (CA-FATA)*. Notre modèle utilise la puissance de l'argumentation en traitant chaque caractéristique comme un argument qui peut soit soutenir, attaquer ou neutraliser une prédiction. De plus, CA-FATA formule l'attribution des caractéristiques comme une procédure d'argumentation, et chaque calcul a une sémantique explicite, ce qui la rend facilement compréhensible. CA-FATA intègre également facilement des informations annexes, telles que les contextes des utilisateurs, aboutissant à des prédictions plus précises. Nos expériences sur deux jeux de données réels démontrent que CA-FATA, ou l'une de ses variantes, surpasse les méthodes existantes basées sur l'argumentation et atteint une performance compétitive par rapport aux méthodes existantes sans contexte et sensibles au contexte.

1 Introduction

Le présent papier résume notre papier accepté à l'atelier CARSs (*Context-aware recommender systems*) de la conférence Recsys 2023 (Zhong et Negre, 2023).

L'attribution de caractéristiques est une pratique de longue date dans le domaine de l'apprentissage automatique pour déterminer la contribution des caractéristiques individuelles ou des variables à la sortie d'un modèle. Cette méthode a également été appliquée dans les modèles de recommandation pour expliquer leurs comportements (Zhong et Negre, 2022a; Rago et al., 2018; Lundberg et Lee, 2017; Ribeiro et al., 2016; Zhong et Negre, 2022b). Le processus

d'attribution de caractéristiques aide à identifier les caractéristiques les plus importantes pour prédire un résultat et les domaines d'amélioration du modèle. L'origine des méthodes d'attribution de caractéristiques remonte aux Modèles Additifs Généraux (GAMs) (Hastie, 2017). Bien que les GAMs soient intrinsèquement interprétables, ils souffrent souvent d'une expressivité limitée (Molnar, 2020). Ces dernières années, des méthodes basées sur les gradients ont été utilisées pour démêler les systèmes d'Intelligence Artificielle (IA) complexes. Ces méthodes déterminent l'importance d'une caractéristique x dans une fonction f en calculant la dérivée de f par rapport à x. Cependant, il se peut que les méthodes basées sur les gradients aient du mal avec des tâches simples qui nécessitent la compréhension d'une région modérément locale (Bilodeau et al., 2022), et l'interprétation de tels gradients peut être un défi pour les non-experts. Pour remédier aux problèmes des méthodes basées sur les gradients, des modèles de substitution tels que LIME (Ribeiro et al., 2016) et SHAP (Lundberg et Lee, 2017) ont émergé en tant que deux méthodes d'explication post-hoc proéminentes. Cependant, les limites de ces méthodes ont été reconnues. LIME souffre intrinsèquement de problèmes de stabilité, et pour SHAP, l'attribution de l'importance des caractéristiques par des propriétés formalisables mathématiquement (c'est-à-dire, précision locale, absence et cohérence) ne peut pas toujours correspondre aux attentes des utilisateurs pour les explications (Kumar et al., 2020).

Ces dernières années, les méthodes basées sur l'argumentation ont gagné une attention significative dans le domaine de l'Intelligence Artificielle Explicable (XAI) (Vassiliades et al., 2021). Cela est dû à leur capacité de représenter les relations de manière claire et compréhensible, telles que le soutien et l'attaque, offerts par les Cadres d'Argumentation (AFs), qui explicite les calculs. Avec les AFs, les processus de prise de décision peuvent être représentés visuellement, et les décisions optimales peuvent être expliquées à l'aide de propriétés bien définies (Vassiliades et al., 2021). Des arguments pondérés sont utilisés pour représenter la force des arguments et les relations dialectiques entre eux, tels que le soutien et l'attaque. La fonction de force des arguments peut être soigneusement conçue pour satisfaire les concepts généralisés de faible équilibre (Rago et al., 2018) et de faible monotonie (Rago et al., 2021), qui caractérisent comment les arguments influencent la prise de décision. Ces méthodes peuvent être utilisées pour expliquer les décisions prises à travers une représentation graphique. Les Systèmes de Recommandation Sensibles au Contexte (CARS) sont un sujet de recherche important dans les systèmes de recommandation. Les CARS peuvent modéliser les préférences des utilisateurs sous différentes situations contextuelles avec une granularité plus fine et générer des recommandations plus personnalisées adaptées aux contextes des utilisateurs. Nous croyons que le contexte est également crucial dans les cadres d'argumentation, car certains arguments considérés comme bons dans un contexte peuvent devenir moins précis dans un autre contexte. Par conséquent, il est important de tirer parti des contextes lors de l'application de l'argumentation (Teze et al., 2018).

Compte tenu des défis d'interprétabilité associés aux méthodes traditionnelles d'attribution de caractéristiques, il est raisonnable d'explorer de nouvelles voies pour améliorer l'explicabilité des modèles d'apprentissage automatique. Puisque l'argumentation offre intrinsèquement de l'interprétabilité, une telle approche consiste à utiliser des techniques d'argumentation pour attribuer l'importance des caractéristiques. Dans ce papier, nous introduisons un nouveau cadre pour l'attribution de caractéristiques.

2 Travaux connexes

Notre recherche est étroitement liée à deux travaux antérieurs : le cadre *Aspect-Item (A-I)* introduit par Rago et al. (2018, 2021) et le *Attribute-aware argumentative recommender (A^3R)* que nous avons proposé (Zhong et Negre, 2022a). Les deux, A-I et A^3R, utilisent l'argumentation pour prédire les scores des utilisateurs envers les articles, traitant les articles et les caractéristiques comme des arguments qui peuvent s'attaquer ou se soutenir mutuellement pour expliquer les recommandations de manière argumentative. Cependant, ces méthodes ne prennent pas en compte l'influence des contextes des utilisateurs.

Le *Faible équilibre* (Rago et al., 2018) et la *faible monotonie* (Baroni et al., 2019) permettent de dériver des explications intuitives de manière argumentative. Essentiellement, le concept de *faible équilibre* concerne l'impact d'un argument sur ses affectés lorsque l'argument est le seul facteur les affectant, tandis que l'idée de *faible monotonie* se concentre sur la manière dont la puissance d'un argument change lorsqu'un de ses affecteurs est réduit au silence par rapport au point neutre.

Faible équilibre : L'intuition derrière cette notion est que si l'affecteur augmente la force de l'affecté, alors il soutient l'affecté. Cette idée a été formalisée comme *weak balance* par Rago et al. (2018). Avec un *faible équilibre*, les relations sous les cadres d'argumentation tels que les attaques (ou soutiens, neutralisations) peuvent être caractérisées comme des connexions entre affecteurs et affectés de la manière suivante : si un affecteur est isolé comme le seul argument qui affecte l'affecté, alors le premier réduit (ou augmente, ne change pas) la note prédite du dernier par rapport au point neutre.

Faible monotonie : L'idée est intuitive : si l'affecteur soutient l'affecté, alors réduire au silence l'affecteur diminuera la force de l'affecté ; si l'affecteur attaque l'affecté, alors réduire au silence l'affecteur augmentera la force de l'affecté ; si l'affecteur neutralise l'affecté, alors réduire au silence l'affecteur ne changera pas la force de l'affecté. Cette intuition a été formalisée comme *weak monotocity* par Baroni et al. (2019).

3 Notre proposition

Dans cette section, nous présentons notre modèle qui fait de l'attribution par argumentation de caractéristiques sensible au contexte : CA-FATA (*Context-Aware Feature Attribution Through Argumentation*), et comment générer des explications intrinsèques au modèle en utilisant CA-FATA, nous comparons ensuite notre méthode avec des travaux existants, puis nous présentons les résultats de nos expériences.

3.1 Attribution par argumentation de caractéristiques sensibles au contexte

La Figure 1 illustre la structure de CA-FATA, qui se compose de trois étapes : (i) calcul de la représentation des utilisateurs cibles dans la situation contextuelle cible pour garantir que les préférences des utilisateurs sont adaptées aux contextes et que les relations dialectiques des arguments tiennent également compte du contexte ; (ii) calcul des scores des utilisateurs vis-à-vis des caractéristiques des articles dans la situation contextuelle cible, qui sont ensuite utilisées pour déterminer les relations dialectiques ; (iii) agrégation des scores obtenus à l'étape précédente pour générer les scores des utilisateurs envers les articles.

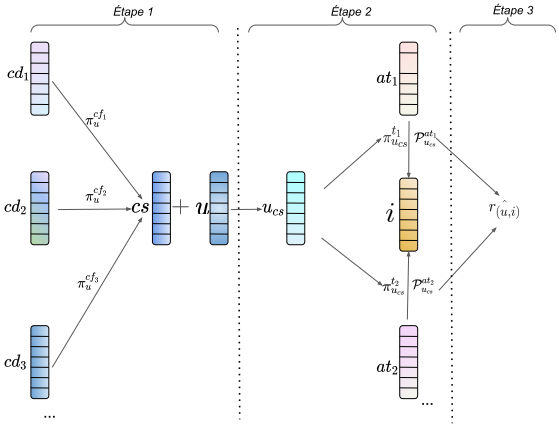


FIG. 1 – Illustration de la structure de CA-FATA.

Calcul de la représentation de l'utilisateur (Étape 1) : La représentation finale d'un utilisateur est déterminée par sa situation contextuelle. Dans cette étape, notre objectif est de calculer la représentation des utilisateurs adaptée à la situation contextuelle cible. Pour ce faire, nous commençons par calculer l'importance de chaque facteur contextuel dans l'Équation 1. L'importance calculée ici est similaire au poids de pertinence proposé par Budán et al. (2020). Cependant, contrairement à ces travaux, où le poids de pertinence du contexte est défini de manière empirique, dans notre travail, l'importance du contexte est apprise de manière pilotée par les données. Intuitivement, π_u^{cf} caractérise la mesure selon laquelle l'utilisateur u souhaite prendre en compte le facteur contextuel cf.

$$\pi_u^{cf} = \frac{exp(LeakyReLU(\langle \mathbf{u}, \mathbf{cf} \rangle))}{\sum_{cf \in C} exp(LeakyReLU(\langle \mathbf{u}, \mathbf{cf} \rangle))} \tag{1}$$

Ensuite, nous calculons la représentation de la situation contextuelle cs en additionnant tous les vecteurs représentant les conditions contextuelles multipliés par π_u^{cf} : $\mathbf{cs} = \sum_{cd \in cs} \pi u^{cf} \mathbf{cd}$, où **cs** est le vecteur qui dénote la situation contextuelle cs. L'étape suivante consiste à agréger la représentation de la situation contextuelle cs avec la représentation de l'utilisateur u pour obtenir une représentation spécifique de l'utilisateur u sous la situation contextuelle cs. Pour éviter d'avoir un nombre excessif de paramètres [1], nous additionnons **u** et **cs**. En agrégeant les informations d'une situation contextuelle cs et d'un utilisateur u, chaque utilisateur u obtient une représentation spécifique $\boldsymbol{u_{cs}}$ sous une situation contextuelle cs : $\boldsymbol{u_{cs}} = \mathbf{u} + \mathbf{cs}$.

Calcul des scores des utilisateurs envers les caractéristiques (Étape 2) : Les types de caractéristiques dans ce papier sont similaires aux relations dans les graphes de connaissances, qui sont des graphes orientés composés de triplets *entité-relation-entité* (Hogan et al., 2021). Par exemple, le triplet $(HarryPotter, aPourDirecteur, MikeNewell)$ indique que le film

1. Notez que d'autres méthodes d'agrégation telles que la concaténation sont également possibles mais induisent plus de paramètres. Nous laissons cette exploration pour des travaux futurs.

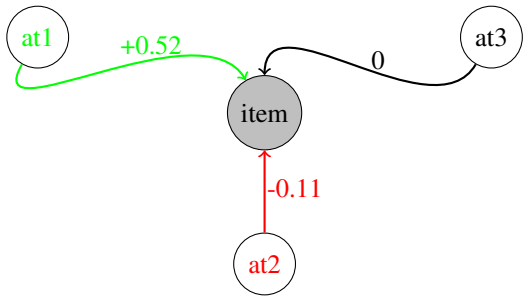


FIG. 2 – Représentation graphique d'une procédure d'argumentation dans un scénario de recommandation. Chaque nœud représente un argument, la valeur sur l'arc indique la force et la polarité de l'argument, +" signifie soutient, "-" signifie attaque et "0" signifie neutralise.

Harry Potter est réalisé par *Mike Newell*. Ici, $aPourDirecteur$ est une relation dans le graphe de connaissances qui concerne les films, et dans ce papier, cela correspond au type de caractéristiques $directeur$. Nous quantifions l'importance de chaque type de caractéristiques en utilisant l'Équation 2.

$$\pi^t_{u_{cs}} = \frac{exp(LeakyReLU(\langle \boldsymbol{u_{cs}}, \mathbf{at} \rangle))}{\sum_{t \in t_i} exp(LeakyReLU(\langle \boldsymbol{u_{cs}}, \mathbf{at} \rangle))} \tag{2}$$

Pour calculer les scores des utilisateurs envers les caractéristiques, nous adoptons à nouveau le produit scalaire : $\mathcal{P}^{at}_{u_{cs}} = g(\boldsymbol{u_{cs}}, \mathbf{at})$. La représentation d'un utilisateur sous un contexte diffère de celle sous un autre contexte. En conséquence, la représentation de l'utilisateur u_{cs} est spécifique à chaque contexte, et l'importance du type de caractéristiques et le score de l'utilisateur envers les caractéristiques sont également sensibles au contexte.

Agrégation des scores envers les caractéristiques (Étape 3) : Après avoir calculé l'importance de chaque type de caractéristiques et les scores des utilisateurs envers chaque caractéristique, le score de u envers i sous cs est :

$$\hat{r}(u,i) = \sum t \in t_i \pi^t_{u_{cs}} \times \frac{\sum_{at \in at^t_i} \mathcal{P}^{at}_{u_{cs}}}{|at^t_i|} \tag{3}$$

où t_i désigne tous les types de caractéristiques de l'article i. Il convient de noter que la valeur réelle du score de l'utilisateur u pour l'article i est un nombre réel entre -1 et 1, comme défini dans des travaux antérieurs tels que Rago et al. (2018, 2021). Il est à noter que l'Équation 3 est remarquablement similaire aux modèles additifs, indiquant que notre modèle appartient à la famille des modèles additifs généralisés. Cette similitude permet d'identifier facilement la contribution de chaque caractéristique.

3.2 Explications contextuelles

Rappelons que le véritable score $r_{(u,i)}$ est un nombre réel compris entre -1 et 1, alors le codomaine de $\mathcal{P}^{at}_{u_{cs}}$ est également censé être entre -1 et 1. Par conséquent, lorsque $\mathcal{P}^{at}_{u_{cs}} > 0$,

alors $\sigma(at) > 0$, indiquant que at est un argument qui soutient rec^i [2] ; quand $\mathcal{P}^{at}_{u_{cs}} = 0$, alors $\sigma(at) = 0$, indiquant at est un argument qui neutralise rec^i ; quand $\mathcal{P}^{at}_{u_{cs}} < 0$, alors $\sigma(at) < 0$, indiquant que at est un argument qui attaque rec^i. Par conséquent, le TAF (Cadre d'argumentation tripolaire) correspondant à une interaction utilisateur-article (u, i) sous cs peut être défini comme suit :

Définition 1. *Le TAF correspondant à* (u, i) *sous cs est un quadruplet :* $< \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \mathcal{R}^0 >$ *tel que :* $\mathcal{R}^- = \{(at, rec^i) | \mathcal{P}^{at}_{u_{cs}} < 0\}$ *;* $\mathcal{R}^+ = \{(at, rec^i) | \mathcal{P}^{at}_{u_{cs}} > 0\}$ *;* $\mathcal{R}^0 = \{(at, rec^i) | \mathcal{P}^{at}_{u_{cs}} = 0\}$.

Selon la Définition 1, $\mathcal{P}^{at}_{u_{cs}}$ détermine la polarité des arguments : si $\mathcal{P}^{at}_{u_{cs}}$ est positif, alors l'argument (caractéristique) soutient la recommandation de l'article i à l'utilisateur u ; si $\mathcal{P}^{at}_{u_{cs}}$ est négatif, alors l'argument attaque la recommandation de l'article i à l'utilisateur u ; si $\mathcal{P}^{at}_{u_{cs}}$ est 0, alors l'argument neutralise la recommandation. En posant $\sigma(at) = \mathcal{P}^{at}_{u_{cs}}$ et $\sigma(rec^i) = \hat{r}_{(u,i)}$, le TAF correspondant à (u, i) sous cs satisfait le *faible équilibre* et la *faible monotonie* (Preuve omise par manque d'espace, veuillez vous référer à notre papier). À titre d'exemple, la Figure 2 présente le TAF pour une interaction utilisateur-article sous la situation contextuelle cs. Dans ce TAF, chaque caractéristique de l'article représente un argument. Le score de l'utilisateur pour chaque caractéristique détermine la force et la polarité de l'argument, reflétant ainsi la préférence de l'utilisateur. La force de l'argument 1 est de $+0.52$, indiquant que l'utilisateur apprécie la caractéristique 1 (par exemple, un réalisateur ou un acteur de film), et cette caractéristique soutient la recommandation de l'article à l'utilisateur. La force de l'argument 2 est de -0.11, indiquant que l'utilisateur n'apprécie pas la caractéristique 2 et que cette caractéristique attaque la recommandation de l'article à l'utilisateur. Enfin, la force de l'argument 3 est de 0, indiquant que cette caractéristique n'influence pas le score de l'utilisateur. Notez que, selon les trois étapes de la Section 3.1, le score de prédiction pourrait varier sous différents contextes, même pour le même utilisateur et le même article. Par conséquent, les TAF correspondants pourraient également différer.

Après avoir mené les analyses ci-dessus, nous proposons trois modèles d'explication, similaires aux trois types d'explication de Rago et al. (2021), mais avec l'inclusion des contextes des utilisateurs. Dans chaque scénario, nous sélectionnons la condition contextuelle la plus influente (telle que déterminée par l'Équation 1). Pour une "recommandation forte", nous proposons de sélectionner les deux arguments les plus solides qui soutiennent la recommandation de l'article. Dans une "recommandation faible", nous proposons de sélectionner l'argument le plus fort qui soutient la recommandation de l'article et le plus fort qui attaque la recommandation de l'article. Dans "non recommandé", nous proposons de sélectionner les deux arguments les plus solides qui attaquent la recommandation de l'article. Chaque modèle inclut des informations contextuelles avec les arguments correspondants qui soutiennent ou attaquent la recommandation de l'article. En résumé, CA-FATA est un modèle polyvalent qui peut être utilisé pour expliquer à la fois les raisons pour lesquelles des articles sont recommandés ainsi que les raisons pour lesquelles certains articles ne devraient pas être recommandés. De plus, les utilisateurs ont la flexibilité de définir leurs propres modèles selon leurs besoins spécifiques.

2. Sémantiquement, l'utilisateur u préfère la caractéristique at.

3.3 Expérimentations

Nous avons mené des expériences sur les jeux de données réels suivants :

Frappé : Ce jeu de données a été collecté par Baltrunas et al. (2015). Il provient de Frappé, un système de recommandation d'applications sensible au contexte. Il comprend 96 303 journaux d'utilisation de 957 utilisateurs dans différentes situations contextuelles, incluant 4 082 applications. Suivant Unger et al. (2020), nous appliquons une transformation logarithmique au nombre d'interactions. En conséquence, le nombre d'interactions est échelonné de 0 à 4.46. Chaque situation contextuelle est composée de 7 conditions contextuelles et de cinq types de caractéristiques.

Yelp : Ce jeu de données contient les avis des utilisateurs sur des bars et restaurants dans des zones métropolitaines aux États-Unis et au Canada. Conformément aux études précédentes de Zhou et al. (2020) et Geng et al. (2022), nous utilisons les enregistrements du 1 janvier 2019 au 31 décembre 2019, qui contiennent 904 648 observations. Il y a 8 facteurs contextuels et trois types de caractéristiques. Pour les deux jeux de données, nous avons adopté le réglage 10-core, suivant Wang et al. (2019), pour assurer la qualité des données. Cela signifie que seuls les utilisateurs avec au moins 10 interactions sont conservés.

Les résultats de nos expériences montrent que CA-FATA surpasse toutes les méthodes de référence sur les jeux de données Yelp et Frappé, indiquant sa supériorité dans le traitement d'informations contextuelles complexes. Voici quelques observations spécifiques : (i) CA-FATA fonctionne bien sur les deux jeux de données, surpassant toutes les méthodes de référence, démontrant sa capacité à modéliser les préférences des utilisateurs dans différents contextes. Un autre avantage de CA-FATA est sa capacité à fournir des explications argumentatives, ce qui n'est pas possible pour ces méthodes de référence. (ii) Comparé à A-I, sur Yelp, CA-FATA réalise une réduction significative du RMSE (*Root Mean Square Error*) et MAE (*Mean Absolute Error*). (iii) Une comparaison horizontale des jeux de données Frappé et Yelp montre que CA-FATA fonctionne mieux sur Frappé que sur Yelp. Nous attribuons cette différence à la rareté du jeu de données, car Yelp reste très clairsemé même après l'application du réglage 10-core, avec une sparsité de 99.84%, tandis que Frappé a une sparsité de 94.47%. Pour résumer, les avantages de CA-FATA sont les suivants : (i) il atteint une performance compétitive par rapport aux méthodes de référence sans contexte et sensibles au contexte. Les méthodes de référence utilisent des méthodes basées sur la factorisation et certaines combinent des réseaux neuronaux, ce qui les rend difficiles à interpréter. D'autre part, CA-FATA fournit une sémantique explicite pour chaque calcul et génère des explications argumentatives ; (ii) comparé à la méthode basée sur l'argumentation A-I, CA-FATA améliore significativement la précision de prédiction et génère des explications sensibles au contexte.

4 Conclusion et perspectives

À la lumière des défis d'interprétabilité associés aux méthodes actuelles d'attribution de caractéristiques, nous présentons un nouveau cadre d'attribution de caractéristiques appelé **C**ontext-**A**ware **F**eature **A**ttribution **T**hrough **A**rgumentation (CA-FATA). CA-FATA est un cadre d'attribution de caractéristiques qui traite les caractéristiques comme des arguments pouvant soutenir, attaquer ou neutraliser une prédiction à l'aide de procédures d'argumentation. Cette approche fournit une sémantique explicite à chaque étape et permet une incorporation

facile du contexte de l'utilisateur pour générer des recommandations et des explications sensibles au contexte. L'armature argumentative dans CA-FATA est conçue pour satisfaire deux propriétés importantes : le *faible équilibre* et la *faible monotonie*, qui mettent en évidence comment les caractéristiques influencent une prédiction. Ces propriétés aident à identifier les caractéristiques importantes et à étudier comment elles influencent la tâche de prédiction. Nous introduisons également trois scénarios d'explication - recommandation forte, recommandation faible et non recommandé, qui peuvent être utilisés pour expliquer pourquoi des articles ont été recommandés ou non. Des investigations supplémentaires montrent que CA-FATA peut être intégré avec des systèmes de recommandation interactifs, qui prennent en compte les retours immédiats des utilisateurs pour améliorer et adapter les recommandations en cours (Veuillez vous référer à notre github pour plus de détails (Zhong et Negre, 2023)). Nos résultats expérimentaux montrent que CA-FATA surpasse plusieurs méthodes de référence solides en termes de RMSE et MAE, soulignant sa capacité à fournir également de la précision. À l'avenir, nous prévoyons d'explorer l'applicabilité de CA-FATA dans d'autres domaines pour vérifier sa généralisabilité. Pour calculer le score du facteur contextuel et du type de caractéristiques, nous avons adopté le produit scalaire (Équation 2). Nous prévoyons d'explorer d'autres fonctions à cette fin. De plus, nous avons l'intention de mener des études utilisateurs pour évaluer et comparer la qualité des explications générées par d'autres méthodes d'explication.

Références

Baltrunas, L., K. Church, A. Karatzoglou, et N. Oliver (2015). Frappe : Understanding the usage and perception of mobile app recommendations in-the-wild. *preprint arXiv :1505.03014*.

Baroni, P., A. Rago, et F. Toni (2019). From fine-grained properties to broad principles for gradual argumentation : A principled spectrum. *International Journal of Approximate Reasoning 105*, 252–286.

Bilodeau, B., N. Jaques, P. W. Koh, et B. Kim (2022). Impossibility theorems for feature attribution. *arXiv preprint arXiv :2212.11870*.

Budán, M. C., M. L. Cobo, D. C. Martinez, et G. R. Simari (2020). Proximity semantics for topic-based abstract argumentation. *Information Sciences 508*, 135–153.

Geng, S., S. Liu, Z. Fu, Y. Ge, et Y. Zhang (2022). Recommendation as language processing (rlp) : A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 299–315.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge.

Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. (2021). Knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge 12*(2), 1–257.

Kumar, I. E., S. Venkatasubramanian, C. Scheidegger, et S. Friedler (2020). Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. *NeurIPS 30*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Rago, A., O. Cocarascu, C. Bechlivanidis, D. Lagnado, et F. Toni (2021). Argumentative explanations for interactive recommendations. *Artificial Intelligence 296*, 103506.

Rago, A., O. Cocarascu, et F. Toni (2018). Argumentation-based recommendations : Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 1949–1955.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). " 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.

Teze, J. C. L., L. Godo, et G. R. Simari (2018). An argumentative recommendation approach based on contextual aspects. In *SUM*, pp. 405–412. Springer.

Unger, M., A. Tuzhilin, et A. Livne (2020). Context-aware recommendations based on deep learning frameworks. *ACM Transactions on Management Information Systems 11*(2), 1–15.

Vassiliades, A., N. Bassiliades, et T. Patkos (2021). Argumentation and explainable artificial intelligence : a survey. *The Knowledge Engineering Review 36*.

Wang, X., X. He, M. Wang, F. Feng, et T.-S. Chua (2019). Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR*, pp. 165–174.

Zhong, J. et E. Negre (2022a). A 3 r : Argumentative explanations for recommendations. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–9. IEEE.

Zhong, J. et E. Negre (2022b). Shap-enhanced counterfactual explanations for recommendations. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pp. 1365–1372.

Zhong, J. et E. Negre (2023). Context-aware feature attribution through argumentation. *arXiv preprint arXiv :2310.16157*.

Zhou, K., H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, et J.-R. Wen (2020). S3-rec : Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1893–1902.

Summary

Feature attribution is a fundamental task in both machine learning and data analysis, which involves determining the contribution of individual features or variables to a model's output. This process helps identify the most important features for predicting an outcome. The history of feature attribution methods can be traced back to General Additive Models (GAMs). In recent years, gradient-based methods and surrogate models have been applied to unravel complex Artificial Intelligence (AI) systems, but these methods have limitations. To address the limitations of existing methods and advance the current state-of-the-art, we define a novel feature attribution framework called **C**ontext-**A**ware **F**eature **A**ttribution **T**hrough **A**rgumentation

(CA-FATA). Our framework harnesses the power of argumentation by treating each feature as an argument that can either support, attack or neutralize a prediction. Additionally, CA-FATA formulates feature attribution as an argumentation procedure, and each computation has explicit semantics, which makes it easily understandable. CA-FATA also easily integrates side information, such as users' contexts, resulting in more accurate predictions. Our experiments on two real-world datasets demonstrate that CA-FATA, or one of its variants, outperforms existing argumentation-based methods and achieves competitive performance compared to existing context-free and context-aware methods.

