International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

La Performance Qui Trompe Valider L'apprentissage Automatique Pour UN Diagnostic Industriel Réellement Généralisable

Professeur Dr. Jean Claude Mukaz Ilunga[1], Ingénieur Olivier Mulagizi Bashebereka[2]

[1]Professeur Dr Ir, Maintenance, Institut Supérieur des Techniques Appliquées (ISTA) Kinshasa
[2]Etudiant en Master 2, Réseaux et cybersécurité/ Ingénierie Biomédicale, Institut Supérieur des Techniques Appliquées (ISTA) Kinshasa

Résumé

Les performances élevées rapportées en diagnostic automatique masquent fréquemment des biais de validation, un déséquilibre extrême des classes et un décalage entre apprentissage et exploitation. Cet article propose une démarche de validation orientée vers l'usage industriel. Un jeu simulé de 2 880 exemples, huit descripteurs, quatre états et douze machines permet de comparer détection d'anomalie, classification supervisée et plusieurs protocoles d'évaluation. La distance de Mahalanobis, apprise uniquement sur l'état sain, détecte 81 % des états dégradés pour 0,67 % de fausses alarmes. La partition aléatoire surestime la performance de 13 points pour un modèle rigide et de 28 points pour un modèle à mémoire, car le modèle reconnaît la machine plutôt que le défaut. Un changement de charge fait chuter la performance de 78,2 % à 66,1 %, tandis qu'un recalage sans étiquette la restaure partiellement. Sous une prévalence de 1 défaut pour 200 états sains, l'AUC ROC reste stable alors que la précision s'effondre. L'apport de l'auteur est le cadre VIGIE: Validation sans fuite, déséquilibre, Généralisation, Interprétabilité et Exploitation qui replace la preuve de robustesse au centre du diagnostic par apprentissage automatique.

Mots-clés : apprentissage automatique ; diagnostic industriel ; fuite d'information ; déséquilibre ; généralisation ; détection d'anomalie ; interprétabilité.

Abstract

High reported performance in automated machinery diagnosis often conceals validation leakage, extreme class imbalance and shifts between training and operating conditions. This article proposes an evaluation methodology designed for industrial use. A simulated dataset of 2,880 examples, eight descriptors, four states and twelve machines is used to compare anomaly detection, supervised classification and validation protocols. Mahalanobis distance trained only on healthy data detects 81% of degraded states with 0.67% false alarms. Random splitting overestimates performance by 13 points for a rigid model and by 28 points for a memory-based model because it recognizes machines rather than faults. A load shift reduces performance from 78.2% to 66.1%, while unsupervised recalibration partly restores it. At a prevalence of one fault per 200 healthy states, ROC AUC remains stable while precision collapses. The author

contributes the VIGIE framework: Leakage-free Validation, Imbalance, Generalization, Interpretability and Exploitation placing evidence of robustness at the centre of machine-learning diagnosis.
Keywords: machine learning; machinery diagnosis; data leakage; class imbalance; generalization; anomaly detection; interpretability.

1. Introduction : l'écart entre laboratoire et exploitation

La littérature du diagnostic par apprentissage automatique compare volontiers réseaux, machines à vecteurs de support, forêts et méthodes de voisinage sur des jeux de données segmentés. Pourtant, la différence entre une performance publiée et une performance réellement obtenue provient souvent moins de l'algorithme que de la façon dont les données sont découpées, des conditions qu'elles représentent et des métriques retenues. Des segments issus du même essai présents à la fois dans l'apprentissage et le test permettent au modèle de reconnaître une installation, un capteur ou un régime plutôt que le défaut [1–4].
La question de recherche est donc : comment évaluer un modèle de diagnostic pour que sa performance mesure sa capacité à reconnaître un défaut sur une machine et des conditions non vues ? L'objectif est de proposer un protocole complet associant formulation progressive du diagnostic, métriques adaptées au déséquilibre, séparation par groupes, épreuve de domaine, interprétabilité et surveillance après déploiement. L'hypothèse est qu'un modèle simple correctement validé est plus utile qu'un modèle complexe dont la performance est un artefact de protocole.

