Caractérisation de la complémentarité des détecteurs d'anomalies par l'analyse des contributions SHAP

Jordan Levy[1,2], Paul Saves[1], Moncef Garouani[1], Nicolas Verstaevel[1], Benoit Gaudou[1]

[1] IRIT, Université Toulouse Capitole
[2] TwinswHeel, Soben

jordan.levy@irit.fr, paul.saves@irit.fr, moncef.garouani@irit.fr, nicolas.verstaevel@irit.fr, benoit.gaudou@irit.fr

Résumé

La détection d'anomalies non supervisée est un problème difficile en raison de la diversité des distributions de données et de l'absence d'étiquettes. Les méthodes ensemblistes sont souvent adoptées pour pallier ces difficultés en combinant plusieurs détecteurs d'anomalies pour réduire les biais individuels et augmenter la robustesse. Cependant, construire un ensemble véritablement complémentaire reste difficile car de nombreux détecteurs reposent sur des critères de discrimination similaires et finissent par produire des scores d'anomalie redondants. Par conséquent, le potentiel de l'apprentissage ensembliste est souvent limité par la difficulté d'identifier des modèles qui capturent vraiment différents types d'irrégularités. Pour remédier à cela, nous proposons une méthodologie pour caractériser les détecteurs d'anomalies à travers leurs mécanismes de décision. En utilisant les explications additives de Shapley (SHAP), nous quantifions comment chaque modèle attribue de l'importance aux caractéristiques d'entrée, et nous utilisons ces profils d'attribution pour mesurer la similarité entre les détecteurs. Nous montrons que les détecteurs ayant des explications similaires ont tendance à produire des scores d'anomalie corrélés et à identifier des anomalies qui se chevauchent largement. Inversement, la divergence des explications indique de manière fiable un comportement de détection complémentaire. Nos résultats démontrent que les métriques basées sur les explications offrent un critère différent, souvent meilleur, des sorties brutes pour sélectionner des modèles dans un ensemble. Cependant, nous démontrons également que la diversité seule est insuffisante ; une performance individuelle élevée des détecteurs d'anomalies reste un prérequis pour des ensembles efficaces. En ciblant explicitement la diversité des explications tout en maintenant la qualité des modèles, nous sommes capables de construire des ensembles plus diversifiés, plus complémentaires et finalement plus efficaces pour la détection d'anomalies non supervisée.

Mots-clés

Détection d'anomalies non supervisée, Modèle ensembliste, Sélection de modèles, Explicabilité.

Abstract

Unsupervised anomaly detection is a challenging problem due to the diversity of data distributions and the lack of labels. Ensemble methods are often adopted to mitigate these challenges by combining multiple anomaly detectors, which can reduce individual biases and increase robustness. Yet building an ensemble that is genuinely complementary remains challenging, since many detectors rely on similar decision cues and end up producing redundant anomaly scores. As a result, the potential of ensemble learning is often limited by the difficulty of identifying models that truly capture different types of irregularities. To address this, we propose a methodology for characterizing anomaly detectors through their decision mechanisms. Using SHapley Additive exPlanations, we quantify how each model attributes importance to input features, and we use these attribution profiles to measure similarity between detectors. We show that detectors with similar explanations tend to produce correlated anomaly scores and identify largely overlapping anomalies. Conversely, explanation divergence reliably indicates complementary detection behavior. Our results demonstrate that explanation-driven metrics offer a different, usually better, criterion than raw outputs for selecting models in an ensemble. However, we also demonstrate that diversity alone is insufficient ; high individual model performance remains a prerequisite for effective ensembles. By explicitly targeting explanation diversity while maintaining model quality, we are able to construct ensembles that are more diverse, more complementary, and ultimately more effective for unsupervised anomaly detection.

Keywords

Unsupervised Anomaly Detection, Ensemble Learning, Model Selection, Explainable AI.

1 Introduction

La détection d'anomalies est un problème difficile, principalement en raison de la nature intrinsèque des anomalies. Les anomalies sont définies comme des déviations par rapport à ce qui est considéré comme un comportement normal [5]. Par conséquent, selon la façon dont cette normalité est définie, un algorithme ou un autre peut être plus adapté pour une

bonne détection. Par exemple, certaines méthodes peuvent définir la normalité en termes géométriques et utiliser des règles basées sur la distance pour repérer les valeurs aberrantes, tandis que d'autres utilisent des hypothèses probabilistes et signalent les instances rares ou à faible probabilité comme des anomalies [19].

Dans de nombreux contextes réels, l'obtention d'anomalies étiquetées est coûteuse ou irréalisable car les événements anormaux sont rares, coûteux à produire ou dangereux à provoquer (*e.g.*, dans l'industrie nucléaire la détection doit fonctionner sans attendre que des défauts se produisent [17]). Par conséquent, les approches semi-supervisées et non supervisées sont souvent préférées. Dans la détection d'anomalies non supervisée (Unsupervised Anomaly Detection, UAD), les spécialistes, ne pouvant pas définir exhaustivement le comportement "normal" à la main, s'appuient généralement sur des méthodes d'apprentissage automatique. Ainsi, au lieu de définir eux-mêmes la normalité, ils utilisent les hypothèses sous-jacentes des algorithmes (géométriques, probabilistes, etc.). Cependant, aucune hypothèse n'est garantie d'être appropriée pour toutes les applications [1], et comme indiqué dans le théorème du "*no free lunch*", aucun détecteur unique ne surpasse systématiquement les autres sur tous les jeux de données [22, 9]. Néanmoins, le choix de l'algorithme reste critique, car une sélection inappropriée peut conduire à de mauvaises performances [1]

Une stratégie courante pour atténuer ce problème consiste à combiner plusieurs détecteurs au sein d'une méthode ensembliste, en tirant parti des atouts de chaque détecteur et de la diversité de leurs hypothèses sur ce qui constitue un comportement normal. L'apprentissage ensembliste a été largement adopté pour cette raison et a démontré de solides performances empiriques dans divers scénarios de détection d'anomalies [1]. En agrégeant des détecteurs avec divers biais inductifs, les approches ensemblistes peuvent capturer une plus grande variété de types d'anomalies. Cette diversité se traduit souvent par une meilleure couverture de détection [21].