Comment les biais cognitifs affectent la prise de décision assistée par l'IA explicable

Rafik Belloum*, Astrid Bertrand**,
James R. Eagan**, Winston Maxwell***

*Univ. Polytechnique Hauts-de-France LAMIH, CNRS, UMR 8201 F-59313 Valenciennes, France
rafik.belloum@uphf.fr
**LTCI, Institut Polytechnique de Paris France
astrid.bertrand@telecom-paris.fr
james.eagan@telecom-paris.fr
***i3, CNRS, Institut Polytechnique de Paris France
winston.maxwell@telecom-paris.fr

Résumé. Ce papier résume une revue de la littérature sur les biais cognitifs influençant la prise de décision assistée par l'IA explicable (XAI). Il va au-delà de la simple identification des biais cognitifs en XAI, offrant une vision stratégique, illustrée par une carte heuristique qui guide le futur développement de systèmes XAI plus en phase avec les processus cognitifs humains. Il convient de noter que ce résumé synthétise un article déjà publié par Bertrand et al. (2022).

1 Introduction

Le domaine de l'Intelligence Artificielle Explicable vise à apporter de la transparence aux systèmes d'IA complexes. Bien qu'il soit généralement considéré comme un domaine essentiellement technique, des efforts ont récemment été déployés pour mieux comprendre les méthodes d'explication humaine des utilisateurs et les contraintes cognitives. Malgré ces avancées, la communauté manque d'une vision générale de la manière dont les biais cognitifs affectent les systèmes d'explicabilité. Cet article, déjà paru et que nous souhaitons résumer ici, comble cette lacune en présentant une cartographie heuristique novatrice, alignant les biais cognitifs humains avec les techniques d'explicabilité issues de la littérature XAI, et structurée autour de la prise de décision assistée par la XAI.

2 Cartographie des Biais Cognitifs en XAI

L'article propose une cartographie de biais cognitifs identifiés dans la littérature XAI, offrant une vue détaillée de leur présence et de leur impact. L'utilisation du guide PRISMA (Moher et al., 2009) pour la revue de la litterature garantit la rigueur méthodologique. Ces biais sont catégorisés en fonction de différents contextes, tels que le type d'explicabilité utilisé, le domaine d'application, la tâche assistée par l'IA et le type d'utilisateur (expert du domaine, expert en IA ou utilisateur lambda).

Biais cognitifs affectant la conception des méthodes XAI. Les résultats du papier soulignent comment certains biais cognitifs influent sur la conception des méthodes XAI. Ces biais (Miller, 2019), présentés dans des boîtes jaunes sur la carte heuristique de la figure 1, sont liés aux heuristiques explicatives que les individus utilisent lors de l'explication ou de la réception d'une explication. Contrairement à d'autres biais, ces heuristiques ne sont pas considérées comme des "erreurs", mais plutôt comme des contraintes à prendre en compte lors de la conception des techniques d'explicabilité.

Biais cognitifs impactant l'évaluation des techniques XAI dans les études utilisateur. Une autre catégorie de biais cognitifs, présentée dans une boîte marron sur la carte heuristique de la figure 1, concerne la distorsion potentielle dans l'évaluation des techniques XAI lors d'études utilisateur. Cela découle des préoccupations croissantes quant à la nécessité de tester les explications avec les utilisateurs. Certains chercheurs insistent sur cette approche, tandis que d'autres la déconseillent, craignant que les biais cognitifs ne faussent les évaluations et trompent le domaine XAI (Herman, 2017).

Atténuation et Exacerbation par les Techniques XAI. L'article identifie également des biais cognitifs qui peuvent être atténués avec succès par les techniques XAI (couleur orange sur la figure 1) (Wang et al., 2019; Bertrand et al., 2023), tout en soulignant que certaines méthodes peuvent également exacerber certains biais (couleur rouge). Cette distinction est cruciale pour guider le développement de futures techniques XAI qui minimisent les effets négatifs sur la prise de décision humaine.

Références

Bertrand, A., R. Belloum, J. R. Eagan, et W. Maxwell (2022). How cognitive biases affect xai-assisted decision-making : A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 78–91.

Bertrand, A., T. Viard, R. Belloum, J. R. Eagan, et W. Maxwell (2023). On selective, mutable and dialogic xai : a review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21.

Herman, B. (2017). The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv :1711.07414*.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence 267*, 1–38.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, et P. Group* (2009). Preferred reporting items for systematic reviews and meta-analyses : the prisma statement. *Annals of internal medicine 151*(4), 264–269.

Wang, D., Q. Yang, A. Abdul, et B. Y. Lim (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15.

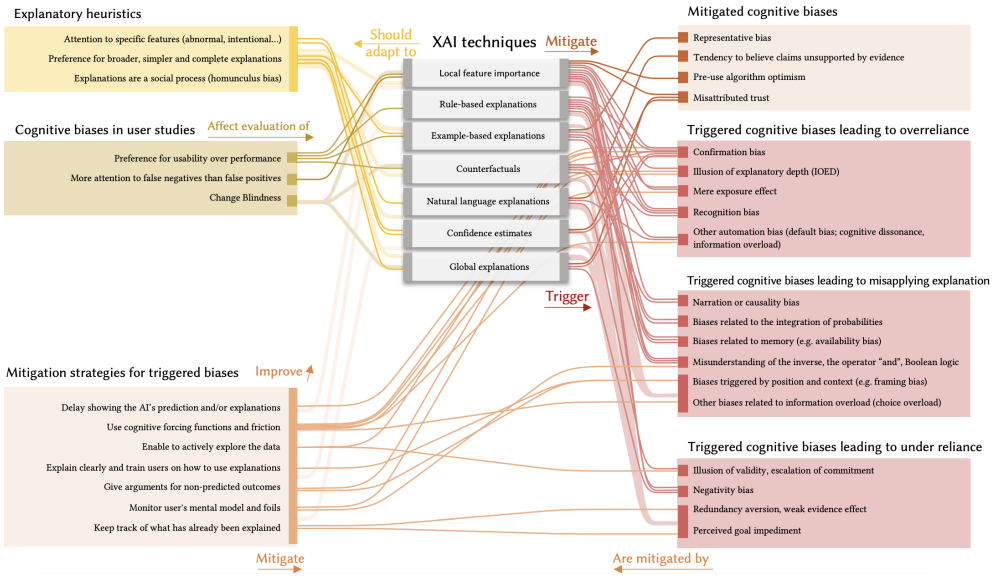


FIG. 1 – *Résumé des contraintes cognitives, des biais et des stratégies d'atténuation identifiés dans les articles inclus dans le corpus (n=37). Ce diagramme présente les différentes catégories de techniques d'explication observées dans le corpus (au centre). Chaque lien représente une connexion établie dans la littérature entre une technique d'explicabilité et un biais cognitif, ou entre un biais cognitif et une technique d'atténuation. Les légendes en couleur soulignées par des flèches indiquent comment et dans quelle direction les liens doivent être lus (par exemple, "Les techniques d'XAI devraient s'adapter aux heuristiques explicatives"). Les liens pâles et larges indiquent que le biais ou la contrainte cognitive s'applique de manière plus générale à toutes les méthodes d'XAI. Il a été identifié davantage de connexions entre les biais et les stratégies d'atténuation, mais seules les plus soutenues sont présentées pour des raisons de concision.*

Summary

This paper summarizes a literature review on cognitive biases influencing XAI-assisted decision-making. It goes beyond mere identification of cognitive biases in XAI, providing a strategic vision, illustrated through a heuristic map, guiding the future development of XAI systems that are more attuned to human cognitive processes. This innovative approach significantly contributes to the evolution of the field by emphasizing alignment with how individuals comprehend and utilize explanations provided by AI systems.

ConceptECGxAI : une approche post-hoc à base de concepts médicaux pour expliquer un modèle d'apprentissage profond d'aide au diagnostic cardiaque

Victoria Bourgeais*, Ahmad Fall**, Alex Lence**, Joe-Elie Salem****,♯
Jean-Daniel Zucker**,***, Edi Prifti**,***, Blaise Hanczar♯♯♯

* LaBRI (CNRS/UMR 5800), Université de Bordeaux, Talence, France
Auteur correspondant : victoria.bourgeais@u-bordeaux.fr
** IRD, Sorbonne Université, UMMISCO, F-93143, Bondy, France
*** Sorbonne Université, INSERM, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France
**** Vanderbilt University Medical Center, Nashville, USA
♯ Centre d'Investigation Clinique Paris-Est, INSERM, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France
♯♯ Laboratoire IBISC (EA 4526), Université Paris-Saclay (Univ Evry), Evry, France

Résumé. Le développement de l'intelligence artificielle (IA) participe à l'émergence d'une nouvelle forme de médecine dite personnalisée, qui vise à mieux prendre en compte les caractérisiques des patients. Dans ce contexte, nous nous intéressons à l'application de l'apprentissage profond pour détecter des maladies cardiovasculaires telles que le syndrome du QT-long à partir d'électrocardiogrammes. Cependant, une préoccupation majeure réside dans la nature souvent opaque de ces modèles d'IA dits "boîte noire". Cet article vise à rendre ces derniers plus interprétables en intégrant des concepts médicaux dans une nouvelle approche post-hoc, ConceptECGxAI, fournissant des explications plus intelligibles aux médecins.

1 Introduction

L'apprentissage profond est une avancée majeure dans le domaine de l'intelligence artificielle (IA) de ces dernières années. Il s'est rapidement imposé comme un nouveau standard dans plusieurs domaines en surpassant les performances des méthodes antérieures considérées comme l'état de l'art. Ses domaines de prédilection sont principalement l'analyse d'images et le traitement du langage naturel. Un des futurs enjeux majeurs de cette approche est lié aux applications dans le domaine de la santé. Au sein du projet **ANR DeepECG4U** [1], nous nous intéressons à l'application des approches d'apprentissage profond pour la détection de maladies cardio-vasculaires comme le syndrome du QT-long qui peut déclencher des arythmies mortelles, telles que la Torsades-de-Pointes (TdP), à partir des données d'électrocardiogrammes (ECG) des patients. La prise de certains médicaments peut induire l'allongement de l'intervalle QT et en être la cause de la TdP, comme peuvent également être certaines mutations

1. https://anr.fr/fr/projets-finances-et-impact/projets-finances/projet/funded/project/anr-20-ce17-0022/

spécifiques (*i.e.*, le QT-long congénital). Nous souhaitons ainsi proposer un outil d'aide au diagnostic à destination des médecins afin de prévenir ces risques. Dans ce sens, un premier outil à base d'apprentissage profond a été développé (Prifti et al., 2021).

Cependant, les modèles d'apprentissage profond, ainsi que d'autres méthodes d'apprentissage automatique comme les machines à vecteurs de support, sont considérés comme des « boîtes noires », dans lesquelles les données de patients sont injectées en entrée, puis une prédiction est retournée en sortie sans aucune explication. Ceci est un gros problème et un point de réflexion actif chez les législateurs. L'Union Européenne a récemment adopté un texte imposant aux utilisateurs d'algorithmes d'apprentissage automatique d'être capables d'expliquer les décisions d'un modèle prédictif (Goodman et Flaxman, 2017). Premièrement, il est important de s'assurer que les modèles d'apprentissage automatique basent leurs prédictions sur des représentations fiables des patients et ne se concentrent pas sur des artefacts non pertinents présents dans les données d'apprentissage, autrement dit qu'ils ne soient pas sensibles aux biais. Deuxièmement, un modèle performant pour la prédiction d'une certaine maladie ou condition, peut avoir identifié une signature dans les données qui pourrait être une piste de recherche pour les médecins, pouvant renseigner sur la physiopathologie de la maladie en question.

Dans l'état de l'art actuel, il existe deux approches principales pour interpréter les réseaux de neurones : en créant des modèles qui sont par essence interprétables (approche dite *ante-hoc* ou *auto-explicative*), ou en ayant recours à une méthode tierce dédiée à l'interprétation du réseau de neurones déjà appris (approche dite *post-hoc*). Quelle que soit la méthode choisie, l'explication fournie consiste généralement en l'identification des variables d'entrée et des neurones importants pour la prédiction. Or, dans le cas d'une application notamment sur les données de santé comme les ECG, cela n'est pas suffisant. Une des pistes d'amélioration pourrait être d'avoir recours à l'utilisation de concepts de plus haut niveau sémantique pour fournir des explications qui utilisent le même langage que celui des médecins. Un premier travail a été réalisé avec une méthode d'occlusion en croisant les explications obtenues avec les connaissances du domaine (Prifti et al., 2021). Une solution alternative est d'intégrer directement ces concepts dans l'approche post-hoc. Ainsi, dans la continuité des travaux précédents sur les données d'image (Kim et al., 2018; Zaeem et Komeili, 2021; Crabbé et van der Schaar, 2022), nous proposons une nouvelle approche post-hoc qui permet d'interpréter n'importe quel type de réseau de neurones boîte noire déjà appris sur des données ECG en intégrant des concepts médicaux ayant un sens sémantique pour les cardiologues. Cette approche se nomme ConceptECGxAI.

2 Méthode

Un ECG enregistre l'activité électrique du cœur à l'aide de plusieurs dérivations, obtenues à partir d'électrodes. La Fig. 1a illustre le tracé normal d'un battement cardiaque, où il est possible d'identifier la position des ondes (P,T) et des pics du complexe QRS (correspondant à l'onde de dépolarisation des ventricules cardiaques). À partir de ces positions, diverses mesures peuvent être réalisées, telles que la durée du complexe QRS, l'intervalle inter-battement R-R, la distance QT, ainsi que les amplitudes des différentes ondes P, T, la tangente de l'onde T, etc. Ces informations permettent aux cardiologues d'évaluer la normalité de l'ECG ou d'identifier d'éventuelles anomalies, comme le QT-long.

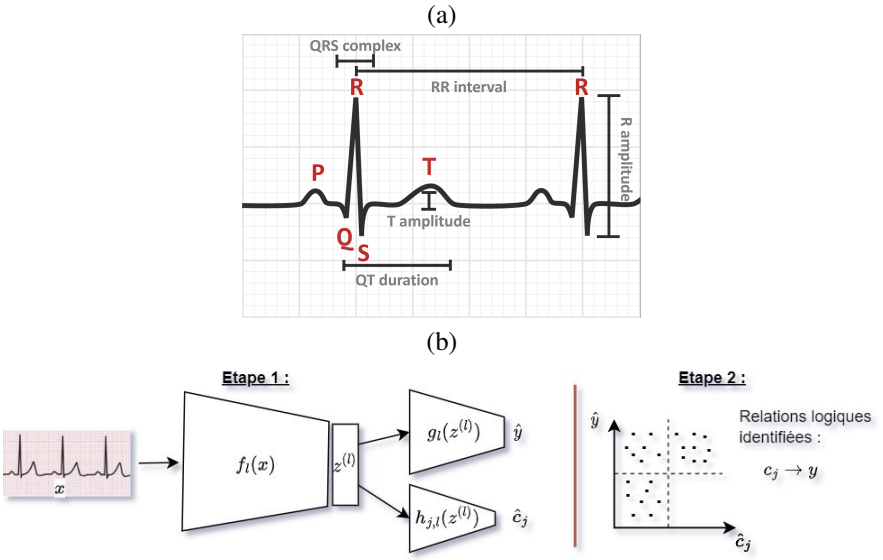


FIG. 1 – (a) Exemple d'un électrocardiogramme de patient annoté. (b) Illustration de ConceptECGxAI pour l'identification d'un concept c_j donné (tel que <qt-duration-long>).

Supposons que nous disposons d'un réseau de neurones prédictif déjà entraîné et qui prend en entrée un électrocardiogramme. ConceptECGxAI vise à accomplir deux objectifs principaux : (i) identifier la présence de concepts significatifs pour le domaine d'application dans les couches cachées du modèle prédictif, puis (ii) établir les relations logiques entre ces concepts et les prédictions du modèle en question. Pour ce faire, nous proposons un pipeline en deux étapes comme illustré dans la Fig. 1b. Une première étape consiste à apprendre par concept une méthode tierce (un interpréteur) permettant d'évaluer à quel point le concept en question est représenté dans une des couches latentes du modèle prédictif. Les concepts sont définis de manière à être compréhensibles pour les utilisateurs finaux, c.-à-d. suffisamment abstraits et représentatifs dans le domaine de compétences de ce dernier. Noter que le modèle prédictif n'a pas été contraint dans son apprentissage à extraire ces concepts. Dans la deuxième étape, nous cherchons à identifier l'existence de relations logiques entre les concepts les plus capturés et les classes de prédiction du modèle prédictif pour construire des règles du type : *"si le concept <QT-long> est capturé dans une des couches latentes du modèle prédictif, alors il est très probable que celui-ci classifie le patient comme étant à risque de <TdP>"*. Les règles ainsi construites pourraient permettre de fournir une interprétation globale du modèle et les interpréteurs permettront de vérifier la présence de ces concepts sur de nouvelles données.

2.1 Description du pipeline

Soit un ensemble de concepts binaires $\{c_j\}_{j=1,\dots,C}$ où C correspond au nombre de concepts. On désigne respectivement par $(x, y, \{c_{x,j}\}_{j=1,\dots,C})$ un profil d'électrocardiogramme d'une classe, l'indicateur booléen d'appartenance à la classe positive à prédire et les indicateurs boo-

léens de présence des concepts $\{c_j\}_{j=1,\ldots,C}$. L'approche ConceptECGxAI est illustrée dans la Fig. 1b.

Étape 1 : On étudie la présence d'un concept c_j dans une des représentations latentes $\{z^{(l)}\}_{l=1,\cdots,L}$ du réseau de neurones avec L la profondeur du réseau de neurones. Le réseau est découpé en deux parties distinctes : la partie avant la couche cachée (l) étudiée ($f_l(x)$), et celle après ($g_l(z^{(l)})$). Le réseau entier est le résultat de la composition de fonctions telles que $\hat{y} = g_l(f_l(x))$ où la notation $\hat{}$ signifie la probabilité de prédiction retournée par un algorithme de classification. f_l et g_l varient en fonction du point de découpe, c.-à-d. la couche l. Le choix de la couche cachée est un hyperparamètre de l'approche. Pour déterminer la présence de concepts, on transforme le problème sous la forme d'une tâche de classification en utilisant une méthode tierce, basée sur un interpréteur h qui prend en entrée une représentation latente $z^{(l)}$. Cet interpréteur peut être un modèle d'apprentissage automatique simple tel qu'un modèle linéaire ou encore plus complexe, tel qu'une machine à vecteurs de support (SVM) ou un réseau de neurones. Ce choix dépendra de la complexité du concept et de la manière dont il est encodé dans les couches cachées. Ainsi, pour chaque concept c_j et chaque représentation latente $z^{(l)}$, on apprend un interpréteur $h_{j,l}$ différent qui retourne une probabilité de prédiction du concept $\hat{c}_j$. Pour entraîner les interpréteurs, on dispose d'une base de concepts provenant d'une base de données, utilisée ou pas dans l'apprentissage du modèle prédictif. On utilise les métriques usuelles d'évaluation des performances en classification (taux d'erreur, F1-score, AUC...) pour déterminer la présence des concepts dans les différentes couches cachées ainsi que le type d'interpréteur le plus performant. Il est possible que les concepts ne soient ni capturés sur la même couche cachée, ni par le même type d'interpréteur. À la fin de cette étape, on dispose d'une liste de taille K de concepts capturés avec $K <= C$.

Étape 2 : On examine ensuite les relations logiques entre les concepts capturés et les prédictions du modèle prédictif. Pour cela, nous nous inspirons des règles d'association afin de construire un ensemble de règles logiques que les cardiologues pourront facilement utiliser. Pour ce faire, étant donné un concept c_j, on peut étudier la répartition dans l'espace des paires de point $(\hat{y}, \hat{c}_{j,x})_{1,\cdots,N}$ (avec N le nombre d'exemples) et calculer différentes mesures d'intérêt utilisées dans les règles d'association telles que la confiance et le support (Agrawal et al., 1994). Le support est la proportion d'exemples dans l'ensemble de données contenant une occurrence particulière, telle que le concept c_j, qu'on pourra exprimer par la probabilité $p(c_j)$. La confiance, quant à elle, se définit comme la probabilité conditionnelle $p(y|c_j)$. On compare généralement ces mesures à un seuil de décision défini par l'utilisateur pour valider ou invalider les règles. Seules les règles respectant ce seuil sont retenues.

2.2 Protocole expérimental

Jeu de données réel Les données proviennent de la cohorte Generepol (NCT00773201) généré par le centre de recherche d'investigation de la Pitié-Salpétrière à Paris (Salem et al., 2017). Les ECGs de 990 sujets sains ont été enregistrés avant et 1, 2, et 3 h après la prise orale d'une dose de 80-mg de Sotalol (connu pour présenter un risque d'induire la TdP). Dans le cadre de ce travail, nous nous sommes concentrés sur la dérivation II, offrant généralement à elle seule un bon aperçu du signal électrique pour la caractérisation du QT-long (Prifti et al.,

2021). Les ECGs sont enregistrés sur une fenêtre de 10 secondes avec une fréquence d'échantillonnage de 500Hz. Après pré-traitement et normalisation, nous disposions de 10292 ECGs répartis en deux classes : Sot- (avant la prise du médicament Sotalol) et Sot+ (après la prise du médicament). Le jeu est découpé de telle sorte que 10% du jeu est réservé pour une évaluation indépendante (*holdout*) et les 90% restants pour l'entraînement général (avec une répartition en 75% apprentissage, 10% validation et 15% test).

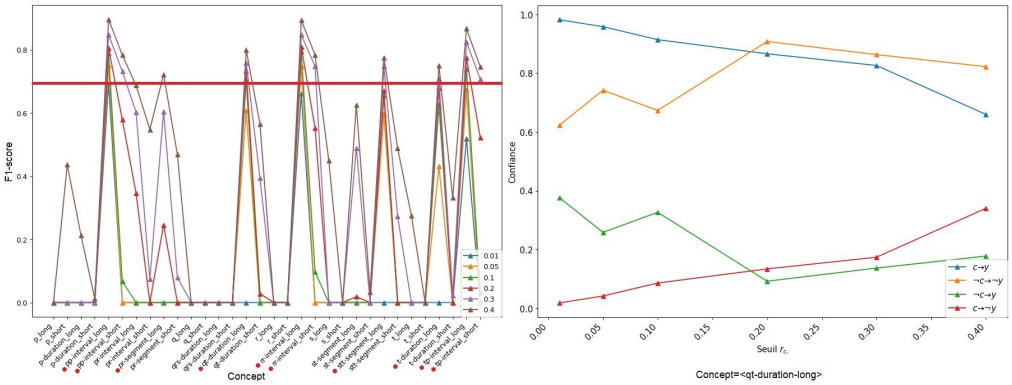
Modèle prédictif Le modèle utilisé (Prifti et al., 2021) est un réseau de neurones convolutif densément connecté (DenseNet) qui prend en entrée un électrocardiogramme de patient de 10 secondes sur 5000 points et retourne une prédiction Sot+ ($\hat{y} > 0.5$) ou Sot- ($\hat{y} \leq 0.5$). Ce réseau est formé de blocs convolutifs denses (DenseBlock) reliés par des transitions. Chaque bloc contient plusieurs couches de convolution qui sont toutes connectées directement les unes aux autres. Le dernier bloc du modèle correspond à un perceptron multicouche totalement connecté. Ce modèle présente un taux d'erreur de 2% sur le jeu *holdout*. Les performances complètes du modèle ainsi que son architecture sont décrites et discutées en détail dans l'article d'origine (Prifti et al., 2021).

Acquisition des concepts La banque de concepts provient du même jeu de données, mais pourrait venir d'une source extérieure. Les concepts c_j sont établis à partir des positions des ondes et des pics sur l'ensemble du jeu de données. À partir de celles-ci, les amplitudes, ainsi que la durée d'intervalles entre deux positions intra ou inter-battement, sont mesurées pour former une base de seize concepts. Pour un ECG et un concept donnés, le concept est mesuré sur tous les battements de l'ECG et on en calcule la médiane pour n'avoir qu'une seule mesure représentative du concept. Pour chaque concept, on trace ensuite la distribution des médianes obtenues sur l'ensemble des ECGs et on seuille pour déterminer les sous-concepts <long> (vs <normal>) ou <court> (vs <normal>). Les sous-concepts sont binarisés de sorte que la valeur 1 encode le sous-concept <long> (resp. <court>) et 0 <normal>. La valeur du seuil r_{c_j} est déterminée à partir de la distribution du concept issu des ECGs de classe Sot- (groupe témoin). Ce seuil peut être fixé a priori ou être considéré comme un hyper-paramètre dont l'impact sur les résultats sera étudié. Chaque ECG du jeu de données Generepol (Sot- ou Sot+) prend donc deux valeurs de sous-concept par concept en fonction du positionnement de la valeur médiane du concept vis-à-vis du seuil choisi. Nous avons ainsi constitué une base de 32 concepts en incluant les sous-concepts. Nous recherchons la présence de ces concepts dans les couches cachées de la partie prédictive du modèle (le perceptron multicouche), après les blocs denses de convolution qui sont en charge d'extraire des motifs abstraits des données.

Interpréteur Rappelons que pour chaque concept étudié c_j, l'interpréteur $h_{j,l}$ prend en entrée la représentation latente choisie $z^{(l)}$ d'un ECG. Ici, chaque interpréteur est un réseau de neurones à une seule couche cachée dont le nombre de neurones (n_{hidden}) dépend du nombre de variables d'entrée (n_{in}) de $z^{(l)}$ selon la règle suivante : $n_{hidden} = \frac{n_{in}}{4}$. Pour apprendre les interpréteurs, nous avons utilisé le même jeu d'entraînement que celui utilisé par le modèle prédictif, sachant que le nombre d'exemples peut légèrement différer dans chacun des sets du fait de l'existence de valeurs manquantes ou aberrantes dans les annotations. D'autres types d'interpréteurs peuvent être considérés. Néanmoins, les premiers tests ont montré qu'un mo-

dèle linéaire retrouvait les concepts bien moins efficacement qu'un modèle non-linéaire (MLP ou SVM).

3 Résultats



(a) (b)

Concept	Règles
pp-interval_long	$c \rightarrow y$ (0,1)
pp-interval_short	$c \rightarrow \neg y$
rr-interval_long	$c \rightarrow y$
rr-interval_short	$c \rightarrow \neg y$ (0,1)
stt-segment_long	$c \rightarrow y$

Concept	Règles
qt-duration_long	$c \rightarrow y$ $y \rightarrow c$
tp-interval_long	$c \rightarrow y$
tp-interval_short	$c \rightarrow \neg y$
t-duration_long	$c \rightarrow y$

(c)

FIG. 2 – (a) Évaluation des performances des interpréteurs de type réseau de neurones sur le jeu de test selon le F1-score en fonction de l'ensemble des 32 concepts et des seuils r_{c_j} (b) Évaluation des différentes relations logiques entre le concept c = <qt-duration-long> et la prédiction y = Sot+ selon le score de confiance en fonction du seuil r_c. $\neg c$ représente le concept <normal> et $\neg y$ représente la classe Sot-. (c) Ensemble des relations logiques extraites avec $r_{c_j} = 0.2$ dans la majorité des cas, sauf à 0.1 pour les cas précisés entre parenthèses.

