La détection d'exemples mal-étiquetés vue comme l'introspection de modèles d'apprentissage: concepts, recensement et étude comparative

Thomas George, Pierre Nodet, Alexis Bondu, Vincent Lemaire

Orange Innovation
Châtillon, France

Résumé. Les données mal-étiquetés sont omniprésentes dans les applications industrielles de l'apprentissage, ce qui appelle au développement de méthodes de détection automatiques. Nous montrons que la plupart des méthodes existantes peuvent être comprises comme l'introspection de modèles d'apprentissage entraînés, dans un cadre paramétré par seulement 4 dimensions. Une bibliothèque Python démontre l'effectivité de ce cadre, qui permet d'implémenter des méthodes existantes, et d'imaginer simplement de nouvelles méthodes. Ceci permet par exemple d'adapter des méthodes initialement développées en apprentissage profond, aux méthodes classiques souvent plus efficaces sur données tabulaires. Nous évaluons les méthodes existantes sur du bruit d'étiquetage Complétement Au Hasard (NCAR, artificiel), ainsi que sur du bruit d'étiquetage Pas Au Hasard (NNAR, plus réaliste) provenant d'une variété de tâches avec des règles d'étiquetage imparfaites. Cette étude comparative fournit de nouvelles perspectives et permet de mettre en lumière les limites des méthodes existantes.

Ce document est un résumé en français de l'article complet : *Mislabeled examples detection viewed as probing machine learning models : concepts, survey and extensive benchmark* (George et al., 2024).

1 Méthodes d'introspection

Les détecteurs basés sur l'introspection examinent s'il existe une différence de traitement entre les exemples bien et mal-étiquetés lors de l'apprentissage d'un *modèle de base*. Cette différence de traitement est mesurée à l'aide de *sondes*, qui peuvent être calculées à plusieurs étapes de l'entraînement (*ensembles* progressifs), ou alors sur plusieurs modèles entraînés sur des sous-ensembles des données d'entraînement (*ensembles* indépendants) Ces différentes valeurs de sondes sont enfin *aggrégées* afin d'obtenir un *score de confiance* pour chacun des exemples.

2 Bibliothèque Python

Une bibliothèque permet d'instancier différentes méthodes existantes en spécifiant les 4 paramètres :

1. Modèle de base : n'importe quel modèle d'apprentissage qui suit l'API de scikit-learn
2. Sonde : parmi des sondes prédéfinies ou alors en implémentant une nouvelle sonde
3. Stratégie d'ensemble : en fonction du modèle entraîné, indépendant ou progressif
4. Méthode d'aggrégation : en fonction de la méthode à implémenter, soit on effectue une moyenne, ou alors plutôt une mesure de différence de sonde entre les différents membres de l'ensemble

Les méthodes de la littérature qui entrent dans ce cadre sont prêtes à l'emploi à l'aide de la bibliothèque, parmi les détecteurs pré-définis. La bibliothèque permet également de proposer et implémenter de nouvelles variations de méthodes en variant un seul des paramètres, ou alors des méthodes complètement nouvelles qui explorent des combinaisons de paramètres.

3 Étude comparative

Nous réalisons des expériences sur des ensembles de données tabulaires et textuelles, labelisées à l'aide de méthodes automatiques imparfaites. Ces méthodes produisent des données bruitées "Pas Au Hasard" où les exemples mal-étiquetés seront plus fréquents dans certaines régions de l'espace des covariables, moins bien couvertes par les règles automatiques. Ce type de bruit, plus difficile, nous permet de comparer les méthodes de l'état de l'art dans une implémentation commune, sur un type de bruit plus *difficile* que du bruit "Complètement Au Hasard". Nous proposons un compte-rendu critique de ces méthodes dans l'article complet.

Références

George, T., P. Nodet, A. Bondu, et V. Lemaire (2024). Mislabeled examples detection viewed as probing machine learning models : concepts, survey and extensive benchmark. *Transactions on Machine Learning Research*.

Summary

Mislabeled examples are ubiquitous in industrial applications of machine learning, advocating for the development of automatic detection methods. We show that most existing such methods can be viewed as probing trained machine learning models, in a framework parameterized by only 4 dimensions. A Python library demonstrates the effectiveness of this framework, allowing implementation existing methods, as well as imagine new methods. This for example permits adapting methods developed for deep learning models to non-deep classifiers for tabular data. We benchmark existing methods on (artificial) Completely At Random (NCAR) as well as (realistic) Not At Random (NNAR) labeling noise from a variety of tasks with imperfect labeling rules. This benchmark provides new insights as well as limitations of existing methods in this setup.

Prétraitement Efficace des Données pour l'Évaluation de la Qualité Écologique dans les Environnements Marins

Houria BRAIKIA*, Sana BEN HAMIDA*
Marta RUKOZ*

*Université Paris Dauphine - PSL, LAMSADE, Pl. du Maréchal de Lattre de Tassigny, 75016 Paris
houria.braikia@dauphine.psl.eu,
https://www.lamsade.dauphine.fr/

Résumé. L'utilisation de l'apprentissage automatique (ML, Machine Learning) pour prédire l'Indice Biotique (IB) et évaluer la Qualité Écologique (QE) des environnements marins à partir de données de métabarcodage d'ADN environnemental (ADNe) marque une avancée majeure dans la biodiversité marine. Cependant, ces données complexes présentent des défis tels que la variabilité systématique et la malédiction de la dimensionnalité, nuisant à la précision des prédictions. Pour surmonter ces obstacles, nous avons développé un pipeline ML intégrant des étapes clés de prétraitement, notamment la normalisation et la réduction dimensionnelle, afin d'optimiser la qualité des prédictions. Nous avons testé le classificateur Random Forest pour prédire les classes d'QE sur sept marqueurs, en évaluant différentes combinaisons de deux techniques de normalisation et de réduction dimensionnelle. Nos expérimentations identifient la combinaison optimale et le nombre idéal de composantes réduites, établissant un protocole standardisé de prétraitement qui améliore significativement la prédiction de l'QE à partir de données de métabarcodage.

1 Introduction

Récemment, l'apprentissage supervisé (SML, Supervised Machine Learning) couplé aux données d'ADN environnemental (ADNe) a été utilisé pour évaluer l'impact des activités humaines sur la biodiversité marine et la santé écologique (Cordier et al. (2017, 2018); Frühe et al. (2021)). L'évaluation de ces activités passe par le calcul d'un Indice Biotique (IB), une métrique largement utilisée pour mesurer la Qualité Écologique (QE), en tenant compte de l'abondance et de la tolérance aux perturbations des espèces présentes (**référées sous le nom d'Unités Taxonomiques Opérationnelles ou OTU**). Les valeurs continues de l'IB sont couramment catégorisées en cinq classes discrètes de QE, allant de "très bonne" à "très mauvaise" (Cordier et al. (2017)). Bien que le SML ait été proposé pour développer des modèles prédictifs précis des valeurs de l'IB, prédire les résultats à partir des données d'ADNe reste un défi en raison de leurs caractéristiques complexes. Des erreurs peuvent survenir à diverses étapes de l'analyse de l'ADNe, introduisant plusieurs sources de variabilité. Un autre problème majeur réside dans la haute dimensionalité des données, où le nombre de caractéristiques dépasse

souvent celui des échantillons, ce qui mène à la rareté des données et au problème de la malédiction de la dimensionalité.

Ce travail vise à améliorer la prédiction de la QE des environnements marins par la conception d'une pipeline de ML générique. Cette amélioration est obtenue en incorporant des étapes de prétraitement importantes, notamment la normalisation et la réduction de la dimensionnalité, et ceci quel que soit le type d'espèce (**référé sous le nom de marqueurs**) testé, incluant les eucaryotes, les bactéries ribosomiques et les foraminifères, constituant ainsi l'entrée pour nos modèles de Forêt Aléatoire (RF, Random Forest). Cet article est organisé en cinq sections suivant l'introduction. La Section 2 résume l'application de l'apprentissage supervisé pour l'analyse de la biodiversité. La Section 3 décrit les étapes de prétraitement visant à améliorer la précision, incluant la normalisation et la réduction de la dimensionnalité. La préparation des données et les résultats sont présentés dans la Section 4. La section finale présente la conclusion.

2 Apprentissage Automatique Supervisé pour le Biomonitoring

Plutôt que de calculer l'IB pour inférer la QE, le ML prédit directement la QE à partir des données d'ADNe, évitant ainsi des calculs complexes et des bases de données de référence incomplètes. Cordier et al. (2018) a utilisé des modèles de RF entraînés sur cinq jeux de données provenant de marqueurs différents, étiquetés avec des valeurs de QE obtenues par méthode conventionnelle. L'objectif était de comparer la précision des modèles prédictifs à travers différents marqueurs pour prédire les valeurs de quatres IB. Leur étude a révélé que tous les marqueurs testés ont donné des modèles prédictifs précis. Frühe et al. (2021) a comparé la prédiction des valeurs de l'Indice Biotique Marin d'AZTI (AMBI) (Borja et al. (2000)) en utilisant la méthode des Valeurs Indicateurs (IndVals) et le SML. Ils ont utilisé les algorithmes de RF et de Machine à Vecteurs de Support sur les marqueurs de bactéries et de ciliés, montrant que l'apprentissage supervisé surpassait l'approche IndVal pour inférer la QE à partir des métabarcodes d'ADNe. Dans une étude récente, Braikia et al. (2024) a présenté une approche simplifiée pour prédire la QE. Ils ont démontré que prédire la classe de QE donne de meilleurs résultats que tenter de prédire une valeur spécifique de l'IB. Il est certain que l'utilisation de l'apprentissage automatique dans le domaine du biomonitoring, en particulier pour prédire la QE, a démontré son efficacité. Cependant, la complexité inhérente des données impacte significativement la qualité des prédictions, ce qui ouvre la voie à des techniques d'amélioration qui seront détaillées dans la section suivante.

3 Étapes de Prétraitement pour Améliorer la Précision

Le défi posé par les données de métabarcodage d'ADNe réside dans la variabilité systématique causée par les différences de profondeur de séquençage, où chaque échantillon est représenté par un nombre variable de lectures (Manor et Borenstein (2015)). Des facteurs techniques, tels que des incohérences dans l'extraction de l'ADN, la manipulation des échantillons, des variations dans la qualité des séquences, ainsi que l'incomplétude des bases de données de référence, en plus de facteurs biologiques comme les différences spécifiques aux échantillons

dans la taille moyenne du génome et la richesse en espèces, peuvent introduire une variabilité systématique (Morgan et al. (2010)). Indépendamment de son origine, la variabilité systématique amplifie les différences entre les échantillons, rendant les méthodes statistiques standard inefficaces ou potentiellement trompeuses dans leurs résultats. Par conséquent, des approches statistiques appropriées adaptées à ces données sont cruciales, nécessitant une étape de normalisation avant l'analyse (Xia (2023)).

Un autre défi provient de la haute dimensionalité des caractéristiques des données, où, dans de nombreux cas, la majorité des espèces ne sont pas observées dans la plupart des échantillons, même au sein du même type d'échantillon. Par conséquent, une table de caractéristiques qui enregistre les comptes de chaque OTU dans chaque échantillon contient souvent de nombreux zéros, indiquant des espèces non observées. Ces tables de caractéristiques, avec une abondance de comptes non observés, sont appelées "sparse" (parses) et posent des défis pour l'analyse statistique et l'apprentissage automatique. De nombreuses caractéristiques deviennent bruyantes ou redondantes et posent le problème de la « malédiction de la dimensionalité », allant de quelques échantillons à des dizaines de milliers, où le nombre de caractéristiques dans les données dépasse souvent celui des échantillons par un facteur de 20 ou plus. De plus, le nombre élevé de caractéristiques peut entraîner des problèmes de temps de traitement lors de l'analyse en aval. Ces défis s'étendent également aux modèles d'apprentissage automatique, conduisant à une performance lente, un surajustement et des résultats peu fiables (Armstrong et al. (2022)).

3.1 Normalisation

Une large gamme de méthodes a été appliquée pour normaliser les données d'abondance des OTU. Ces méthodes consistent principalement à mettre à l'échelle les données en estimant des facteurs spécifiques à chaque échantillon pour corriger les abondances des OTU. Une approche courante consiste à dériver le facteur de mise à l'échelle basé sur le nombre total de gènes observés dans chaque échantillon. Cela permet de tenir compte des différences de profondeur de séquençage, qui peuvent varier considérablement. Cependant, le nombre total de gènes est fortement dominé par les gènes les plus abondants, de sorte que leur variabilité peut avoir un impact majeur sur le facteur de mise à l'échelle (Pereira et al. (2018)).

Pour résoudre ce problème, des méthodes de normalisation plus robustes ont été proposées. Les méthodes de normalisation par médiane et par quartile supérieur estiment les facteurs de mise à l'échelle respectivement en fonction des 50% et 75% des percentiles de la distribution des comptes de gènes. En utilisant ces valeurs de percentile, l'impact des gènes très abondants est réduit.

Par ailleurs, des efforts considérables ont été investis pour comparer et résumer diverses méthodes de normalisation. Xia (2023) propose une perspective complète, catégorisant ces méthodes en quatre groupes en fonction de leur approche technique ou statistique et des types de données qu'elles visent à normaliser : méthodes de normalisation basées sur les données écologiques, méthodes traditionnelles de normalisation, méthodes de normalisation basées sur les données de séquençage ARN, et méthodes de normalisation basées sur les données de microbiome. Dans une étude de Pereira et al. (2018), où neuf méthodes de normalisation ont été évaluées sur des données de microbiome, il a été observé que les méthodes Trimmed Mean of M-values (TMM) (Robinson et Oshlack (2010)) et Relative Log Expression (RLE) ont montré la meilleure performance globale. En conséquence, ces méthodes sont recommandées pour l'analyse des données d'abondance de gènes. De plus, pour des tailles d'échantillons plus

grandes, la méthode Cumulative Sum Scaling (CSS) (Paulson et al. (2013)) a montré une performance satisfaisante.

Sur la base de ces travaux, nous avons sélectionné les techniques de normalisation CSS et TMM pour le prétraitement de nos données afin d'améliorer la prédiction de la QE. Le paragraphe suivant rappelle la définition de ces techniques.

3.1.1 Cumulative Sum Scaling (CSS)

Considérons les données brutes sous forme de matrice de comptage $M(m, n)$, où m et n sont respectivement le nombre de caractéristiques et d'échantillons. Les données brutes dans cette matrice sont représentées par des comptes c_{ij}, indiquant le nombre de fois où la caractéristique i a été observée dans l'échantillon j. Nous définissons la somme des comptes pour l'échantillon j comme dans l'Éq.1. Nous notons le l-ième quantile de l'échantillon j comme q_j^l, ce qui signifie que dans l'échantillon j, il y a l caractéristiques avec des comptes inférieurs à q_j^l. La somme des comptes pour l'échantillon j jusqu'au l-ième quantile est définie dans l'Éq.2. Enfin, nous calculons le compte normalisé comme dans l'Éq.3.

$$s_j = \sum_i c_{ij} \tag{1}$$

$$s_j^l = \sum_{i|c_{ij} \leq q_j^l} (c_{ij}) \tag{2}$$

$$\tilde{c}ij = \frac{cij}{s_j^l} \cdot N \tag{3}$$

Où N est une constante de normalisation choisie de manière appropriée.

En appliquant la normalisation CSS, l'impact des gènes très abondants et variables est minimisé, ce qui améliore la comparabilité et la fiabilité des données d'abondance de gènes pour les analyses ultérieures.

3.1.2 Trimmed Mean of M-values (TMM)

La méthode TMM compare les abondances des gènes au sein des échantillons en sélectionnant un échantillon de référence, généralement un des échantillons de l'étude. La méthode repose sur l'hypothèse que la majorité des gènes ne présentent pas de différences d'abondance significatives et devraient avoir une abondance moyenne similaire entre les échantillons (Robinson et Oshlack (2010)).

Pour effectuer la normalisation TMM, un échantillon de référence, noté r, est sélectionné. Pour chaque échantillon j, les gènes sont filtrés en fonction de leur abondance moyenne et du changement de fold (fold-change) par rapport à l'échantillon de référence. Les changements de fold restants sont ensuite pondérés par l'inverse de leur variance, et leur moyenne est calculée. Cette valeur est appelée le facteur d'ajustement $f_j^{(r)}$ (Pereira et al. (2018)).

Le facteur de normalisation N_j est ensuite donné par l'Éq.4, où Y_{ij} représente les comptes du gène $i = 1, \ldots, m$ dans l'échantillon $j = 1, \ldots, n$.

$$N_j = f_j^{(r)} \sum_{i=1}^{m} Y_{ij} \tag{4}$$

3.2 Réduction de Dimensionnalité (RD)

L'objectif de la réduction de dimensionnalité (RD) est de transformer un ensemble de données à haute dimension en une représentation avec moins de dimensions, tout en préservant les relations essentielles entre les échantillons de l'ensemble de données complet. Ce processus rend l'analyse plus gérable et facilite la création de visualisations compréhensibles par l'homme, servant de base pour les analyses ultérieures.

Plusieurs revues ont résumé et comparé ces méthodes dans leurs études. Armstrong et al. (2022) s'est concentré sur leur application aux données de microbiome [1], Nanga et al. (2021) s'est spécifiquement concentré sur la révision des méthodes de réduction dimensionnelle, Kabir et al. (2023) les ont utilisées pour des modèles d'apprentissage automatique dans la prédiction du cancer. De plus, certaines revues ont exploré leur utilité comme étape de prétraitement pour la prédiction génomique (Manthena et al. (2022)), ainsi que des études liées à l'analyse des données de séquençage ARN de cellule unique (scRNA-seq) (Xiang et al. (2021), Sun et al. (2019)).

Dans les études sur le microbiome, la réduction de dimensionnalité permet de visualiser et de réduire la dimensionnalité en raison de la présence de regroupements dans les données. Cela permet aux chercheurs d'explorer les motifs de regroupement, d'identifier des groupes différenciés et de mettre en évidence les micro-organismes spécifiques responsables de ces distinctions.

Bien que l'Analyse en Composantes Principales (ACP) soit largement utilisée pour la réduction de dimensionnalité, son efficacité est limitée dans les données de microbiome en raison de problèmes tels que la composition (variations dans les comptes entre les échantillons), conduisant à des résultats erronés. Pour y remédier, la Décomposition en Valeurs Singulières (SVD, Singular Value Decomposition) (Stewart (1993)) et l'ACP Robuste (RPCA, Robust Principal Component Analysis) (Martino et al. (2019)) ont été proposées comme solutions. L'Analyse des Coordonnées Principales (PCoA, Principal Coordinate Analysis) (Kruskal et Wish (1978)), également connue sous le nom de mise à l'échelle multidimensionnelle métrique (MDS, Metric Multidimensional Scaling), est une autre technique couramment utilisée dans l'analyse des données de microbiome. Contrairement à l'ACP, la PCoA fonctionne sur des matrices de distance, capturant les relations entre les points de données en fonction des dissimilarités par paires, utilisant généralement des distances écologiquement significatives telles que UniFrac (Lozupone et al. (2007)). Certaines méthodes comme l'Analyse de Correspondance Canonique (CCA, Canonical Correspondence Analysis) (ter Braak (1986)) peuvent utiliser à la fois des tables de caractéristiques et des métadonnées d'échantillons, estimant conjointement des représentations d'échantillons de faible dimension et la contribution des vecteurs de métadonnées.

Dans cette étude, nous appliquons la SVD comme technique centrale soutenant les méthodes mentionnées ci-dessus. Elle est couramment utilisée dans la littérature pour gérer efficacement les données OTU. La SVD est une technique de factorisation de matrice. Elle permet d'éliminer les composants moins importants afin d'approximer la représentation avec un nombre de dimensions souhaité. Dans l'algorithme SVD, une matrice A peut être représentée comme un produit de trois matrices, comme indiqué dans Éq.5 : U, Σ, et V (YongchangWang

1. Le microbiome fait référence à l'ensemble des micro-organismes formant une communauté vivant dans un environnement spécifique (Armstrong et al. (2022)).

et Zhu (2017)).

$$A = U\Sigma V^T \tag{5}$$

U et V sont des matrices avec des colonnes orthogonales. Σ est une matrice diagonale avec des valeurs réelles non négatives, représentant les valeurs singulières de A. Ces valeurs sont ordonnées de la plus grande à la plus petite et diminuent très rapidement. Souvent, un petit pourcentage (comme 10% ou même 1%) des valeurs singulières contribue de manière significative à 99% de la somme totale des valeurs singulières. En termes simples, nous pouvons n'utiliser que les premières valeurs singulières pour fournir une bonne approximation de la matrice (YongchangWang et Zhu (2017)).

4 Étude Expérimentale

4.1 Données

Dans notre étude, nous avons utilisé un jeu de données provenant de Cordier et al. (2018), qui comprend des tableaux OTU pour cinq marqueurs : eucaryotes (V1V2, V4 et V9), bactériens ribosomiques (V3V4) et foraminifères (37F). Le deuxième jeu de données, provenant de Frühe et al. (2021), inclut des tableaux OTU pour deux marqueurs, Bactéries (V3V4) et Ciliés (V9). Ces jeux de données sont accessibles sur les dépôts Zenodo et GitHub, respectivement [2]. Toutes les échantillons ont été collectés autour des fermes aquacoles de saumon en Norvège.

Ces fichiers contiennent les noms des échantillons ainsi que l'abondance de chaque OTU dans chaque échantillon. Le tableau ci-dessous 1 présente les dimensions des jeux de données utilisés, y compris le nombre d'OTUs (lignes) et le nombre d'échantillons (colonnes) pour chaque marqueur.

Marqueur	**Dimensions (lignes, colonnes)**
V1V2	(941375, 148)
V3V4	(3630, 145)
V4	(1050816, 144)
37F	(12332, 127)
V9	(3829, 148)
Ciliates	(2650, 167)
Bacteria	(38845, 165)

TAB. 1 – *Dimensions des différentes matrices OTU-échantillon utilisées dans l'analyse.*

Les jeux de données diffèrent en termes de diversité microbienne qu'ils capturent, de la résolution de l'identification taxonomique et du nombre de caractéristiques (dimensionnalité) en raison des variations dans les marqueurs utilisés (V1V2, V3V4, V4, 37F et V9).

Un autre fichier est attaché à chaque jeu de données, contenant les métadonnées des échantillons. Ces fichiers incluent une colonne pour l'AMBI, calculé à l'aide de la méthode traditionnelle. Ces informations seront utiles pour notre analyse ultérieure.

2. `https://zenodo.org/record/1286477`, `https://github.com/lafrue/Salmon_ISM`

4.2 Prétraitement des données

Le prétraitement des données suit deux étapes : le nettoyage et la transformation. L'étape de nettoyage vise à filtrer les OTUs rares. Les OTUs rares peuvent résulter de chimères (fusion de séquences différentes) formées pendant les étapes de métabarcodage en raison de la contamination, des erreurs d'amplification et des erreurs de substitution pendant le séquençage. Cette étape de filtrage est nécessaire pour éviter le bruit dans les données d'entraînement et est résumée dans les lignes 3 à 12 de l'algorithme 1. La transformation concerne la normalisation et la réduction de dimensionnalité décrites dans la section 3 et est résumée dans les lignes 13 et 14.

Algorithm 1 Algorithme de prétraitement des données

1: **Entrée :** OTU_table, τ_sample, τ_OTU
2: **Sortie :** Fil_OTU_tab
3: **for** i = 1 à m **do**
4: **if** sum($OTU\_table[i,:]) \geq \tau\_sample$ **then**
5: Ajouter $OTU\_table[i,:]$ à $Filtered\_samples$
6: **end if**
7: **end for**
8: **for** j = 1 à n **do**
9: **if** sum($Filtered\_samples[:,j]) \geq \tau\_OTU$ **then**
10: Ajouter $Filtered\_samples[:,j]$ à Fil_OTU_tab
11: **end if**
12: **end for**
13: Normaliser le Fil_OTU_tab
14: Réaliser la réduction de dimensionnalité en utilisant SVD
15: Transformer la colonne $AMBI$ en classes QE puis la fusionner avec Fil_OTU_tab

L'étape de prétraitement prend en entrée un tableau OTU $OTU\_table[n, m]$ et un tableau de métadonnées, et produit les données d'apprentissage en sortie pour être utilisées lors de l'étape d'entraînement. Tout d'abord, nous avons effectué un filtrage sur les lignes des échantillons et les colonnes des OTUs. Nous avons conservé uniquement les échantillons dont la profondeur de séquençage dépasse la moyenne totale des lectures par échantillon (τ_sample) (lignes 3 à 7), qui était de 10 000 lectures pour les marqueurs (V1V2, V3V4, V4, V9 et 37F), et de 1 000 pour les bactéries et les ciliés. Ensuite, nous avons supprimé les OTUs rares (lignes 8 à 12), définis comme ceux ayant une abondance totale (τ_OTU) inférieure à 100 pour les marqueurs (V1V2, V3V4, V4, 37F), inférieure à 1 000 pour V9, et inférieure à 10 pour les bactéries et les ciliés. Ce seuil de filtrage est conforme à la méthodologie décrite dans les codes de Cordier et al. (2018); Frühe et al. (2021).

Ensuite, le tableau OTU filtré Fil_OTU_tab subit une normalisation et une réduction de dimensionnalité (lignes 13 et 14). Dans cette étude, nous cherchons à évaluer les performances de l'apprentissage automatique pour prédire les classes QE à partir des données normalisées avec CSS et TMM (expliquées dans la section 3.1) et réduites avec SVD (expliqué dans la section 3.2). Pour la normalisation CSS, nous avons utilisé la fonction cumNorm du package R metagenomeSeq. Pour la normalisation TMM, la fonction calcNormFactors du package R edgeR a été utilisée. L'utilisation du langage R pour importer les packages de normalisation dans le code Python a été rendue possible grâce au package rpy2 de Python. La logique derrière

le choix du nombre de composants à conserver pour la réduction de dimensionnalité via SVD est discutée dans la section 4.3.

Dans le prétraitement du tableau de métadonnées, nous avons converti les valeurs AMBI en cinq classes allant de 1 pour "très bon" à 5 pour "très mauvais", en utilisant des seuils prédéfinis d'AMBI. Ensuite, cette nouvelle colonne du tableau de métadonnées a été fusionnée avec le tableau OTU prétraité, en fonction de la colonne des noms des échantillons (ligne 15). Seuls les échantillons ayant des valeurs BI correspondantes ont été conservés. Les données d'apprentissage préparées fil_OTU_tab sont divisées en ensembles d'entraînement et de test. Ensuite, un modèle RF est utilisé pour générer les classes QE prédites pour chaque échantillon de l'ensemble de test. Le modèle a été construit à l'aide de la bibliothèque Python Scikit-Learn avec 200 arbres dans la forêt (hyperparamètre $n\_estimators$). La performance du modèle a été évaluée à l'aide des métriques kappa et Recall. Les résultats pour chaque jeu de données, ainsi que les meilleurs paramètres, sont présentés dans la section 4.4.

4.3 Choix du nombre de composants pour la SVD

Le choix du nombre optimal de composants avant la réduction des données repose sur deux objectifs : maximiser à la fois la variance entre les caractéristiques et la précision du modèle RF avec le minimum de composants possible. Pour évaluer la variance conservée par chaque nombre de composants, nous avons calculé la variance expliquée cumulée [3] pour les deux méthodes de normalisation des données, de 1 jusqu'au nombre de composants qui retient 99% de la variance, afin de mesurer l'impact de la normalisation sur la variance des données.

Pour la normalisation CSS, nous avons conservé les composants qui préservent 90% de la variance, afin de trouver un équilibre entre la capture de la variance et la stabilité du modèle, car l'ajout de composants supplémentaires n'améliorait pas les performances. Pour la normalisation TMM, le premier composant seul capture plus de 92% de la variance pour la plupart des marqueurs. Étant donné qu'un seul composant était insuffisant pour le modèle RF, nous avons utilisé un nombre fixe de 20 composants pour garantir la capture de plus de 90% de la variance tout en maintenant une précision robuste du RF.

4.4 Résultats et Discussion

La performance de la prédiction des classes QE à l'aide de RF, évaluée en termes de métriques de Rappel et de Kappa, est présentée dans le Tableau 2. Les valeurs de Kappa variant de 0,01 à 0,2 suggèrent un accord faible, tandis que des valeurs comprises entre 0,8 et 1 indiquent un accord presque parfait. Le Rappel est calculé pour chaque classe individuellement en divisant le nombre d'instances correctement classées dans cette classe par le nombre total d'instances de cette classe, un Rappel plus élevé indiquant une meilleure performance.

Les meilleurs résultats sont indiqués en gras. L'analyse couvre à la fois les données originales et les données réduites, en utilisant la SVD. De plus, la normalisation des données a été effectuée à l'aide des méthodes CSS et TMM, permettant une comparaison entre ces deux approches de normalisation. Le tableau fournit également des informations sur le nombre de composants, noté N, utilisés pour la prédiction pour chaque marqueur.

3. La Variance Expliquée Cumulée représente la proportion cumulée de la variance du jeu de données expliquée par chaque composant, offrant des informations sur l'efficacité de la réduction de dimensionnalité.

		Données originales			Données réduites (SVD)		
	Marker	N	Kappa	Recall	N	Kappa	Recall
CSS	V1V2	2680	**0.956**	[1., 0., 1., 1.]	49	**0.956**	[1., 0., 1., 1.]
	V3V4	2391	**0.834**	[1., 0., 0.667, 1.]	25	**0.834**	[1., 0., 0.667, 1.]
	V4	2031	**0.916**	[1., 0., 1., 1.]	54	0.826	[1., 0., 1., 0.75]
	37F	1594	0.838	[1., 0., 0.667, 1.]	51	**0.893**	[1., 1., 0.667, 1.]
	V9	3588	0.804	[0.952, 0., 0.875, 0.889]	15	0.882	[0.952, 0.5, 1., 0.889]
	Ciliates	1943	0.779	[1., 0., 0.875, 0.75]	91	0.772	[1., 0., 0.625, 0.917]
	Bacteria	9165	0.892	[1., 0., 0.857, 1.]	92	0.892	[1., 0., 1., 0.923]
TMM	V1V2	2680	**0.956**	[1., 0., 1., 1.]	20	**0.956**	[1., 0., 1., 1.]
	V3V4	2391	**0.834**	[1., 0., 0.667, 1.]	20	**0.834**	[1., 0., 0.667, 1.]
	V4	2031	**0.916**	[1. 0., 1., 1.]	20	**0.916**	[1., 0., 1., 1]
	37F	1594	0.838	[1., 0., 0.667, 1.]	20	0.889	[1., 1., 0.833, 1.]
	V9	3588	**0.92**	[1, 0., 1, 1]	20	0.881	[1, 0.5, 0.875, 0.889]
	Ciliates	1943	**0.856**	[1., 0., 0.875, 0.917]	20	0.819	[0.957, 0., 0.875, 0.917]
	Bacteria	9165	0.892	[1., 0., 0.857, 1]	20	**0.929**	[1., 0., 1., 1.]

TAB. 2 – *Kappa et Recall des résultats de la prédiction QE avec le modèle RF pour les combinaisons des méthodes de normalisation et de réduction de dimension.*

L'intégration de la réduction de dimensionnalité via SVD, accompagnée des méthodes de normalisation CSS et TMM, améliore l'exactitude de la prédiction des classes QE. En général, les données réduites ont surpassé les données originales pour divers marqueurs, à l'exception de V4 dans le cas de la normalisation CSS et de V9 et des Ciliates avec la normalisation TMM. Pour les marqueurs V1V2, V3V4, Ciliates et Bacteria, les résultats du modèle RF étaient identiques avec les données originales ou réduites normalisées avec CSS. L'utilisation de SVD a amélioré les résultats pour 37F et V9, tandis que les données originales ont mieux performé pour V4.

Concernant la normalisation TMM, les résultats étaient équivalents pour V1V2, V3V4 et V4, tandis que SVD surpassait les données originales pour 37F et Bacteria. Nous avons ainsi réussi à réduire la dimensionnalité du jeu de données, passant de 2000 à 9165 caractéristiques à un nombre plus réduit, tout en obtenant des résultats égaux ou meilleurs que ceux des données originales. De plus, les résultats du modèle RF pour tous les marqueurs montrent un accord parfait (Kappa > 0.8) dans le cadre de la normalisation TMM. La variation du rappel à travers les classes est principalement due aux déséquilibres de classe, le modèle RF réussissant bien pour la classe majoritaire mais ayant des difficultés à distinguer les classes moins fréquentes, soulignant ainsi le défi de différencier des classes presque identiques.

Dans certains cas, les valeurs de Kappa restent identiques car le processus de réduction de dimensionnalité préserve la structure essentielle et la séparabilité des classes des données. Cela suggère que les caractéristiques supprimées n'étaient pas essentielles à la capacité du modèle à obtenir un accord cohérent. De tels résultats soulignent la robustesse du modèle, démontrant sa capacité à maintenir la performance même lorsqu'il est appliqué à des jeux de données réduits.

Chaque méthode de normalisation et de réduction de dimensionnalité présente des avantages distincts et doit être choisie avec soin en fonction de critères spécifiques. Des techniques largement utilisées, comme CSS et TMM, sont sélectionnées en raison de leur efficacité prou-

vée dans les études récentes, en particulier pour leur capacité à réduire les biais et à prendre en compte les variations systématiques. De même, des méthodes de réduction de dimensionnalité comme SVD sont choisies pour leur capacité à mettre en évidence les motifs clés dans les données tout en minimisant l'impact des caractéristiques non pertinentes ou bruyantes.

Comparés aux résultats de référence de Cordier et al. (2018); Frühe et al. (2021) présentés dans le Tableau 3, qui n'utilisaient que la normalisation CSS, nos meilleurs résultats, obtenus grâce à la combinaison TMM et SVD, surpassent les références pour tous les marqueurs, à l'exception de V3V4 et V9, où nos résultats étaient très proches et acceptables. Notre approche a démontré une performance systématiquement supérieure dans les sept ensembles de données, surpassant ainsi l'état de l'art.

Marqueurs	Résultats Kappa de la littérature	Nos résultats Kappa
V1V2	0.866	**0.956**
V3V4	**0.918**	0.834
V4	0.877	**0.916**
37F	0.832	**0.889**
V9	**0.927**	0.881
Ciliates	0.8	**0.819**
Bacteria	0.53	**0.929**

TAB. 3 – *Comparaison des résultats de Kappa*

Ce que nous retenons : Nous recommandons donc l'utilisation des données ADNe traitées avec SVD et TMM. Cette approche combinée est suggérée pour être intégrée en tant qu'entrée dans le modèle RF afin de classifier efficacement les catégories QE.

5 Conclusion et Perspectives

En conclusion, notre approche démontre un potentiel significatif en tant qu'outil précieux pour la surveillance écologique avec des données ADNe, ouvrant la voie au développement d'outils basés sur l'intelligence artificielle qui améliorent la biomonitoring et permettent une surveillance plus complète des environnements marins. Les recherches futures exploreront l'intégration de modèles d'apprentissage automatique supplémentaires, l'incorporation de données provenant de différentes régions, ainsi que l'application de notre cadre à différents contextes écologiques afin de valider davantage sa généralisabilité.

Références

Armstrong, G., G. Rahman, C. Martino, D. McDonald, A. Gonzalez, G. Mishne, et R. Knight (2022). Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. *Frontiers in Bioinformatics 2*.

Borja, A., J. Franco, et V. Pérez (2000). A marine biotic index to establish the ecological quality of soft-bottom benthos within european estuarine and coastal environments.

Marine Pollution Bulletin 40(12), 1100–1114. doi:https://doi.org/10.1016/S0025-326X(00)00061-8.

Braikia, H., S. Ben Hamida, et M. Rukoz (2024). Random forest classifier for marine biodiversity analysis. In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1–8.

Cordier, T., P. Esling, F. Lejzerowicz, J. Visco, A. Ouadahi, C. Martins, et et al. (2017). Predicting the ecological quality status of marine environments from edna metabarcoding data using supervised machine learning. *Environ Sci Technol 51*(16), 9118–9126. doi:10.1021/acs.est.7b01518.

Cordier, T., D. Forster, Y. Dufresne, C. Martins, T. Stoeck, et J. Pawlowski (2018). Supervised machine learning outperforms taxonomy-based environmental dna metabarcoding applied to biomonitoring. *Mol Ecol Resour 18*(6), 1381–1391. doi:10.1111/1755-0998.12926.

Frühe, L., T. Cordier, V. Dully, H. Breiner, G. Lentendu, J. Pawlowski, et et al. (2021). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using edna metabarcodes. *Mol Ecol 30*(13), 2988–3006. doi:10.1111/mec.15434.

Kabir, M. F., T. Chen, et S. A. Ludwig (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics 3*, 100125.

Kruskal, J. et M. Wish (1978). *Multidimensional Scaling*. Number 07-011 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Newbury Park : Sage Publications.

Lozupone, C. A., M. Hamady, S. T. Kelley, et R. Knight (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology 73*(5), 1576–1585.

Manor, O. et E. Borenstein (2015). MUSiCC : a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biology 16*(1), 53.

Manthena, V., D. Jarquín, R. K. Varshney, M. Roorkiwal, G. P. Dixit, C. Bharadwaj, et R. Howard (2022). Evaluating dimensionality reduction for genomic prediction. *Frontiers in Genetics 13*.

Martino, C., J. T. Morton, C. A. Marotz, L. R. Thompson, A. Tripathi, R. Knight, et K. Zengler (2019). A novel sparse compositional technique reveals microbial perturbations. *mSystems 4*, e00016–19.

Morgan, J. L., A. E. Darling, et J. A. Eisen (2010). Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *PLOS ONE 5*(4), e10209. Publisher : Public Library of Science.

Nanga, S., A. T. Bawah, B. A. Acquaye, M.-I. Billa, F. D. Baeta, N. A. Odai, S. K. Obeng, et A. D. Nsiah (2021). Review of Dimension Reduction Methods. *Journal of Data Analysis and Information Processing 9*(3), 189–231. Number : 3 Publisher : Scientific Research Publishing.

Paulson, J. N., O. C. Stine, H. C. Bravo, et M. Pop (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods 10*(12), 1200–1202. Number : 12 Publisher : Nature Publishing Group.

Pereira, M. B., M. Wallroth, V. Jonsson, et E. Kristiansson (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics 19*(1), 274.

Robinson, M. D. et A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology 11*(3), R25.

Stewart, G. W. (1993). On the Early History of the Singular Value Decomposition. *SIAM Review 35*(4), 551–566. Publisher : Society for Industrial and Applied Mathematics.

Sun, S., J. Zhu, Y. Ma, et X. Zhou (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology 20*(1), 269.

ter Braak, C. J. F. (1986). Canonical correspondence analysis : A new eigenvector technique for multivariate direct gradient analysis. *Ecology 67*(5), 1167–1179.

Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome cancer research. *Gut Microbes 15*(2), 2244139. PMID : 37622724.