Un défi clé dans l'apprentissage ensembliste est la sélection de détecteurs appropriés. Bien qu'une approche ensembliste robuste nécessite théoriquement des détecteurs à la fois diversifiés et performants [21], l'identification de tels comportements complémentaires reste un problème ouvert [13].

Pour relever ce défi, nous proposons une nouvelle méthodologie qui caractérise le comportement des détecteurs d'anomalies en utilisant les explications additives de Shapley (SHAP) [12]. Contrairement aux approches reposant uniquement sur les sorties, nous nous concentrons sur la compréhension des mécanismes de décision internes des détecteurs d'anomalies. Notre étude révèle que les détecteurs ayant des schémas d'explication similaires ont tendance à produire des scores d'anomalie redondants, tandis que la divergence des explications est un indicateur fort de complémentarité. Par conséquent, nous démontrons que la sélection de détecteurs basée sur leur comportement d'explication conduit à une amélioration des performances du modèle ensembliste. Cela conduit à trois contributions majeures :

- Une analyse des algorithmes d'UAD basée sur leurs explications SHAP, qui démontre que leurs comportements sont corrélés à la similarité de leurs sorties.
- Une comparaison entre les explications SHAP et les similarités des sorties de modèles pour sélectionner des modèles de détection d'anomalies diversifiés.
- Une étude empirique quantifiant l'importance relative de la diversité par rapport à la performance individuelle, démontrant que la qualité du modèle reste un prérequis critique.

Cet article est structuré comme suit : la Section 2 présente les travaux connexes. La méthodologie de notre approche est décrite dans la Section 3. Les résultats expérimentaux sont présentés en Section 4. Enfin, nous concluons l'article dans la Section 5.

2 Travaux Connexes

2.1 Algorithmes de détection d'anomalies non supervisée

L'UAD est un problème largement étudié [19] avec un grand nombre d'algorithmes, chacun fondé sur des hypothèses différentes concernant la nature des anomalies. Des bibliothèques comme PyOD [23] ont été introduites pour standardiser l'utilisation de ces modèles. Parmi ceux-ci, les méthodes basées sur la distance (KNN, LOF, CBLOF) caractérisent les anomalies en fonction de leur éloignement des points voisins. De leur côté, les algorithmes basés sur la reconstruction (AutoEncoder, PCA) détectent les anomalies en apprenant à reconstruire des modèles de données normaux. D'autres approches, telles que HBOS, ECOD et COPOD, exploitent des hypothèses probabilistes, tandis que OCSVM et DeepSVDD utilisent la classification à une classe. Enfin, des algorithmes comme Isolation Forest et LODA reposent respectivement sur le partitionnement aléatoire et la projection spatiale.

2.2 Sélection de modèles pour la détection d'anomalies non supervisée

La sélection d'un modèle approprié pour une tâche d'UAD est communément appelée Sélection de Modèle de Valeurs Aberrantes Non Supervisée (Unsupervised Outlier Model Selection, UOMS). L'UOMS est un problème particulièrement difficile qui a gagné en importance à mesure que le nombre de modèles disponibles, comprenant diverses familles algorithmiques et configurations d'hyperparamètres, continue de s'étendre avec l'apparition de nouvelles méthodes dans la littérature [19].

Une stratégie courante en UOMS consiste à estimer la qualité du modèle directement à partir de données non étiquetées de manière non supervisée, pour sélectionner les meilleurs modèles. Ces approches se divisent en deux catégories principales. Les méthodes *autonomes* (stand-alone) [8, 16] calculent un score non supervisé pour chaque détecteur indépendamment, tandis que les méthodes *basées sur le consensus* (consensus-based) évaluent les détecteurs en mesurant l'accord au sein d'un groupe de modèles et en sélectionnant ceux qui se conforment le mieux au groupe [6, 11]. Les preuves empiriques indiquent que les critères d'évaluation

autonomes échouent souvent à fournir des résultats cohérents ou fiables en pratique. En revanche, les approches basées sur le consensus, malgré leur coût de calcul plus élevé, sont apparues comme une stratégie plus efficace et prometteuse pour l'UOMS [13].

Il existe peu de travaux dans la littérature scientifique sur l'UOMS pour l'apprentissage d'ensemble. Dans [21], les auteurs montrent que la diversité des hypothèses algorithmiques tend à donner de meilleurs résultats. Pour caractériser la diversité, les auteurs utilisent une corrélation de Pearson pondérée entre les scores d'anomalie des modèles. Cependant, ils soulignent également que la diversité est importante, mais que les algorithmes choisis doivent déjà avoir de bonnes performances sur le jeu de données pour obtenir un modèle d'ensemble performant. À partir de l'hypothèse précédente, les auteurs de [18] ont introduit SELECT, qui, au lieu d'essayer de sélectionner des modèles en fonction de leur diversité, sélectionne des modèles en fonction de leurs performances à partir d'une pseudo-étiquette de vérité terrain créée à partir des scores d'anomalie. Plus récemment, dans [3], les auteurs font de la détection d'anomalies dans les séries temporelles en utilisant un ensemble de plusieurs auto-encodeurs convolutionnels. Ils optimisent la diversité en intégrant une métrique basée sur la dissimilarité des sorties de reconstruction. Ils démontrent également que cette stratégie axée sur la diversité améliore les performances de détection.

