Effet de la complexité du réseau LSTM sur l'explicabilité en Maintenance Prédictive

Lamine Mouhamadou Ndao, Genane Youness, Ndèye Niang, Gilbert Saporta

Effet de la complexité du réseau LSTM sur l'explicabilité en Maintenance Prédictive

Lamine NDAO [1,2] & Genane YOUNESS [1,2] & Ndeye NIANG [2] & Gilbert SAPORTA [2]

[1] *Laboratoire LINEACT CESI, Nanterre, IDFC*
[2] *Laboratoire Cedric-MSDMA-CNAM, Paris, France*
{mlndao ; gyouness}@cesi.fr ; {n-deye.niang ; gilbert.saporta}@cnam.fr

Résumé. La nature complexe des données en maintenance prédictive impose souvent l'utilisation de modèles d'apprentissage profonds. Malgré leur efficacité dans la prédiction du RUL (durée de vie résiduelle des machines), ces « boites noires » fournissent des résultats qui ne sont pas directement compréhensibles. Ainsi, des méthodes XAI post hoc sont généralement utilisées pour les expliquer. La modélisation inclut habituellement le choix de la complexité du modèle telle que la profondeur du réseau. Par ailleurs, on pourrait se demander si une complexité élevée du modèle ne freine pas la capacité des méthodes XAI dans l'explication des prédictions. Cette étude examine l'effet de la profondeur du réseau LSTM sur la qualité des explications des méthodes XAI, post hoc, locales, LIME, SHAP et L2X, utilisant huit métriques d'évaluation. Les résultats obtenus montrent que la qualité des explications peut suivre une certaine tendance en fonction de la complexité du réseau et selon la propriété de l'explication évaluée. Ces résultats ont montré également le manque de concordance entre les métriques d'évaluation, impliquant ainsi un besoin de cadre consensuel plus fiable dans l'évaluation des méthodes XAI.

Mots-clés Maintenance prédictive, Méthodes post-hoc locales, XAI, LSTM, métriques d'évaluation de XAI.

Abstract. The complexity of predictive maintenance data often requires the use of deep learning models. Despite their effectiveness in predicting RUL (Remain Useful Life), these "black boxes" provide results that are not directly comprehensible. Thus, post-hoc XAI methods are generally used to explain them. Modeling usually includes the choice of model complexity, such as network depth (number of layers). On the other hand, it may be asked whether high model complexity inhibits the ability of XAI methods to explain predictions. This study examines the effect of LSTM network depth on the explanation quality of XAI, post hoc, local, LIME, SHAP and L2X methods, using eight evaluation metrics. The results obtained show that the quality of explanations can follow a certain trend depending on the complexity of the network and the property of the explanation being evaluated. These results also showed a lack of agreement between evaluation metrics, implying a need for a more reliable consensus framework in the evaluation of XAI methods.

Keywords. Predictive maintenance, local post-hoc methods, XAI, LSTM, XAI evaluation metrics.