Xiang, R., W. Wang, L. Yang, S. Wang, C. Xu, et X. Chen (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in Genetics 12*.

YongchangWang et L. Zhu (2017). Research and implementation of svd in machine learning. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 471–475.

Summary

The integration of machine learning (ML) to predict the Biotic Index (BI) for assessing Ecological Quality (EQ) of marine environments using environmental DNA (eDNA) metabarcoding data represents a significant advance in biomonitoring. However, the complexity of these data types poses challenges like systematic variability and the curse of dimensionality, impacting prediction quality. To address these issues, we propose a ML pipeline with crucial preprocessing steps, including normalization and dimensionality reduction techniques, to enhance prediction quality. Our work involves comparing the performance of the Random Forest Classifier for predicting EQ classes across seven markers. We examine various combinations of two normalization techniques and a suitable dimensionality reduction with an optimal number of reduced features. Through rigorous experimentation, we identify the most effective combination and optimal number of components to retain, establishing a standardized preprocessing protocol for this data type. This robust framework for preprocessing metabarcoding data significantly advances EQ prediction.

Interpolation pour l'augmentation de données géo-référencées : Application à la prédiction des adventices de la canne à sucre à La Réunion

Frédérick Fabre-Ferber[***], Dominique Gay[**], Jean-Christophe Soulie[**], Jean Diatta[*], Odalric-Ambrym Maillard[***],

[*]LIM EA2525, Université de La Réunion
frederick.fabre@univ-reunion.fr dominique.gay@univ-reunion.fr jean.diatta@univ-reunion.fr
[**]UPR Recyclage et risque, CIRAD
jean-christophe.soulie@cirad.fr frederick.fabre-ferber@cirad.fr
[***]SCOOL, Inria
odalric.maillard@inria.fr

Résumé. L'augmentation de données est essentielle pour développer des modèles d'apprentissage supervisé robustes, notamment avec des petits jeux de données. Cette étude évalue des techniques d'interpolation géo-référencées pour prédire la présence de Commelina benghalensis L. dans des parcelles de canne à sucre à La Réunion ; avec pour objectifs, la performance prédictive en fonction du nombre de données ajoutées et la cohérence spatiale des données augmentées. Deux approches ont été comparées : les processus gaussiens (GPs) avec différents noyaux et le krigeage avec divers variogrammes. Les GPs, notamment avec noyaux combinés, améliorent significativement les performances tout en nécessitant moins de données, tandis que le krigeage, bien que légèrement moins performant, offre une couverture spatiale plus homogène.

1 Introduction

Dans le domaine de l'intelligence artificielle centrée sur les données, l'augmentation de données est essentielle pour enrichir les ensembles d'apprentissage sans collecte supplémentaire (Zha et al., 2023; Cui et al., 2024). Dans la littérature, les méthodes développées utilisent e.g., les autoencodeurs variationnels (VAE), les réseaux antagonistes génératifs (GANs) ou les modèles de diffusion (Li et al., 2021; Zhu et al., 2020; Kotelnikov et al., 2023). Nous explorons ici l'augmentation de données appliquée à la prédiction de l'espèce Commelina benghalensis L. dans des parcelles de canne à sucre à La Réunion. Ce contexte présente des contraintes spécifiques : collecte chronophage, volume limité de données et dimension spatiale. Pour relever ces défis, nous testons des méthodes géo-référencées, comme le krigeage et les processus gaussiens (GPs), adaptées à la modélisation des dépendances spatiales. Cette étude vise à (i) évaluer l'impact de ces méthodes sur les performances prédictives, (ii) analyser leur évolution en fonction des données ajoutées, et (iii) vérifier la préservation de la cohérence spatiale. Les sections suivantes détaillent ces concepts, la méthodologie, les résultats et leur discussion.

2 Méthode proposée

L'interpolation est une méthode efficace pour enrichir des données géo-référencées, notamment lorsque les observations disponibles sont limitées. Cette étude compare le krigeage et les processus gaussiens (GPs) pour augmenter un jeu de données dédié à la prédiction du recouvrement de Commelina benghalensis L. dans des parcelles de canne à sucre à La Réunion. Les objectifs principaux sont d'identifier la méthode d'interpolation la plus performante, d'étudier la convergence des performances selon le nombre de données ajoutées, et d'évaluer la cohérence spatiale des distributions interpolées. Le jeu de données comprend 745 observations avec des variables explicatives (altitude, pluviométrie, etc.) récupérées via le service Meteor [1] et une variable cible de recouvrement (0-100).

Les GPs sont testés avec plusieurs noyaux (linéaire, RBF, quadratique, voir Table 1) et des combinaisons optimales fondées sur le BIC (Duvenaud, 2014). Le co-krigeage, extension du krigeage classique, est utilisé pour intégrer ces variables auxiliaires. Le protocole évalue les performances prédictives de divers algorithmes de régression, dont la régression linéaire (LR), Ridge (RR), SVR, forêts aléatoires (RF), Gradient Boosting (GB), K-NN et MLP, via l'erreur quadratique moyenne (MSE). Trois expérimentations sont menées : (i) comparaison des performances avec et sans augmentation (+200 points), (ii) analyse des performances selon le nombre de points ajoutés, et (iii) comparaison des cartes de densité de recouvrement pour les jeux augmentés (+300 points). Les paramètres des GPs sont optimisés par descente de gradient sur la vraisemblance avec `GPy` [2], tandis que ceux des variogrammes sont ajustés avec `PyKrige` [3]. Les hyperparamètres des algorithmes de régression sont issus des valeurs par défaut de `scikit-learn` Pedregosa et al. (2011).

Méthode	Acronyme
P. Gauss. - noyau linéaire	GP-LIN
P. Gauss. - noyau polynomial	GP-POLY
P. Gauss. - noyau RBF	GP-RBF
P. Gauss. - comb. de noyaux	GP-COMB
Krige - variog. linéaire	COKR-LIN
Krige - variog. exponentiel	COKR-EXP
Krige - variog. gaussien	COKR-GAU
Krige - variog. sphérique	COKR-SPHE

TAB. 1 – Méthodes d'interpolation utilisées

3 Resultats et interprétation

Les résultats de la table 2 présentent les erreurs quadratiques moyennes (MSE) pour divers algorithmes de régression appliqués à des jeux de données augmentés via différentes techniques d'interpolation. La figure 1 montre l'évolution de la performance pour un algorithme particulier plusieurs nombre de points rajoutés par les différentes interpolations. Enfin, la table 3 montre les différences moyennes de densité pour plusieurs zones de l'île avec les jeux de données augmentés selons plusieurs méthodes d'interpolations. Les résultats de cette étude mettent en évidence l'impact des techniques d'interpolation, telles que les processus gaussiens (GPs) et le krigeage, sur l'amélioration des performances prédictives et la répartition spatiale des données géo-référencées. Globalement, l'ajout de données synthétiques via ces méthodes améliore systématiquement les performances des algorithmes de régression par rapport au jeu

1. https ://smartis.re/METEOR
2. https ://gpy.readthedocs.io/en/deploy/
3. https ://geostat-framework.readthedocs.io/projects/pykrige/en/stable/

de données initial. Les GPs, en particulier avec le noyau combiné (GP-COMB), se distinguent par leur capacité à maximiser rapidement les performances prédictives, nécessitant moins de points supplémentaires pour converger vers des résultats optimaux. Ces méthodes s'avèrent particulièrement adaptées aux algorithmes comme les forêts aléatoires, le gradient boosting et K-NN, grâce à leur aptitude à capturer des relations complexes dans les données.

À l'inverse, le krigeage, bien que légèrement moins performant sur certains modèles, offre une couverture spatiale plus homogène, ce qui peut être un atout dans des zones géographiques moins denses en observations. Les résultats montrent également que l'évolution des performances varie selon le nombre de points ajoutés : les GPs convergent plus rapidement, tandis que le krigeage nécessite un effort d'augmentation plus important pour atteindre des résultats comparables. En termes de répartition spatiale, les GPs produisent des prédictions globales parfois moins cohérentes avec les données d'origine, tandis que le krigeage préserve mieux la structure spatiale initiale. Ces différences soulignent l'importance de choisir une méthode en fonction des objectifs spécifiques : les GPs sont idéaux pour maximiser rapidement la précision globale, tandis que le krigeage est préférable pour garantir une fidélité spatiale.

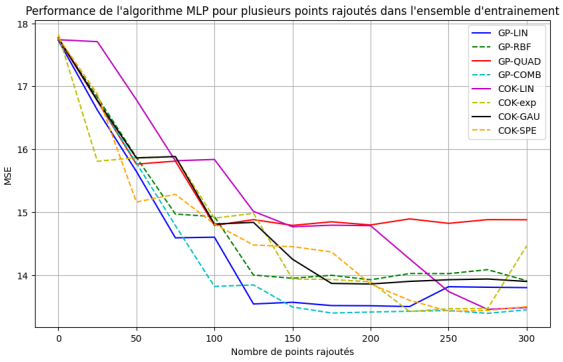


FIG. 1 – Performance en terme de MSE de l'algoritme MLP pour 0 à 300 points rajoutés par différentes techniques d'interpolation.

Modèle	Base	GP-RBF	GP-LIN	GP-QUAD	GP-COMB	CoK-LIN	CoK-EXP	CoK-GAU	CoK-SPHE
LR	18.80	13.68	13.71	**13.67**	14.57	14.58	14.78	14.66	14.56
RR	14.84	13.68	13.71	**13.67**	13.57	14.57	14.58	14.66	14.56
SVR	16.80	14.09	**13.74**	13.93	14.98	14.87	15.12	15.11	15.11
RF	23.83	14.05	14.05	14.01	**13.35**	13.35	13.62	13.53	13.55
BG	36.45	13.65	13.55	13.57	**13.27**	13.34	13.28	13.28	13.34
KNN	16.30	14.44	14.55	14.44	**13.17**	13.18	13.21	13.17	13.17
MLP	17.17	13.47	13.55	13.54	**13.41**	13.49	13.47	13.44	**13.38**

TAB. 2 – Résultats des modèles de régression avec différentes méthodes d'interpolation pour 200 points générés

	GP-COMB	GP-LIN	GP-RBF	COK-SPHE	COK-EXP
Nord-Nord Est	**+1.1**	**+1.27**	**+1.9**	-0.2	-0.2
Est	**+1.56**	**+1.32**	-0.4	+0.1	+0.1
Ouest	+0.2	+0.2	**-0.8**	-0.14	-0.14
Sud-Sud Est	**-1.2**	-0.24	**+0.8**	-0.1	-0.1

TAB. 3 – Difference de recouvrement moyens entre les points du jeu de données sans augmentation et les jeux de données aumgentés par les différentes techniques d'interpolation. Les différences signicatives sont en gras.

4 Conclusion

Dans cette étude, nous avons exploré diverses techniques d'interpolation pour augmenter des données géo-référencées et évalué leur efficacité à travers plusieurs algorithmes de régression. Nous avons également analysé leur impact sur la répartition spatiale de *Commelina benghalensis L.*. Deux approches principales ont été examinées : les processus gaussiens (GP) avec différents noyaux et le krigeage avec divers variogrammes. Les résultats montrent que GP-COMB et GP-LIN améliorent significativement les performances prédictives tout en nécessitant moins de points pour converger. Bien que le krigeage soit moins performant et converge plus lentement, il offre un recouvrement spatial plus homogène.

Références

Cui, L., H. Li, K. Chen, L. Shou, et G. Chen (2024). Tabular data augmentation for machine learning : Progress and prospects of embracing generative ai. *arXiv preprint arXiv :2407.21523*.

Duvenaud, D. (2014). The kernel cookbook : Advice on covariance functions. *URL https ://www. cs. toronto. edu/duvenaud/cookbook*.

Kotelnikov, A., D. Baranchuk, I. Rubachev, et A. Babenko (2023). Tabddpm : Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pp. 17564–17579. PMLR.

Li, G., Z. Sun, L. Qian, Q. Guo, et W. Hu (2021). Rule-based data augmentation for knowledge graph embedding. *AI Open 2*, 186–196.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Zha, D., Z. P. Bhat, K. Lai, F. Yang, Z. Jiang, S. Zhong, et X. Hu (2023). Data-centric artificial intelligence : A survey. *CoRR abs/2303.10158*.

Zhu, Y., M. R. Min, A. Kadav, et H. P. Graf (2020). S3VAE : Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation. pp. 6538–6547.

Summary

Data augmentation is essential to develop robust supervised learning models, especially with limited datasets. This study evaluates geo-referenced interpolation techniques to predict the presence of Commelina benghalensis L. in sugarcane plots in La Réunion. Two approaches were compared: Gaussian processes (GPs) with different kernels and kriging with different variograms. Objectives included identifying the methods offering the best predictive performance, analyzing performance according to the number of data items added, and assessing the spatial consistency of augmented sets. GPs, especially with combined kernels, significantly improve performance while requiring less additional data, while kriging, although slightly less efficient, offers a more homogeneous spatial coverage.

Découverte de Contraintes Monotones pour la Prédiction de Propriétés Physiques des Matériaux

Thamer Mecharnia*, Mathieu d'Aquin*, Liudmyla Klochko*

*LORIA, Université de Lorraine, CNRS, INRIA 54506, Vandœuvre-lès-Nancy, France
prenom.nom@loria.fr

Résumé. Cette étude vise à découvrir des contraintes monotones reliant les propriétés physiques des matériaux, permettant d'améliorer la prédiction et l'analyse des propriétés physiques des matériaux, notamment leur conductivité thermique. Des règles de monotonicité sont extraites en analysant les relations entre les propriétés physiques et les méthodes employées pour leur mesure, permettant ainsi d'identifier des contraintes de monotonicité croissante ou décroissante entre ces propriétés. Cette approche a permis d'identifier des règles prédictives qui expliquent comment les variations des propriétés physiques des matériaux dépendent des changements de méthode d'estimation. En proposant une méthode pour explorer et analyser les relations entre les propriétés physiques des matériaux, nous avons pu générer des règles monotones précises qui prédisent les variations monotones de ces propriétés.

1 Introduction

Dans cet article, nous présentons une approche visant à découvrir des contraintes monotones entre les propriétés physiques des matériaux. Les contraintes monotones représentent des relations cohérentes, telles que des relations directes ou inverses entre des variables. Ces contraintes sont utilisées pour guider la prédiction, et la rendant ainsi plus précise.

Pour cela, nous nous appuyons sur une approche pour extraire des connaissances d'adaptation à partir de bases de cas (voir, par exemple, d'Aquin et al. (2007)). Nous l'appliquons pour découvrir des règles de variations liant les changements de valeurs de certaines variables à d'autres et les interprétant ainsi comme des contraintes monotones.

Une approche de la découverte de matériaux consiste à estimer les propriétés physiques d'un grand nombre de matériaux plausibles mais pas encore observés, afin d'identifier ceux qui sont les plus susceptibles d'avoir les valeurs souhaitées pour ces propriétés. La conductivité thermique est un bon exemple, car les matériaux à faible conductivité thermique ont de nombreuses applications. Cependant, un problème clé est que la conductivité thermique des matériaux est extrêmement difficile à estimer. Les meilleures méthodes basées sur la simulation peuvent prendre des millions d'heures de calcul et dépendent toujours de paramètres précisément choisis et calculés. Plusieurs modèles de calcul plus simples ont été développés (par exemple AGL Toher et al. (2014)), mais avec une précision nettement inférieure. En conséquence, il existe plusieurs ensembles de données, contenant entre quelques dizaines et quelques

milliers de matériaux avec des valeurs estimées de conductivité thermique obtenues par différentes méthodes. Ces ensembles de données se chevauchent souvent partiellement, avec une grande ambiguïté sur la manière dont les matériaux sont identifiés. De plus, leurs estimations peuvent varier considérablement, ce qui rend difficile la recherche en science des matériaux axée sur les matériaux à faible conductivité thermique. Pour cette raison, nous sommes en train de construire un graphe de connaissances (GC) avec l'objectif de : 1- créer des liens fiables reliant des matériaux équivalents trouvés dans différents ensembles de données, et 2- inclure des (méta)informations claires sur les méthodes et les sources utilisées pour estimer les valeurs de conductivité thermique pour chacun des matériaux. Ce faisant, nous visons à permettre une meilleure compréhension de la manière dont les estimations de conductivité thermique peuvent être exploitées dans la découverte de matériaux, et à créer de nouveaux modèles combinant ces méthodes existantes pour un criblage plus précis des matériaux candidats.

Dans cet article, nous utilisons une version initiale du graphe de connaissances pour se concentrer spécifiquement sur cette dernière partie. L'objectif est d'obtenir un modèle efficace de prédiction de la conductivité thermique. Des tests initiaux ayant montré que des relations existent entre les descripteurs utilisés pour les matériaux et la conductivité thermique, nous voulons en particulier montrer : 1- comment l'extraction de règles d'adaptation peut mettre en avant ces relations, et 2- comment l'exploitation de ces règles interprétées comme contraintes monotones peut permettre d'améliorer la précision de la prédiction.

Le reste de cet article est organisé comme suit. Dans la section 2, nous présentons des travaux connexes. Dans la section 3, nous résumons le processus utilisé pour créer le GC pour la conductivité thermique. Ensuite, dans la section 4, nous présentons notre approche d'apprentissage pour extraire les contraintes monotones et entraîner un modèle d'apprentissage exploitant ces contraintes. La section 5 présente les résultats obtenus en appliquant cette méthode sur le GC de conductivité thermique créé. Enfin, la section 6 présente les conclusions.

2 Travaux connexes

La méthode utilisée ici pour extraire des contraintes monotones est fortement inspirée de l'acquisition de connaissances d'adaptation en raisonnement à partir de cas. Ces méthodes permettent d'obtenir à partir d'une base de cas (une base d'exemples) des connaissances liant les variations sur les descripteurs de problèmes aux variations sur les descripteurs de solutions. Pour y parvenir, plusieurs méthodes ont été proposées pour apprendre les connaissances d'adaptation, que ce soit sous la forme de règles d'Aquin et al. (2022) ou par le biais d'approches d'apprentissage automatique (par exemple, comme dans d'Aquin et al. (2022); Ye et al. (2021), via des réseaux neuronaux).

Ces méthodes d'extraction de règles d'adaptation se sont révélées efficaces mais s'appliquent généralement dans des scénarios où il existe une distinction claire entre un espace de problèmes et un espace de solutions, et où les deux sont clairement décrits en termes d'un ensemble fixe de variables bien définies. Dans d'Aquin et al. (2006) par exemple, elle a été appliquée à l'oncologie, où les problèmes correspondent aux descriptions des patients et la solution au traitement applicable.

D'autres travaux, tels que Kotłowski et Słowiński (2009), ont exploré l'apprentissage de règles avec des contraintes monotones, une approche dans laquelle on suppose que le label de classe doit augmenter avec les valeurs croissantes des attributs. L'objectif principal de cette

méthode est d'exploiter les contraintes monotones pour améliorer l'apprentissage des règles en classification. Le processus débute par la monotonisation des données à l'aide d'une procédure de classification non paramétrique, afin d'adapter les données aux contraintes monotones, et vise à générer un ensemble de règles qui soit cohérent avec l'ensemble d'apprentissage.

Nos travaux s'appuient sur la disponibilité d'un GC pour les matériaux étudiés, les méthodes d'estimation et la conductivité thermique de ces matériaux. Les GC Hogan et al. (2021) sont des représentations d'informations à base de graphes où le contenu du graphe reçoit une signification bien définie en utilisant une ontologie Staab et Studer (2013). L'ontologie fournit plus qu'un schéma pour le graphe, car elle identifie clairement les concepts et leur donne des définitions logiques, harmonisant ainsi sémantiquement la compréhension de ces concepts à travers les différentes sources de données. En intégrant toutes les sources dans un graphe commun, en suivant une ontologie commune, des connexions peuvent être établies entre les sources, ce qui permet d'interroger, de raisonner et d'explorer l'ensemble de l'espace de données Adamou et d'Aquin (2020).

Notre objectif ici est de montrer comment ces connexions peuvent être exploitées dans un processus d'apprentissage combinant les deux idées : l'intégration sémantique d'informations sous un GC et l'exploration de contraintes monotones. Nous fournissons brièvement ci-dessous un aperçu général du GC et de l'ontologie pour notre étude de cas, avant de décrire plus en détail l'approche proposée telle qu'illustrée dans cette étude de cas.

3 Vers un graphe de connaissances sur la conductivité thermique

Nous résumons ici brièvement la construction d'un GC qui relie différentes sources d'informations sur les matériaux et leur conductivité thermique estimée. En guise d'aperçu, nous avons collecté dans la littérature sur la science des matériaux un certain nombre d'ensembles de données contenant des estimations des valeurs de conductivité thermique des matériaux, construit une petite ontologie pour représenter les liens entre les matériaux, les méthodes, les valeurs et les sources, et les mappages entre les ensembles de données collectés et l'ontologie. Nous avons également créé des scripts pour relier les matériaux identifiés de différentes manières, dans différents ensembles de données, les uns aux autres.

3.1 Ensembles de données

La conductivité thermique est un phénomène complexe en physique, et il n'entre pas dans le cadre de cet article d'entrer dans les détails. Cependant, cette complexité explique pourquoi, depuis plusieurs décennies, de nombreuses recherches ont été menées sur des modèles mathématiques et informatiques pour le calculer ou l'approximer, qui varient considérablement en termes de précision et d'exigences de calcul. Par exemple, certaines méthodes sont basées sur des simulations à l'échelle microscopique (reposant généralement sur la théorie fonctionnelle de la densité, DFT), qui peuvent être appliquées à différentes résolutions et volumes. D'autres méthodes, telles que l'AGL Toher et al. (2014), s'appuient sur des modèles mathématiques plus simples utilisant des descripteurs ou des matériaux relativement faciles à obtenir. En outre, il

existe de nombreuses variantes de chacune de ces grandes catégories de méthodes et différents niveaux d'approximation qui peuvent être pris en compte (par exemple, la valeur totale de la conductivité thermique ou la conductivité thermique minimale du réseau). Enfin, la conductivité thermique peut également être estimée expérimentalement pour tout matériau observable, c'est-à-dire existant dans la nature ou en cours de synthèse.

En examinant la littérature sur la conductivité thermique et les modèles d'apprentissage automatique pour prédire la conductivité thermique, nous avons identifié plusieurs ensembles de données contenant des estimations de la conductivité thermique par diverses méthodes. Cependant, une autre difficulté est que les matériaux ne sont pas facilement identifiés sans ambiguïté. En fait, de nombreux ensembles de données ne fournissent que la formule du matériau, par exemple Ag_1Br_1. Cependant, il existe plusieurs matériaux qui peuvent correspondre à cette formule, même si elle est relativement simple. La différence entre ceux-ci est la structure dans laquelle les atomes qui les composent sont organisés, et cette structure est un facteur important dans la conductivité thermique du matériau. Par conséquent, nous n'incluons pour le moment dans les graphes de connaissances que des informations provenant de sources qui incluent un moyen sans ambiguïté et précis d'identifier les matériaux correspondants. Le tableau 1 résume les ensembles de données inclus, y compris un résumé de la méthode d'estimation de la conductivité thermique (TC) utilisée dans cet ensemble de données.

ID du jeu de données	**Description**	**Source**	**Méthode d'estimation TC**	**Taille**
miyazaki_2021_hh	Composés Half-Heusler	Miyazaki et al. (2021)	DFT direct (méthode 1)	143
aflow_agl	Depuis AFLOWLib	Toher et al. (2014)	AGL (méthode 2)	4937
te_designlab_spreadsheet	Depuis le site web TEDesignLab	Gorai et al. (2016)	DFT+modèle semi-empirique ad-hoc (méthode 3)	2701
te_designlab_json	Depuis la sortie JSON de TEDesignLab	Gorai et al. (2016)	DFT+modèle semi-empirique ad-hoc (méthode 4)	2292
minltc_clarke	Méthode de Clarke, enregistrée dans la base de données Material Project	De Jong et al. (2015)	minLTC depuis DFT (méthode 5)	6285

TAB. 1 – *Résumé des jeux de données inclus. La taille est en nombre de matériaux.*

3.2 Ontologie

Afin de représenter les informations sur les matériaux et leur estimation de conductivité thermique, nous avons conçu une petite ontologie. Un aperçu de sa structure de base est présenté dans la Figure 1.

La classe principale de cette ontologie est celle de `Material`. Un matériau est une structure atomique clairement identifiée, qui peut être liée à ses éléments chimiques constitutifs (les types d'atomes inclus dans sa structure) et à son groupe spatial, qui est une façon de décrire le type de structure et les symétries qui existent au sein de sa structure. De plus, un certain nombre de propriétés de données sont associées aux matériaux, y compris les chaînes représentant leur formule, ou d'autres caractéristiques numériques qui peuvent être disponibles dans les jeux de données (volume, densité, etc.).

Le cœur de l'ontologie est cependant l'association entre un matériau et une (éventuellement plusieurs) estimation(s) de sa conductivité thermique. Chaque estimation est représentée par la classe `TCApproximation`. Une telle approximation, associée à un matériau donné, doit au minimum spécifier la méthode d'approximation et une valeur obtenue. Une petite hiérarchie des types de méthodes d'approximation est représentée dans la Figure 1 (cette hiérarchie peut être beaucoup plus détaillée en fonction des jeux de données inclus). De plus, certains paramètres de la méthode d'approximation peuvent également être inclus, le plus courant étant la température, puisque la conductivité thermique varie en fonction de la température de l'environnement dans lequel elle est mesurée. Enfin, une propriété de données est disponible pour lier l'approximation à sa source, i.e. l'article ou le jeu de données d'où elle a été extraite, et qui peut donner plus de détails sur son calcul.

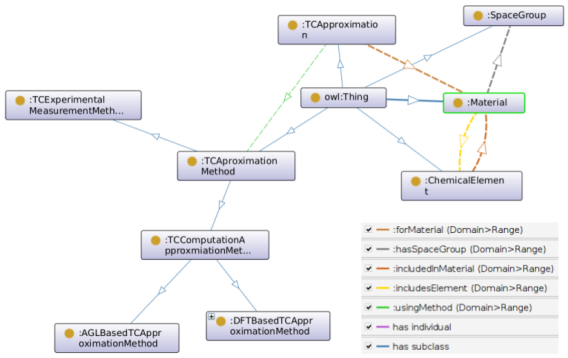


FIG. 1 – *Aperçu de l'ontologie (basé sur la visualisation OntoGraf dans Protégé).*

Bien que cette ontologie ad hoc soit très simple, une discussion peut être utile ici sur le choix fait pour représenter le fait que différentes sources ont des valeurs différentes pour une même caractéristique et l'incertitude que cela représente. En effet, nous choisissons ici de représenter cela en indiquant que le matériau était l'entrée d'un processus d'estimation de la conductivité thermique, et que ce processus a conduit à certains résultats. En faisant ce choix, nous avons considéré d'autres options. Par exemple, chacun des jeux de données intégrés dans le GC est inclus dans le triplestore utilisé (graphDB) sous forme de RDF séparé nommé graph [1]. Cela nous permet d'inclure des métadonnées sur le graphe et sa provenance. Nous aurions donc pu inclure les informations sur la méthode de génération de l'approximation de la conductivité thermique dans l'ensemble du graphe une seule fois, au niveau des métadonnées du graphe, au lieu de les inclure pour chaque approximation et, par conséquent, de simplement relier les matériaux à leur valeur approximée de conductivité thermique. Cependant, nous aurions également pu établir un tel lien direct et ajouter des informations sur la méthode utilisée pour produire la valeur approximative, en utilisant des instructions RDF-star [2] sur chacun des triplets de conductivité thermique. Cela aurait conduit, par exemple, à des triplets dans RDF-star de la forme :

```
<< <material> tco:thermalConductivity value >> tco:producedBy <method>
```

1. https://www.w3.org/TR/rdf11-concepts/##dfn-named-graph
2. https://w3c.github.io/rdf-star/cg-spec/editors_draft.html

Cependant, nous privilégions ici une approche qui relie indirectement la valeur estimée de la conductivité thermique, qui a peu de signification en elle-même, aux caractéristiques du matériau. En d'autres termes, nous disons seulement qu'une méthode a estimé une valeur, et non que la valeur de la conductivité thermique pour ce matériau est celle indiquée. De plus, l'approche proposée correspond à des modèles de conception d'ontologie générale qui peuvent être réutilisés de manière cohérente pour d'autres propriétés du matériau à intégrer dans le GC à l'avenir.

3.3 Instancier le graphe de connaissances

Pour instancier le GC, chacun des ensembles de données répertoriés dans le Tableau 1 est d'abord transformé, si nécessaire, en un format tabulaire. Un mappage entre chaque tableau et une représentation RDF suivant l'ontologie décrite ci-dessus est ensuite créé à l'aide de l'outil OntoRefine [3] associé à graphDB. Dans ces mappages, un nouvel individu correspondant à la méthode d'approximation est créé, qui instancie la classe pertinente dans la hiérarchie de l'ontologie, et la structure correspondante est créée pour chaque matériau. Chaque graphe RDF généré à partir de ces mappages est stocké dans un graphe nommé et les métadonnées pertinentes sont ajoutées.

De plus, comme mentionné ci-dessus, l'identification et la liaison des matériaux peuvent être compliquées. À ce stade, nous n'avons inclus que les ensembles de données pour lesquels des identifiants clairs sont disponibles. Ces identifiants peuvent provenir soit de la base de données Material Project [4], soit de l'ICSD [5] (la base de données sur les structures cristallines inorganiques). Nous créons donc des URI de matériaux en utilisant ces identifiants lorsqu'ils sont disponibles. Nous utilisons ensuite l'API de la base de données Material Project pour récupérer les identifiants ICSD pour tous les identifiants Material Project lorsqu'ils sont disponibles et générer les liens `owl:sameAs` correspondants.

Pour illustrer l'ampleur des écarts entre les différentes sources d'approximations de conductivité thermique dans les ensembles de données sélectionnés, la Figure 2 montre l'écart sur les matériaux superposés à partir de paires d'ensembles de données. Cela correspond à l'erreur absolue moyenne en pourcentage lors de l'utilisation de l'ensemble de données de la colonne en remplacement de celui de la ligne.

4 La découverte de contraintes monotones

Dans cette section, nous décrivons notre approche pour apprendre des contraintes monotones à partir des cas représentés dans le GC. Cette approche est capable de générer une règle liant les variations d'attributs de matériaux et les variations de conductivité thermique selon les différentes méthodes d'approximation. Ces contraintes sont extraites du GC et peuvent être liées à un concept, une propriété de donnée ou une propriété d'objet susceptible d'influencer la conclusion de la règle. Dans notre cas, la conclusion de la règle correspond à la valeur obtenue de la conductivité thermique "`TCApproximation`". La variation de cette valeur dépend de la méthode de mesure et de la variation des propriétés physiques du matériau mesuré :

3. https://www.ontotext.com/products/ontotext-refine/
4. https://materialsproject.org/
5. https://www.psds.ac.uk/icsd

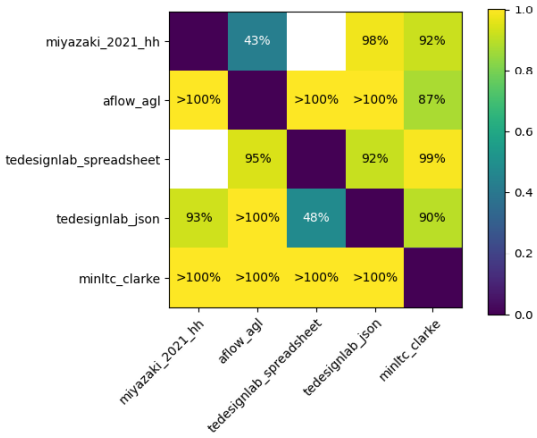


FIG. 2 – *Écart des approximations de conductivité thermique dans les matériaux qui se chevauchent dans des paires d'ensembles de données.*

— "Bulk Modulus" [6] : mesure de la résistance d'une substance à la compression sous pression. Il quantifie le rapport entre la variation de pression et la variation fractionnelle de volume par unité de volume.
— "Debye Temperature" [7] : une température caractéristique d'un matériau solide qui décrit ses vibrations atomiques. Elle représente la température à laquelle les vibrations du réseau cristallin dans le matériau sont excitées au maximum.
— Densité [8] : mesure de la masse par unité de volume d'une substance. Elle quantifie la quantité de masse contenue dans un volume donné.
— Nombre d'Atomes : correspond au nombre total d'atomes présents dans la maille élémentaire du matériau.
— "Shear Modulus" [9] : mesure de la résistance d'une substance à la déformation par cisaillement. Il quantifie le rapport entre la contrainte de cisaillement et la déformation de cisaillement dans un matériau.
— Volume : correspond à la quantité d'espace occupée par une cellule du matériau.

Chaque contrainte monotone, qui est liée à une de ces propriétés physiques, peut prendre l'une des valeurs suivantes : 1 si c'est une contrainte croissante (sa valeur et la valeur de `TCApproximation` évoluent dans le même sens), -1 si c'est une contrainte décroissante (sa valeur et la valeur de `TCApproximation` évoluent dans des sens opposés), ou 0 si elle n'est pas liée de façon monotone à la valeur de la conductivité thermique. Comme la dépendance entre le changement des valeurs de propriétés physiques et la valeur de `TCApproximation` dépend de la méthode de mesure, chaque règle est associée à une de ces méthodes. Ainsi, notre objectif est d'acquérir des règles de monotonicité avec des contraintes sur chaque propriété physique. Chaque règle est liée à une des 5 méthodes de mesure associées aux jeux de données

6. https://www.britannica.com/science/bulk-modulus
7. https://www.sciencedirect.com/topics/chemistry/debye-temperature
8. https://www.britannica.com/science/density
9. https://www.britannica.com/science/shear-modulus

	Cas 1	Cas 2	Les contraintes de monotone
Materiau	mp-16314	mp-1018135	
La méthode de mesure	1	1	
“Bulk Modulus”	103,20	46,59	-1
“Debye Temperature”	241,60	288,68	1
Densité	11,41	5,04	-1
Nombre d’Atomes	/	/	0
“Shear Modulus”	49,59	29,67	-1
Volume	70,69	63,95	-1
`TCApproximation`	5,86	16,49	

TAB. 2 – *L’extraction des contraintes de monotone à partir de cas.*

représentés dans la Figure 2 comme suit :

— méthode 1 appliquée dans miyazaki_2021_hh
— méthode 2 appliquée dans aflow_agl
— méthode 3 appliquée dans tedesignlab_spreadsheet
— méthode 4 appliquée dans tedesignlab_json
— méthode 5 appliquée dans minltc_clarke

Notre approche comporte trois éléments majeurs qui sont étroitement liés : (1) la génération/l’enrichissement de la base de cas, (2) la génération de règles de monotonie et (3) l’application des règles.

L’étape de *génération/enrichissement de la base de cas* consiste à générer la base de cas ou à la mettre à jour avec de nouveaux cas. Chaque cas est un matériau, ses caractéristiques et sa conductivité thermique, associé à une des 5 méthodes de mesure, comme indiqué dans le Tableau 2. Ce tableau montre que la valeur obtenue de la `TCApproximation` varie dans le même sens que la variation de “debye Temperature”, et donc elle est considérée comme une contrainte croissante. Dans le cas de “bulk modulus”, densité, “shear modulus” et volume, la valeur de `TCApproximation` varie dans un sens opposé à celle de ces propriétés et donc elles sont des contraintes décroissantes. Le nombre d’atomes n’est pas mentionné dans ces deux cas, et donc aucune contrainte n’est associée à cette propriété.

Le processus de *génération de règles de monotonicité* consiste à analyser la base de cas générée. Pour chaque cas, nous effectuons une comparaison par paire des deux valeurs possibles pour chaque propriété physique. Selon la comparaison entre la variation de chaque propriété entre les cas et la variation de la valeur de `TCApproximation`, une contrainte monotone sera associée à cette propriété (1, -1 ou 0). En outre, un facteur de certitude, qui représente la confiance accordée à la règle, est attribué à chaque règle en fonction de sa fréquence de prédictions correctes. Ainsi, plus l’association entre le segment antécédent et le segment conséquent de la règle est fréquente, plus sa confiance est élevée.

L’*application des règles* consiste à intégrer les contraintes monotones dans un modèle prédictif, c’est-à-dire que les règles découvertes sont incorporées comme contraintes dans le modèle XGBoost [10] dans cette étude. Cela permet d’aligner les prédictions avec les relations phy-

10. “Extreme Gradient Boosting” : `https://xgboost.readthedocs.io/en/latest/tutorials/model.html`

siques sous-jacentes (les relations entre les propriétés physiques et la valeur de conductivité thermique). Le choix du modèle XGBoost a été motivé par sa capacité à traiter efficacement nos données complexes comportant de nombreuses propriétés, ainsi que les relations complexes entre ces dernières, tout en assurant une grande robustesse face au bruit et aux valeurs manquantes.

5 Expérimentations

Le GC utilisé contient 901 524 triplets et décrit 23 251 matériaux ainsi que leur conductivité thermique. Chaque valeur de conductivité thermique (TC) a été estimée par une méthode spécifique et est associée à une seule valeur. Les conductivités thermiques dans ce GC ont été évaluées à l'aide des 5 méthodes décrites plus haut.

L'objectif de l'expérimentation est d'apprendre des règles de monotonicité à partir d'un sous-ensemble des valeurs de TC des matériaux, formant un ensemble d'apprentissage. Ensuite, évaluer la qualité des prédictions des TC sur les matériaux restants dans l'ensemble de test.

Pour évaluer notre système, nous avons divisé les données du GC en utilisant la méthode 20/80. Ainsi, 80% des matériaux ont été utilisés durant la phase d'apprentissage pour générer les règles de monotonicité, tandis que les 20% restants ont servi à tester ces règles. Ainsi, nous avons évalué la valeur prédite de la conductivité thermique pour chaque matériau du jeu de test en calculant l'erreur quadratique moyenne (MSE).

L'application de notre approche sur ce GC a permis de découvrir 114 règles de monotonicité avec une confiance supérieure à $0,6$. Le Tableau 3 présente la répartition des règles extraites selon les plages de confiance. Nous remarquons que notre approche a trouvé 83 règles avec une confiance d'au moins $0,9$ et 99 règles avec une confiance d'au moins $0,7$.

Intervalle de confiance	$\in [0.6, 0.7[$	$\in [0.7, 0.8[$	$\in [0.8, 0.9[$	$\in [0.9, 1]$
# Règles	15	16	0	83

TAB. 3 – *Répartition des règles selon les intervalles de confiance.*

Pour trouver l'impact des règles générées, nous avons entraîné le modèle XGBoost sur l'ensemble de données d'apprentissage avec et sans les contraintes de monotonicité. Ensuite, nous avons évalué ses performances sur l'ensemble de données de test en calculant la MSE. Le Tableau 4 présente les résultats obtenus pour chaque méthode. L'ajout de contraintes monotones améliore les performances du modèle, en réduisant la MSE. Cette réduction est significative pour certaines méthodes comme la méthode 5, la MSE est réduite de 33,45 à **16,83**. Ce tableau montre aussi les contraintes utilisées pour chaque méthode, les attributs sont associés dans l'ordre à ["Bulk Modulus", "Debye Temperature", Densité, Nombre d'Atomes, "Shear Modulus", Volume]. Ces contraintes proviennent de la règle de monotonicité appliquée qui est évaluée selon sa confiance. Les règles avec une confiance élevées tendent à obtenir une faible valeur MSE. Ces résultats montrent l'impact de l'utilisation des règles sur la performance globale du modèle.