Concernant les résultats de l'étape 1, rappelons qu'en faisant varier à la fois le concept c_j, la couche (l) et le seuil r_{c_j}, cela emmène à réapprendre systématiquement un nouvel interpréteur. Nous exposons en Fig. 2 les résultats obtenus avec l correspondant à la couche d'entrée du perceptron multicouche du modèle prédictif. En effet, des tests préliminaires ont montré que cette couche permettait de mieux prédire les concepts contrairement aux couches plus profondes du perceptron multicouche. La Fig. 2a présente les performances des interpréteurs de type réseau de neurones pour chaque concept en faisant varier également le seuil r_{c_j}. On remarque que les concepts entre deux battements <pp-interval-{long,short}, rr-interval-{long,short}, tp-interval-{long,short}> et intra-battement <qt-duration-long, stt-segment-long, t-duration-long> sont capturés avec un F1-score supérieur à 0.7. Aucun concept sur les ampli-

tudes semble être présent. Les performances varient également avec le seuil r_{c_j}. On peut par exemple noter une différence de performances de 0.2 sur le concept <rr-interval-long> entre deux valeurs de seuil différent (0.4 et 0.01). Concernant les résultats de l'étape 2, l'objectif est d'établir les relations logiques sur la liste K de concepts retenus à l'étape précédente (F1-score > 0.7). Les résultats préliminaires sur le concept $j =$ <qt-duration-long> sont présentés dans la Fig.2b en utilisant la confiance comme métrique. Cela permet d'avoir un premier aperçu des implications logiques intéressantes. Cette figure illustre l'évolution du score de confiance en fonction du seuil r_c de la règle inspectée. Ce score, variant entre 0 et 1, indique la force de l'association, où un score plus élevé signifie une association plus forte. On observe que les règles logiques $(c \rightarrow y)$ et $(\neg c \rightarrow \neg y)$ se démarquent avec un score supérieur à 0.6. Deux tendances émergent également : avec un seuil plus élevé, le score de confiance tend à augmenter pour les règles $\{(\neg c \rightarrow \neg y), (c \rightarrow \neg y)\}$, tandis qu'il a tendance à diminuer pour les règles $\{(c \rightarrow y), (\neg c \rightarrow y)\}$. Il semble y avoir un compromis à trouver dans le choix du seuil. Un seuil élevé signifie un équilibre entre les données d'entraînement présentant et ne présentant pas le concept. En revanche, un seuil bas signifie que peu de données le présentent. Dans le cas exposé, le seuil 0.2 semble être le plus approprié, où les règles $(c \rightarrow y)$ et $(\neg c \rightarrow \neg y)$ ont toutes deux un score de confiance entre 0.8 et 0.9. Cette analyse peut être répétée pour tous les concepts retenus à l'étape précédente, permettant ainsi de définir un ensemble de règles logiques pour fournir une interprétation globale du modèle. Une première estimation des règles est présentée dans le tableau 2c.

4 Conclusion et perspectives

Ces premières expériences ont démontré que notre approche est efficace pour détecter la présence de concepts dans les couches cachées du modèle prédictif et établir des relations logiques entre les concepts et les prédictions. Sur le plan méthodologique, des expériences approfondies sont nécessaires pour valider l'approche, notamment en explorant différents types d'interpréteurs non-linéaires tels qu'un SVM. Ensuite, il existe plus d'une vingtaine de mesures dans la littérature (Lenca et al., 2007) pour évaluer les règles d'association, comme l'indépendance et l'absence de contre-exemples. Celles-ci pourraient être utilisées pour choisir les relations logiques les plus adaptées. L'utilisation d'algorithmes d'extraction d'items-sets fréquents [2], tels qu'Apriori (Agrawal et al., 1994), permettrait également de déduire des règles de combinaison de concepts. La validation de l'approche sur d'autres ensembles de données et types de réseaux neuronaux est aussi essentielle. Enfin, il est important d'évaluer et de valider l'interprétation construite de manière qualitative et quantitative par des critères spécifiques (Islam et al., 2020), mais aussi auprès des cardiologues pour vérifier l'alignement des règles logiques avec les connaissances du domaine. Notons déjà que la présence du concept <qt-duration-long> dans les règles est cohérente avec les connaissances du domaine.

Financement

Cette étude a été soutenue par le financement ANR-20-CE17-0022 DeepECG4U de l'Agence Nationale de la Recherche.

2. Par définition, un item-set est un ensemble d'items correspondant dans notre cas aux concepts.

Références

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, Volume 1215.

Crabbé, J. et M. van der Schaar (2022). Concept activation regions : A generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 2590–2607.

Goodman, B. et S. Flaxman (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine 38*(3), 50–57.

Islam, S. R., W. Eberle, et S. K. Ghafoor (2020). Towards quantification of explainability in explainable artificial intelligence methods. In *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference*, pp. 75–81. AAAI Press.

Kim, B., M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, et R. Sayres (2018). Interpretability beyond feature attribution : Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, pp. 2673–2682. PMLR.

Lenca, P., B. Vaillant, P. Meyer, et S. Lallich (2007). Association rule interestingness measures : Experimental and theoretical studies. In *Quality Measures in Data Mining*, Studies in Computational Intelligence, pp. 51–76. Springer.

Prifti, E., A. Fall, G. Davogustto, A. Pulini, I. Denjoy, C. Funck-Brentano, Y. Khan, A. Durand-Salmon, F. Badilini, Q. S. Wells, A. Leenhardt, J.-D. Zucker, D. M. Roden, F. Extramiana, et J.-E. Salem (2021). Deep learning analysis of electrocardiogram for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome. *European Heart Journal 42*(38), 3948–3961.

Salem, J.-E., M. Germain, J.-S. Hulot, P. Voiriot, B. Lebourgeois, J. Waldura, D.-A. Tregouet, B. Charbit, et C. Funck-Brentano (2017). Genome wide analysis of sotalol-induced ikr inhibition during ventricular repolarization,"generepol study" : Lack of common variants with large effect sizes. *PLoS One 12*(8), e0181875.

Zaeem, M. N. et M. Komeili (2021). Cause and effect : Concept-based explanation of neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2730–2736.

Summary

The development of artificial intelligence contributes to the emergence of a new form of personalized medicine, aiming to better consider patients' characteristics. In this context, we focus on the application of deep learning to detect cardiovascular diseases, such as Long QT syndrome, from electrocardiograms. However, a major concern lies in the often opaque nature of these so-called "black box" AI models. This article aims to make these models more interpretable by integrating medical concepts into a new post-hoc approach, ConceptECGxAI, providing more understandable explanations to physicians. Experiments have shown that the approach can identify the presence of meaningful concepts in the hidden layers of the predictive model and establish logical relationships between these concepts and predictions.

Espace latent pour l'explicabilité en sous-titrage d'images

Sofiane Elguendouze*, Adel Hafiane**
Marcilio C.P. de Souto*, Anaïs Halftermeyer*

*LIFO, 45000 Orléans, France
prénom.nom@univ-orleans.fr
**PRISME, 18022 Bourges, France
prénom.nom@insa-cvl.fr

Résumé. Cet article se concentre sur l'espace latent dans les architectures de sous-titrage d'images pour développer des méthodes d'explication avec différentes portées - une méthode basée sur la substitution avec des explications locales (LIME) - et - une méthode basée sur la rétropropagation de la pertinence (LRP) avec une portée globale. Pour évaluer la qualité des explications obtenues, nous proposons le concept de masquage latent (Latent Ablation), qui opère dans l'espace latent, permettant d'éviter les incohérences et les informations tronquées que l'on trouve dans le masquage classique. Nos expérimentations montrent que les deux méthodes donnent des résultats comparables. La portée de la méthode d'explication s'est avérée moins décisive dans la quête d'une qualité d'explication supérieure, mais joue plutôt un rôle dans la détermination de la granularité/subtilité des explications produites. L'article est un résumé de (Elguendouze et al., 2023).

1 Introduction

Le sous-titrage d'images est l'une des tâches vision-langage visant à générer des descriptions textuelles pour les images. La plupart des modèles sont conçus dans le cadre Encodeur-Décodeur, généralement composé d'un réseau neuronal convolutif (CNN) comme encodeur et d'un autre récurrent (RNN) comme décodeur. Les caractéristiques visuelles sont extraites par l'encodeur, puis traduites par le décodeur en légendes textuelles. Un mécanisme d'attention, dont le rôle est de guider le modèle pour se concentrer sur les informations pertinentes de l'image lors du décodage, est souvent inséré entre les deux composants.

Malgré leurs performances élevées, le fonctionnement interne de ces architectures, principalement basées sur des réseaux de neurones profonds, fait qu'elles sont considérées comme des boîtes noires, ce qui rend le processus de décision difficile à comprendre. Très peu de travaux ont abordé la question de l'explicabilité en sous-titrage d'images dans la littérature, la plupart d'entre eux cherchant à établir une relation de causalité perceptive entre la sortie et l'entrée. La question de savoir si le choix d'une approche d'explication, basé sur des critères tels que sa portée [1], peut influencer significativement la qualité des explications, n'a pas

1. Mesure dans laquelle la méthode d'explication explore et saisit les facteurs sous-jacents, englobant des aspects tels que la couverture des éléments par le processus d'explication (instances locales ou comportement global, etc.).

fait l'objet de recherches approfondies jusqu'à présent. Malgré la diversité des techniques de l'explicabilité, de simples substitutions comme LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) à des approches plus complexes et plus exigeantes sur le plan computationnel basées sur la rétropropagation comme LRP (Layer-wise Relevance Propagation) (Bach et al., 2015), l'usage de ces dernières ne garantit pas systématiquement des explications de meilleure qualité. Dans le contexte global de l'explicabilité, il est crucial de considérer le compromis entre le coût de calcul des explications et leur qualité. Les méthodes chronophages exigent souvent des ressources computationnelles importantes, limitant leur applicabilité en temps réel ou sur de vastes ensembles de données. Des études comparatives sont nécessaires pour comprendre la relation entre la complexité des méthodes d'explication et leur contribution à la compréhension des modèles de sous-titrage.

2 Littérature

La prise de conscience de la nécessité de développer des modèles de sous-titrage plus précis a conduit à des progrès exponentiels en termes de complexité d'architectures. Toutefois, les travaux sur l'explicabilité du sous-titrage d'images demeurent limités, la plupart d'entre eux adhérant au paradigme post-hoc. Les auteurs dans (Sun et al., 2022) ont proposé plusieurs méthodes d'explication incluant une version adaptée de LRP. Celle-ci génère des explications visuelles qui mettent en évidence les pixels soutenant ou s'opposant à la prédiction d'un mot cible dans la légende. Dans (Sahay et al., 2021), les auteurs ont utilisé LIME pour approximer le modèle de sous-titrage par un modèle linéaire plus simple pour chaque instance de données. L'intuition est de générer un ensemble d'instances perturbées localement autour d'une image en utilisant des opérations (comme le floutage) sur l'entrée originale (pixels), puis d'évaluer le changement en prédiction. Une fois le modèle entraîné, il attribue un poids à chaque région de l'entrée en fonction de sa capacité à préserver ou non la prédiction d'un mot de la légende.

Se fondant sur les atouts distinctifs constatés dans l'espace latent, nous avons conçu deux méthodes d'explication, BU-LIME et BU-LRP, ainsi qu'une nouvelle approche d'évaluation, appellée Masquage Latent. L'objectif est d'étudier la causalité possible entre la qualité des explications et la portée/fonctionnement des méthodes explicatives. En fonction de la granularité des concepts manipulés dans l'espace latent (partiels/complets), nous expérimentons deux versions de BU-LIME et du masquage latent. L'objectif est d'étudier l'impact de cette granularité sur la conception et l'évaluation des méthodes d'explication. Alors que la majorité des travaux existants utilisent des CNNs simples pour l'encodage des images, nous nous intéressons ici à des caractéristiques de type de bas en haut, qui sont plus répandues dans ce domaine. Cela se traduit par la manipulation d'éléments de bas niveau considérées comme très proches de la théorie interne du modèle afin de caractériser son fonctionnement.

3 Méthode

3.1 Architecture standard de sous-titrage

Nous utilisons l'architecture standard Ada-LSTM de (Sun et al., 2022) qui repose sur la la topologie Encodeur-Attention-Décodeur (figure 1). Elle utilise des caractéristiques d'image de

bas en haut, générées par un module de détection d'objets Faster-RCNN. Bien qu'il existe des architectures plus sophistiquées telles que celles basées sur les transformers, elles pourraient être moins adaptées à notre étude en raison de leur complexité trop élevée par rapport à leurs performances ordinaires.

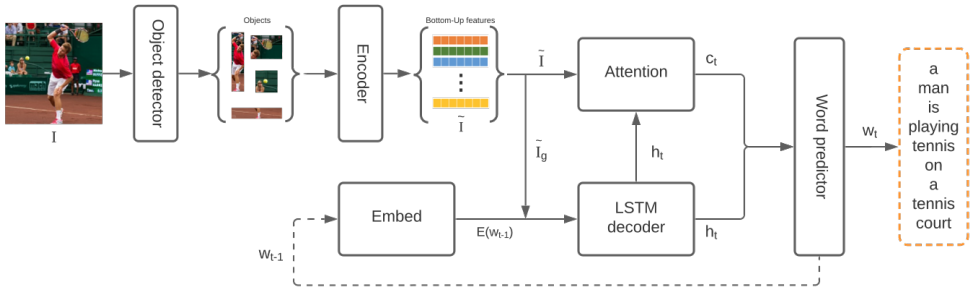


FIG. 1 – Architecture de sous-titrage de type bas en haut.

Étant donné une paire alignée (I, S) comprenant une image I et des légendes de référence S, l'encodeur d'images extrait un ensemble $\tilde{I}$ de V caractéristiques visuelles sous forme de vecteurs latents $(v_i)_{i=1}^{V}$, de dimension d_v chacun. La moyenne des caractéristiques visuelles est ensuite calculée pour obtenir une caractéristique globale $\tilde{I}g$. Celle-ci est combinée avec l'état caché précédent h_{t-1} et le plongement du mot précédent dans la séquence $E(w_{t-1}) \in \mathbb{R}^{d_w}$. Les informations résultantes sont ensuite transmises au décodeur LSTM pour générer l'état caché courant h_t. Ici, d_w et d_h représentent respectivement les dimensions des vecteurs de plongements de mots et d'états cachés. Le module d'attention attribue des coefficients de focalisation à chaque caractéristique visuelle, générant une représentation du contexte c_t de dimension d_c. À chaque étape t, les états cachés et les vecteurs de contexte sont utilisés par le prédicteur de mots (de type LSTM) pour générer le mot suivant w_t dans la séquence de sortie C, L étant la longueur maximale de la légende.

3.2 Méthode d'explication basée sur LRP

Le concept de LRP est similaire à celui de la rétropropagation. La méthode redistribue la prédiction finale (sortie) le long d'un réseau de neurones en attribuant récursivement un score de pertinence à chaque neurone, jusqu'à ce que l'entrée soit atteinte. Cela peut être réalisé selon plusieurs règles d'affectation définies par (Bach et al., 2015). Nous adaptons LRP pour correspondre à l'architecture de sous-titrage de type bas en haut. Nous initialisons les scores de pertinence des mots avec les logits de la dernière couche du prédicteur de mots. Le score de pertinence de chaque mot $R(w_t)$ est rétro-propagé le long de l'architecture. Nous nous arrêtons à la sortie de l'encodeur pour obtenir les pertinences des caractéristiques visuelles (figure 2).

Les scores de pertinence $R(c_t)$, $R(h_t)$, $R(\tilde{I}_g)$, $R(\tilde{I})$ désignent respectivement l'importance des vecteurs de contexte, des états cachés, des caractéristiques d'image globales et des caractéristiques d'image. Nous sommes uniquement intéressés par l'exploration de la pertinence des caractéristiques visuelles pour la prédiction du mot final. Le score global de pertinence pour une caractéristique visuelle est alors la somme/moyenne de tous les éléments $R(v_{ij})$ du vecteur

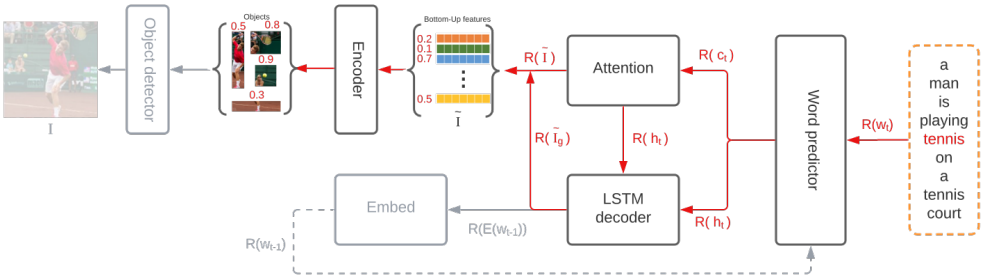


FIG. 2 – Flux de rétropropagation de BU-LRP.

de pertinence. Nous obtenons un vecteur de coefficient $\hat{\alpha} \in \mathbb{R}^V$ comme résultat final d'explication, représentant la contribution de toutes les caractéristiques visuelles à la prédiction d'un mot donné dans la légende.

3.3 Méthode d'explication basée sur LIME

LIME est une méthode perturbative qui peut être utilisée pour expliquer n'importe quel modèle boîte noire. Elle fonctionne en perturbant l'entrée originale pour générer des instances voisines, qui sont utilisées pour prédire de nouvelles sorties et évaluer le changement par rapport à la sortie originale. Cette évaluation est réalisée par l'apprentissage d'un modèle linéaire dont les entrées et les sorties sont les instances perturbées et leurs sorties correspondantes. Les coefficients du modèle linéaire fournissent des explications locales sous la forme d'attributions d'importance à chaque caractéristique d'entrée pour prédire une sortie spécifique.

Nous générons P instances de voisinage $\Gamma = \{\acute{\tilde{I}}^{(p)}\}_{p=1}^{P}$ autour d'une image donnée I en appliquant des perturbations Gaussiennes intrinsèques graduelles à un sous-ensemble de ses caractéristiques visuelles.Un vecteur binaire $X^{(p)} \in \{0,1\}^V$ dont les éléments sont fixés à '1' pour les caractéristiques perturbées, '0' ailleurs, est associé à chaque instance perturbée $\acute{\tilde{I}}^{(p)}$. Les instances perturbées sont ensuite introduites dans le modèle de sous-titrage. L'ensemble des vecteurs binaires $X = \{X^{(p)}\}_{p=1}^{P}$ est désigné par la matrice binaire $X \in \{0,1\}^{P \times V}$. Nous nous intéressons aux poids maximaux générés de tous les mots appartenant au vocabulaire lors de la prédiction de la légende (logits de la dernière couche du LSTM). À l'issue de cette étape, nous obtenons une matrice de poids $\Delta = (\delta_{pq}) \in \mathbb{R}^{P \times Q}$, Q étant la taille du vocabulaire. Le modèle de régression linéaire $Y = X \cdot \beta + \gamma$ est entraîné en utilisant les données appariées (matrice binaire X, vecteur de poids y) comme ensemble d'apprentissage, $y = \delta_{.q} \in \mathbb{R}^P$ étant la q^{th} colonne de Δ. Une instance d'apprentissage est représentée par une paire (X_p, y_p). $\beta \in \mathbb{R}^V$ est le vecteur des coefficients à estimer et représente les explications générées pour le mot w_q du vocabulaire. L'approche est illustré dans la figure 3.

Sachant que les régions/objets d'une image sont souvent encodés par plusieurs caractéristiques visuelles impliquant une redondance/ambiguïté d'information, nous proposons d'étudier son effet par une deuxième version de LIME (BU-LIME-*N*-OBJ) impliquant la perturbation d'objets complets (l'ensemble des caractéristiques qui leur correspondent) plutôt que des caractéristiques individuelles. Un maximum de *N* objets sont perturbés simultanément.

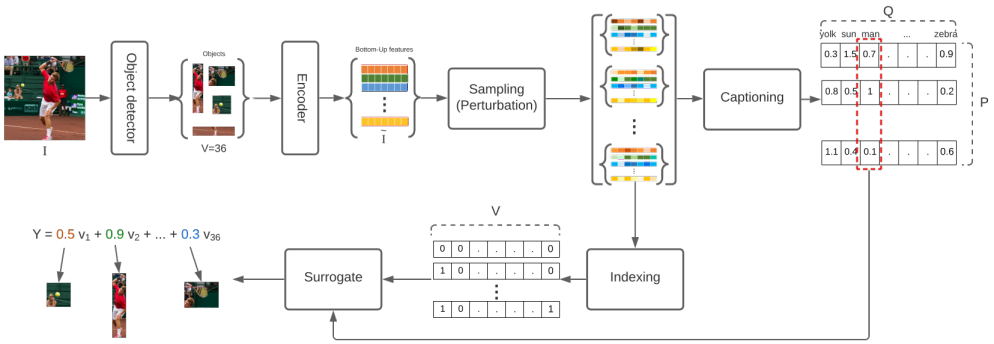


FIG. 3 – Aperçu de la méthode BU-LIME

3.4 Masquage latent

Les travaux existants sur l'évaluation de l'explicabilité du sous-titrage d'images sont principalement basés sur une évaluation qualitative (Xu et al., 2015; Han et Choi, 2018) et/ou quantitative (Sun et al., 2022) de la propriété de fidélité des explications générées par les techniques d'explication perceptuelles (cartes à chaleur etc.), qui comprend soit une évaluation visuelle de l'exactitude de l'explication, soit la mesure de l'écart entre une nouvelle prédiction (en occultant l'élément d'explication de l'entrée), et la prédiction originale.

Nous proposons une méthode d'évaluation basée sur le principe de masquage latent en deux versions : masquage des k caractéristiques visuelles ou des k objets entiers, les plus importants. Cela permet de déterminer le rôle des objets dans l'évaluation des explications. Ce masquage latent est effectué en utilisant diverses magnitudes telles que des perturbations Gaussiennes ou une saturation avec les valeurs minimales/maximales globales sur les vecteurs de caractéristiques visuelles. Nous régénérons ensuite la légende et évaluons la présence du mot à expliquer. L'évaluation ne prend en compte que les mots représentant des objets, et renvoie le pourcentage de mots manquants comme mesure de fidélité de l'explication.

4 Expérimentations et résultats

Nous expérimentons avec l'ensemble de données MSCOCO2017. La figure 4 montre un exemple d'une bonne explication avec BU-LRP pour le mot "girafe" dans la légende d'une image représentant un troupeau de girafes. La figure 4a représente l'importance de chaque élément du vecteur de la caractéristique visuelle v_0, ainsi que l'importance globale (moyenne) de la caractéristique en haut de la carte à chaleur. Nous obtenons 36 vecteurs d'importance pour chaque mot de la légende affichés sous forme de matrices de tailles identiques $64 * 32$, chacun correspondant à une caractéristique visuelle (figure 4b).

Les valeurs d'explication du même vecteur/caractéristique semblent homogènes, à l'exception de quelques valeurs aberrantes suggérant une contribution disproportionnée de certaines dimensions. Dans la figure 4b, certaines caractéristiques ont un impact plus important sur la prédiction du mot "girafe" (une carte à chaleur de couleur plus chaude) : 9^{eme}, 10^{eme} etc. qui

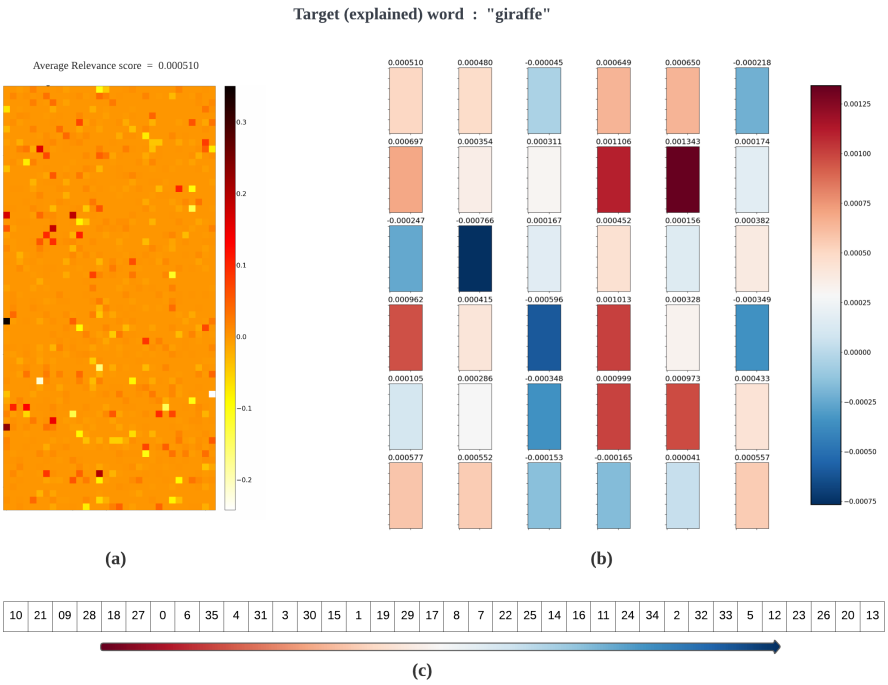


FIG. 4 – Explications BU-LRP. Couleurs plus chaudes → plus grande importance et vice versa.

correspondent respectivement aux objets/régions avec les étiquettes "girafe", "toit", etc.). Les caractéristiques classées par score de pertinence sont présentées dans la figure 4c.

Le tableau 1a montre les résultats de masquage des **caractéristiques individuelles**. La magnitude de masquage "normal" correspond à la perturbation des caractéristiques par l'ajout de valeurs Gaussiennes avec une variance $std = 1,5$, les amplitudes "min" et "max" correspondent à la substitution des valeurs des caractéristiques par les valeurs minimales et maximales de la dimension vectorielle concernée. Nous incluons également une ligne de référence de masquage aléatoire (explications aléatoires) à des fins de comparaison. Le tableau 1b montre les résultats du masquage latent pour la version **objet** rapportés en termes de pourcentage de mots manquants et de deux mesures supplémentaires : la baisse de probabilité moyenne, qui exprime l'ampleur de la baisse de la probabilité de prédiction d'un mot donné dans la nouvelle légende (après masquage) par rapport à celle d'avant, et sa fréquence. Notons que par souci de concision, seule le masquage avec la valeur "min" a été rapportée dans le tableau 1b.