Introduction et travaux antérieurs

En maintenance prédictive, les réseaux de neurones récurrents de type LSTM sont de plus en plus utilisés. Cela peut s'expliquer par leurs bonnes performances dans ce domaine (Gouriveau et al. (2013)). Cependant, malgré ces performances grandissantes, ces méthodes sont souvent considérées comme des "boîtes noires" en raison de leurs structures internes complexes. Cette dernière implique le manque de transparence dans le processus de prédiction qui ne permet pas une explication directe des prédictions comme dans le cas d'une régression linéaire. D'ailleurs, ce manque de transparence a suscité de nombreuses questions relatives à la confiance en l'IA. En réponse, l'IA eXplicable (XAI) a été présentée comme une solution (DARPA Gunning and Aha (2019)) Depuis, une diversité de méthodes pour expliquer les résultats de ces modèles qualifiés de "boites noires" a été proposée. Ces méthodes peuvent être classées en différentes catégories, selon leur approche, le type d'explication qu'elles fournissent ou la portée de l'explication fournie. Ainsi, elles peuvent être : intrinsèques ou post hoc, selon qu'elles interviennent pendant ou après l'apprentissage du modèle ; globales, de cohorte ou locales, selon leur capacité à expliquer la prédiction d'une observation, d'un groupe d'observations ou l'ensemble des observations. Plusieurs auteurs soutiennent la pertinence des méthodes XAI. Toutefois, on peut s'interroger sur la fiabilité de ces méthodes XAI. On peut également se demander comment évaluer qualitativement et quantitativement leurs résultats et sur quelle base elles pourront être comparées. Pour répondre à ce besoin, diverses approches ont été proposées. Ces approches comprennent des méthodes qualitatives basées sur l'appréciation humaine ainsi que des approches quantitatives visant à évaluer quantitativement certaines des propriétés qu'une explication devrait satisfaire. Cette évaluation quantitative se fonde sur l'analyse de la relation entre les données et les explications (Honegger (2018)) ou entre les données, les prédictions et les explications (Solís-Martín et al. (2023)). Lorsqu'on parcourt la littérature, on constate que peu d'études s'intéressent à la relation entre la qualité des explications fournies par une méthode XAI et la complexité du modèle d'analyse (ex. nombre de couches, nombre de cellules de neurones). Dans le cadre des réseaux de neurones, on sait qu'augmenter le nombre de couches cachées (rajouter de la complexité) peut améliorer la performance du modèle d'analyse. Ainsi, on peut se demander si ce rajout de complexité ne limite-t-il pas la capacité d'une méthode XAI dans le processus d'extraction d'explications, sachant que les méthodes XAI fournissent un lien approximatif entre les prédictions du modèle d'analyse et les variables.

Dans cette étude, on se propose d'analyser empiriquement le lien entre la complexité d'un modèle d'analyse de type LSTM et la qualité des explications fournies par trois méthodes XAI LIME (Ribeiro et al. (2016)), SHAP (Lundberg and Lee (2017)) et L2X (Chen et al. (2018)) évaluée au moyen de huit métriques d'évaluation.

1 Méthodologie

Notations :

— N : Nombre d'observations (nombre de moteurs)

— $X = (x_i^t)_{(i \in N, t \in T)}$ l'ensemble des observations avec i le moteur et t une date donnée.

— Y_t : le RUL observée à la date t
— f : la fonction de prédiction du modèle d'analyse
— $\epsilon = \{\epsilon_i\}_{(i \in N)}$: poids des variables ("feature importance) dans l'explication de la prédiction du $\mathbf{RUL}_i$.
— ρ : coefficient de Spearman

Long Short-Term Memory (LSTM) : LSTM (Hochreiter and Schmidhuber (1997)) est un réseau de neurones de type récurrent. Il s'agit d'un modèle performant dans le traitement des données temporelles avec des structures complexes, résolvant le problème du gradient optimal de la rétro propagation pour modifier les poids du réseau.

L'eXplicabilité en Intelligence Artificielle (XAI) : Dans le contexte de la prédiction du RUL, on s'intéresse particulièrement aux parties du moteur qui sont responsables de la dégradation de la durée de vie résiduelle d'un moteur donné. Ainsi, dans notre analyse, nous nous concentrerons sur trois méthodes XAI locales post hoc : LIME (Local Interpretable Model-Agnostic Explanations), SHAP et L2X (Learning to Explain) qui sont considérées comme des méthodes d'explication locale basées sur les perturbations. Nous notons (x, y) une observation dans (X, Y), et g la fonction d'apprentissage du modèle de substitution (e.g. la régression linéaire).

— Pour l'approche LIME, l'idée principale est de créer un ensemble d'observations à partir de X_i à partir de la distribution de h. Ensuite, un modèle linéaire g est entrainé sur cet échantillon avec une contrainte d'éparpillement. Enfin, les coefficients de régression phi_i sont utilisés comme l'effet des différentes variables impliquées dans la prédiction.
— KernelSHAP utilise la valeur Shapley issue de la théorie des jeux pour attribuer une valeur, appelée valeur SHAP, à chaque variable, décrivant sa contribution à la prédiction finale. Par souci de simplicité, nous écrirons SHAP en référence à KernelSHAP.
— L2X cherche le sous-ensemble de variables le plus informatif en termes de prédiction correspondante pour cette instance. Le sous-ensemble est déterminé par un sélecteur de variables, par approximation variationnelle, qui est optimisée de manière à maximiser l'information mutuelle MI entre les caractéristiques et la prédiction correspondante.