2. Méthodologie et cas d'étude

Le jeu de données simulé contient 2 880 observations décrites par huit descripteurs issus de la vibration, de l'enveloppe, de la température et d'indicateurs statistiques. Quatre états sont représentés : sain, défaut de roulement, défaut d'engrenage et balourd. Les données proviennent de douze machines, chacune affectée d'un biais de montage propre. Cette structure permet de mesurer explicitement la fuite liée à l'identité de la machine.
L'étude compare des modèles d'anomalie appris sur l'état sain, des classifieurs supervisés de flexibilité différente, deux stratégies de partition aléatoire ou par machine et un changement de charge. Les performances sont analysées par matrice de confusion, rappel, spécificité, précision, courbes ROC et précision–rappel, calibration et importance par permutation. Les résultats absolus décrivent la simulation ; les écarts entre protocoles décrivent des mécanismes généralisables.

3. Trois niveaux de diagnostic, trois exigences de données

Question	Sortie	Données nécessaires	Déploiement réaliste
Détection	Normal ou anormal	Historique sain	Dès la disponibilité d'une référence
Identification	Mode de défaillance	Exemples de chaque mode	Après accumulation d'événements
Sévérité	Stade ou grandeur continue	Trajectoires complètes étiquetées	Après retour d'expérience suffisant

Tableau 1 Formulations progressives du diagnostic. Source : synthèse de l'auteur.

Un mode absent de l'apprentissage ne peut pas être identifié : il sera forcé vers une classe connue, parfois avec forte confiance. La détection d'anomalie constitue donc le point de départ le plus robuste. À mesure que les défauts sont documentés, le système peut être enrichi d'une identification, puis d'une évaluation de sévérité. Cet ordre évite de promettre des fonctions pour lesquelles les données n'existent pas.

4. Détection d'anomalie à partir de l'état sain

La distance de Mahalanobis mesure l'écart à la moyenne saine en tenant compte de la covariance entre descripteurs. Elle détecte ainsi un point qui paraît normal variable par variable mais viole leur relation habituelle. Sous hypothèse gaussienne, sa valeur quadratique suit approximativement une loi du khi-deux ; un seuil peut être fixé à partir d'un taux de fausses alarmes visé, sans exemple de défaut.

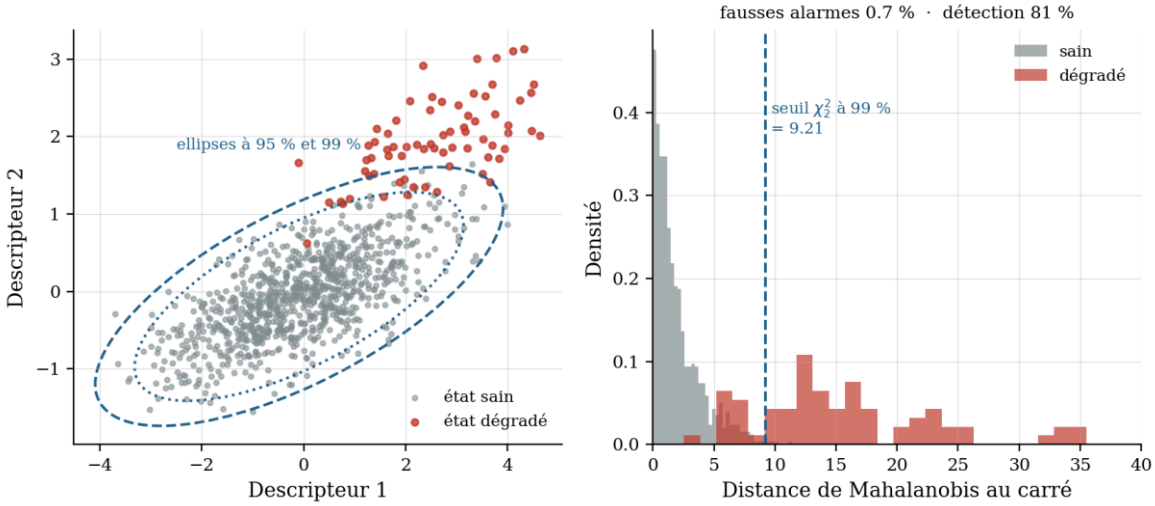

Figure 1 Détection par distance de Mahalanobis : seuil théorique, fausses alarmes et états dégradés. Source : simulation de l'auteur.

