Prédiction des défaillances d'équipements industriels par apprentissage automatique

Application à la maintenance prédictive

Steeve NDONG

Modélisation Random Forest · Jeu de données AI4I-PMDI · Explicabilité SHAP

Résumé

La maintenance prédictive vise à exploiter les données issues des équipements industriels afin d'anticiper les situations de défaillance et d'améliorer la prise de décision en maintenance. Cette étude présente une démarche reproductible d'apprentissage automatique appliquée à la classification binaire des défaillances d'équipements à partir du jeu de données **AI4I-PMDI**, dérivé du jeu AI4I 2020 et enrichi d'irrégularités représentatives de contextes industriels plus complexes. AI4I-PMDI a été proposé par Autran et al. afin notamment d'introduire des données manquantes et des irrégularités d'acquisition.

Le protocole expérimental porte sur 10 000 observations, réparties de manière stratifiée entre un jeu d'entraînement de 8 000 observations et un jeu de test indépendant de 2 000 observations. Trois familles de modèles ont été comparées : régression logistique, arbre de décision et Random Forest, dans leurs configurations initiales et optimisées. Le **Random Forest optimisé** a été retenu comme modèle champion. Ses hyperparamètres ont été sélectionnés par `RandomizedSearchCV` parmi 40 configurations candidates, selon une validation croisée stratifiée à cinq plis utilisant l'*Average Precision* comme métrique d'optimisation. La meilleure configuration comporte 500 arbres, une profondeur maximale de 8, min_samples_split = 20, min_samples_leaf = 5, max_features = None, class_weight = balanced_subsample et bootstrap = True.

Une procédure distincte de validation croisée *out-of-fold* (OOF), limitée au jeu d'entraînement, a ensuite permis de sélectionner un seuil décisionnel de **0,6350330879616113** en maximisant le score F2. Sur le jeu de test indépendant, ce seuil conduit à une accuracy de **98,80 %**, une précision de **76,74 %**, un rappel de **94,29 %**, un F1-score de **84,62 %**, un F2-score de **90,16 %**, une ROC-AUC de **0,9813** et une Average Precision de **0,9444**. La matrice de confusion comporte 1 910 vrais négatifs, 20 faux positifs, 4 faux négatifs et 66 vrais positifs.

L'analyse de l'importance intrinsèque, de l'importance par permutation et des contributions SHAP complète l'évaluation quantitative par une lecture globale et locale du comportement du modèle. La démarche est orchestrée sous Kedro et tracée sous MLflow. Les résultats établissent la faisabilité méthodologique d'une chaîne reproductible associant modélisation, optimisation, sélection du seuil hors test, évaluation indépendante et explicabilité. Ils ne constituent cependant pas une validation industrielle externe : celle-ci nécessiterait des données opérationnelles, longitudinales et représentatives des conditions réelles d'exploitation.

Mots clés

maintenance prédictive ; apprentissage automatique ; Random Forest ; classification binaire ; défaillance industrielle ; F2-score ; SHAP ; explicabilité ; Kedro ; MLflow.

Mots clés

Predictive maintenance aims to leverage data generated by industrial equipment to anticipate failure conditions and improve maintenance decision-making. This study presents a reproducible machine-learning approach to binary equipment-failure classification using the AI4I-PMDI dataset. The experimental protocol comprises 10,000 observations, stratified into a training set of 8,000 observations and an independent held-out test set of 2,000 observations. Three model families were compared -Logistic Regression, Decision Tree, and Random Forest- in both baseline and tuned configurations. The tuned Random Forest was selected as the champion model. Its hyperparameters were optimized using RandomizedSearchCV over 40 candidate configurations with five-fold stratified cross-validation and Average Precision as the optimization metric. The optimal configuration comprises 500 trees, a maximum depth of 8, min_samples_split = 20, min_samples_leaf = 5, max_features = None, class_weight = balanced_subsample, and bootstrap = True.

A separate out-of-fold (OOF) cross-validation procedure, restricted to the training set, was subsequently used to select a decision threshold of **0.6350330879616113** by maximizing the F2-score. On the independent test set, the resulting model achieved an accuracy of **98.80%**, precision of **76.74%**, recall of **94.29%**, an F1-score of **84.62%**, an F2-score of **90.16%**, a ROC-AUC of **0.9813**, and an Average Precision of **0.9444**. The confusion matrix yielded **1,910 true negatives**, **20 false positives**, **4 false negatives**, and **66 true positives**.

Intrinsic feature importance, permutation importance, and SHAP-based explanations complement the quantitative evaluation by providing global and local insights into model behavior. The experimental workflow is orchestrated with Kedro, while MLflow ensures experiment tracking and model traceability. The results demonstrate the methodological feasibility of a reproducible pipeline integrating model development, hyperparameter optimization, test-independent threshold selection, held-out evaluation, and explainability. However, these findings should not be interpreted as external industrial validation. Such validation would require operational, longitudinal data representative of actual industrial operating conditions.

Keywords

predictive maintenance; machine learning; Random Forest; binary classification; industrial failure; F2-score; SHAP; explainability; Kedro; MLflow.

Liste des abréviations, sigles et acronymes

Sigle / Abréviation	Signification
AI	Artificial Intelligence - Intelligence artificielle
AI4I	Artificial Intelligence for Industry
AI4I-PMDI	AI4I – Predictive Maintenance Dataset with Irregularities
AP	Average Precision
CV	Cross-Validation - Validation croisée
F1	F1-score
F2	F2-score
FN	False Negative - Faux négatif
FP	False Positive - Faux positif
GMAO	Gestion de Maintenance Assistée par Ordinateur
IA	Intelligence artificielle
IIoT	Industrial Internet of Things - Internet industriel des objets
ML	Machine Learning - Apprentissage automatique
MLOps	Machine Learning Operations
OOF	Out-of-Fold
PR	Precision–Recall - Précision–Rappel
RF	Random Forest - Forêt aléatoire
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
RUL	Remaining Useful Life - Durée de vie utile restante
SHAP	SHapley Additive exPlanations
TN	True Negative - Vrai négatif
TP	True Positive - Vrai positif

1. Introduction

La maintenance constitue une fonction essentielle des systèmes industriels. La disponibilité, la sûreté de fonctionnement, la continuité de production et la maîtrise des coûts dépendent directement de la capacité d'une organisation à détecter les dégradations et à intervenir avant qu'elles ne produisent des conséquences significatives. Le développement de l'Internet industriel des objets, des systèmes cyberphysiques et de l'apprentissage automatique a progressivement élargi les possibilités d'exploitation des données de fonctionnement à des fins de maintenance.

La maintenance prédictive s'inscrit dans cette évolution. Elle mobilise des données relatives à l'état et au fonctionnement des équipements afin de produire des informations susceptibles de soutenir une décision de maintenance. La littérature montre toutefois que la maintenance prédictive recouvre des problématiques diverses : détection d'anomalies, diagnostic de défauts, classification d'états de fonctionnement, pronostic ou estimation de durée de vie restante (*Remaining Useful Life*, RUL) (Carvalho et al., 2019 ; Zonta et al., 2020).