En général, la génération des explications est basée sur la perturbation des variables et la sélection de voisinage. Dans le contexte de cette étude, étant donné que nous traitons de séries temporelles, nous adoptons l'approche de perturbation proposée par Solís-Martín et al. (2023), qui est plus appropriée pour les séries temporelles. La qualité des explications générées est évaluée par des métriques d'évaluation XAI qui vérifient certaines propriétés que ces explications doivent respecter, telles que la robustesse ou la stabilité.

Évaluation des méthodes XAI : Doshi-Velez and Kim (2017) ont décrit trois catégories d'approches d'évaluation pour des méthodes XAI : **"Human-grounded Evaluation"** qui englobe les méthodes basées sur l'appréciation humaine ; **"Application-grounded Evaluation"** qui implique des approches basées sur l'appréciation humaine spécifique à une application particulière, avec un accent prédominant sur les opinions d'experts dans le domaine concerné et **"Functionally-grounded Evaluation"** qui concerne les approches utilisant

des fonctions mathématiques ou proxy pour évaluer quantitativement la qualité des modèles post-hoc.

Dans ce travail, on s'intéresse particulièrement à la dernière approche d'évaluation. Il s'agit de métriques ayant comme but d'évaluer certaines propriétés d'une "bonne explication". Par exemple, on pourrait s'intéresser à la robustesse d'une explication. Autrement dit, on pourrait voir si les observations qui se ressemblent ont tendance à avoir des explications de leur prédication ressemblantes. La Table 1, présente les huit métriques utilisées dans cette étude. Un travail d'état de l'art nous a permis de réaliser ce tableau. Pour chaque métrique, nous avons fourni une description détaillée, sa base théorique et les propriétés qu'elle cherche dans l'évaluation de la performance des méthodes XAI.

Métrique	**Propriétés**	**Formule**	**Description**
Identité	Fidélité	$d(x_i, x_j) = 0 \Longrightarrow d(\epsilon_i, \epsilon_j) = 0$	Deux observations identiques au regard de d doivent recevoir des explications identiques au regard de d aussi (Honegger (2018)).
Séparabilité	Fidélité	$(x_i, x_j) \neq 0 \Longrightarrow d(\epsilon_i, \epsilon_j) > 0$	Deux observations (i, j) différentes au regard de la distance d ne peuvent pas recevoir 2 explications identiques au regard de d(Honegger (2018)).
Stabilité	Précision/Fidélité	$\rho_i = \rho(XX_i, E_i)$, $\rho_i > 0\ \forall i \in N$	La stabilité évalue si 2 observations similaires en termes de variable explicative X sont également similaires en termes d'explications ϵ. Un ρ_i élevé indique une interprétation plus intuitive(Honegger (2018)).
Congruence	Cohérence	$\delta = \sqrt{\frac{\sum(\alpha_i - \bar{\alpha})^2}{N}}$	La congruence évalue la variabilité de la cohérence. Une valeur δ plus petite indique une cohérence plus stable dans les prédictions (Doshi-Velez and Kim (2017)).
completeness	Représentativité	$\gamma_i = \frac{e_e^i}{p_e^i}$	La complétude évalue le ratio entre l'erreur de prédiction initiale et l'erreur après une perturbation des données initiales. Une valeur proche de 1 indique une meilleure qualité de l'explication (Doshi-Velez and Kim (2017)).
Cohérence	Cohérence	$\alpha_i = \lvert p_e^i - e_e^i \rvert$	La cohérence évalue la différence entre l'erreur de prédiction avec les données réelles et l'erreur de prédiction après perturbation des variables non importantes selon la méthode XAI. Une faible valeur d'α indique une cohérence élevée dans l'approche d'explicabilité (Doshi-Velez and Kim (2017)).
Acumen	Robustesse	$\omega = 1 - \frac{\sum_{f_i \in \mathcal{I}} \frac{p_a(f_i)}{N}}{M}$	L'acumen évalue si l'importance d'une variable selon la méthode XAI dépend de sa position dans les données. Une valeur d'acumen élevée indique une stabilité de l'importance des variables(Solís-Martín et al. (2023)).
Sélectivité	Sélectivité		La selectivité évalue la perturbation des variables les plus importantes. Elle consiste à ordonner les variables selon leur importance selon une méthode d'explicabilité, puis à perturber les données en substituant des variables aléatoires pour les variables les plus importantes, et enfin à calculer l'erreur de prédiction pour chaque perturbation (Laugel (2020)).