Le seuil au quantile 99 % produit 0,67 % de fausses alarmes et détecte 81 % des états dégradés. La méthode reste limitée par la non-gaussianité, les régimes multiples et l'instabilité de la covariance lorsque la dimension approche l'effectif. Des modèles conditionnés au régime, une covariance régularisée ou des méthodes non paramétriques peuvent être nécessaires.

4.1 Le seuil est une décision d'exploitation

Un taux unitaire de 1 % paraît faible, mais une évaluation toutes les dix minutes produit environ 52 000 décisions par an et plus de 500 alertes par point. La persistance exiger plusieurs franchissements consécutifs réduit fortement les fausses alertes au prix d'un retard mesurable. Le seuil doit résulter du coût relatif de la vérification inutile et de la non-détection, non d'une convention statistique isolée.

5. Classification supervisée sous faible effectif

Famille	Atout industriel	Risque principal	Contexte favorable
Linéaire / LDA	Peu de paramètres ; explicable	Frontière trop simple	Faible effectif, descripteurs physiques

Famille	Atout industriel	Risque principal	Contexte favorable
SVM à noyau	Bonne marge en dimension élevée	Réglage et calibration	Petits jeux non linéaires
Forêts / boosting	Interactions, robustesse tabulaire	Importance biaisée, surajustement	Descripteurs hétérogènes
Réseaux profonds	Apprentissage de représentations	Apprend le banc et le régime	Données abondantes et domaines proches
k-NN	Simplicité et mémoire des exemples	Très sensible à la fuite	Analyse exploratoire, démonstration

Tableau 2. Positionnement des familles de modèles. Source : synthèse de l'auteur.

Sur des descripteurs tabulaires et peu de défaillances, la flexibilité n'est pas automatiquement un avantage. Les descripteurs issus de la physique agissent comme une restriction utile : une amplitude à une fréquence de défaut calculée conserve une signification lorsque la vitesse change, alors qu'un motif librement appris peut correspondre à une particularité du banc.

6. Déséquilibre : quand 99,5 % d'exactitude ne détecte rien

Avec un défaut pour 200 états sains, prédire toujours « sain » obtient 99,5 % de bonnes réponses tout en ayant un rappel nul. L'exactitude globale devient donc trompeuse. La matrice de confusion, le rappel, la spécificité et surtout la précision doivent être rapportés au point de fonctionnement choisi. La courbe précision–rappel est plus informative que la ROC lorsque la classe positive est rare [5].