Bien qu'il ait été démontré que les sorties des modèles comme les scores d'anomalies peuvent refléter la diversité, dans ce travail, nous étudions une méthode de consensus basée sur les valeurs SHAP, qui caractérisent le comportement des modèles de détection d'anomalies sur différents jeux de données.

2.3 Explicabilité

L'Intelligence Artificielle Explicable (Explainable Artificial Intelligence, XAI) cherche à rendre les modèles complexes transparents en produisant des représentations interprétables par l'homme sur la manière dont les entrées, la structure du modèle et l'incertitude produisent des sorties particulières [14]. Parmi les méthodes d'explication locales et agnostiques au modèle, les explications additives de Shapley (SHAP) sont largement utilisées car elles fournissent des attributions de caractéristiques axiomatiques au niveau de l'instance [12].

Au-delà de l'interprétabilité, les représentations SHAP ont été exploitées pour comparer des modèles et former des représentations informatives : des travaux récents utilisent des attributions SHAP agrégées pour identifier des familles de modèles ou pour servir de caractéristiques pour des tâches d'apprentissage en aval [20, 7]. À notre connaissance, l'exploitation de SHAP pour la comparaison et la sélection de modèles n'a pas été systématiquement étudiée dans le cadre de l'UAD, où l'absence d'étiquettes rend la sélection des modèles particulièrement difficile.

3 Méthodologie

3.1 Cadre du problème

Nous considérons le problème d'UAD, où l'objectif est d'identifier des échantillons anormaux au sein d'un jeu de données $\mathcal{D} = \{x_1, \ldots, x_n\}$ contenant n instances dans un espace de caractéristiques à d dimensions. Soit $\mathcal{M}$ un ensemble de m modèles de détection d'anomalies $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$. Pour un jeu de données donné, chaque modèle M_i produit un vecteur de score d'anomalie $s_i = M_i(\mathcal{D}) \in \mathbb{R}^n$ où $s_i^{(k)}$ est le score d'anomalie du $k^{\text{ième}}$ point d'entrée selon le $i^{\text{ième}}$ modèle. Chaque modèle produit également un vecteur de prédictions binaires $a_i \in \{0,1\}^n$, obtenu par un seuillage du vecteur de scores d'anomalie, indiquant directement la présence ou l'absence d'anomalies. Plus précisément pour chaque instance k, la prédiction est $a_i^{(k)} = 0$ si $s_i^{(k)} < \tau_i$, et 1 sinon. Le paramètre τ_i est un seuil de décision interne, propre à chaque modèle i. Notre objectif est d'analyser le comportement de ces modèles et d'identifier des groupes de modèles qui partagent des structures interprétatives similaires.

3.2 Similarité des modèles à partir des explications

Pour chaque modèle M_i, nous calculons sa matrice d'explication SHAP $Sh_i \in \mathbb{R}^{n \times d}$, où $Sh_i^{(k)} \in \mathbb{R}^d$ est le vecteur représentant la contribution de chaque caractéristique au score d'anomalie de l'instance x_k.
Nous définissons la similarité de comportement entre deux modèles M_i et M_j comme la corrélation de Pearson moyenne par instance entre leurs vecteurs SHAP :

$$\rho_{ij}^{\text{PS}} = \frac{1}{n} \sum_{k=1}^{n} \text{corr}\big(Sh_i^{(k)}, Sh_j^{(k)}\big).$$

Pour capturer la cohérence du classement entre les importances des caractéristiques plutôt que les magnitudes brutes, nous calculons également une similarité basée sur le Gain Cumulé Actualisé Normalisé (Normalized Discounted Cumulative Gain, NDCG) :

$$\rho_{ij}^{\text{NDCG}} = \frac{1}{2n} \sum_{k=1}^{n} \Big(\text{NDCG}\left(|Sh_i^{(k)}|, |Sh_j^{(k)}|\right) + \text{NDCG}\left(|Sh_j^{(k)}|, |Sh_i^{(k)}|\right) \Big),$$

où le NDCG évalue l'accord dans l'importance classée des caractéristiques entre deux détecteurs, et $|\cdot|$ correspond à la valeur absolue. Si les deux modèles attribuent une importance SHAP élevée aux mêmes caractéristiques, leur valeur NDCG sera proche de 1, indiquant un comportement explicatif similaire [2].

3.3 Lier les similarités des explications aux sorties des détecteurs

Pour comparer la similarité des explications des modèles avec la similarité des sorties, nous introduisons deux matrices supplémentaires : la matrice des corrélations de scores et la matrice des similarités de Jaccard.