Selon le tableau 1a, les approches d'explication ne semblent pas présenter d'avantages explicites par rapport à la baseline aléatoire, BU-LRP étant légèrement plus performante que les approches basées sur LIME. Ces résultats peuvent être attribués à l'artefact de la manipulation de caractéristiques visuelles individuelles qui représentent souvent un objet partiel plutôt qu'un concept entier. En effet, les informations retirées peuvent être récupérées en utilisant le reste des caractéristiques correspondant au même objets. Le masquage latent d'objets entiers prend en compte cette préoccupation, montrant une augmentation du "pourcentage de mots man-

quants" et de "la baisse de probabilité moyenne" pour les différentes approches d'explication, avec une différence notable entre les méthodes proposées et la baseline aléatoire.

k	Méthode	Magnitude de masquage		
		normal	min	max
1	Random	0.1159	0.2823	0.2735
	BU-LRP	**0.1211**	**0.2840**	**0.2756**
	BU-LIME-1-2	0.1057	0.2797	0.2619
	BU-LIME-5	0.1089	0.2807	0.2648
3	Random	0.1403	0.5116	0.6094
	BU-LRP	**0.1445**	0.5170	**0.6125**
	BU-LIME-1-2	0.1160	**0.5189**	0.6015
	BU-LIME-5	0.1224	0.5154	0.5997
6	Random	0.1779	0.6881	0.8452
	BU-LRP	**0.1948**	0.6892	**0.8456**
	BU-LIME-1-2	0.1616	**0.6944**	0.8364
	BU-LIME-5	0.1559	0.6847	0.8340
9	Random	0.2338	0.7700	**0.9019**
	BU-LRP	**0.2418**	**0.7744**	0.9012
	BU-LIME-1-2	0.2029	0.7725	0.8945
	BU-LIME-5	0.1985	0.7703	0.8910

(a) Masquage latent de caractéristiques individuelles.

k	Méthode	Baisse Prob Moy	Freq Baisse Prob	% mots manquants
1	Random	1.3521	0.9028	0.4460
	BU-LRP	2.1259	**0.9311**	0.5931
	BU-LIME-1-2	**2.1548**	0.9242	0.5922
	BU-LIME-5	2.1322	0.9224	**0.5960**
	BU-LIME-5-OBJ	1.3781	0.9011	0.4568
	BU-LIME-8-OBJ	1.1813	0.8908	0.4191
3	Random	2.8782	0.9507	0.7309
	BU-LRP	3.4346	0.9595	0.8033
	BU-LIME-1-2	**3.4776**	**0.9600**	0.8068
	BU-LIME-5	3.4396	0.9589	**0.8091**
	BU-LIME-5-OBJ	2.4333	0.9417	0.6598
	BU-LIME-8-OBJ	2.1437	0.9374	0.6284
5	Random	3.4810	0.9332	0.8061
	BU-LRP	3.8915	0.9623	0.8617
	BU-LIME-1-2	**3.9280**	**0.9640**	**0.8677**
	BU-LIME-5	3.9019	0.963 3	0.8654
	BU-LIME-5-OBJ	2.8814	0.9486	0.7414
	BU-LIME-8-OBJ	2.5787	0.9458	0.7097

(b) Masquage latent d'objets.

TAB. 1 – Scores de cohérence des explications.

Le modèle BU-LIME-N a obtenu les meilleurs résultats dans la plupart des cas, tandis que les modèles BU-LIME-N-OBJ basés sur la perturbation de l'objet montrent des performances inférieures aux attentes. Cela signifie que la manipulation d'objets complets plutôt que de caractéristiques visuelles individuelles est plus utile pour l'évaluation basée sur le masquage latent que pour l'utilisation intrinsèque dans la construction de modèles d'explication. La manipulation d'objets entiers lors de la création de modèles d'explication linéaires est potentiellement à l'origine de pertes d'informations conduisant à une incohérence dans les poids du modèle pendant la phase d'entraînement. Ainsi, elle pourrait plutôt être utilisée pour des évaluations post-hoc telles que le masquage latent, sans pour autant être utilisée de manière intrinsèque lors de la conception d'approches explicatives.

Dans l'ensemble, les explications basées sur LRP et LIME présentent une qualité satisfaisante par rapport à la seule référence aléatoire existante. Les deux techniques semblent être très similaires en termes de justesse de leurs explications. Leur portée ne semble pas avoir un impact significatif sur la qualité des explications. Toutefois, cela pourrait dépendre de la finesse de l'explication recherchée. Ces résultats pourraient conduire à mettre davantage l'accent sur les méthodes locales, étant considérées comme plus réalisables dans la pratique que les méthodes globales, qui exigent des approches plus complexes et coûteuses en temps.

5 Conclusion

Nous avons conçu deux méthodes d'explication pour le sous-titrage d'images distinctes dans leur portée/fonctionnement, LRP et LIME, et avons réalisé une étude comparative pour évaluer la qualité de leurs explications, constatant des résultats similaires en termes d'exactitude. Nous avons introduit le concept de masquage latent pour évaluer la qualité des explications, offrant une meilleure manipulation des objets et différents niveaux d'altération pour

une évaluation plus précise. Pour explorer l'impact de l'exhaustivité des concepts de l'image sur le développement et l'évaluation de méthodes de l'explicabilité, nous avons conçu deux versions de LIME et de masquage latent. La première version est basée sur les caractéristiques individuelles et la seconde est basée sur les objets. Nos résultats ont montré que la portée de la méthode d'explication n'est pas cruciale pour une meilleure qualité d'explication, et que la manipulation des objets n'améliore pas systématiquement la robustesse du modèle d'explication linéaire, mais est un élément clé dans l'évaluation de la qualité des explications.

Références

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, et W. Samek (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE 10*(7), 1–46.

Elguendouze, S., A. Hafiane, M. C. de Souto, et A. Halftermeyer (2023). Explainability in image captioning based on the latent space. *Neurocomputing 546*, 126319.

Han, S.-H. et H.-J. Choi (2018). Explainable image caption generator using attention and bayesian inference. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 478–481.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). "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*, KDD '16, New York, NY, USA, pp. 1135–1144. Association for Computing Machinery.

Sahay, S., N. Omare, et K. K. Shukla (2021). An approach to identify captioning keywords in an image using lime. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 648–651.

Sun, J., S. Lapuschkin, W. Samek, et A. Binder (2022). Explain and improve : Lrp-inference fine-tuning for image captioning models. *Information Fusion 77*, 233–246.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, et Y. Bengio (2015). Show, attend and tell : Neural image caption generation with visual attention. In F. Bach et D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, Volume 37 of *Proceedings of Machine Learning Research*, Lille, France, pp. 2048–2057.

Summary

We focus on the latent space in image captioning architectures to develop explanation methods with different scopes - a surrogate-based method with local explanations (LIME) - and - a method based on relevance backpropagation (LRP) with a global scope. We propose the new concept of latent ablation to assess the quality of the explanations obtained which, by operating on latent space, avoids inconsistencies and truncated information in conventional ablation. The two methods produce comparable results and the scope of the explanation method proved less decisive in the quest for superior explanation quality, but rather plays a role in determining the granularity/subtlety of the explanations produced. The article is a summary of (Elguendouze et al., 2023).

Explication basée concepts pour les modèles de langage - Application à la classification de notes cliniques

Charlotte Claye[*,**], Dylan Vellas [***] Yves Allenbach [***], Florence Tubach [***], Raphaèle Seror [****], Julien Duquesne[*,♯], Céline Hudelot [**,♯], Wassila Ouerdane[**,♯],

[*]Scienta Lab
prenom.nom@scientalab.com,
[**]Université Paris-Saclay, CentraleSupélec, MICS
prenom.nom@centralesupelec.fr
[***]Hôpital Pitié-Salpêtrière
[****]Hôpital Bicêtre
[♯]Co-derniers auteurs

Résumé. L'utilisation de modèles de langage dans le domaine de la santé offre de vastes possibilités, notamment dans le cadre de la recherche médicale. Dans ce contexte, de nombreuses initiatives de recherche ont été mises en place pour concevoir des modèles de langage spécifiques au domaine médical, comme par exemple le modèle DrBert pour la langue française. Cependant, une des principales limitations à l'utilisation de ces modèles est l'opacité de leur fonctionnement. Nous présentons dans ce papier un cadre modulaire d'explicabilité post-hoc par concepts appris sans supervision. Nous appliquons ensuite ce cadre à une tâche de classification de notes cliniques.

1 Introduction

Les modèles de langage sont de plus en plus performants et offrent de nombreuses opportunités, notamment en médecine pour des applications cliniques, de recherche ou d'éducation (Thirunavukarasu et al. (2023)). Cependant, les gains en performance s'accompagnent d'une augmentation de la complexité des modèles. Ainsi, une de leurs limitations connues est qu'ils sont opaques dans leur fonctionnement (Ray (2023)), ce qui est un frein important à leur utilisation en routines cliniques. Améliorer l'explicabilité de ces modèles est nécessaire pour assurer leur utilisation en pratique et tirer parti de leur potentiel (Imrie et al. (2023)).

Dans nos travaux, nous nous positionnons dans le champ des approches d'explicabilité dîtes post-hoc, c'est-à-dire visant à expliquer un modèle après son entraînement. C'est un champ de recherche très actif pour les modèles de langage et de nombreuses méthodes voient le jour (Zhao et al. (2023a)). Parmi ces approches, nous distinguons cinq grandes familles. **Les méthodes d'attribution** (Atanasova et al. (2020)) associent un poids à chaque élément d'entrée en fonction de son importance pour la prédiction, en se basant sur des perturbations, sur les gradients, ou encore sur l'attention. **Les méthodes basées exemples** (Poch'e et al. (2023)) présentent des échantillons issus du jeu de données ou générés pour fournir une explication, il peut s'agir d'exemples similaires, de contre-factuels, de semi-factuels, de prototypes,

d'exemples influents ou encore de visualisation de neurones. **Les méthodes par sondage** (Belinkov et al. (2020)) consistent à entraîner un modèle simple à effectuer une tâche à partir des représentations internes du modèle complexe à expliquer. Un lien est ensuite fait entre les performances de ce modèle et l'information contenue dans les représentations. **Les méthodes basées modèles** consistent à approximer le modèle boite noire par un modèle intrinsèquement interprétable, par exemple avec l'extraction de règles (Ribeiro et al. (2018)). Enfin, **les méthodes basées concepts**, principalement utilisées en vision (Kim et al. (2018), Fel et al. (2023b)), sont également de plus en plus appliquées au texte (Jourdan et al. (2023), Zhao et al. (2023b)). Elles consistent à expliquer un modèle à partir de concepts haut-niveaux qui ont un sens pour l'utilisateur.

Nos travaux se positionnent dans cette dernière famille d'approches. Notre objectif dans ce papier est de présenter un cadre modulaire d'explication par concepts pour les données médicales textuelles. Nous appliquons ce cadre à un cas pratique de classification de texte médical de grande dimension.

Le papier est organisé comme suit. La Section 2 introduit l'explicabilité par concepts ainsi que les limites actuelles de son application aux modèles de langage. Dans la Section 3, nous proposons un cadre pour l'explicabilité par concepts post-hoc non-supervisée. Dans la Section 4, nous présentons un nouveau modèle de classification de notes cliniques que nous utilisons comme cas d'application et présentons les tous premiers résultats obtenus. Enfin, dans la Section 5, nous discutons des prochaines étapes.

2 Explicabilité par concepts

Dans cette section, nous présentons l'explicabilité par concepts en définissant la notion de concept et les différentes approches. Enfin, nous identifions les défis et limites actuelles de son application aux modèles de langage.

2.1 Définition

L'explicabilité par concepts consiste à expliquer le fonctionnement d'un réseau de neurones à travers des concepts haut-niveaux qui ont un sens sémantique pour l'utilisateur et qui sont encodés par le modèle. En vision, Kim et al. (2018) montre comment le concept "rayures" influence la prédiction de la classe "zèbre". Pour des données ARN, Zarlenga et al. (2023) retrouve comme concepts des ensembles de gènes qui codent pour une fonction biologique simple. Pour du texte médical, un concept pourrait être "Atteinte du système respiratoire".

Dans la littérature, l'explicabilité par concepts est accompagnée d'un ensemble de propriétés souhaitées décrites en Figure 1. La signification et la cohérence (Ghorbani et al. (2019)) sont des propriétés sémantiques liées au principe de concept et qui garantissent la compréhension des concepts par l'utilisateur. L'importance (Ghorbani et al. (2019)), la fidélité (Chefer et al. (2023)) et l'exhaustivité (Yeh et al. (2020)) garantissent la validité des concepts vis-à-vis du modèle prédictif. Enfin la concision (Vielhaben et al. (2023)) et la robustesse (Chefer et al. (2023)) sont des propriétés souhaitées plus généralement pour respectivement l'explication et la méthode d'explicabilité.

L'avantage de l'explicabilité par concepts est d'obtenir des explications plus interprétables car plus proches de la façon de raisonner de l'utilisateur, et moins sujettes aux biais de confir-

mation (Fel et al. (2023b)). Il est aussi plus immédiat d'obtenir des explications à la fois locales et globales, contrairement aux méthodes d'attribution locale par exemple.

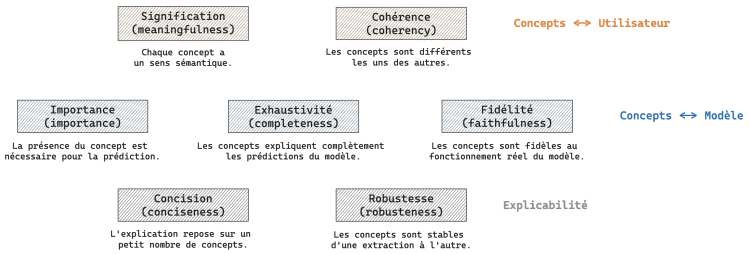


FIG. 1 – *Propriétés souhaitées pour l'explicabilité par concepts.*

2.2 Principales approches

Deux grandes familles d'explicabilité par concepts ont été proposées dans la littérature (Marconato et al. (2023)) : les **modèles basés concepts** (CBMs Concept-based Models) sont des modèles qui extraient des concepts depuis l'entrée puis utilisent ces concepts pour leur prédiction. Les concepts sont appris pendant l'entraînement, de manière supervisée ou non-supervisée. Les **explicateurs basés concepts** (CBEs : Concept-Based Explainers) sont des méthodes post-hoc qui consistent à extraire les concepts d'une couche d'activation d'un modèle déjà entraîné. Cette approche peut également être divisée en deux méthodes (Schwalbe (2022)) : La **localisation** de concepts lorsque l'extraction est faite de manière supervisée et l'**exploration** de concepts lorsque l'extraction est faite de manière non-supervisée. Dans ce papier, nous nous concentrons sur cette dernière approche.

2.3 Application aux modèles de langage

L'explicabilité par concepts a d'abord été utilisée en vision, et plus récemment pour le langage (Jourdan et al. (2023)). Les spécificités du langage impliquent des défis particuliers.

Sémantique Dans le domaine de la vision, la notion de concept est bien définie, il peut s'agir d'un objet, d'une partie ou d'une caractéristique d'un objet, du thème de l'image (Schwalbe (2022)). Cependant, pour le texte, la définition n'est pas encore bien établie. Une piste est de s'intéresser aux concepts utilisés en linguistique : la phonologie (prononciation), la morphologie (structure d'un mot), la syntaxe (combinaison de mots), la sémantique (façon dont le langage transmet du sens) et enfin la pragmatique (style du texte).

Visualisation Un deuxième défi pour l'explicabilité par concepts pour le texte est la visualisation des données. La visualisation de prototypes par exemple est souvent utilisée pour comprendre le sens d'un concept en vision (Fel et al. (2023b)). Toutefois, s'il est relativement facile de visualiser plusieurs images et d'en extraire des caractéristiques communes, il peut être plus fastidieux de lire plusieurs paragraphes de texte et d'identifier les similarités, d'autant

plus lorsque le texte est long. Ainsi, une attention particulière doit être donnée à l'interface homme-machine pour permettre de communiquer efficacement l'information.

Méthodes d'attribution. Les méthodes d'attributions entre le texte d'entrée et un concept sont beaucoup utilisées en complément des prototypes pour comprendre le sens des concepts (Jourdan et al. (2023)). Toutefois, ces méthodes pour les modèles de langage posent encore des questions de validité : les méthodes d'occlusion tendent à générer des données hors-distribution (Zhao et al. (2023a)) et les scores par élément d'entrée ne sont pas forcément valides ou utiles s'ils ne prennent pas en compte l'information redondante entre éléments (Deb et al. (2023)). Une piste pourrait être de les complémenter d'outils d'analyse contrefactuelle (*What if"*) afin que l'utilisateur puisse tester l'attribution si besoin.

3 Notre proposition : un cadre modulaire pour l'explicabilité par concepts

Nous proposons dans cette section un cadre pour l'explicabilité par concepts, post-hoc et non-supervisée en particulier, composé de 5 blocs modulaires et présenté en Figure 2. Ce cadre présente les avantages suivants :

— Il permet de regrouper de manière systématique les techniques proposées dans la littérature, les limites actuelles pour chaque module et les pistes d'amélioration ;
— Il permet de définir des critères d'évaluation pour chaque module afin de les évaluer individuellement ou de comparer plusieurs méthodes au sein d'un même bloc ;
— Il est modulaire et permet d'ajouter facilement de nouvelles méthodes.

Nous donnons dans la suite une description des différents blocs.

3.1 Formalisation du problème

Notre cadre d'explicabilité de situe dans les approches dîtes CBEs non supervisées comme expliqué dans la Section 2.2. Nous introduisons les notations de la Figure 2 et formalisons le problème comme suit :

Soit x un échantillon du jeu de données. Dans le cadre de données textuelles, cet échantillon correspond à une séquence d'éléments du vocabulaire identifiés par un entier, ainsi $x \in \mathbb{N}^k$ avec k la taille maximale de la séquence d'entrée pour le modèle.

Soit $\hat{y} \in \mathbb{R}^p$ la sortie du modèle pour l'échantillon x. La sortie est de taille p, qui correspond par exemple au nombre de classes pour un modèle de classification.

Soit f le réseau de neurones étudié, $f : \mathbb{N}^k \to \mathbb{R}^p$. Soit L la couche d'activation étudiée et dans laquelle les concepts sont extraits. Le modèle est ainsi composé de deux parties : $f = \phi_L \circ \psi_L$. La première fonction $\psi_L : \mathbb{N}^k \to \mathbb{R}^d$ prend en entrée x et calcule l'activation $a \in \mathbb{R}^d$ de la couche L. L'activation, de dimension d, correspond à une représentation du texte d'entrée complet. Ensuite, la fonction $\phi_L : \mathbb{R}^d \to \mathbb{R}^p$ prend en entrée l'activation a et calcule la prédiction $\hat{y}$. Ainsi, pour un échantillon, on a $\hat{y} = f(x) = \phi_L(\psi_L(x))$.

L'extraction de concepts s'effectue au sein de $A = (a^1, a^2, ..., a^n) \in \mathbb{R}^{n \times d}$ pour un ensemble de n échantillons $X = (x^1, x^2, ..., x^n) \in \mathbb{R}^{n \times k}$.

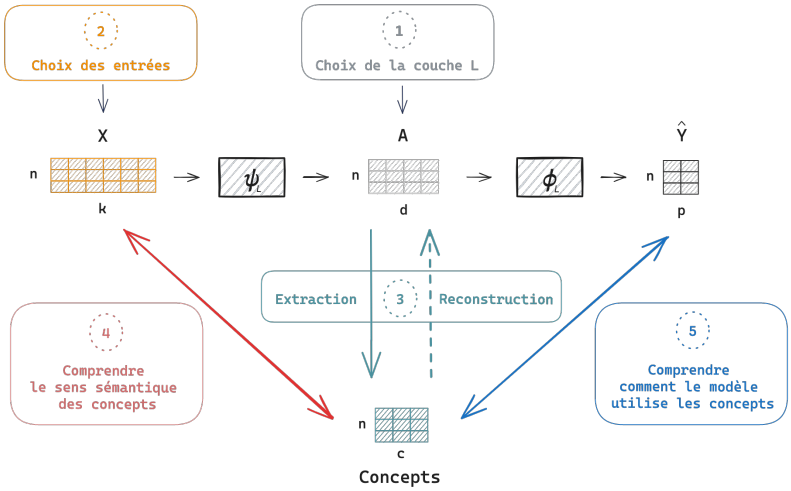


FIG. 2 – *Un cadre modulaire d'explicabilité post-hoc par concepts.*

3.2 Description des modules

Dans cette partie, nous présentons les différents modules, leur rôle ainsi que les techniques proposées dans la littérature.

Bloc 1. Choix de la couche d'activation L. Le choix de la couche d'activation dans laquelle les concepts sont extraits influence la qualité de l'explication sous deux aspects. Premièrement, les concepts doivent avoir un sens pour l'utilisateur. Par exemple, les modèles de traduction encodent des propriétés linguistiques de bas niveau (morphologie, relations locales) dans les premières couches puis des propriétés haut-niveau dans les dernières couches (sémantique, relations longue portée) (Belinkov et al. (2020)). Deuxièmement, l'utilisateur doit pouvoir comprendre comment le modèle utilise les concepts pour la prédiction. Plus la couche est proche de la sortie et moins d'opérations non-linéaires il y a entre les concepts et la prédiction.

Plusieurs stratégies sont adoptées dans l'état de l'art. Le plus souvent, les concepts sont extraits dans la dernière couche avant la prédiction (Jourdan et al. (2023)). D'autres méthodes consistent à choisir itérativement la meilleure couche en partant de la fin (Yeh et al. (2020)), où encore à extraire les concepts à plusieurs niveaux (Fel et al. (2023b)).

Bloc 2. Choix des données d'entrée. Le choix des données d'entrée a également un impact sur les concepts extraits. Deux stratégies complémentaires sont souvent utilisées. Premièrement, il peut être intéressant d'extraire les concepts dans un sous-ensemble du jeu de données, par classe par exemple, ce qui permet d'avoir des concepts plus précis et spécifiques à ce sous-ensemble (Vielhaben et al. (2023), Fel et al. (2023b)). La deuxième stratégie consiste à recadrer l'entrée, par exemple en ne prenant qu'une phrase dans un texte complet (Jourdan et al. (2023)). Ce traitement permet d'isoler les concepts pour les extraire plus facilement, mais pose

aussi des questions de validité des entrées pour le modèle, qui pourraient être hors-distribution (Vielhaben et al. (2023)).

Bloc 3. Extraction des concepts. L'extraction des concepts consiste à identifier dans l'espace des activations des régions qui correspondent à des concepts. Afin de les identifier de manière non-supervisée, l'extraction doit être guidée par des biais (Marconato et al. (2023)). Le premier choix est un a-priori sur la forme que prend le concept. Le plus souvent, les concepts sont définis comme étant des combinaisons linéaires de neurones. Un concept est alors représenté par un vecteur qui pointe dans la direction du concept. Cependant, il n'est pas garanti que le modèle encode les concepts de cette manière. Au lieu de vecteurs, Vielhaben et al. (2023) considère des concepts en plusieurs dimensions, définis par des bases, ce qui permet d'obtenir des concepts plus fidèles au modèle et des explications plus concises. Enfin, il n'est pas non plus garanti que les concepts soient linéairement séparables, il faut alors considérer des structures non-linéaires.

Un second choix important est le nombre c de concepts à extraire. Jourdan et al. (2023) utilise un jeu de données annoté pour comparer les performances selon le nombre de concepts. Zhao et al. (2023b) réalise une étude humaine pour étudier l'effet du nombre de concepts sur leur qualité.

Les méthodes les plus répandues sont celles utilisant du clustering d'activations (Vielhaben et al. (2023)), de la factorisation de matrice comme la PCA ou la NMF [1] (Zhang et al. (2020), Jourdan et al. (2023), Fel et al. (2023a)), et des auto-encodeurs (Cunningham et al. (2023), Zhao et al. (2023b), Yeh et al. (2020)) avec des fonctions de coût optimisant certains biais, comme la sparsité.

Bloc 4. Sens sémantique des concepts. Pour être interprétable, un concept doit avoir un sens sémantique aligné avec l'utilisateur, par exemple "Atteinte du système respiratoire". Pour comprendre le sens d'un concept, la méthode la plus utilisée consiste à extraire des exemples (prototypes, contre-factuels, semi-factuels), souvent agrémentés d'attribution par concepts. En vision, les techniques de *feature visualization* (Zeiler et Fergus (2013)) par concept sont aussi utilisées. Ces techniques visent à produire des exemples fictifs qui maximisent l'activation du concept.

La mise à disposition d'exemples peut être suffisante pour certaines applications, mais il peut être intéressant de décrire textuellement les concepts à partir de ces exemples. L'approche la plus répandue consiste à visualiser et annoter manuellement les concepts. Plus récemment, des méthodes automatiques utilisent des modèles de deep learning pour générer des descriptions (Kalibhat et al. (2023)).

Bloc 5. Utilisation des concepts. La dernière étape est de comprendre comment le modèle utilise les concepts pour parvenir à ses prédictions. L'attribution entre la couche de concepts et la sortie est souvent utilisée. Celle-ci nécessite d'avoir accès à la reconstruction de l'activation à partir des concepts. De plus, l'explication est locale. Afin de comprendre les mécanismes à une échelle plus globale, une première technique est d'agréger les explications locales (par exemple, le score TCAV [2] de Kim et al. (2018) est une agrégation par classe qui correspond

1. NMF : Non-negative matrix factorization
2. TCAV : Testing with Concept Activation Vectors

à la fraction d'exemples d'une classe positivement influencée par le concept), une autre est d'utiliser des outils de visualisation pour explorer facilement l'ensemble du jeu de données et les explications locales (Fel et al. (2023a)). D'autres techniques telles que l'extraction de règles et la distillation de modèle peuvent être envisagées (Dadvar et al. (2023)).

4 Application : classification de notes cliniques

Dans cette section, nous présentons un jeu de données ainsi qu'un modèle de classification de notes cliniques que nous avons développés. Ce cas d'application permet de valider l'approche en conditions réelles, avec notamment plusieurs défis.

4.1 Modèle de classification des myosites

Contexte. Les myopathies inflammatoires idiopathiques, aussi appelées myosites, sont des maladies auto-immunes affectant les muscles. Plusieurs types de myosites sont définis par des critères cliniques, biologiques et histologiques. Dans un but de construction de cohortes pour la recherche sur ce groupe de maladies, nous avons développé un modèle afin d'identifier les patients atteints de ces maladies à partir de leurs compte-rendus médicaux.

Base de données. Le modèle est entraîné de manière supervisée à partir d'un jeu de données composé d'un ensemble de dossiers patients revus et annotés par des médecins spécialistes des myosites. Les annotations correspondent aux 6 sous-types de myosites décrits ci-dessous ainsi qu'à la classe "Absence de myosite". La base finale est constituée de : **myosite à inclusions** (122 dossiers), **dermatomyosite** (144 dossiers), **myopathie nécrosante à médiation immunitaire** (112 dossiers), **myosite chevauchante** (147 dossiers), **myosite induite par les inhibiteurs de checkpoint** (75 dossiers), **syndrome des antisynthétases** (141 dossiers) et **absence de myosite** (1825 dossiers). 30% du jeu est réservé pour la validation.

Format de l'entrée. Les données disponibles pour chaque patient sont des comptes-rendus médicaux issus de leurs visites à l'hôpital. Afin de nettoyer et de réduire la taille des données, des concepts médicaux (à ne pas confondre avec les concepts issus de l'explicabilité qui sont plus haut-niveau) sont extraits des comptes-rendus grâce à un outil d'extraction développé par Scienta Lab, et sont mis en entrée du modèle. Un exemple est donné en Figure 3. Chaque patient est représenté par une séquence de visites, chaque visite étant une séquence de concepts médicaux.

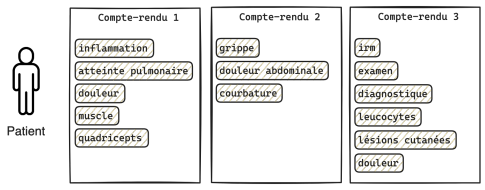


FIG. 3 – *Exemple de données (concepts médicaux) pour un patient.*

Modèle de classification. Le modèle que nous proposons pour la classification des notes cliniques, présenté en Figure 4, encode chaque concept médical, puis les agrège pour obtenir une représentation par visite, et enfin agrège les visites pour obtenir une représentation par patient. L'agrégation est une somme des représentations, pondérée par un mécanisme d'attention simple qui calcule un score pour chaque représentation indépendemment des autres. La classification est ensuite obtenue à partir de la représentation de patient. Afin de pouvoir tester l'extraction de concepts sur les sous-types de myosites dans la prochaine section, nous entraînons un modèle de classification binaire, les différentes classes de myosites étant réunies en une seule classe "Présence de myosite".