La présente étude porte spécifiquement sur une tâche de classification binaire. Il ne s'agit donc ni de déterminer la date future d'une panne ni d'estimer une RUL, mais d'estimer, à partir d'un ensemble de caractéristiques X, la probabilité associée à une classe de défaillance :

$$P(Y = 1 \mid X)$$

Cette distinction est importante. Une performance élevée en classification ne suffit pas à démontrer qu'un système constitue à lui seul une solution industrielle complète de maintenance prédictive.

L'évaluation des modèles soulève en outre une difficulté méthodologique lorsque les défaillances représentent une classe minoritaire. Une accuracy élevée peut alors masquer une capacité insuffisante à identifier les événements rares. Les mesures Precision–Recall sont particulièrement informatives dans les situations de classification déséquilibrée (Saito & Rehmsmeier, 2015).

Un second enjeu concerne la transparence des modèles. Les méthodes d'ensemble telles que Random Forest peuvent atteindre de bonnes performances tout en rendant leur comportement moins immédiatement interprétable qu'un modèle linéaire ou un arbre unique. SHAP fournit à cet égard un cadre d'attribution permettant d'étudier la contribution des variables aux prédictions d'un modèle (Lundberg & Lee, 2017).

Enfin, l'évaluation expérimentale doit éviter toute utilisation du jeu de test lors de la sélection du modèle ou du seuil décisionnel. L'objectif du présent travail est donc double : obtenir une classification performante des défaillances et construire une procédure suffisamment structurée pour distinguer clairement apprentissage, optimisation, sélection du seuil, évaluation finale et interprétabilité.

La question de recherche est ainsi formulée :

> ***Dans quelle mesure une chaîne reproductible d'apprentissage automatique, associant optimisation d'un Random Forest, sélection d'un seuil décisionnel hors test et méthodes d'explicabilité, permet-elle de détecter efficacement les défaillances présentes dans le jeu AI4I-PMDI ?***

2. Cadre conceptuel et état de l'art

2.1. Maintenance industrielle et maintenance prédictive

Les stratégies de maintenance peuvent être distinguées selon la logique qui déclenche l'intervention :

- **La maintenance corrective** intervient après défaillance ;
- **la maintenance préventive systématique** repose sur des échéances ou des usages prédéterminés ;
- **la maintenance conditionnelle** mobilise l'état observé de l'équipement ;
- **la maintenance prédictive** cherche à exploiter les données disponibles pour anticiper les situations nécessitant une intervention.

L'évolution vers l'Industrie 4.0 renforce cette dernière approche en rendant accessibles des volumes croissants de données issues de capteurs, d'automates, de systèmes de supervision et de plateformes industrielles. La littérature souligne toutefois le caractère multidisciplinaire de la maintenance prédictive, à l'intersection de l'ingénierie, des données et des systèmes d'information (Zonta et al., 2020).

2.2. Apprentissage automatique appliqué à la maintenance

L'apprentissage automatique permet de construire des fonctions de décision à partir d'exemples historiques. Dans une classification binaire :

$$f : X \rightarrow \{0, 1\}$$

ou, lorsque le classifieur produit des probabilités :

$$f : X \rightarrow [0, 1]$$

Carvalho et al. (2019) montrent la diversité des techniques d'apprentissage automatique appliquées à la maintenance prédictive ainsi que la nécessité d'adapter les méthodes au type de données et au problème industriel étudié.

Le Random Forest constitue un ensemble d'arbres de décision combinant randomisation des observations et/ou des variables afin de réduire notamment la dépendance à un arbre unique. Il a été formalisé par Breiman (2001).

2.3. AI4I et AI4I-PMDI

Le jeu AI4I original a été présenté par Matzka dans le contexte d'applications de maintenance prédictive et d'intelligence artificielle explicable. L'article associé présente un dataset synthétique destiné à fournir un benchmark reproductible à la communauté (Matzka, 2020).

AI4I-PMDI constitue une extension de ce corpus. Autran et al. (2024) ont introduit des irrégularités destinées à mieux représenter certaines difficultés des environnements industriels, notamment les données manquantes et l'acquisition irrégulière. L'article correspondant a été publié dans *Procedia Computer Science*, volume 246, pages 1201–1209. Le dataset est également publié sur Figshare sous licence CC BY 4.0.

2.4. Classification déséquilibrée

Dans le jeu de test de cette étude, seules 70 observations sur 2 000 appartiennent à la classe positive, soit :

$$\frac{70}{2000} = 3,5\%.$$

Dans un tel contexte, l'accuracy ne suffit pas à caractériser la capacité du modèle à détecter les défaillances. Les métriques utilisées comprennent donc Precision, Recall, F1, F2, ROC-AUC et Average Precision.

Le score Fβ est défini par :

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 Precision + Recall}$$

Avec β = 2 :

$$F_2 = 5\frac{Precision \times Recall}{4Precision + Recall}$$

Le rappel reçoit ainsi davantage de poids que la précision.

2.5. Explicabilité

L'explicabilité complète l'évaluation prédictive en étudiant le comportement du modèle. Trois approches complémentaires sont retenues :

- l'importance intrinsèque du Random Forest ;
- l'importance par permutation ;
- les valeurs SHAP.

SHAP (*SHapley Additive exPlanations*) attribue aux variables une contribution à la sortie du modèle pour une prédiction donnée (Lundberg & Lee, 2017).

Une contribution SHAP doit toutefois être interprétée comme une explication du **comportement du modèle**, et non comme la preuve d'une relation causale physique.

3. Matériels et méthodes

3.1. Jeu de données et variable cible

L'étude utilise AI4I-PMDI, publié par Autran et dérivé du dataset AI4I 2020.

Le corpus expérimental comprend :

$$N = 10\,000 \text{ observations}$$

La variable cible est machine_failure, avec :

$$Y = \begin{cases} 0, & \text{absence de défaillance} \\ 1, & \text{défaillance.} \end{cases}$$

Le corpus est séparé de manière stratifiée en :

$$N_{train} = 8\,000 \;\; et \;\; N_{test} = 2\,000$$

Le jeu de test représente donc 20 % du corpus et reste hors du processus de sélection du seuil décisionnel.

3.2. Variables explicatives

Après exclusion des variables susceptibles d'introduire une fuite de cible, dix variables explicatives sont utilisées.

Les sept variables numériques sont :

- Air temperature (K) ;
- Process temperature (K) ;
- Rotational speed (rpm) ;
- Torque (Nm) ;
- Tool wear (min) ;
- temperature_difference ;
- mechanical_power.

Les trois variables catégorielles sont :

- System ;
- Control ;
- Type.

Deux caractéristiques sont dérivées des variables initiales. La différence de température est calculée par :

$$\Delta T = T_{process} - T_{air}$$

La puissance mécanique est obtenue par :

$$P = \tau \cdot \omega,$$

Avec :

$$\omega = \frac{2\pi n}{60},$$

d'où :

$$P = \tau \frac{2\pi n}{60}$$

où :

- P est la puissance mécanique en watts ;

- τ le couple et n la vitesse de rotation en Nm ;
- n est la vitesse de rotation en tr/min ;
- ω est la vitesse angulaire en rad/s.