Nous avons aussi comparé notre approche avec une baseline qui applique l'apprentissage statistique basé sur les cas. Cette baseline calcule les corrélations monotones en utilisant le

	MSE du modèle sans contraintes	**MSE du modèle avec contraintes**	**Contraintes monotones**	**Confiance de la règle de monotonicité**
Méthode 1	28,34	**13,15**	[0, 1, 1, 0, 1, 1]	99,67%
Méthode 2	26,27	**15,68**	[1, 0, 0, 0, 0, 0]	91,31%
Méthode 3	231,46	**195,43**	[0, 0, 0, 0, 0, 1]	99,67%
Méthode 4	5,61	**4,28**	[1, 0, 0, 0, 0, 1]	99,67%
Méthode 5	33,45	**16,83**	[0, 0, 0, 0, 0, -1]	69,91%

TAB. 4 – *Comparaison entre le modèle avec et sans contraintes monotones.*

	MSE du modèle sans contraintes	**MSE du modèle avec les contraintes des règles**	**MSE du modèle avec les contraintes de la baseline**
Méthode 1	28,34	**13,15**	27,34
Méthode 2	26,27	**15,68**	15,70
Méthode 3	231,46	**195,43**	481,01
Méthode 4	5,61	**4,28**	8,67
Méthode 5	33,45	**16,83**	17,68

TAB. 5 – *Comparaison entre le modèle avec les contraintes des règles, avec les contraintes de la baseline et sans contraintes.*

coefficient de Spearman ρ pour chaque propriété en mesurant la relation monotone entre la propriété et la valeur de `TCApproximation`, afin de déterminer la direction des contraintes monotones. L'inférence des règles monotones suit la méthode suivante :

— $\rho > 0$: monotonicité croissante.
— $\rho < 0$: monotonicité décroissante.
— $\rho = 0$: aucune contrainte.

Le Tableau 5 compare les performances de notre approche basée sur la génération de règles avec la baseline, ainsi qu'avec le modèle sans contraintes monotones. Les résultats montrent que notre approche réduit significativement la MSE pour toutes les méthodes, ce qui indique une meilleure précision dans la prédiction des valeurs de conductivité thermique.

En particulier, pour la Méthode 1, le modèle utilisant les contraintes dérivées des règles atteint un MSE de **13,15**, surpassant à la fois la baseline (27,34) et le modèle sans contraintes (28,34). Une amélioration similaire est observée pour les autres méthodes. Par exemple, pour la Méthode 4, notre approche réduit le MSE de 5,61 (sans contraintes) à **4,28**, tandis que la baseline obtient un MSE bien plus élevé de 8,67. Ces résultats mettent en évidence l'efficacité de notre approche dans l'intégration des contraintes monotones pour améliorer les performances du modèle par rapport à la baseline et au modèle sans contraintes.

La Figure 3 illustre la relation entre la confiance des règles de monotonicité et la MSE associée aux prédictions effectuées. Cette figure représente la distribution de 30 règles générées pour la méthode 4. Chaque point représente une règle de monotonicité appliquée, avec sa confiance et la MSE correspondante. On observe que les règles avec une confiance élevée (entre 0,6 et 1) tendent à produire des MSE plus faibles, avec une MSE minimale de 4,29. À l'inverse, les règles avec une faible confiance (inférieure à 0,5) présentent des MSE plus élevées. Ces observations confirment que les règles ayant une confiance plus élevée sont généralement plus fiables pour prédire avec précision la conductivité thermique, tandis que celles ayant une faible

confiance sont moins précises et génèrent plus d'erreurs.

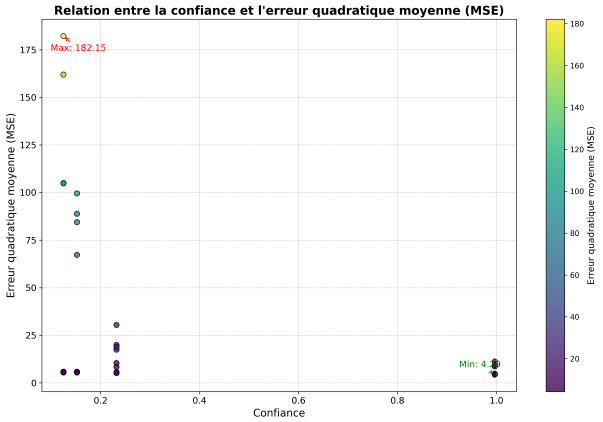


FIG. 3 – *La relation entre la confiance des règles de monotonicité sur la méthode 4 et l'erreur quadratique moyenne (MSE) associée aux prédictions effectuées par XGBoost.*

6 Conclusion

Dans cet article, nous avons montré comment une méthode s'appuyant sur l'extraction de règles d'adaptation inspirée du raisonnement à partir de cas peut permettre d'identifier des contraintes monotones dans les données, permettant elle-même d'améliorer les capacités de prédiction d'un modèle d'apprentissage. Nous avons illustré ces approches sur un cas d'études s'appuyant sur des données en science des matériaux issues d'un graphe de connaissances, montrant que les contraintes monotones permettent, dans ce cas, de réduire l'erreur de prédiction obtenue, et cela de façon plus efficace qu'une méthode naïve d'extraction de contraintes monotones. L'application de cette même méthode à d'autres cas pourrait permettre de confirmer que l'exploitation de contraintes monotones automatiquement extraites des données peut fournir des connaissances utiles à guider l'apprentissage d'un modèle de prédiction.

Pour aller plus loin, en plus d'étendre l'étude de cas pour découvrir des connaissances d'adaptation reliant différentes propriétés physiques des matériaux, plusieurs pistes peuvent être explorées pour améliorer l'approche d'extraction de règles, notamment des solutions visant à la rendre plus évolutive, ainsi que des heuristiques pour orienter le choix des propriétés à inclure dans les cas.

Références

Adamou, A. et M. d'Aquin (2020). Linked data principles for data lakes. In *Data Lakes*, pp. 145–169. Wiley].

d'Aquin, M., F. Badra, S. Lafrogne, J. Lieber, A. Napoli, et L. Szathmary (2007). Case base mining for adaptation knowledge acquisition. In *Twentieth International Joint Conference on Artificial Intelligence-IJCAI'07*, pp. 750–755.

d'Aquin, M., J. Lieber, et A. Napoli (2006). Adaptation knowledge acquisition : A case study for case-based decision support in oncology. *Computational intelligence 22*(3-4), 161–176.

De Jong, M., W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. Van Der Zwaag, J. J. Plata, et al. (2015). Charting the complete elastic properties of inorganic crystalline compounds. *Scientific data 2*(1), 1–13.

d'Aquin, M., E. Nauer, et J. Lieber (2022). A factorial study of neural network learning from differences for regression. In *International Conference on Case-Based Reasoning*, pp. 289–303. Springer.

Gorai, P., D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanović, et E. S. Toberer (2016). Te design lab : A virtual laboratory for thermoelectric material design. *Computational Materials Science 112*, 368–376.

Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. (2021). Knowledge graphs. *ACM Computing Surveys (Csur) 54*(4), 1–37.

Kotłowski, W. et R. Słowiński (2009). Rule learning with monotonicity constraints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 537–544.

Miyazaki, H., T. Tamura, M. Mikami, K. Watanabe, N. Ide, O. M. Ozkendir, et Y. Nishino (2021). Machine learning based prediction of lattice thermal conductivity for half-heusler compounds using atomic information. *Scientific reports 11*(1), 13410.

Staab, S. et R. Studer (2013). *Handbook on ontologies*. Springer Science & Business Media.

Toher, C., J. J. Plata, O. Levy, M. De Jong, M. Asta, M. B. Nardelli, et S. Curtarolo (2014). High-throughput computational screening of thermal conductivity, debye temperature, and grüneisen parameter using a quasiharmonic debye model. *Physical Review B 90*(17), 174107.

Ye, X., D. Leake, V. Jalali, et D. J. Crandall (2021). Learning adaptations for case-based classification : A neural network approach. In *Case-Based Reasoning Research and Development : 29th International Conference, ICCBR 2021, Salamanca, Spain, September 13–16, 2021, Proceedings 29*, pp. 279–293. Springer.

Summary

This study aims to discover monotonic constraints linking the physical properties of materials to improve the prediction and analysis of the physical properties of materials, in particular their thermal conductivity. Monotonicity rules are extracted by analyzing the relationships between physical properties and the methods used for their measurement, thus identifying constraints of increasing or decreasing monotonicity between these properties. This approach has made it possible to identify predictive rules that explain how variations in the physical properties of materials depend on changes in the estimation method. By providing a method to explore and analyze the relationships between the physical properties of materials, we were able to generate accurate monotonic rules that predict the monotonic change in these properties.

Sélection de données par leurs difficulté pour réduire la quantité nécessaire à la recherche d'hyperparamètres*

Gustavo Rodrigues dos Reis*,**, Mario Cortes Cornax*
Adrian Mos** Cyril Labbé*

*Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG,
700 Av. Centrale, 38401 Saint-Martin-d'Hères, France
{mario.cortes-cornax,cyril.labbe}@univ-grenoble-alpes.fr ,
**NAVER LABS Europe,
6 Chemin de Maupertuis, 38240 Meylan, France
{gustavo.rodrigues-dosreis,adrian.mos}@naverlabs.com

Résumé. Dans ce travail, nous décrivons une méthode de recherche d'hyperparamètres (*Hyperparameter Optimisation* ou *HPO*) qui utilise moins de données en mesurant la difficulté d'adaptation du modèle à un échantillon donné. Une partie réduite des données est sélectionnée en respectant la proportion des difficultés dans l'ensemble du jeu de données. Ce raisonnement est inspiré par des résultats de recherche sur l'apprentissage machine par *curriculum learning*. Une évaluation de l'approche est évaluée sur les tâches de reconnaissance d'images et de reconnaissance d'entités nommées (*Named Entity Recognition–NER*). Les expériences montrent que la quantité de données nécessaire pour l'HPO peut être réduite jusqu'à 60 % pour obtenir le même choix d'hyperparamètres par rapport à l'utilisation de la totalité des données disponibles.

1 Introduction

Les tendances actuelles en matière d'apprentissage profond conduisent à des modèles plus performants en augmentant leur taille (nombre de paramètres). Cela augmente de plusieurs ordres de grandeur la quantité de données nécessaire à l'apprentissage. Des concepteurs de système disposant de peu de ressources doivent adapter de ces modèles à leurs besoins. La recherche d'hyperparamètres est une tâche importante que ces concepteurs doivent effectuer à un moment donné. Chaque nouveau processus de HPO augmente les coûts et l'impact environnemental de l'apprentissage (Rafat et al., 2023). Ces processus HPO doivent donc être aussi légers que possible.

Les recherches sur le HPO (He et al., 2021) (Neutatz et al., 2022) peuvent conduire à réduire la quantité de données nécessaire Adadi (2021). L'objectif de l'HPO est de trouver une configuration optimale pour l'adaptation postérieure du modèle. La pratique courante consiste à utiliser toutes les données d'entraînement disponibles pour cette recherche. Nous nous posons la question suivante *La recherche d'hyperparamètres peut-elle être moins coûteuse grâce à l'utilisation de la réduction des données en s'inspirant de l'apprentissage par curriculum ?*

Nous abordons cette question en quantifiant la *difficulté* avec laquelle un modèle produit le résultat attendu pour un échantillon donné. Pour chaque échantillon, la sortie d'un état du

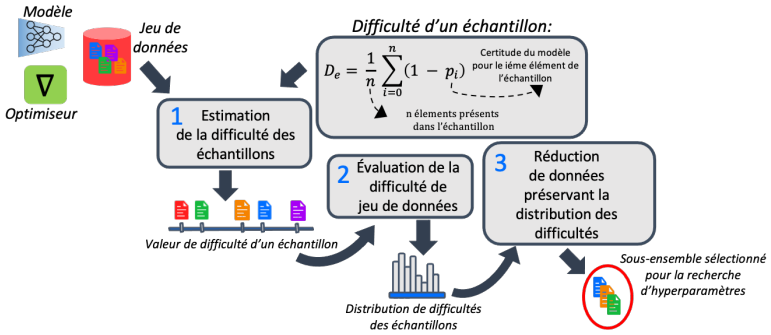


FIG. 1 – *Les différentes étapes de la méthode sont illustrées avec les éléments d'entrée et de sortie, ainsi que la notion de difficulté d'un échantillon utilisée pour la sélection.*

modèle (celui des premières étapes de l'optimisation), est utilisée comme indicateur de la difficulté à produire le résultat correct.

La méthode que nous proposons est motivée par les résultats de Schwartz et al. (2020), sur l'importance que chaque échantillon possède dans le processus d'optimisation d'un modèle, et de Soviany et al. (2022) sur l'apprentissage guidée par curriculum. Le point principal de l'apprentissage par curriculum est d'organiser les données dans un ordre significatif, en classant les éléments de faciles à difficiles. L'importance des échantillons est traduite par leur impact sur la vitesse de changement des paramètres du modèle.

L'hypothèse que nous évaluons dans ce travail est de savoir si les sorties d'un modèle peuvent servir à ordonner les échantillons d'un jeu de données, en créant une distribution de l'importance des données en fonction de la difficulté d'adaptation du modèle à ces échantillons. Nous décrivons une nouvelle méthode de sélection des données pour l'optimisation d'hyperparamètres qui réduit le nombre d'échantillons utilisé tout en préservant les proportions originales de difficulté de l'ensemble total d'échantillons. Nous fournissons une évaluation préliminaire pour l'adaptation des modèles basés sur des architectures de type *transformers*.

2 Méthode pour réduire la quantité de données pour la recherche d'hyperparamètres

L'objectif est d'obtenir le bon choix d'hyperparamètres avec une réduction du nombre d'échantillons utilisés. La méthode se compose de trois étapes, comme le montre la Figure 1.

Dans une première phase, tous les échantillons sont nécessaires pour estimer la difficulté de chacun d'eux. La difficulté d'un échantillon est liée à la confiance que le modèle donne au résultat attendu (voir formule figure 1). Une fois cette évaluation achevée, le jeux de données est séparé en sous-ensembles ayant des éléments de difficulté similaires. On obtient ainsi une distribution de difficulté sur l'ensemble total du jeu de données. La dernière étape consiste à utiliser cette distribution pour sélectionner un ensemble réduit mais représentatif des données ainsi qu'un ordre pour une approche par curriculum.

Nous évaluons cette méthodologie de sélection de données pour le processus de HPO en adaptant deux modèles pré-entraînés de type *transformers* (sciBERT et ViT). Ce choix est motivé par le fait que l'architecture transformers est dominante dans les grands *fondation models*

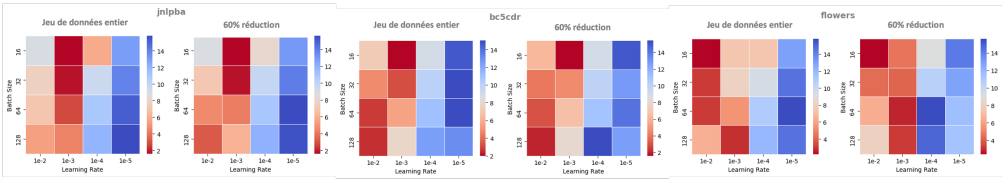


FIG. 2 – *Recherche d'hyperparamètres pour des jeux de données textuelles (jnlpba, bc5cdr) et d'images (flowers). La configuration d'hyperparamètres en rouge foncées (F1-score élevée) sont identiques pour la totalité des données et la réduction à l'aide de la difficulté.*

qui sont très coûteux. Ceci nous permet de tester l'approche pour deux tâches différentes qui nécessitent une optimisation des modèles à chaque identification de nouveaux concepts.

Les ensembles de données que nous utilisons ici ont été sélectionnés parmi ceux présentés dans les travaux (Beltagy et al., 2019) et (Dosovitskiy et al., 2021). Les tâches ciblées sont respectivement l'identification de concepts dans des images et l'identification de termes scientifique dans des textes (NER).

Nous illustrons en Figure 1 la définition de la difficulté des échantillons pour les tâches que nous évaluons ainsi que son utilisation pour caractériser, selon différents niveaux de difficulté, les éléments du jeu de données. Cette distribution est un élément clé de notre méthode, car elle définie la sélection et l'ordre d'utilisation des sous-ensembles d'échantillons.

Pour chacune des tâches et chacun des jeux de données, nous avons testé les résultats de l'exécution du HPO avec l'ensemble total du jeu de données afin de fournir une référence. Ensuite, nous procédons avec un pourcentage réduit d'échantillons sélectionnées au hasard et terminons par la recherche de paramètres avec le même pourcentage mais en respectant les proportions de difficulté produites et une approche par curriculum.

3 Évaluation de la méthode pour des tâches de reconnaissance d'images et de termes scientifiques

La figure 2 montre les résultats des classements des meilleurs hyperparamètres obtenus à partir des exécutions de processus de HPO en utilisant différents jeux de données. En utilisant l'optimiseur Adam, nous recherchons les hyperparamètres suivants : taux d'apprentissage et de la taille du *batch*. Ces hyperparamètres sont directement liés à la consommation de ressources lors de l'adaptation du modèle.

Les jeux de données utilisés sont : la reconnaissance de **termes scientifiques** sur les deux corpus présentés dans Collier et al. (2004); Li et al. (2016) et la **reconnaissance d'images** avec le jeu de donnée présenté par Nilsback et Zisserman (2008). Ce dernier jeux de données comporte moins d'échantillons par classe que les jeux de données textuelles.

Parmi tous les choix possibles d'hyperparamètres testés, ceux en rouges présentent des scores F1 systématiquement plus élevés.

Avec cette méthode, nous avons démontré, à travers deux cas d'usage, qu'il est possible de réduire significativement la quantité de données tout en obtenant un choix d'hyperparamètres similaire. Ces résultats sont prometteurs. D'autres cas d'usage doivent être explorés. La mesure de difficulté peut être définie en fonction d'autres paramètres, tels que la quantité de données disponible, le modèle utilisé ou la nature de la tâche.

4 Conclusion

Ce travail propose une méthode de sélection des données pour réduire la quantité des données utilisées dans la recherche des hyperparamètres. La sélection est réalisée à l'aide d'une *difficulté* mesurée pour chaque échantillon. Il est possible de remplacer la totalité des données d'apprentissage par un sous-ensemble de données qui ressemble suffisamment à la totalité en préservant la distribution des niveaux de difficulté. Lors de l'apprentissage, les données sont présentées au modèle de la plus facile à la plus difficile. Nous prévoyons de poursuivre ce travail en étudiant d'autres mesures de difficulté ainsi que d'autres stratégies pour ordonner les données.

Références

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *J Big Data 8*, 24.

Beltagy, I., K. Lo, et A. Cohan (2019). SciBERT: A Pretrained Language Model for Scientific Text. In R. H. Sebastian Padó (Ed.), *Proc. of the 2019 Conference EMNLP-IJCNLP*. ACL.

Collier, N., T. Ohta, et al. (2004). Introduction to the Bio-entity Recognition Task at JNLPBA. In N. Collier, P. Ruch, et A. Nazarenko (Eds.), *Proc. of the Int. Joint Workshop NLPBA/BioNLP*. COLING.

Dosovitskiy, A. et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *12th Int. Conf. on Learning Representations (ICLR)*. OpenReview.

He, X., K. Zhao, et X. Chu (2021). AutoML: A Survey of the State-of-the-Art. *Knowledge-Based Systems 212*, 106622.

Li, J., Y. Sun, et al. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*.

Neutatz, F., B. Chen, Y. Alkhatib, J. Ye, et Z. Abedjan (2022). Data Cleaning and AutoML: Would an Optimizer Choose to Clean? *Datenbank Spektrum 2*, 121–130.

Nilsback, M.-E. et A. Zisserman (2008). Automated flower classification over a large number of classes. In A. Majumdar et S. K. Mitra (Eds.), *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729. IEEE.

Rafat, K., S. Islam, A. A. Mahfug, M. I. Hossain, F. Rahman, S. Momen, S. Rahman, et N. Mohammed (2023). Mitigating carbon footprint for knowledge distillation based deep learning model compression. *PLoS ONE 18*, e0285668.

Schwartz, R., J. Dodge, N. A. Smith, et O. Etzioni (2020). Green AI. *Comm. ACM 12*, 54–63.

Soviany, P., R. T. Ionescu, P. Rota, et N. Sebe (2022). Curriculum Learning: A Survey. *International Journal of Computer Vision 130*, 1573–1405.

Summary

In this work, we describe a hyperparameter search methodology (*HPO*) that uses less data by measuring the difficulty of fitting the model to a given sample. A reduced portion of the data is selected, respecting the proportion of difficulty in the whole dataset. This reasoning is inspired by research results on machine learning using *curriculum learning*. An evaluation of the approach is illustrated on the tasks of image recognition and Named Entity Recognition (NER). Experiments show that the amount of data required for HPO can be reduced by up to 60% to obtain the same choice of hyperparameters, compared with using all available data.

Representation Learning pour la codification des parcours thérapeutiques de patientes atteintes de cancer du sein à partir de données de remboursement : un benchmark pour des tâches de clustering

Marie Guyomard*, Anne-Déborah Bouhnik*, Louis Tassy** Raquel Ureña *

*SESSTIM, INSERM, IRD, AMU, Marseille, France
marie.guyomard@univ-amu.fr,
raquel.urena@univ-amu.fr,
**Département d'oncologie médicale, Institut Paoli Calmettes, Marseille, France

Résumé. La disponibilité croissante de dossiers médicaux informatisés (*Electronic Health Records*, EHRs) offre de nouvelles opportunités de recherche notamment en Intelligence Artificielle à des fins cliniques. Bien que les données de remboursement de soins de santé ne constituent pas à proprement parler des EHRs, elles fournissent néanmoins des informations précieuses quant aux parcours thérapeutiques des patients, incluant des informations sur les diagnostics, les procédures médicales et la prescription de médicaments. Cependant, la complexité et la dimensionnalité inhérentes à ce type de données posent des défis majeurs pour l'application directe de techniques d'apprentissage automatique (*Machine Learning*, ML). Par conséquent, des méthodes non supervisées permettant de codifier les données médicales des patients tout en réduisant leur dimensionnalité sont indispensables. Dans ce but, des méthodes de *Representation Learning* (RL) sont développées afin de créer des représentations fiables des patients.

Nous proposons dans cet article une évaluation comparative approfondie de trois méthodes couramment utilisées de RL. Nous codifions et regroupons les parcours thérapeutiques de patientes âgées d'un cancer du sein incident à partir de données de remboursement extraites du Système National des Données de Santé (SNDS). Nos résultats mettent en évidence les limites d'une évaluation des approches de RL uniquement basée sur des métriques de performance d'outils de ML. Nous soulignons dans nos travaux la nécessité de réaliser une évaluation de la fiabilité des espaces latents issus des apprentissages de RL, à travers notamment des analyses statistiques.

1 Introduction

Selon l'Institut National du Cancer [1], le cancer du sein (*Breast Cancer*, BC) est le cancer le plus fréquent mais aussi ayant le taux de mortalité le plus élevé chez les femmes, représentant 14% des décès par cancer féminin en 2018. Entre 1990 et 2018, l'incidence annuelle des nouveaux cas de BC en France a presque doublé, passant de 29 970 à 58 400 cas par an. Il est donc essentiel de développer des algorithmes de support d'aide à la décision clinique (ADS), dédiés à la prise en charge du BC.

Ces dernières années, les systèmes de Dossiers de Santé Électroniques (*Electronic Health Records*, EHRs) ont été de plus en plus adoptés dans les hôpitaux. Un cas particulier est celui de la base du Système National des Données de Santé (SNDS) française. Conçu initialement à des fins de remboursement de soins, le SNDS constitue un vaste répertoire de données de santé, permettant de suivre les parcours thérapeutiques des patients représentés sous la forme d'une séquence de visites au fil du temps. Chaque visite peut inclure plusieurs concepts médicaux représentés par des codes médicaux, comprenant des diagnostics, des actes médicaux et des traitements prescrits. L'abondance des données médicales qu'elle contient offre de nouvelles opportunités de recherche (Moulis et al., 2015), notamment en intelligence artificielle (IA) pour mieux comprendre les parcours thérapeutiques des patients.

Néanmoins, exploiter les EHRs pose plusieurs défis (Si et al., 2021; Shickel et al., 2017). Premièrement il est crucial de capter les dynamiques temporelles contenues dans ces données, tant les dépendances temporelles sont essentielles dans la pratique clinique. Le deuxième grand challenge de ces données réside dans leur caractère multimodal ; pour une seule hospitalisation plusieurs procédures médicales peuvent être réalisées et plusieurs traitements prescrits pour un seul diagnostic. Enfin, les codes médicaux contenus dans les EHRs sont très nombreux, à l'échelle des patients mais aussi des informations médicales contenues, rendant ces données hautement complexes.

Ainsi, l'application directe de méthodes d'apprentissage automatique (*Machine Learning*, ML) ou profondes (*Deep Learning*, DL) à ce format de données non structurées est d'une part très complexe et d'autre part peut aboutir à de faibles performances et à une interprétation difficile si ce n'est impossible des résultats. Les tâches de *Representation Learning* (RL) (Si et al., 2021; Shickel et al., 2017) des patients consistent à extraire des informations significatives d'une représentation mathématique dense d'un patient au sein d'un espace latent (*embedding*) de plus petite dimension. La qualité de cet embedding influence naturellement les performances des algorithmes qui en découlent : une mauvaise représentation entraîne nécessairement de faibles précisions pour les algorithmes prédictifs entraînés sur ces embeddings. Cependant la qualité et la fiabilité de ces apprentissages dans l'état de l'art sont principalement évaluées via les performances issues de tâches de prédiction, en se basant principalement sur des métriques de classification (Choi et al., 2016a,c; Miotto et al., 2016; Li et al., 2020; Rasmy et al., 2021; De Oliveira et al., 2022).

Cet article vise à évaluer différentes stratégies de RL à travers des tâches de clustering. D'une part, le clustering peut permettre d'obtenir des trajectoires types de soins des patients, ce qui est particulièrement important pour les études sur le cancer. Cette tâche vise à identifier des schémas communs entre les patients dans l'évolution de leur cancer, à travers leurs parcours thérapeutiques mais aussi leurs comorbidités spécifiques. D'autre part, notre objectif

1. Source : www.e-cancer.fr - Cancer du sein

étant de développer une représentation générale des parcours thérapeutiques des patients, indépendamment d'une tâche de prédiction spécifique, le clustering semble être la méthode la plus appropriée pour évaluer l'espace latent issu des tâches de RL. Dans cet article nous appliquons alors des méthodes de RL couramment employées dans la littérature à l'étude des parcours thérapeutiques des femmes atteintes de cancer du sein. Plus précisément, nous examinons les différents clusters appris à partir des espaces latents pour comparer leur capacité à conserver des informations importantes concernant la réalité clinique des patients, comme les traitements thérapeutiques par exemple, pouvant impacter la précision des ADS.

Cet article est structuré de la manière suivante ; La Section 2 fournit une vue d'ensemble des travaux de l'état de l'art impliquant le développement ou l'utilisation d'outils de RL. La Section 3 présente les données et décrit les critères de sélection des patientes atteintes d'un cancer du sein pour notre étude. La Section 4 introduit les méthodes de RL testées ainsi que la méthodologie de clustering utilisée pour évaluer à la fois leur précision et leur fiabilité. La Section 5 compare la précision et la fiabilité des espaces latents issus des diverses méthodes de RL appliqués à des tâches de clustering. Enfin la Section 6 conclut l'article.

2 État de l'art

Compte tenu de la vaste quantité de données disponibles, des méthodes de Deep Learning (DL), reconnues pour leur capacité à comprendre et réduire des données complexes, ont été explorées et mises en œuvre pour des tâches de RL.Parmi les stratégies d'apprentissage profond proposées dans la littérature, trois principales catégories émergent : celles basées sur des outils de traitement automatique du langage naturel (*Natural Langage Processing*, NLP) (Choi et al., 2016d,e,b; Pham et al., 2016; Beam et al., 2019), celles utilisant des réseaux de neurones autoencodeurs (Miotto et al., 2016; Landi et al., 2020; Baytas et al., 2017; De Oliveira et al., 2022), et celles reposant sur des Transformers (Li et al., 2020; Rasmy et al., 2021).

Ces stratégies de RL ont été principalement développées à des fins de prédiction. La majorité des algorithmes proposés se concentrent sur la représentation de codes médicaux (carrés colorés sur la Figure 1) afin de prédire des maladies (Choi et al., 2016c,e; Li et al., 2020; Rasmy et al., 2021), ce qui aboutit à un vecteur pondéré pour chaque code médical unique. Une autre stratégie consiste à estimer une représentation des visites (contours noirs sur la Figure 1) pour prédire une pathologie (Rasmy et al., 2021) ou la prochaine visite médicale (Choi et al., 2016b,a). Dans ce cas, l'espace latent résultant a la même taille que la longueur de la séquence initiale pour chaque patient. Enfin, le dernier type de représentation vise à créer une représentation complète d'un patient (Pham et al., 2016; Miotto et al., 2016; Baytas et al., 2017; Landi et al., 2020). Cela signifie que l'ensemble des données brutes d'un patient est condensé dans un vecteur d'encodage, ou embedding englobant à la fois les codes médicaux et les informations temporelles.

3 Données

Les données proviennent de l'étude VICAN (Bouhnik et al., 2015), une enquête nationale sur les survivants du cancer en France. Douze types de cancers ont été pris en compte, mais pour notre étude, nous nous concentrons uniquement sur les patientes atteintes d'un cancer du

sein (BC). La méthodologie décrite dans Dumas et al. (2022) a été utilisée pour préparer le jeu de données. Seules les femmes diagnostiquées avec un cancer du sein, ayant atteint l'âge de la majorité et ayant subi une intervention chirurgicale, ont été incluses dans l'étude. Les patientes diagnostiquées avec un autre type de cancer pendant la période d'inclusion ont été exclues intentionnellement. Cependant, contrairement à la méthodologie décrite dans Dumas et al. (2022), les femmes atteintes d'un cancer du sein métastatique de stade IV ont été incluses, représentant 28% des patientes. De plus, seuls les événements (c'est-à-dire les consultations médicales) survenus dans les deux années suivant la date de diagnostic ont été retenus. La date de diagnostic et les caractéristiques cliniques des patientes (Tableau 5.3) ont été déterminées selon la méthodologie également proposée dans Dumas et al. (2022).

Au final, le jeu de données résultant est composé de 1 304 361 événements pour 6 111 patientes. En moyenne, chaque patiente a eu recours à 213 visites, avec un minimum de 4 et un maximum de 1 111 visites. L'âge des patientes à l'année du diagnostic varie entre 21 et 82 ans, avec un âge moyen de 51 ans. Il est important de noter que tous les codes médicaux sont conservés pour chaque patiente, et pas uniquement ceux directement liés aux séquences thérapeutiques pour le traitement du BC.

Le jeu de données initial contient 5 467 codes médicaux uniques, incluant 2 447 codes de diagnostic (classification CIM-10), 1 977 procédures (classification *Anatomical Therapeutic Chemical*, ATC) et 1 043 médicaments (Classification Commune des Actes Médicaux, CCAM). En suivant la méthodologie décrite dans Choi et al. (2016d) et Choi et al. (2016e), nous regroupons les codes médicaux en fonction de leur structure hiérarchique. Ce processus vise à réduire le nombre de chiffres contenus dans les codes médicaux afin de les regrouper dans des niveaux hiérarchiques supérieurs, réduisant ainsi le nombre total de codes uniques. Cela entraîne certes une perte de détails spécifiques, mais diminue grandement la complexité des données. Nous avons décidé de ne conserver que les 2 premiers chiffres pour tous les codes médicaux, ce qui réduit le jeu de données à 3 407 codes médicaux uniques.

4 Méthodologie

L'objectif de cet article est d'évaluer la précision et de mesurer la fiabilité de différentes méthodes de RL pour codifier les parcours thérapeutiques de patientes atteintes d'un BC dans le cadre de tâches de clustering. Soit $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ le dictionnaire contenant tous les codes médicaux uniques. Sans perte de généralité, toutes les notations et algorithmes sont présentés dans cette section pour un seul patient fixe afin de simplifier les formulations. Un EHR est défini comme une séquence de n visites $V = \{v_1, \ldots, v_n\}$, chaque visite contenant un sous-ensemble de codes médicaux. La j-ème visite $v_j = \{d_1^j, \ldots, d_{k_j}^j\}$ est définie par une séquence de k_j codes médicaux, avec $v_j \subseteq \mathcal{C}$. Ainsi, le patient dispose d'un total de L codes médicaux, avec $L = \sum_{t=1}^{n} |v_t|$.

4.1 Méthodes de Deep Representation Learning

Notre objectif principal est d'apprendre un espace latent ou embedding via une tâche de RL (Figure 1) codifiant les parcours thérapeutiques des patients dans un espace de taille $m \in \mathbb{R}_+$ avec des méthodes de RL $f_C : \mathbb{R}^L \to \mathbb{R}_+^m$.

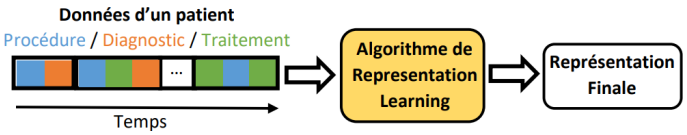


FIG. 1 – *Schéma d'application du Representation Learning pour un patient.*

Dans le benchmark proposé, nous nous concentrons sur trois algorithmes conçus pour apprendre des représentations profondes, Skip-Gram, Med2Vec et Deep Patient. Il convient de noter que, comme indiqué précédemment, cette contribution vise à évaluer le RL uniquement sur des tâches de clustering. Par conséquent, les approches supervisées basées sur des Transformers proposées dans Li et al. (2020); Rasmy et al. (2021) ne sont pas considérées.

4.1.1 Skip-Gram

Dans Choi et al. (2016d) et Choi et al. (2016e), les auteurs proposent d'utiliser Skip-Gram (Mikolov et al., 2013) pour apprendre les codifications des codes médicaux. L'application de Skip-Gram aux EHRs suppose qu'un concept cible (un code médical) est défini par ses voisins, appelés vecteurs de contexte. Ainsi, pour chaque code médical unique, un vecteur de concept $\nu(c)$ est modélisé et appris en maximisant la probabilité logarithmique moyenne suivante :

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{-w\leq j\leq w, j\neq 0}\log p(c_{t+j}|c_t), \tag{1}$$

avec w représentant la taille de la fenêtre contextuelle, c'est-à-dire le nombre de voisins et $p(c_{t+j}|c_t)$ la probabilité logarithmique définie par

$$p(c_{t+j}|c_t) = \frac{\exp(\nu(c_{t+j})^T\nu(c_t))}{\sum_{c=1}^{|\mathcal{C}|}\exp(\nu(c)^T\nu(c_t))}. \tag{2}$$

Il résulte de cette approche un embedding $\nu(c)$ pour chacun des $\mathcal{C}|$ codes médicaux uniques. Afin d'obtenir des représentations de patients, les auteurs Choi et al. (2016e) proposent d'additionner tous les vecteurs de concepts médicaux apparaissant pour un échantillon. La représentation du patient qui en résulte peut alors être définie comme suit :

$$e^{SG} = \sum_{t=1}^{n}\sum_{j=1}^{k_t}\nu(d_j^t) \in \mathbb{R}^m. \tag{3}$$

4.1.2 Med2Vec

L'algorithme proposé dans Choi et al. (2016b) est un réseau de neurones perceptron multicouche. Cette méthode utilise en entrée des représentations vectorielles binaires des EHRs modélisés par $\bar{V} = \{0,1\}^{|\mathcal{C}|\times n}$, tel que $\bar{V}_{i,j} = 1$ si lors de la visite j, le code médical i est présent. Cette méthode apprend conjointement les représentations des codes et des visites. Elle commence par modéliser une représentation intermédiaire de visite $u_t \in \mathbb{R}^{m'}$:

$$u_t = \phi(W_c\bar{v}_t + b_c), \tag{4}$$

avec $\bar{v}_t = \bar{V}_{.,t}$ la visite, $\phi(x) = \max\{0, x\}$ la fonction d'activation ReLU, $W_c \in \mathbb{R}^{m' \times |\mathcal{C}|}$ la matrice de poids du code et $b_c \in \mathbb{R}^{m'}$ les biais. Les auteurs concatènent ensuite cette représentation intermédiaire de visites avec les informations démographiques $d_t \in \mathbb{R}^d$ afin d'obtenir la représentation finale des visites de dimension m :

$$\nu_t = \phi(W_v[u_t, d_t] + b_v), \tag{5}$$

tel que $W_v \in \mathbb{R}^{m \times (m' \times d)}$ est la matrice de poids de visite et $b_v \in \mathbb{R}^m$ le vecteur de biais. Les deux vecteurs de représentation sont appris conjointement en minimisant l'erreur d'entropie croisée. Comme pour la méthode précédente, la représentation au niveau du patient peut être obtenue, mais cette fois en additionnant les représentations des visites :

$$e^{Med} = \sum_{t=1}^{n} \nu_t \in \mathbb{R}^m. \tag{6}$$

Cette méthodologie a été utilisée à la fois pour prédire une pathologie (Choi et al., 2016c) ou la prochaine visite d'un patient (Choi et al., 2016a).

4.1.3 Deep Patient

Les Autoencodeurs (AE) sont reconnus pour permettre de réduire la dimension des EHRs (Miotto et al., 2016; Baytas et al., 2017; Landi et al., 2020). Pour comprendre cette approche, considérons l'entrée $\tilde{V} \in \mathbb{R}^L$ définie comme la concaténation de toutes les séquences de visites $\tilde{V} = v_1 \oplus \cdots \oplus v_n$. Un AE transforme la séquence de visites $\tilde{V}$ en une représentation latente $y\ dans \mathbb{R}^m$ au travers d'un encodeur f_θ paramétré par $\theta = \{W, b\}$ et défini par :

$$y = f_\theta(\tilde{V}) = s(W\tilde{V} + b), \tag{7}$$

avec $s(\cdot)$ une transformation non linéaire, $W \in \mathbb{R}^{m \times L}$ une matrice de poids et $b \in \mathbb{R}^m$ un biais. Un décodeur $g_{\theta'}$ paramétré par $\theta' = \{W', b'\}$, avec $W' \in \mathbb{R}^{L \times m}$ et $b' \in \mathbb{R}^m$ est ensuite appliqué à la représentation latente y afin d'obtenir le vecteur reconstruit $z \in \mathbb{R}^L$:

$$z = g_{\theta'}(y) = s(W'y + b'). \tag{8}$$

Les paramètres θ et θ' sont optimisés en minimisant l'erreur moyenne de reconstruction. L'embedding du patient résultant de cette architecture est la représentation latente (7).