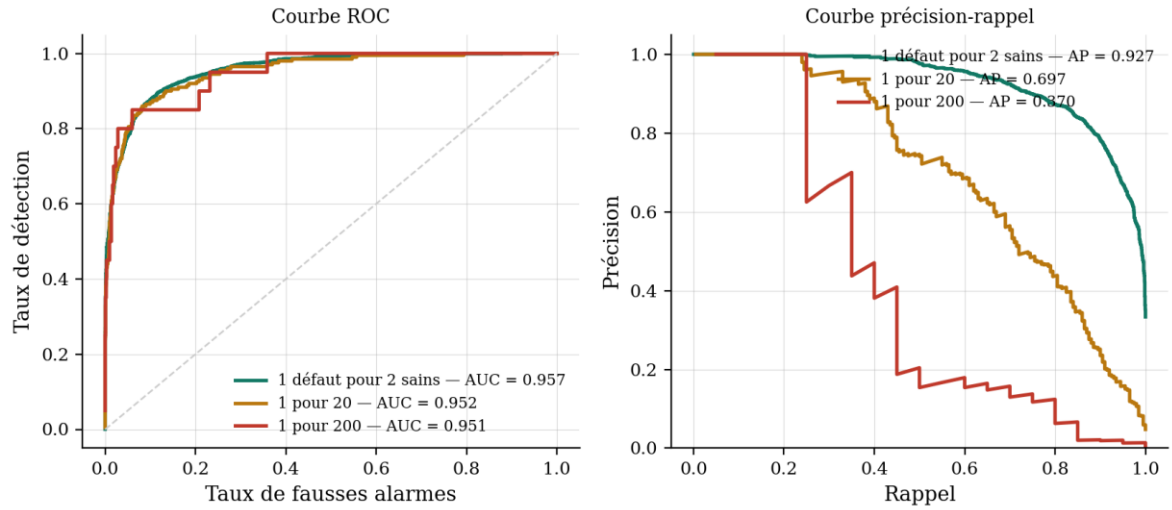

Figure 2. Effet de la prévalence : ROC stable, effondrement de la précision. Source : simulation de l'auteur.

La pondération des classes explicite le coût plus élevé des erreurs rares. Le déplacement du seuil est souvent préférable au rééchantillonnage : il préserve les données, la calibration et permet de modifier le

compromis en exploitation sans réentraîner. La synthèse d'exemples rares doit rester prudente lorsque les modes de défaillance sont physiquement différents.

7. Calibration : une confiance de 90 % doit signifier 90 %

Un classifieur peut bien classer tout en étant excessivement confiant. Or la confiance intervient directement dans la décision, la fusion et la hiérarchisation des interventions. Un diagramme de fiabilité compare la confiance annoncée à la fréquence réelle de succès ; un écart sous la diagonale signale une surestimation. Des méthodes de recalibration doivent être ajustées sur un échantillon distinct du test [6].

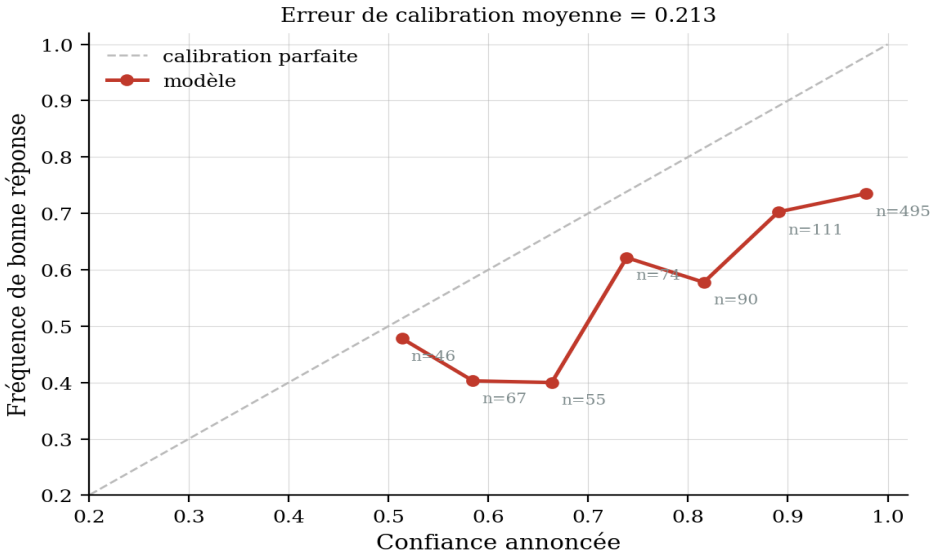

Figure 3. Diagramme de fiabilité et erreur de calibration. Source : simulation de l'auteur.

8. Résultat critique : la fuite d'information par le découpage

Une partition aléatoire de segments mélange des observations provenant des mêmes machines entre apprentissage et test. Un modèle à mémoire peut alors exploiter le biais de montage, le bruit ou la fonction de transfert propres à la machine. La performance mesure une reconnaissance d'identité. Pour un usage sur de nouvelles machines, la partition doit maintenir chaque machine dans un seul sous-ensemble.