3.3. Prétraitement

Les variables numériques font l'objet d'une imputation par la médiane et d'une standardisation. Les variables catégorielles sont imputées selon la modalité la plus fréquente puis encodées par *one-hot encoding*, avec prise en charge des catégories inconnues lors de l'inférence.

La chaîne de traitement utilise également des indicateurs de valeurs manquantes. La reproductibilité des opérations stochastiques est assurée par :

$$random\_state = 42$$

3.4. Protocole expérimental

La Figure 1 représente la séparation stricte des différentes étapes expérimentales.

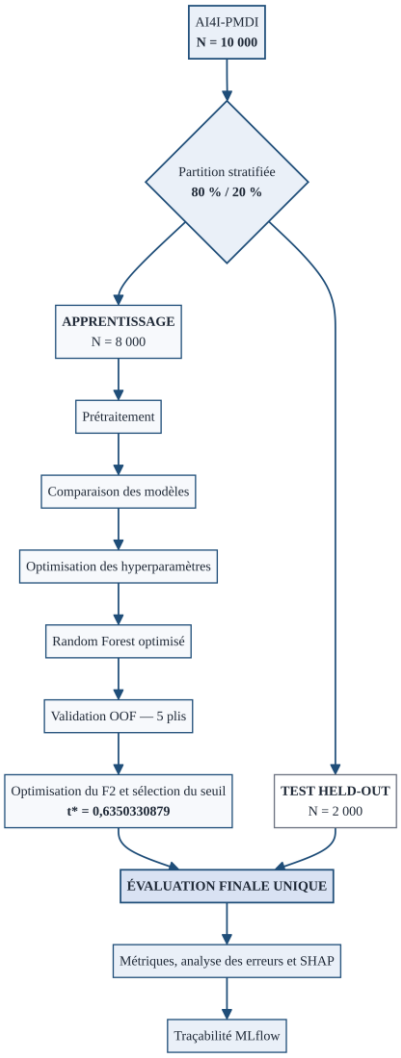

Figure 1 — Protocole expérimental de l'étude. *Le jeu de test est maintenu hors des opérations d'optimisation des hyperparamètres et de sélection du seuil.*

3.5. Modèles candidats

Trois familles de classifieurs sont étudiées :

- régression logistique ;
- arbre de décision ;
- Random Forest.

Pour chacune, une configuration initiale et une configuration optimisée sont évaluées.

3.6. Optimisation du Random Forest

Les hyperparamètres du Random Forest sont optimisés par RandomizedSearchCV.

Quarante configurations candidates sont évaluées selon une validation croisée stratifiée à cinq plis. L'Average Precision constitue la fonction de score de la recherche.

La meilleure Average Precision moyenne obtenue en validation croisée est :

$$AP\_CV = 0{,}8982431945578817$$

Hyperparamètre	Valeur retenue
n_estimators	500
max_depth	8
min_samples_split	20
min_samples_leaf	5
max_features	None
class_weight	balanced_subsample
bootstrap	True

Tableau 2 — Hyperparamètres du Random Forest optimisé. *Ces valeurs correspondent à la configuration effectivement sélectionnée par le pipeline parmi 40 configurations candidates.*

3.7. Sélection du seuil décisionnel

La sélection des hyperparamètres et celle du seuil constituent deux opérations distinctes.

Après optimisation du Random Forest, des probabilités *out-of-fold* sont générées sur le seul jeu d'entraînement au moyen d'une validation croisée stratifiée à cinq plis.

Pour chaque seuil candidat t :

$$\hat{y}_i(t) = \begin{cases} 1, & \hat{p}_i \geq t \\ 0, & \hat{p}_i < t \end{cases}$$

Le seuil maximisant le score F2 est retenu :

$$t^* = 0{,}6350330879616113$$

Le jeu de test n'intervient pas dans cette optimisation.

3.8. Évaluation finale

Après gel du seuil, le modèle est évalué une seule fois sur le jeu de test indépendant.

Les métriques étudiées sont Accuracy, Precision, Recall, F1, F2, ROC-AUC et Average Precision, complétées par la matrice de confusion.

3.9. Interprétabilité

L'analyse post-hoc comporte :

- l'importance intrinsèque des variables ;
- l'importance par permutation ;
- une analyse SHAP globale ;
- une analyse SHAP locale des faux négatifs.

3.10. Reproductibilité et traçabilité

Le workflow est orchestré sous Kedro et le suivi expérimental est assuré par MLflow.

Le run canonique associé aux résultats finals est : *8fecf4f65fe24632a9486bb3e1b569d5*

Le modèle enregistré est le **Random Forest Tuned**, version **4**, avec l'alias : champion.

4. Résultats

4.1. Comparaison des modèles

Modèle	Version	Accuracy	Precision	Recall	F1	ROC-AUC	AP
Régression logistique	Initiale	0,8880	0,2308	0,9429	0,3708	0,9397	0,2863
Régression logistique	Optimisée	0,8850	0,2297	0,9714	0,3716	0,9469	0,3228
Arbre de décision	Initiale	0,9750	0,5877	0,9571	0,7283	0,9718	0,9269
Arbre de décision	Optimisée	0,9750	0,5862	0,9714	0,7312	0,9804	0,9363
Random Forest	Initiale	0,9550	0,4359	0,9714	0,6018	0,9779	0,9292
Random Forest	Optimisée	0,9825	0,6804	0,9429	0,7904	0,9813	0,9444

Tableau 1 — Performances comparées des modèles (seuil conventionnel de 0,50). *Le Random Forest optimisé présente la meilleure Average Precision, le meilleur ROC-AUC et le meilleur F1 parmi les configurations comparées. Il est par conséquent retenu pour la suite du protocole.*

4.2. Résultats OOF et sélection du seuil

Indicateur	Valeur
Seuil décisionnel	0,6350330879616113
Precision	0,7423
Recall	0,8705
F1	0,8013
F2	0,8414
TP	242
FP	84
TN	7 638
FN	36
Observations	8 000

Tableau 3 — Performances out-of-fold au seuil décisionnel sélectionné.

Les valeurs exactes sont :

$$Precision_{OOF} = 0{,}7423312883$$

$$Recall_{OOF} = 0{,}8705035971$$

$$F_{1,OOF} = 0{,}8013245033$$

$$F_{2,OOF} = 0{,}8414464534$$

4.3. Évaluation sur le jeu de test indépendant

Métrique	Valeur
Accuracy	0,9880
Precision	0,7674
Recall	0,9429
F1	0,8462
F2	0,9016
ROC-AUC	0,9813
Average Precision	0,9444

Tableau 4 — Performances finales du Random Forest optimisé sur le jeu de test indépendant.

Au seuil décisionnel sélectionné hors test, le modèle atteint ainsi une accuracy de **98,80 %**, une précision de **76,74 %** et un rappel de **94,29 %**.