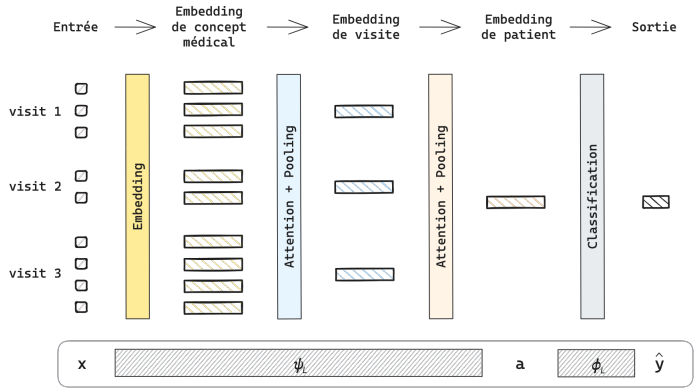


FIG. 4 – *Architecture du modèle de classification des myosites.*

Les performances sur le jeu de validation, contenant 771 dossiers, sont présentées dans le Tableau 1. Les résultats sont satisfaisants au regard de la difficulté de diagnostique des myosites et de l'utilisation de données de vie réelle mais la sensibilité devra être améliorée pour une utilisation en pratique.

F1	Sensibilité (recall)	Specificité	PPV (précision)	NPV
0.74	0.70	0.91	0.79	0.86

TAB. 1 – *Performances du modèle de classification binaire.*

4.2 Pipeline d'explicabilité

Dans cette section nous présentons les choix effectués pour les différents modules pour le modèle de classification des myosites, et les les résultats préliminaires de cette mise en oeuvre. Le modèle étudié est un modèle de classification binaire. Cependant, les annotations correspondant aux sous-types de myosites sont disponibles. Retrouver ces sous-types au sein des concepts extraits permettrait de valider la pipeline d'extraction non-supervisée des concepts. Les différents modules implémentés sont les suivants :

Bloc 1. Choix de la couche d'activation. Ce papier se concentre sur la dernière couche avant la prédiction, qui correspond aux représentations de patients. Toutefois, il pourrait aussi être intéressant d'extraire les concepts dans les représentations de visites.

Bloc 2. Choix des données d'entrée. Le jeu de données contient beaucoup de dossiers négatifs, extraire les concepts dans l'ensemble des échantillons pourrait mener à une sur-représentation des concepts liés aux dossiers négatifs. Il est plus intéressant de se concentrer sur les dossiers de patients atteints de myosites. Un filtre est appliqué pour sélectionner les dossiers prédits positifs par le modèle. Enfin, les données ne sont pas recadrées.

Bloc 3. Extraction des concepts. La méthode utilisée est la factorisation par NMF (Lee et Seung (1999)) avec extraction de 10 concepts. Les représentations de patients sont positives par application d'une ReLu. Le nombre de concepts devra faire l'objet d'une étude plus approfondie.

Bloc 4. Sens sémantique des concepts. Les méthodes actuelles d'extraction de concepts sur du texte sont limitées à des entrées de petite taille, de quelques mots à quelques phrases. Dans notre cas d'étude, une entrée peut représenter plusieurs milliers d'éléments, répartis au sein de plusieurs visites. Cela représente un défi pour la visualisation des prototypes pour comprendre le sens de chaque concept. La stratégie que nous avons choisie dans ce travail est la suivante : 5 prototypes par concept sont identifiés en prenant les échantillons qui activent le plus ce concept. Les visites qui ont le plus d'impact sur le concept sont ensuite identifiées en utilisant l'attribution par occlusion. Il existe plusieurs stratégies d'occlusion, nous choisissons pour commencer de masquer avec l'élément *unknown* mais ce choix devra faire l'objet d'une évaluation. Tous les éléments d'une visite sont masqués et on observe la variation du poids du concept. Enfin, les éléments des 3 visites les plus importantes de chaque prototype du concept sont regroupés et affichés dans un nuage de mots spécifique au concept. Les 30 éléments qui apparaissent le plus dans l'ensemble des prototypes sont filtrés afin de réduire le bruit. Les mots sont ensuite comparés à une liste de mots-clés définis par les spécialistes de la maladie pour chaque sous-type. Il est à noter que l'utilisation de nuages de mots entraîne une perte d'information importante, d'autres méthodes seront implémentées à l'avenir.

Bloc 5. Utilisation des concepts Afin d'estimer l'importance des concepts pour le modèle, on utilise Shap (Lundberg et Lee (2017)) pour calculer une attribution par concept et par prototype. Ce choix devra également être évalué.

Résultats préliminaires Suite à la mise en oeuvre de notre pipline, nous avons obtenus les premiers résultats synthétisés dans la Figure 5. Dans un premier temps, il est intéressant de vérifier que les représentations des patients pour chaque sous-type forment des groupes assez distincts en projetant les représentations avec UMAP pour les visualiser. La Figure 5.A montre que certains types de myosites forment en effet des groupes, tandis que d'autres sont moins distincts.

Ensuite l'analyse des nuages de mots obtenus sur les 10 concepts extraits, 4 semblent liés à des sous-types de myosites. En particulier, le concept 3 correspond au syndrome des anti-synthétases (Figure 5.E) et le concept 7 aux myosites à inclusions (Figure 5.F). Ce sont aussi

les deux sous-types les plus distincts sur la Figure 5.A. D'après la Figure 5.B, les 4 concepts liés aux sous-types sont aussi ceux qui ont le plus d'impact sur la prédiction du modèle. Ces résultats préliminaires sont intéressants et encourageant car permettent d'apprécier le fonctionnement des premières briques implémentées. Il est clair que nous devons poursuivre cette phase d'expérimentation pour une meilleure validation et évaluation du cadre proposé. Ce travail est en cours de réalisation.

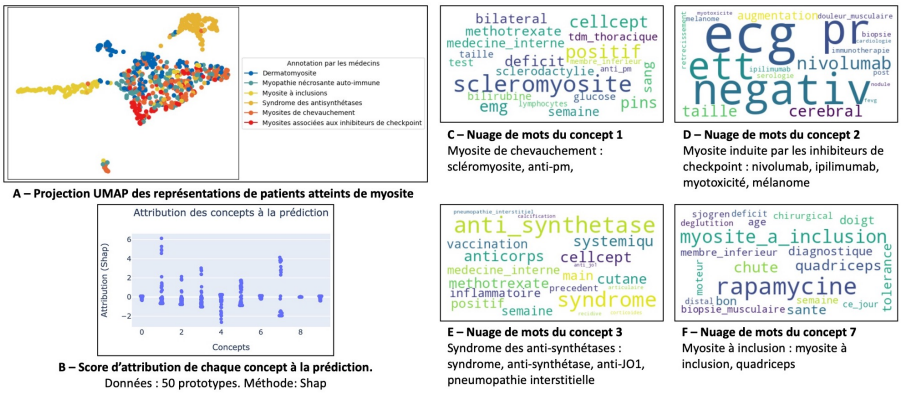


FIG. 5 – *Résultats préliminaires de l'extraction de concepts au sein du modèle de classification des myosites.*

5 Conclusion et perspectives

Ce papier présente un cadre pour l'explicabilité post-hoc non-supervisée par concepts avec l'identification de 5 blocs modulaires. La cadre est ensuite mis en oeuvre sur un cas réel de classification de dossiers patients, impliquant plusieurs défis. Les premiers résultats obtenus sont encourageants et les prochaines étapes vont se concentrer sur : (1) l'amélioration de chacun des blocs pour mieux comprendre le sens des concepts et comment le modèle les utilise pour sa prédiction, (2) l'évaluation, une étape importante dans la validation de l'approche. Chaque module nécessite une évaluation et plusieurs aspects doivent être évalués. Certains peuvent reposer sur une évaluation de type métrique, comme la fidélité de l'explication vis-à-vis du modèle. D'autres plus difficiles, comme la clarté du concept d'un point de vue sémantique, vont nécessiter la mise en place de protocoles d'évaluation avec une intervention humaine.

6 Contributions et remerciements

CC a effectué les recherches, implémenté les modèles et écrit l'article. JD, WO, CH ont relu l'article et guidé les recherches. JD a construit la base de dossiers patients ainsi que l'interface pour l'annotation. DV a relu et annoté les dossiers patients pour le modèle de classification. YA, FT, RS ont fourni une expertise médicale pour le modèle de classification. Nous remercions l'ANRT pour la thèse CIFRE n°[2023/0005].

Références

Atanasova, P., J. G. Simonsen, C. Lioma, et I. Augenstein (2020). A diagnostic study of explainability techniques for text classification. In B. Webber, T. Cohn, Y. He, et Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 3256–3274. Association for Computational Linguistics.

Belinkov, Y., N. Durrani, F. Dalvi, H. Sajjad, et J. Glass (2020). On the linguistic representational power of neural machine translation models. *Computational Linguistics 46*(1), 1–52.

Chefer, H., O. Lang, M. Geva, V. Polosukhin, A. Shocher, M. Irani, I. Mosseri, et L. Wolf (2023). The hidden language of diffusion models. *ArXiv abs/2306.00966*.

Cunningham, H., A. Ewart, L. Riggs, R. Huben, et L. Sharkey (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv :2309.08600*.

Dadvar, V., L. Golab, et D. Srivastava (2023). Poem : Pattern-oriented explanations of convolutional neural networks. *Proc. VLDB Endow. 16*, 3192–3200.

Deb, M., B. Deiseroth, S. Weinbach, M. Brack, P. Schramowski, et K. Kersting (2023). Atman : Understanding transformer predictions through memory efficient attention manipulation. *ArXiv abs/2301.08110*.

Fel, T., V. Boutin, M. Moayeri, R. Cadène, L. Bethune, M. Chalvidal, T. Serre, et al. (2023a). A holistic approach to unifying automatic concept extraction and concept importance estimation. *arXiv preprint arXiv :2306.07304*.

Fel, T., A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, et T. Serre (2023b). Craft : Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721.

Ghorbani, A., J. Wexler, J. Y. Zou, et B. Kim (2019). Towards automatic concept-based explanations. In *Neural Information Processing Systems*.

Imrie, F., R. Davis, et M. V. D. Schaar (2023). Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. *Nature Machine Intelligence 5*, 824 – 829.

Jourdan, F., A. Picard, T. Fel, L. Risser, J.-M. Loubes, et N. Asher (2023). COCKATIEL : COntinuous concept ranKed ATtribution with interpretable ELements for explaining neural net classifiers on NLP. In A. Rogers, J. Boyd-Graber, et N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics : ACL 2023*, Toronto, Canada, pp. 5120–5136. Association for Computational Linguistics.

Kalibhat, N., S. Bhardwaj, C. B. Bruss, H. Firooz, M. Sanjabi, et S. Feizi (2023). Identifying interpretable subspaces in image representations. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, et J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 15623–15638. PMLR.

Kim, B., M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, et R. Sayres (2018). Interpretability beyond feature attribution : Quantitative testing with concept activation vectors (TCAV). In *ICML*, pp. 2673–2682.

Lee, D. D. et H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Marconato, E., A. Passerini, et S. Teso (2023). Interpretability is in the mind of the beholder : A causal framework for human-interpretable representation learning. *Entropy 25*(12), 1574.

Poch'e, A., L. Hervier, et M. C. Bakkay (2023). Natural example-based explainability : a survey. In *xAI*.

Ray, P. P. (2023). Can llms improve existing scenario of healthcare ? *Journal of hepatology*.

Ribeiro, M. T., S. Singh, et C. Guestrin (2018). Anchors : High-precision model-agnostic explanations. In *Proceedings of AAAI'18/IAAI'18/EAAI'18*, pp. 1527–1535. AAAI Press.

Schwalbe, G. (2022). Concept embedding analysis : A review. *arXiv preprint arXiv :2203.13909*.

Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, et D. S. W. Ting (2023). Large language models in medicine. *Nature Medicine 29*, 1930 – 1940.

Vielhaben, J., S. Bluecher, et N. Strodthoff (2023). Multi-dimensional concept discovery (MCD) : A unifying framework with completeness guarantees. *Trans. on ML Research*.

Yeh, C.-K., B. Kim, S. Arik, C.-L. Li, T. Pfister, et P. Ravikumar (2020). On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems 33*, 20554–20565.

Zarlenga, M. E., M. E. Nelson, B. Kim, et M. Jamnik (2023). Tabcbm : Concept-based interpretable neural networks for tabular data.

Zeiler, M. D. et R. Fergus (2013). Visualizing and understanding convolutional networks. *ArXiv abs/1311.2901*.

Zhang, R., P. Madumal, T. Miller, K. A. Ehinger, et B. I. P. Rubinstein (2020). Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *AAAI Conference on Artificial Intelligence*.

Zhao, H., H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, et M. Du (2023a). Explainability for large language models : A survey. *arXiv preprint arXiv :2309.01029*.

Zhao, R., S. Joty, Y. Wang, et T. Wang (2023b). Explaining language models' predictions with high-impact concepts. *arXiv preprint arXiv :2305.02160*.

Summary

Large language models offer many possibilities in the medical field, particularly for medical research. In this context, numerous initiatives of research have been set up to design language models specific to the medical field, such as the DrBert model for the French language. However, one of the main limitations of the use of these models is the opacity of their behavior. In this paper, we present a modular framework for post-hoc explainability using concepts learned without supervision. We then apply this framework to a classification task of clinical notes.

IA neuro-symbolique explicable : état de l'art et perspectives pour la détection de fraude

Rita Sleiman*, Fatoumata Dama*

*Centre de Recherche et d'Innovation de Talan, France
rita.sleiman@talan.com,
fatoumata.dama@talan.com,
https://talan.com/a-propos/centre-recherche-innovation/

Résumé. L'IA neuro-symbolique est une nouvelle approche d'explicabilité qui allie la capacité prédictive des modèles neuronaux à la transparence des modèles symboliques. Les systèmes basés sur cette approche ont obtenu des résultats prometteurs dans la littérature pour des applications en classification d'images. Ce papier présente un état de l'art préliminaire des systèmes d'IA neuro-symbolique pour l'explicabilité et soulève la question de l'adaptation de cette approche à d'autres domaines tels que la détection de fraude.

1 Introduction

Les systèmes d'intelligence artificielle (IA), et plus précisément les réseaux de neurones profonds, ont atteint des performances élevées dans beaucoup d'applications telles que le traitement d'image, le traitement de langage naturel et la conduite autonome. Cela est principalement dû à la conception d'architectures de plus en plus sophistiquées, donc de plus en plus opaques, sacrifiant ainsi l'aspect « explicabilité ». Hors, l'explicabilité constitue un enjeu majeur en matière de responsabilité, d'éthique et de confiance, tout particulièrement lorsqu'il s'agit de domaines critiques tels que la santé ou la finance. D'où l'intérêt croissant de la communauté scientifique pour l'IA explicable au cours des dernières années (Adadi et Berrada, 2018; Tjoa et Guan, 2020).

Il existe des différences dans la manière dont les diverses communautés d'IA perçoivent le concept d'explicabilité. Une analyse de la littérature, réalisée par Doran et al. (2017), a permis d'identifier trois notions quant à l'explicabilité des systèmes d'IA : les systèmes complètement **opaques**, les systèmes **interprétables** et les systèmes **compréhensibles**. Les systèmes complètement opaques sont des boîtes noires qui ne donnent aucune information sur le mécanisme de prise de décision. À l'opposé, les systèmes interprétables sont des modèles transparents qui offrent une visibilité claire sur la démarche technique et mathématique des algorithmes utilisés. Enfin, les systèmes compréhensibles fournissent des résultats avec des éléments explicatifs (tels que des visualisations) permettant aux utilisateurs de formuler leurs propres explications sur la base de leurs propres connaissances.

Dans leur analyse, Doran et al. (2017) introduisent également un quatrième concept : les systèmes « **réellement explicables** ». Ces systèmes intègrent un raisonnement automatisé pour

la génération d'explications en s'appuyant sur des connaissances expertes. Les auteurs insistent sur l'importance primordiale de l'intégration directe du raisonnement afin qu'un modèle soit réellement explicable. En combinant les approches connexionnistes (réseaux de neurones) et symboliques, l'**IA neuro-symbolique** s'inscrit dans le cadre des systèmes « réellement explicables ».

L'**IA neuro-symbolique** est une nouvelle approche prometteuse qui permet de doter les modèles d'apprentissage profond de la capacité de raisonner sur un problème bien déterminé (Hitzler et Sarker, 2022; Sarker et al., 2021). Et plus précisément, l'utilisation de l'IA neuro-symbolique pour l'explicabilité des modèles d'IA permet de surmonter le compromis typique « Performance versus Interprétabilité », en combinant les capacités prédictives des modèles connexionnistes avec la transparence inhérente aux modèles symboliques. En effet, bien que les modèles d'IA symbolique manquent de la capacité de généralisation et soient moins performants que les approches connexionnistes, ils sont toujours utilisés et préférés dans de nombreux domaines d'application grâce à leur pouvoir explicatif. Toutefois, les méthodes connexionnistes sont critiquées pour leur opacité malgré leur précision.

Dans la section 2, nous décrivons trois travaux de la littérature qui ont proposé des solutions d'IA neuro-symbolique pour l'explicabilité. Ensuite, la section 3 présente les perspectives et les défis restant à couvrir afin d'appliquer ces solutions à d'autres domaines tels que la détection de fraude. Enfin, la dernière section conclut le papier.

2 État de l'art

Face aux limites des méthodes existantes d'IA explicable résultant en des systèmes mathématiquement interprétables ou légèrement compréhensibles, la nécessité de doter les modèles d'apprentissage profond de la capacité d'auto-expliquer leurs décisions est désormais une évidence absolue. Ceci nous permet d'obtenir un système « réellement explicable » et d'identifier les biais dans les modèles améliorant ainsi leur fidélité et leur robustesse. L'IA neuro-symbolique est une nouvelle approche qui vise à créer des systèmes « réellement explicables » en combinant la capacité prédictive des réseaux de neurones avec la transparence des modèles symboliques. Nous avons identifié trois travaux intéressants qui ont proposé des systèmes explicables basés sur ce principe (Díaz-Rodríguez et al., 2022; Bennetot et al., 2019, 2022).

2.1 Le système X-NeSyL

Le système X-NeSyL (pour eXplainable Neural-Symbolic Learning) a été introduit par Díaz-Rodríguez et al. (2022) dans le but de traiter le défi consistant à rendre les modèles neuronaux interprétables tout en fournissant des explications universelles à la fois pour les utilisateurs finaux et les experts du domaine. La méthodologie proposée par les auteurs associe des représentations symboliques et profondes en introduisant une métrique d'explicabilité pour évaluer l'alignement des explications de la machine avec celles de l'expert humain.

X-NeSyL est composé de trois composants clés. Le premier composant comprend une unité de traitement symbolique, qui utilise des graphes de connaissances pour modéliser les connaissances explicites des experts du domaine. Le deuxième composant, une unité de traitement neuronal appelée EXPLANet, utilise une architecture profonde compositionnelle pour

classer les objets en fonction des parties détectées. Enfin, le troisième et dernier composant est une procédure d'entraînement utilisant un mécanisme de "*feedback*" d'explicabilité, appelé SHAP-Backprop, conçue pour guider le modèle dans l'alignement de ses sorties avec des explications symboliques.

Dans le module SHAP-Backprop, les auteurs utilisent l'analyse de Shapley et une fonction de pénalisation (*misattribution function*) pour ajuster le modèle en pénalisant les désalignements avec les graphes de connaissances. La mesure de l'explicabilité introduite dans cet article, SHAP Graph Edit Distance (SHAP GED), évalue l'interprétabilité du modèle en mesurant l'alignement entre les représentations symboliques (de l'expert) et les représentations neuronales (de la machine).

Les auteurs ont illustré leur système sur une application de classification de styles architecturaux de monuments en utilisant l'ensemble de données **MonuMAI** (Lamas et al., 2021). Les résultats expérimentaux démontrent que X-NeSyL améliore non seulement la capacité d'explication des réseaux neuronaux convolutionnels (CNN), mais aussi leurs performances. Cependant, une telle méthodologie ne pourra pas être utilisée dans des applications où les connaissances des experts sont indisponibles, rendant ainsi la construction des graphes de connaissance difficile.

2.2 Le système proposé par Bennetot et al. (2019)

Dans la méthodologie proposée par Bennetot et al. (2019), au lieu de fournir une base de connaissances externe (comme c'est le cas dans le système X-NeSyL), les auteurs suggèrent d'extraire des règles symboliques directement à partir des données afin de construire une base de connaissances. Cela permet d'obtenir des explications en langage naturel directement à partir de la boîte noire, ce qui permet de comprendre les problèmes potentiels liés au raisonnement du modèle, de mettre en évidence la possibilité d'un biais dans l'ensemble des données ou dans l'entraînement, et d'améliorer la précision du modèle en général.

Les auteurs émettent l'hypothèse que l'utilisation de fonctions de perte ayant une signification concrète peut rendre les explications plus accessibles que l'entropie croisée traditionnelle, moins intuitive. Les auteurs soulignent également l'importance de deux conditions préalables à la conception d'un système réellement explicable : la base de connaissances doit provenir directement de la boîte noire, ce qui garantit un vrai alignement symbolique avec le processus d'apprentissage du modèle, et la partie symbolique doit contraindre la partie connexionniste afin d'améliorer les performances de prédiction.

2.3 Le système Greybox XAI

En suivant le même principe de la génération de base de connaissances, « Greybox XAI », la méthodologie proposée par Bennetot et al. (2022), consiste à extraire une base de connaissances de l'ensemble de données, utilisée pour former un modèle transparent, en particulier une régression logistique. Simultanément, une architecture d'encodage-décodage est entraînée sur des images pour générer une sortie ressemblant à la base de connaissances utilisée par le modèle transparent. Ces modèles formés indépendamment, la régression logistique et le codeur-décodeur, sont ensuite intégrés de manière à constituer un modèle prédictif explicable. Particulièrement conçue pour la classification d'images, cette architecture garantit la transparence en combinant un codeur-décodeur pour créer un espace latent explicable, utilisé ensuite

par la régression logistique pour en tirer un nouveau graphe de connaissance pour représenter le lien entre les attributs et les classes, facilitant ainsi leur explication. Cette intégration permet alors une compréhension transparente des raisons qui expliquent la classification d'une image. Les auteurs ont validé le modèle proposé en l'appliquant sur deux ensembles de données (MonuMai et PASCAL-Part). Les résultats montrent la supériorité de leur modèles sur d'autres modèles explicables.

3 Perspectives et travaux futurs

L'IA neuro-symbolique est une nouvelle approche d'explicabilité qui allie la capacité prédictive des modèles neuronaux à la transparence des modèles symboliques. Les systèmes basés sur cette approche, proposés dans l'état de l'art, ont fourni des résultats prometteurs. En effet, les résultats expérimentaux démontrent la capacité de ces systèmes à améliorer non seulement la transparence des modèles neuronaux (habituellement qualifiés de boîtes noires) mais permettent également d'améliorer leurs performances prédictives.

Cependant, les travaux réalisés dans la littérature abordent l'application spécifique de la classification d'images et considèrent uniquement les modèles neuronaux adaptés à cette tâche (e.g., CNN). Il serait donc intéressant d'étudier d'autres architectures de réseaux neuronaux adaptées à d'autres types de données (e.g., données tabulaires et séries temporelles). D'autre part, il serait également intéressant d'étendre l'application de l'IA neuro-symbolique à des domaines sensibles tels que le médical et la finance.

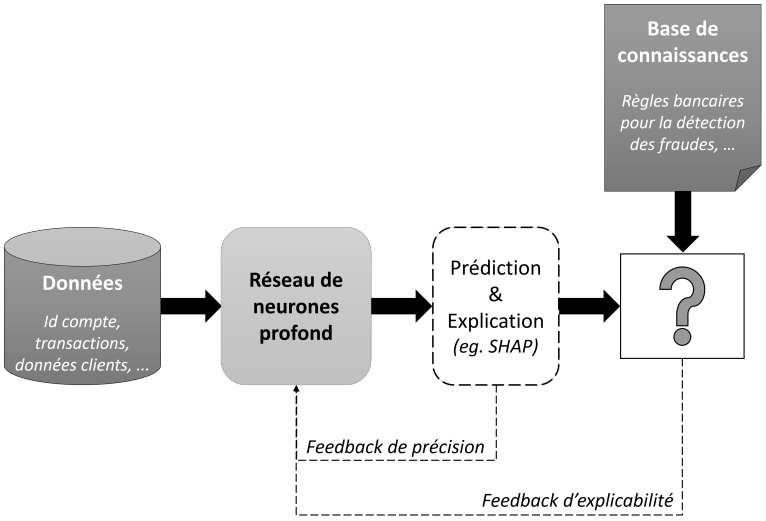


FIG. 1 – *Proposition d'un modèle neuro-symbolique explicable pour la détection de fraude*

Nous pensons que la détection de fraude est une application particulièrement intéressante pour laquelle une base de connaissance conséquente est déjà disponible. En effet, les banques utilisent de nombreuses règles de détection de fraude dans leurs dispositifs de surveillance des

transactions. Ces règles peuvent donc être utilisées pour construire la partie symbolique du système. La figure 1 présente l'architecture d'un système d'IA neuro-symbolique pouvant être appliquée à la détection de fraude. Notons que, le principal défi de la conception de ce système consistera à proposer une métrique qui évalue l'alignement des explications de la machine avec les règles expertes (représentée par la boite avec un point d'interrogation dans la figure 1). L'explication des prédictions du réseau de neurones et la définition d'un mécanisme de *feedback* d'explicabilité constituent également des questions importantes à considérer.

4 Conclusion

Avec l'utilisation croissante de l'IA dans de nombreuses applications pratiques, en particulier celles ayant des implications critiques pouvant influencer significativement les prises de décision, le besoin d'une IA digne de confiance continue à augmenter. Ainsi, des efforts sont consacrés au développement des systèmes d'IA explicable pouvant générer des explications sur leur raisonnement d'une manière compréhensible pour les humains. Dans ce papier, nous présentons une étude préliminaire de l'état de l'art des systèmes d'IA Neuro-symbolique explicable : une approche d'explicabilité prometteuse mais faiblement explorée dans la littérature. Il a été démontré que l'intégration des approches connexionnistes et symboliques améliore à la fois les performances du système, sa robustesse ainsi que sa capacité à fournir des explications claires et compréhensibles. Pour toutes ces raisons, nous pensons qu'il serait pertinent d'étudier l'application de cette approche à d'autres domaines (autres que l'image) tels que la détection de fraude. Cette perspective fera l'objet de nos travaux futurs.

Références

Adadi, A. et M. Berrada (2018). Peeking inside the black-box : a survey on explainable artificial intelligence (xai). *IEEE access 6*, 52138–52160.

Bennetot, A., G. Franchi, J. Del Ser, R. Chatila, et N. Diaz-Rodriguez (2022). Greybox xai : A neural-symbolic learning framework to produce interpretable predictions for image classification. *Knowledge-Based Systems 258*, 109947.

Bennetot, A., J.-L. Laurent, R. Chatila, et N. Díaz-Rodríguez (2019). Towards explainable neural-symbolic visual reasoning. *arXiv preprint arXiv :1909.09065*.

Díaz-Rodríguez, N., A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, et F. Herrera (2022). Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs : The monumai cultural heritage use case. *Information Fusion 79*, 58–83.

Doran, D., S. Schulz, et T. R. Besold (2017). What does explainable ai really mean ? a new conceptualization of perspectives. *arXiv preprint arXiv :1710.00794*.