TABLE 1 – L'ensemble des métriques d'évaluation utilisées dans cette analyse

2 Expérimentations

Données : Cette analyse s'est basée sur les données C-MAPSS (Commercial Modular Aero-Propulsion System Simulation (Saxena et al. (2008)). Ces données enregistrent la durée de fonctionnement jusqu'à la défaillance des moteurs d'avions en simulant un large éventail de conditions opérationnelles réalistes, de paramètres de défaillance et de tendances de

dégradation dans différentes sections du système moteur.

La Figure 1 montre une vue abstraite de haut niveau de l'ensemble de données C-MAPSS. Essentiellement, les données de chaque moteur sont des séries temporelles multivariées (STM) qui consistent en des mesures prises au fil du temps à partir de différents capteurs montés sur le moteur. Chaque intervalle de temps correspond à un cycle de fonctionnement du moteur. L'objectif est de prédire la durée de vie utile restante (RUL), c'est-à-dire le nombre de cycles de fonctionnement restants pour le moteur, compte tenu de l'historique des mesures de ses capteurs. Les données de chaque moteur sont représentées à l'aide d'un total de 21 capteurs et de 3 modes de fonctionnement. Elles comprennent 4 parcs (flottes) de moteurs FD01, FD002, FD003, FD004. On considérera le groupe de moteurs FD004 dans cette analyse. Alors que les données d'entrainement enregistrent les parcours jusqu'à la défaillance, les données de test contiennent les mesures historiques des capteurs des moteurs jusqu'à un certain moment, avec une durée de vie résiduelle connue.

Pré-traitement : Dans cette étude, le prétraitement des séries temporelles est effectué en trois phases : lissage exponentiel, fenêtre temporelle et RUL rectifiée.

Normalisation : On commence par une normalisation des données en utilisant l'approche **min-max** dans chaque groupe de moteurs ayant les mêmes conditions opérationnelles. Cela permet de mettre les données issues des différents capteurs sur une même échelle. La formule est donnée par :

$$\text{norm}\left(x_{i,j}^{t}\right) = \frac{x_{i,j}^{t} - \min\left(x_j\right)}{\max\left(x_j\right) - \min\left(x_j\right)} - 1 \tag{1}$$

Lissage exponentiel des données : Pour produire une estimation précise de la RUL malgré la présence de bruit dans les données, un processus de lissage exponentiel est appliqué. Le lissage exponentiel attribue différents poids aux observations historiques en fonction de leur récence. Le choix du paramètre de lissage α dans le lissage exponentiel détermine le niveau d'importance accordé aux observations récentes. Les valeurs ajustées utilisent le paramètre de lissage α selon l'équation suivante :

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)\hat{y}_{t|t-1} \tag{2}$$

$\alpha = 1$ indique un apprentissage rapide, ce qui signifie que les prévisions sont basées sur les valeurs les plus récentes, tandis que $\alpha = 0$ indique un apprentissage lent.

Fenêtre temporelle et RUL rectifié : Après avoir réduit le bruit par lissage exponentiel, une fenêtre temporelle glissante de longueur fixe TW est appliquée pour convertir les données de séries temporelles multivariées. En fixant une durée de vie résiduelle à un certain seuil dit RUL rectifié RUL_{early}, le système est considéré comme "sain" jusqu'à ce qu'il atteigne ce point prédéfini. Cela permet au modèle de se concentrer sur l'apprentissage à partir du cycle RUL_{early}, quelle que soit la durée de vie antérieure du moteur. Cette valeur est à définir de façon optimale par le concepteur.