Le score :

$$F2 = 0{,}9016393443$$

traduit la capacité du classifieur à maintenir un rappel élevé tout en conservant une précision substantielle.

ROC-AUC et Average Precision sont calculées à partir des scores probabilistes et ne dépendent donc pas du choix particulier du seuil final.

4.4. Matrice de confusion

Classe réelle	Prédit : absence	Prédit : défaillance	Total
Absence de défaillance	1 910	20	1 930
Défaillance	4	66	70
Total	1 914	86	2 000

Tableau 5 — Matrice de confusion du Random Forest optimisé sur le jeu de test indépendant.

La cohérence est vérifiée :

$$1910 + 20 + 4 + 66 = 2000$$

Par ailleurs :

$$66 + 4 = 70 \text{ défaillances réelles}$$

	Prédit 0	Prédit 1
Réel 0	1910	20
Réel 1	4	66

Figure 2 — Matrice de confusion du Random Forest optimisé. *Matrice de confusion du Random Forest optimisé sur le jeu de test indépendant au seuil t* = 0,635033.*

Le rappel peut être reconstruit directement :

$$Recall = 66 / (66 + 4) = 0{,}942857$$

Le modèle identifie donc 66 des 70 défaillances présentes dans le test.

4.5. Analyse des faux négatifs

Observation	Classe réelle	Classe prédite	P(défaillance)	Seuil
442	1	0	0,281611	0,635033
883	1	0	0,017819	0,635033
1588	1	0	0,275008	0,635033
1976	1	0	0,017819	0,635033

Tableau 6 — Faux négatifs du Random Forest optimisé.

Toutes les probabilités sont cohérentes avec la règle : $P(Y = 1) < t^*$

Les erreurs ne présentent toutefois pas toutes le même profil : les observations 442 et 1588 reçoivent des probabilités intermédiaires, tandis que les observations 883 et 1976 sont classées avec une probabilité de défaillance particulièrement faible.

5. Interprétabilité

5.1. Importance intrinsèque des variables

L'importance intrinsèque du Random Forest permet d'étudier la contribution relative des variables aux divisions opérées dans les arbres.

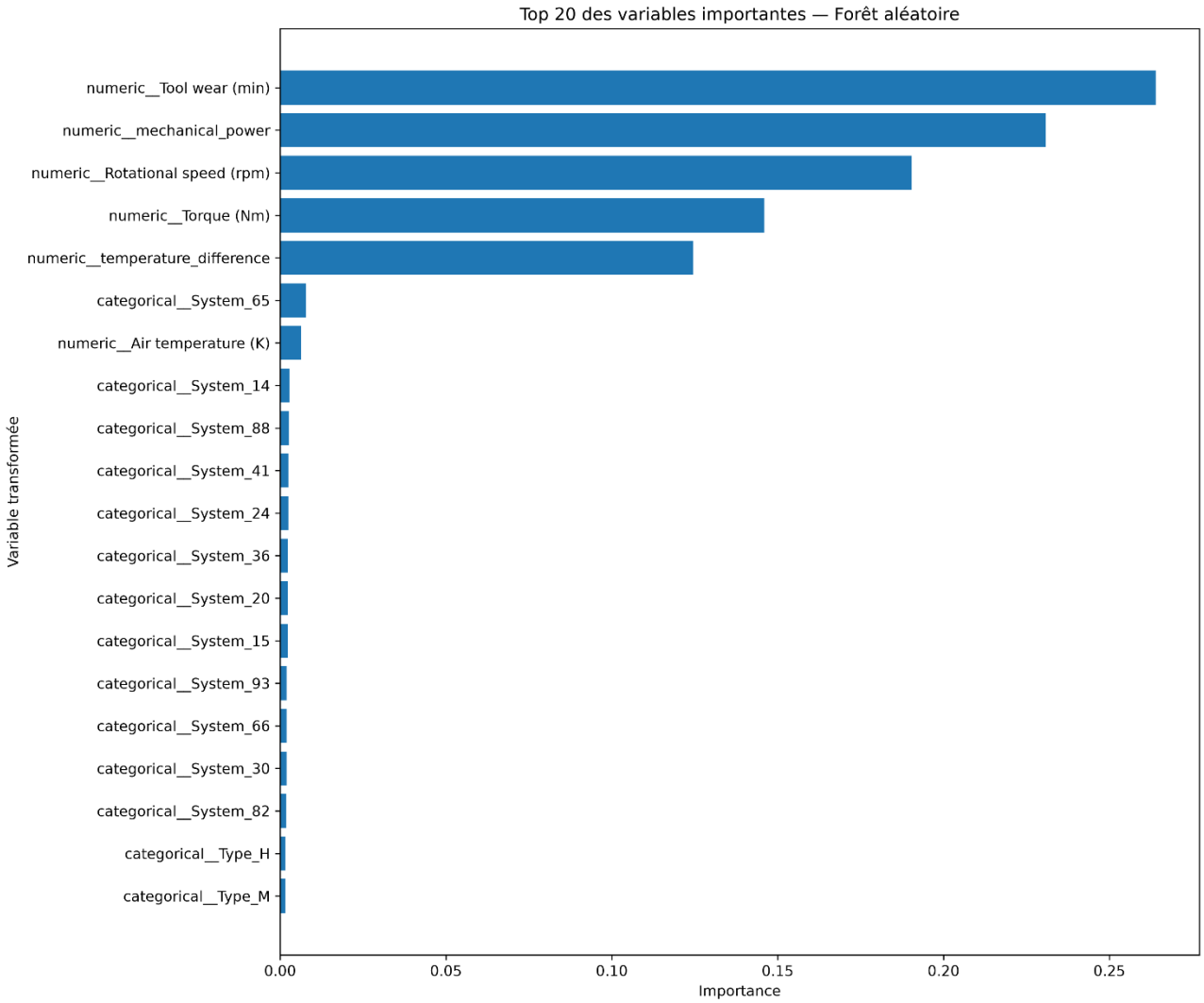

Figure 3 — Importance intrinsèque des variables du Random Forest optimisé. *Cette mesure renseigne sur l'utilisation relative des variables dans les arbres du modèle ; elle ne démontre pas une relation causale entre les variables et les défaillances.*

5.2. Importance par permutation

L'importance par permutation mesure la dégradation de la performance lorsque les valeurs d'une variable sont perturbées aléatoirement.