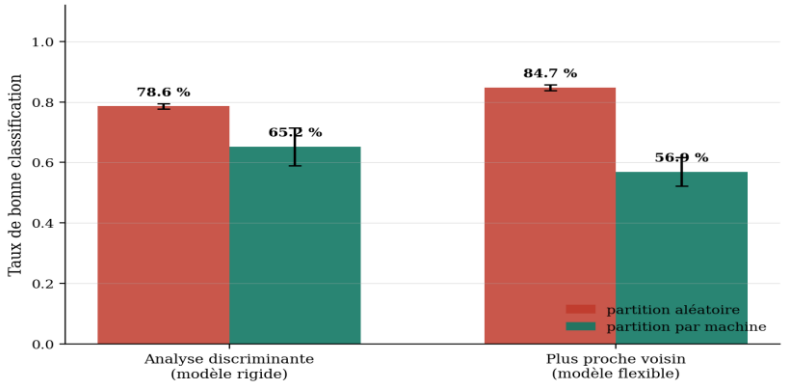

Figure 4. Performance selon la partition : surestimation de 13 à 28 points. Source : simulation de l'auteur.

Le modèle rigide perd 13 points lorsque la partition devient réaliste ; le modèle flexible en perd 28. Ainsi, un protocole permissif favorise artificiellement les algorithmes les plus capables de mémoriser. Une

hiérarchie publiée sous partition aléatoire peut être l'inverse de celle observée sur machines non vues. L'unité de partition doit correspondre à l'unité réelle de généralisation : machine, site, campagne ou patient selon l'application [3,7].

9. Protocole de validation sans fuite

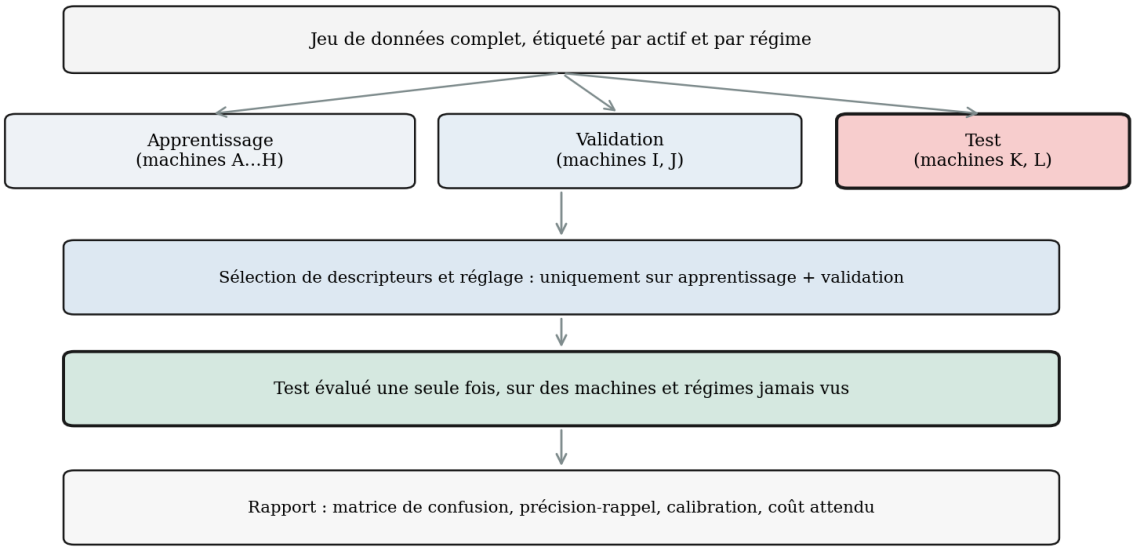

Figure 5. Séparation apprentissage, validation et test selon les groupes réels. Source : conception de l'auteur.

La sélection des descripteurs, la normalisation, l'imputation, le réglage des hyperparamètres et la calibration consomment de l'information. Toutes ces opérations doivent être apprises uniquement sur le sous-ensemble d'entraînement, puis appliquées sans recalcul aux autres. Le test reste scellé jusqu'à la décision finale. Lorsque l'effectif est faible, une validation croisée groupée fournit une estimation plus stable, à condition que toutes les observations d'une même machine restent dans le même pli.