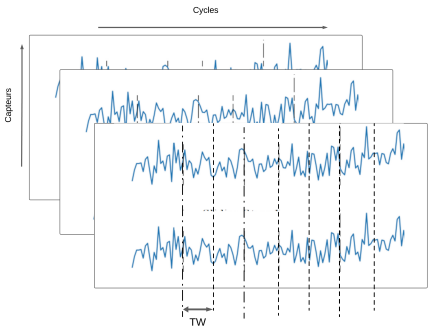

FIGURE 1 – Format des données et leur découpage en séquence suivant TW fixé

Processus d'analyse : L'objectif de cette étude est d'analyser empiriquement l'effet de la complexité du modèle d'analyse LSTM sur la qualité des explications fournies par les méthodes LIME, KernelSHAP et L2X. Pour ce faire, nous avons adopté le processus suivant :

1. Un modèle LSTM à une couche a été optimisé. Cette optimisation a été faite en choisissant les paramètres de prétraitement (α, TW et RUL_{early}) qui donnaient les meilleures performances en termes des deux critères $S-score$ et $RMSE$.
2. On répète 5 fois les étapes suivantes :
   — Une explication des prédictions sur un échantillon de 20 moteurs est obtenue à l'aide des 3 méthodes LIME, KernelSHAP et L2X.
   — L'évaluation des explications fournies est faite en utilisant les 8 métriques présentées dans la Table 1
3. L'étape (2) est répétée en ajoutant à chaque fois une couche aux réseaux de neurones LSTM jusqu'à avoir un modèle à 4 couches.

3 Résultats et discussions

On commencera par présenter les performances des modèles d'analyse. Ensuite, on présentera la qualité des explications en fonction de la profondeur du réseau LSTM. Le dernier paragraphe portera sur l'analyse du lien entre les métriques.

Performances versus profondeur du réseau LSTM : Une première étape de cette analyse portait sur l'optimisation du modèle à 1 couche en choisissant les paramètres de prétraitement optimaux. Ainsi, nous avons retenu le triplet de paramètres ($TW = 40$, $\alpha = 0.5$, $RUL_{early} = 100$) qui a donné les meilleures performances en termes de $RMSE$ et $S-score$. La structure du modèle d'analyse de base est présentée dans la Table 2. La Table 3 montre la performance du modèle en fonction du nombre de couches cachées. L'analyse de cette Table montre qu'en passant de 1 à 2 couches cachées, on arrive à obtenir de meilleures performances : $RMSE$ qui passe de 13.58 à 9.88 ; $S-score$ passe de 832.8 à 445. Cependant, cette amélioration du modèle n'est pas observée lorsqu'on passe de 2 à 3 couches cachées ou de 3 à 4 couches cachées. Ceci permet de noter que, dans le cadre de cette étude, le meilleur

modèle en termes de performances, n'est pas le modèle ayant le plus grand nombre de couches cachées.

Hyperparamètres	**valeurs**
Couches	1
Nœuds	64
Dropout	0.2
Batch-size	120
Learning-rate	10^{-3}

TABLE 2 – Les hyper-paramètres de LSTM

Couches	RMSE	S-score
1	13.58	832.8
2	**9.88**	**445.61**
3	10.56	479.67
4	10.22	515.54

TABLE 3 – Performance des modèles en fonction du nombre de couches cachées

Cependant, plus de couches cachées, implique une liaison moins linéaire entre la variable cible y et les variables dépendantes X. De surcroît, certaines méthodes XAI comme LIME, lient, de façon linéaire, X et y. LIME se base sur un modèle de substitution comme la régression linéaire à la place du modèle d'analyse afin d'utiliser les coefficients de régression pour expliquer l'influence de chaque X_j dans la prédication de $\bar{y}$. Ainsi, on se questionne sur la fiabilité des explications fournies par ces méthodes dites post hoc. Pour avoir une réponse de façon empirique, nous avons évalué la qualité des explications fournies par 4 modèles avec des nombres de couches cachées différentes (1 à 4) sur la base de 8 métriques d'évaluation.