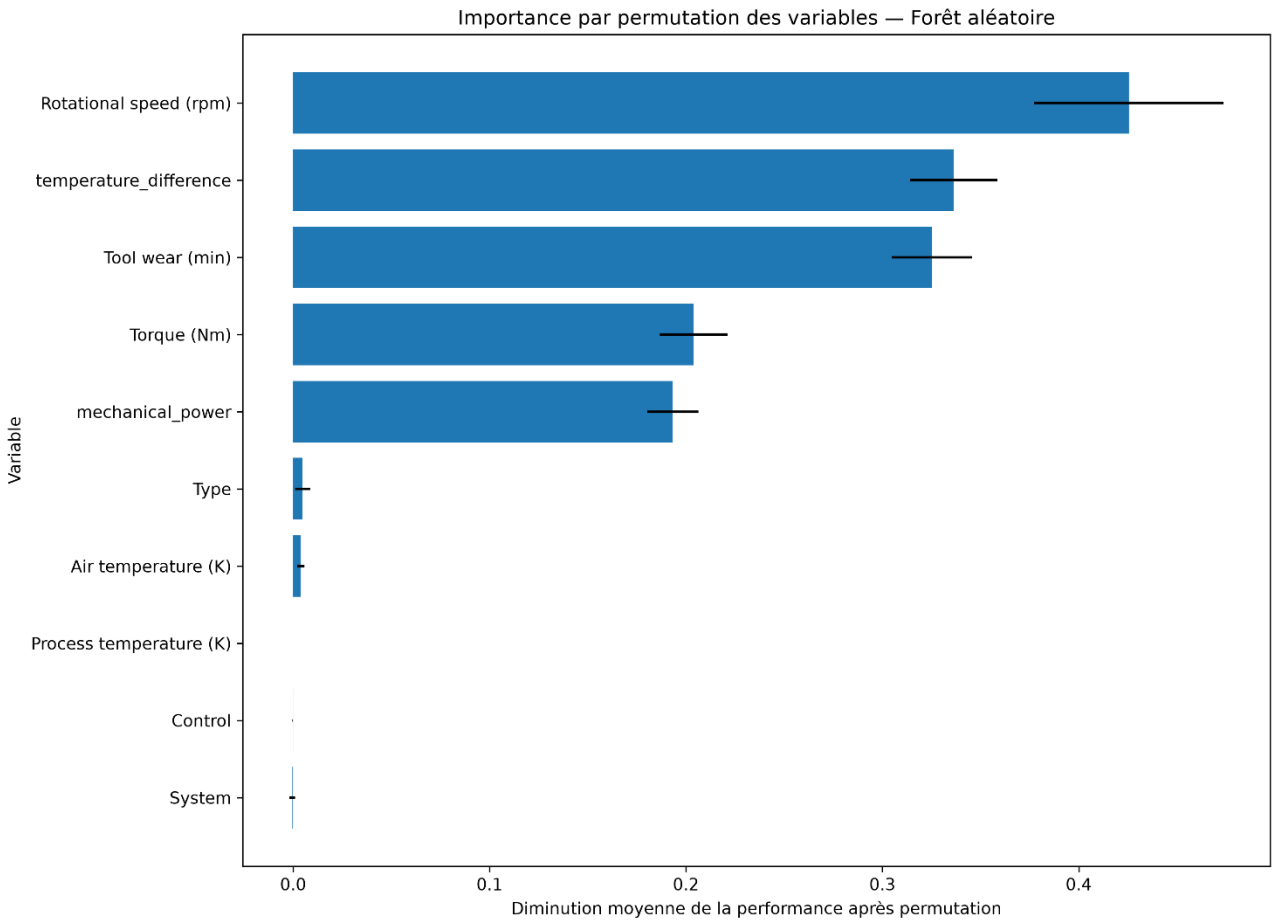

Figure 4 — Importance des variables par permutation. *La méthode fournit un contrôle complémentaire de l'importance intrinsèque en quantifiant l'effet de la perturbation d'une variable sur la performance du modèle.*

Les deux méthodes ne mesurent pas exactement le même objet ; leur confrontation est donc plus informative que l'utilisation isolée d'un seul classement.

5.3. Explicabilité SHAP globale

SHAP fournit une représentation additive des contributions des variables à la sortie du modèle (Lundberg & Lee, 2017).

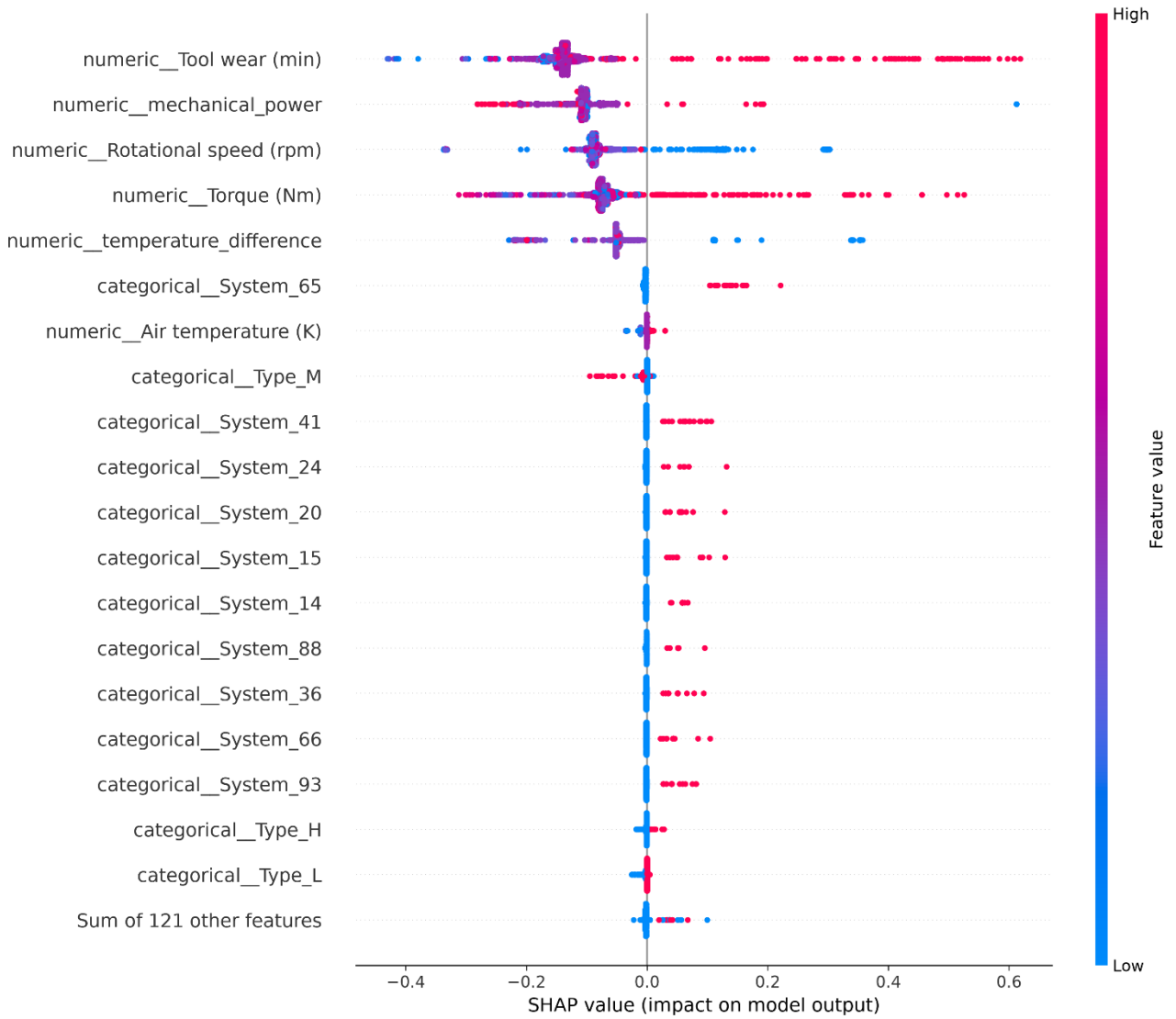

Figure 5 — Distribution globale des contributions SHAP. *Le beeswarm plot représente simultanément l'importance relative des variables, la distribution des valeurs SHAP et le sens de leurs contributions aux sorties du modèle. Les valeurs SHAP expliquent le comportement prédictif du classifieur ; elles ne constituent pas une démonstration de causalité physique.*

5.4. Explicabilité locale des faux négatifs

Les quatre faux négatifs sont particulièrement importants du point de vue de la maintenance : ils représentent des défaillances réelles que le modèle n'a pas signalées.