Dans le contexte de notre benchmark, nous utilisons l'architecture Deep Patient (Miotto et al., 2016). Cette architecture est basée sur l'empilement de plusieurs AE de manière consécutive. Le premier est un *denoising AE*, entraîné sur une version bruitée de l'entrée pour limiter le sur-apprentissage.

4.2 Evaluation des méthodes de RL : Clustering

Comme mentionné précédemment, la majorité des travaux de l'état de l'art évaluent la performance des approches de RL en fonction de leur précision dans les tâches de prédiction pour lesquelles elles sont conçues. Étant donné que nous nous concentrons sur des méthodes non supervisées de RL apprenant des embeddings généraux, nous avons décidé leur performance

générale sur des tâches de clustering. Le clustering vise à regrouper les données en clusters similaires. Afin d'évaluer à la fois la performance et la fiabilité des espaces latents générés pour les parcours médicaux des patients, deux méthodes de clustering sont employées. La méthode K-means segmente l'espace d'entrée, c'est-à-dire les représentations des patients, en K sous-groupes. L'une des principales limitations de l'algorithme K-means est qu'un point donné appartient à un et un seul cluster. Pour surmonter cette limitation potentielle, nous avons également implémenté le modèle de mélange gaussien (*Gaussian Mixture Model*, GMM). Le GMM utilise K mélanges gaussiens et estime la probabilité qu'un échantillon appartienne à l'une des K distributions gaussiennes.

5 Expériences numériques

Dans cette section, les expériences réalisées sur les méthodes de RL appliquées à la population de patientes atteintes de cancer du BC sont détaillées. Tout d'abord, la configuration des expériences est détaillée. Ensuite, la qualité et la fiabilité des clusters obtenus sont évaluées.

5.1 Configuration des expériences

5.1.1 Méthodes

Tout d'abord, nous avons implémenté Skip-Gram en Pytorch, étant donné que l'outil RL décrit dans Choi et al. (2016d) et Choi et al. (2016e) n'est pas disponible en open-source. Comme suggéré dans Steiger et Kroll (2023), pour chaque code médical, nous avons généré des voisins réels mais aussi des faux voisins sélectionnés de manière aléatoire. Pour les deux autres stratégies de RL testées, Med2Vec et Deep Patient, le code est disponible sur les GitHub des auteurs [2, 3]. Cependant, ces packages ne fournissent pas suffisamment d'outils pour notre objectif, qui est de valider la qualité et la fiabilité des clusters. Ainsi, nous avons décidé de compléter leurs fonctionnalités (marquées par une étoile sur la Figure 2) et de fournir ces outils complémentaires sur notre GitHub [4]. Sachant que Deep Patient et Skip-Gram n'incluent pas de données démographiques, l'entraînement de Med2Vec n'a pas inclus ces informations.

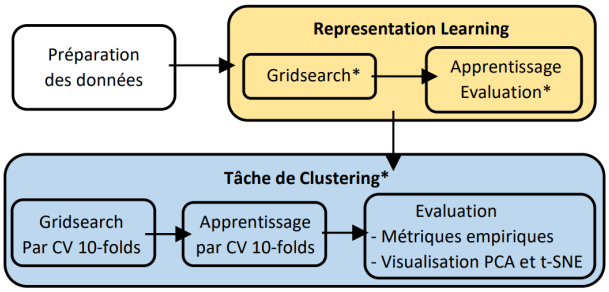


FIG. 2 – *Configuration des expériences. *Les outils complémentaires fournis sur notre Github.*

2. ; Med2Vec Github
3. ; Deep Patient Github
4. ; Deep Representation Learning and Clustering

5.1.2 Entraînement et Validation des méthodes de RL

Plusieurs paramètres doivent être ajustés pour tous les algorithmes testés. Nous avons implémenté une fonction de gridsearch pour toutes les méthodes. Nous commençons par diviser le jeu de données en un ensemble d'entraînement (80%) et un ensemble de validation (20%). Ensuite, les modèles sont entraînés avec différentes combinaisons de paramètres. L'ensemble optimal d'hyperparamètres est celui qui minimise la fonction de coût sur l'échantillon de validation afin d'éviter le sur-apprentissage.

Une fois les hyperparamètres optimaux obtenus, nous les utilisons pour apprendre les méthodes de RL. Nous avons implémenté pour chaque modèle une fonction d'évaluation afin de vérifier la généralisation de l'espace d'embedding sur un nouvel échantillon. Pour ce faire, nous ajustons les modèles sur un ensemble d'entraînement (80%) et les appliquons sur un ensemble de validation (20%). Pour chaque méthode testée, les codes permettant d'évaluer la fonction de coût sur un échantillon de validation sont fournis sur notre Github.

5.1.3 Clustering

Nous utilisons les fonctions de scikit-learn pour implémenter les K-means et les GMMs. Comme pour l'entraînement de la tâche RL, une étape de gridsearch est nécessaire pour optimiser le nombre de clusters. Nous conservons le paramètre qui minimise le score de silhouette moyen obtenu lors d'une validation croisée à 10 folds (CV) sur l'échantillon de test. Le modèle optimal est ensuite entraîné sur une CV à 10 folds. Bien que la validation croisée ne soit généralement pas utilisée dans le cadre des méthodes de clustering, notre objectif est d'explorer la fiabilité et, par conséquent, la stabilité de nos espaces latents.

Pour évaluer la qualité des clusters, nous calculons le score de silhouette et l'indice de Davies-Bouldin. Le tableau 1 résume la moyenne (et l'écart-type) de ces métriques sur les 10 folds. Nous examinons également la nature discriminante des clusters en les visualisant à l'aide de deux techniques de réduction dimensionnelle : une analyse en composantes principales (PCA) et un t-SNE (Figure 3). Enfin, nous effectuons un test du Chi-carré pour évaluer la fiabilité des clusters par rapport à la réalité clinique des patients (Tableau 5.3).

5.2 Résultats empiriques

Pour tous les embeddings testés, le clustering K-means est plus précis que le modèle GMM (Tableau 1). En se concentrant sur le score de silhouette, la représentation de Deep Patient (7) semble mieux adaptée à notre jeu de données sur le BC. En effet, les scores de silhouette des espaces d'embedding de Skip-Gram (3) et de Med2Vec (6) sont très faibles. Par conséquent, selon l'évaluation basée sur le score de silhouette, les représentations apprises ne parviennent pas à fournir des informations permettant de séparer correctement la population. Ces résultats étaient attendus, car ces deux méthodes ne génèrent pas directement les représentations des patients. En effet, la représentation finale est construite dans Skip-Gram (3) et Med2Vec (6) à partir des représentations des codes et des visites respectivement. Lorsqu'on estime directement la représentation complète du patient, comme le fait Deep Patient, la tâche de clustering qui lui est appliquée devient plus pertinente ; le score de silhouette est égal à 0.98 pour le K-means, aussi bien sur l'échantillon d'entraînement que de validation. De plus, l'indice moyen de Davies-Bouldin obtenu sur la représentation de Deep Patient par K-means est plus faible

(0.13) que ceux calculés sur les deux autres représentations des patients (supérieurs à 0.3). Ainsi, la première conclusion que l'on peut tirer est qu'apprendre la représentation du patient, plutôt que de la construire à partir des représentations des codes ou des visites, s'avère plus performant.

	Training Sample		**Validation Sample**	
	Silhouette Score	Davies-Bouldin ind.	Silhouette Score	Davies-Bouldin ind.
K-means				
SG	0.6 (0.005)	0.34 (0.005)	0.6 (0.006)	0.344 (0.02)
M2V	0.55 (0.004)	0.3 (0)	0.54 (0.006)	0.31 (0.005)
DP	0.98 (0)	0.13 (0.005)	0.98 (0 002)	0.13 (0.007)
Gaussian Mixture Model				
SG	0.37 (0.01)	0.52 (0.008)	0.35 (0.01)	0.52 (0.01)
M2V	0.06 (0.06)	1.1 (0.4)	0.3 (0.09)	0.8 (0.2)
DP	0.9 (0)	0.62 (0.01)	0.9 (0.005)	0.6 (0.09)

TAB. 1 – *Résultats moyens (écart-type) obtenus par le clustering par CV* $10-$*folds, pour Skip-Gram (SG (3)), Med2Vec (M2V (6)) et Deep Patient (DP (7)).*

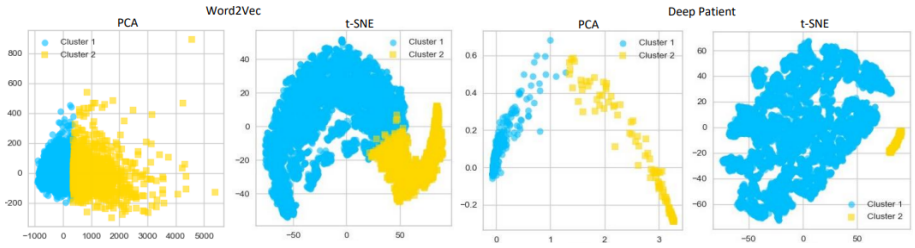


FIG. 3 – *Visualisation par PCA et t-SNE des clusters appris sur la représentation du patient résultant des méthodes de RL Skip-Gram (Gauche) et Deep Patient (Droite).*

Sur la Figure 3, les visualisations PCA et t-SNE nous retrouvons les clusters appris par K-means sur les embeddings résultant de Skip-Gram (Gauche) et de Deep Patient (Droite). Comme l'indiquent les résultats empiriques du Tableau 1, et plus précisément les scores de silhouette, les clusters issus des représentations de Deep Patient sont plus faciles à séparer et à distinguer que ceux issus de Skip-Gram.

5.3 Comparaison sur la fiabilité clinique

Nous avons effectué un test du Chi-2 sur le clustering K-means en fonction des caractéristiques du cancer des patientes. Les résultats moyens (et l'écart-type) obtenus sur 5 sous-échantillons aléatoires sont affichés dans le Tableau 5.3. Une valeur de p faible ($< \ 0.05$) signifie que l'hypothèse d'indépendance entre les clusters n'est pas acceptée, ce qui suggère que les clusters discriminent bien par rapport à cette variable. D'après ce test statistique, nous pouvons conclure que la tâche de clustering réalisée sur les représentations des patients Med2Vec

est fortement alignée à la réalité clinique des patients, contrairement à celle réalisée sur les représentations de Deep Patient. En effet, bien que l'outil RL Deep Patient semble plus précis en termes de score de silhouette (Tableau 1), le clustering obtenu n'est pas fiable en termes de réalité clinique.

	SG	**M2V**	**DP**
Partial Mastectomy	<0.05 (0)	0.07 (0.04)	<0.05 (0.02)
Mastectomy	<0.05 (0)	0.37 (0.13)	<0.05 (0.01)
Axillary Surgery	<0.05 (0)	<0.05 (0)	0.7 (0.23)
Chemotherapy Y/N	<0.05 (0)	<0.05 (0)	0.5 (0.27)
Chemotherapy Setting	<0.05 (0)	<0.05 (0)	<0.05 (0.03)
Chemotherapy Regimen	<0.05 (0)	<0.05 (0)	0.1 (0.22)
Targeted Therapy Y/N	0.87 (0.12)	<0.05 (0)	0.6 (0.31)
Targeted Therapy Setting	0.7 (0.01)	<0.05 (0)	0.7 (0.2)
Targeted therapy Regimen	0.34 (0.12)	<0.05 (0)	0.6 (0.31)
Radiotherapy Y/N	<0.05 (0.3)	<0.05 (0)	0.4 (0.23)
Radiotherapy Setting	<0.05 (0.21)	<0.05 (0)	<0.05 (0)
Endocrine Therapy Y/N	<0.05 (0.01)	<0.05 (0)	0.2 (0.2)
Endocrine Therapy Setting	<0.05 (0.03)	<0.05 (0)	<0.05 (0)
Endocrine Therapy Regimen	<0.05 (0)	<0.05 (0)	<0.05 (0)
BC Sub Type	<0.05 (0)	<0.05 (0)	0.2 (0.12)
Nodal status	<0.05 (0.01)	<0.05 (0)	0.06 (0.07)
Metastatic	<0.05 (0)	<0.05 (0)	<0.05 (0)

TAB. 2 – *Moyenne (écart-type) des p-values obtenues par le test du chi-2 entre les clusters K-means et les variables caractérisant le BC, sur* 5 *des sous-échantillons aléatoires.*

6 Conclusion

Dans cet article, nous avons réalisé une évaluation comparative des trois approches de *Representation Learning* (RL) couramment utilisées, dans le but de codifier les parcours thérapeutiques des patients à partir de données de remboursement des soins issues du SNDS, sur des tâches de clustering. Notre étude montre qu'évaluer les méthodes de RL uniquement aux travers de métriques empiriques est insuffisant pour juger de la qualité des espaces latents obtenus. En effet, lorsque nous considérons des tâches non supervisées comme le clustering, il devient évident qu'une valeur élevée du score de silhouette n'implique pas nécessairement une grande fiabilité de l'algorithme face à la réalité clinique. Par conséquent, il est essentiel de développer des métriques d'évaluation qui prennent en compte à la fois la performance et la cohérence des outils d'apprentissage de RL.

Références

Baytas, I. M., C. Xiao, X. Zhang, F. Wang, A. K. Jain, et J. Zhou (2017). Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international*

conference on knowledge discovery and data mining, pp. 65–74.

Beam, A. L., B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, et I. S. Kohane (2019). Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing 2020*, pp. 295–306. World Scientific.

Bouhnik, A.-D., M.-K. Bendiane, S. Cortaredona, L. S. Teyssier, D. Rey, C. Berenger, V. Seror, P. Peretti-Watel, V. Group, et al. (2015). The labour market, psychosocial outcomes and health conditions in cancer survivors : protocol for a nationwide longitudinal survey 2 and 5 years after cancer diagnosis (the vican survey). *BMJ open 5*(3), e005971.

Choi, E., M. T. Bahadori, A. Schuetz, W. F. Stewart, et J. Sun (2016a). Doctor ai : Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR.

Choi, E., M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, et J. Sun (2016b). Multi-layer representation learning for medical concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1495–1504.

Choi, E., M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, et W. Stewart (2016c). Retain : An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems 29*.

Choi, E., A. Schuetz, W. F. Stewart, et J. Sun (2016e). Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv :1602.03686*.

Choi, Y., C. Y.-I. Chiu, et D. Sontag (2016d). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings 2016*, 41.

De Oliveira, H., P. Martin, L. Ludovic, A. Vincent, et X. Xiaolan (2022). Explaining predictive factors in patient pathways using autoencoders. *Plos one 17*(11), e0277135.

Dumas, E., L. Laot, F. Coussy, B. Grandal Rejo, E. Daoud, E. Laas, A. Kassara, A. Majdling, R. Kabirian, F. Jochum, et al. (2022). The french early breast cancer cohort (fresh) : a resource for breast cancer research and evaluations of oncology practices based on the french national healthcare system database (snds). *Cancers 14*(11), 2671.

Landi, I., B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, et R. Miotto (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine 3*(1), 96.

Li, Y., S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, et G. Salimi-Khorshidi (2020). Behrt : transformer for electronic health records. *Scientific reports 10*(1), 7155.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems 26*.

Miotto, R., L. Li, B. A. Kidd, et J. T. Dudley (2016). Deep patient : an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports 6*(1), 1–10.

Moulis, G., M. Lapeyre-Mestre, A. Palmaro, G. Pugnet, J.-L. Montastruc, et L. Sailler (2015).

French health insurance databases : what interest for medical research ? *La Revue de médecine interne 36*(6), 411–417.

Pham, T., T. Tran, D. Phung, et S. Venkatesh (2016). Deepcare : A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining : 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20*, pp. 30–41. Springer.

Rasmy, L., Y. Xiang, Z. Xie, C. Tao, et D. Zhi (2021). Med-bert : pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine 4*(1), 86.

Shickel, B., P. J. Tighe, A. Bihorac, et P. Rashidi (2017). Deep ehr : a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics 22*(5), 1589–1604.

Si, Y., J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, et K. Roberts (2021). Deep representation learning of patient data from electronic health records (ehr) : A systematic review. *Journal of biomedical informatics 115*, 103671.

Steiger, E. et L. E. Kroll (2023). Patient embeddings from diagnosis codes for health care prediction tasks : Pat2vec machine learning framework. *JMIR AI 2*, e40755.

Summary

The increasing availability of Electronic Health Records (EHRs), offers new research opportunities, particularly in Artificial Intelligence (AI) for clinical purposes. Although healthcare reimbursement data do not strictly constitute EHRs, they nonetheless provide valuable insights into patients' therapeutic pathways, including information on diagnoses, medical procedures, and medication prescriptions. However, the complexity and high dimensionality inherent to this type of data pose significant challenges for the direct application of Machine Learning (ML) techniques. Consequently, computing unsupervised methods for encoding patients' medical data while reducing their dimensionality are essential. To this end, Representation Learning (RL) methods have been developed to create reliable patient representations.

In this paper, we present a benchmark evaluation of three commonly used RL methods. We encode and cluster the therapeutic pathways of patients diagnosed with incident breast cancer using reimbursement data extracted from the French Nationwide Healthcare Database (SNDS). Our results highlight the limitations of evaluating RL approaches solely based on ML performance metrics. Our findings emphasize the necessity of assessing the reliability of the latent spaces produced by RL methods, particularly through statistical analyses.

Utilisation de Réseau de Neurone Bayésien pour la Prédiction de la Résistance aux Antibiotiques

Résumé. La résistance aux antibiotiques est reconnue par l'Organisation Mondiale de la Santé comme une menace majeure pour la santé mondiale. L'identification précise de la sensibilité bactérienne aux antibiotiques est cruciale, mais nécessite souvent plusieurs jours. Dans les systèmes d'aide à la décision médicale, comme celui proposé dans cette contribution, il est essentiel d'évaluer l'incertitude du modèle lors de la prédiction. Dans ce travail, nous proposons un réseau de neurones bayésien pour prédire la résistance aux antibiotiques à différentes étapes du processus d'identification bactérienne, pour un ensemble de 47 antibiotiques. D'excellents résultats ont été obtenus, tout en fournissant une mesure de l'incertitude épistémique. Pour permettre une utilisation clinique de l'approche proposée en tant que système d'aide à la décision, le modèle a été intégré dans une application web accessible à la fois sur téléphone mobile et ordinateur.

1 Introduction

La résistance aux antibiotiques a été déclarée par l'Organisation mondiale de la Santé (OMS) comme l'une des menaces les plus graves pour la santé mondiale (World Health Organization, 2020). La meilleure façon de contrôler la propagation de cette résistance réside dans la prescription du traitement le plus approprié. Un traitement antimicrobien initial est souvent prescrit de manière empirique, en attendant les résultats de la culture bactérienne et de l'antibiogramme, un processus pouvant prendre plusieurs jours, ce qui représente un défi, même pour les médecins expérimentés. Nous proposons ici une première approche combinant les métadonnées des patients et les informations bactériennes pour prédire la résistance à chaque étape de l'identification bactérienne et à un large éventail d'antibiotiques (Urena et al., 2024). L'apprentissage profond offre une solution viable pour cette tâche de prédiction, mais dans un domaine tel que la médecine, il est crucial d'être prudent et précis. Il reste difficile de comprendre le processus de décision d'un modèle et d'évaluer son incertitude. Différents algorithmes d'apprentissage automatique ont été proposés pour aborder le problème de la résistance aux antibiotiques (Yelin et al., 2019; Ren et al., 2021). Dans ce travail, nous proposons l'utilisation des réseaux de neurones bayésiens (Bayesian Neural Networks, BNNs), une approche

probabiliste de l'apprentissage profond dans laquelle les poids d'un réseau de neurones classique sont remplacés par des distributions de probabilité apprises via l'inférence bayésienne (Neal, 2012; MacKay, 1992). Les BNNs permettent de quantifier l'incertitude épistémique, offrant ainsi au modèle la capacité d'indiquer son niveau d'incertitude dans ses prédictions. L'utilité de l'incertitude en apprentissage automatique appliqué à la santé suscite un intérêt croissant (Begoli et al., 2019; Kompa et al., 2021). Pour répondre aux besoins des applications en conditions réelles, le modèle développé peut produire des prédictions à chaque étape du processus d'identification.

2 Méthodes

2.1 Données

Le jeu de données comprends 91 061 informations de cultures bactériennes collectées entre 2014 et 2022 dans deux hôpitaux de Marseille. Il inclut des métadonnées, des informations sur les espèces bactériennes et leur sensibilité à 47 antibiotiques. L'un des principaux défis consistait à structurer le processus de prise de décision en cinq étapes, de manière analogue avec la pratique clinique. Voici un aperçu de ces étapes : lorsqu'un médecin reçoit un patient souffrant d'une infection bactérienne, il prélève un échantillon qu'il envoie au laboratoire pour une analyse bactériologique (stade 1). À la réception de l'échantillon, le bactériologiste effectue une observation directe (stade 2) avant de mettre l'échantillon en culture. Sur une période pouvant aller jusqu'à 2 jours, la culture permet aux bactéries de se développer, et les résultats sont ensuite transmis au médecin (stade 3). Après quelques jours supplémentaires, l'espèce exacte peut être identifiée (stade 5), bien que parfois seul le genre soit disponible (stade 4). La dernière étape est l'antibiogramme, qui détermine si la bactérie est résistante ou non à un ensemble d'antibiotiques, ce qui constitue nos variables cibles. En alignement avec la pratique médicale, le modèle proposé peut fournir des prédictions à chaque étape de ce processus, facilitant ainsi une prescription empirique. En outre, des métadonnées ont été extraites et ajoutées en tant que variables supplémentaires (à savoir, le site de l'infection, les antécédents du patient concernant le portage de bactéries multirésistantes (BMR) et l'unité médicale).

2.2 Réseaux de Neurones Bayésiens

Les BNNs peuvent être compris comme une combinaison de réseaux de neurones et d'inférence bayésienne. Plus précisément, ils sont définis comme des réseaux neuronaux artificiels stochastiques entraînés à l'aide de techniques d'inférence bayésienne (Jospin et al., 2022), une méthode statistique basée sur le théorème de Bayes. Dans l'apprentissage profond traditionnel, les poids sont des valeurs fixes (initialement aléatoires) qui sont itérativement mises à jour par descente de gradient pour minimiser une fonction de coût. En revanche, les BNNs apprennent des distributions sur les poids, et les paramètres de ces distributions sont optimisés via des méthodes d'inférence bayésienne. Cette approche présente plusieurs avantages notables, tels que l'atténuation du surapprentissage, l'apprentissage à partir de petits ensembles de données, et la possibilité de mesurer d'incertitude sur les prédictions (Qinghui Yu et al.). Diverses techniques permettent d'apprendre des distributions de probabilité sur les poids ; ce travail se base sur l'inférence variationnelle (Blei et al., 2017), qui permet d'approximer la distribution à posteriori

par minimisation de la divergence via descente de gradient. Une fois le modèle entraîné, pour effectuer des prédictions, la prédiction est calculée avec une approche de Monte Carlo : on calcule la prédiction $\hat{y}$ par rapport à une entrée x en prenant la moyenne sur un échantillon de la distribution prédictive : $\hat{y} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{F}_{w_k}(x)$ où K représente le nombre d'échantillons et $\mathcal{F}_{w_k}$ est le modèle avec un ensemble de poids w_k échantillonné à partir de la distribution a posteriori apprise. Cette approche, consistant à calculer la moyenne empirique sur un échantillon de la distribution prédictive, est analogue à un cas spécifique d'apprentissage par ensemble. Chaque propagation avant correspond au résultat d'un modèle unique avec un ensemble de poids distinct. La variance de la distribution prédictive constitue alors une mesure de l'incertitude épistémique. Chaque $\mathcal{F}_{w_k}$ dans l'ensemble correspond à une règle de décision différente. Par conséquent, si le modèle est confiant dans ses prédictions, toutes décisions seront similaires et à l'inverse on observera une plus grande variance dans les prédicitions.

2.2.1 Le modèle proposé

Après un ajustement des hyperparamètres à l'aide de Hyperband (Li et al., 2018), les meilleurs résultats ont été obtenus avec trois couches variationnelles denses comportant 300 unités chacune, et de la batch normalization entre chaque couche. Dans les réseaux de neurones bayésiens entraînés par inference variationelle, la distribution a priori joue un rôle de régularisation. Nous avons utilisé une loi normale multivariée avec une matrice de covariance diagonale, une approche équivalente à la régularisation L2. Concernant la distribution a posteriori, l'utilisation d'une loi normale indépendante est une pratique courante, à la fois pour sa commodité mathématique et pour permettre l'utilisation du *reparametrization trick* (Jospin et al., 2022). Le modèle doit produire des prédictions de sensibilité pour un ensemble de 47 antibiotiques. Par conséquent, la dernière couche est une couche entièrement connectée avec une activation sigmoïde, permettant de générer des prédictions indépendantes pour chaque antibiotique. Pour une tâche de classification binaire, la fonction de coût utilisée est la fonction d'entropie croisée binaire. Le modèle a été validé par validation croisée en 5 folds, chaque fold étant entraîné pendant 30 époques. Pour évaluer les performances du modèle, l'Aire sous la courbe ROC (AUROC) a été utilisée comme métrique principale.

TAB. 1 – *AUROC et AUSE sur l'ensemble de test pour le réseau de neurones bayésien et le réseau de neurones récurrent de référence, pour chaque étape.*

	Stade 1	Stade 2	Stade 3	Stade 4	Stade 5	Moy.
BNN	0.669	0.815	0.861	0.886	0.886	0.8234
RNN	0.676	0.82	0.865	0.896	0.896	0.831
AUSE	0.13	0.066	0.044	0.034	0.034	0.061

2.3 Résultats

Les résultats du BNN ont été comparés à ceux du meilleur modèle d'apprentissage profond traditionnel développé, un bi-LSTM. Les deux modèles présentent des comportements identiques à travers les étapes, ce qui indique que le BNN converge presque de manière identique au bi-LSTM. Cependant, une différence apparaît au niveau des scores, légèrement inférieurs

pour le BNN. Cette perte de performance est attendue dans le cadre de l'inférence bayésienne. Les AUROC pour chaque étape, moyennées sur tous les antibiotiques, sont présentées dans le Tableau 1. De manière générale, l'AUROC augmente au fil des étapes, ce qui est cohérent, car davantage d'informations sont obtenues tout au long du processus d'identification de la bactérie. La première étape est la moins performante (AUROC 0.67) et présente la plus forte variance par rapport aux autres étapes, en raison de l'absence d'informations sur la bactérie. À l'étape 2, on observe une augmentation significative de l'AUROC (+14%). À partir de cette étape, l'écart entre les étapes se réduit progressivement pour atteindre un plateau à 0.896. Il n'y a pas de différence entre les étapes 4 et 5, où les informations sur le genre et l'espèce sont introduites. Le modèle ne tire pas de bénéfice supplémentaire de ces deux caractéristiques, car l'espèce est une version plus détaillée du genre, bien que les deux soient des caractéristiques proches. Cependant, le genre a été inclus en raison de son utilité potentielle dans des cas rares. Pour évaluer la qualité des estimations d'incertitude, les *sparsification plots* ont été utilisés. L'*Area Under the Sparsification Error* (AUSE), introduite dans Ilg et al. (2018), quantifie la différence entre un oracle et les les *sparsification plots* du modèle en calculant l'aire entre les deux courbes. Idéalement, l'AUSE doit être la plus faible possible. Comme indiqué dans le Tableau 1, cette valeur est faible. De manière similaire à l'AUROC, l'AUSE diminue à chaque étape, ce qui indique que l'incertitude est de mieux en mieux calibrée à mesure que des informations supplémentaires sur la bactérie sont acquises. En résumé, la mesure d'incertitude est pertinente, car elle est plus élevée lorsque le modèle commet des erreurs, ce qui, dans une certaine mesure, constitue un comportement satisfaisant.

3 Discussion

Les résultats obtenus sont satisfaisants, avec des AUROC atteignant jusqu'à 0,9 pour les dernières étapes (Tableau 1). Nous estimons que disposer d'une mesure de l'incertitude épistémique peut avoir un impact significatif en pratique, comme le montrent les incertitudes bien calibrées reflétées par l'AUSE. Cependant, il est important de noter que l'incertitude tend à être plus faible pour les faux positifs que pour les faux négatifs, ces derniers étant les erreurs que nous devons éviter en priorité. Il est plausible que certaines prédictions incorrectes résultent de combinaisons rares et complexes de caractéristiques propres au jeu de données, ce qui suggère l'intérêt d'estimer l'incertitude aléatoire en complément des prédictions. Par ailleurs, l'utilisation d'un jeu de données plus large, avec des caractéristiques supplémentaires, pourrait améliorer la qualité des prédictions tout en réduisant l'incertitude globale. Malgré plusieurs tentatives avec différents réseaux de neurones traditionnels, les scores similaires obtenus indiquent que le jeu de données a probablement atteint sa capacité prédictive maximale. Idéalement, nous viserions une AUSE encore plus faible, en particulier pour les premières étapes. Cependant, compte tenu des informations limitées disponibles à ces étapes, une calibration parfaite semble difficile à atteindre. Néanmoins, le modèle demeure utilisable en pratique et démontre son utilité malgré ces défis. Même si l'on constate une légère perte de performance par rapport au bi-LSTM, celle-ci reste suffisamment minimale pour être acceptable, et l'ajout de l'incertitude constitue un gain significatif, justifiant largement ce compromis.

4 Conclusion

Nous présentons un Réseau de Neurones Bayésien pour prédire la résistance aux antibiotiques, en exploitant les métadonnées des patients ainsi que les informations bacteriologiques. Notre modèle est capable de générer des prédictions de résistance à chaque étape du processus typique d'identification des bacteries. La nature bayésienne du modèle permet une quantification de l'incertitude qui, après une évaluation minutieuse, s'avère bien calibrée. Nous estimons que cette mesure d'incertitude constitue une valeur ajoutée significative, en particulier dans le contexte de son utilisation prévue en pratique médicale. De plus, le modèle a été intégré dans une application web, permettant aux professionnels de le tester et de l'évaluer.

5 Remerciements

Ce travail a été publié lors d'AIME 2024 : 22nd International Conference of AI in Medicine, à Salt Lake City. Ce projet a été soutenu par une subvention cofinancée par l'Institut des Sciences de la Santé Publique Aix-Marseille (ISSPAM) et l'Institut de Recherche pour le Développement (IRD). Les auteurs remercient les bactériologistes des laboratoires de l'Hôpital Européen et de l'Hôpital Saint-Joseph pour avoir réalisé les tests de sensibilité aux antimicrobiens et extrait le jeu de données des antibiogrammes.

Références

Begoli, E., T. Bhattacharya, et D. Kusnezov (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence 1*(1), 20–23.

Blei, D. M., A. Kucukelbir, et J. D. McAuliffe (2017). Variational inference : A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Ilg, E., Özgün Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, et T. Brox (2018). Uncertainty estimates and multi-hypotheses networks for optical flow.

Jospin, L. V., H. Laga, F. Boussaid, W. Buntine, et M. Bennamoun (2022). Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine 17*(2), 29–48.

Kompa, B., J. Snoek, et A. L. Beam (2021). Second opinion needed : communicating uncertainty in medical machine learning. *npj Digital Medicine 4*(1).

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, et A. Talwalkar (2018). Hyperband : A novel bandit-based approach to hyperparameter optimization.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation 4*(3), 448–472.

Neal, R. M. (2012). *Bayesian learning for neural networks*, Volume 118. Springer Science & Business Media.

Qinghui Yu, J., E. Creager, D. Duvenaud, et J. Bettencourt. Bayesian neural networks.

Prédiction de l'antibiorésistance à l'aide de BNN.

Ren, Y., T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, A.-C. Hauschild, O. Schwengers, et D. Heider (2021). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics 38*(2), 325–334.

Urena, R., S. Camiade, Y. Baalla, M. Piarroux, L. Vouriot, P. Halfon, J. Gaudart, J.-C. Dufour, et S. Rebaudet (2024). Proof of concept study on early forecasting of antimicrobial resistance in hospitalized patients using machine learning and simple bacterial ecology data. *Scientific Reports 14*(1), 22683.

World Health Organization (2020). Antibiotic resistance. Last accessed 30/05/2023.

Yelin, I., O. Snitser, G. Novich, R. Katz, O. Tal, M. Parizade, G. Chodick, G. Koren, V. Shalev, et R. Kishony (2019). Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature Medicine 25*(7), 1143–1152.

Summary

Antimicrobial resistance is recognized by the World Health Organization as a significant global health threat. The accurate identification of bacterial susceptibility to antibiotics is crucial, but it often takes several days. On the other hand, in medical decision support systems, such as the one proposed in this contribution, it is crucial to assess the uncertainty of the model when a decision is provided. In this work, we propose a model based on a Bayesian Neural Network to predict antibiotic resistance at different stages of the antibiogram process for a set of 47 antibiotic therapies. Excellent results were achieved, with the area under the receiver operating curve reaching up to 0.9 at the final stage, while also providing a measure of the epistemic uncertainty. To enable clinical usage of the proposed approach as a decision support system, the model has been integrated into a user-friendly and responsive web application accessible on both mobile phones and desktops.

Pipeline d'Aide à la Découverte et l'Utilisation de Données Ouvertes basées sur les LLM

Antoine Dupuy*, Nathalie Aussenac-Gilles*
Christophe Baehr**,** Cassia Trojahn***

*IRIT, Université de Toulouse, UT2, CNRS, Toulouse, France
firstname.lastname@irit.fr,
https://www.irit.fr/
**CNRM UMR-3589, Université de Toulouse, Météo-France, CNRS, Toulouse, France
christophe.baehr@meteo.fr
https://www.umr-cnrm.fr/
***Université Grenoble Alpes
cassia.trojahn-dos-santos@univ-grenoble-alpes.fr

Résumé. L'utilisation des grands modèles de langues (LLM) et d'agents conversationnels pour la réalisation de tâches spécifiques telles que la programmation, la recherche d'information dans des documents, les systèmes de questions/réponses et les systèmes de recommandations, permettent des améliorations significatives dans la réponse aux besoins des utilisateurs. Dans le domaine de la découverte de données ouvertes, en particulier, la compréhension des besoins des utilisateurs finaux est un défi majeur. Des travaux proposent notamment des systèmes de génération augmentée par récupération (RAG) impliquant des ontologies associées à des agents conversationnels pour améliorer la qualité des réponses aux requêtes utilisateur, données par un LLM. Dans cette optique, cet article propose un pipeline d'aide à la découverte de jeux de données basés sur leurs métadonnées, décrites par les ontologies, dont des métadonnées sur utilisation par d'autres utilisateurs. Ce pipeline s'appuie sur des agents conversationnels basés sur le large modèle de langage Llama 3.1 70B et s'appuyant sur une base de connaissance ontologique, DATA-FW. Les résultats sont prometteurs mais de nouveaux travaux doivent être réalisés pour améliorer le système, notamment sur l'extraction de données depuis les plateformes publiques.

1 Introduction

Combler l'écart entre les besoins en données des utilisateurs finaux et les données partagées par les producteurs, qui les génèrent principalement pour leurs propres usages, constitue un défi majeur dans de nombreux domaines. Prenons l'exemple des données météorologiques : le producteur Météo France utilise ses propres données pour prévoir le temps, étudier le changement climatique, analyser l'environnement et créer des produits pour divers secteurs tels que l'agriculture, l'aviation, le ferroviaire et la santé. Cependant, les utilisateurs finaux peuvent avoir des besoins spécifiques et une compréhension particulière des données, comme une entreprise

utilisant des grues à Toulouse, souhaitant connaître le nombre de jours dans l'année où la vitesse du vent est trop élevée pour manœuvrer les grues. Cette divergence entre les besoins des producteurs et ceux des utilisateurs soulève une question clé : comment aligner efficacement les besoins en données des utilisateurs finaux avec les informations issues du vocabulaire des producteurs ?

Les données produites par ces producteurs sont généralement mises à disposition sous forme de volumes importants de données ouvertes sur le web. Elles peuvent être consultées sous des licences ouvertes via différents portails, tels que les portails gouvernementaux pour les données publiques (par exemple, data.gouv en France [1] ou data.gov [2] aux États-Unis, des portails européens comme le Portail Européen des Données [3]), des portails de services publics (par exemple, la Bibliothèque nationale de France [4]), ou encore des portails de données scientifiques comme Copernicus [5] pour les sciences de la Terre. Cependant, ces données sont souvent publiées avec des métadonnées insuffisantes, ce qui complique la tâche d'identifier les jeux de données adaptés aux besoins des utilisateurs, comme indiqué par Ahmad et al..

Pour répondre à cet objectif, nous proposons un système composé de quatre agents conversationnels basés sur le LLM Llama 3.1 70B et sur l'ontologie de métadonnées DATA-FW [6] qui sert de base de connaissances pour représenter les métadonnées des jeux de données ainsi que sur les informations des portails de données ouvertes (Dupuy et al.).

La suite de l'article est organisée comme suit. La Section 2 introduit un exemple illustratif de recherche de données, en prenant le cas spécifique d'une entreprise de construction utilisant des grues pour ses activités. La Section 3 spécifie l'implémentation technique du système. La Section 4 indique les différentes étapes de création du système et le pipeline de données. La Section 5 présente les premiers résultats du système. La Section 6 présente les outils et travaux réalisés utilisant le framework RAG pour répondre aux demandes d'utilisateurs dans divers domaines. Enfin, la Section 7 résume les contributions de l'article et discute des perspectives pour les travaux futurs.

2 Exemple d'illustration

Prenons l'exemple d'une entreprise de construction exerçant son activité dans la zone de Toulouse Métropole et recherchant des données sur la force du vent pour calculer le nombre de jours où les grues de l'entreprise ne seront pas utilisables. Dans ce cas, l'utilisateur peut formuler une requête en langage naturel, telle que : "Je recherche des données d'observation fiables de force du vent pour savoir combien de jours dans l'année l'activité de mes grues sera interrompue."

Notre système commence par analyser cette requête à l'aide du LLM pour en extraire les critères essentiels : le domaine (météorologie), le paramètre recherché (la force ou la vitesse du vent), la qualité (les certifications), et l'usage (projets scientifiques, usages similaires). Il peut demander des informations complémentaires à l'utilisateur, comme par exemple la loca-

1. https://www.data.gouv.fr/fr/
2. https://www.data.gov/
3. https://ec.europa.eu/info/statistics/eu-open-data-portal_en
4. https://data.bnf.fr/
5. https://www.copernicus.eu/en/access-data
6. DATA-FW est disponible ici : https://w3id.org/data-fw

lisation des données recherchées ou la période qu'il souhaite analyser. Ensuite, il explore les métadonnées descriptives pour trouver des jeux de données pertinents en cherchant les critères sélectionnés dans la requête utilisateur et filtre les résultats en se basant sur les métadonnées de qualité et d'usage. Cet exemple montre comment l'intégration d'une ontologie et d'agents conversationnels peut transformer la recherche de données.