Nous définissons la similarité entre deux modèles M_i et M_j sur la base de la corrélation de leurs scores d'anomalie. Plus précisément, nous calculons la corrélation de Pearson moyenne par instance entre leurs vecteurs de scores comme suit :

$$\rho_{ij}^{\text{Score}} = \frac{1}{n}\sum_{k=1}^{n} \text{corr}\left(s_i^{(k)}, s_j^{(k)}\right).$$

Il en résulte une matrice symétrique qui reflète la similarité par paires entre les modèles vis-à-vis des scores attribués.
Pour comparer directement les prédictions de deux modèles i et j, la similarité de Jaccard entre les deux vecteurs de prédictions a_i et a_j est calculée comme suit :

$$J_{ij} = \frac{|a_i \cap a_j|}{|a_i \cup a_j|},$$

avec $|\cdot|$ le cardinal de l'ensemble. Cette métrique mesure le chevauchement entre les modèles : une valeur de 1 implique des prédictions identiques, tandis que 0 implique des ensembles disjoints d'anomalies détectées.
Chaque matrice de similarité $P \in \{\rho^{PS}, \rho^{NDCG}, \rho^{Scores}, J\}$ peut également être transformée en une matrice de dissimilarité $D \in \{\delta^{PS}, \delta^{NDCG}, \delta^{Scores}, \delta^{J}\}$. Ces matrices sont calculées selon la relation $D = 1 - P$, où chaque élément représente la distance entre deux détecteurs.
Afin de quantifier la relation entre les différentes mesures, nous utilisons le test de Mantel [15] pour déterminer si deux matrices de dissimilarité données sont statistiquement corrélées. Le coefficient de Mantel (r_M) est calculé comme la corrélation de Pearson entre les éléments triangulaires supérieurs des matrices, la signification statistique étant établie par des tests de permutation.

4 Expérimentations

Nous avons mené des expériences pour répondre aux questions suivantes : (1) La similarité des explications implique-t-elle une similarité des prédictions ? (2) Les métriques basées sur SHAP surpassent-elles les sorties brutes pour quantifier la diversité ? et (3) Dans quelle mesure cette diversité impacte-t-elle la précision et la robustesse de l'ensemble résultant ?

4.1 Configuration expérimentale

Dans les expériences, 14 algorithmes UAD ont été utilisés : COF, KNN, LOF, IForest, PCA, CBLOF, LODA, HBOS, MCD, OCSVM, DAGMM, DeepSVDD, COPOD et ECOD tels qu'implémentés dans la bibliothèque PyOD [23]. Les hyperparamètres définis par les auteurs de chaque modèle ont été conservés et aucune indication du pourcentage d'anomalie n'a été fournie pour aucun jeu de données. Au niveau de l'explicabilité, nous utilisons l'implémentation de la méthode agnostique Kernel SHAP disponible dans la bibliothèque Python `shap`. Le choix de cette méthode était nécessaire pour assurer un cadre unifié d'explication à travers nos divers algorithmes. De plus, pour assurer une approximation stable, le jeu de données d'arrière-plan a été résumé en utilisant l'algorithme de partitionnement k-means avec $k = 50$ centroïdes.

Pour augmenter la robustesse des résultats, chaque jeu de données a été divisé à cinq reprises en un ensemble d'entraînement et un ensemble de test, l'ensemble d'entraînement représentant 80 % des données. La graine aléatoire pour chaque division a été fixée égale à l'indice d'itération pour assurer la reproductibilité. Le code source utilisé dans les expériences est disponible sur GitHub [1].

4.2 Jeux de données considérés

En raison du coût de calcul élevé associé à SHAP, la sélection des jeux de données a été restreinte aux 50 % plus petits jeux de données disponibles dans ADBench [9]. De plus, les jeux de données contenant plus de 20 caractéristiques ont été supprimés pour maintenir la faisabilité des calculs. Ce processus de filtrage initial a abouti à un ensemble de 16 jeux de données provenant de diverses applications : annthyroid (AN), breastw (BR), glass (GL), Hepatitis (HE), Lymphography (LY), mammography (MA), PageBlocks (PB), Pima (PI), Stamps (ST), thyroid (TH), vertebral (VE), vowels (VO), WBC (WB), Wilt (WL), wine (WN) et yeast (YE).

4.3 Corrélation entre les matrices de similarité

Nous avons calculé les quatre matrices de similarité ρ^{PS} (corrélation linéaire entre les SHAP), ρ^{NDCG} (similarité des classements d'importance des caractéristiques SHAP), ρ^{Scores} (corrélation linéaire entre les scores d'anomalie), J (indice de Jaccard entre les prédictions d'anomalie) pour chaque jeu de données. Les matrices moyennes sur tous les jeux de données sont présentées dans la Figure 1, où un partitionnement hiérarchique a été appliqué pour optimiser l'ordre. Visuellement, deux groupes de modèles émergent. Premièrement, COPOD et ECOD se regroupent systématiquement, ce qui est attendu car les deux algorithmes reposent sur des hypothèses de distribution de données similaires. Deuxièmement, un groupe plus large comprenant OCSVM, AutoEncoder, IForest, PCA, KNN, CBLOF et GMM présente une forte corrélation. Ces algorithmes partagent des caractéristiques sous-jacentes liées aux métriques de distance et aux stratégies d'encodage des données.
Le test de Mantel est utilisé pour évaluer si les matrices de similarité sont statistiquement corrélées. Le Tableau 1 rapporte les corrélations de Mantel moyennes sur les jeux de données. Nous observons une forte corrélation ($r_M = 0.83$) entre δ^{PS} et δ^{NDCG}, confirmant que la similarité basée sur SHAP est cohérente en termes de magnitude et de classement de l'importance des caractéristiques. De même, δ^{Scores} et δ^{J} montrent une forte corrélation ($r_M = 0.76$), ce qui est intuitif puisque les modèles avec des distributions de scores similaires ont tendance à produire des prédictions binaires similaires. Enfin, la corrélation entre δ^{PS} et δ^{J} ($r_M = 0.67$) suggère que les détecteurs partageant des schémas de raisonnement similaires (tels que capturés par SHAP) ont également tendance à donner des prédictions d'anomalie similaires. Cependant, la similarité entre les modèles dépend du jeu de données. Plus précisément, le Tableau 2 montre les corrélations

