Dans un contexte marqué par un développement accéléré de l’ IA l’industrie est contrainte de mettre à niveau.

L’utilisation de toute cette technologie récente a permis une modernisation de l’industrie. Ce développement est marqué entre autre, par l’autonomie des systèmes tels que les bâtiments autonomes ou voiture autonome et une multiplication des sources de données grâce aux objets connectés (iot INTERNET OF THINGS) : on parle de l’industrie 4.0. Cette situation a permis une disponibilité des données qui peuvent être intéressante dans le cadre de la prise de décision notamment dans le cadre de la maintenance. La maintenance permet de garder les composante d’un système comme les machines dans leur condition optimale de fonctionnement. En effet, elle permet de prévenir les pannes de système qui peuvent être très couteuses aux industries \citep[ran2019survey].

Il existe généralement trois principales stratégies de maintenance selon \citep{ ran2019survey}. La première, dite maintenance proactive ou réactive est une stratégie où l’intervention aura lieux que quand une panne est observée. Ceci peut causer des coûts assez élevé liés aux interruption ou panne inattendues. La deuxième stratégie est appelée la maintenance préventive. Il s’agit d’une stratégie qui consiste à planifier des interventions de façon régulière. Elle peut être couteuse lorsque l’intervention est planifiée alors qu’il n’y a pas de panne. La maintenance prédictive, quant à elle, est une stratégie qui essaye de trouver le juste milieu entre les deux premières approches. Elle se base sur la prédiction de la santé de la machine comme la durée de vie résiduel des machines (RUL) à travers une modélisation permettant de planifier de façon utile les interventions de réparation. Dans la recherche, trois type de modèles sont généralement utilisés pour prédire le RUL.

Il s’agit des modèles physiques qui se base sur la compréhension du fonctionnement de dégradation des machines qui est formulé via des modèles mathématiques permettant de calculer le RUL. Le deuxième type de modèle est appelé modèle hybride qui combine les modèles physiques et les connaissances acquises (données empiriques) au cours du fonctionnement des machines pour mettre à jour les paramètre du modèle mathématique dans le calcul du RUL.

L’approche data driven se base sur les données historiques disponibles pour prédire le RUL via les modèle d’apprentissage comme les réseaux de neurones. Ainsi, elle permet une exploitation de la masse d’information disponible grâce aux capteurs afin de prédire la durée de vie résiduel des machines (RUL) indispensable dans la gestion des pannes.

Dans cette approche, la littérature montre que les modèle de l’IA dit DEEP LEARNING fournissent souvent les meilleures performances (mettre les références). En effet, ceci peut s’expliquer par le fait que les capteurs fournissent souvent des données complexes comme les images ou les signaux qui sont des séries temporelles multivariées. A cause de leur complexité, les paternes contenues dans ces dernières sont mieux pris en compte par les modèles de Deep Learning.

Cependant, ces types de modèles d’apprentissage profond, malgré leurs bonnes performances dans ce domaine, fournissent des résultats qui sont peu interprétables. Ceci ne responsabilise pas les preneurs de décision qui se basent sur les résultats de ces IA. En outre, il est important de connaitre les capteurs qui interviennent le plus dans la prédiction du RUL pour une intervention plus optimale dans le cadre de la maintenance. D’ailleurs, pour une IA de confiance et une meilleure transparence, DARPA (mettre la ref) a introduit le concept d’IA explicable (XAI).

Ce concept est devenu un domaine à part entier de l’IA. Il promeut une meilleure considération de l’IA explicable à travers des méthodes dite d’explicabilité (méthode XAI). Pour répondre à ce besoin, des auteurs ont proposé une variété d’approches permettant d’expliquer les résultats fournis par les modèles d’apprentissage moins interprétables comme les modèles d’ensemble ou les modèles d’apprentissage profond appelés souvent boites noires. [Linardatoss] a présenter plus de 20 méthodes d’explicabilité.

Cependant, ces méthodes utilisent des algorithmes souvent indépendants de l’apprentissage (e.g. méthodes dites post-hoc). Ce qui ne permet pas d’avoir une vision sur les conséquences (direct ou indirecte) des aspects de la modélisation sur la qualité de l’explication fournies par les méthodes d’XAI.

Cette modélisation peut etre décomposée en 3 phases : le prétraitement des données, le choix et l’entrainement du modèle et l’évaluation des résultats. A chaque phase, nous sommes emmenés à faire des choix de méthodologie. Par exemple, lors du prétraitement dans le cadre des series temporelles multivariée, on doit s’assurer de selectionner les « bonnes features », une facon optimale de lissage etc ; lors de l’entrainement, on est emené à faire de choix d’hyper parametre du modele, de fixer les critères d’évaluation etc. Ainsi, tous ces choix de méthodes (e.g. méthode de standardisation, taille des fenêtres dans le cadre des séries temporelles) peuvent influer de façon directe ou indirecte sur la qualité des explications obtenues à l’aide des méthodes XAI.

Face à cette variété de méthodes d’explicabilité, il apparait un besoin de cadre de comparaison qui permettra d’arbitrer entre les méthodes d’XAI pour une meilleure explication des résultats des modèles de boite noire. C’est d’ailleurs ce qui a poussé certains auteurs à proposer différentes approches pour évaluer les méthodes d’XAI. Ces approches peuvent etre regroupées en trois groupes : human grounded, application grounded et functionnal grounded.