Dans la Section 3, nous proposons une implémentation technique combinant l'utilisation de l'ontologie DATA-FW et un agent conversationnel à base de grand modèle de langue (LLM).

3 Implementation Technique

La génération augmentée par récupération (RAG) est un framework d'intelligence artificielle qui améliore les capacités des grands modèles de langue en intégrant des sources de connaissances externes (Wiratunga et al., Gao et al., Jeong et al., Hu et Lu, Zhao et al., Alaofi et al., Edwards). Le fonctionnement de RAG repose sur la récupération d'informations pertinentes à partir d'une base de connaissances, que l'on utilise ensuite pour enrichir l'entrée du modèle de langage, afin que le modèle génère des réponses plus précises, actualisées et contextuellement pertinentes. Cette approche permet de dépasser des limites des LLM telles que le risque d'hallucinations dans les résultats. Notre implémentation repose sur (i) l'ontologie DATA-FW en tant que base de connaissances qui modélise les métadonnées des jeux de données à l'aide de l'ontologie DATA-FW, permettant de modéliser les métadonnées des jeux de données et (ii) quatre agents conversationnels basés sur un LLM qui captent et affinent la requête utilisateur (**?**), la compare aux métadonnées représentées dans l'ontologie, récupèrent les données sur les plateformes publiques et construisent une réponse composée des données et d'explications sur le choix des données effectuées par les agents.

3.1 Architecture Proposée

Quatre agents interagissent pour répondre à la requête utilisateur et fournir des données pertinentes dans son cas d'utilisation. Chaque agent a un rôle spécifique dans le traitement de la requête (Cheng et al.).

Un agent d'interaction utilisateur analyse la requête initiale de l'utilisateur et identifie les informations manquantes, comme des précisions sur la localisation, le format des données (CSV, JSON, GeoJSON, etc.), les critères de qualité ou l'objectif d'utilisation des données. Si ces éléments ne sont pas présents dans la requête, l'agent engage une conversation pour les demander explicitement à l'utilisateur, assurant ainsi une recherche plus ciblée et pertinente.

Un agent de génération de requêtes SPARQL traduit les besoins exprimés par l'utilisateur en requêtes SPARQL en testant les valeurs des critères sélectionnés dans la requête utilisateur sur des triplets présents dans la base de connaissance basée sur l'ontologie DATA-FW. Ces requêtes permettent d'interroger les métadonnées des jeux de données pour identifier celles qui correspondent aux critères de l'utilisateur.

Un agent d'extraction des données se charge d'interagir avec les services associés, tels que des API ou des connecteurs externes, référencés dans l'ontologie. Son rôle est de récupérer uniquement les données nécessaires pour répondre à la demande de l'utilisateur, réduisant ainsi les volumes inutiles et optimisant les ressources.

Enfin, un agent de construction de réponse assemble les résultats sous une forme compréhensible et informative pour l'utilisateur. Il explique les choix effectués par le système, en justifiant pourquoi certains jeux de données ou données spécifiques ont été recommandés. Cette transparence renforce la confiance de l'utilisateur dans les réponses fournies.

Ces interactions entre l'utilisateur, les agents et l'ontologie DATA-FW sont illustrés en Figure 1.

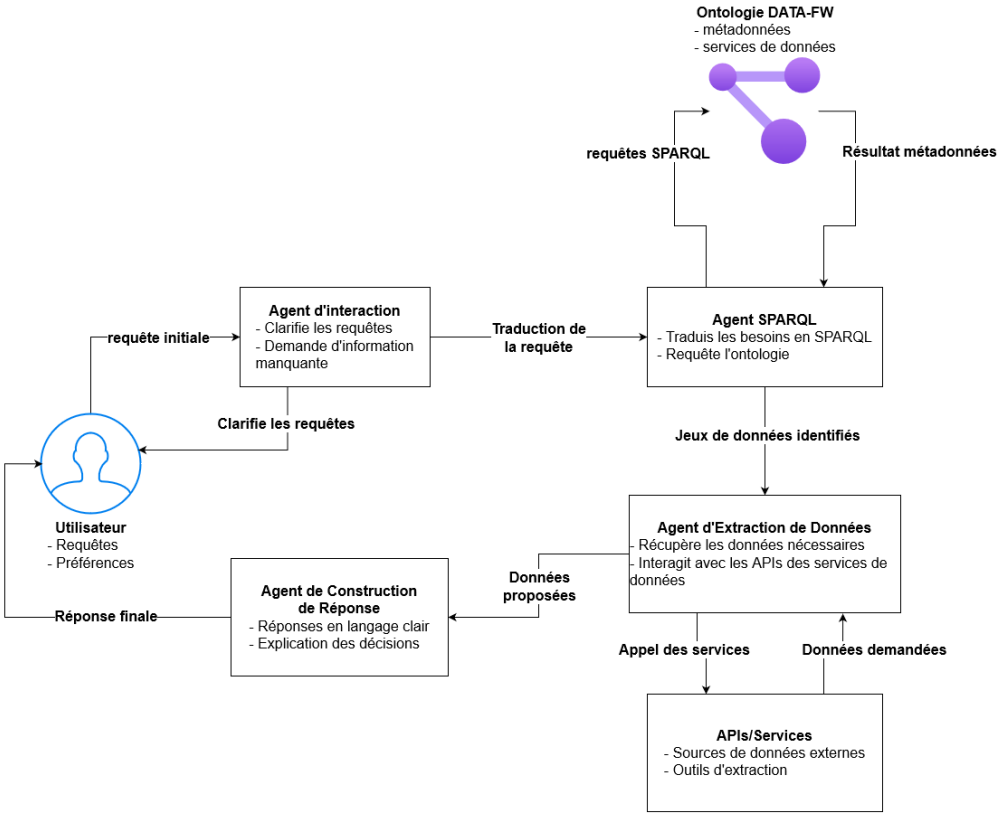


FIG. 1 – *Interactions entre l'utilisateur, les agents conversationnels et l'ontologie DATA-FW.*

3.2 Ontologie DATA-FW

L'ontologie DATA-FW, proposée par Dupuy et al., a été modélisée en utilisant Web Ontology Language (OWL). Elle inclut des classes et des propriétés représentant quatre dimensions de métadonnées : les métadonnées de base (descriptive, de provenance, de droits d'accès et d'historique de version), structurelles, qualitatives et d'usage, ainsi que des relations entre ces classes (par exemple, un jeu de données est lié à son producteur ou à des outils compatibles) (Annane et al., Barbosa et al.). Pour représenter ces dimensions, elle réutilise plusieurs vocabulaires et ontologies pour la description des métadonnées des jeux de données et leur

réutilisation : Data Catalog Vocabulary (DCAT 3), RDF Data Cube, Data Quality Vocabulary (DQV), Dataset Usage Vocabulary (DUV) et Friend Of A Friend (FOAF). Les métadonnées représentées par l'ontologie DATA-FW ne sont pas spécifiques à un domaine d'étude. Elle peut être étendue par des ontologies de domaine, comme par exemple l'ontologie Semantic Sensor Network (SSN[7]) dans le cadre de données d'observations de la Terre via la représentation des capteurs et des observations (Haller et al. (2018)).

Dans la Section 4, nous expliquons le pipeline de recherche de données, en détaillant les étapes qui permettent la réalisation de l'outil de recherche de données.

4 Pipeline

Le pipeline du système proposé se déroule en plusieurs étapes clés : l'intégration des jeux de données dans l'ontologie DATA-FW, l'interaction de l'utilisateur avec des agents basés sur le grand modèle de langage Llama 3.1 70B, la recherche et le filtrage des jeux de données à partir de la requête de l'utilisateur et la génération de réponses, comme illustré par la Figure 2.

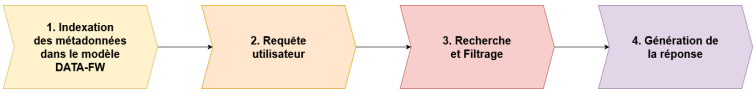


FIG. 2 – *Etapes Clés du Pipeline.*

4.1 Intégration des Jeux de Données

L'intégration des métadonnées des jeux de données est la première étape du pipeline d'aide à la découverte de jeux de données. Lors de cette étape, les informations relatives aux jeux de données seront extraites de plateformes publiques via les interfaces de programmation d'application (APIs). Le catalogue de chaque plateforme est extrait dans un fichier JSON, comme illustré sur la Figure 3.

Les fichiers JSON sont ensuite traités par un script Python pour structurer et insérer les métadonnées selon le modèle de l'ontologie DATA-FW, que nous présentons dans la Section 3.2.

4.2 Interaction Utilisateur

L'utilisateur interagit avec l'outil en formulant ses besoins en langage naturel. Cette requête peut être raffinée pour obtenir des informations complémentaires. Cette interaction entre l'utilisateur et l'outil permet de construire une requête regroupant les différents besoins de l'utilisateur pour les comparer aux métadonnées structurés dans l'ontologie DATA-FW.

4.3 Recherche et Filtrage

Les agents conversationnels du prototype, basés sur le modèle Llama 3.1, interprètent cette requête pour en identifier les critères principaux, recherchent dans l'ontologie les métadonnées des jeux de données qui correspondent au besoin de l'utilisateur, interroge les services

7. https://w3c.github.io/sdw-sosa-ssn/ssn/

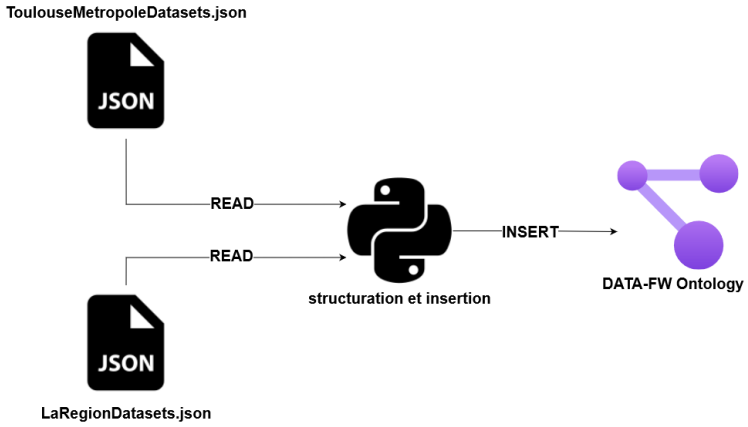


FIG. 3 – *Extraction des Métadonnées des Jeux de Données des Plateformes de Toulouse Métropole et de la Région Occitanie*

de données pour trouver les jeux de données correspondant. Les résultats sont filtrés et classés en fonction de paramètres spécifiques, par exemple si il y a un besoin exprimé sur des certifications de qualité.

4.4 Génération de Réponses

L'agent de construction de réponses génère une réponse en langage clair, expliquant les choix effectués et proposant des recommandations personnalisées. Ce processus a pour but que chaque utilisateur reçoit des suggestions adaptées à ses attentes, tout en offrant des options pour ajuster les critères ou explorer de nouvelles pistes.

5 Résultats Initiaux

Les résultats initiaux obtenus sont prometteurs. L'utilisation de l'ontologie s'avère efficace pour restreindre la recherche aux données pertinentes contenues dans celle-ci, réduisant ainsi les risques d'hallucinations du système. Les données utilisées pour le prototype sont celles des plateformes de Toulouse Métropole [8] et de La Région Occitanie [9]. Elles ont été récupérées via les APIs de ces plateformes [10] [11]. L'agent d'interaction utilisateur contribue à une meilleure spécification des requêtes en guidant l'utilisateur pour compléter les informations manquantes (Chapman et al., Gwizdka, Lafia et al.), ce qui permet de filtrer les résultats et de produire des requêtes SPARQL plus ciblées vers l'ontologie. Par ailleurs, l'agent de construction de réponse fournit des réponses détaillées tout en justifiant les choix effectués par le système. Ces explications permettent à l'utilisateur de mieux comprendre le fonctionnement de l'outil

8. https://data.toulouse-metropole.fr
9. https://data.laregion.fr/pages/accueil/
10. https://data.toulouse-metropole.fr/api/v1/console/datasets/1.0/search/
11. https://data.laregion.fr/api/explore/v2.1/console

et obtenir une meilleure visualisation des données proposées, facilitant ainsi ses recherches futures (Peña et al., Wu et al.). Cependant, des limites ont été identifiées, notamment avec l'agent de construction de requête SPARQL dont les requêtes SPARQL renvoient des résultats erronés et l'agent d'extraction des données, qui tend à proposer des résultats non pertinents et nécessite une révision de ses instructions. Par exemple, lors d'un test, un utilisateur demandant des données sur la force du vent a reçu des suggestions portant sur des données relatives aux musées parisiens, illustrant un problème de cohérence entre la requête et les données extraites dû à un exemple de requête SPARQL sur la base de connaissance Wikidata de l'agent de construction de requêtes SPARQL.

6 Travaux Liés

Plusieurs systèmes de découverte de données ont été développés ces dernières années, comme Google Dataset SearchBrickley et al.Benjelloun et al.Sostek et al.. Ces solutions se concentrent principalement sur les métadonnées descriptives, ce qui limite leur capacité à fournir des recommandations adaptées aux contextes d'usage spécifiques. Notre approche enrichit cette base en intégrant des dimensions comme la qualité des données, leur structure, et les usages antérieurs.

En parallèle, des recherches récentes sur les modèles de langage, notamment dans le cadre des architectures RAG, ont montré leur potentiel pour la génération de réponses complexes et adaptées au contexte de l'utilisateur (Li et al.). Il existe également des travaux dans divers domaines, comme la recherche d'information dans des documents et des bases de données, notamment dans le domaine du droit (Lála et al., Wiratunga et al., Yang et al. (b)), dans la recherche de littérature scientifique (Zhang et Kotanko), la programmation informatique (Yang et al. (a)) ou dans l'éducation et le développement des activités commerciales (Posedaru et al., Alqahtani et al.), où des systèmes basés sur des agents conversationnels et des modèles de langage sont utilisés pour extraire des informations pertinentes et répondre à des requêtes complexes.

7 Conclusion et Futurs Travaux

Ce système basé sur une architecture RAG offre une solution pour la découverte de jeux de données. En s'appuyant sur l'ontologie DATA-FW et les agents conversationnels alimentés par Llama 3.1, il permet une recherche simplifiée par rapport aux plateformes de données traditionnelles, par le biais d'une recherche en langage naturel, une personnalisation accrue des recommandations par l'intégration des paramètres de qualité et d'usage, et une meilleure compréhension des besoins des utilisateurs par l'intégration d'agents d'intelligence artificielle pour préciser la demande de l'utilisateur.

Cependant, plusieurs axes d'amélioration et de développement sont envisagés. Un benchmark entre différents modèles de langage open source permettra de comparer leurs performances pour optimiser les recommandations. Une collaboration avec la Région Occitanie permettra de définir des cas d'utilisation concrets pour adapter le système à des besoins territoriaux et tester son efficacité en situation réelle. L'intégration de nouvelles plateformes sectorielles ou internationales, comme Eurostat ou des catalogues spécialisés, viendra enrichir le catalogue de

données disponibles. Par ailleurs, les limites actuelles, telles que la dépendance à la qualité des métadonnées initiales et les biais potentiels des modèles de langage, nécessitent une attention particulière pour améliorer la fiabilité des recommandations. Enfin, l'ajout d'agents spécialisés, comme ceux dédiés à la génération de code ou à la recherche d'informations contextuelles sur les données, renforcera les fonctionnalités et l'utilité du système dans divers scénarios.

Références

Ahmad, R. A., J. D'Souza, M. Zloch, W. Otto, G. Rehm, A. Oelen, S. Dietze, et S. Auer. Toward FAIR semantic publishing of research dataset metadata in the open research knowledge graph.

Alaofi, M., N. Arabzadeh, C. L. A. Clarke, et M. Sanderson. Generative information retrieval evaluation. Version Number : 2.

Alqahtani, T., H. A. Badreldin, M. Alrashed, A. I. Alshaya, S. S. Alghamdi, K. Bin Saleh, S. A. Alowais, O. A. Alshaya, I. Rahman, M. S. Al Yami, et A. M. Albekairy. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *19*(8), 1236–1242.

Annane, A., M. Kamel, C. Trojahn, N. Aussenac-Gilles, C. Comparot, et C. Baehr. SYNOP data evaluation using FAIR maturity model.

Barbosa, L., K. Pham, C. Silva, M. R. Vieira, et J. Freire. Structured open urban data : Understanding the landscape. *2*(3), 144–154.

Benjelloun, O., S. Chen, et N. Noy. Google dataset search by the numbers. Version Number : 1.

Brickley, D., M. Burgess, et N. Noy. Google dataset search : Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pp. 1365–1375. ACM.

Chapman, A., E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, et P. Groth. Dataset search : a survey. *29*(1), 251–272.

Cheng, Y., C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, et X. He. Exploring large language model based intelligent agents : Definitions, methods, and prospects. Version Number : 1.

Dupuy, A., C. Trojahn, N. Aussenac-Gilles, et C. Baehr. Data-fw : An ontology network for annotating open datasets. ACM.

Edwards, C. Hybrid context retrieval augmented generation pipeline : LLM-augmented knowledge graphs and vector database for accreditation reporting assistance. Version Number : 1.

Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, et H. Wang. Retrieval-augmented generation for large language models : A survey. Version Number : 5.

Gwizdka, J. *Direct, Orienting, and Scenic Paths : How Users Navigate Search in a Research Data Archive*. ACM Conferences. Association for Computing Machinery.

Haller, A., K. Janowicz, S. J. Cox, M. Lefrançois, K. Taylor, D. Le Phuoc, J. Lieberman, R. García-Castro, R. Atkinson, et C. Stadler (2018). The Modular SSN Ontology : A Joint

W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web – Interoperability, Usability, Applicability 10*(1), 9–32.

Hu, Y. et Y. Lu. RAG and RAU : A survey on retrieval-augmented language model in natural language processing. Version Number : 1.

Jeong, S., J. Baek, S. Cho, S. J. Hwang, et J. C. Park. Adaptive-RAG : Learning to adapt retrieval-augmented large language models through question complexity. Version Number : 2.

Lafia, S., A. Million, et L. Hemphill. Exploratory and directed search strategies at a social science data archive. *48*(1).

Li, X., K. Lv, H. Yan, T. Lin, W. Zhu, Y. Ni, G. Xie, X. Wang, et X. Qiu. Unified demonstration retriever for in-context learning. Version Number : 2.

Lála, J., O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, et A. D. White. PaperQA : Retrieval-augmented generative agent for scientific research. Version Number : 2.

Peña, O., U. Aguilera, et D. López-de Ipiña. Linked open data visualization revisited : A survey.

Posedaru, B.-S., F.-V. Pantelimon, M.-N. Dulgheru, et T.-M. Georgescu. Artificial intelligence text processing using retrieval-augmented generation : Applications in business and education fields. *18*(1), 209–222.

Sostek, K., D. M. Russell, N. Goyal, T. Alrashed, S. Dugall, et N. Noy. Discovering datasets on the web scale : Challenges and recommendations for google dataset search.

Wiratunga, N., R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, et B. Fleisch. CBR-RAG : Case-based reasoning for retrieval augmented generation in LLMs for legal question answering. Version Number : 1.

Wu, M., F. Psomopoulos, S. J. Khalsa, et A. De Waard. Data discovery paradigms : User requirements and recommendations for data repositories. *18*, 3.

Yang, K., J. Liu, J. Wu, C. Yang, Y. R. Fung, S. Li, Z. Huang, X. Cao, X. Wang, Y. Wang, H. Ji, et C. Zhai. If LLM is the wizard, then code is the wand : A survey on how code empowers large language models to serve as intelligent agents. Version Number : 2.

Yang, X., Z. Wang, Q. Wang, K. Wei, K. Zhang, et J. Shi. Large language models for automated q&a involving legal documents : a survey on algorithms, frameworks and applications. *20*(4), 413–435.

Zhang, H. et P. Kotanko. #1506 uremic toxicity : gaining novel insights through AI-driven literature review. *39*, gfae069–0657–1506.

Zhao, P., H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, et B. Cui. Retrieval-augmented generation for AI-generated content : A survey. Version Number : 6.

Summary

The use of Large Language Models (LLMs) and conversational agents for specific tasks such as programming, information retrieval from documents, question-and-answer systems, and recommendation systems enable significant improvements in meeting user needs. In the

Pipeline d'Aide à la Découverte et l'Utilisation de Données Ouvertes basées sur les LLM

field of data search, understanding end-user needs is a major challenge. Some works propose Retrieval-Augmented Generation (RAG) systems that incorporate knowledge graphs, such as ontologies, combined with conversational agents powered by LLMs to enhance the quality of responses to user queries. In this context, our approach introduces a pipeline to facilitate dataset discovery based on dataset metadata, described by the ontologies, and their usage by other users associated with conversational agents from the large language model Llama 3.1 70B supported by an ontological knowledge base, DATA-FW. The results are promising, but further work is needed to improve the system, particularly in data extraction from public platforms.

Détection d'anomalies et assurance qualité pour les diagnostics du Tokamak WEST

Feda Almuhisen * Mathias Couraud* Paulo Puglia* Jorge Morales* Rémi Dumont* Dorian Midou*, et l'équipe WEST**

*CEA, IRFM, Saint-Paul-lez-Durance, F-13108, France
** http://west.cea.fr/WESTteam

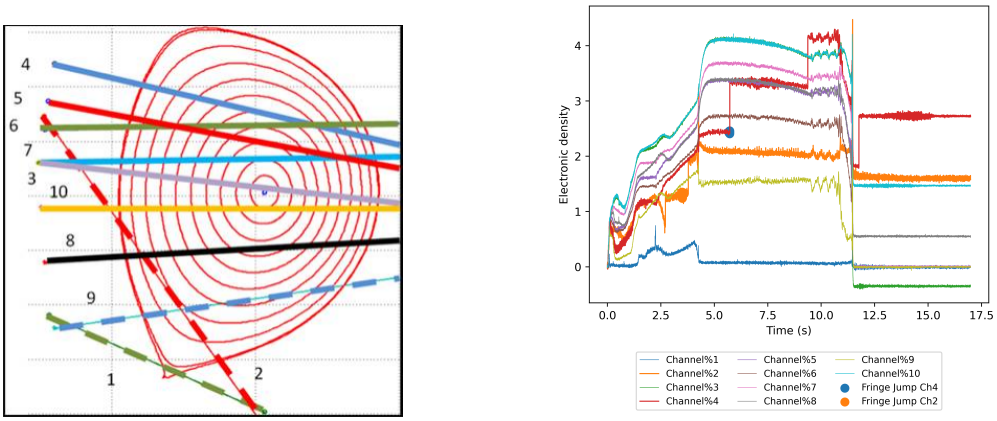
1 Introduction

1.1 Contexte du Tokamak WEST

Les tokamaks sont des dispositifs dédiés à la recherche sur la fusion nucléaire, où un plasma chaud est maintenu. Ils sont équipés de divers systèmes de diagnostic permettant de surveiller le comportement du plasma. WEST, « Tungsten Environment in Steady-state Tokamak » (où « W » désigne le tungstène), est consacré au soutien des opérations ITER et à l'exploration de décharges plasma de longue durée dans un environnement entièrement composé de tungstène (Bucalossi et al., 2014). Il réalise de nombreuses décharges plasma surveillées par plus de 50 systèmes de diagnostic. Cependant, la gestion de la qualité des données représente un défi majeur en raison de la complexité et du volume des informations collectées. Ces difficultés peuvent entraîner des interprétations erronées du comportement du plasma. C'est pourquoi la mise en place d'une chaîne automatisée dédiée à l'assurance qualité des données et à la détection des anomalies est cruciale pour garantir la fiabilité des données utilisées par les chercheurs.

1.2 Diagnostics d'interférométrie plasma

La caractérisation du plasma dans les tokamaks repose sur des mesures précises de la densité et de de la température électroniques. L'une des méthodes les plus courantes pour mesurer la densité électronique est l'interférométrie, qui se base sur le fait que l'indice de réfraction du plasma est proportionnel à la densité des électrons. Cette technique utilise deux faisceaux laser : l'un traverse le plasma tandis que l'autre sert de référence. La différence de phase entre les deux faisceaux, causée par le plasma, permet de mesurer la densité électronique du plasma. Le laser utilisé opère dans la bande de fréquence des infrarouges lointains (Gil et al., 2019). Au sein du tokamak WEST, ce dispositif de diagnostic comprend huit canaux équatoriaux et deux canaux semi-verticaux dédiés aux mesures en bord de plasma. Un exemple des lignes de visée interférométriques et des signaux correspondants est illustré dans la figure 1.



(a) Lignes de visée à l'intérieur de WEST (b) Signaux mesurés pour la décharge 59703

FIG. 1 – Lignes de visée interférométriques et signaux correspondants

2 Sauts de franges dans l'interférométrie Tokamak

Lors du processus de mesure de la différence de phase, une erreur de saut de frange peut conduire à des valeurs de phase et de densité incorrectes. Cela se produit lorsque la différence de phase $\Delta\phi(t)$ entre le faisceau de référence et un faisceau de mesure est, à tort, augmentée ou diminuée d'un entier N de 2π. Cette erreur peut résulter de variations brusques dans la densité électronique n_e (pour plus de détails, voir (Zabeo et al., 2004)). La différence totale de phase $\Delta\phi(t)$ se définit comme suit : $\Delta\phi(t) = k \int_L n_e(l,t)\, dl$ où : $k = \frac{2\pi}{\lambda}$ est le nombre d'onde du laser (avec λ comme longueur d'onde), $n_e(l,t)$ est la densité électronique intégrée le long de la ligne, dans le cas du plasma, en fonction de la position l parcourue par le faisceau sur le trajet L, quant à L il représente le chemin optique du faisceau. Dans un environnement expérimental, des changements soudains dans les gradients de densité (causés, par exemple, par des disruptions), du bruit ou des pertes de données peuvent entraîner un échec de la résolution de la différence de phase. Si la différence de phase dépasse un multiple de 2π,le système de diagnostic risque de surestimer ou de sous-estimer la valeur réelle, ce qui se traduit par une perte de données. On décrit cette erreur par l'équation suivante : $\Delta\phi_{\text{mesurée}}(t) = \Delta\phi_{\text{vraie}}(t) + 2\pi \cdot N$ où $N \in \mathbb{Z}$, avec $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ est un entier correspondant au nombre de sauts de phase de 2π omis ou ajoutés, entraînant l'erreur de saut de frange. La figure 2 montre plusieurs sauts de franges, qui ont une caractéristique de point unique, instantané (marqué en rouge) dans la décharge 59703, survenant autour de 2.5 s et 4s, 4.5s. En revanche, les fausses anomalies en vert consistent en plusieurs points consécutifs et montrent une déviation suivie d'un retour au niveau précédent. Notons que les sauts de franges dus à des défaillances du diagnostic peuvent influencer le contrôle de la densité du plasma, l'exploitation et potentiellement la sécurité de la machine (Blanken et al., 2018). La détection des sauts de franges est différente des anomalies typiques des séries temporelles, car elle consiste à identifier des changements brusques et ponctuels de phase qui peuvent se produire alors que le signal reste dans sa plage normale et physiquement valide avant et après le saut. Cela nécessite de détecter des ruptures franches dans la dynamique du signal plutôt que de simples écarts d'amplitude, rendant inefficaces les

méthodes reposant exclusivement sur des seuils.

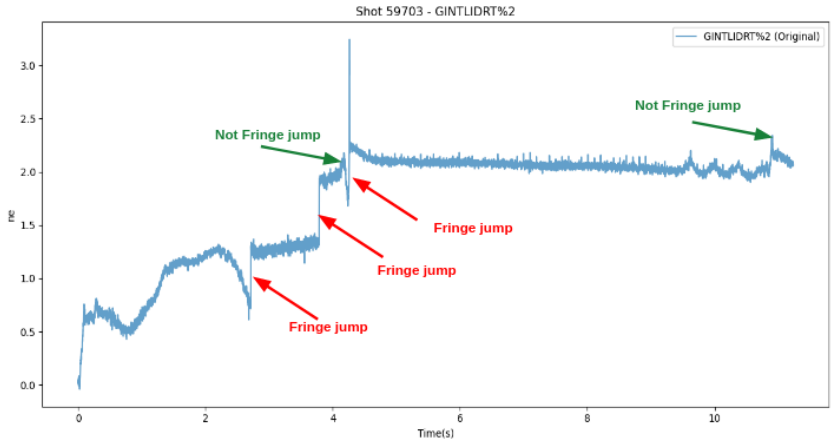


FIG. 2 – Sauts de franges observés lors de la décharge 59703

3 Détection d'anomalies pour les sauts de franges en interférométrie

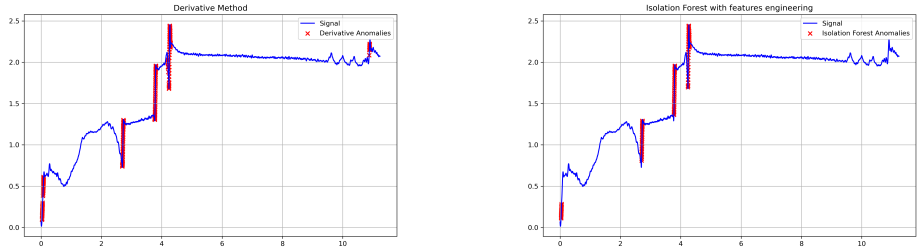
Les méthodes traditionnelles pour détecter et corriger les sauts de franges s'appuient souvent sur une modélisation physique approfondie ou sur des vérifications et corrections manuelles (Zabeo et al., 2004). Récemment,, les approches d'apprentissage automatique (ML) ont montré leur potentiel dans ce domaine. Par example, dans (Siddiqui et al., 2022), les auteurs utilisent des réseaux neuronaux profonds de détection d'objets pour classifier les franges produites par l'interférométrie à auto-mélange (SMI) basée sur la rétroaction laser. Dans cette étude, nous cherchons à développer un cadre d'apprentissage automatique exploitant des mesures expérimentales pour détecter automatiquement les sauts de franges et fournir des alertes de qualité de données pour chaque décharge. En outre, ce cadre permettra de constituer des jeux de données annotés par des experts, servant à l'entraînement de modèles ML avancés capables de détecter et de classer les causes de ces sauts (par exemple, des défaillances matérielles ou des phénomènes liés au plasma). Notre analyse porte sur 341 décharges issues de la campagne C9 du diagnostic d'interférométrie du tokamak WEST. Chaque décharge fournissait des mesures issues de dix canaux d'interférométrie (voir la figure 1), permettant d'enregistrer la densité électronique tout au long de la décharge du plasma, depuis différentes positions, à un taux d'échantillonnage de 1kHz.

3.1 Méthodes de détection d' anomalies

Dans notre ensemble de données, plusieurs difficultés se posent : l'absence de labels permettant d'indiquer précisément où se produisent les sauts de franges, ainsi que la présence de bruit et d'oscillations. Afin de remédier à ces problèmes, nous commençons par normaliser les signaux de densité afin de standardiser leur amplitude tout en préservant leur structure temporelle. Nous appliquons ensuite un filtre de Savitzky-Golay qui permet de lisser les signaux tout en

maintenant les transitions brusques, indispensables à la détection des sauts de franges. L'impact de ces opérations sur le contenu fréquentiel sera quantifié afin de minimiser la distorsion.

Approche par dérivée et seuils À titre de méthode de référence, nous avons implémenté une méthode de détection basée sur la dérivée afin d'identifier les sauts de franges. Cette approche consiste à calculer la dérivée des signaux prétraités, puis à relever les points où le taux de variation dépasse un seuil prédéfini. Bien qu'elle saisisse efficacement les changements dans le signal, elle se montre également sensible aux petites fluctuations et peut générer un nombre important de faux positifs, surtout en présence de bruit ou de signaux perturbés. Par exemple, comme l'illustre la figure 3a, des faux positifs apparaissent à la fin et au début du signal. Si cette méthode assure de ne manquer aucune transition significative, elle manque toutefois de robustesse dans certains scénarios.



(a) Sauts de franges détectés par la méthode des dérivées

(b) Sauts de franges détectés par Isolation Forest

FIG. 3 – Comparaison des méthodes de détection des sauts de franges

Approche par apprentissage automatique En alternative, nous avons appliqué une méthode non supervisée, Isolation Forest (Liu et al., 2008), associée à de l'ingénierie de caractéristiques. Nous avons entraîné des modèles distincts pour chaque canal, en utilisant les mêmes caractéristiques et hyperparamètres. Les caractéristiques exploitées sont issues des signaux de densité (par exemple, l'amplitude et certaines propriétés statistiques). En raison de l'absence de données annotées, nous avons procédé à une vérification manuelle d'un sous-ensemble d'anomalies détectées et comparé ces résultats avec la méthode de référence. Le modèle Isolation Forest s'est révélé plus robuste, tandis que la méthode par dérivée se montrait sensible aux variations mineures. Par exemple, comme illustré dans la figure 3b, Isolation Forest identifie avec succès les transitions brusques tout en réduisant les faux positifs, notamment en fin de décharge. Toutefois, quelques faux positifs subsistent en début de signal, comme pour la méthode de référence. Nous avons réalisé un réglage préliminaire des paramètres du modèle Isolation Forest sur un sous-ensemble de données. Les premiers essais montrent que les hyperparamètres sélectionnés demeurent relativement stables et que de légères variations n'affectent pas significativement la performance. Bien que les résultats initiaux indiquent une amélioration qualitative de l'utilisation d'Isolation Forest par rapport à la méthode par dérivée, nous reconnaissons la nécessité d'une évaluation quantitative. Les travaux futurs porteront sur la levée des limitations de cette étude. Nous prévoyons notamment de mettre en place un mécanisme permettant aux experts du domaine de réviser les anomalies détectées et de confirmer les sauts de franges, afin de

constituer des données annotées. De plus, nous comptons étendre notre comparaison à d'autres techniques de regroupement non supervisées (par exemple DBSCAN) et à des approches de deep learning (telles que les autoencodeurs). Nous évaluerons également la généralisation du modèle à l'aide de méthodes de validation croisée et en intégrant des données issues de plusieurs campagnes expérimentales. La figure 4 présente une architecture préliminaire du cadre que nous proposons pour l'assurance qualité des données, intégrant le prétraitement, la détection d'anomalies par IA et des boucles de rétroaction destinées à automatiser le contrôle qualité dans les diagnostics Tokamak.

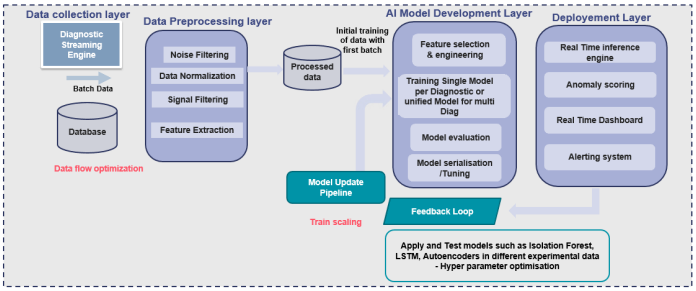


FIG. 4 – Architecture préliminaire du cadre proposé pour l'assurance qualité des données

Références

Blanken, T., F. Felici, C. Rapson, M. De Baar, W. Heemels, A.-U. team, et al. (2018). Control-oriented modeling of the plasma particle density in tokamaks and application to real-time density profile reconstruction. *Fusion Engineering and Design 126*, 87–103.

Bucalossi, J., M. Missirlian, P. Moreau, F. Samaille, E. Tsitrone, D. van Houtte, T. Batal, C. Bourdelle, M. Chantant, Y. Corre, X. Courtois, L. Delpech, L. Doceul, D. Douai, H. Dougnac, F. Faïsse, C. Fenzi, F. Ferlay, M. Firdaouss, L. Gargiulo, P. Garin, C. Gil, A. Grosman, D. Guilhem, J. Gunn, C. Hernandez, D. Keller, S. Larroque, F. Leroux, M. Lipa, P. Lotte, A. Martinez, O. Meyer, F. Micolon, P. Mollard, E. Nardon, R. Nouailletas, A. Pilia, M. Richou, S. Salasca, et J.-M. Travère (2014). The west project : Testing iter divertor high heat flux component technology in a steady state tokamak environment. *Fusion Engineering and Design 89*(7), 907–912. Proceedings of the 11th International Symposium on Fusion Nuclear Technology-11 (ISFNT-11) Barcelona, Spain, 15-20 September, 2013.

Gil, C., G. Colledani, M. Domenes, D. Volpe, A. Berne, F. Faisse, C. Guillon, J. Morales, P. Moreau, et B. Santraine (2019). Renewal of the interfero-polarimeter diagnostic for west. *Fusion Engineering and Design 140*, 81–91.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. IEEE.

Siddiqui, A. A., U. Zabit, et O. D. Bernal (2022). Fringe detection and displacement sensing for variable optical feedback-based self-mixing interferometry by using deep neural networks. *Sensors 22*(24), 9831.

Zabeo, L., A. Murari, P. Innocente, et al. (2004). Different approaches to real-time correction of fringe jumps in interferometers for nuclear fusion. *Review of Scientific Instruments 75*, 3881–3889.

Sélection de variables utilisant une nouvelle approche interactive d'apprentissage par renforcement profond

XX

XX

Abstract. La sélection des variables est essentielle pour optimiser les sous-ensembles de variables en apprentissage automatique. Les avancées récentes, telles que les méthodes interactives de sélection de variables avec apprentissage par renforcement, montrent un potentiel prometteur, mais se concentrent souvent sur des métriques uniques comme la précision, négligeant ainsi les objectifs contradictoires. Par ailleurs, s'appuyer sur un formateur unique limite l'efficacité en raison d'une guidance répétitive. Pour relever ces défis, cet article propose une approche d'apprentissage par renforcement profond interactif multi-agents et multi-objectifs. Cette méthode redéfinit la sélection de variables comme un problème multi-objectifs, intégrant un mécanisme de récompense novateur pour équilibrer l'aire sous la courbe (AUC) et le nombre de variables sélectionnées. Une stratégie de formateurs diversifiée est également proposée, exploitant plusieurs formateurs pour améliorer l'exploration des paramètres et réduire la redondance tout en encourageant l'auto-exploration des agents. La méthode fournit des solutions sous forme de Front de Pareto, offrant une flexibilité aux décideurs pour sélectionner les ensembles de variables optimaux. Les expériences montrent des performances supérieures par rapport aux méthodes de pointe, équilibrant efficacement l'AUC et le nombre de variables tout en améliorant l'efficacité et la robustesse de la sélection des variables.