1. https://github.com/jordanlv/Analyzing-SHAP-UOMS

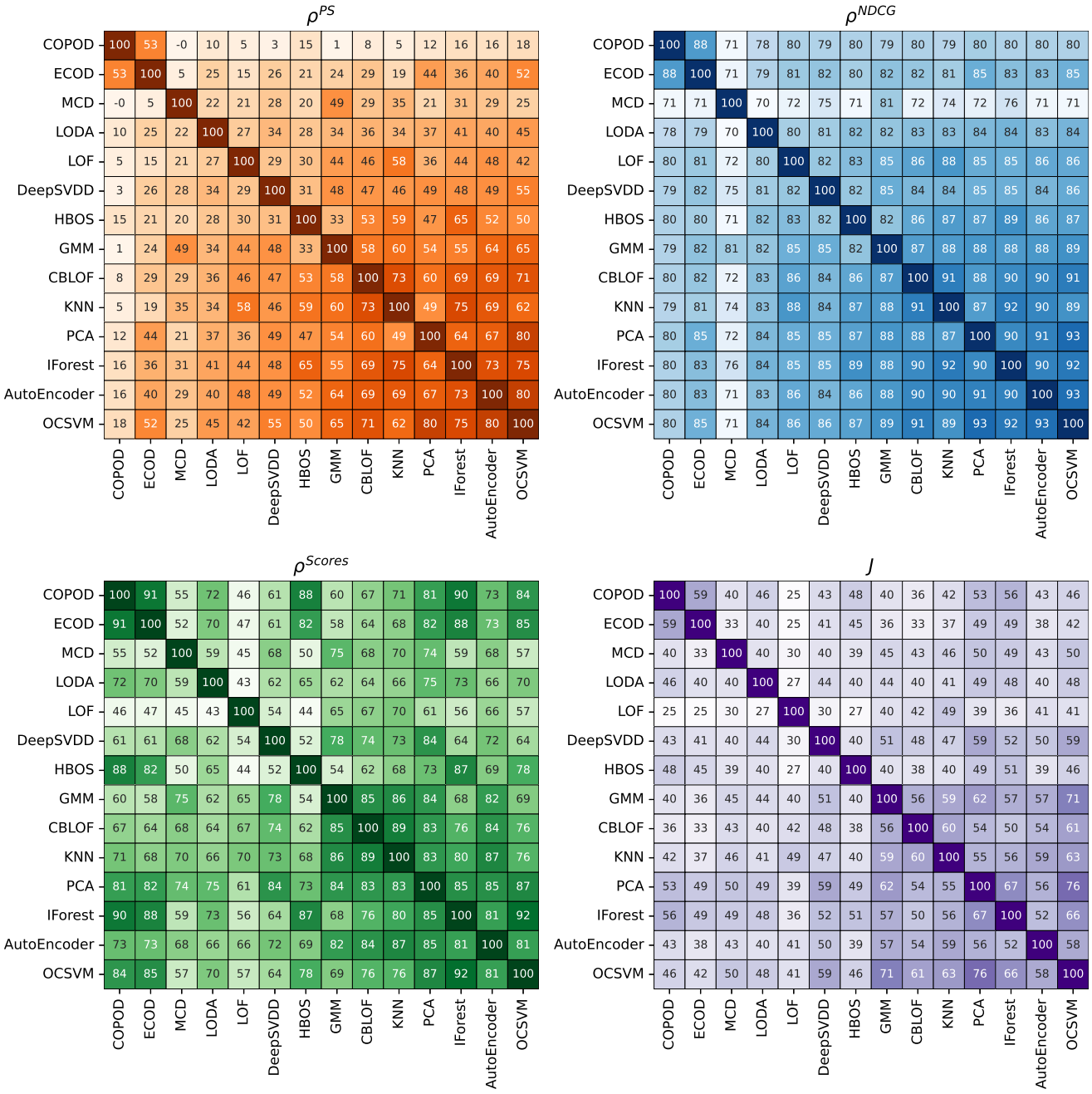


FIGURE 1 – Similarité moyenne entre les modèles sur tous les jeux de données. Disposition : Haut-Gauche : corrélations des valeurs SHAP ; Haut-Droite : NDCG des SHAP ; Bas-Gauche : corrélations des scores d'anomalie ; et Bas-Droite : similarités de Jaccard.

entre δ^{PS} et δ^{Scores}, et entre δ^{PS} et δ^{J} pour chaque jeu de données. Habituellement, les deux paires sont corrélées. Certains jeux de données montrent de fortes corrélations entre les matrices ($r_M > 0.5$) comme AN, PI, TH ou VO, tandis que d'autres ont une corrélation modérée ou faible entre les deux matrices. Enfin, certains jeux de données intéressants sont BR, LY ou YE où $r_M(\delta^{PS}, \delta^{Scores}) \ll r_M(\delta^{PS}, \delta^{J})$.

Dans l'ensemble, la similarité des modèles est corrélée avec leurs prédictions. Dans les sections suivantes, nous étudions la combinaison de modèles éloignés dans les matrices de similarité pour améliorer les résultats globaux d'un ensemble.

4.4 Agrégation des prédictions des modèles

L'objectif d'une méthode ensembliste est de combiner les prédictions de plusieurs détecteurs pour créer une méthode plus robuste. Un défi dans ce type de méthode est de savoir comment agréger chaque prédiction. Chaque détecteur d'anomalies produit un score d'anomalie et une prédiction indiquant si un point de données est normal ou anormal. Pour tirer véritablement parti de la spécialité de chaque modèle, nous utilisons les scores d'anomalie comme entrée

TABLE 1 – Corrélations de Mantel moyennes entre les matrices de distance, moyennées sur les jeux de données. Toutes les corrélations sont significatives ($p \leq 0.004$).