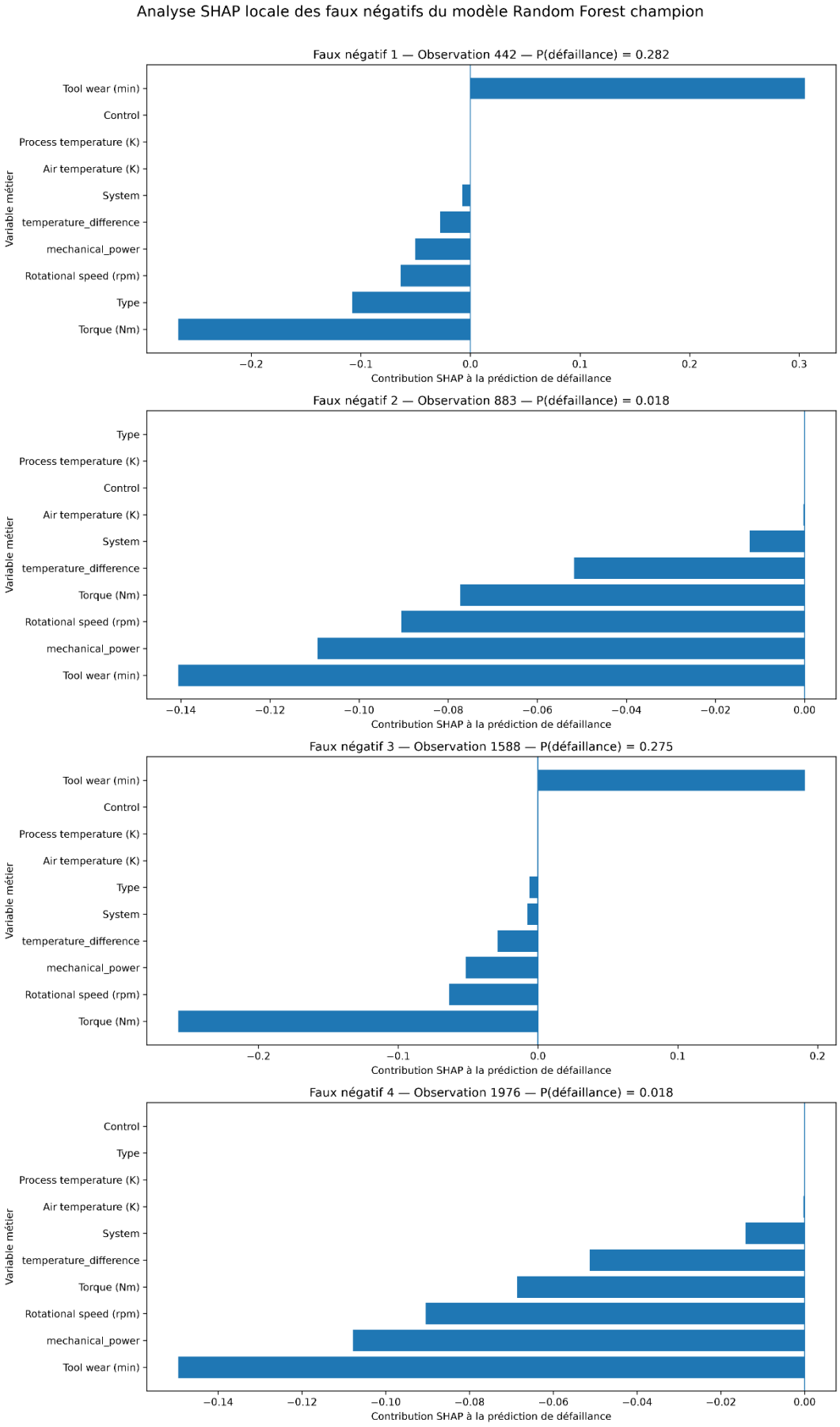

Figure 6 — Analyse SHAP locale des faux négatifs. *La figure présente les contributions locales associées aux quatre observations de défaillance non détectées au seuil t* = 0,635033. Elle permet d'identifier les variables ayant orienté la sortie du modèle vers ou à l'opposé de la classe positive.*

5.5. Complémentarité des méthodes

Les trois familles d'explications répondent à des questions distinctes. L'importance intrinsèque décrit l'utilisation des variables par les arbres ; la permutation mesure leur importance relativement à la performance ; SHAP examine leurs contributions aux sorties du modèle.

Une variable peut donc occuper des positions différentes selon la méthode sans que cela constitue nécessairement une contradiction.

6. Discussion

6.1. Performance globale

Le Random Forest optimisé présente une capacité discriminante élevée :

$$ROC\text{-}AUC = 0{,}9813 \quad et \quad AP = 0{,}9444$$

La matrice de confusion confirme cette performance sur la classe minoritaire puisque 66 des 70 défaillances sont correctement détectées.

L'accuracy de 98,80 % doit néanmoins être interprétée avec prudence compte tenu du taux de défaillance de seulement 3,5 % dans le jeu de test. Cette situation justifie l'utilisation conjointe de Precision, Recall, F2 et Average Precision. La littérature montre précisément l'intérêt des mesures Precision–Recall pour les problèmes déséquilibrés.

6.2. Effet du seuil décisionnel

Au seuil conventionnel de 0,50, le Random Forest optimisé présente :

$$Precision = 0{,}6804\ ;\ Recall = 0{,}9429\ ;\ F1 = 0{,}7904$$

Au seuil final de 0,635033 :

$$Precision = 0{,}7674\ ;\ Recall = 0{,}9429\ ;\ F1 = 0{,}8462$$

Sur ce jeu de test particulier, l'augmentation du seuil améliore donc la précision et le F1 sans diminuer le rappel observé.

Il serait cependant incorrect d'en déduire une propriété générale de l'augmentation du seuil. Ce comportement dépend de la distribution particulière des probabilités produites par le modèle.

6.3. OOF et test indépendant

Une distinction essentielle doit être maintenue entre :

$$F_{2,OOF} = 0{,}841446$$

et

$$F_{2,test} = 0{,}901639$$

Le premier est associé à la procédure utilisée pour sélectionner le seuil sur le jeu d'entraînement. Le second est la performance observée après gel du seuil sur le test indépendant.