Hitzler, P. et M. K. Sarker (2022). Neuro-symbolic artificial intelligence : The state of the art.

Lamas, A., S. Tabik, P. Cruz, R. Montes, Á. Martínez-Sevilla, T. Cruz, et F. Herrera (2021). Monumai : Dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification. *Neurocomputing 420*, 266–280.

Sarker, M. K., L. Zhou, A. Eberhart, et P. Hitzler (2021). Neuro-symbolic artificial intelligence. *AI Communications 34*(3), 197–209.

Tjoa, E. et C. Guan (2020). A survey on explainable artificial intelligence (xai) : Toward medical xai. *IEEE transactions on neural networks and learning systems 32*(11), 4793–4813.

Summary

Neuro-symbolic AI is a new approach to explainable AI that combines the predictive potential of neural models with the transparency of symbolic models. Systems based on this approach have achieved promising results in the literature for image classification applications. This paper presents a preliminary state-of-the-art of Neuro-symbolic AI systems for explicability and raises the question of adapting this approach to other domains such as fraud detection.

Pizzaïolo Dataset : Des Images Synthétiques Ontologiquement Explicables

Grégory Bourguin*, Arnaud Lewandowski**

LISIC, Laboratoire Informatique Signal et Image de la Côte d'Opale
50 rue Ferdinand Buisson - 62228 - Calais CEDEX
https://lisic-prod.univ-littoral.fr/
* gregory.bourguin@univ-littoral.fr
** arnaud.lewandowski@univ-littoral.fr

Résumé. Les travaux issus du mouvement XAI ont souligné la nécessité d'inventer des modèles d'IA explicables du point de vue de leurs utilisateurs. Une approche prégnante vise à concevoir des modèles qui réifient les concepts et raisonnements du domaine d'application. En Vision par Ordinateur, les travaux qui expérimentent la création de ces nouvelles IA ont souligné l'intérêt de datasets synthétiques dédiés à des domaines spécifiques, proposant des représentations visuelles simples, et mettant en oeuvre des concepts et raisonnements non ambigus, explicites, et plus ou moins complexes. Nos travaux en XAI s'inscrivent dans cette mouvance et nous avons créé le *Pizzaïolo Dataset* : un ensemble d'exemples générés à partir d'une ontologie sous forme d'images synthétiques de pizzas annotées pour expérimenter la conception d'IA explicables. L'objectif de ce papier est de présenter le *Pizzaïolo Dataset*, et de le mettre librement à disposition de la communauté des chercheurs.

1 Introduction

Les travaux issus du mouvement XAI (eXplainable AI) ont permis de mettre en exergue la nécessité d'inventer des modèles d'IA qui soient non seulement interprétables, mais aussi explicables du point de vue de leurs utilisateurs. Les utilisateurs ne sont généralement pas des spécialistes en IA, mais des spécialistes de leur domaine d'application. Pour qu'un utilisateur puisse comprendre une explication, il faut qu'elle concorde avec sa connaissance, c'est à dire qu'elle manipule des concepts qui lui sont accessibles, voire qu'elle utilise un raisonnement en adéquation avec sa propre façon de penser.

Nous nous intéressons au domaine de la Vision par Ordinateur. Pour répondre à ce besoin, de nombreux travaux de recherche visent à expliquer les modèles d'IA existants en faisant ressortir les concepts du domaine auxquels ils sont dédiés (Fel et al., 2023). D'autres approches ont pour but de créer des modèles d'IA directement conçus pour manipuler les concepts du domaine visé. On peut en particulier citer les approches CBM (Concept Based/Bottleneck Models) (Losch et al., 2019; Koh et al., 2020), ainsi que les approches Neuro-Symboliques dont l'objectif est de mettre en oeuvre un raisonnement plus explicable du point de vue des utilisateurs (Barbiero et al., 2023; Han et al., 2023).

Les expérimentations autour de ces nouvelles IA explicables nécessitent des jeux de données (datasets) particuliers : ils doivent contenir des exemples qui reflètent exclusivement et explicitement des raisonnements qui manipulent des concepts spécifiques au domaine d'application. Si certains travaux utilisent des jeux de données réels (Liu et al., 2015), ceux qui focalisent sur le raisonnement sémantique tirent généralement parti de datasets synthétiques, c.à.d. contenant des exemples générés qui mettent en oeuvre des propriétés visuelles simples (Kottur et al., 2019; Seo et al., 2018) pour représenter des concepts et raisonnements non ambigus (Barbiero et al., 2023). Plusieurs travaux récents ont démontré l'intérêt résultant d'un dataset construit à partir d'une ontologie (Ribeiro et Leite, 2021; Agafonov et Ponomarev, 2022). En effet, les ontologies ont pour objet la réification des connaissances (concepts et raisonnements) liées à un domaine utilisateur. Une ontologie permet donc de générer un dataset contrôlé qui représente explicitement les concepts et raisonnements souhaités.

Dans le cadre de nos propres travaux sur la construction d'IA explicables (Bourguin et al., 2021), nous avons créé le *Pizzaïolo Dataset* (Bourguin et Lewandowski, 2023), un jeu de données permettant la classification d'images synthétiques de pizzas généré à partir d'une variation de l'ontologie des Pizzas (Horridge, 2011). Sa création a été motivée par le fait que nous n'avons pas pu trouver d'autre dataset qui réponde à nos besoins, c'est à dire : libre, totalement basé sur (et fourni avec) une ontologie aux concepts et descriptions de classes plus ou moins complexes, mettant en oeuvre différentes propriétés ontologiques, et proposant plusieurs types d'annotations (labels, segmentation sémantique, bounding boxes, segmentation d'instances) qui permettent des expérimentations dans diverses tâches (classification multi-class/multi-label, génération d'explications, ...).

L'objectif de ce papier est de présenter le *Pizzaïolo Dataset*[1], et de le mettre publiquement à disposition de la communauté des chercheurs pour qu'ils puissent librement l'utiliser dans leurs propres travaux. Nous fournissons en sus le module Python *Pizzaiolo*[2] que nous avons créé pour l'occasion et avec lequel ce dataset a été généré.

2 Le Pizzaiolo

La figure 1 donne un aperçu des images synthétiques qui composent le *Pizzaïolo Dataset*. D'une manière générale, on peut constater que ces images sont étiquetées à partir de définitions du domaine (définitions ontologiques de classes : recettes, caractéristiques). Ces étiquettes sont destinées à expérimenter la création d'IA (ontologiquement explicables) dans des tâches de classification multi-classes (ex. nom/recette de pizza : **Fiorentina**, etc.), et/ou multi-label (ex. nom de pizza + caractéristiques : **VegeratianPizza**, **SpicyPizza**, etc.).

Les images sont constituées d'un assemblage d'éléments graphiques libres[3] qui représentent la base/pâte d'une pizza recouverte d'un ensemble d'icônes correspondant à des toppings (anchois, olives, etc.), et éventuellement d'une icône de drapeau qui indique le pays d'origine de la (recette de) pizza. Toutes les images ont été générées grâce au module Python *Pizzaiolo*[2] que nous avons développé et rendu disponible sur *Github*. Cette partie vise à expliciter les principaux mécanismes du processus de génération mis en oeuvre.

1. https://zenodo.org/records/10165941
2. https://github.com/SysReIC/pizzaiolo
3. https://www.flaticon.com/

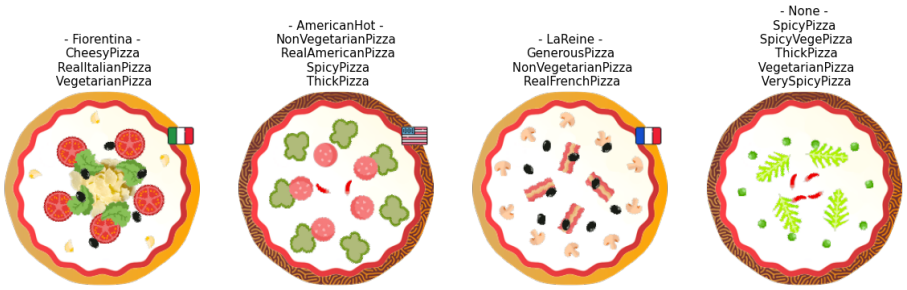


FIG. 1 – *Exemples d'images synthétiques de pizzas générées par Pizzaiolo. Les étiquettes ont été calculées selon les descriptions du domaine (classes ontologiques) et à partir des concepts (base, toppings, pays d'origine) présents sur l'image : la 1ère étiquette correspond au nom (recette) de la pizza générée, les autres étiquettes en soulignent diverses caractéristiques. Les 3 premières images ont été générées à partir d'une "recette ontologique" (classes* ***Fiorentina****,* ***AmericanHot*** *et* ***LaReine****), la 4ème résulte d'une génération aléatoire (étiquetée par* ***None****).*

2.1 Ontologies des Pizzas Allégées : *Pizzaïolo Ontology*

2.1.1 Concepts Généraux

La génération des exemples est contrôlée par la *Pizzaïolo Ontology*, une adaptation de la fameuse ontologie des *Pizzas* (Horridge, 2011). Cette ontologie introduit le concept de **Pizza** constituée d'une base (**hasBase.PizzaBase**) et de toppings (**hasTopping.PizzaTopping**) plus ou moins épicés (**hasSpiciness.Spiciness**). En résultent les descriptions suivantes :

Pizza $\equiv$ Food $\sqcap$ ($\exists$ **hasBase.PizzaBase**) $\sqcap$ ($\exists$ **hasTopping.PizzaTopping**)
PizzaBase $\sqsubseteq$ Food
PizzaTopping $\sqsubseteq$ Food $\sqcap$ $\exists$ **hasSpiciness.Spiciness**
Mild $\sqsubseteq$ Spiciness ; **Medium** $\sqsubseteq$ Spiciness ; **Hot** $\sqsubseteq$ Spiciness

La classe **PizzaBase** est spécialisée en **2 sous-classes** :

DeepPanBase $\sqsubseteq$ PizzaBase
ThinAndCrispyBase $\sqsubseteq$ PizzaBase

La classe **PizzaTopping** est spécialisée en **6 catégories principales** qui permettent de caractériser/étiqueter les pizzas qui les utilisent (pizza au fromage, végétarienne, épicée, etc.) :

CheeseTopping $\sqsubseteq$ PizzaTopping
FruitTopping $\sqsubseteq$ PizzaTopping
MeatTopping $\sqsubseteq$ PizzaTopping
SeafoodTopping $\sqsubseteq$ PizzaTopping
VegetableTopping $\sqsubseteq$ PizzaTopping
SpicyTopping $\equiv$ PizzaTopping $\sqcap$ (($\exists$ hasSpiciness.**Medium**) $\sqcup$ ($\exists$ hasSpiciness.**Hot**))

Ces catégories de toppings donnent elles-mêmes naissance à un total de **16 sous-classes** supplémentaires de **PizzaTopping** qui correspondent aux types des éléments visuels (anchois, olives, etc.) qui sont disposés sur les images de pizzas générées. Ces 16 sous-classes sont issues d'une sélection parmi les 32 sous-classes de **PizzaTopping** de l'ontologie des Pizzas originelle

que nous avons allégée en retirant les classes de toppings (comme les coulis) qui auraient inutilement complexifié les images synthétiques. La liste complète est donnée en figure 4.

On trouve ainsi par exemple :

AnchovyTopping $\sqsubseteq$ SeafoodTopping $\sqcap$ ($\exists$ hasSpiciness.Mild)
PeperoniSausageTopping $\sqsubseteq$ MeatTopping $\sqcap$ ($\exists$ hasSpiciness.Medium)
JalapenoPepperTopping $\sqsubseteq$ VegetableTopping $\sqcap$ ($\exists$ hasSpiciness.Hot)

On peut enfin ajouter que l'ontologie utilise de plus la propriété **hasCountryOfOrigin** qui a pour *domain* **Pizza** et pour *range* une instance de **Country** parmi **America**, **England**, **France** et **Italy**, ce qui permet d'indiquer le pays d'origine d'une "recette" de pizza.

2.1.2 Recettes de Pizzas (sous-classes de NamedPizza)

A partir de ces concepts et propriétés, l'ontologie décrit 15 "recettes" qui correspondent à **15 sous-classes de NamedPizza** (elle-même sous-classe de **Pizza**). Chacune de ces classes possède une description précise utilisée par *Pizzaiolo* pour générer les images du dataset. Inversement, ces descriptions permettent d'inférer si (et d'expliquer pourquoi) une pizza "inconnue" est une instance de ces sous-classes. La liste complète est donnée en figure 4.

A titre d'exemple, on trouve la description de la pizza **Napoletana** définie par :

Napoletana $\equiv$ Pizza $\sqcap$ ($\forall$ hasCountryOfOrigin.Italy)
$\sqcap$ ($\exists$ hasTopping.AnchovyTopping) $\sqcap$ ($\exists$ hasTopping.OliveTopping)
$\sqcap$ ($\forall$ hasTopping.(AnchovyTopping $\sqcup$ OliveTopping))
$\sqcap$ ($\exists$ hasBase.ThinAndCrispyBase) $\sqcap$ ($\forall$ hasBase.ThinAndCrispyBase)

2.1.3 Étiquettes Ontologiques Supplémentaires

L'ontologie proposée décrit **11 sous-classes de Pizza supplémentaires** qui permettent de caractériser des instances de pizza par inférence. La liste de ces 11 classes supplémentaires est donnée en figure 4. Leur intérêt est de fournir des étiquettes correspondant à des raisonnements ontologiques plus complexes que la "simple" identification des recettes à partir des instances de concepts (base, toppings, pays) constituant une instance de pizza.

Certaines définitions impliquent un raisonnement basé sur les super-classes (catégories) de toppings (**MeatTopping**, **SpicyTopping**, etc.) citées dans la partie 2.1.1.

Par exemple :

VegetarianPizza $\equiv$ Pizza
$\sqcap \neg$ ($\exists$ hasTopping.SeafoodTopping)
$\sqcap \neg$ ($\exists$ hasTopping.MeatTopping)
SpicyPizza $\equiv$ Pizza $\sqcap$ ($\exists$ hasTopping.SpicyTopping)
SpicyVegePizza $\equiv$ VegetarianPizza $\sqcap$ SpicyPizza

D'autres définitions forcent à prendre en compte le nombre de toppings générés :

GenerousPizza $\equiv$ Pizza $\sqcap$ ($\geq$ 8 hasTopping.OliveTopping)
VerySpicyPizza $\equiv$ SpicyPizza
$\sqcap$ (($\geq$ 8 hasTopping.SpicyTopping) $\sqcup$ ($\geq$ 4 hasTopping.($\exists$ hasSpiciness.Hot))

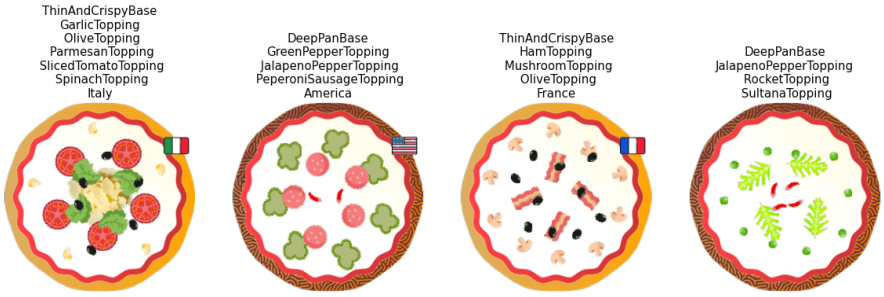


FIG. 2 – *Exemples de représentation visuelle des concepts constituant les pizzas.*

2.2 Génération d'Image Ontologique et Aléatoire

La génération des exemples par *Pizzaiolo* se base sur les concepts de la *Pizzaïolo Ontology* et sur des tirages aléatoires "contrôlés" pour obtenir des variations conformes aux descriptions ontologiques du domaine. L'ontologie est manipulée grâce à Owlready2 (Lamy, 2017).

2.2.1 Les Images

Les images générées ont une taille de 224*224 pixels. Elles utilisent toutes le même fond pour aider une IA à focaliser sur les concepts présents (base, toppings, pays). La figure 2 présente des exemples de représentations visuelles des concepts constituant les pizzas. Les sous-classes de **PizzaBase** sont caractérisées par des contours différents. Les toppings sont générés aléatoirement en suivant quelques règles pour garantir la lisibilité. Les éléments visuels pouvant être amenés à se chevaucher, nous avons fixé un ordre de génération pour les sous-classes de **PizzaTopping** de manière à ce que les élément les plus gros (ex. **ParmesanTopping**) ne recouvrent pas les plus petits (ex. **OliveTopping**). Leur position est tirée aléatoirement entre 2 cercles (min et max) aux rayons fixés en fonction de la classe de topping. Le nombre d'instances est choisi aléatoirement dans un intervalle aux bornes fixées en fonction de la classe de topping. L'orientation est définie aléatoirement. Lorsqu'un pays d'origine (**Country**) est spécifié, il est représenté par une icône de drapeau apposée sur le bord de la pizza.

2.2.2 Annotations

Chaque image générée est accompagnée d'annotations destinées à la création d'IA mettant en oeuvre diverses modalités. Pour ce faire, nous avons intégré plusieurs types d'annotations qui sont représentées dans l'exemple en figure 3. On trouvera pour chaque exemple :

- le **type de pizza** générée (sous-classe de **NamedPizza**, ou ***None*** si pizza aléatoire)
- l'ensemble des **étiquettes ontologiques** qui caractérisent la pizza (sous-classes de **Pizza**)
- la liste des **concepts** constituants (sous-classe de **PizzaBase**, sous-classes de **PizzaTopping**, instance de **Country**)
- l'ensemble des **bounding boxes** (pour la détection d'instances de concepts)
- l'ensemble des **contours** (pour la segmentation d'instances de concepts)
- le masque de **segmentation sémantique** pour tous les concepts constituants

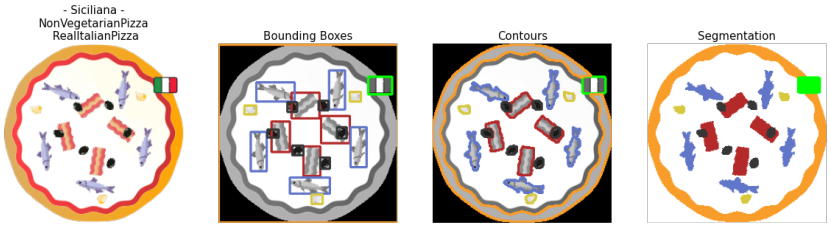


FIG. 3 – *Les différents types d'annotations générés pour chaque pizza.*

L'ensemble des annotations prend en compte le clipping entre instances de concepts. Par exemple, dans le cas où un **AnchovyTopping** se trouve coupé en 2 par le chevauchement d'un **HamTopping**, cet **AnchovyTopping** donne naissance à 2 bounding boxes et 2 contours, chacun(e) soulignant une partie de l'élément initial.

3 Le Pizzaïolo Dataset

Le *Pizzaïolo Dataset* (Bourguin et Lewandowski, 2023) est constitué de **4800 exemples** générés par *Pizzaiolo* et étiquetés conformément à la *Pizzaïolo Ontology*. Ce dataset contient :

- la ***Pizzaïolo Ontology*** en *OWL 2* : format *.xml*
- les **images** synthétiques de pizzas : format *.png*
- les **bounding boxes** des concepts : 1 fichier *.json* par image
- les **contours** des concepts : 1 fichier *.json* par image
- la **segmentation sémantique** globale : 1 fichier *.txt* par image
- le **descriptif détaillé** de l'ensemble des exemples (étiquettes, types de concepts, décomptes, noms des fichiers d'annotations, etc.) : 1 fichiers *.csv*

La figure 4 offre le détail de la répartition des étiquettes et des concepts au sein du dataset. Le site du projet propose une répartition équilibrée en 3 sous-ensembles pour l'entraînement, la validation et les tests. Il est à noter que le module *Pizzaiolo* permet de transformer aisément les bounding boxes du *Pizzaïolo Dataset* vers le format YOLO.

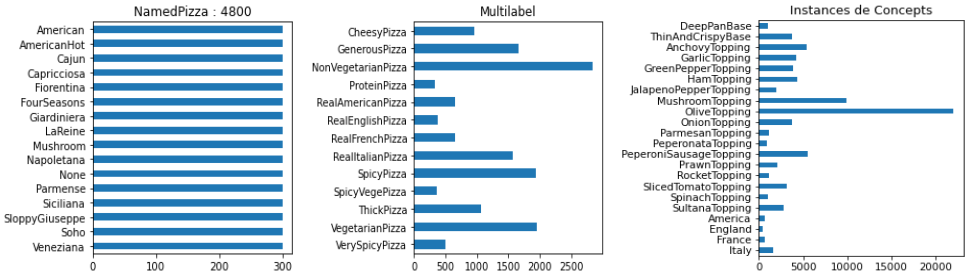


FIG. 4 – *Répartition des étiquettes et des instances de concepts au sein du Pizzaïolo Dataset.*

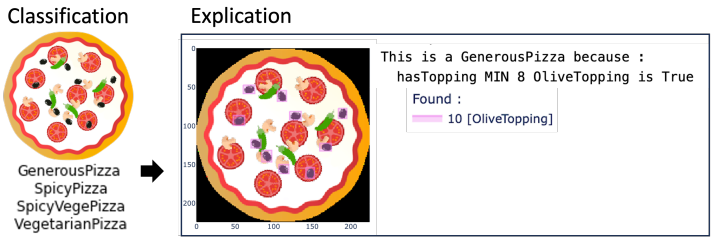


FIG. 5 – *Exemple d'utilisation du Pizzaïolo Dataset dans le projet OntoClassifier (***?***).*

4 Conclusion

Dans le but d'aider la recherche en XAI nous proposons le *Pizzaïolo Dataset* fondé sur la *Pizzaïolo Ontology*. Ce dataset présente des étiquettes qui correspondent à des descriptions de classes simples et complexes. Les annotations disponibles permettent d'envisager des tâches diverses et des modalités variées. Nous fournissons en sus le module *Pizzaiolo* qui permet au besoin de générer de nouveaux exemples. Nous avons utilisé ces outils avec succès dans nos propre travaux pour la création de classifieurs automatiques ontologiquement explicables : l'approche est exposée dans (Bourguin et al., 2021) et la figure 5 donne un exemple des derniers résultats obtenus. Nous espérons que ce dataset et ses outils aideront les chercheurs à inventer de nouvelles IA qui soient explicables du point de vue de leur domaine d'application.

Références

Agafonov, A. et A. Ponomarev (2022). An Experiment on Localization of Ontology Concepts in Deep Convolutional Neural Networks. In *The 11th International Symposium on Information and Communication Technology, SoICT 2022, Hanoi, Vietnam, December 1-3, 2022*, pp. 82–87. ACM.

Barbiero, P., G. Ciravegna, F. Giannini, M. E. Zarlenga, L. C. Magister, A. Tonda, P. Lio, F. Precioso, M. Jamnik, et G. Marra (2023). Interpretable Neural-Symbolic Concept Reasoning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, et J. Scarlett (Eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 1801–1825. PMLR.

Bourguin, G. et A. Lewandowski (2023). Pizzaïolo Dataset (1.0.0) [Data set]. Zenodo. https ://doi.org/10.5281/zenodo.10165941.

Bourguin, G., A. Lewandowski, M. Bouneffa, et A. Ahmad (2021). Towards Ontologically Explainable Classifiers. In I. Farkas, P. Masulli, S. Otte, et S. Wermter (Eds.), *Artificial Neural Networks and Machine Learning - ICANN 2021 - 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14-17, 2021, Proceedings, Part II*, Volume 12892 of *Lecture Notes in Computer Science*, pp. 472–484. Springer.

Fel, T., A. Picard, L. Béthune, T. Boissin, D. Vigouroux, J. Colin, R. Cadènc, et T. Serre (2023). CRAFT : Concept Recursive Activation FacTorization for Explainability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 2711–2721. IEEE.

Han, Z., L.-W. Cai, W.-Z. Dai, Y.-X. Huang, B. Wei, W. Wang, et Y. Yin (2023). Abductive subconcept learning. *Science China Information Sciences 66*(2), 122103.

Horridge, M. (2011). Protégé OWL Tutorial | OWL research at the University of Manchester. http ://owl.cs.manchester.ac.uk/publications/talks-and-tutorials/protg-owl-tutorial/.

Koh, P. W., T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, et P. Liang (2020). Concept bottleneck models. In H. D. III et A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR.

Kottur, S., J. M. F. Moura, D. Parikh, D. Batra, et M. Rohrbach (2019). CLEVR-Dialog : A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog.

Lamy, J.-B. (2017). Owlready : Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine 80*, 11–28.

Liu, Z., P. Luo, X. Wang, et X. Tang (2015). Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society.

Losch, M. M., M. Fritz, et B. Schiele (2019). Interpretability Beyond Classification Output : Semantic Bottleneck Networks. *CoRR abs/1907.10882*.

Ribeiro, M. d. S. et J. Leite (2021). Aligning Artificial Neural Networks and Ontologies towards Explainable AI. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 4932–4940. AAAI Press.

Seo, P. H., A. Lehrmann, B. Han, et L. Sigal (2018). Visual Reference Resolution using Attention Memory for Visual Dialog.

Summary

Research works in the XAI movement have emphasized the need to invent AI models that are explainable from their users' perspective. A prevalent approach aims to design models that reify the concepts and reasoning of the application domain. In Computer Vision, studies experimenting with the creation of these new AIs have highlighted the importance in working with synthetic datasets. These datasets propose simple visual representations, and implement explicit non-ambiguous concepts and reasoning of varying complexity. Our own work in XAI aligns with this trend, and we have created the *Pizzaïolo Dataset*, a set of samples generated from an ontology in the form of synthetic annotated pizzas images. This dataset is designed for experimenting with the development of explainable AI. The purpose of this paper is to present the *Pizzaïolo Dataset* and to make it freely available to the research community.

RFIViz : Random Forest Interactive Visualisation, un outil simple pour comprendre les modèles

Résumé. Les forêts aléatoires (FA) constituent un outil d'apprentissage automatique largement adopté pour les tâches de classification en raison de leur facilité d'interprétation par rapport à des modèles complexes tels que les réseaux neuronaux. Cependant, dans les scénarios où les FR sont construits dans des espaces de caractéristiques étendus et des échantillons abondants, leur interprétabilité diminue au fur et à mesure que les paramètres précédents augmentent. Cette perte d'interprétabilité empêche les utilisateurs de comprendre la logique qui sous-tend les classifications d'échantillons.

Notre recherche relève ce défi en proposant un outil de visualisation pour une meilleure compréhension des modèles RF, quelle que soit leur taille, particulièrement adapté aux utilisateurs non experts. Nous reconnaissons la nécessité d'outils interactifs qui facilitent une compréhension plus approfondie du raisonnement algorithmique employé par les modèles de radiofréquence et en particulier les systèmes de filtrage et de recherche pour surmonter l'espace de caractéristiques de haute dimension.

1 Introduction

Les forêts aléatoires (RF) sont utilisées dans divers domaines d'applications comme outil de classification ou de régression. Les arbres de décision et les RF sont généralement interprétables. Cependant, avec des données à haute dimension, il peut être difficile d'examiner chaque caractéristique et chaque arbre de décision individuellement. Comme pour la plupart des applications, l'utilisateur final n'est pas toujours un expert de l'algorithme en question, d'où la nécessité d'obtenir des résultats facilement interprétables. Pour les plus initiés aux RF, la visualisation permet d'identifier les arbres les plus faibles ou les plus forts et de les traiter en conséquence.