	δ^{PS}	δ^{NDCG}	δ^{Scores}	δ^{J}
δ^{PS}	1.00	0.83	0.55	0.67
δ^{NDCG}		1.00	0.54	0.57
δ^{Scores}			1.00	0.76
δ^{J}				1.00

TABLE 2 – Corrélations de Mantel ($\times 10^2$) entre les matrices de distance pour chaque jeu de données.

Jeu de données	$r_M(\delta^{PS}, \delta^{Scores})$	$r_M(\delta^{PS}, \delta^{J})$
AN	75	78
BR	22	51
GL	42	24
HE	26	51
LY	36	84
MA	56	51
PB	33	66
PI	55	52
ST	56	53
TH	72	76
VE	26	-4
VO	56	58
WB	18	15
WL	49	49
WN	26	30
YE	38	62

pour notre fonction d'agrégation. Les méthodes d'agrégation connues pour ces scores incluent l'utilisation du maximum, de la moyenne arithmétique, ou la moyenne des rangs (une méthode consistant à substituer les scores bruts par leur classement d'anormalité avant de les moyenner) [1]. Nous avons évalué ces fonctions sur tous les ensembles possibles de 3 modèles distincts parmi notre groupe de 14 modèles (résultant en 364 ensembles) sur chaque jeu de données. Compte tenu du déséquilibre de classe de chaque jeu de données, nous utilisons l'Aire Sous la Courbe Précision-Rappel (AUCPR) pour évaluer les performances. Comme le montre le Tableau 3, l'agrégation par rang donne des résultats supérieurs sur 11 des 16 jeux de données. Par conséquent, nous adoptons l'agrégation par rang pour le reste de cette étude.

4.5 Complémentarité

Pour sélectionner les modèles les plus dissimilaires, nous utilisons les matrices de dissimilarité D, ce qui donne des matrices sur la distance entre deux détecteurs. De nouveau, à partir du groupe de 14 modèles, nous avons construit des ensembles de taille $n = 3$. Le Tableau 4 présente les corrélations entre l'AUCPR des ensembles et les distances de diversité pour chaque matrice de dissimilarité. Une corrélation positive entre la diversité et la performance implique que l'augmentation de la diversité de l'ensemble conduit à des résultats de détection supérieurs.

TABLE 3 – AUCPR moyen ($\times 10^2$) entre tous les 364 ensembles avec les 3 principales stratégies d'agrégation. Les meilleurs résultats sont en gras.

Jeu de données	Rang	Max	Moyenne
AN	**26**	21	9
BR	**98**	59	49
GL	18	**22**	19
HE	**84**	64	47
LY	**100**	68	36
MA	**30**	8	4
PB	**60**	35	12
PI	**55**	45	42
ST	**62**	35	22
TH	**58**	28	5
VE	18	**28**	**28**
VO	**41**	11	8
WB	**97**	41	25
WL	6	**9**	7
WN	49	**53**	31
YE	37	**40**	39
moyenne	**52**	35	24

TABLE 4 – Corrélation ($\times 10^2$) entre la distance des modèles et l'AUCPR des ensembles. Les plus grands résultats sont en gras.

Jeu de données	δ^{PS}	δ^{NDCG}	δ^{Scores}	δ^{J}
AN	-20	**39**	-1	-32
BR	**-0**	-17	-90	-59
GL	**5**	-2	-33	-37
HE	**-28**	-32	-48	-31
LY	15	-0	-29	**16**
MA	**-9**	-68	-18	3
PB	-18	**21**	-17	-4
PI	**37**	14	30	24
ST	**40**	15	15	36
TH	-38	**19**	-12	-48
VE	-44	-30	-12	**-7**
VO	-66	-54	**-0**	-3
WB	**51**	36	-11	-28
WL	55	**62**	46	55
WN	43	**48**	7	1
YE	29	37	34	**45**
moyenne	3	**5**	-9	-4

Le tableau présente trois informations clés. Premièrement, la sélection de la diversité à partir des valeurs SHAP tend

à donner de meilleures performances que les sorties des modèles (scores et prédictions). Sur 11 des 16 jeux de données, l'utilisation de δ^{PS} et δ^{NDCG} comme diversité tend à donner de meilleurs résultats que l'utilisation de δ^{Scores} et δ^{J}. Deuxièmement, comme indiqué précédemment, la diversité contribue à l'apprentissage d'ensemble en élargissant la gamme des anomalies détectées. Cependant, une sélection de modèles efficace doit également tenir compte de la performance individuelle des modèles, un facteur non explicitement optimisé dans cette étude. Cette limitation explique probablement pourquoi les corrélations entre les métriques de distance et la performance restent modérées. De plus, dans certains jeux de données, ces corrélations sont systématiquement négatives, suggérant que la diversité n'est pas toujours bénéfique. Ce phénomène se produit notamment lorsqu'un seul modèle est plus performant que les autres. Dans de tels cas, la condition principale pour un ensemble réussi est l'inclusion de ce modèle spécifique. De plus, des corrélations négatives peuvent survenir dans des jeux de données très complexes où tous les modèles présentent de mauvaises performances. Dans ces scénarios, même une grande diversité ne peut pas compenser le manque de détection significative, entraînant un ensemble inefficace. Enfin, il est intéressant de noter que les SHAP et les scores ne montrent pas la même diversité car les corrélations varient entre les jeux de données. Par exemple, sur le jeu de données WB, les corrélations entre δ^{Scores} et δ^{J} sont négatives, tandis que δ^{PS} et δ^{NDCG} sont positives. Par conséquent, les deux métriques mettent en évidence des diversités différentes.