Le résultat test, supérieur au résultat OOF, ne doit donc pas être utilisé rétroactivement pour justifier le seuil.

6.4. Faux négatifs

Les quatre faux négatifs représentent :

$$\frac{4}{70} \times 100 \approx 5,71\%$$

des défaillances réelles du test.

Leur analyse révèle deux catégories de situations : des observations relativement plus proches de la frontière de décision et deux observations auxquelles le modèle attribue une probabilité de défaillance très faible.

Cette hétérogénéité montre l'intérêt de compléter les métriques agrégées par une analyse locale.

6.5. Interprétabilité et causalité

Les méthodes d'interprétabilité ne permettent pas, à elles seules, d'établir des mécanismes causaux. Une contribution SHAP positive signifie qu'une variable contribue à augmenter la sortie du modèle relativement à une valeur de référence ; elle ne démontre pas que cette variable a physiquement causé la défaillance.

Cette distinction est particulièrement importante dans les environnements industriels, où une interprétation causale exige généralement une combinaison de connaissances physiques, d'expertise métier et de protocoles expérimentaux adaptés.

6.6. Limites du dataset

AI4I-PMDI améliore le réalisme du dataset AI4I en introduisant notamment des données manquantes et des irrégularités d'acquisition. Autran et al. montrent précisément que ces modifications peuvent affecter les performances des modèles.

Le corpus demeure cependant dérivé d'un benchmark synthétique. Les performances obtenues ne peuvent donc pas être assimilées à celles qui seraient nécessairement obtenues sur des équipements industriels réels.

6.7. Absence de validation temporelle et externe

Le partitionnement stratifié permet une évaluation held-out cohérente pour la classification considérée, mais ne constitue pas une validation temporelle.

Or, dans un environnement industriel, plusieurs phénomènes peuvent apparaître :

- vieillissement des équipements ;
- modification des régimes de fonctionnement ;
- changement de capteurs ;
- dérive des distributions ;

- modification des procédures de maintenance ;
- apparition de nouveaux modes de défaillance.

Une validation externe sur d'autres équipements, périodes ou sites serait donc nécessaire avant toute généralisation.

6.8. Classification versus pronostic

Le modèle estime :

$$P(Y = 1 / X)$$

Il n'estime pas :

RUL(t)

et ne prédit pas directement l'instant futur de défaillance.

L'étude doit ainsi être interprétée comme une **classification de défaillances appliquée au domaine de la maintenance prédictive**, et non comme un système complet de pronostic industriel.

6.9. Contribution scientifique

La contribution principale du travail n'est pas la proposition d'un nouvel algorithme. Elle réside dans l'intégration cohérente de plusieurs exigences méthodologiques :

- comparaison de plusieurs modèles ;
- optimisation reproductible ;
- prévention des fuites de cible ;
- séparation apprentissage/test ;
- sélection du seuil sur prédictions OOF ;
- évaluation finale indépendante ;
- analyse des erreurs ;
- explicabilité ;
- traçabilité expérimentale.

7. Perspectives et architecture cible

7.1. Passage du prototype au système industriel

Une industrialisation nécessiterait de dépasser le traitement d'un dataset statique pour mettre en place une chaîne de données opérationnelle reliant équipements, acquisition, traitement, modèle et systèmes de maintenance.

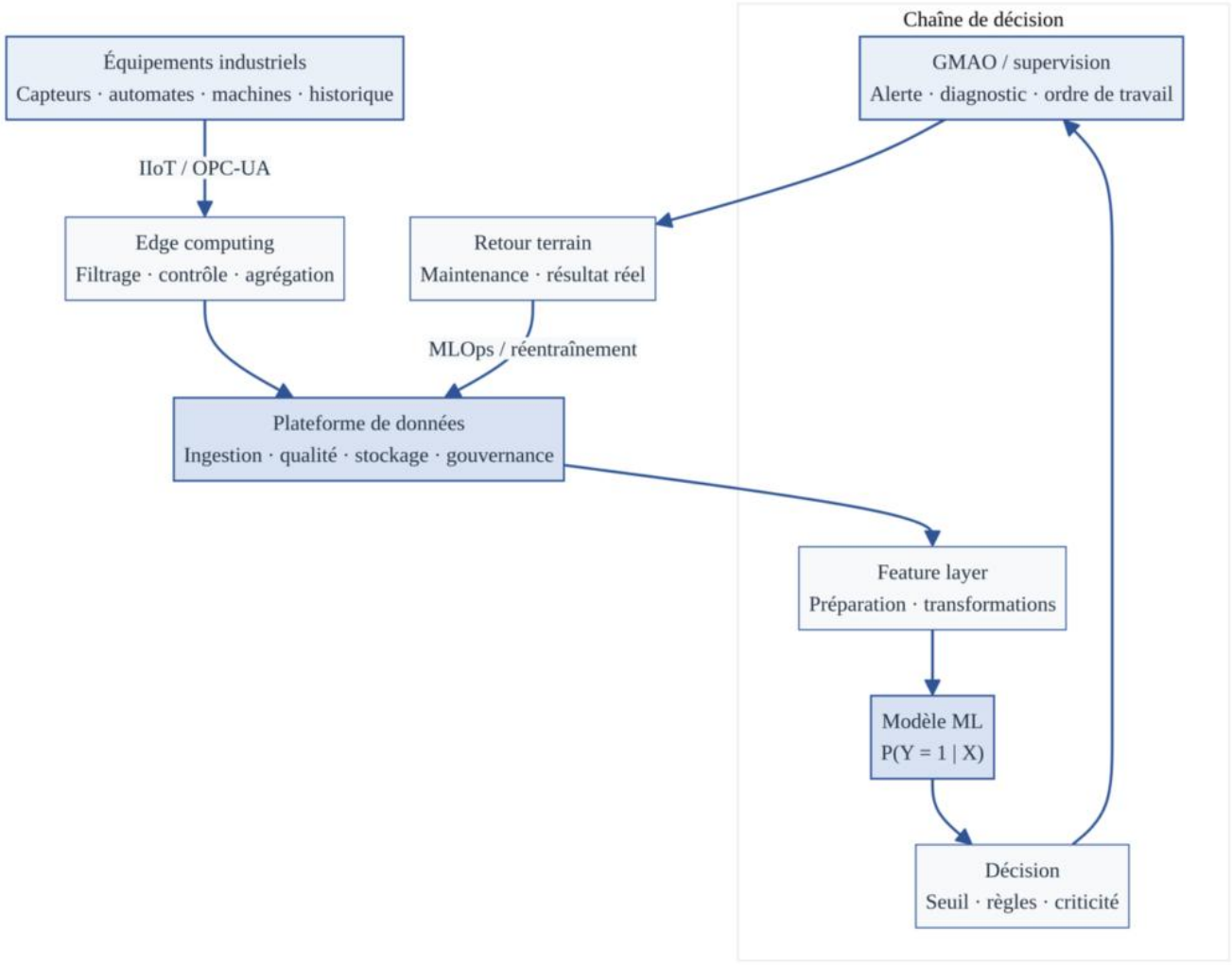

Figure 7 — Architecture conceptuelle cible d'un système industriel de maintenance prédictive. *Cette architecture constitue une perspective d'industrialisation ; elle n'a pas été implémentée ni évaluée dans l'expérience présentée.*

7.2. IIoT et Edge Computing

L'IIoT permettrait l'acquisition des mesures opérationnelles. Un niveau Edge pourrait effectuer les contrôles de qualité, filtrages ou agrégations nécessitant une faible latence avant transmission vers une plateforme centralisée.