L'objectif de ce travail est de se concentrer sur les scores et les informations pertinentes d'un modèle RF entraîné afin de comprendre et in fine d'améliorer la classification de certains échantillons. Dans cet article, nous allons présenter : (2) les travaux existant pour la visualisation des RF, (3) les tâches de l'utilisateur que l'on cherche à incorporer dans l'outil, (4) les visualisations créées pour trois niveaux différents : les

arbres de décision, la forêt et les échantillons d'entrée, (5) et pour finir l'application avec un jeu de données portant sur la réussite des étudiants.

2 État de l'art

Quelques approches de visualisation des RF ont été réalisées. Nous distinguerons ici les approches statiques et interactives. Les méthodes statiques visent à représenter graphiquement les structures, les relations et les caractéristiques des RF, permettant aux utilisateurs de mieux comprendre leur fonctionnement et leurs performances. Elles fournissent des informations précieuses sur les modèles, sans nécessiter d'interfaces utilisateur complexes et interactives.

Certaines méthodes statiques utilisent des approches de visualisation communes utilisées en statistiques telles que t-SNE ou le tracé des valeurs propres. Des travaux plus récents font appel à des outils plus complexes tels que RF-PHATE Rhodes et al. (2020) et Explanable Matrix Neto et Paulovich (2020), mais ces deux méthodes rendent difficile une visualisation détaillée de chaque arbre de la forêt.

Forest Floor est un autre outil qui permet de représenter des RF Welling et al. (2016). Il représente la structure du modèle RF par des projections géométriques, tout en mettant l'accent sur les scores importants. L'inconvénient de cette visualisation est qu'elle présente une image beaucoup trop compliquée pour les utilisateurs novices. L'encombrement et l'approche géométrique rendent difficile l'interprétation des modèles RF à haute dimension et ne constituent donc pas une solution adéquate.

Colourful Trees Nsch et al. (2019) adopte une approche plus évolutive en représentant la structure, les paramètres et les caractéristiques des arbres par des couleurs, des angles de branches et des tailles. Cela permet une vue d'ensemble facile d'un arbre donné, quelle que soit sa taille, mais ne permet pas une analyse plus approfondie des composants d'un arbre (scores, échantillons utilisés pour la construction, seuils).

Pour surmonter les inconvénients des méthodes précédentes, notamment due à la complexité et aux connaissances requises, certaines méthodes introduisent des outils dynamiques pour simplifier l'interprétation des FR. Toutefois, la visualisation RAFT Cutler et Breiman (Random Forest Tool) est obsolète et très complexe. Une autre visualisation très complexe est PaintingClass Teoh et Ma (2003). Bien que codée pour les arbres de décision, elle peut être étendue aux FR. Cet outil est adapté à la construction d'arbres et en donne un aperçu ciblé, mais le manque d'informations apparentes sur la visualisation rend difficile à la fois la compréhension d'un arbre et la prise en main de cet outil.

Enfin, iForest Zhao et al. (2019) est un outil très intuitif qui met l'accent sur les voies qui ont mené à une décision, mais qui reste compliqué pour les utilisateurs novices en raison de l'utilisation de méthodes de traçage obscures.

3 Tâches utilisateur

Après une étude exhaustive des outils actuels utilisés pour la visualisation de RF, ainsi que des outils populaires actuels utilisés pour d'autres algorithmes d'apprentissage

automatique, nous avons retenu les tâches suivantes qui nous paraissent importantes. Nous supposons que le modèle a été construit et que cet outil de visualisation est utilisé pour les données de test/validation.

3.1 Tâche de l'utilisateur 1 : Comprendre le modèle en tant qu'utilisateur non averti

Comme nous l'avons vu dans les travaux connexes, la plupart des outils de visualisation supposent une bonne connaissance de l'algorithme (RF) ou de l'environnement nécessaire à la visualisation (codage, ingénierie). En outre, l'une des principales limites de ces visualisations existantes est la complexité visuelle, en particulier lorsqu'il s'agit de représenter un grand nombre d'arbres. Le besoin d'un outil simple mais explicable est évident. il nous parait essentiel de proposer à un utilisateur de notre outil, qui cherche à classifier des données, de pouvoir parcourrir tous les arbres du modèle pour pouvir voir précisément comment chaque élément de son jeu de données a été classé, et notamment si la classification est compliquée (les arbres de la forêt ne sont pas tous d'accord entre eux) ou au contraire sans ambiguïté (les arbres de la forêt sont majoritairement unanimes).

3.2 Tâche de l'utilisateur 2 : Détecter les arbres les plus faibles

Certaines variantes des RF Zhang et Wang (2009)Yang et al. (2012)Kulkarni et Sinha (2012) nécessitent la suppression d'arbres afin d'améliorer les performances globales. Des approches utilisent des calculs complexes de seuils, mais il est essentiel de trouver un moyen rapide et visuellement simple d'identifier ces arbres. Nous souhaitons donc proposer un mode de visualisation qui permet de représenter tous les arbres de la forêt et d'immédiatement voir ceux qui sont le moins efficace. Il est également nécessaire, pour ces arbres aux performances mauvaises, de voir quelles caractéristiques y ont utilisées.

3.3 Tâche de l'utilisateur 3 : Comprendre la relation entre les caractéristiques, les prédictions et le modèle

Se débarrasser de la boîte noire des algorithmes d'apprentissage automatique est une tâche essentielle de la visualisation algorithmique moderne. L'affichage d'un flux d'information clair entre l'entrée (les caractéristiques) et la sortie (les prédictions) permet de justifier les classifications et d'ajuster éventuellement les valeurs des caractéristiques pour obtenir une classe préférée différente (cf section 5).

4 Approche proposée

Sur la base de la section précédente et des travaux antérieurs dans ce domaine, nous proposons une approche interactive principalement axée sur un utilisateur non-expert des RF. Notre outil -RFIViz- est conçu pour une compréhension rapide d'un modèle de FR et la classification de ses échantillons. Pour simplifier les RF, nous avons décidé

d'adopter une approche "Détails à la demande", en affichant les informations dans un flux en cascade pour permettre à tout utilisateur de comprendre une forêt donnée à différents niveaux. Pour surmonter la difficulté d'extraire des informations importantes (telles que l'importance des caractéristiques ou la précision de l'arbre), nous cherchons à utiliser des filtres dans les différentes scènes de visualisation de notre outil. Notre visualisation est disponible sur PyPi via "*pip install RFIViz*".

Pour faciliter l'utilisation, nous avons choisi de mettre en œuvre une visualisation à trois niveaux : une vue de l'échantillon, une vue de la forêt et une vue de l'arbre. Toutes ces vues permettent à l'utilisateur de disposer de flux d'information différents et adaptés à ses besoins.

4.1 Vue des échantillons

L'exemple de vue tel qu'il apparait dans la figure 1 (A) affiche tous les échantillons passés par le modèle et la classification de chaque échantillon par chaque arbre de la forêt.

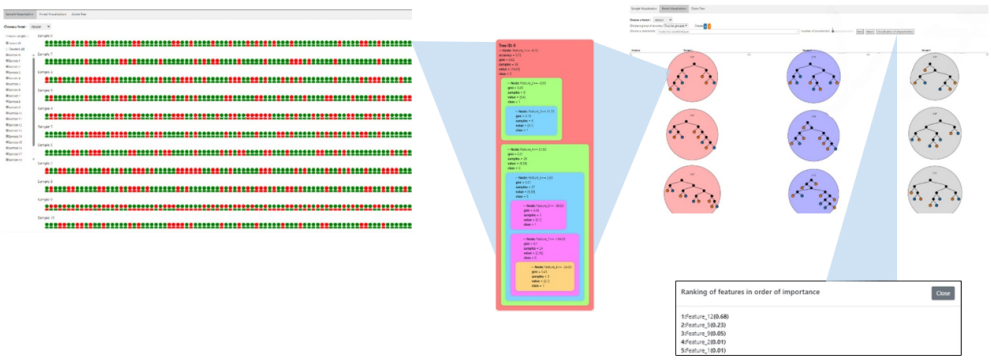


FIG. 1 – *Aperçu de l'outil de visualisation Random Forest. (A) Échantillons représentés par un cercle rempli et leurs classes prédites représentées par un rectangle rempli par rapport à leurs classes réelles. (B) Visualisation des structures arborescentes individuelles et de leurs informations, telles que la population divisée, le seuil de valeur et la classe prédite de chaque nœud. La classification de l'échantillon sélectionné peut également être affichée. (C) Une vue globale de tous les arbres construits dans le modèle forestier, leur précision correspondante et les caractéristiques utilisées. L'utilisateur peut également afficher tous les arbres utilisant les caractéristiques sélectionnées. (D) Une liste ordonnée de toutes les caractéristiques classées par ordre d'importance.*

Lorsque l'on traite des échantillons introduits dans un modèle, il est nécessaire de comprendre pourquoi certains échantillons peuvent être classés d'une manière ou d'une autre. Les utilisateurs plus avertis doivent généralement parcourir chaque arbre du modèle et afficher la classe choisie, ce qui est toujours une tâche fastidieuse. Ici, l'utilisateur peut utiliser cette vue pour des tâches multiples :

— En ce qui concerne la précision des arbres, les arbres qui prédisent rarement la bonne classe sont fréquents dans les RF, et certains travaux Yang et al. (2012) se débarrassent de ces arbres pour augmenter les performances de l'ensemble du RF ;
— Comprendre pourquoi l'échantillon peut être mal classé. Comme l'explique la figure 2, la classification incorrecte/correcte est immédiatement reconnue par le rectangle coloré, ce qui permet à l'utilisateur de passer rapidement en revue les échantillons qui sont très mal classés. L'utilisateur peut ensuite recueillir des informations supplémentaires en consultant l'arbre correspondant (voir la section 4.3). ;

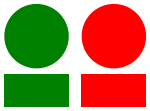


FIG. 2 – *Vue des échantillons utilisés dans le modèle, les cercles colorés représentent la classe prédite par l'arbre de la forêt (ici seulement deux couleurs dans cet exemple de classification binaire), les rectangles colorés sous les cercles indiquent si chaque arbre de la forêt a donné une bonne réponse (couleur verte) ou une mauvaise réponse (couleur rouge).*

4.2 Vue de la forêt

La vue suivante montre la forêt (Fig. 1 (C)) dans son ensemble et centre ses mécanismes sur le filtrage des caractéristiques. Chaque bulle représente un arbre de la forêt, trié horizontalement de gauche à droite en fonction de la précision de la prédiction. Une option de filtrage des caractéristiques est possible. L'utilisateur peut cliquer sur l'un des nœuds, ce qui met en évidence tous les autres nœuds de la forêt qui utilisent la même caractéristique pour diviser les échantillons, comme le montre la figure 4. Il peut également choisir d'afficher les arbres contenant certaines caractéristiques, choisies manuellement ou en fonction de leur importance. Cela permet de comprendre quelles sont les caractéristiques les plus importantes pour la prédiction. Un outil couramment utilisé dans divers progiciels RF consiste à afficher la liste des caractéristiques classées par ordre d'importance (Fig1 (D)), qui est également accessible dans cette vue. Enfin, l'utilisateur peut également sélectionner une option pour afficher les arbres en utilisant uniquement les N premières caractéristiques.

4.3 Vue de l'arbre

La dernière vue, l'arborescence (Fig. 1 (B), est une visualisation courante dans la plupart des implémentations RF. L'un des principaux avantages de notre interface est l'onglet consacré à la visualisation approfondie des informations d'un arbre. Cette fonctionnalité permet de comprendre en profondeur le fonctionnement d'un arbre spécifique, en offrant la possibilité de sonder chaque aspect de sa structure. Cette vue est construite de manière hiérarchique, chaque nœud est représenté par un rectangle, et les nœuds suivants sont construits à l'intérieur du rectangle précédent. Le nom du nœud (c'est-à-dire la caractéristique utilisée pour la division particulière) est placé en premier. Chaque nœud de l'arbre est détaillé, mettant en évidence des informations essentielles telles que :

- L'indice de Gini (score typique de notation des caractéristiques utilisé dans les RF), offrant un aperçu de l'impureté du nœud ;
- Le nombre total d'échantillons qui sont passés par ce nœud ;
- La valeur, qui indique le nombre d'échecs et de réussites à ce stade précis ;
- Le seuil calculé des valeurs des caractéristiques ;
- La classe prédite par le nœud, ce qui permet de comprendre la décision prise par le nœud sur la base des données qu'il a traitées.

De même, lorsque l'on appelle cette vue à partir de la vue Échantillon, l'arbre met en évidence le chemin choisi par l'échantillon spécifique, comme le montre la figure 3. Cela permet aux utilisateurs d'étudier plus en détail le choix de la classification pour le modèle sélectionné.

5 Scénario d'utilisation

Pour ce scénario, nous utilisons l'ensemble de données publié dans un article sur la prédiction de la réussite des étudiants Cortez et Silva (2008). Il s'agit d'une compilation de données sur les étudiants dans le but de prédire la réussite de la note de fin d'année.

Lors du test du modèle pour cet ensemble de données particulier, nous avons obtenu un taux de précision de prédiction global de 70%. Nous examinons d'abord tous les échantillons dans la première scène de la visualisation (section 4.1). Nous remarquons ici que la majorité des échantillons présentent des cercles et des rectangles verts (représentant la classe de réussite correctement prédite), mais lorsqu'il y a un plus grand nombre de cercles rouges (classe d'échec), il y a également un plus grand nombre de rectangles rouges et verts (prédictions incorrectes et correctes). Cela indique que la forêt a tendance à classer facilement les élèves qui réussissent, mais qu'il y a plus de divergences pour les élèves qui échouent.

Nous avons choisi d'examiner un autre aspect, un échantillon 155 (voir Fig. 5, un échantillon est un étudiant) est classé avec précision pour la plupart des arbres, à l'exception d'un seul. D'un point de vue plus fonctionnel, étant donné le contexte de l'ensemble de données, cet élève peut être en danger d'échec en fonction de la raison pour laquelle l'arbre de décision a mal classé cet échantillon. Dans ce cas, et pour cet arbre, l'élève est considéré comme "en échec" en raison de la caractéristique

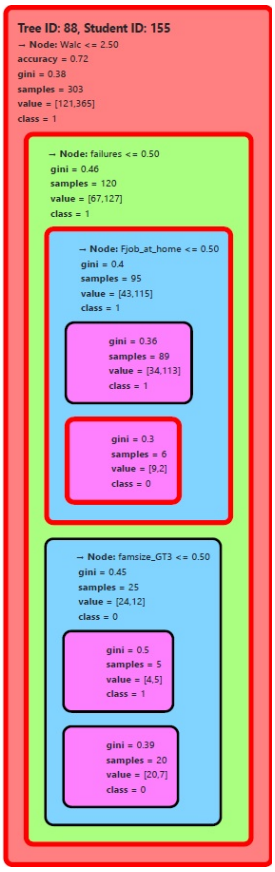


FIG. 3 – *Vue d'un arbre singulier lorsqu'un échantillon a été préalablement sélectionné. Le chemin de l'échantillon classé est représenté par un rectangle rouge entourant les rectangles des nœuds.*

"Fjob-at-home" (le parent n'a pas d'emploi). L'utilisateur peut alors décider si cette caractéristique favorise ou non la réussite de l'élève.

Une autre tâche consisterait à examiner les caractéristiques les plus importantes. En examinant la liste d'importance des caractéristiques (voir Fig.1 (D)), la caractéristique "G2" apparaît comme la plus importante pour les divisions. Après un examen plus approfondi, en sélectionnant la caractéristique et en mettant en évidence sa présence dans tous les autres arbres, nous remarquons que la plupart des arbres dont la précision est supérieure à 80% utilisent cette caractéristique pour les divisions (voir la Fig.4).

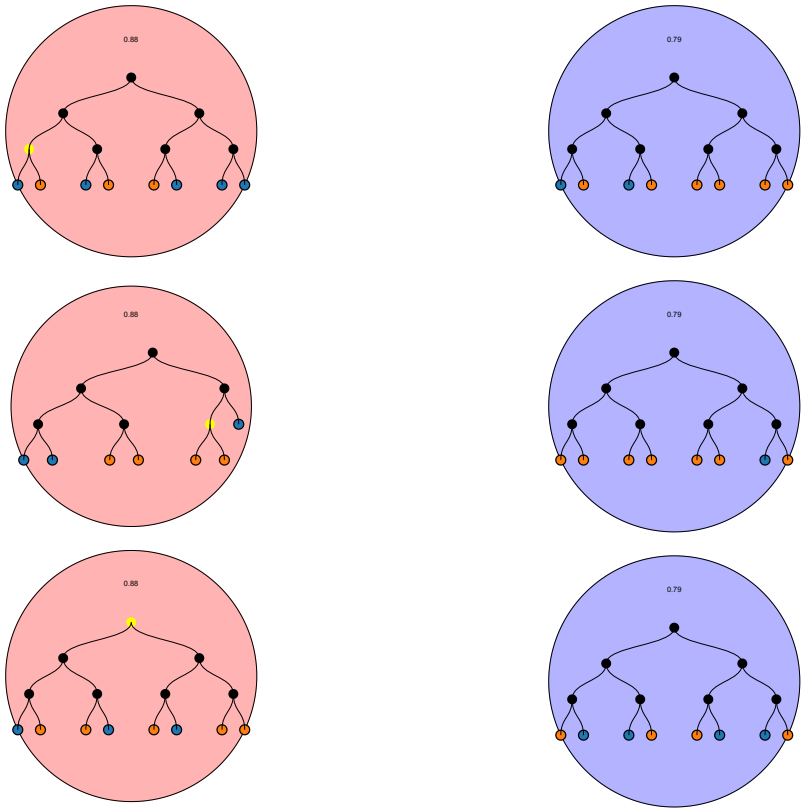


FIG. 4 – *Vue globale de la forêt (seuls quelques arbres sont représentés ici). À gauche, les arbres dont la précision est supérieure à 80% et à droite, les arbres dont la précision est comprise entre 60% et 80%. L'utilisateur a sélectionné une caractéristique, qui est surlignée en jaune lorsqu'elle est présente.*

6 Discussion et travaux futurs

Les travaux futurs permettront d'afficher davantage de valeurs telles que les scores f, l'entropie, afin que l'utilisateur puisse avoir une meilleure idée de la performance du modèle sélectionné. La visualisation sera également étendue à d'autres types de méthodes d'ensemble d'arbres de décision : Extremely Randomized Trees Geurts et al. (2006), XGBoost Chen et Guestrin (2016).

Nous prévoyons également d'ajouter une connexion entre la visualisation et l'instanciation RF, permettant à l'utilisateur d'interagir directement avec le modèle construit. Cela permettrait de mettre en œuvre Yang et al. (2012) et d'étendre la tâche de l'utilisateur 3.3 en interagissant avec l'entrée et le modèle sur la base d'un résultat requis. Un autre problème souvent rencontré avec les approches interactives est le temps de

calcul, en particulier pour la construction de la vue de la forêt (en raison de la création des nœuds et des branches). Nous cherchons à améliorer ce point en réduisant le temps de latence du client lors de la génération des éléments web.

Enfin, la visualisation résout les tâches de l'utilisateur énumérées dans la section 3, d'autres tâches de l'utilisateur ont été traitées dans d'autres outils tels que Zhao et al. (2019). Une compilation des différentes tâches des RF existantes permettra d'obtenir un outil plus complet. Cependant, notre objectif est toujours de faire en sorte que RFIViz soit destiné à des utilisateurs non experts, c'est pourquoi il faut encore travailler à la simplification des flux d'information.

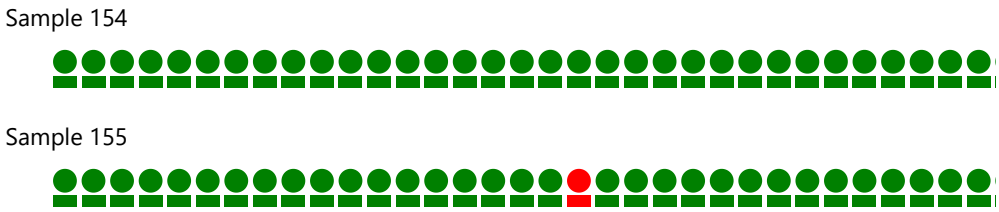


FIG. 5 – *Prédiction sur 2 échantillons de l'ensemble de données Cortez et Silva (2008). Un échantillon est correctement prédit par tous les arbres, l'autre est incorrectement prédit par tous les arbres sauf un.*

7 Conclusion

Dans cette étude, nous avons présenté RFIViz, un outil de visualisation interactif et convivial conçu pour combler le fossé entre la complexité des modèles de forêt aléatoire et la capacité d'interprétation de l'utilisateur. En combinant les connaissances des méthodes existantes et les besoins des utilisateurs, RFIViz offre une approche "Détails à la demande" à travers trois vues clés - échantillon, forêt et arbre - permettant une compréhension complète à des niveaux de profondeur variables.

RFIViz simplifie les subtilités des modèles complexes, tant pour les experts que pour les novices, en aidant à identifier les arbres faibles ayant un impact sur les performances globales et en démêlant la nature de la boîte noire des algorithmes d'apprentissage automatique.

L'application à un ensemble de données sur la réussite des étudiants a mis en évidence les prouesses de RFIViz dans l'identification des modèles de classification, l'offre d'informations exploitables et l'aide à la prise de décision.

Les améliorations futures incluront de nouvelles mesures, l'extension à d'autres méthodes d'ensemble, et une intégration plus profonde avec l'instanciation du modèle, réaffirmant notre engagement à affiner RFIViz en tant qu'outil accessible et complet pour l'interprétation des modèles.

Références

Chen, T. et C. Guestrin (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Cortez, P. et A. Silva (2008). Using data mining to predict secondary school student performance. *EUROSIS*.

Cutler, A. et L. Breiman. Raft (random forest tool).

Geurts, P., D. Ernst, et L. Wehenkel (2006). Extremely randomized trees. *Mach. Learn. 63*(1), 3–42.

Kulkarni, V. Y. et P. K. Sinha (2012). Pruning of random forest classifiers : A survey and future directions. In *2012 International Conference on Data Science Engineering (ICDSE)*, pp. 64–68.

Neto, M. et F. Paulovich (2020). Explainable matrix—visualization for global and local interpretability of random forest classification ensembles. *27*(2), 1427–1437.

Nsch, R. H., P. Wiesner, S. Wendler, et O. Hellwich (2019). Colorful trees : Visualizing random forests for analysis and interpretation. pp. 294–302.

Rhodes, J. S., A. Cutler, G. Wolf, et K. R. Moon (2020). Supervised visualization for data exploration. *ArXiv abs/2006.08701*.

Teoh, S. et K. Ma (2003). PaintingClass : Interactive construction, visualization and exploration of decision trees. pp. 667–672.

Welling, S., H. Refsgaard, P. Brockhoff, et L. Clemmensen (2016). Forest floor visualizations of random forests.

Yang, F., W. hang Lu, L. kai Luo, et T. Li (2012). Margin optimization based pruning for random forest. *Neurocomputing 94*, 54–63.

Zhang, H. et M. Wang (2009). Search for the smallest random forest. *Statistics and its interface 2*, 381.

Zhao, X., Y. Wu, et D. Lee (2019). iForest : Interpreting random forests via visual analytics. *25*(1), 407–416.

Summary

Random Forests (RFs) stand as a widely adopted machine learning tool for classification tasks due to their interpretability relative to complex models like neural networks. However, in scenarios where RFs are constructed within expansive feature spaces and abundant samples, their interpretability diminishes as the previous parameters grow. This loss of interpretability impedes users from comprehending the rationale behind sample classifications. Our research addresses this challenge by proposing a visualisation tool for better understanding of RF models whatever their size, particularly tailored for non-expert users. We recognize the necessity for interactive tools that facilitate a deeper grasp of the algorithmic reasoning employed by RF models and especially filter and find systems to overcome the high dimensional feature space.

Utilization of Autoencoder and Bayesian Network for Anomaly Detection on Wind Turbine Main Bearings

Ahmed MABROUK*,

*4 Rue Joséphine Baker
ahmed.mabrouk@engie.com,

Résumé. This study addresses challenges in wind turbine operation, with a focus on main bearing anomalies. Leveraging SCADA data, we construct an autoencoder model to detect these anomalies, and a probabilistic graphical model (or Bayesian network) to explore relationships among system components and explain failure causes. The approach demonstrates robustness through numerical experiments, providing advantages over traditional fault detection methods

1 Introduction

The growing demand for renewable energy, particularly from wind turbines, brings control and maintenance challenges due to harsh operating conditions, making these turbines susceptible to various faults. Unscheduled maintenance resulting from unexpected faults can incur significant costs. To enhance wind energy performance and reduce maintenance costs, a thorough evaluation of wind turbine component reliability throughout its lifecycle is imperative. In the last decade, numerous efforts have been devoted to overcome the limitation of current maintenance strategies, which are mainly based on preventive and corrective maintenance. These approaches allow real-time monitoring rather than costly time interval intervention and performance tune-ups. Feedback from various industries shows that condition monitoring techniques can detect anomalies before they turn into system-critical faults, allowing therefore maintenance to be well scheduled. In the wind turbine industry, various condition monitoring techniques are utilized, such as acoustic monitoring Bouno et al. (2005), oil monitoring Zhu et al. (2013), thermal monitoring Guo et al. (2011), etc. The main drawbacks of these methods are the high cost related to the installation of custom sensors and the complex communication infrastructure and protocols needed to manage their data. To tackle these issues, several data-driven approaches based on SCADA data have been proposed. For example, in Encalada-Dávila et al. (2021), an unsupervised approach for predicting main bearing faults is introduced. This method exclusively uses healthy SCADA data, employing an artificial neural network to forecast the low-speed shaft temperature. An anomaly detection threshold is established based on residuals between predicted and actual values. Another approach, discussed in Tutivén et al. (2022), suggests a semi-supervised based on a one-class support vector machine classifier to identify anomalies in the main bearing. However, existing unsupervised anomaly detection methods, such as those presented in Campoverde-Vilela et al. (2023); Li et al. (2020); Zhao et al. (2018), lack the capability to pinpoint the sensor observations responsible for failures—a critical aspect in planning diagnostic and maintenance actions. A supervised machine learning

approach proposed in Elasha et al. (2019) focuses on detecting gearbox failures. In this case, the training process requires labeled historical data, a time-consuming and error-prone task. Additionally, the highly imbalanced nature of gearbox fault observations in SCADA data may pose performance challenges.

In this study, we shall focus our discussion on the main bearing component. Therefore, a special attention should be given to this component in order to prevent severe consequences. We propose a two-phase anomaly detection algorithm. The first phase aims to train an autoencoder model using only safe SCADA data. In the second phase, we propose an algorithm based on the Bayesian network (BN) model to interpret and explain the anomaly results returned by the autoencoder. In this context, beyond the wish of using the BN to identify the main sources of the anomaly, there is also the will to understand and validate the already detected failure in the main-bearing component. The rest of the paper is organized as follows : Section 2 introduces the wind turbine system. Then, we discuss the used SCADA data in section 3. In section 4, we present a new approach based on an autoencoder and a BN for detecting and explaining main bearing failures. Its effectiveness is proven through experiments in section 5. Finally, some concluding remarks are given.