XAI et PHM

La confiance dans les algorithmes de l'IA soulève de plus en plus de questions depuis un certain temps surtout dans le domaine de la maintenance predictive (\citet{nor2021explainable, ran2019survey}).

En PhM, lrd modeles de DL sont souvent les plus utilisés. Cependant, leur aspect, souvent complexe, ou incompréhensible par les décideurs ou acteurs concernés soulève de nombreuses interrogations, voire jusqu'à leur attribuer l'appellation "boîtes noires".

Pour répondre à ce besoin de compréhension des modèles de ML, des auteurs (e.g. \cite{ribeiro2016should, lundberg2017unified, chen2018learning}) ont proposé une variété de propositions d'algorithmes dits XAI methods permettant d'expliquer les résultats de ces. Ces méthodes peuvent être regroupées selon leur capacité à expliquer tout type de modèle ML ou un type bien spécifique de modèles (agnostique, spécifique), selon le porté d'explication (globale ou locale) et selon le type de données utilisée durant l'apprentissage (images, tabulaires, temporelles, etc.) ou même selon le mécanisme (durant l'apprentissage : intrinsèque, après l'apprentissage post-hoc).

Une approche est dite globales lorsqu’elle fournit une explication générale des résultats d’un modèle de ML (\cite{molnar2022}). Elle porte souvent sur l’explication d’une ou de plusieurs variables en termes d’apport sur l’ensemble des résultats du modèle à expliquer.

Nous opposons les méthodes XAI globale aux celles dites locales qui fournit l’explication d’un résultat portant sur un ou plusieurs individus ayant une modalité spécifique sur une variable donnée. Par exemple, dans le cadre de la prédiction du RUL, une explication locale consistera à expliquer pour chaque machine le capteur qui informe le plus sur la dégradation de la machine en question.

Lorsque le modèle d'explicabilité opère sans tenir compte des calculs faits par la boîte noire pour produire une explication sur ce dernier, on dit qu'il s'agit d'un modèle post-hoc. Dans le cas contraire, on parle de modèle intrinsèque.

Ces méthodes sont de plus en plus utilisées dans le cadre de la prédiction du RUL en PhM. Par exemple, \citet{nair2023predicting}, pour sélectionner les meilleures variables permettant une meilleure prédiction du RUL des Li-ion des batteries a utilisé la méthode SHAP \cite. \citet{9318912}, pour expliquer les résultats de ses prédictions du RUL sur les données du CMAPSS, a utilisé la même méthode locale post hoc (shap). Plus récemment, genane et al, a fait recours à la méthode.

Au regard de cet usage progressive des méthodes XAI en PdM, on pourrait se poser certaines questions, notamment relative à la fiabilité de ces méthodes. Pour ce faire, des auteurs comme (hoonneger, solis, doshi, etc.) se sont pencher sur comment évaluer la fiabilité des résultats fournis par les méthodes XAI. Ces derniers soulignent un nombre de propriétés que doivent respecter que doivent respecter les explications fournies par ces modèles. De ce fait, ils ont proposé un cadre permettant la comparaison des résultats des méthodes d'explicabilité au regard de ces propriétés.

XAI evaluation metric

Pour évaluer les resultats des méthodes d’explicabilité, doshi et al . Distinguent dans son revu trois grande approches.

Human grounded :

Applications grounded

Functionnaly grounded.

Une des questions qui concernent surtout le dernier groupe d'approches, c'est comment évaluer quantitativement la qualité d'une explication. Partant des définitions de l’interprétabilité et d’explicabilité existantes dans la littérature, des propriétés qui devrait respecter une « bonne explication » ont été identifiées (survey metrics). Par exemple, une explication doit être fidèle aux données d'entrainement, claire, robuste, etc. Ces propriétés ont été transformées en objectifs à atteindre par les modèles XAI. Ainsi, une métrique d'évaluation a pour but d'évaluer l'atteinte de ces objectifs de façon quantitative.

Problématique de l’étude

Au regard de l’existence des méthodes XAI et les moyens (metrics) pour les évaluer on pourrait s’intéresser aux conséquences faites lors de la mise en place d’un modèle d’analyse sur la qualité des explications fournies par les méthodes XAI. En effet, la modélisation inclut de nombreux aspect qui affectent directement ou indirectement les relations entre les variables prédites $Y$ et les variables explicatives $X$ dans le cadre de l’apprentissage supervisé par exemple. Par exemple, plus le modèle d’apprentissage est complexe, plus les relations entre $X$ et $Y$ est non linéaire. Par ailleurs, certaines méthodes XAI comme LIME considère une relation plutôt linéaire dans leur processus de recherche d’explication. De ce fait, donner un lien linéaire entre $X$ et $Y$, alors qu’elles ne sont pas liées de façon linéaire peut entrainer un biais dans les explications fournies par méthodes XAI. Cette question n’est pas souvent traitée par la littérature même si les auteurs s’intéressent de plus en plus à l’évaluation des méthodes XAI. Dans cette étude, nous proposons d’analyse l’impact de certains choix de modélisation sur la qualité des explications fournies par les méthodes d’XAI. Dans nos analyses antérieures, on s’est intéressé que des aspects concernant la phase preprocessing. Dans cette analyse, on s’intéressera aux trois phases de la modélisation à savoir Pre-In and Post-procession et post-procession.