4.6 Diversité et performances individuelles

Bien que nous ayons montré que la diversité améliore la performance de l'ensemble en élargissant la couverture des anomalies, notre analyse précédente a négligé la précision individuelle des modèles. Ici, nous affinons nos résultats pour démontrer que malgré la valeur de la diversité, la qualité individuelle de chaque modèle reste un facteur critique. Pour les besoins de cette section, nous utilisons δ^{PS} pour quantifier la diversité.

La Figure 2 illustre l'importance de la performance individuelle de chaque modèle pour obtenir un ensemble efficace. Plus précisément, la Figure 2a présente les résultats pour le jeu de données LY. Une corrélation claire entre la diversité, la performance individuelle et la précision de l'ensemble est observable sur ce jeu de données. Cependant, cette corrélation n'est pas toujours bénéfique. Dans certains scénarios, imposer la diversité peut être préjudiciable à l'ensemble. Par exemple, la Figure 2b révèle que pour le jeu de données VO, la diversité offre un gain négligeable et peut même conduire à des résultats sous-optimaux.

Pour évaluer l'impact relatif de la qualité du modèle par rapport à la diversité, nous avons effectué une régression linéaire pour prédire le gain de performance de l'ensemble en utilisant la performance individuelle moyenne et les scores de diversité. Les poids résultants, présentés dans le Tableau 5, indiquent que bien que la performance individuelle soit généralement le facteur dominant, la diversité joue un rôle complémentaire crucial. Dans 12 des 16 jeux de données, le coefficient de diversité est positif, confirmant sa valeur en tant qu'amplificateur de performance. Notamment, pour les jeux de données WB et LY, le poids de la diversité rivalise ou dépasse même celui de la performance individuelle (ratios de 1,2 et 0,8 respectivement). Dans l'ensemble, avec un ratio moyen de 0,2, la diversité joue un rôle secondaire mais précieux dans la conception d'ensembles UAD.

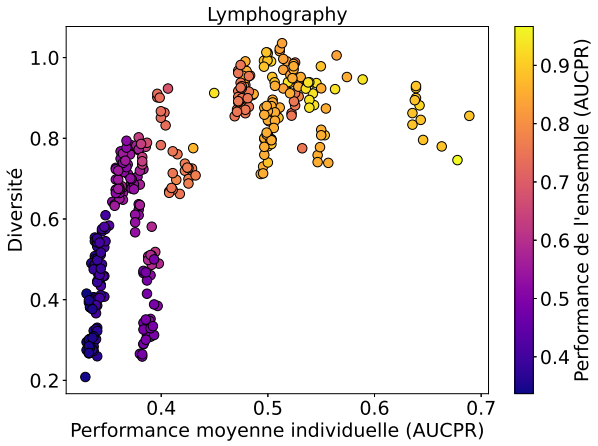
TABLE 5 – Poids de régression linéaire ($\times 10^2$) prédisant la performance de l'ensemble à partir de la performance individuelle moyenne et de la diversité, ainsi que leur ratio (vert : positif ; rouge : négatif).

Jeu de données	Perf. Indiv.	Diversité	Ratio
AN	2.7	0.5	0.2
BR	6.3	2.0	0.3
GL	1.1	0.1	0.1
HE	6.5	0.5	0.1
LY	10.7	8.4	0.8
MA	3.4	0.7	0.2
PB	5.0	2.2	0.4
PI	1.7	0.2	0.1
ST	5.6	0.3	0.0
TH	5.7	1.7	0.3
VE	0.1	-0.0	-0.1
VO	7.5	-0.6	-0.1
WB	2.4	2.8	1.2
WL	0.0	0.0	0.4
WN	8.0	-0.1	-0.0
YE	0.3	-0.0	-0.0
moyenne	4.2	1.2	0.2

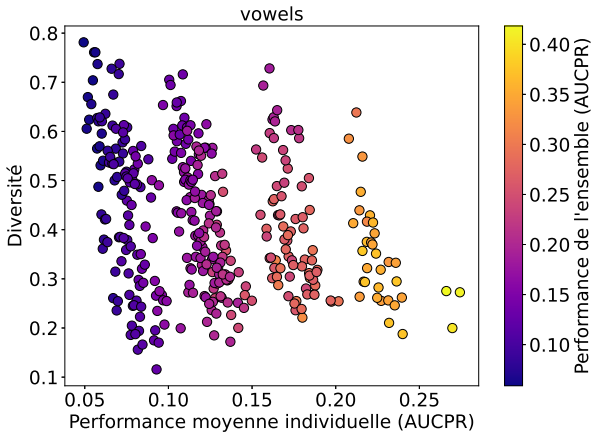
5 Conclusion

Nous avons présenté une méthodologie pour sélectionner des modèles dans des ensembles UAD, basée sur la similarité de leurs explications. Nous avons démontré que la diversité s'avère bénéfique et permet d'améliorer les résultats. Nous avons analysé quatre métriques de diversité : deux fondées sur les explications SHAP et deux directement sur les sorties des modèles. Ces métriques ont conduit à des résultats différents lors de la sélection, indiquant qu'elles capturent des formes de diversité distinctes. De manière générale, les métriques basées sur SHAP ont affiché des résultats supérieurs à ceux basés sur les sorties. Enfin, nous avons établi que malgré les avantages de la diversité, la performance individuelle des modèles reste le facteur décisif : des modèles faibles mais diversifiés ne peuvent pas surpasser des modèles forts mais similaires. Cette recherche tend à montrer que l'explicabilité devrait être davantage prise en compte dans l'UOMS, car elle fournit de nouvelles informations sur le comportement des modèles.