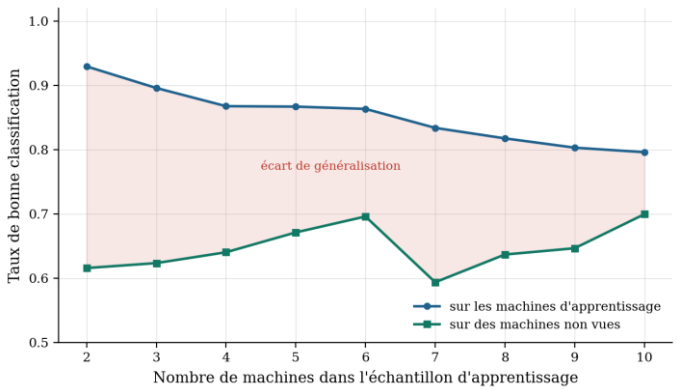

Figure 6. Écart entre apprentissage et machines non vues selon le nombre de machines disponibles. Source : simulation de l'auteur.

L'écart de généralisation se réduit avec le nombre de machines, non seulement avec le nombre de segments. Mille segments supplémentaires de la même machine apportent moins de diversité qu'une

nouvelle machine. La planification de la collecte doit donc viser la variété des installations et des conditions.

10. Décalage de domaine et hybridation physique–données

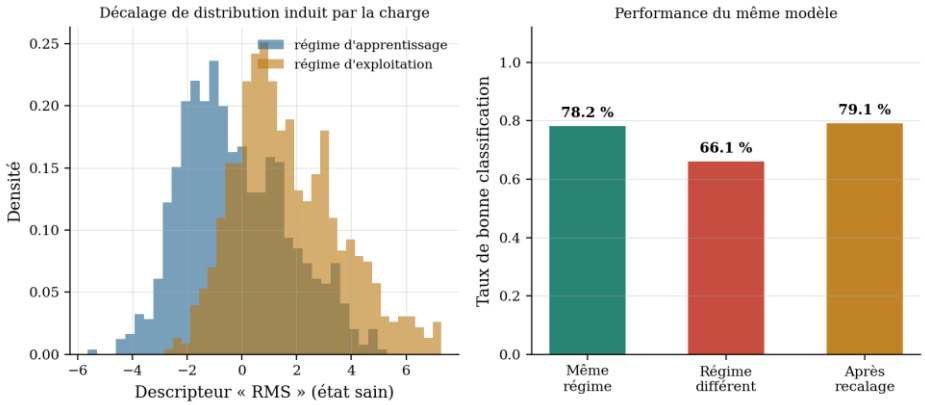

Figure 7. Chute sous changement de charge et restauration partielle par recalage. Source : simulation de l'auteur.

Le même modèle passe de 78,2 % en régime identique à 66,1 % sous charge différente. Un recentrage et une remise à l'échelle estimés sur des données saines du nouveau domaine restaurent une partie de la performance sans étiquette. Ce résultat montre qu'une simple adaptation statistique peut être rentable, mais seulement si la relation entre descripteurs et défaut reste stable.

La stratégie la plus robuste injecte la physique en amont : fréquences recalculées avec la vitesse, ordres, résidus par rapport à un modèle thermique ou électrique, démodulation orientée par la structure. Les données simulées peuvent compléter les défauts rares, mais introduisent un écart simulation–réalité qui doit être mesuré. L'apprentissage contraint par monotonie ou invariance constitue une perspective prometteuse [8,9].