Complexité du réseau versus qualité des explications : Les résultats de cette évaluation (Table 4) montre que, de manière générale, SHAP fournit de meilleurs résultats au regard des métriques d'évaluation utilisées. En effet, elle donne une meilleure qualité des explications au regard de 5 des 8 métriques d'évaluation utilisées (Cohérence : 0.20, Congruence : 0.24, Sélectivité : 0.79, Acumen : 0.5 ; Congruence 0,42). On note également, qu'en général, la valeur de la métrique sélectivité augmente lorsqu'on augmente le nombre de couches cachées. On peut le voir sur les méthodes comme LIME (1 couche : 0.58, 2 couches : 0.72, 3 couches : 0.73, 4 couches : 0.78). Cette tendance est observée sur l'ensemble des 3 méthodes XAI.

Nous avons également analysé le comportement des valeurs prises par les 6 métriques d'évaluation en excluant Id et Sep, vu qu'elles ne présentent pas variations (Figure 2). Dans cette analyse, on cherche à déceler empiriquement l'effet de la profondeur du modèle (complexité du modèle) sur la qualité des explications fournies. L'analyse montre que pour certaines métriques (e.g. sélectivité), on arrive à déceler une tendance de la qualité des explications dans certaines méthodes XAI lorsque le nombre de couches augmente. Cela montre que la performance du modèle d'analyse n'est forcément liée à la qualité de l'explication. De même, pour certaines méthodes XAI, (ex. SHAP), lorsqu'on augmente le nombre de couches, la qualité de l'explication s'améliore selon certaines métriques (ex. Sélectivité), et se dégrade selon d'autres métriques (cohérence, completeness, congruence).

Méthodes	Couches	Id ↑	Sep ↑	St ↑	Co ↓	Con ↓	Sel ↑	Ac ↑	Com ↑
LIME	1	1.0(0.0)	1.0(0.0)	0.98(0.04)	0.24(0.11)	0.27(0.06)	0.58(0.04)	0.07(0.04)	0.14(0.2)
	2	1.0(0.0)	1.0(0.0)	**1.0(0.0)**	**0.22(0.08)**	**0.26(0.06)**	0.72(0.06)	**0.12(0.01)**	0.07(0.04)
	3	1.0(0.0)	1.0(0.0)	0.98(0.04)	**0.22(0.08)**	**0.26(0.06)**	0.73(0.06)	0.09(0.04)	**0.21(0.38)**
	4	1.0(0.0)	1.0(0.0)	**1.0(0.0)**	0.22(0.09)	0.26(0.06)	**0.78(0.06)**	0.09(0.03)	0.19(0.35)
SHAP	1	1.0(0.0)	1.0(0.0)	**0.96(0.05)**	**0.20(0.08)**	**0.24(0.06)**	0.64(0.06)	0.42(0.12)	**0.42(0.1)**
	2	1.0(0.0)	1.0(0.0)	0.88(0.16)	0.22(0.08)	0.26(0.06)	0.77(0.07)	**0.50(0.06)**	0.19(0.13)
	3	1.0(0.0)	1.0(0.0)	0.94(0.09)	0.22(0.08)	0.26(0.06)	0.78(0.05)	0.43(0.07)	0.11(0.18)
	4	1.0(0.0)	1.0(0.0)	0.94(0.09)	0.22(0.09)	0.26(0.06)	**0.79(0.07)**	0.34(0.05)	0.17(0.19)
L2X	1	1.0(0.0)	1.0(0.0)	0.96(0.05)	0.25(0.13)	**0.27(0.07)**	0.59(0.05)	**0.04(0.02)**	**0.38(0.22)**
	2	1.0(0.0)	1.0(0.0)	0.96(0.05)	**0.24(0.1)**	0.28(0.08)	0.72(0.06)	0.03(0.01)	0.21(0.36)
	3	1.0(0.0)	1.0(0.0)	**1.0(0.0)**	0.26(0.09)	0.29(0.07)	0.68(0.06)	0.03(0.01)	0.06(0.07)
	4	1.0(0.0)	1.0(0.0)	0.98(0.04)	0.25(0.1)	0.29(0.07)	**0.74(0.05)**	0.03(0.01)	0.27(0.41)