1 Introduction

La sélection de variables (SV) consiste à identifier un sous-ensemble des variables les plus pertinentes à utiliser dans un modèle d'apprentissage automatique. Cette approche permet d'améliorer les performances du modèle en réduisant le surapprentissage et en renforçant sa capacité de généralisation. Récemment, des techniques avancées de sélection de variables basées sur l'apprentissage par renforcement (RL) Liu et al. (2019); Fan et al. (2020) ont suscité un intérêt croissant. Ces méthodes modélisent la sélection de variables comme un problème d'apprentissage par renforcement profond multi-agents (MADRL), où chaque variable est associée à un agent chargé de décider de son inclusion ou exclusion. Les approches MADRL visent généralement à optimiser les systèmes selon un objectif unique, aboutissant ainsi à une solution unique. Cependant, les problèmes de sélection de variables impliquent souvent des objectifs contradictoires, générant un ensemble de solutions viables appelée Front de Pareto (PF). Ces solutions satisfont tous les objectifs mais ne sont pas mutuellement comparables. À ce jour, aucune recherche n'a abordé le problème de la sélection de variables

renforcée comme un défi multi-objectifs. Dans ce travail, nous reformulons le problème en utilisant l'apprentissage par renforcement profond multi-objectifs pour répondre à deux objectifs conflictuels : minimiser le nombre de variables sélectionnées tout en maximisant le score AUC. Dans le cadre du MADRL, s'appuyer uniquement sur l'auto-exploration ne suffit pas pour garantir une exploration des variables à la fois efficace et performante. Récemment, l'apprentissage par renforcement interactif (IRL) a attiré l'attention en tant que méthode inspirée des processus d'apprentissage biologiques, tirant parti de connaissances externes pour accélérer l'exploration en apprentissage par renforcement Lin et al. (2020). Ce paradigme ouvre la voie à l'utilisation d'agents enseignants ou de guides pour une exploration et un apprentissage plus efficaces dans les tâches de sélection de variables. Par exemple, Fan et al. (2020) a introduit une approche IRL utilisant des connaissances externes via des formateurs. Cependant, leur stratégie applique des conseils uniquement pour un nombre spécifique d'itérations et repose sur un seul formateur, qui fournit souvent des recommandations similaires lorsqu'il rencontre des variables actives similaires. De plus, leur approche manque d'adaptabilité, car le rôle et les conseils du formateur restent statiques tout au long de la période définie.

Pour surmonter ces limitations, notre approche introduit une stratégie d'enseignement innovante qui intègre diversité et adaptabilité en s'appuyant sur plusieurs formateurs. À chaque itération, les conseils du meilleur formateur sont sélectionnés, garantissant que les agents reçoivent des orientations variées et efficaces tout au long du processus d'apprentissage. Cette approche dynamique améliore à la fois l'exploration et l'efficacité globale de l'apprentissage dans les tâches de sélection de variables. Dans cet article, nous présentons une méthode efficace nommée Apprentissage par Renforcement Profond Interactif Multi-Objectifs avec Agents Multiples (MOIDRL-MA) [1], spécifiquement conçue pour la sélection de variables. Nos contributions principales sont les suivantes : (1) Nous redéfinissons la sélection de variables pour fournir un éventail de solutions équilibrant la sélection et les performances de classification, offrant ainsi flexibilité et adaptabilité aux décideurs. (2) Une approche novatrice utilisant des formateurs diversifiés pour réduire les répétitions et améliorer l'efficacité de l'apprentissage. (3) MOIDRL-MA surpasse les méthodes RL classiques et les techniques traditionnelles, validant la pertinence de la stratégie multi-formateurs.

2 État de l'Art

De nombreuses études ont exploré l'application de l' RL pour la sélection de variables. Par exemple, les auteurs de Khurana et al. (2018) ont proposé une méthode basée sur le RL utilisant un graphe de transformation pour naviguer dans l'espace des variables, afin d'optimiser les performances en identifiant les variables essentielles. Également, l'approche RL présentée dans Fard et al. (2013) a introduit un critère basé sur la moyenne des récompenses pour évaluer les variables en fonction de leur influence sur les transitions d'état, associé à un algorithme itératif pour déterminer le meilleur sous-ensemble de variables en combinant les méthodes de sélection par filtre et par enveloppe. Dans une autre étude, Kim et al. (2022) a présenté une technique RL multi-agents où chaque variable est associée à un duo d'agents, principal et guide. Ces agents collaborent : les agents principaux mettent à jour les valeurs Q pour optimiser la sélection de variables, guidés par des critères fournis par les agents guides. Dans Liu

[1] Cette méthode a été acceptée pour présentation à la conférence internationale Neural Information Processing (ICONIP 2024).

et al. (2023), une approche RL mono-agent a été proposée, exploitant une stratégie interactive basée sur les récompenses et les niveaux d'entraînement pour améliorer l'efficacité grâce à des conseils externes. Les approches de l'RL profond, notamment celles employant le MADRL pour la sélection de variables, ont également suscité un intérêt croissant. Par exemple, Liu et al. (2019) a modélisé la sélection de variables comme un problème de RL profond, en assignant chaque variable à un agent. Ils ont exploré trois méthodes de représentation des états : descriptions statistiques, autoencodeurs et réseaux convolutifs de graphes (GCNs), pour enrichir le processus d'apprentissage. Ils ont également examiné des métriques visant à améliorer la coordination entre les variables en affinant les calculs de récompenses. S'appuyant sur ces travaux, Fan et al. (2020) a introduit un cadre MADRL avec une stratégie d'enseignement hybride inspirée de IRL. Cette approche utilisait différents formateurs à diverses étapes, offrant aux agents des orientations variées. Cependant, malgré son caractère novateur, la stratégie souffrait de redondance : un formateur unique pouvait fournir des conseils répétitifs, et elle manquait d'adaptabilité en raison de la durée fixe des phases d'enseignement. Les méthodes MADRL se concentrent généralement sur l'optimisation de systèmes à objectif unique, visant une solution unique. Bien que des approches d'apprentissage par renforcement profond multi-objectifs (MODRL) existent, elles n'ont pas encore été appliquées à la sélection de variables dans une perspective multi-objectifs. Les cadres MODRL associent chaque objectif à un signal de récompense distinct, formant un vecteur de récompenses au lieu d'un simple scalaire. Cela permet aux agents de prendre des décisions équilibrant plusieurs objectifs et d'obtenir des solutions optimales. Pour faire progresser l'état de l'art et répondre aux limites des méthodes existantes, en particulier Fan et al. (2020), nous proposons MOIDRL-MA. Notre approche intègre un apprentissage par renforcement profond interactif multi-objectifs avec agents multiples, adoptant un paradigme multi-politiques. Ce design permet aux agents d'opérer selon des politiques distinctes et de poursuivre des stratégies diversifiées, équilibrant efficacement les objectifs contradictoires tout en renforçant l'adaptabilité et l'efficacité de la sélection de variables.

3 Méthode Proposée

L'architecture du cadre MOIDRL-MA est illustrée dans la Figure 1. Initialement, chaque variable destinée à l'exploration est attribuée à un agent correspondant. Ces agents ont pour rôle de décider si leur variable associée doit être incluse dans le sous-ensemble final. Avant de finaliser ce sous-ensemble, qui constitue l'environnement, chaque agent reçoit des conseils de formateurs. Ces conseils permettent d'ajuster leurs décisions en fonction des actions réalisées lors des épisodes précédents (Section 3.1). Pour chaque formateur, un score est calculé sur la base des variables sélectionnées lors des itérations antérieures (agents actifs). Le formateur ayant obtenu le score le plus élevé est désigné comme conseiller et fournit des recommandations exclusivement aux agents indécis qui choisissent de ne pas sélectionner leurs variables. Le sous-ensemble final des variables sélectionnées constitue alors l'environnement dans lequel les agents interagissent. Au fur et à mesure que l'exploration de l'espace des sous-ensembles progresse, le nombre de variables sélectionnées évolue, entraînant des vecteurs de représentation des états de longueur variable. Pour résoudre ce problème, l'algorithme utilise une méthode de représentation des états basée sur des statistiques descriptives, introduite dans Liu et al. (2019), assurant une représentation cohérente des états quel que soit le nombre de

variables sélectionnées. Simultanément, les actions entreprises par les agents contribuent à une récompense cumulative, redistribuée entre les agents participants. Comme MOIDRL-MA fonctionne dans un cadre multi-objectifs, la récompense globale est évaluée en tenant compte à la fois du AUC et du nombre de variables sélectionnées pour la tâche aval (Section 3.2). Le cadre inclut également une étape de contrôle axée sur l'entraînement des agents, le stockage des expériences et l'accélération du processus d'apprentissage. Pour améliorer l'efficacité de l'entraînement, chaque agent maintient une unité de mémoire où il enregistre des tuples composés de l'état, de l'action, de la récompense et du prochain état après chaque épisode. Ces expériences stockées sont ensuite utilisées pour entraîner les politiques des agents. Chaque agent utilise son propre réseau Q profond (DQN) pour affiner ses stratégies en échantillonnant aléatoirement des mini-lots à partir de sa mémoire (Section 3.3). Grâce à des mises à jour itératives basées sur ces expériences, les agents améliorent progressivement leur processus de prise de décision, conduisant finalement à la convergence vers un espace optimal de variables finales. Les composantes clés de notre cadre MOIDRL-MA sont les suivantes : **Multi-agent :** Nous considérons un ensemble de N variables, chacune étant attribuée à un agent dédié. Chaque agent décide si sa variable doit être sélectionnée ou non et s'entraîne avec son propre DQN pour prendre ces décisions. **Actions :** Chaque agent dispose de deux actions possibles : sélectionner ou ne pas sélectionner sa variable assignée. **Environnement :** L'environnement correspond au sous-ensemble actuel des variables sélectionnées. Lorsqu'un agent inclut ou exclut une variable, l'état de l'espace des variables est mis à jour en conséquence. L'environnement inclut également l'AUC calculée à partir du sous-ensemble de variables sélectionnées précédemment, jouant un rôle crucial dans l'attribution des récompenses. **État :** L'état s représente le sous-ensemble actuel des variables sélectionnées. Étant donné que la longueur du vecteur d'état varie à chaque itération, contrairement aux scénarios DQN standard où la forme de l'entrée est fixe, nous utilisons des statistiques méta-descriptives (comme décrit dans Liu et al. (2019)) pour représenter s. **Récompense :** La récompense motive l'exploration de l'espace des variables. Elle est calculée en tenant compte à la fois du score AUC et du nombre de variables sélectionnées pour la tâche aval. La récompense est ensuite répartie équitablement entre les agents ayant sélectionné des variables lors de l'itération en cours. **Formateur :** Le formateur agit comme une source de guidance externe pour les agents, fournissant des conseils sur la manière d'ajuster leurs décisions en fonction des actions réalisées lors des épisodes précédents.

3.1 Approche multi-formateurs

Pour résoudre le problème des conseils répétitifs d'un seul formateur, en particulier lorsqu'il s'agit de variables d'entrée similaires, nous nous inspirons de Fan et al. (2020) et proposons une approche pédagogique améliorée. Cette nouvelle méthode utilise des algorithmes de sélection de variables pour permettre aux formateurs externes de fournir des conseils plus efficaces aux agents. Les éléments clés de notre stratégie multi-formateurs sont les suivants :

Agents de variables actives/inactives : Pour introduire de la variabilité dans les entrées fournies aux formateurs, nous ajustons dynamiquement ces entrées en identifiant les variables appelées "actives". Les variables actives sont celles sélectionnées par les agents lors de l'étape précédente. Par exemple, si à l'étape $i-1$, les agents sélectionnent j_2 et j_5, les variables actives à l'étape i sont j_2 et j_5. Les agents associés à ces variables actives sont appelés "agents actifs" J_a, tandis que les agents restants, liés aux variables non sélectionnées (inactives), sont classés comme "agents inactifs". Ces agents inactifs ont pour tâche de décider s'ils doivent

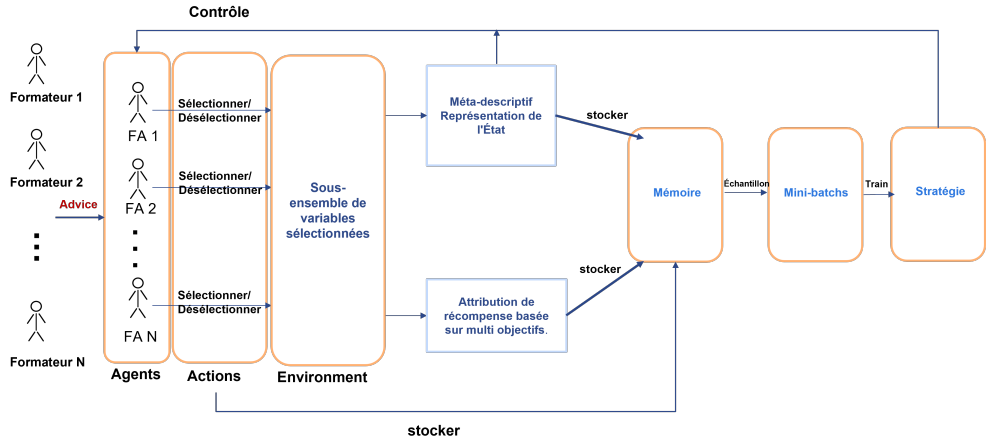


FIG. 1 – *Architecture MOIDRL-MA*

sélectionner ou désélectionner leurs variables inactives, assurant ainsi une exploration continue et un affinage du sous-ensemble de variables.

Agents de variables confiantes/indécises : Nous catégorisons dynamiquement les variables actives en deux groupes : les variables confiantes et les variables indécises, qui sont associées respectivement aux agents confiants et indécis. À chaque étape i, les variables sélectionnées par les agents sont étiquetées comme variables confiantes, tandis que celles non sélectionnées sont classées comme variables indécises. Par exemple, si à l'étape i, les variables confiantes incluent j_2 et j_5, et que les agents ag_2 et ag_5 décident de sélectionner j_2 et de désélectionner j_5, alors j_2 est classée comme caractéristique confiante avec ag_2 comme agent correspondant, tandis que j_5 est marquée comme caractéristique indécise avec ag_5 comme agent indécis correspondant.

Conseils et Sélections Initiales : À chaque étape, les agents utilisent leurs réseaux de politiques pour prendre des décisions initiales. Ils recherchent ensuite des conseils auprès des formateurs externes et ajustent leurs actions en conséquence, appelées "actions conseillées". Seuls les agents indécis suivent les conseils du formateur et exécutent les actions conseillées, tandis que les agents confiants conservent leurs choix initiaux et ne modifient pas leurs décisions en fonction des conseils du formateur.

Avant d'identifier les indices des agents nécessitant des ajustements, nous collectons d'abord leurs entrées et évaluons une métrique de performance (par exemple, le score F1) pour chaque formateur en fonction des performances des agents impliqués (J_p). Notre approche prend en charge un ensemble de formateurs défini par l'utilisateur, tels que K-meilleur, LASSO et l'élimination récursive des variables (RFE). À l'étape actuelle j, l'ensemble des formateurs est initialisé comme $T = T_1, \ldots, T_n$. Chaque formateur T_i est appliqué à J_p pour calculer sa métrique de performance. Le formateur ayant obtenu le meilleur score de performance est sélectionné comme "formateur gagnant" et fournit des conseils aux agents indécis. Cette sélection garantit que le formateur le plus efficace de l'ensemble T est choisi à chaque itération, améliorant ainsi le processus de guidance et favorisant une meilleure exploration. Lors

de l'exploration, des choix similaires peuvent être sélectionnés de manière aléatoire, ce qui peut amener un formateur à donner des conseils répétitifs. Pour y remédier, notre méthode favorise une exploration diversifiée en utilisant plusieurs formateurs, chacun apportant une perspective unique. De plus, se fier exclusivement à un seul formateur tout au long du processus pourrait entraîner des résultats sous-optimaux, exposant potentiellement les agents à des conseils biaisés ou inefficaces. En intégrant les contributions de plusieurs formateurs hautement performants, notre approche atteint un équilibre, réduisant les risques de mauvaises recommandations. Une dépendance excessive aux conseils des formateurs pourrait nuire aux capacités d'auto-apprentissage des agents. Pour atténuer cela, les agents commencent progressivement à s'orienter vers une exploration et une prise de décision indépendantes après avoir atteint la moitié du nombre d'épisodes. À ce stade, ils se concentrent sur la sélection des variables qu'ils identifient de manière autonome avec confiance, favorisant ainsi l'apprentissage et l'exploration autonomes.

3.2 Méthode d'Allocation des Récompenses Basée sur MODRL

Notre framework MOIDRL-MA étend les principes du MADRL. Dans cette approche, chaque agent exploite indépendamment son propre réseau de politique pour déterminer si la caractéristique correspondante doit être incluse ou exclue. Ce design adopte un paradigme multi-politique, où les agents sont équipés de politiques uniques $\pi_1, \ldots, \pi_N$, permettant à chacun des L agents d'adopter des stratégies distinctes. Pour intégrer le cadre multi-objectifs dans le problème de sélection de variables basé sur MADRL et l'adapter à MODRL, nous introduisons des préférences dans la fonction d'attribution des récompenses. Souvent, les critères d'évaluation largement utilisés combinent un score AUC avec une pénalité liée au nombre de variables sélectionnées Cavanaugh and Neath (2019). MOIDRL-MA surveille à la fois l'AUC et le nombre de variables à chaque itération, en comparant ces métriques avec leurs valeurs de l'itération précédente. Si le score AUC s'améliore, tous les agents correspondant aux variables sélectionnées reçoivent une récompense égale, indépendamment du nombre total de variables sélectionnées. En revanche, si le score AUC diminue, une pénalité est répartie de manière égale entre tous les agents de variables, indépendamment du nombre de variables. Lorsque le score AUC reste inchangé par rapport à l'itération précédente, la récompense ou la pénalité dépend du nombre de variables : une augmentation du nombre de variables entraîne une pénalité, tandis qu'une diminution donne lieu à une récompense pour tous les agents. La méthode de détermination des pénalités dans la sélection de variables multi-objectifs varie. Alors que Zheng and Wang (2018) utilise un hyperparamètre β multiplié par l'entropie, nous choisissons d'utiliser la somme du gain d'information (IG) à la place. Ce choix évite une suppression excessive des récompenses causée par l'entropie, ce qui pourrait fausser la distribution des récompenses dans l'RL. La somme du gain d'information $M_{IG}(\text{input})$ est définie comme suit :

$$M_{IG}(MF) = \sum_i IG(P_i; W) \tag{1}$$

where MF represents the set of selected features, $IG(P_i; W)$ denotes the information gain between the feature P_i and the target variable W. Accordingly, the penalty is calculated using the following formula:

$$\text{Pénalité} = \text{AUC} - \beta * M_{IG}(MF) \tag{2}$$

Nous suivons les vecteurs de récompense pour l'AUC et les variables sélectionnées tout au long de l'exécution de MOIDRL-MA. Une fois que le nombre maximum d'épisodes est atteint, nous extrayons les solutions du Front de Pareto à partir des vecteurs d'entraînement collectés.

3.3 Exploration du sous-ensemble des variables par Rejeu d'Expériences

Le Rejeu d'Expériences (ER) améliore l'efficacité de l'apprentissage par renforcement. Après chaque action, l'expérience la plus récente de l'agent i à l'épisode t, comprenant l'action (a_i^t), la récompense (r_i^t), l'état actuel (s_i^t) et l'état suivant (s_i^{t+1}), est stockée en mémoire en remplaçant l'expérience la plus ancienne. Les agents indécis utilisent l'ER pour entraîner leurs politiques en échantillonnant aléatoirement des mini-lots à partir de cette mémoire. Chaque agent entraîne ensuite son DQN en utilisant ces mini-lots afin d'optimiser les récompenses à long terme, comme décrit ci-dessous :

$$Q(s_i^t, a_i^t) = r_i^t + \delta \; max \; Q(s_i^t, a_i^{t+1}) \tag{3}$$

où δ représente le facteur d'actualisation. L'algorithme continue son exécution jusqu'à atteindre le nombre maximal d'épisodes.

4 Configuration Expérimentale

Nous abordons les questions de recherche (QR) suivantes à l'aide des ensembles de données présentés dans le tableau 1 : **QR1 :** Comment MOIDRL-MA, avec sa composante multi-objective et ses solutions du Front de Pareto, se compare-t-il aux méthodes existantes d'apprentissage par renforcement et de sélection de variables traditionnelles ? **QR2 :** Comment la nouvelle stratégie de multi-formateur améliore-t-elle la diversité des conseils et renforce-t-elle l'efficacité de MOIDRL-MA ?

TAB. 1 – *Datasets de test*

Dataset	**# Attributs**	**# Exemples**	**# Classes**
Cancer du sein de Wisconsin (WBC)	32	569	2
Type de couverture forestière (FCT)	54	15120	7
Spambase (Spam)	57	4601	2
Toxicité	1203	171	2
Cancer du côlon (Colon)	2000	62	2
Leucémie (LEU)	5148	72	2
Leucémies à lignage mixte (MLL)	12534	72	3
Cancer du Poumon (LUNG)	12601	203	5
Expression génétique du cancer par séquençage (RNA)	16383	801	5
TCGA Cancers du rénaux(KIRC)	20532	1024	2

Pour répondre à la QR1, nous avons comparé MOIDRL-MA avec plusieurs méthodes issues de la littérature Kim et al. (2022), Fan et al. (2020), Liu et al. (2019), Khurana et al. (2018), et Fard et al. (2013), ainsi qu'avec quatre algorithmes de référence : (i) mRMR, en utilisant le nombre moyen de variables des solutions du Front de Pareto (PF) de MOIDRL-MA ; (ii) RFE, également basé sur le nombre moyen de variables des solutions du PF de MOIDRL-MA ; (iii) LASSO avec un paramètre de régularisation $\lambda = 1.0$; et (iv) la sélection génétique des variables (GFS) avec une probabilité de croisement de 0,7, une probabilité de mutation de 0,1, et 50 générations. Cette comparaison vise à démontrer l'efficacité de MOIDRL-MA en l'évaluant à la fois face à des approches existantes basées sur l'apprentissage par renforcement pour la sélection de variables et à des algorithmes de référence standards. Pour répondre à la QR2, nous avons effectué une comparaison entre MOIDRL-MA et IRFS-HT Fan et al. (2020), en remplaçant GCN par les statistiques méta-descriptives pour la représentation des états. Nous avons utilisé une taille de mini-lot de 64, AdamOptimizer avec un taux d'apprentissage de 0,01, un facteur d'actualisation $\gamma = 0.99$, et une mémoire de taille 2000 pour le rejeu d'expériences. Notre architecture DQN comprend deux couches ReLU avec respectivement 64 et 8 nœuds. Les expériences ont été menées sur 1000 épisodes pour les ensembles de données LEU, MLL, LUNG, RNA et KIRC, et sur 5000 épisodes pour WBC, FCT, Spam, Toxicity et Colon. Pour les tâches en aval, nous avons utilisé la régression logistique pour WBC, la forêt aléatoire pour KIRC, XGBoost pour Spam, Toxicity, MLL et RNA, et SVM pour Colon, LEU et LUNG, tous avec les paramètres par défaut de scikit-learn. Les données ont été divisées de manière aléatoire en ensembles d'entraînement (70%) et de test (30%). Toutes les expériences ont été exécutées sur un framework Python avec GPU, 128 Go de RAM, et un système d'exploitation Linux 64 bits. Les détails de mise en œuvre et les paramètres sont disponibles dans notre code open-source : `https://anonymous.4open.science/r/MOIDRL-MA-6C5F`.

Dans les tableaux 2 et 3, les nombres entre parenthèses indiquent le nombre de variables sélectionnées, sauf dans la colonne "Dataset" où ils représentent les variables initiales. Seules les études Kim et al. (2022) et Fard et al. (2013) rapportent le nombre de variables sélectionnées dans le tableau 2, qui compare les ensembles de données en fonction de l'Exactitude (%), avec les résultats manquants marqués comme "-". Le tableau 3 présente les résultats basés sur l'AUC. Notre méthode utilise les entraîneurs K-best et les arbres de décision. Nous sélectionnons deux entraîneurs pour les comparer à IRFS-HT. Pour les ajustements d'index : **K-best :** Si une caractéristique indécise surpasse au moins la moitié des variables actives, elle passe du statut "non sélectionnée" à "sélectionnée". Le nombre de variables confiantes est e, et celui des variables indécises est f. Nous fixons $d = \lceil 2e + f \rceil$ et sélectionnons les d meilleures variables en utilisant K-best, ajustant les agents dans J_{KBest}. **Arbre de décision :** Nous entraînons le modèle sur J_a et calculons l'importance des variables. L'importance des variables indécises J_i est notée IMP_i, et celle des variables confiantes J_c est notée IMP_c. Si l'importance d'une caractéristique indécise dépasse la médiane de IMP_c, l'agent s'ajuste en conséquence.

5 Résultats et Discussions

Le tableau 2 montre que MOIDRL-MA surpasse systématiquement les autres méthodes de sélection de variables basées sur le RL dans la majorité des solutions de Front de Pareto. Pour le jeu de données Colon, MOIDRL-MA atteint une précision maximale de 95,8% avec seulement 8 variables, contre 94,7% avec 38 variables pour Kim et al. (2022), et 73% avec 40 variables

pour Fard et al. (2013). Cela met en évidence l'efficacité de la conception de MOIDRL-MA, qui combine apprentissage profond, RL interactif et approche multi-objective. En revanche, Kim et al. (2022) utilise un RL multi-agent simple, ce qui peut entraîner des décisions moins précises. De plus, Fard et al. (2013) s'appuie sur un agent unique qui doit explorer 2^N possibilités à chaque itération, conduisant souvent à des optima locaux. MOIDRL-MA évite ce problème en permettant à chaque agent de simplement sélectionner ou désélectionner des variables. Pour le jeu de données SPAM, parmi ses 22 solutions de Front de Pareto, MOIDRL-MA surpasse Fan et al. (2020) (93,3%) grâce à sa stratégie de diversification des entraîneurs à chaque épisode, au lieu d'utiliser un entraîneur unique tout au long de l'entraînement. Cette approche améliore considérablement le processus d'apprentissage et les performances de l'algorithme. Comparé à Fard et al. (2013) (82,9%, 20 variables), notre méthode montre des performances supérieures grâce à son composant multi-agent. Kim et al. (2022) rapporte une précision de 96,3% avec 20 variables, et Khurana et al. (2018) atteint 96,1%. Bien que notre approche soit légèrement inférieure avec une précision de 93,2%, elle sélectionne seulement 16 variables, offrant toujours des solutions compétitives dans le Front de Pareto. Pour le jeu de données FCT, les solutions de Front de Pareto générées par notre méthode n'atteignent pas la précision maximale rapportée par Kim et al. (2022) (88%), mais elles dépassent celles de Fan et al. (2020) (80%) et sont proches de Liu et al. (2019) (87,3%). Nous avons identifié des solutions avec une précision légèrement inférieure (86%), mais avec une réduction significative du nombre de variables : 27 pour MOIDRL-MA contre 33 pour Kim et al. (2022). Cela met en avant la valeur de la flexibilité multi-objective, qui propose des solutions pertinentes pour les décideurs, même avec une légère diminution de la précision.

Egalement, le tableau 3 montre que MOIDRL-MA surpasse les méthodes de référence dans presque tous les cas, à la fois pour les deux objectifs. Par exemple, dans le jeu de données Toxicity, MOIDRL-MA a atteint un AUC de 88,3% avec 31 variables, tandis que mRMR avait un AUC de 61,5%, RFE 76,8%, GFS 75,7% avec 624 variables, LASSO 75,3% avec 879 variables, et le scénario sans utilisation a produit un AUC de 65,2%. Notre approche a réduit de manière significative l'ensemble des variables, passant de 1203 à seulement 31, tout en obtenant le meilleur AUC. De même, pour des jeux de données plus volumineux comme KIRC, MOIDRL-MA a obtenu un AUC de 91%, contre 90,9% pour mRMR, 81,8% pour RFE, 90,1% pour GFS avec 10 185 variables, 89,7% pour LASSO avec 852 variables, et 81,8% pour le scénario sans utilisation. De plus, il a réduit l'ensemble des variables de 20 532 à seulement 17 variables finales sélectionnées. En conclusion, en réponse à **QR1**, MOIDRL-MA montre un grand potentiel pour surpasser les méthodes existantes (profondes) de renforcement et les méthodes traditionnelles de sélection de variables. En utilisant l'optimisation multi-objectifs, nous avons réussi à générer des solutions de Front de Pareto offrant aux décideurs une large gamme d'options. Cela améliore non seulement la flexibilité et l'adaptabilité de MOIDRL-MA, mais met également en évidence sa pertinence pratique dans les applications réelles. Bien que certaines méthodes puissent occasionnellement obtenir une précision ou un AUC plus élevés, notre approche se distingue par sa capacité à réduire de manière significative le nombre de variables dans la plupart des cas.

Le tableau 3 démontre l'efficacité de notre nouvelle stratégie d'enseignement par rapport à IRFS-HT. Dans le jeu de données RNA, notre méthode atteint un AUC comparable de 99,8% à travers les solutions du PF, tandis qu'IRFS-HT atteint 99,7%. Cependant, notre approche réduit de manière significative le nombre de variables à 63, tandis qu'IRFS-HT utilise 8207 variables.

MOIDRL-MA

Dataset	#Front de Pareto	Méthode proposée MOIDRL-MA	(Kim et al., 2022)	(Fan et al., 2020)	(Liu et al., 2019)	(Khurana et al., 2018)	(Fard et al., 2013)
WBC (32)	8	92.3 (1), 97.9 (4), 96.5 (2), 97.2 (3), 97.2 (3), 96.5 (2), 97.9 (5), 98.6 (9)	98.2 (17)	-	-	-	-
FCT (54)	17	85.4 (25), 85.2 (23), 85.4 (24), 85.7 (26), 54.7 (1), 85.1 (18), 84.0 (10), 75.1 (8), 87.1 (39), 69.8 (5), 68.9 (4), 59.3 (2), 71.1 (6), 77.7 (9), 86.2 (27), 86.7 (32), 72.7 (7)	88 (33)	80	87.3	-	-
Spam (57)	22	95.8 (33), 95.8 (37), 95.9 (41), 95.7 (29), 95.7 (32), 95.9 (39), 95.3 (25), 94.9 (24), 95.3 (26), 80.4 (1), 85.8 (3), 93.2 (16), 92.2 (10), 92.7 (15), 90.0 (9), 87.3 (4), 87.6 (6), 83.2 (2), 92.5 (13), 89.5 (8), 94.1 (20), 89.4 (7)	96.3 (20)	93	-	96.1	82.9 (20)
Colon (2000)	5	83.3 (2), 95.8 (8), 87.5 (7), 87.5 (4), 79.1 (1)	94.7 (38)	-	-	-	73 (40)

TAB. 2 – *Comparaison des méthodes MOIDRL-MA et RL (profonde) (Précision (%))*

Dans KIRC, notre méthode surpasse IRFS-HT avec un AUC de 91%, contre 89,7% pour IRFS-HT, et utilise seulement 17 variables, contre 10 165 pour IRFS-HT. Ces améliorations sont dues à l'intégration du concept de "trainer gagnant", qui fournit des conseils adaptatifs plutôt que de s'appuyer sur un seul formateur. Cette approche dynamique évolue en continu tout au long du processus, améliorant la performance et l'adaptabilité de l'algorithme. De plus, tandis qu'IRFS-HT fournit une seule solution basée sur le plus haut AUC, MOIDRL-MA offre 17 solutions PF non dominées pour KIRC, offrant aux décideurs une variété d'options parmi lesquelles choisir en fonction de leurs priorités.

En conclusion, et en réponse à **QR2**, MOIDRL-MA montre un potentiel prometteur pour surpasser IRFS-HT, comme en témoigne une réduction significative du nombre de variables tout en maintenant une performance élevée en termes d'AUC. Cela soutient l'efficacité de notre stratégie d'enseignement proposée. Cependant, un défi clé de notre méthode réside dans sa consommation élevée de mémoire et de temps en raison de la création d'un DQN pour chaque caractéristique, entraînant des coûts computationnels importants à mesure que le nombre de variables augmente. Pour y remédier, nous développons un processus d'apprentissage sélectif qui se concentre sur l'entraînement uniquement des agents indécis, dans le but de réduire l'utilisation des ressources et le temps d'exécution.

6 Conclusion

MOIDRL-MA combine le RL profond interactif avec l'optimisation multi-objectifs, et excelle dans la sélection de variables en utilisant les solutions du Pareto Front pour équilibrer l'AUC et le nombre de variables, offrant ainsi aux experts des options personnalisées. Sa principale force réside dans la génération de solutions sur mesure, mais elle présente des coûts élevés en termes de mémoire et de temps en raison du grand nombre d'agents basés sur DQN. Les travaux futurs se concentreront sur la mise en œuvre de l'apprentissage sélectif pour les agents indécis afin de réduire les besoins en ressources, l'application de la méthode aux données multiomiques dans le projet ANR RECORDS (qui inclut des ensembles de données métabolomiques et transcriptomiques avec jusqu'à 69 000 variables), et le développement

Dataset	# Front de Pareto	Méthode proposée MOIDRL-MA	IRFS-HT	mRMR	RFE	GFS	LASSO	Pas d'usage
WBC (32)	8	92.3 (1), 97.9 (4), 96.7 (2), 97.3 (3), 97.3 (3), 96.7 (2), 98.1 (5), 98.7 (9)	98.6 (11)	97.6 (4)	93.9 (4)	97.8 (14)	97.2 (28)	94.0
FCT (54)	17	85.4 (25), 85.2 (23), 85.4 (24), 85.7 (26), 54.7 (1), 85.1 (18), 84.0 (10), 75.1 (8), 87.1 (39), 69.8 (5), 68.9 (4), 59.3 (2), 71.1 (6), 77.7 (9), 86.2 (27), 86.7 (32), 72.7 (7)	86.5 (34)	82.9 (16)	54.7 (16)	86.2 (32)	85.5 (52)	85.5
Spam (57)	22	95.9 (33), 95.9 (37), 96.0 (41), 95.8 (29), 95.8 (32), 95.9 (39), 95.3 (25), 95.0 (24), 95.4 (26), 80.5 (1), 86.0 (3), 93.2 (16), 92.3 (10), 92.7 (15), 90.1 (9), 87.4 (4), 87.7 (6), 83.3 (2), 92.5 (13), 89.6 (8), 94.1 (20), 89.5 (7)	96.2 (34)	93.1 (18)	93.9 (18)	95.2 (34)	96.5 (55)	95.4
Toxicity (1203)	9	91.4 (588), 84.2 (10), 90.1 (556), 88.5 (109), 75.0 (1), 81.1 (2), 83.0 (8), 88.3 (31), 84.3 (21)	86.9 (957)	61.5 (147)	76.8 (147)	75.7 (624)	75.3 (879)	65.2
Colon (2000)	5	85.7 (2), 96.4 (8), 89.3 (7), 87.9 (4), 77.9 (1)	95.8 (957)	70.8 (4)	66.6 (4)	91 (990)	95.8 (541)	83.3
LEU (5148)	4	80.4 (1), 97.1 (4), 87.0 (3), 84.1 (2)	96.5 (2489)	89.6 (4)	89.7 (4)	98.5 (2520)	93.1 (338)	96.5
MLL (12534)	9	92.8 (19), 97.7 (113), 95.4 (57), 95.2 (47), 95.1 (36), 90.8 (7), 85.9 (3), 76.1 (1), 76.1 (1)	96.1 (6144)	88.4 (32)	92.3 (32)	98.1 (6205)	84.6 (500)	92.3
LUNG (12600)	13	98.6 (23), 97.4 (12), 91.3 (3), 85.3 (2), 92.3 (5), 96.2 (8), 98.0 (14), 71.3 (1), 71.3 (1), 99.7 (37), 98.3 (18), 95.1 (7), 99.1 (24)	99 (6142)	97.6 (17)	98 (17)	99.1 (6263)	99.5 (1365)	95.1
RNA (16383)	19	99.7 (55), 96.1 (10), 99.1 (31), 98.1 (21), 96.9 (12), 95.6 (8), 94.0 (6), 74.1 (2), 67.3 (1), 99.2 (32), 91.7 (4), 88.5 (3), 99.8 (63), 97.9 (15), 99.8 (102), 99.2 (35), 99.6 (47), 98.9 (23)	99.7 (8207)	99.7 (30)	99.3 (30)	99.7 (8253)	99.7 (2910)	99.5
KIRC (20532)	17	94.7 (493), 93.5 (210), 97.9 (4525), 85.7 (13), 81.1 (9), 85.4 (10), 70.0 (2), 95.7 (3952), 78.5 (7), 86.8 (16), 91.3 (70), 66.1 (1), 75.5 (3), 81.1 (9), 92.6 (92), 91 (17), 77.6 (4), 98.9 (23)	89.7 (10165)	90.9 (555)	81.8 (555)	90.1 (10185)	89.7 (852)	81.8

TAB. 3 – *Comparaison de MOIDRL-MA, IRFS-HT et des algorithmes traditionnels de sélection de variables (AUC (%))*

de solutions du Pareto Front pour le diagnostic de la septicémie afin d'optimiser à la fois la précision et le nombre de biomarqueurs.

Remerciements

Cette étude a été financée par le "Programme d'Investissements d'Avenir (PIA)" dans le cadre du programme France 2030 (ANR-18-RHUS-0004) et fait partie de l'initiative Federation Hospitalo-Universitaire (FHU) Saclay et Paris Seine Nord SEPSIS. Elle a également

bénéficié du financement ANR PIA (ANR-20-IDEES-0002). Ce travail a été réalisé en utilisant les ressources HPC de GENCI-IDRIS (Grant 2023-AD011014619) et également de HPC Grid'5000.

References

Cavanaugh, J. E. and A. A. Neath (2019). The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics 11*(3), e1460.

Fan, W., K. Liu, H. Liu, P. Wang, Y. Ge, and Y. Fu (2020). Autofs: Automated feature selection via diversity-aware interactive reinforcement learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1008–1013. IEEE.

Fard, S. M. H., A. Hamzeh, and S. Hashemi (2013). Using reinforcement learning to find an optimal set of features. *Computers & Mathematics with Applications 66*(10), 1892–1904.

Khurana, U., H. Samulowitz, and D. Turaga (2018). Feature engineering for predictive modeling using reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32.

Kim, M., J. Bae, B. Wang, H. Ko, and J. S. Lim (2022). Feature selection method using multi-agent reinforcement learning based on guide agents. *Sensors 23*(1), 98.

Lin, J., Z. Ma, R. Gomez, K. Nakamura, B. He, and G. Li (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access 8*, 120757–120765.

Liu, K., Y. Fu, P. Wang, L. Wu, R. Bo, and X. Li (2019). Automating feature subspace exploration via multi-agent reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 207–215.

Liu, K., D. Wang, W. Du, D. O. Wu, and Y. Fu (2023). Interactive reinforced feature selection with traverse strategy. *Knowledge and Information Systems 65*(5), 1935–1962.

Zheng, K. and X. Wang (2018). Feature selection method with joint maximal information entropy between features and class. *Pattern Recognition 77*, 20–29.