11. Résultats diagnostiques et interprétabilité

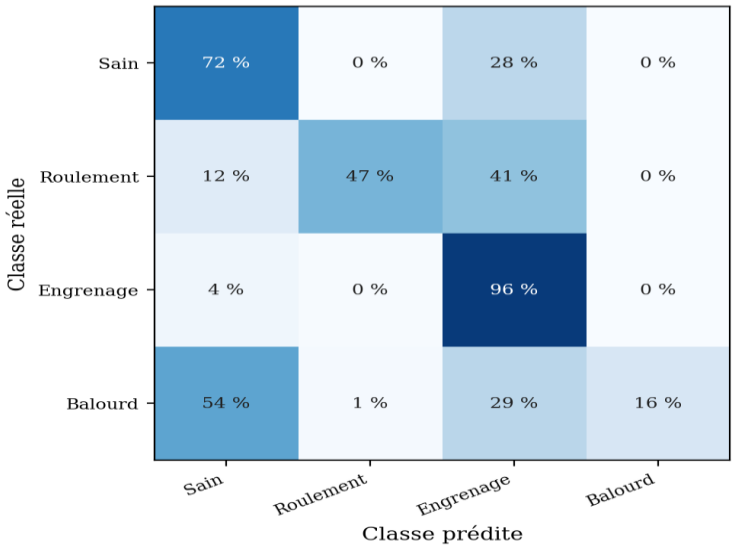

Figure 8. Matrice de confusion sous partition par machine. Source : simulation de l'auteur.

Les taux par classe se situent entre 86 et 89 %. Les roulements sont mieux reconnus grâce à une signature spécifique ; les confusions dominantes opposent l'état sain à l'engrenage ou au balourd. La matrice rend

visibles deux conséquences différentes : une fausse alarme mobilise inutilement, tandis qu'un défaut classé sain peut conduire à une panne. Un taux global masque cette asymétrie.

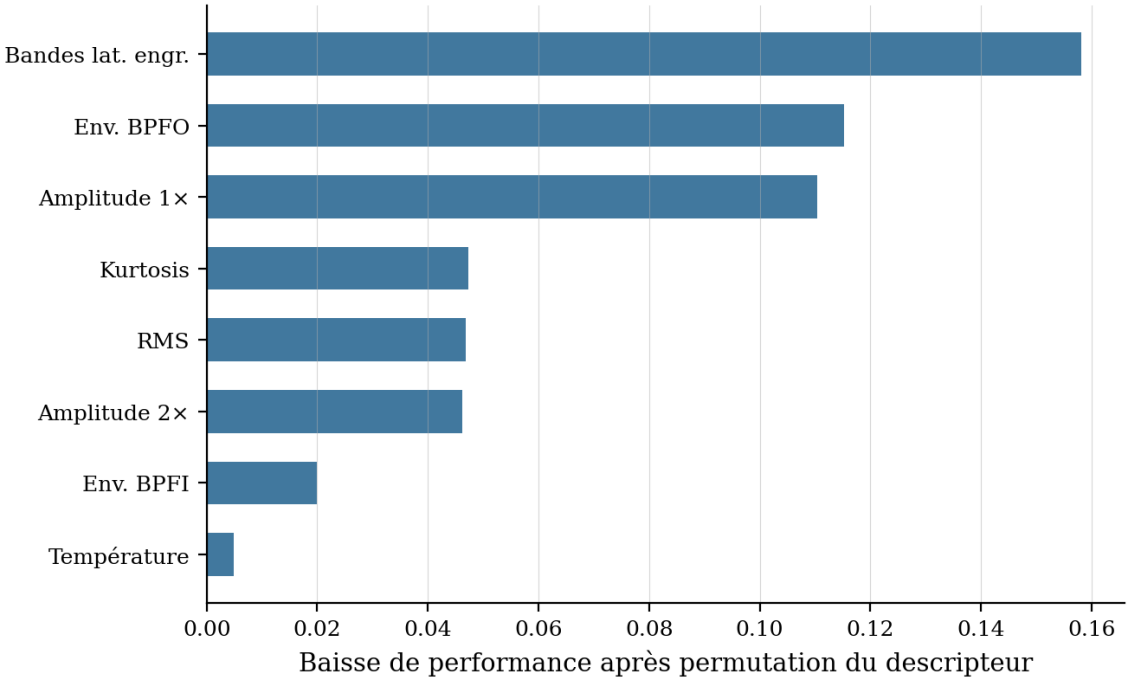

Figure 9. Importance par permutation des huit descripteurs. Source : simulation de l'auteur.

Trois descripteurs concentrent l'information : bandes latérales d'engrènement, amplitude à la rotation et amplitude d'enveloppe de bague externe. Cette cohérence avec la physique valide le protocole. L'importance doit toutefois être lue avec la corrélation : deux variables redondantes peuvent paraître faibles séparément alors que leur information commune est essentielle.