TABLE 4 – Moyenne (écart-type) des métriques d'évaluation en fonction de la profondeur du réseau LSTM. Les moyennes et les écarts-type ont été obtenus en répétant 5 fois la prédiction, l'explication puis l'évaluation (Id : Identité, Sep : Séparabilité, St : Stabilité, Co : Cohérence : Con : Congruence, Sel : Sélectivité : Ac : Acumen, Com : Completeness)

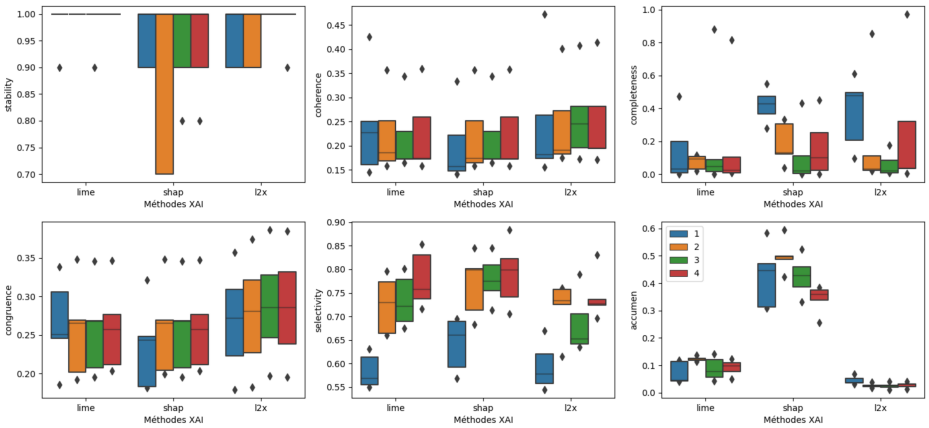

FIGURE 2 – Variation des métriques en fonction du nombre de couches cachées (couleurs), par méthode(en abscisse).

Cependant, si certaines ont décelé une tendance positive en fonction du nombre de couches cachées (Sélectivité), d'autres révèlent une tendance plutôt négative de la qualité de l'explicabilité en fonction du nombre de couches. Ceci montre qu'une complexité grandissante du modèle ne limite pas forcément la qualité de l'explication lorsque certaines propriétés de l'explication sont évaluées. Sélectivité évalue la capacité de la méthode XAI à sélectionner des variables pertinentes dans le processus de l'explication. Ainsi, on pourrait conclure qu'une complexité grandissante du modèle permet à LIME et SHAP de bien sélectionner les variables pertinentes dans le processus de l'explication.

Relation entre les métriques d'évaluation : L'analyse de la relation entre les métriques a permis de noter que toutes les métriques ne sont pas corrélées à un même axe (Figure 3). Ceci est compréhensible dans la mesure où elles ne sont pas censées évaluer les mêmes propriétés (Table 1). Par exemple, sélectivité évalue la capacité d'une méthode XAI à sélectionner les variables pertinentes dans l'explication des résultats d'une "boite noire",

alors que la cohérence évalue sa capacité à commettre des erreurs similaires lorsque les variables pertinentes ne sont pas perturbées (cohérence). Ainsi, on note que la cohérence et la congruence sont positivement corrélées au premier axe, car évaluant la même propriété, tandis qu'acumen, la sélectivité et la stabilité sont plutôt fortement liées au deuxième axe. La corrélation forte et positive entre acumen et sélectivité peut permettre de dire que les deux propriétés qu'elles évaluent (la robustesse et la sélectivité) peuvent être placées sur une même dimension.