Résumé

Cet article propose une méthode de sélection de variables utilisant un RL profond multi-objectif. Elle équilibre l'AUC et le nombre de variables grâce à des récompenses et une exploration autonome, produisant des solutions sur un Front de Pareto et surpassant les méthodes existantes.

DCGAN pour la Complétion et la Génération d'Objets 3D Cassés à partir de Jeux de Données Réduits

Yahia Hamdi*, Nicolas Andrialovanirina*,** Pierre-Alexandre Hébert*,
Kélig Mahé**, Émilie Poisson Caillault*

*Univ. Littoral Cote d'Opale, LISIC UR 4491, F-62228 Calais, France
emilie.poisson@univ-littoral.fr,
https://lisic-prod.univ-littoral.fr/author/epoissoncaillault/
**IFREMER - LRH, F-62200 Boulogne-sur-mer
kelig.mahe@ifremer.fr
https://annuaire.ifremer.fr/cv/16959/

Résumé. La génération et la complétion d'objets 3D représentent un défi transformateur dans le domaine de la vision par ordinateur. Notre étude explore l'utilisation des réseaux antagonistes génératifs (GANs) pour la génération et la complétion d'objets 3D scannés et fracturés. Notre approche s'appuie sur des GANs convolutifs 3D profonds (DCGANs) pour générer des modèles 3D de haute qualité et reconstruire des objets incomplets ou endommagés. En entraînant les DCGANs sur des vecteurs latents, nous permettons la génération réaliste de formes 3D ainsi que la complétion d'objets partiels. De plus, nous évaluons la capacité du modèle à identifier et remplir des trous de tailles variées et comparons ses performances à celles de méthodes existantes. Les résultats quantitatifs et qualitatifs mettent en évidence l'efficacité du DCGAN proposé pour traiter de petits ensembles de données, gérer des données 3D complexes et produire des structures cohérentes et biologiquement plausibles. Le code et les données sont disponibles à l'adresse suivante : https://github.com/yahyahamdi-lab/3D-DCGAN.

1 Introduction

Aujourd'hui la classification des objets 3D reste un défi majeur dans le domaine de la vision par ordinateur, principalement en raison de la rareté de bases de données vastes et complètes ou de la taille limitée des bases existantes. Il est possible d'utiliser une base de données contenant des objets endommagés. L'un des principaux défis réside dans le fait que le calcul des descripteurs de Fourier sphériques, couramment utilisés pour l'analyse et la classification d'objets 3D, nécessite des formes complètes. Pour surmonter cette limitation, il est nécessaire de développer des méthodes capables d'exploiter efficacement toutes les données disponibles, même si elles sont incomplètes, afin de reconstruire ou d'approcher la forme originale des objets. Cela permet ensuite de calculer de manière fiable les descripteurs de Fourier sphériques et de classer avec précision les objets, malgré leur état fragmenté.

Plusieurs approches peuvent être envisagées, telles que les méthodes de reconstruction 3D qui permettent de combler les parties manquantes des objets. Cela inclut des techniques comme la modélisation basée sur la symétrie Dang et al. (2014), Hähnlein et al. (2022), ou l'utilisation de réseaux neuronaux profonds, tels que les GANs Wu et al. (2016) ou les autoencodeurs Jolicoeur-Martineau (2018), capables d'apprendre à prédire et restaurer les formes manquantes à partir des fragments disponibles.

Les technologies de deep learning, en particulier les réseaux de neurones convolutifs (CNNs) et les GANs (Generative Adversarial Networks), ont révolutionné le remplissage d'images en permettant des processus automatiques et hautement précis. Des techniques avancées, telles que les mécanismes d'attention Rabhi et al. (2024), Hamdi et al. (2023) et Rabhi et al. (2021), ainsi que l'extraction de caractéristiques multi-échelles, améliorent encore leur capacité à gérer des contextes d'images complexes.

La plupart des méthodes actuelles de reconstruction de formes 3D reposent sur une supervision complète, où des formes partielles d'un type donné sont utilisées en combinaison avec les formes complètes correspondantes pour l'entraînement Huang et al. (2020). Une approche non supervisée, appelée pcl2pcl, a été présentée par Chen et al. (2020). Elle exploite des données non appariées, telles que des formes complètes provenant de modèles 3D et des scans partiels issus du monde réel. Cette méthode consiste à entraîner deux autoencodeurs distincts : l'un pour reconstruire les formes complètes et l'autre pour les formes partielles. Elle apprend également une correspondance entre l'espace latent des formes partielles et celui des formes complètes. De plus, Zhang et al. (2021) a introduit une méthode non supervisée pour générer des formes 3D complètes en utilisant l'inversion de GAN. Cette méthode commence par un GAN pré-entraîné destiné à générer des formes complètes, puis affine le code latent afin que le GAN produise une forme complète correspondant à l'entrée partielle fournie. Elle peut atteindre des performances comparables à celles des techniques supervisées mentionnées.

Plus récemment, un modèle de GAN en cascade 3D basé sur une architecture encodeur-décodeur a été proposé par Alhamazani et al. (2024) pour reconstruire des formes à partir d'images de profondeur obtenues sous un seul angle de vue. Ce modèle vise à sélectionner des codes importants, à capturer les relations non locales dans l'espace latent 3D et à intégrer une couche d'auto-attention pour stabiliser l'apprentissage du modèle et affiner les formes reconstruites. Inspirés par le succès des modèles GAN dans la génération et la reconstruction d'objets 3D, nous proposons une architecture profonde basée sur des couches convolutionnelles 3D, conçue pour la complétion et la génération de données 3D de haute qualité. Nous évaluons ses performances dans les tâches de complétion et de génération à l'aide de diverses métriques, telles que la distance de Chamfer (Chamfer Distance, CD), la distance de Hausdorff (Hausdorff Distance, HD) et la distance du cantonnier (Earth Mover's Distance, EMD), sur les ensembles de données de référence ShapeNet Chang et al. (2015) et Otolith Andrialovanirina et al. (2024).

2 Méthode

Les tentatives précédentes pour faire évoluer les GANs en utilisant des CNNs afin de générer des images 3D haute résolution ont généralement été fructueuses, bien que certaines limitations subsistent. En réponse à ces défis, Wu et al. (2016) a proposé une méthode qui génère des objets 3D à partir d'un espace probabiliste en utilisant des GANs et des réseaux

convolutionnels volumiques (3DGAN). Cette approche permet au générateur de mapper un espace probabiliste de faible dimension vers l'espace des objets 3D, facilitant ainsi l'échantillonnage d'objets sans nécessiter d'images de référence. De plus, le discriminateur adversarial agit comme un descripteur puissant et non supervisé de formes 3D. Cette méthode a été étendue en employant la distance de Wasserstein normalisée avec une pénalisation de gradient comme objectif d'apprentissage dans le modèle RaGANs Jolicoeur-Martineau (2018).

Bien que ces architectures soient efficaces, elles ne répondent pas entièrement à nos besoins, en particulier pour les formes ovoïdes ou irrégulières, comme les otolithes, qui peuvent présenter des surfaces avec des dépressions, des bosses et des lignes variant d'une espèce à l'autre.

Grâce à une expérimentation approfondie avec divers modèles, nous avons identifié une architecture spécifique (voir Figure 1) qui a permis un apprentissage stable sur plusieurs ensembles de données, facilitant ainsi la conception de modèles génératifs plus profonds et à plus haute résolution.

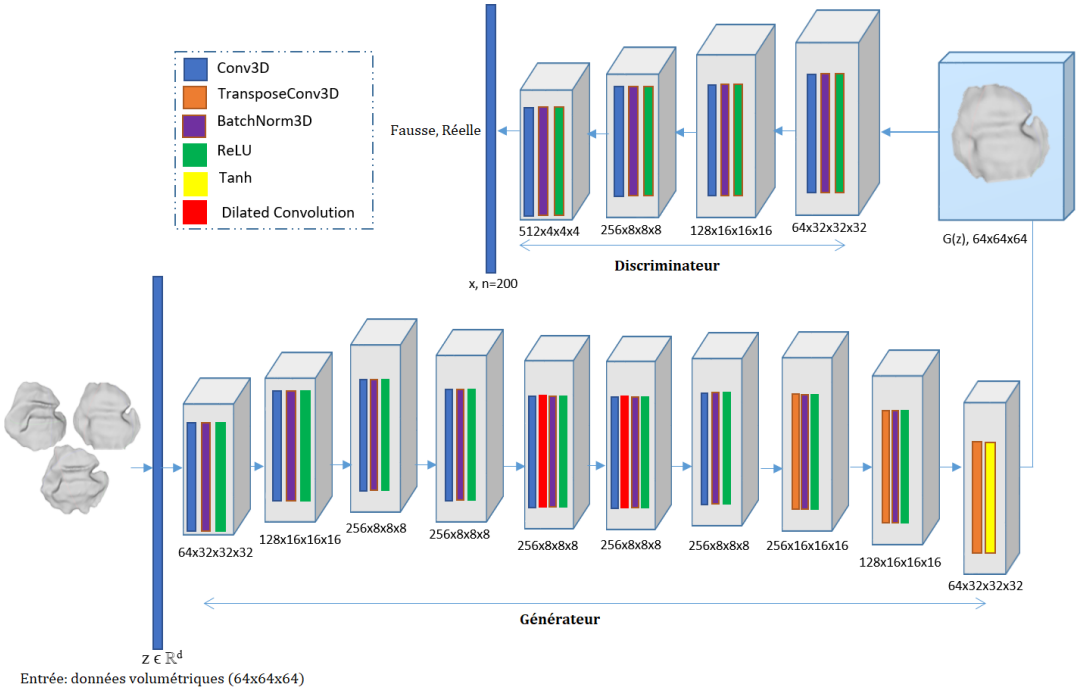


FIG. 1 – *Architecture DCGAN proposée.*

2.1 Détails de l'architecture

Pour traiter des objets 3D plus complexes tant dans les tâches de complétion que de génération, nous avons exploité l'architecture développée par Jolicoeur-Martineau (2018). Comme présenté dans la Figure 1, l'architecture proposée se concentre sur des couches profondes et

complexes tant pour le Générateur que pour le Discriminateur, ce qui améliore potentiellement leur capacité à capturer des caractéristiques complexes dans les données 3D.

Générateur (G) : L'architecture se compose de 14 blocs basés sur des couches convolutionnelles. Elle utilise une combinaison de couches convolutionnelles et de couches convolutionnelles transposées pour augmenter progressivement les dimensions spatiales des données d'entrée. Elle commence par des couches convolutionnelles 3D (*Conv3d*) utilisant une taille de noyau de 4 et un stride de 2 pour la réduction des dimensions, chacune suivie d'une normalisation par lots (*BatchNorm3d*) et d'une activation *ReLU* au lieu de *LeakyReLU*. Ces techniques de normalisation peuvent atténuer les gradients explosifs ou évanescents en maintenant les activations dans un intervalle stable. Le générateur augmente progressivement le nombre de filtres (de 1 à 256) tout en réduisant les dimensions spatiales. De plus, des convolutions dilatées avec des taux de dilatation croissants (4 et 8) sont utilisées pour capturer des caractéristiques à plusieurs échelles sans réduction spatiale supplémentaire. Nous avons utilisé des taux de dilatation pour ajuster le champ réceptif, permettant de capturer les dépendances spatiales dans les données volumétriques. En effet, les faibles taux de dilatation excellent dans la capture des détails fins, mais peuvent échouer à saisir le contexte global. À l'inverse, des taux de dilatation élevés permettent de gérer les dépendances à longue portée et les grandes parties manquantes, mais risquent de perdre des détails fins ou d'introduire des artefacts. Pour obtenir des meilleurs résultats, nous avons utilisé une combinaison équilibrée de petits et grands taux de dilatation, intégrant ainsi les contextes local et global. Les défis, tels que les goulets d'étranglement en mémoire et les instabilités d'entraînement, ont été atténués grâce à l'utilisation de techniques d'ajustement du taux d'apprentissage.

Pour augmenter la taille des matrices de caractéristiques, des couches convolutionnelles transposées (*ConvTranspose3d*) sont appliquées, également accompagnées de normalisation par lots et d'activation *ReLU*. La couche finale applique une fonction d'activation *Tanh* pour produire des sorties dans l'intervalle [-1, 1], adaptées aux représentations voxel de la forme $x_c \in \mathbb{R}^{m\times 3}$ générées à partir d'un vecteur latent z $\in \mathbb{R}^d$.

Discriminateur (D) : Tout comme le générateur, le discriminateur comprend plusieurs couches *Conv3d*, atteignant jusqu'à 512 filtres, qui réduisent les dimensions spatiales de l'entrée tout en appliquant la normalisation par lots et les activations *ReLU*, augmentant ainsi la profondeur. La sortie de la dernière couche convolutionnelle est aplatie et passée à travers une couche linéaire entièrement connectée, suivie d'une fonction d'activation *Sigmoid* pour générer une probabilité indiquant si l'entrée est réelle ou générée. La taille de l'entrée pour la couche linéaire est définie dynamiquement à l'aide d'un tenseur simulé, garantissant ainsi la flexibilité pour différentes dimensions d'entrée.

2.2 Détails de l'entraînement et de l'implémentation du modèle

Le modèle a été développé en utilisant PyTorch 1.10.1 avec Python 3.7 et CUDA version 11.3. Il a été entraîné avec des pertes d'entropie croisée binaire (BCE-loss) en définissant des paramètres empiriques tels que le taux d'apprentissage de 1e-4 pour *G* et *D*, la taille du lot de 128 et le nombre d'époques de 5000. En effet, la valeur de 128 a été choisie comme taille de lot optimale pour notre architecture, après plusieurs tests expérimentaux. Cette valeur a montré les meilleurs résultats en termes de performance et de stabilité d'entraînement, offrant un bon compromis entre l'efficacité de l'entraînement et l'utilisation de la mémoire. Typiquement, la performance du modèle a atteint un plateau après 2000 époques. L'optimiseur Adam est utilisé

avec les paramètres de moment *beta1 = 0.5* et *beta2 = 0.9*, définis sur les valeurs par défaut adaptées à un entraînement stable. Nous avons alterné l'entraînement du discriminateur et du générateur, permettant au modèle d'améliorer sa capacité à générer des images réalistes au fil du temps.

3 Résultats

Dans ce travail, nous démontrons la capacité de notre modèle à générer et à compléter des formes partielles, surpassant ainsi les modèles précédents tels que ceux mentionnés dans Wu et al. (2016) et Jolicoeur-Martineau (2018). Pour cette raison, nous décrivons d'abord les ensembles de données utilisés, suivis des métriques d'évaluation et des résultats obtenus.

3.1 Base de données

Deux bases de données distinctes sont utilisées pour évaluer notre modèle. Nous utilisons les sous-ensembles Airplane et Chair, extraits des benchmarks ShapeNet Chang et al. (2015), pour le premier ensemble de données, comme présenté dans des études précédentes Jolicoeur-Martineau (2018), Miao et al. (2023). Le deuxième ensemble de données contient des formes d'otolithes issues de 691 spécimens individuels de rouget-barbet (*Mullus barbatus*), comme décrit dans Andrialovanirina et al. (2024).

Pour les tâches de génération, nous avons entraîné le modèle proposé en utilisant des maillages voxel représentant 80 % des formes des ensembles de données fournis, les 20 % restants étant réservés aux tests. Pour la complétion de formes, nous générons aléatoirement des nuages de points partiels en débruitant 65 % des images de profondeur disponibles, puis effectuons des tests de complétion de formes sur des formes partielles correspondant à chaque échantillon de référence.

3.2 Métriques d'évaluation

Tel que présenté dans des études précédentes, notamment Dang et al. (2014), Alhamazani et al. (2024), Miao et al. (2023), nous utilisons CD, HD et EMD comme métriques principales pour évaluer la qualité de complétion et de génération en 3D. En effet, CD et HD mesurent la similarité entre deux ensembles de points, tandis que EMD calcule le coût minimal de transformation d'une distribution de probabilité en une autre. Par ailleurs, nous utilisons le Taux de Rétention en Pourcentage des Points à partir de la forme d'entrée (PRR) comme un indicateur de qualité, de fidélité structurelle et d'efficacité pour les modèles de reconstruction 3D.

3.3 Génération de Formes

Dans cette sous-section, nous évaluons notre modèle proposé, DCGAN, en le comparant à des études ayant utilisé des couches convolutionnelles basées sur des GAN pour les nuages de points.

La Table 1 présente les résultats quantitatifs pour chaque base de test, tandis que la Figure 2 offre une comparaison visuelle obtenue à partir d'un objet réel, complet (vérité terrain). À partir de la table, nous observons que notre modèle surpasse systématiquement à la fois

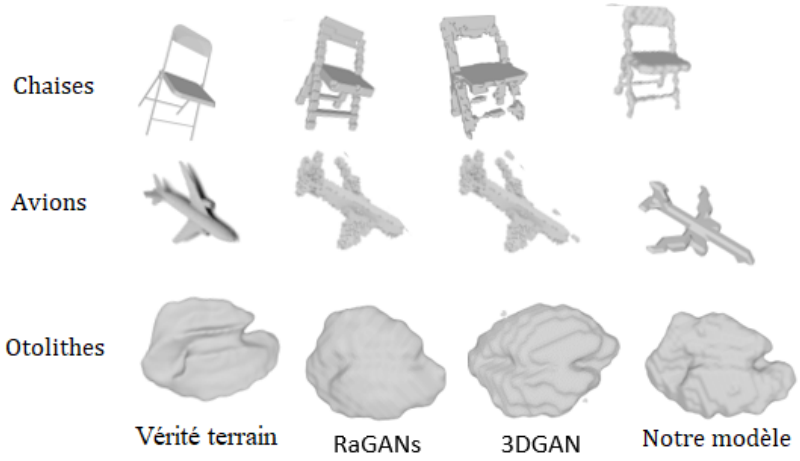


FIG. 2 – *Comparaisons des résultats de génération obtenus par différentes méthodes sur les formes de chaises, d'avions et d'otolithes.*

Catégories	Modèles	CD	HD	EMD	PRR
Avions	3DGAN	19.8	68.1	67.1	85.1
	RaGANs	16.4	63.2	64.5	89.5
	DCGAN (ours)	**14.5**	**62.1**	**60.4**	**95.4**
Chaises	3DGAN	16.9	65.1	63.1	90.9
	RaGANs	15.1	64.2	60.3	92.1
	DCGAN (ours)	**13.3**	**61.6**	**58.2**	**96.5**
Otolithes	3DGAN	31.2	82.5	71.0	87.1
	RaGANs	28.9	76.6	69.7	89.8
	DCGAN (ours)	**25.5**	**71.2**	**65.5**	**94.6**

TAB. 1 – *Résultats de génération sur les ensembles de données Chaises, Avions et Otolithes en utilisant les métriques CD ($\times 10^2$), HD ($\times 10^3$), EMD ($\times 10$), et PRR (%) comparés aux méthodes de l'état de l'art.*

3DGAN et RaGANs dans la génération de formes 3D, comme en témoignent ses scores inférieurs en CD, HD et EMD dans toutes les catégories. Son PRR plus élevé suggère également une meilleure capacité à préserver la structure de l'objet original. Cela indique que DCGAN offre des résultats de génération d'objets 3D de meilleure qualité, plus précis et visuellement plus réalistes que les autres modèles testés.

3.4 Complétion de Formes

Dans de nombreuses applications graphiques, les utilisateurs n'ont pas accès à tous les points de vue possibles d'un objet. Il est souvent nécessaire de compléter une forme 3D lorsque seules des informations partielles, telles qu'une seule carte de profondeur, sont disponibles. Par

conséquent, la capacité à reconstruire ou à compléter des formes partielles devient extrêmement précieuse en pratique. Dans cette section, nous évaluons également notre modèle de complétion de formes en utilisant la même architecture décrite précédemment.

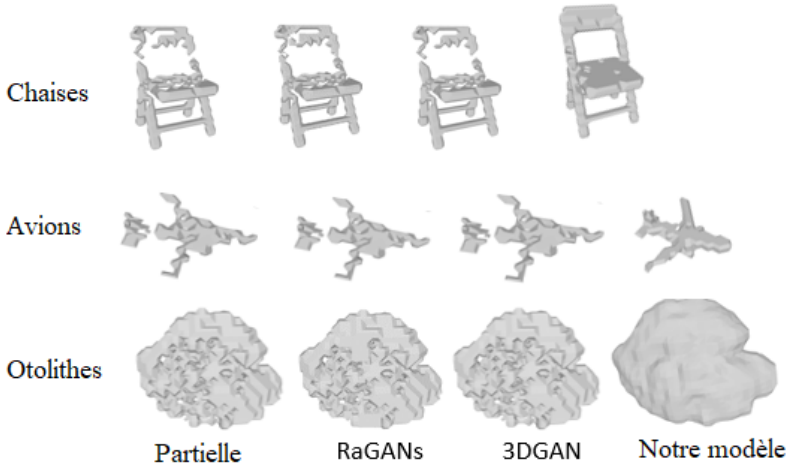


FIG. 3 – *Comparaisons des résultats de complétion obtenus par différentes méthodes sur les formes de chaises, d'avions et d'otolithes.*

Catégories	Modèles	CD	HD	EMD	PRR
Avions	3DGAN	46.7	72.8	84.0	77.2
	RaGANs	41.2	71.3	81.1	85.7
	DCGAN (ours)	**27.1**	**67.6**	**65.5**	**94.9**
Chaises	3DGAN	43.2	76.0	80.4	82.5
	RaGANs	41.1	73.2	79.6	83.4
	DCGAN (ours)	**26.0**	**65.4**	**61.2**	**95.0**
Otolithes	3DGAN	54.8	94.1	81.9	80.0
	RaGANs	52.6	87.6	79.1	82.3
	DCGAN (ours)	**41.7**	**82.1**	**69.4**	**94.0**

TAB. 2 – *Résultats de complétion sur les ensembles de données Chaises, Avions, et Otolithes en utilisant les métriques CD ($\times 10^2$), HD ($\times 10^3$), EMD ($\times 10$), et PRR (%) comparés aux méthodes de l'état de l'art.*

La Table 2 résume les résultats quantitatifs pour la complétion d'objets 3D sur chaque ensemble de données de test, tandis que la Figure 3 fournit une comparaison visuelle d'un objet réel sous sa forme artificiellement perforée (partielle). Les résultats montrent que notre modèle DCGAN surpasse systématiquement 3DGAN et RaGANs dans la complétion des formes 3D, comme en témoignent ses scores plus faibles en CD, HD et EMD dans toutes les catégories.

De plus, ses valeurs élevées de PRR reflètent la capacité du modèle à préserver l'intégrité structurelle des objets, garantissant des complétions réalistes et cohérentes même pour des entrées fortement complexes ou incomplètes. Cette robustesse souligne son adéquation pour des applications réelles nécessitant une reconstruction 3D fiable.

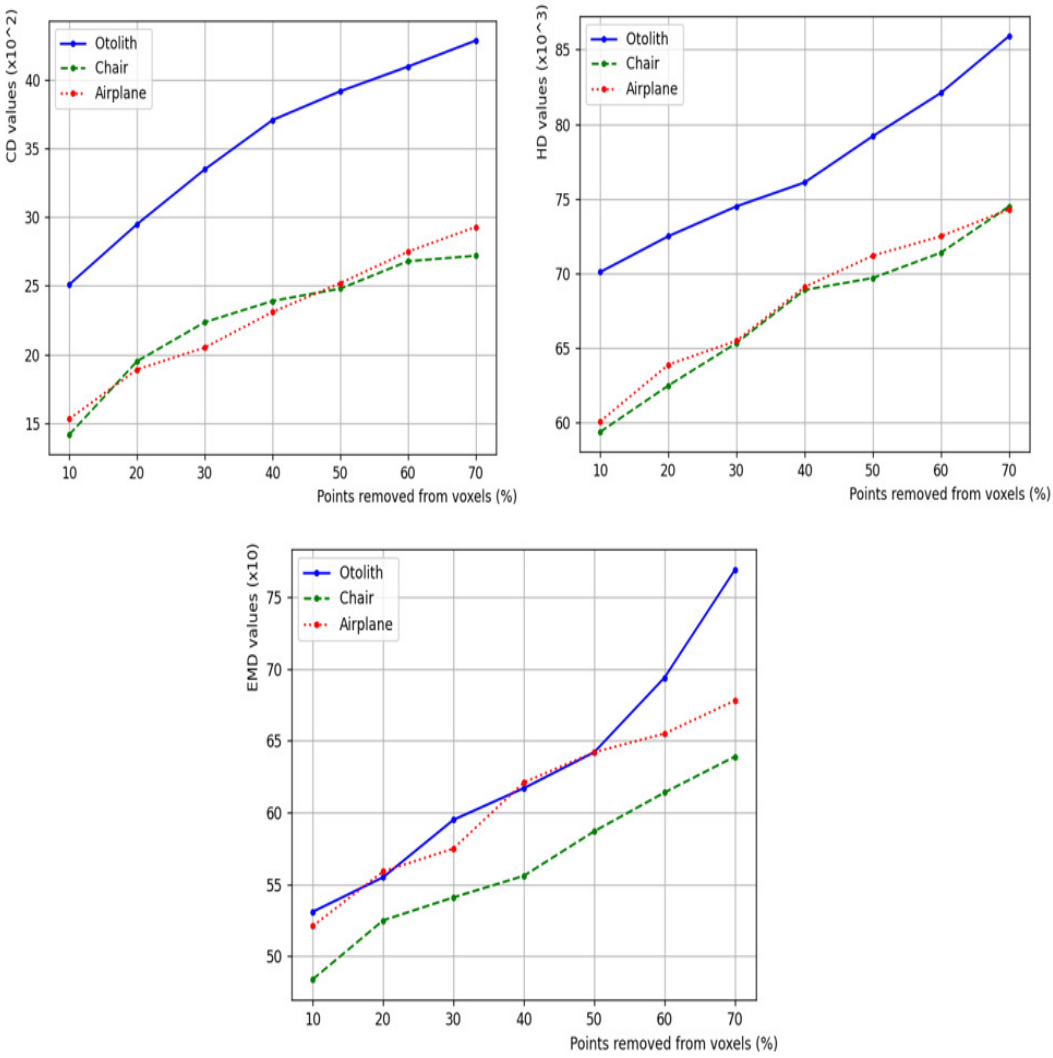


FIG. 4 – *Pourcentage de points supprimés des voxels et les valeurs correspondantes de CD, HD et EMD.*

La Figure 4 illustre l'impact du pourcentage de points supprimés sur la précision de la reconstruction des objets 3D (Otolithes, Chaises et Avions) en utilisant les métriques CD, HD et EMD. Les résultats révèlent qu'à mesure que le pourcentage de points supprimés augmente,

les valeurs des métriques correspondantes augmentent, indiquant une baisse de la précision de la reconstruction. Cela démontre la capacité du modèle proposé à compléter efficacement des objets 3D, même lorsque jusqu'à 70 % des voxels ont été supprimés. En effet, les performances se dégraderaient en cas d'incomplétude sévère, en particulier si de tels cas ne faisaient pas partie de la distribution des données d'entraînement.

D'une part, cette performance met en évidence la robustesse du modèle à travers différentes catégories d'objets, incluant des géométries complexes comme les otolithes ainsi que des structures plus simples telles que les chaises et les avions. D'autre part, les otolithes, en raison de leurs structures complexes et très détaillées, montrent la plus forte augmentation des valeurs des métriques, ce qui reflète une sensibilité accrue à la perte de données. À l'inverse, les objets plus simples et mieux structurés, tels que les chaises et les avions, font preuve d'une plus grande résilience, avec des taux de croissance des valeurs des métriques plus modérés.

Par ailleurs, ce modèle peut se généraliser à divers ensembles de données, notamment en imagerie médicale (par exemple, IRM, CT scans) et pour des objets aux formes irrégulières, moyennant certaines modifications. Ces adaptations incluent des ajustements de prétraitement pour gérer des résolutions et des structures spécifiques, ainsi que des modifications architecturales, telles que l'intégration de convolutions anisotropes, le traitement multi-échelles ou des mécanismes d'attention pour capturer à la fois les détails fins et le contexte global.

4 Conclusion et Travaux futurs

Dans cet article, nous avons adopté les couches convolutionnelles basées sur les GAN existantes pour les tâches de génération et de complétion. La spécificité de l'architecture proposée réside dans l'utilisation de multiples blocs de couches Conv3D, Deconv et DCL dans les modules du générateur et du discriminateur. Ce modèle est structuré sous forme d'un réseau encodeur-décodeur avec une méthode d'extraction progressive des caractéristiques, où chaque couche affine les matrices de caractéristiques et capture des informations géométriques de plus en plus complexes, renforçant ainsi la capacité du modèle à générer des complétions précises et réalistes. Les résultats quantitatifs et qualitatifs démontrent l'efficacité de notre modèle pour inférer et combler les lacunes dans la forme 3D, produisant des reconstructions réalistes même lorsqu'il est fourni avec des informations visuelles limitées. Cependant, cette architecture peut être étendue pour traiter des grilles de voxels 3D de résolution plus élevée afin d'obtenir des reconstructions plus détaillées. Cela engendre néanmoins des défis computationnels et architecturaux significatifs. En particulier, des réseaux plus profonds sont nécessaires pour traiter ces données haute résolution, ce qui augmente les risques de gradients évanescents et de sur-apprentissage. Les stratégies pour relever ces défis incluent des techniques telles que le traitement hiérarchique multi-résolution, les convolutions éparses, l'entraînement distribué et l'application de méthodes de régularisation adaptées. Par ailleurs, nous suggérons d'exploiter les objets générés et complétés pour des tâches de classification, en s'appuyant sur le descripteur de Fourier sphérique afin d'améliorer la performance et l'efficacité de l'analyse des formes 3D.

Ce travail fait partie de l'école doctorale IFSEA, qui bénéficie de la subvention ANR-21-EXES-0011, financée par l'Agence Nationale de la Recherche (ANR), dans le cadre du programme France 2030.

Références

Alhamazani, F., Y.-K. Lai, et P. L. Rosin (2024). A coarse-to-fine point completion network with details compensation and structure enhancement. *Scientific Reports 14*, 1991.

Andrialovanirina, N., L. Poloni, R. Laffont, Émilie Poisson Caillault, S. Couette, et K. Mahé (2024). 3d meshes dataset of sagittal otoliths from red mullet in the mediterranean sea. *Scientific Data 11*.

Chang, A. X., T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q.-X. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, et F. Yu (2015). Shapenet: An information-rich 3d model repository. *ArXiv abs/1512.03012*.

Chen, X., B. Chen, et N. J. Mitra (2020). Unpaired point cloud completion on real scans using adversarial training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dang, Q.-V., S. Mouysset, et G. Morin (2014). Symmetry-based alignment for 3d model retrieval. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6.

Hähnlein, F., Y. Gryaditskaya, A. Sheffer, et A. Bousseau (2022). Symmetry-driven 3d reconstruction from concept sketches. In M. Nandigjav, N. J. Mitra, et A. Hertzmann (Eds.), *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pp. 19:1–19:8. ACM.

Hamdi, Y., B. Rabhi, T. Dhieb, et A. M. Alimi (2023). Multi-head self-attention and BGRU for online arabic grapheme text segmentation. In N. E. B. Amara, A. Sourin, O. Sourina, et C. Rosenberger (Eds.), *International Conference on Cyberworlds, CW 2023, Sousse, Tunisia, October 3-5, 2023*, pp. 78–85. IEEE.

Huang, Z., Y. Yu, J. Xu, F. Ni, et X. Le (2020). Pf-net: Point fractal network for 3d point cloud completion. *CoRR abs/2003.00410*.

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard GAN. *CoRR abs/1807.00734*.

Miao, Y., C. Jing, W. Gao, et X. Zhang (2023). 3dcascade-gan: Shape completion from single-view depth images. *Computers & Graphics 115*, 412–422.

Rabhi, B., A. Elbaati, H. Boubaker, Y. Hamdi, A. Hussain, et A. M. Alimi (2021). Multilingual character handwriting framework based on an integrated deep learning based sequence-to-sequence attention model. *Memetic Comput. 13*(4), 459–475.

Rabhi, B., A. Elbaati, Y. Hamdi, H. Dhahri, U. Pal, H. Chabchoub, K. Ouahada, et A. M. Alimi (2024). A novel multi-head attention and long short-term network for enhanced inpainting of occluded handwriting. *Cognitive Computation 17*.

Wu, J., C. Zhang, T. Xue, B. Freeman, et J. Tenenbaum (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, et R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 82–90.

Zhang, J., X. Chen, Z. Cai, L. Pan, H. Zhao, S. Yi, C. K. Yeo, B. Dai, et C. C. Loy (2021). Un-

supervised 3d shape completion through GAN inversion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 1768–1777. Computer Vision Foundation / IEEE.

Summary

The generation and completion of 3D objects is a transformative challenge in computer vision. Our study explores Generative Adversarial Networks (GANs) for generating and completing fractured 3D scanned objects. Our approach leverages Deep 3D Convolutional GANs (DCGANs) to generate high-quality 3D models and reconstruct incomplete or damaged objects. By training DCGANs on latent vectors, we enable realistic 3D shape generation and completion of partial objects. Additionally, we evaluate the model's ability to recognize and fill holes of varying sizes and compare its performance against existing methods. Quantitative and qualitative results highlight the effectiveness of the proposed DCGAN in handling small datasets, processing complex 3D data, and generating coherent, biologically plausible structures. The code and data supporting this work are publicly available at:

https://github.com/yahyahamdi-lab/3D-DCGAN.

Évaluation des stratégies de Mixup sur des ensembles de données déséquilibrés de petite taille

Asmaa LAGRID *, Sébastien FOURNIER*

*Aix-Marseille université
asmaa.lagrid@univ-amu.fr,
sebastien.fournier@univ-amu.fr

Résumé. La classification de texte est une tâche cruciale en traitement automatique du langage naturel (TALN), souvent compliquée par le déséquilibre significatif des classes dans les jeux de données réels. Ce déséquilibre entrave l'efficacité des classifieurs en apprentissage profond, en particulier pour les classes minoritaires, qui tendent à être sous-représentées.

Pour pallier ces défis, diverses stratégies telles que la re-pondération de la fonction de perte et les techniques d'augmentation des données ont été explorées. Ces approches ont prouvé leur utilité dans des contextes de déséquilibre des classes, mais leur impact sur des petits jeux de données déséquilibrés reste peu exploré.

Notre étude évalue de manière systématique ces méthodologies sur des jeux de données de classification de texte bien établis que nous avons artificiellement déséquilibrés. Nous examinons également leur scalabilité et performance à travers des expérimentations sur des grands jeux de données déséquilibrés réduits par échantillonnage, ainsi que sur un jeu de données naturellement déséquilibré et de petite taille. L'objectif est de déterminer l'efficacité de ces techniques spécifiquement dans le contexte de petits ensembles de données déséquilibrés, pour mieux comprendre leur applicabilité et optimiser leur mise en œuvre.

1 Introduction

Les modèles d'apprentissage profond ont prouvé leur efficacité dans de nombreuses applications de traitement du langage naturel (TALN), telles que la génération de résumés automatiques, la classification de texte, l'extraction de texte et la traduction. Cependant, un problème courant dans l'apprentissage profond supervisé est le déséquilibre des données, où les classes majoritaires prédominent. Ce déséquilibre peut engendrer un biais dans l'apprentissage, compromettant ainsi la capacité des modèles à traiter efficacement les classes minoritaires et réduisant leur généralisation.

Pour pallier ce problème, des stratégies comme le réajustement et le ré-échantillonnage des classes ont été développées. Les méthodes de réajustement ajustent les poids des classes dans la fonction de coût, utilisant soit l'inverse de la fréquence des classes (Huang et al., 2016), soit le nombre effectif d'exemples par classe (Cui et al., 2019). Des approches telles que la balanced-focal-loss et la balanced-cross-entropy sont fréquemment utilisées pour équilibrer

les contributions des différentes classes. La technique LDAM-loss (Label Distribution-Aware Margin) (Cao et al., 2019) propose une solution innovante en attribuant une marge plus importante aux classes minoritaires, favorisant ainsi une meilleure représentation et généralisation.

Le rééchantillonnage comprend principalement le sur-échantillonnage (Kubat et al., 1997), qui augmente le nombre d'exemples des classes minoritaires, et le sous-échantillonnage, qui réduit celui des classes majoritaires. Les techniques de sur-échantillonnage (Kubat et al., 1997), incluent la duplication simple d'exemples ou l'utilisation de méthodes d'augmentation des données, telles que l'EDA (Easy Data Augmentation) (Wei et Zou, 2019) et le backtranslation (Kobayashi, 2018). Ces techniques introduisent des variations linguistiques naturelles et préservent l'intégrité des informations originales.

De plus, des techniques d'augmentation basées sur des modèles pré-entraînés (Wu et al., 2019) ou des modèles génératifs (Witteveen et Andrews, 2019), (Kurt Pehlivanoğlu et al., 2024) sont utilisées pour enrichir les ensembles de données tout en prenant en compte le contexte des phrases. Ces méthodes utilisent des architectures avancées pour générer des modifications textuelles qui maintiennent la cohérence et la pertinence du contenu tout en diversifiant l'ensemble des données.

Outre l'augmentation traditionnelle des données, des approches innovantes comme la génération de données synthétiques par interpolation dans l'espace des représentations vectorielles sont utilisées. Le DeepSMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002), (Dablain et al., 2022), qui génère des données synthétiques en interpolant les exemples des classes minoritaires, et le Mixup (Zhang, 2017), (Guo et al., 2019), qui combine des exemples de différentes classes pour former des instances nouvelles. Le Mixup, conçu initialement pour augmenter la taille des jeux de données, a prouvé son efficacité particulièrement dans les ensembles de petite taille. Il a également été démontré, comme le souligne (Cheng et al., 2023), que le Mixup améliore la généralisation des modèles dans le cas des données déséquilibrées en réduisant implicitement l'écart entre les classes majoritaires et minoritaires en enrichissant l'espace des caractéristiques par une interpolation judicieuse. Bien que ses variantes dédiées au probleme de déséquilibre des données ont été validé dans des contextes de vastes ensembles de données, le Mixup s'avère prometteur pour enrichir des ensembles plus restreints, en réduisant le biais de classe et augmentant la diversité, essentiels pour la généralisation des modèles.

Ce travail se concentre sur l'évaluation de la technique de Mixup pour des jeux de données textuelles déséquilibrés de petite taille, en comparaison avec des méthodes conventionnelles telles que la balanced-focal-loss, le LDAM-loss, et la technique EDA (Easy Data Augmentation). L'objectif principal est de vérifier si le Mixup peut offrir une amélioration supérieure en termes de généralisation des modèles face à ces méthodes établies dans des contextes spécifiques de déséquilibre.