Une limite de notre approche réside dans le coût de calcul de



(a) Jeu de données Lymphography (LY).



(b) Jeu de données Vowels (VO).

FIGURE 2 – Relation entre la diversité de l'ensemble (donnée par δ^{PS}) et la performance individuelle moyenne. Chaque point représente un ensemble de modèles. L'échelle de couleurs indique la performance globale de l'ensemble (AUCPR).

SHAP, qui peut devenir prohibitif pour les jeux de données comportant un grand nombre d'instances ou de caractéristiques. Cependant, le coût de calcul peut être atténué en utilisant des techniques d'approximation ou des modèles de substitution. De plus, notre stratégie est agnostique quant à la méthode d'explication, permettant l'utilisation de techniques d'interprétabilité moins coûteuses si nécessaire. Par exemple, dans [4], les auteurs ont récemment démontré comment l'agrégation de profils de dépendance partielle (PDP) sur un ensemble de modèles quasi-optimaux peut fournir des métriques fiables pour l'incertitude et la robustesse des explications.

Bien que nous ayons modélisé avec succès la diversité, l'optimisation des modèles individuels constitue une perspective d'évolution naturelle de ces travaux. L'ajustement fin des hyperparamètres permettrait de maximiser le potentiel de chaque détecteur, offrant ainsi un levier supplémentaire pour améliorer la performance globale de l'ensemble.

Concernant les travaux futurs, nous envisageons d'étudier la divergence entre les similarités basées sur SHAP et celles issues des sorties brutes afin d'affiner le processus de sélection de modèles. Par ailleurs, la performance individuelle étant critique, l'intégration de méthodes d'estimation de la qualité des modèles au sein de notre méthodologie constitue une priorité. Enfin, nous visons à étendre cette stratégie à l'UAD pour les séries temporelles, un domaine où la complexité accrue rend les approches ensemblistes particulièrement pertinentes [10].

Remerciements

La recherche présentée dans cet article a bénéficié d'un financement de l'Association Nationale de la Recherche et de la Technologie sous le numéro de subvention CIFRE 2023/1398 et de la société Soben. Les auteurs remercient également l'Agence Nationale de la Recherche pour le financement du projet MIMICO sous le numéro de subvention ANR-24-CE23-0380.

Références

[1] Charu C Aggarwal. *Outlier Analysis*. Springer, 2016.

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[3] David Campos, Tung Kieu, Chenjuan Guo, Feiteng Huang, Kai Zheng, Bin Yang, and Christian S Jensen. Unsupervised time series outlier detection with diversity-driven convolutional ensembles. *Proceedings of the VLDB Endowment*, 15(3) :611–623, 2021.

[4] Mustafa Cavus, Jan N van Rijn, and Przemysław Biecek. Beyond the single-best model : Rashomon partial dependence profile for trustworthy explanations in automl. In *International Conference on Discovery Science*, pages 445–459. Springer, 2025.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3) :1–58, 2009.

[6] Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv :1905.12614*, 2019.

[7] Moncef Garouani, Ayah Barhrhouj, and Olivier Teste. Xstacking : An effective and inherently explainable framework for stacked ensemble learning. *Information Fusion*, 2025.

[8] Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms ? *arXiv preprint arXiv :1607.01152*, 2016.

[9] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench : anomaly detection benchmark. *NeurIPS*, 35 :32142–32159, 2022.

[10] Jordan Levy, Clément Blanco-Volle, Nicolas Verstaevel, Benoit Gaudou, and Vincent Talon. Timeciel : Contextual interactive ensemble learning for time series classification. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 316–327. Springer, 2025.

[11] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. InfoGAN-CR and ModelCentrality : Self-supervised Model Training and Selection for Disentangling GANs. In *International conference on machine learning*, pages 6127–6139. PMLR, 2020.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NeurIPS*, 30, 2017.

[13] Martin Q Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. The need for unsupervised outlier model selection : A review and evaluation of internal evaluation strategies. *ACM SIGKDD Explorations Newsletter*, 25(1) :19–35, 2023.

[14] Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. Interpretability needs a new paradigm. *arXiv*, 2024.

[15] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27 :209–220, 1967.

[16] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *TKDD*, 14(4) :1–42, 2020.

[17] Sang Won Oh, Hye Seon Jo, Ho Jun Lee, Man Gyun Na, SW Oh, HS Jo, HJ Lee, and MG Na. Anomalies detection by unsupervised learning using explainable artificial intelligence in nuclear power plants. In *Transactions of the Korean Nuclear Society Spring Meeting Jeju, Korea*, 2022.

[18] Shebuti Rayana and Leman Akoglu. Less is more : Building selective anomaly ensembles. *TKDD*, 10(4) :1–33, 2016.

[19] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5) :756–795, 2021.

[20] Paul Saves, Pramudita Satria Palar, Muhammad Daffa Robani, Nicolas Verstaevel, Moncef Garouani, Julien Aligon, Benoit Gaudou, Koji Shimoyama, and Joseph Morlier. Surrogate modeling and explainable artificial intelligence for complex systems : A workflow for automated simulation exploration. *arXiv preprint*, 2025.

[21] Erich Schubert, Remigius Wojdanowski, Arthur Zimek, and Hans-Peter Kriegel. On evaluation of outlier rankings and outlier scores. In *International conference on data mining*, pages 1047–1058. SIAM, 2012.

[22] David H Wolpert and William G Macready. No free lunch theorems for optimization. *Transactions on evolutionary computation*, 2002.

[23] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod : A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96) :1–7, 2019.