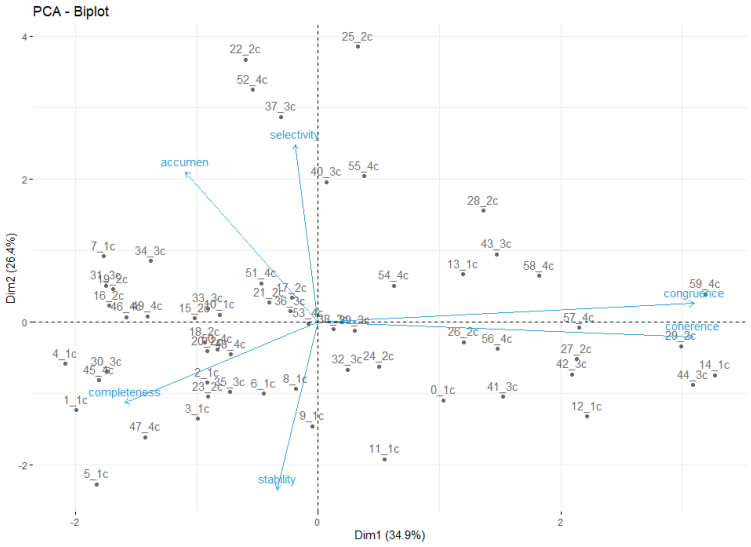

FIGURE 3 – Nuage des individus associé à l'affichage de la corrélation entre les métriques d'évaluation

Conclusion et perspectives

Dans cette étude, il était question d'analyser l'effet de la complexité du réseau LSTM sur la qualité des explications fournies par trois méthodes XAI post-hoc LIME, SHAP et L2X, mesurée par huit proxys. Les résultats ont montré que la complexité du réseau, en termes de nombres de couches, ne limite pas en général la capacité des méthodes XAI à fournir de bonnes explications. Ainsi, il n'y a pas de compromis à faire entre la complexité du réseau en termes de nombre de couches et la transparence lorsqu'une méthode post-hoc est utilisée. Ces résultats ont également permis de noter que certaines métriques, par leur définition comme identité et séparabilité, donnaient en général une valeur de 1, et que d'autres comme la sélectivité donnait des valeurs très liées à la complexité du réseau. Par ailleurs, nos résultats ont montré une concordance entre les métriques et lient fortement et positivement deux métriques (cohérence et congruence) censées évaluer la même propriété (cohérence). Ils ont permis par contre de noter qu'une explication ne peut pas avoir à la fois les propriétés fidélité, précision et robustesse, car les métriques qui les évaluent sont directement opposées par le deuxième axe factoriel. Ceci pourrait être problématique quant au choix de la métrique à considérer pour l'évaluation des méthodes post-hoc.

Dans nos travaux futurs, on pourrait envisager de remédier à ce manquement en redéfinissant certaines métriques pour améliorer leur pertinence dans l'évaluation des méthodes XAI. Il sera également pertinent de proposer une métrique synthétique pouvant prendre en compte l'ensemble des métriques utilisées aujourd'hui. Cela permettra d'éviter le dilemme sur le choix de la métrique à considérer dans l'évaluation des méthodes XAI post-hoc en maintenance prédictive.

Références

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain : An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pages 883–892. PMLR, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv :1702.08608*, 2017.

Rafael Gouriveau, Kamal Medjaher, Emmanuel Ramasso, and Noureddine Zehrouni. Phm – prognostics and health management de la surveillance au pronostic de défaillances de systèmes complexes. *Techniques de l'ingénieur Maintenance*, base documentaire : TIP095WEB.(ref. article : mt9570), 2013. doi : 10.51257/a-v1-mt9570. fre.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2) :44–58, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8) :1735–1780, 1997.

Milo Honegger. Shedding light on black box machine learning algorithms : Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv :1808.05054*, 2018.

Thibault Laugel. *Interprétabilité locale post-hoc des modèles de classification" boites noires"*. PhD thesis, Sorbonne université, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management*, pages 1–9. IEEE, 2008.

David Solís-Martín, Juan Galán-Páez, and Joaquín Borrego-Díaz. On the soundness of xai in prognostics and health management (phm). *Information*, 14(5) :256, 2023.