7.3. Intégration à la GMAO

La sortie probabiliste du modèle ne devrait pas nécessairement déclencher directement une intervention. Elle pourrait alimenter un moteur de décision combinant :

$$P(Y = 1 \mid X)$$

criticité de l'équipement, conséquences de la défaillance, historique de maintenance et règles opérationnelles.

La décision pourrait ensuite être transmise à la GMAO pour créer une alerte, une demande d'inspection ou un ordre de travail.

7.4. MLOps et gouvernance

L'exploitation continue nécessiterait notamment :

- versionnement des données ;
- versionnement des modèles ;
- surveillance de la dérive ;
- monitoring des performances ;
- gestion des seuils ;
- traçabilité des décisions ;
- contrôle des accès ;
- procédures de validation des nouvelles versions.

7.5. Séries temporelles et RUL

Une extension scientifique naturelle consisterait à exploiter explicitement la dimension temporelle.

Des séquences longitudinales pourraient permettre d'étudier :

$$P(Y_{t+h} = 1 \mid X_{\leq t}) \quad ou \quad RUL(t).$$

Ces problématiques sont plus proches du pronostic au sens strict que la classification instantanée étudiée ici.

Domaine	Composante	Fonction principale
Acquisition	Capteurs, automates, IIoT	Collecte des variables opérationnelles
Edge	Filtrage, contrôle, agrégation	Prétraitement à proximité des actifs
Données	Pipeline d'ingestion et stockage	Centralisation et historisation
Qualité	Contrôles et règles de validation	Fiabilisation des données
Modélisation	Feature engineering et modèle ML	Estimation de P(Y=1|X)
Décision	Seuil, criticité, règles métier	Conversion du score en action potentielle
Maintenance	GMAO	Gestion des inspections et interventions
MLOps	Registry, monitoring, réentraînement	Gestion du cycle de vie du modèle
Gouvernance	IAM, audit, traçabilité	Contrôle et responsabilité
Cybersécurité	Segmentation, chiffrement, supervision	Protection de la chaîne industrielle

Tableau 7 — Composantes proposées pour l'industrialisation.

Conclusion générale

Cette étude a développé et évalué une chaîne reproductible d'apprentissage automatique destinée à la classification des défaillances d'équipements industriels à partir du dataset AI4I-PMDI.

Sur les 10 000 observations considérées, 8 000 ont été affectées au développement du modèle et 2 000 conservées pour l'évaluation finale indépendante. Trois familles de classifieurs ont été comparées. Le **Random Forest optimisé** a été retenu comme modèle champion.

L'optimisation par RandomizedSearchCV, portant sur 40 configurations et une validation croisée stratifiée à cinq plis, conduit à un modèle comprenant 500 arbres, une profondeur maximale de 8, min_samples_split = 20, min_samples_leaf = 5, max_features = None, class_weight = balanced_subsample et bootstrap = True. La meilleure Average Precision moyenne obtenue pendant cette optimisation est de :

$$0{,}898243$$

Une seconde procédure, indépendante de l'optimisation des hyperparamètres, a ensuite utilisé les prédictions OOF du jeu d'entraînement pour maximiser le F2-score. Le seuil sélectionné est :

$$t^* = 0{,}6350330879616113$$

Au seuil retenu, la performance OOF est :

$$F_{2,OOF} = 0{,}841446$$

L'application du modèle et du seuil figé au jeu de test indépendant conduit à :

$$Accuracy = 98{,}80\,\% \; ; \; Precision = 76{,}74\,\% \; ; \; Recall = 94{,}29\,\%$$

$$F1 = 84{,}62\,\% \; ; \; F2 = 90{,}16\,\%$$

$$ROC\text{-}AUC = 0{,}9813 \; ; \; Average\ Precision = 0{,}9444$$

La matrice de confusion finale comporte 1 910 vrais négatifs, 20 faux positifs, 4 faux négatifs et 66 vrais positifs. Le modèle identifie ainsi 66 des 70 défaillances du jeu de test.

L'analyse d'interprétabilité complète ces résultats en mobilisant importance intrinsèque, importance par permutation et SHAP. L'étude des quatre faux négatifs montre en particulier que toutes les erreurs ne présentent pas le même profil probabiliste, ce qui confirme l'intérêt d'une analyse locale au-delà des seules métriques agrégées.

La démarche présente néanmoins plusieurs limites. AI4I-PMDI reste dérivé d'un benchmark synthétique ; le test indépendant provient du même corpus que l'apprentissage ; aucune validation externe ou temporelle n'a été réalisée ; enfin, le problème étudié est une classification binaire et non une estimation de RUL. Les performances rapportées constituent donc une **preuve de faisabilité expérimentale**, et non la démonstration d'une performance équivalente dans un environnement industriel réel.

La contribution principale réside dès lors dans la cohérence de la chaîne expérimentale : préparation des données, comparaison des modèles, optimisation, validation OOF, sélection du seuil sans exploitation du test, évaluation held-out, analyse des erreurs, explicabilité et traçabilité MLflow.

La poursuite de ces travaux devra porter prioritairement sur des données industrielles longitudinales et sur une validation externe. L'intégration à une architecture IIoT–Edge–Data–ML–GMAO permettrait ensuite d'étudier le passage d'un modèle expérimental à un véritable système d'aide à la décision de maintenance.

Bibliographie

Autran, J.-V., Kuhn, V., Diguet, J.-P., Dubois, M., & Buche, C. (2024). AI4I-PMDI: Predictive maintenance datasets with complex industrial settings' irregularities. Procedia Computer Science, 246, 1201–1209.
https://doi.org/10.1016/j.procs.2024.09.546

Autran, J.-V. (2024). AI4I-PMDI: Predictive Maintenance Dataset with Irregularities [Dataset]. Figshare.
https://doi.org/10.6084/m9.figshare.24718428.v1

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
https://doi.org/10.1023/A:1010933404324

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P. T. V., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024.
https://doi.org/10.1016/j.cie.2019.106024

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Matzka, S. (2020). Explainable artificial intelligence for predictive maintenance applications. In 2020 Third International Conference on Artificial Intelligence for Industries (AI4I) (pp. 69–74). IEEE.
https://doi.org/10.1109/AI4I49448.2020.00023

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.
https://doi.org/10.1371/journal.pone.0118432

Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. Computers & Industrial Engineering, 150, 106889.
https://doi.org/10.1016/j.cie.2020.106889