2 Wind turbines overview

A wind turbine is a device that converts the kinetic energy of wind into mechanical energy. The mechanical energy is then transformed into electrical energy by a generator. The electrical energy produced by a wind turbine depends mainly on three factors : the design and size of the blades, wind speed, and outdoor temperature, which directly affect the density of the air. The wind power P (in Watts) is computed as follows : $P = \frac{1}{2}A\rho w^3$, where A is the rotor swept area in m^2, ρ is the air density in Kg/m^3 and w is the wind speed in m/s. The wind turbine consists of three major parts : a tower, a nacelle, and rotor blades. It is also equipped with a pitch control system to reduce failures while improving safety and reliability. In fact, the pitch control is used to adjust the angle of the blades by rotating them to achieve specific rotor speeds and power output. Moreover, it serves as a protective mechanism by ensuring the safety of the wind turbine during high winds, loss of electrical load, or other extreme conditions.

The wind farm SCADA data provides a rich source of continuous-time observations about different components of the system, as well as the environmental conditions. Together with the pitch system, the SCADA data can be used to ensure wind turbine performance.

3 Used data

The SCADA data in this study offers continuous-time observations on various condition variables from wind turbines located in a northeast France wind farm. Datasets include environmental, electrical, component temperature, hydraulic, and control variables. Environmental variables like ambient temperature, nacelle temperature, wind speed, and turbulence index are correlated with electrical and component temperature variables. Wind speed notably influences subsystems ; below rated speed, higher wind speeds increase rotor speed, output power, and component temperatures. Electrical variables cover active power, phase voltage, power factor, reactive power, and electric network frequency, impacting power generation and efficiency.

Component temperature variables focus on key nacelle locations, with a primary emphasis on detecting anomalies in the main bearing. Hydraulic variables describe observations of the general accumulator, brake pressure, blade pressure, and hydraulic group pressure. Control variables ensure safe operation, power optimization, and structural longevity, including blade pitch, yaw, rotor and generator speeds. In addition to SCADA data, maintenance intervention reports provide further insights. These information are crucial for model training.

4 Proposed approach

The main bearing failure detection pipeline is structured around four essential steps. Initially, the SCADA data is preprocessed and filtered. Subsequently, the chosen data is partitioned into training, validating, and testing datasets. In the third step, normality models for each wind turbine in the wind farm are constructed using two complementary models : autoencoder and Bayesian network. Finally, a novel method for anomaly detection and explanation is introduced.

4.1 SCADA data process

The SCADA dataset used in this study includes a large number of parameters with varying outliers, incomplete observations, and mismatches in time and date stamps. Therefore, preprocessing operations are crucial to obtain a clean dataset that is suitable for machine learning algorithms. The initial step involves selecting a set of variables crucial for model training, emphasizing features that offer valuable insights into main bearing behaviors. Domain experts recommend utilizing mean values of environmental measurements, rotor speed, and main bearing temperatures. To address strong correlations among component states in wind turbine systems, exogenous variables are preferred. The dataset is enhanced by generating new features related to the median of main bearing temperatures across the wind farm and the difference between this median and the current temperature of the main bearing under investigation. A substantial deviation between these temperatures signals a potential main bearing failure. After the feature selection process, a data cleaning pre-processing step is executed to eliminate outliers and sensor measurement errors in the selected features. These abnormal observations do not contain valuable guidance in detecting anomalies. In our study, the cleaning process is initiated by removing out-of-range values for each selected variable. The second step employs the quartile method to eliminate any remaining outliers. This involves sorting and dividing wind speed values into small, regularly spaced intervals. The lower and upper quartiles (Q_1 and Q_3) are computed, and the interquartile range (IQR) is determined as $IQR = Q_3 - Q_1$. Outliers in a wind speed interval are identified by comparing each value (x_i) with $Q_1 - 1.5IQR$ (considered too small) or $Q_3 + 1.5IQR$ (considered too large). After identifying the outliers, they are treated as missing values and filled using a cubic Hermite interpolating polynomial (CHP) Lu et al. (2018). In situations where the dataset contains missing values at the boundaries, we suggest the use of the nearest available values before or after the missing values. Finally, we deal with the very different magnitudes of the selected features. For a fair training process, every variable X is scaled using the min-max normalization process : $x_i^{'} = (x_i - min(X))/(max(X) - min(X))$. Carrying out these preprocessing procedures, we end up with an effective SCADA data quality that can be used by machine learning algorithms.

4.2 SCADA data splitting

Selecting datasets for training, validation, and testing is a critical step in developing an accurate main-bearing normality model. It is important to consider various operational and environmental conditions, including different wind velocities and seasonal variations. For training and validating we use SCADA data without main bearing failures. The model has been also tested on data containing main bearing anomaly observations. Testing the model against these diverse observations ensures its ability to accurately distinguish between normal and abnormal conditions, irrespective of season and environmental factors.

4.3 Train an Auto-encoder

The Autoencoder (AE) is an unsupervised and symmetrical neural network used for diagnostic tasks, providing a flexible model capable of representing complex functions without labeled data. The AE architecture consists of an encoder (f) that compresses input data ($\mathbf{x}_i$) into a latent representation ($\mathbf{h}$) and a decoder (g) that maps the latent representation back to the original data ($\tilde{\mathbf{x}}_i$). In our case, we used an undercomplete AE architecture for main bearing anomaly detection. Model parameters are estimated by minimizing the reconstruction error (RE) using mean square error (MSE). The five-layer architecture with layers of 10, 6, 3, 6, and 10 neurons is optimized through Bayesian hyper-parameter optimization. The Keras implementation of the adam optimizer, ELU activation function, learning rate of 0.001, 20 epochs, and a mini-batch size of 64 samples are used.

Although the AE is often effective in detecting anomalies in SCADA data, it is nevertheless unable to reveal which features are the root causes of the anomaly. Roughly speaking, the deviation of only one of the considered input features is enough to cause the propagation of the error through the AE network, resulting in a significant reconstruction error in most other features as well. To address this issue, a Bayesian network (BN) is used to identify features responsible for anomalies. The BN is probabilistic graphical model that represents a high-dimensional distribution over complex systems using a graph-based structure.

4.4 Train a Bayesian network

To address concerns about false positive alarms triggered by the deviation of sensors unrelated to the main bearing component, we propose an anomaly explanation model based on the Bayesian network (BN) model Pearl (1988). In contrast to autoencoders, BN is a probabilistic graphical model that uses a graph-based representation to compactly encode high-dimensional distributions in complex systems.

A Bayesian network is defined as a pair $(\mathbf{G}, \Theta)$, where $\mathbf{G} = (\mathbf{V}, \mathcal{A})$ is a directed acyclic graph (DAG), $\mathbf{V}$ is a set of random variables, $\mathcal{A}$ is a set of arcs, and $\Theta = \theta_{X_i|\mathbf{Pa}(X_i)} X_i \in \mathbf{V}$ represents the conditional probability distributions (CPD) of nodes X_i given their parents $\mathbf{Pa}(X_i)$. The joint probability over $\mathbf{V}$ is encoded as $P(\mathbf{V}) = \prod_{X_i \in \mathbf{V}} P(X_i|\mathbf{Pa}(X_i))$.

BNs encode an independence model through their graphical structure, characterized by the *d-separation* property Pearl (1988). Several approaches for learning the BN graph have been proposed in the literature. These algorithms can be divided into 3 classes : i) the search-based approaches that focus on optimizing the scoring function of the structure Heckerman et al. (1995) ; ii) the constraint-based approaches that exploit statistical independence tests to find the

best structure Spirtes et Glymour (1991) ; iii) the hybrid methods that exploit a combination of both Tsamardinos et al. (2006). However, these methods may not always identify all causal directions in the graph. To overcome this, we adopt an alternative approach that combines expert knowledge with a score-based approach for causal discovery optimization. Experts provide knowledge about causal relations between variables, expressed as hard structural constraints (the BN learning algorithm must not modify). Expert knowledge is also utilized after learning to adjust causal relations. The proposed failure detection method integrates both AE and BN models for anomaly detection and explanation.

4.5 Anomaly detection and explanation using AE and BN models

The initial step involves utilizing the Autoencoder (AE) to identify anomalies in SCADA observations. This detection relies on the reconstruction error (RE), calculated as the absolute difference between the reconstructed and original data. The threshold ϵ is set based on the distribution of RE values computed using the validation dataset. In this approach, observations are deemed anomalous if the reconstruction error for a given input set surpasses three times the standard deviation above the mean of computed REs. After detecting the anomaly using the AE, we rely on the inference engine of the BN to perform a detailed analysis of the sensors that are responsible for the anomaly.

To perform variable analysis, the conditional probability query is used. The probability of a query variable $Q = q$ given evidence $\mathbf{E} = \mathbf{e}$ is calculated as follows :

$$P(Q = q|\mathbf{E} = \mathbf{e}) = \frac{P(Q = q, \mathbf{E} = \mathbf{e})}{P(\mathbf{E} = \mathbf{e})} \tag{1}$$

During variable analysis, we assume that the evidence includes only safe measurements. In the context of wind turbine anomaly diagnosis, environmental variables seem to be a good choice for this purpose, as their observations are unaffected by anomalies in the system, and can be easily collected and verified. We also follow variables ordering within the causal Bayesian network during diagnosis, where cause variables are examined before consequences.

We propose to use a probabilistic framework to locate the conditional probability associated with every sensor observation q_i of a sensor Q w.r.t. the theoretical $P(Q|\mathbf{E} = \mathbf{e})$ calculated using the BN [1]. For a given observation q_i of Q and the computed posterior $P(Q|\mathbf{E} = \mathbf{e})$, we collect all posterior value's probabilities that are less or equal to $P(Q = q_i|\mathbf{E} = \mathbf{e})$: $F(q_i) = \{q_j \in \Omega_Q : P(Q = q_j|\mathbf{E} = \mathbf{e}) \leq P(Q = q_i|\mathbf{E} = \mathbf{e})\}$.
Given the set of elements in $F(q_i)$ we calculate the probability $G(q_i)$ which is correspond to the probability that $P(Q|\mathbf{E} = \mathbf{e})$ being less or equal to $P(Q = q_i|\mathbf{E} = \mathbf{e})$:

$$G(q_i) = \sum_{q_j \in F(q_i)} P(Q = q_j|\mathbf{E} = \mathbf{e}) \tag{2}$$

If $G(q_i) \leq \tau$, where τ is set to 0.3, then the observation q_i is classified as anomalous (or unusual).

1. This posterior distribution represents the distribution of sensor Q given environment variable values (wind speed, outdoor temperature, etc.) in normal functioning conditions of the main bearing

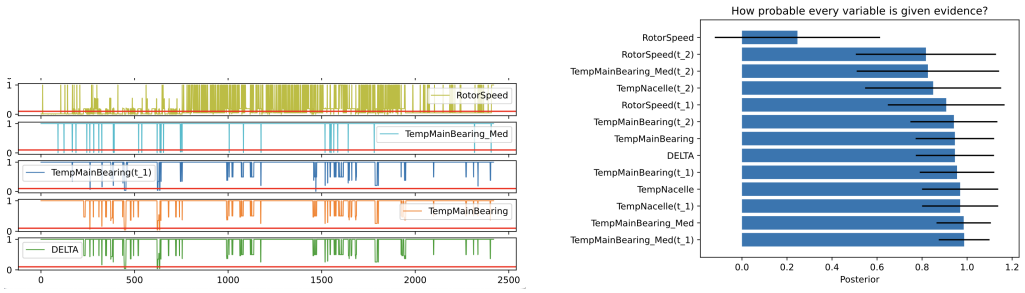
Dataset (WT)	Train	Validation	Test
$\mathbf{WT}_1$	0.99	0.95	**0.88**
$\mathbf{WT}_2$	0.84	0.85	0.85
$\mathbf{WT}_3$	0.85	0.86	0.86
$\mathbf{WT}_4$	0.85	0.85	0.87
$\mathbf{WT}_5$	0.86	0.85	0.86

(a)

Method	Safe (MB)	Anomaly (MB)	AUC
Isolation Forest	0.79	0.69	0.71
LocalOutlierFactor	**0.98**	0.11	0.55
Gradient Boosting	0.82	**0.91**	0.82
Auto-encoder	0.95	0.80	**0.87**

(b)

FIG. 1 – *(a) Anomaly detection in the studied wind farm (b) Comparison of several approaches*



(a) (b)

FIG. 2 – *An example of false positive anomaly explanation : data with a safe main bearing, and added noise to the rotor speed (a,b).*

5 Experiments

We conducted a comparison using SCADA data from a wind farm in France. The wind farm contains 5 wind turbines. According to the maintenance report, an anomaly on the main bearing appeared in wind turbine 1 ($\mathbf{WT}_1$) in 2014. In our experimentation, data from 2011, 2012, and 2014 have been respectively used for training, validating, and testing. Table 1 illustrates the results of the auto-encoder and compares them with other machine-learning models. As can be seen, the use of the autoencoder outperforms the results of other models.

To prove the efficiency of using BN models, we consider a situation where the data includes a healthy main bearing but introduces noise in the rotor speed observations (which is not our focus component).

In this instance, even though the autoencoder (AE) suggests an anomaly in the main bearing, the BN clearly indicates that this anomaly is not related to the main bearing sensor, but instead to the rotor speed (see Fig. 2.(a),(b)). Then, the use of the BN helps to prevent false positive alarms in this scenario, this demonstrating the effectiveness of our approach in detecting the false positive anomalies on the main bearing.

6 Conclusion

The paper introduces a novel method for detecting and explaining main bearing anomalies, utilizing autoencoder (AE) and Bayesian network (BN) models. The approach effectively

combines these heterogeneous models, leveraging BN properties to enhance anomaly interpretation. The experimental results demonstrate a high correctness score with the SCADA data. Future work aims to develop a generic approach capable of addressing multiple anomalies simultaneously and incorporating inherent uncertainty in SCADA data through a variational autoencoder.

Références

Bouno, T., T. Yuji, T. Hamada, et T. Hideaki (2005). Failure forecast diagnosis of small wind turbine using acoustic emission sensor. *KIEE International Transaction on Electrical Machinery and Energy Conversion Systems 5*(1), 78–83.

Campoverde-Vilela, L., M. C. Feijóo, Y. Vidal, J. Sampietro, et C. Tutivén (2023). Anomaly-based fault detection in wind turbine main bearings. *Wind Energy Science Discussions 2023*, 1–29.

Elasha, F., S. Shanbr, X. Li, et D. Mba (2019). Prognosis of a wind turbine gearbox bearing using supervised machine learning. *Sensors 19*(14).

Encalada-Dávila, , B. Puruncajas, C. Tutivén, et Y. Vidal (2021). Wind turbine main bearing fault prognosis based solely on scada data. *Sensors 21*(6).

Guo, P., D. Infield, et X. Yang (2011). Wind turbine generator condition-monitoring using temperature trend analysis. *IEEE Transactions on sustainable energy 3*(1), 124–133.

Heckerman, D., D. Geiger, et D. M. Chickering (1995). Learning bayesian networks : The combination of knowledge and statistical data. *Machine learning 20*, 197–243.

Li, M., S. Wang, S. Fang, et J. Zhao (2020). Anomaly detection of wind turbines based on deep small-world neural network. *Applied Sciences 10*(4), 1243.

Lu, S., Y. Wang, et Y. Wu (2018). Novel high-precision simulation technology for high-dynamics signal simulators based on piecewise hermite cubic interpolation. *IEEE Transactions on Aerospace and Electronic Systems 54*(5), 2304–2317.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan kaufmann.

Spirtes, P. et C. Glymour (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review 9*(1), 62–72.

Tsamardinos, I., L. E. Brown, et C. F. Aliferis (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning 65*, 31–78.

Tutivén, C., Y. Vidal, A. Insuasty, L. Campoverde-Vilela, et W. Achicanoy (2022). Early fault diagnosis strategy for wt main bearings based on scada data and one-class svm. *Energies 15*(12), 4381.

Zhao, H., H. Liu, W. Hu, et X. Yan (2018). Anomaly detection and fault analysis of wind turbine components based on deep learning network. *Renewable energy 127*, 825–834.

Zhu, J., J. Yoon, D. He, B. Qiu, et E. Bechhoefer (2013). Online condition monitoring and remaining useful life prediction of particle contaminated lubrication oil. *2013 IEEE Conference on Prognostics and Health Management (PHM)*, 1–14.

Xplique
A Deep Learning Explainability Toolbox

Thomas Fel[1,3,4*] Lucas Hervier[2*] **Antonin Poche**[2]
David Vigouroux[2] Justin Plakoo[2] Remi Cadene[3] Mathieu Chalvidal[1,3]
Julien Colin[1,3] Thibaut Boissin[1,2] Louis Bethune[1] Agustin Picard[1,2] Fréderic Boisnard[1,5] Claire Nicodeme[4]
Laurent Gardes[4] Gregory Flandin[1,2] Thomas Serre[1,3]

[1]Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
[2] Institut de Recherche Technologique Saint-Exupery, France
[3]Carney Institute for Brain Science, Brown University, USA
[4] Innovation & Research Division, SNCF , [5] Renault AMPERE

Abstract

Les modèles d'apprentissage automatique les plus avancés d'aujourd'hui sont faiblement compréhensibles. Le défi principal des méthodes d'explicabilité est d'aider les chercheurs à ouvrir ces boîtes noires en révélant la stratégie ayant conduit à une décision donnée, en caractérisant leurs états internes ou en étudiant la représentation sous-jacente des données. Pour relever ce défi, nous avons développé ***Xplique****: une bibliothèque logicielle d'explicabilité qui inclut des méthodes d'explication caractéristiques ainsi que des métriques d'évaluation associées. Elle s'interface avec les bibliothèques d'apprentissage les plus populaires : Tensorflow et PyTorch, ainsi qu'avec d'autres bibliothèques telles que scikit-learn et Theano. Le code est sous licence MIT et est disponible gratuitement sur* `github.com/deel-ai/xplique`*."*

Cet article a initialement été publié dans un workshop CVPR en 2022 [17].

1. Introduction

Deep neural networks [35, 50] are widely used in many applications including medicine, transportation, security, and finance, with broad societal implications [7, 30, 44]. Yet, these networks have become almost impenetrable. Furthermore, in most real-world scenarios, these systems are used to make critical decisions, often without any explanation. A growing body of research thus focuses on making those systems more trustworthy via the development of explainability methods to make their predictions more interpretable [11]. Such methods will find broad societal uses and will help to fulfill the "right to explanation" that European laws guarantee to its citizens [28]. Hence, it is important for explainability methods to be made widely available. Indeed, several libraries have already been proposed including Captum [32] for Pytorch.

In this work, we propose the first of such libraries – based on Tensorflow [1]. Our library includes all main explainability approaches including (1) attribution methods (and their associated metrics), (2) feature visualization methods, and (3) concept-based methods.

1.1. Attribution methods

These techniques aim to produce so-called saliency maps or more simply, heatmaps, to explain models' decisions. These maps reveal the discriminating input variables used by the system to arrive at a given decision. The score assigned to a region of an image (or a word in a sentence) reflects its importance for the prediction of the model. We have re-implemented more than 16 representative explanation methods [2, 9, 15, 16, 20, 37, 38, 41, 43, 45, 48, 49, 51, 53, 54, 58–60]. We provide support for different data types such as images, tabular data, and time series as well different tasks, including classification, regression, object detection, and semantic segmentation. As one can imagine, the large number of explanation methods available has brought to the forefront a major issue: the urgent need for metrics to evaluate explanations. Indeed, inconsistencies produced across these methods have raised questions about their legitimacy [2,3,6,8,10,19,23,24,26,27,34,36,46,52,55,56]. Our implementation thus also includes several common metrics associated with these attribution methods.

1.2. Feature Visualization

Even though attribution methods are sometimes useful to understand a decision, they leave aside the global study of

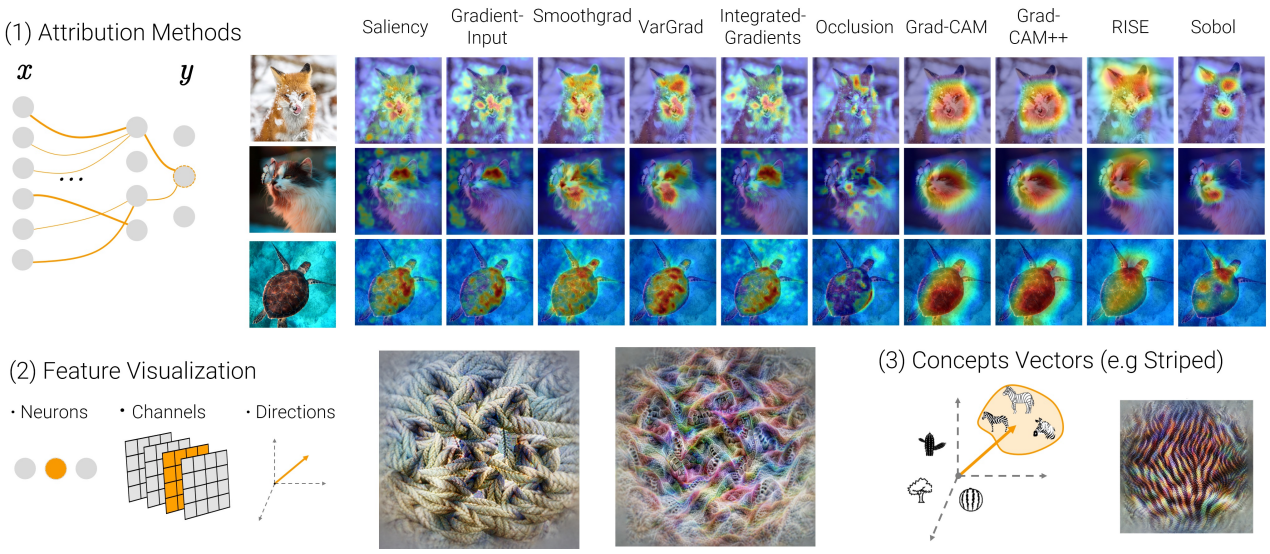


Figure 1. **Xplique modules.** The library contains 3 main modules: **(1)** an "Attribution Methods" module, **(2)** a "Feature Visualization" module and **(3)** a "Concepts" module .

a Deep Learning model. Several methods attempt to tackle this issue including feature visualization methods for studying the internal representations learned by a model.

The method proposed in [39, 40, 42] is a popular technique employed to explain the internal representations of a model. This method aims to find an interpretable input (or stimulus) that maximizes the response of a given neuron, a set of neurons (e.g., a channel), or a direction in an internal space of the model. Thus, the corresponding stimulus is a prototype of what the neuron responds to. We provide an API able to optimize such input by targeting a layer, a channel, a direction, or combinations of these objectives. The optimization tool leverages the latest advances in the field (e.g., Fourier preconditioning, robustness to transformations). This method was further improved in [13], allowing it to scale to larger and more modern neural networks (e.g. ViT [12], FlexiViT [5], BeiT [4]).

1.3. Concept-based methods

Nevertheless, the interpretation of feature visualization methods is left to the user. Fortunately, another approach consists of letting the user derive concept vectors that are meaningful to them: Concept-based methods.

[21, 22, 25, 29, 31, 47, 57] work on high-level features interpretable by humans. This includes a method to retrieve Vectors of Activations of these human Concepts (CAV) [29]. These vectors help to make the passage between human concepts and a vector base formed by the neurons of a model at a specific layer. In addition, we have also re-implemented TCAV, which then tests how important these human vectors are to the model's decisions.

The library also proposes unsupervised concept extraction with Craft [14, 18], abstracting the methods from the dependence on concepts that are manually defined by humans (see Fig. 2).

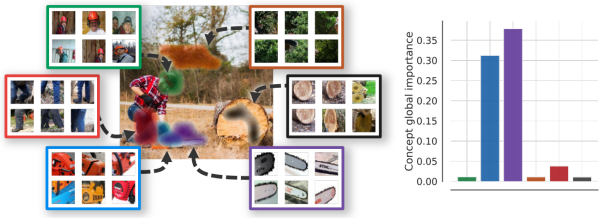


Figure 2. **Craft.** Example of concepts detected for the "chainsaw" class from ImageNet [33] as presented in [18].

Finally, the library also allows interactions between all 3 modules such that one can leverage the feature visualization module to visualize the extracted CAV (see Fig. 1) or the feature attribution module to visualize the location of the CAV on an image. A major effort has been made to facilitate the use of the software and various examples are provided as notebooks for each of the modules.

2. Acknowledgments

This work was conducted as part of the DEEL project[1]. Funding was provided by ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). Additional support provided by ONR grant N00014-19-1-2029 and NSF grant IIS-1912280. Support for computing hardware provided by Google via the TensorFlow Research Cloud (TFRC) program and by the Center for Computation

[1]https://www.deel.ai/

and Visualization (CCV) at Brown University (NIH grant S10OD025181).

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. 1

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 1

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[5] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023. 2

[6] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 1

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 2018. 1

[8] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 2019. 1

[9] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018. 1

[10] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[11] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[13] Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Laurent Gardes, et al. Unlocking feature visualization for deeper networks with magnitude constrained optimization. 2023. 2

[14] Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Mathieu Chalvidal, Thomas Serre, et al. A holistic approach to unifying automatic concept extraction and concept importance estimation. 2023. 2

[15] Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[16] Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. Don't lie to me! robust and efficient explainability with verified perturbation analysis. *Workshop on Formal Verification of Machine Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2022. 1

[17] Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Bethune, Agustin Picard, Claire Nicodeme, Laurent Gardes, Gregory Flandin, and Thomas Serre. Xplique: A deep learning explainability toolbox. *Workshop on Explainable Artificial Intelligence for Computer Vision (CVPR)*, 2022. 1

[18] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[19] Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 1

[20] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[21] Jessica Zosa Forde, Charles Lovering, George Konidaris, Ellie Pavlick, and Michael L Littman. Where, when & which concepts does alphazero learn? lessons from the game of hex. In *AAAI Workshop on Reinforcement Learning in Games*, 2022. 2

[22] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W Picard. Dissect: Disentan-

gled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*, 2021. 2

[23] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 1

[24] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the IEEE International Conference on data science and advanced analytics (DSAA)*, 2018. 1

[25] P Hitzler and MK Sarker. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342:337, 2022. 2

[26] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[27] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and methods for explanation through robustness analysis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[28] Margot E Kaminski. The right to explanation, explained. In *Research Handbook on Information Law and Governance*. Edward Elgar Publishing, 2021. 1

[29] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. Proceedings of the International Conference on Machine Learning (ICML), 2018. 2

[30] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. 1

[31] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2

[32] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020. 1

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2

[34] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. In *Workshop on Correcting and Critiquing Trends in Machine Learning, Advances in Neural Information Processing Systems (NIPS)*, 2019. 1

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. 1

[36] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 1

[37] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 1

[38] Sabine Muzellec, Leo Andeol, Thomas Fel, Rufin VanRullen, and Thomas Serre. Gradient strikes back: How filtering out high frequencies improves explanations, 2023. 1

[39] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *Visualization for Deep Learning workshop, Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2

[40] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. *arXiv preprint arXiv:1904.08939*, 2019. 2

[41] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[42] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. 2

[43] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1

[44] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019. 1

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016. 1

[46] Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. In *Workshop, Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1

[47] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv e-prints*, pages arXiv–2106, 2021. 2

[48] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[49] Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order

partial derivative. In *Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 1

[50] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual review of vision science*, 2019. 1

[51] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1

[52] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 1

[53] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1

[54] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1

[55] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1

[56] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[57] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. 2

[58] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014. 1

[59] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014. 1

[60] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011. 1