L'explication globale vérifie ce que le modèle a appris ; l'explication locale justifie une alerte particulière ; l'explication contrastive indique la modification minimale qui aurait changé la décision. L'interprétabilité n'est donc pas seulement un outil d'acceptation : elle sert à détecter qu'un modèle s'appuie sur un artefact avant son déploiement.

12. Apport scientifique : le cadre VIGIE

> **Contribution originale : cadre VIGIE:** Validation sans fuite : partitionner selon l'usage réel. Imbalance : évaluer précision, rappel et coût à la prévalence d'exploitation. Généralisation : tester machines, charges et sites non vus. Interprétabilité : vérifier la cohérence physique globale et locale. Exploitation : surveiller le décalage, collecter les retours et versionner.

VIGIE transforme la validation en cycle de vie. Avant déploiement, il exige une partition par groupes et un test scellé. Au choix du seuil, il réintroduit la prévalence et les coûts. À l'épreuve de généralisation, il vérifie les domaines non vus et recherche des descripteurs invariants. Dans l'explication, il confronte l'importance du modèle aux signatures physiques. Après mise en service, il surveille les distributions, documente chaque alerte et marque chaque version.

Le cadre répond à une lacune fréquente : les articles optimisent une métrique sur un jeu fixe, alors que l'industrie exploite un système évolutif. Une performance n'est crédible que si son domaine de validité, son point de fonctionnement, sa calibration et son historique de versions sont connus. Le modèle devient ainsi un composant maintenu, non un livrable figé.

13. Limites et perspectives

Le cas simulé ne fixe pas une performance de référence et ne contient qu'un défaut à la fois. Les effets quantifiés dépendent de l'intensité des biais inter-machines et du décalage de charge. La coexistence de plusieurs défauts appelle une formulation multi-étiquette. Les futures études devraient valider VIGIE sur des parcs multisites, intégrer des intervalles d'incertitude, comparer des stratégies d'apprentissage continu et mesurer le coût total des fausses alertes et non-détections.

14. Conclusion

La valeur d'un diagnostic par apprentissage automatique se juge d'abord à la question posée et au protocole qui produit sa performance. Détection, identification et sévérité n'exigent pas les mêmes données. Sous déséquilibre, l'exactitude et parfois l'AUC peuvent masquer un système inexploitable. Sous partition aléatoire, un modèle flexible peut reconnaître la machine plutôt que le défaut. Sous changement de charge, une performance valable au laboratoire peut disparaître.

Les résultats montrent qu'une validation par machine, une évaluation adaptée à la prévalence, des descripteurs physiquement invariants et un suivi continu sont plus déterminants qu'un gain marginal d'architecture. Le cadre VIGIE proposé organise ces exigences en une méthode cohérente. Il défend une idée simple et exigeante : un modèle industriel n'est pas celui qui obtient le meilleur chiffre sur les données disponibles, mais celui dont on sait précisément pourquoi, où et jusqu'à quand le chiffre reste vrai.

Bibliographie

1. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579
2. Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the CWRU bearing fault dataset. Mechanical Systems and Signal Processing, 169, 108732. https://doi.org/10.1016/j.ymssp.2021.108732
3. Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mechanical Systems and Signal Processing, 64–65, 100–131. https://doi.org/10.1016/j.ymssp.2015.04.021
4. ISO. (2025). ISO 13379-1:2025, Condition monitoring and diagnostics of machine systems: Data interpretation and diagnostics techniques, Part 1: General guidelines. International Organization for Standardization.
5. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of ICML, 233–240. https://doi.org/10.1145/1143844.1143874
6. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of ICML, 1321–1330.
7. Roberts, D. R., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881
8. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138, 106587. https://doi.org/10.1016/j.ymssp.2019.106587

9. Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3, 422–440. https://doi.org/10.1038/s42254-021-00314-5
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer.
11. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
12. Antoni, J., & Randall, R. B. (2004). Unsupervised noise cancellation for vibration signals. Mechanical Systems and Signal Processing, 18(1), 89–101. https://doi.org/10.1016/S0888-3270(03)00012-8
13. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
14. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). Dataset shift in machine learning. MIT Press.
15. Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28.