Pour atteindre cet objectif, nous avons procédé à une modification artificielle des rapports de classe sur des jeux de données fréquemment utilisés, tels que SST-2, SST-5, SUBJ, TREC, et MR, pour simuler des déséquilibres de type long-tailed. En outre, des jeux de données naturellement déséquilibrés tels que R8 et NG20 ont été ajustés en réduisant leur volume pour correspondre à des scénarios de petite taille. Nous avons également inclus des expériences sur un jeu de données naturellement déséquilibré et de petite taille pour évaluer de manière approfondie l'applicabilité et la performance des techniques envisagées.

2 Préliminaires

2.1 Approches basées sur le réajustement

2.1.1 Focal loss (FL)

La Focal Loss (Cui et al., 2019) est une adaptation de la fonction de perte de cross-entropy, initialement conçue pour l'application en détection d'objets, où elle a montré des résultats significatifs. Cette fonction modifie la perte standard en incorporant un terme de focalisation qui ajuste la contribution de chaque exemple selon sa probabilité de classification correcte. Ce mécanisme diminue l'impact des exemples correctement classés sur la perte globale, permettant ainsi une attention accrue aux exemples mal classés. Toutefois, comme le souligne (Cao et al., 2019), son efficacité diminue dans les tâches de classification d'images caractérisées par un fort déséquilibre entre les classes.

2.1.2 Label-distribution-aware-margin loss (LDAM)

La fonction de perte LDAM (Cao et al., 2019) est une technique de régularisation qui vise à élargir les marges pour les classes minoritaires en étendant la perte de marge douce standard afin de permettre un meilleur compromis entre les marges de toutes les classes, conduisant à une meilleure généralisation du modèle sur des jeux de données déséquilibrés, mais elle est connu pour la difficulté de son optimisation lors de la phase d'entraînement.

2.2 Approches basées sur le rééchantillonage

2.2.1 Easy-Data-Augmentation (EDA)

EDA (Wei et Zou, 2019) est une technique basée sur des règles pour l'augmentation de données textuelles. Pour chaque phrase d'un ensemble de données d'entraînement, la technique EDA sélectionne aléatoirement une parmi les opérations suivantes : remplacement de synonymes, insertion aléatoire, permutation aléatoire ou suppression aléatoire. Cette approche vise à diversifier le corpus de formation en introduisant des variations lexicales et syntaxiques, contribuant ainsi à améliorer la robustesse et la généralisation des modèles de traitement de langage naturel. Néanmoins, il convient de noter que l'application de EDA peut entraîner un sur-apprentissage, particulièrement lorsque la diversité initiale des données est restreinte. En outre, les modifications induites par EDA peuvent compromettre l'efficacité de la classification si elles affectent de manière significative le contexte sémantique des phrases, ce qui est particulièrement problématique dans des domaines nécessitant une grande précision lexicale.

2.2.2 Mixup et ses variantes

Mixup (Guo et al., 2019) : est une méthode d'augmentation de données qui interpole linéairement deux exemples aléatoires et leurs étiquettes associées, initialement développée pour la classification d'images mais applicable à d'autres domaines. Les exemples (x_i, y_i) et (x_j, y_j) sont tirés aléatoirement de l'ensemble des données d'entraînement, et les étiquettes sont au format one-hot. Le coefficient λ est sélectionné de manière aléatoire suivant une distribution Beta(α, α), où $\alpha > 0$. Les nouvelles instances et leurs étiquettes sont calculées comme suit :

Mixup pour les petits ensembles de données déséquilibrés

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

Cette méthode génère des étiquettes "soft" et augmente la diversité des données, améliorant ainsi la généralisation et réduisant le risque de surajustement. Mixup peut être appliqué aux représentations vectorielles à différents niveaux dans les modèles de traitement du langage, incluant la couche de *Embeddings* des mots de chaque phrase ainsi que la couche avant-dernière précédant la fonction softmax, qui offre une représentation intégrale de la phrase *(Sentence Embedding)*.

Non-linear Mixup (Guo, 2020) : La technique **Nonlinear Mixup** est une version avancée du Mixup, qui se distingue par son application spécifique aux *embeddings* de mots. Elle utilise une politique de mélange distincte pour chaque dimension des mots dans une phrase, employant une matrice Λ, où chaque élément de cette matrice est indépendamment tiré d'une distribution Beta(α, α). Cette approche produit des échantillons synthétiques dans un espace plus vaste que celui du Mixup traditionnel, améliorant ainsi la régularisation et imposant des contraintes plus fortes. En outre, le Nonlinear Mixup intègre une *fonction de mappage* qui permet de calculer la matrice de mixage des représentations vectorielles des étiquettes (label-embeddings). Les nouvelles représentations vectorielles des exemples générées et leurs étiquettes correspondantes sont définies comme suit :

$$\tilde{B}^{ij} = \Lambda \odot B^i + (1 - \Lambda) \odot B^j$$

Où $\tilde{B}^{ij}$ représente la représentation mixte des caractéristiques de l'exemple généré, B^i et B^j sont les matrices des caractéristiques des exemples originaux i et j respectivement, et Λ est la matrice des coefficients de mélange pour les caractéristiques, appliquée élément par élément (element-wise).

$$\tilde{z}^{ij} = \Phi z^i + (1 - \Phi) z^j$$

$\tilde{z}^{ij}$ est l'étiquette mixte généré, z^i et z^j sont les matrice des caractéristiques des étiquettes originaux (label embeddings) i et j, tandis que Φ est la matrice des coefficients de mélange pour les étiquettes, choisie indépendamment de Λ et calculé a partir de $\tilde{B}^{ij}$.

Remix (Chou et al., 2020) : La technique Remix représente une amélioration de la technique Mixup pour l'adapter au scenario des jeux de données déséquilibrés en dissociant les facteurs de mélange pour les caractéristiques et les étiquettes. Cette séparation permet d'attribuer un poids supérieur aux classes minoritaires dans les étiquettes, favorisant ainsi leur représentation. Remix offre une stratégie ciblée pour améliorer la généralisation des modèles face à des déséquilibres importants entre les classes, en ajustant les contributions des classes selon leur fréquence relative dans le jeu de données.

$$\tilde{x}^{RM} = \lambda_x x_i + (1 - \lambda_x) x_j$$

$$\tilde{y}^{RM} = \lambda_y y_i + (1 - \lambda_y) y_j$$

Le coefficient λ_x est échantillonné à partir d'une distribution Beta. Le coefficient λ_y est ensuite défini par l'équation ci-dessous, où n_i et n_j désignent respectivement le nombre d'exemples appartenant aux classes i et j :

$$\lambda_y = \begin{cases} 0, & \text{if } \frac{n_i}{n_j} \geq \kappa \text{ and } \lambda < \tau, \\ 1, & \text{if } \frac{n_i}{n_j} \leq \frac{1}{\kappa} \text{ and } 1 - \lambda < \tau, \\ \lambda, & \text{otherwise.} \end{cases}$$

Un échantillon (x_i, y_i) est considéré comme κ-*majoritaire* par rapport à un échantillon (x_j, y_j) si $\frac{n_i}{n_j} \geq \kappa$. Les paramètres κ et τ sont des hyperparamètres de cette méthode. Cette formulation permet d'ajuster λ_y afin de refléter la proportion de chaque classe dans l'échantillon.

Mamix (Cheng et al., 2023) : est aussi une adaptation du Mixup traditionnel pour le contexte de déséquilibre des donnés en introduisant une modulation de λ_y basée sur la représentativité des classes dans les échantillons impliqués, visant à améliorer la représentation des classes minoritaires. Cette approche est similaire à la technique **Remix**, mais elle propose une relation entre λ_y et λ_x. Le mélange est effectué comme suit :

$$\tilde{x}^{MAM} = \lambda_x x_i + (1 - \lambda_x) x_j$$

Pour les étiquettes, λ_y est ajusté afin d'équilibrer de manière optimale les contributions des classes, en accordant une marge plus large aux classes minoritaires, conformément à l'équation suivante :

$$\tilde{y}^{MAM} = \lambda_y y_i + (1 - \lambda_y) y_j$$

$$\lambda_y = \begin{cases} (1 - \lambda_x) \times 0.5 & \text{if } \lambda_x \geq \frac{\eta_j}{(\eta_i + \eta_j)}, \\ \frac{\eta_i}{(\eta_i + \eta_j)} & \text{if } \lambda_x < \frac{\eta_j}{(\eta_i + \eta_j)}. \end{cases}$$

3 Expérimentations

Dans le cadre de notre étude, nous avons appliqué diverses techniques d'augmentation de données basées sur le concept de Mixup pour évaluer leur efficacité en classification de texte. Nous avons adapté les méthodes suivantes pour fonctionner tant au niveau des embeddings de mots qu'au niveau des embeddings de phrases :

— Mixup : Implémenté pour les Embeddings des mots et des phrases, tel que décrit par (Guo et al., 2019).
— Remix : Adapté du travail de (Bellinger et al., 2020) initialement présenté pour la classification d'images, et appliqué à la classification de textes, conformément à l'approche décrite par (Chou et al., 2020).
— MaMix : Adapté de la méthode proposée par (Cheng et al., 2023) pour la classification de textes, issue à l'origine d'une technique de classification d'images.
— Mixup Non-linéaire : Adapté de l'approche de (Guo, 2020), initialement conçue pour les Embeddings de mots, et modifiée pour être applicable aux Embeddings de phrases, avec des ajustements spécifiques à notre contexte d'étude

3.1 Jeux de données utilisées

Pour évaluer ces méthodes, nous avons mené des expérimentations sur des jeux de données initialement équilibrés, dont les caractéristiques sont présentées dans le tableau 1.

1. **SUBJ** (Pang et Lee, 2004) : Ce jeu de données classifie les phrases comme subjectives ou objectives.
2. **TREC** (Li et Roth, 2002) : Ce corpus catégorise des questions en six types distincts.
3. **SST-5** (Socher et al., 2013) : Il s'agit du Stanford Sentiment Treebank, qui catégorise les phrases en cinq classes de sentiments (très positif, positif, neutre, négatif, très négatif). Les données proviennent de critiques de films et incluent des annotations émotionnelles.
4. **SST-2** (Socher et al., 2013) : C'est une variante binaire du Stanford Sentiment Treebank, avec uniquement deux étiquettes (positive et négative).
5. **MR** (Pang et Lee, 2005) : Ce jeu de données classifie les critiques de films en deux catégories (positive et négative).

Data	Label	Train	Test	Val
SUBJ	2	8000	1000	1000
MR	2	8528	1068	1066
SST-2	2	6920	1821	872
SST-5	5	8544	2210	1101
TREC	6	4952	500	X

TAB. 1 – *Caractéristiques des jeux de données équilibrés utilisés. Le symbole **X** indique l'absence de jeu de validation distinct ; un sous-ensemble (10%) du jeu d'entraînement a été échantillonné pour servir de validation.*

Nous avons artificiellement déséquilibrés ces cinq datasets avec différents ratios de déséquilibre (imbalance ratio) ($ir = \{10, 50, 100\}$) de type "long-tailed".

Nous avons également évalué ces méthodes sur des jeux de données de grande taille et naturellement déséquilibrés, tels que **NG20** et **R8**. Nous avons ensuite réduit leur taille en appliquant divers ratios d'échantillonnage ($sr = \{10, 50\}$) afin de tester la robustesse des approches dans différents contextes de déséquilibre. Enfin, nous avons examiné un petit ensemble de données spécifiquement déséquilibré, constitué de 2000 phrases analysant la légitimité de start-up, afin d'évaluer l'applicabilité de ces méthodes dans un contexte hautement spécialisé. Les caractéristiques initiales de ces jeux de données sont présentées dans le tableau 2.

Data	Label	Train	Test	ir
NG20	20	11293	7528	1.59
R8	8	5485	2189	69.26
Legitim	2	1528	382	11.16

TAB. 2 – *Caractéristiques des jeux de données déséquilibré utilisés. ir : représente le ratio de déséquilibre initial des datasets*

3.2 Détails d'implémentation

Pour nos expérimentations, nous avons utilisé la variante *BERT-large* avec un lot (*batch size*) de 8, une longueur maximale de 256 tokens et un entraînement sur 20 époques, assorti d'un *early stopping*. Chaque configuration a été exécutée 5 fois avec différentes seed aléatoirs. Nous avons employé l'optimiseur *Adam* avec un taux d'apprentissage de $2e-5$, un *warmup ratio* de 0,1 et un *weight decay* de $1e-5$.

— **Mixup** : $\alpha = 0.1$, conformément à l'article de référence.
— **Remix** : k-majority $= 3$ et $\tau = 0.5$, suivant les spécifications de l'article original.
— **MaMix** : ratio $= -0.25$, suivant les recommandations de l'article original.
— **Non-linear Mixup** : dimension des étiquettes fixée à 100, telle que décrite dans la méthode de référence.

3.3 Résultats

Cette section examine les performances des différentes variantes de Mixup comparées aux approches classiques (EDA, Focal Loss, LDAM) sur divers jeux de données de classification de texte. Les résultats montrent que les méthodes basées sur Mixup, en particulier Sent-Mixup et Sent-Remix, surpassent de manière significative les performances de BERT seul ainsi que des configurations BERT+FL et BERT+LDAM. Ces approches démontrent une grande robustesse, même face à des ratios de déséquilibre élevés (ir) et sur des ensembles de données comme SUBJ et SST-5. Par exemple, dans ces jeux de données, Sent-Remix affiche des performances supérieures, même pour des ratios de déséquilibre atteignant 100, où d'autres approches enregistrent des baisses marquées de leurs scores F1-macro.

En ce qui concerne les ensembles de données avec des ratios modérés de déséquilibre, tels que NG20, Word-Mixup et Sent-Mixup maintiennent des scores compétitifs, bien que les résultats montrent une légère supériorité de Sent-Mixup dans les scénarios des ensemebles de données les plus déséquilibrés et de petites tailles (sr=10). Par contraste, des méthodes comme BERT+EDA, bien qu'elles améliorent la robustesse par rapport à BERT seul, n'atteignent pas la même stabilité de performance dans les cas de déséquilibre extrême. Cela est particulièrement visible dans des ensembles comme MR et SUBJ, où les améliorations apportées par EDA se limitent aux ratios faibles de déséquilibre.

Les expérimentations sur des jeux de données déséquilibrés échantillonnés confirment que les variantes de Mixup, telles que Word-Mixup, Sent-Mixup, Word-Remix, et Sent-Remix, offrent des performances robustes et constantes. Cela est particulièrement remarquable pour des ensembles de données tels que Legitim, où Sent-Remix affiche une nette amélioration par rapport aux méthodes traditionnelles, même dans des scénarios où l'équilibre entre classes est extrêmement défavorable.

En revanche, bien que les approches traditionnelles comme BERT+FL et BERT+LDAM se montrent efficaces dans certains cas spécifiques (par exemple, sur des ensembles équilibrés comme SUBJ), elles n'offrent pas une amélioration systématique et leur performance tend à se dégrader significativement avec l'augmentation du ratio de déséquilibre. À titre d'exemple, sur SST-5 et MR, les scores de F1-macro pour BERT+LDAM diminuent drastiquement lorsque le déséquilibre passe à un ratio élevé (ir=100), soulignant ainsi la limitation de ces approches dans ces contextes.

Enfin, les résultats confirment que l'enrichissement du contexte sémantique grâce à des techniques comme BERT+EDA peut offrir une amélioration modérée de la robustesse pour les ensembles de données de petite taille, mais que les variantes de Mixup restent les plus performantes globalement. Ces dernières, en combinant des augmentations basées sur le contexte et des stratégies de rééquilibrage efficaces, représentent une solution prometteuse pour gérer des scénarios de déséquilibre extrême dans les tâches de classification de texte.

Dataset	ir	BERT	BERT+ FL	BERT+ LDAM	BERT+ EDA	Word-Mixup	Sent-Mixup	Word-Remix	Sent-Remix	Word-Mamix	Sent-Mamix	Word-Nonlinear	Sent-Nonlinear
SST-2	org	93.23 ± 0.003	93.15 ± 0.004	92.77 ± 0.003	93.12 ± 0.002	92.51 ± 0.005	$\mathbf{93.62 \pm 0.002}$	92.88 ± 0.003	93.31 ± 0.003	82.66 ± 0.14	92.55 ± 0.006	88.77 ± 0.06	93.18 ± 0.005
	10	87.11 ± 0.009	87.31 ± 0.01	87.77 ± 0.01	48.91 ± 0.06	84.06 ± 0.05	76.94 ± 0.2	86.80 ± 0.02	$\mathbf{88.40 \pm 0.01}$	44.92 ± 0.2	86.45 ± 0.01	41.71 ± 0.1	87.01 ± 0.01
	50	67.48 ± 0.05	76.11 ± 0.06	69.13 ± 0.05	48.65 ± 0.06	78.41 ± 0.06	74.05 ± 0.1	81.05 ± 0.03	79.61 ± 0.04	35.44 ± 0.02	$\mathbf{82.19 \pm 0.04}$	33.29 ± 0.0001	33.32 ± 0.0009
	100	57.85 ± 0.05	67.44 ± 0.06	49.54 ± 0.17	47.11 ± 0.07	67.03 ± 0.19	$\mathbf{77.55 \pm 0.01}$	56.35 ± 0.22	66.08 ± 0.19	34.009 ± 0.009	59.27 ± 0.24	33.43 ± 0.003	33.3 ± 0.0005
SST-5	org	53.49 ± 0.008	46.95 ± 0.14	51.78 ± 0.007	50.11 ± 0.05	50.32 ± 0.05	52.67 ± 0.009	52.39 ± 0.02	$\mathbf{53.96 \pm 0.004}$	34.62 ± 0.1	52.03 ± 0.004	29.61 ± 0.11	51.41 ± 0.01
	10	41.54 ± 0.01	47.57 ± 0.009	47.19 ± 0.02	15.99 ± 0.04	46.5 ± 0.02	47.87 ± 0.02	45.82 ± 0.01	$\mathbf{48.45 \pm 0.01}$	21.91 ± 0.08	34.65 ± 0.1	14.14 ± 0.08	25.92 ± 0.03
	50	23.5 ± 0.003	$\mathbf{32.44 \pm 0.03}$	31.4 ± 0.06	16.21 ± 0.02	28.15 ± 0.07	33.96 ± 0.08	23.59 ± 0.1	32.41 ± 0.05	14.01 ± 0.07	19.79 ± 0.05	$7.5 \pm 1e-05$	23.11 ± 0.003
	100	19.88 ± 0.06	24.07 ± 0.09	25.2 ± 0.01	15.70 ± 0.03	23.55 ± 0.1	28.006 ± 0.03	25.89 ± 0.03	$\mathbf{29.32 \pm 0.02}$	8.61 ± 0.01	21.88 ± 0.06	7.5 ± 0.0	22.36 ± 0.008
TREC	org	95.97 ± 0.01	83.19 ± 0.3	96.41 ± 0.005	96.29 ± 0.005	96.05 ± 0.004	95.94 ± 0.005	96.42 ± 0.008	$\mathbf{97.12 \pm 0.006}$	91.18 ± 0.02	95.54 ± 0.009	62.33 ± 0.3	94.66 ± 0.009
	10	$\mathbf{96.2 \pm 0.009}$	95.09 ± 0.01	92.44 ± 0.02	18.63 ± 0.09	94.18 ± 0.01	94.78 ± 0.01	94.52 ± 0.01	94.89 ± 0.01	66.6 ± 0.2	92.00 ± 0.02	8.22 ± 0.06	88.72 ± 0.1
	50	79.08 ± 0.1	88.37 ± 0.07	86.21 ± 0.08	12.13 ± 0.03	65.44 ± 0.2	91.54 ± 0.01	80.09 ± 0.07	$\mathbf{92.11 \pm 0.01}$	25.91 ± 0.1	53.94 ± 0.3	8.23 ± 0.06	20.23 ± 0.007
	100	53.12 ± 0.1	$\mathbf{81.47 \pm 0.06}$	69.47 ± 0.2	18.87 ± 0.08	39.33 ± 0.2	70.96 ± 0.1	56.75 ± 0.2	80.78 ± 0.07	11.15 ± 0.03	25.46 ± 0.06	5.27 ± 0.0	16.04 ± 0.06
SUBJ	org	97.4 ± 0.002	97.18 ± 0.004	97.29 ± 0.02	96.85 ± 0.002	97.03 ± 0.003	97.17 ± 0.003	96.98 ± 0.001	$\mathbf{97.45 \pm 0.001}$	93.33 ± 0.02	96.48 ± 0.006	96.33 ± 0.004	97.1 ± 0.002
	10	$\mathbf{96.19 \pm 0.001}$	96.03 ± 0.003	95.98 ± 0.003	96.26 ± 0.003	95.68 ± 0.003	95.9 ± 0.001	95.58 ± 0.005	95.55 ± 0.007	92.4 ± 0.01	94.89 ± 0.007	85.51 ± 0.06	95.8 ± 0.002
	50	93.34 ± 0.008	94.28 ± 0.008	92.09 ± 0.02	70.79 ± 0.09	93.84 ± 0.01	93.46 ± 0.01	93.42 ± 0.01	$\mathbf{94.33 \pm 0.006}$	89.53 ± 0.02	93.27 ± 0.01	32.82 ± 0.001	72.24 ± 0.2
	100	88.41 ± 0.03	92.30 ± 0.01	70.79 ± 0.09	78.85 ± 0.07	92.05 ± 0.01	92.8 ± 0.02	74.83 ± 0.2	$\mathbf{92.35 \pm 0.01}$	59.91 ± 0.2	91.33 ± 0.01	32.99 ± 0.003	42.21 ± 0.2
MR	org	86.93 ± 0.006	87.36 ± 0.004	87.10 ± 0.005	86.55 ± 0.008	86.42 ± 0.01	86.59 ± 0.005	86.15 ± 0.01	86.55 ± 0.004	77.41 ± 0.1	86.16 ± 0.003	86.42 ± 0.01	$\mathbf{87.96 \pm 0.006}$
	10	80.66 ± 0.01	78.41 ± 0.04	82.09 ± 0.01	44.02 ± 0.07	75.66 ± 0.05	82.13 ± 0.01	71.89 ± 0.1	$\mathbf{82.33 \pm 0.02}$	50.13 ± 0.1	80.12 ± 0.02	33.44 ± 0.001	79.88 ± 0.01
	50	58.38 ± 0.03	71.08 ± 0.05	58.74 ± 0.1	44.85 ± 0.09	55.96 ± 0.2	$\mathbf{72.57 \pm 0.06}$	61.72 ± 0.1	68.55 ± 0.03	36.4 ± 0.04	56.95 ± 0.2	33.33 ± 0.0	33.37 ± 0.0009
	100	43.19 ± 0.09	53.00 ± 0.1	50.88 ± 0.1	47.04 ± 0.1	42.59 ± 0.1	$\mathbf{69.76 \pm 0.05}$	56.84 ± 0.1	41.96 ± 0.1	33.81 ± 0.009	49.07 ± 0.1	33.37 ± 0.0009	33.37 ± 0.0009

TAB. 3 – *Table de comparaison de F1-macro pour les différents modèles sur des données artificiellement déséquilibrées avec différent rapport de déséquilibre (ir)*

Dataset	sr	BERT	BERT+ FL	BERT+ LDAM	BERT+ EDA	Word-Mixup	Sent-Mixup	Word-Remix	Sent-Remix	Word-Mamix	Sent-Mamix	Word-Nonlinear	Sent-Nonlinear
NG20	org	84.6 ± 0.004	$\mathbf{84.7 \pm 0.004}$	84.16 ± 0.003	84.68 ± 0.002	83.6 ± 0.005	84.23 ± 0.006	83.29 ± 0.004	84.54 ± 0.003	40.53 ± 0.3	81.71 ± 0.01	51.38 ± 0.2	83.17 ± 0.003
	50	81.07 ± 0.007	81.43 ± 0.002	80.47 ± 0.008	$\mathbf{82.24 \pm 0.006}$	79.3 ± 0.009	81.22 ± 0.005	79.27 ± 0.005	81.38 ± 0.004	63.4 ± 0.05	77.16 ± 0.01	28.01 ± 0.2	77.7 ± 0.01
	10	53.21 ± 0.2	67.26 ± 0.02	56.25 ± 0.2	$\mathbf{75.16 \pm 0.008}$	43.33 ± 0.1	64.55 ± 0.02	48.28 ± 0.07	61.92 ± 0.05	14.95 ± 0.04	31.13 ± 0.1	11.49 ± 0.007	13.14 ± 0.06
R8	org	95.01 ± 0.004	94.61 ± 0.005	94.24 ± 0.009	95.18 ± 0.001	93.51 ± 0.01	94.37 ± 0.07	94.86 ± 0.01	94.61 ± 0.003	77.07 ± 0.3	93.65 ± 0.006	54.85 ± 0.2	$\mathbf{95.28 \pm 0.004}$
	50	93.65 ± 0.007	94.05 ± 0.006	92.83 ± 0.005	94.23 ± 0.003	93.09 ± 0.02	94.12 ± 0.008	93.92 ± 0.006	$\mathbf{94.29 \pm 0.005}$	70.46 ± 0.05	92.53 ± 0.009	25.28 ± 0.05	79.36 ± 0.1
	10	32.51 ± 0.05	57.45 ± 0.09	56.21 ± 0.1	$\mathbf{90.37 \pm 0.01}$	47.45 ± 0.1	60.95 ± 0.1	48.07 ± 0.06	54.70 ± 0.1	23.45 ± 0.06	34.16 ± 0.1	$8.27 \pm 1.1e-05$	20.73 ± 0.01
Legitim	org	57.54 ± 0.09	59.45 ± 0.1	67.79 ± 0.03	48.65 ± 0.01	56.94 ± 0.09	59.7 ± 0.02	62.36 ± 0.09	$\mathbf{68.74 \pm 0.1}$	$47.68 \pm 6.2e-17$	56.12 ± 0.1	47.66 ± 0.0003	$47.68 \pm 6.2e-17$

TAB. 4 – *Table de comparaison de F1-macro pour les différents modèles sur des données déséquilibrées échantillonnées avec différent ratio (sr)*

4 Discussion

Bien que les techniques de Mixup aient montré une efficacité notable dans la gestion des déséquilibres de classes, elles présentent certaines limites importantes liées à l'incertitude des étiquettes, ce qui peut justifier l'instabilité des variantes de Mixup dans les expérimentations sur les données déséquilibrées de petites tailles. En effet, dans le Mixup, les étiquettes sont dérivées par interpolation linéaire plutôt que par des déductions basées sur une logique explicite. Cette méthode d'interpolation peut introduire un biais dans la décision du modèle, représentant un défi réel dans l'application de cette technique, comme souligné par (Xie et al., 2023). Cette ambiguïté dans les étiquettes peut compromettre la clarté des frontières de classe, ce qui est particulièrement problématique dans le traitement des données déséquilibrées de petite taille.

Pour pallier ces défis, il est crucial d'améliorer la gestion de l'ambiguïté des étiquettes dans les applications de Mixup. Il est également essentiel de calibrer soigneusement le jeu de données en manipulant le paramètre d'interpolation pour éviter la dilution des caractéristiques des classes minoritaires. La sélection aléatoire des exemples à mixer peut en effet conduire à une représentation inadéquate des classes minoritaires, exacerbant ainsi le problème de déséquilibre initial. Ainsi, une approche plus stratégique dans le choix des paires d'exemples à combiner pourrait aider à maintenir une représentation équilibrée des classes au sein du jeu de données synthétisé. Les recherches futures doivent donc explorer ces enjeux en développant des stratégies pour atténuer l'impact négatif de l'ambiguïté des données synthétiques et améliorer la précision de la classification dans des contextes d'application réels.

5 Conclusion

Ce travail a abordé l'efficacité des techniques de Mixup sur des ensembles de données textuelles déséquilibrés de petite taille, en démontrant que des adaptations telles que Word-Mixup, Sent-Mixup, et leurs variantes comme Remix et MaMix, surpassent les méthodes traditionnelles de traitement des déséquilibres de classes. Ces techniques se distinguent par leur capacité à améliorer la généralisation des modèles face à des défis tels que la réduction du volume des données et des déséquilibres significatifs entre les classes.

Toutefois, l'interpolation des étiquettes par ces méthodes introduit une certaine incertitude qui peut affecter la clarté des décisions du modèle. Il est donc essentiel de poursuivre le développement et l'affinement de ces approches pour optimiser leur efficacité et leur applicabilité, en particulier dans des scénarios complexes où les données sont à la fois limitées et déséquilibrées.

Références

Bellinger, C., R. Corizzo, et N. Japkowicz (2020). Remix : Calibrated resampling for class imbalance in deep learning. *arXiv preprint arXiv :2012.02312*.

Cao, K., C. Wei, A. Gaidon, N. Arechiga, et T. Ma (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems 32*.

Chawla, N. V., K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer (2002). Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Cheng, W.-C., T.-H. Mai, et H.-T. Lin (2023). From smote to mixup for deep imbalanced classification. In *International Conference on Technologies and Applications of Artificial Intelligence*, pp. 75–96. Springer.

Chou, H.-P., S.-C. Chang, J.-Y. Pan, W. Wei, et D.-C. Juan (2020). Remix : rebalanced mixup. In *Computer Vision–ECCV 2020 Workshops : Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 95–110. Springer.

Cui, Y., M. Jia, T.-Y. Lin, Y. Song, et S. Belongie (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277.

Dablain, D., B. Krawczyk, et N. V. Chawla (2022). Deepsmote : Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems 34*(9), 6390–6404.

Guo, H. (2020). Nonlinear mixup : Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 34, pp. 4044–4051.

Guo, H., Y. Mao, et R. Zhang (2019). Augmenting data with mixup for sentence classification : An empirical study. *arXiv preprint arXiv :1905.08941*.

Huang, C., Y. Li, C. C. Loy, et X. Tang (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384.

Kobayashi, S. (2018). Contextual augmentation : Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv :1805.06201*.

Kubat, M., S. Matwin, et al. (1997). Addressing the curse of imbalanced training sets : one-sided selection. In *Icml*, Volume 97, pp. 179. Citeseer.

Kurt Pehlivanoğlu, M., R. T. Gobosho, M. A. Syakura, V. Shanmuganathan, et L. de-la Fuente-Valentín (2024). Comparative analysis of paraphrasing performance of chatgpt, gpt-3, and t5 language models using a new chatgpt generated dataset : Paragpt. *Expert Systems 41*(11), e13699.

Li, X. et D. Roth (2002). Learning question classifiers. In *COLING 2002 : The 19th International Conference on Computational Linguistics*.

Pang, B. et L. Lee (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Pang, B. et L. Lee (2005). Seeing stars : Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Socher, R., J. Bauer, C. D. Manning, et A. Y. Ng (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pp. 455–465.

Wei, J. et K. Zou (2019). Eda : Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv :1901.11196*.

Witteveen, S. et M. Andrews (2019). Paraphrasing with large language models. *arXiv preprint arXiv :1911.09661*.

Wu, X., S. Lv, L. Zang, J. Han, et S. Hu (2019). Conditional bert contextual augmentation. In *Computational Science–ICCS 2019 : 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pp. 84–95. Springer.

Xie, X., L. Yangning, W. Chen, K. Ouyang, Z. Xie, et H.-T. Zheng (2023). Global mixup : Eliminating ambiguity with clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 37, pp. 13798–13806.

Zhang, H. (2017). mixup : Beyond empirical risk minimization. *arXiv preprint arXiv :1710.09412*.

Summary

Text classification is a crucial task in natural language processing (NLP), often complicated by significant class imbalances in real datasets. This imbalance hampers the effectiveness of deep learning classifiers, especially for minority classes, which tend to be underrepresented.

To address these challenges, various strategies such as loss function re-weighting and data augmentation techniques have been explored. These approaches have proven useful in contexts of class imbalance, but their impact on small datasets remains underexplored.

Our study systematically evaluates these methodologies on well-established text classification datasets that we have artificially imbalanced. We also assess their scalability and performance through experiments on large imbalanced datasets reduced by sampling, as well as on a naturally imbalanced and small dataset. The aim is to determine the effectiveness of these techniques specifically in the context of small imbalanced datasets, to better understand their applicability and optimize their implementation.

Prédiction de la Peur de la Récidive du Cancer du sein à partir des données de remboursement de soins de santé

Mamoudou KOUME*, Lorène SEGUIN**
Anne-Déborah BOUHNIK*, Raquel URENA*

*Marseille, France
prenom.nom@univ-amu.fr,
**Marseille, France
seguinl@ipc.unicancer.fr

Résumé. Le cancer du sein est le cancer le plus fréquent et la première cause de mortalité chez les femmes en France, avec des répercussions profondes sur le bien-être physique, émotionnel et psychologique. Malgré les avancées thérapeutiques, la peur de la récidive du cancer (PRC) persiste chez les survivantes, entraînant une augmentation potentielle de l'utilisation des soins de santé et une diminution du bien-être global. Cependant, la prédiction précise de la probabilité de PRC à l'aide de modèles statistiques traditionnels et d'algorithmes d'apprentissage automatique (ML) reste un défi en raison de la complexité des interactions des données de santé au fil du temps. Dans cette étude, nous proposons une approche basée sur des modèles prédictifs utilisant des réseaux de neurones et les données de remboursement de soins de santé pour prédire la probabilité de PRC chez les femmes cinq ans après le diagnostic. Notre méthode intègre des informations temporelles et exploite à la fois des données médico-administratives à l'aide d'architectures de réseaux de neurones (NN) et de techniques d'apprentissage supervisé. Cette approche offre une perspective prometteuse pour des évaluations personnalisées du risque et des stratégies d'intervention adaptées aux survivantes du cancer du sein. Les résultats expérimentaux en cours, obtenus à l'aide de modèles ML et NN, montrent des performances encourageantes, suggérant que les réseaux de neurones pourraient améliorer davantage la précision des prédictions.

1 Introduction

Le cancer du sein est l'un des cancers les plus répandus dans le monde et la première cause de mortalité chez les femmes en France, avec des impacts physiques, émotionnels et psychologiques significatifs, notamment une anxiété et une détresse liées à la récidive du cancer. Ces préoccupations peuvent compromettre le bien-être mental et la qualité de vie globale des survivantes. Malgré les progrès réalisés dans les traitements oncologiques, la peur de la récidive du cancer (PRC) demeure une inquiétude persistante, en particulier dans les années qui suivent le diagnostic Magnani et al. (2022). Comprendre et atténuer cette PRC est crucial pour améliorer le bien-être global des survivantes. Des études ont montré que la PRC peut persister de nombreuses années après le traitement, entraînant une utilisation accrue des services de santé, une diminution de l'observance des rendez-vous de suivi et une dégradation du bien-être général Otto et al. (2018). Par ailleurs, Roorda et al. (2013) observe une augmentation de l'utilisation des soins primaires chez les femmes ayant des antécédents de cancer du sein, particulièrement chez les femmes âgées. Identifier les individus à risque élevé de PRC et mettre en œuvre des interventions ciblées est essentiel pour en atténuer les conséquences négatives et améliorer les résultats post-cancer.

Bien que de nombreuses études aient exploré les facteurs contribuant à la PRC, il existe encore un manque dans la prédiction efficace de la probabilité de souffrir de PRC Koume et al. à l'aide de techniques computationnelles avancées. Dans ce contexte, l'utilisation de modèles d'apprentissage profond (DL) pour analyser les informations de santé des patients, telles que les dossiers de santé électroniques (EHRs), a montré des avantages significatifs par rapport aux approches statistiques traditionnelles et aux algorithmes de ML, souvent limités par la complexité des facteurs cliniques et démographiques au fil du temps. D'autres études ont utilisé des réseaux neuronaux pour prédire le risque de maladie, le pronostic et les résultats des patients Miotto et al. (2018, 2016); Kourou et al. (2014). Par exemple, TabMLPNet, un modèle de réseau neuronal, a été conçu pour analyser les antécédents cliniques et identifier les immunodéficiences primaires chez les adultes à partir de données nationales de remboursement de soins médicaux Papanastasiou et al. (2023).

Dans le cadre de l'analyse des EHRs, les réseaux neuronaux récurrents (RNN) ont été utilisés pour prédire divers résultats, tels que les diagnostics futurs, les traitements médicamenteux et les actes médicaux lors des visites médicales suivantes, en utilisant des codes médicaux comme entrées. Ces modèles permettent des prédictions multilabels basées sur les antécédents des patients Choi et al. (2015). Un modèle appelé Timeline Bai et al. (2018), conçu pour la modélisation prédictive des EHRs à partir des données de remboursement de soins médicaux, intègre une architecture DL interprétable. En apprenant les facteurs de décroissance temporelle pour chaque code médical et les poids d'attention pour les visites, ce modèle prédit avec précision les catégories de diagnostic primaire pour les futures visites hospitalières tout en fournissant des informations sur l'évolution temporelle des conditions des patients. De plus, dans Finch et al. (2021), quatre modèles attentionnels ont été développés à partir des dossiers des patients pour évaluer leurs performances dans la prédiction de la mortalité et des hospitalisations.

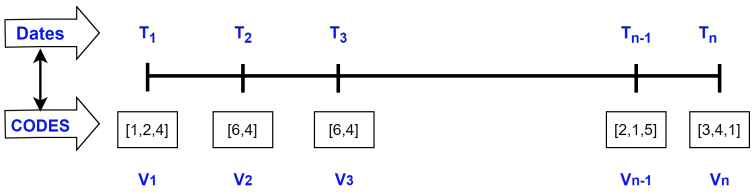


FIG. 1 – *La représentation des visites médicales d'un patient, notées $V_{i=1,\ldots,n}$, correspond aux visites ayant lieu aux dates $T_{i=1,\ldots,n}$. À chaque visite, le patient peut recevoir des codes médicaux associés aux médicaments, aux actes biologiques et médicaux.*

Cependant, à notre connaissance, aucune étude antérieure ne s'est spécifiquement concentrée sur le développement d'un modèle basé sur les réseaux neuronaux pour prédire la PRC à partir des données de remboursement de soins de santé. Cette étude vise à combler cette lacune en proposant un modèle prédictif qui intègre des informations temporelles et exploite des données médico-administratives en utilisant des architectures RNN et des techniques d'apprentissage supervisé. Notre hypothèse suggère que la PRC est liée à l'anxiété et à la surconsommation médicale, qui peuvent se manifester par une utilisation fréquente de médicaments anxiolytiques et une surutilisation des actes biologiques et médicaux.

2 Matériels et Méthodes

Dans notre étude, nous avons utilisé un ensemble de données anonymisées comprenant les remboursements de soins de santé de survivantes du cancer du sein, collectés sur une période de cinq ans après le diagnostic et incluant la variable cible, la PRC. Ces données de remboursement de soins ont été obtenues à partir de l'enquête VICAN-5 (VIe après le CANcer), qui inclut une cohorte diversifiée de survivantes du cancer du sein diagnostiquées entre 2009 et 2011. Réalisée cinq ans après le diagnostic initial, cette enquête couvre un échantillon représentatif au niveau national de 4 174 survivants adultes résidant en France, âgées de 18 à 82 ans au moment de leur diagnostic initial INCa (2013, 2014). Les critères d'inclusion comprenaient les patients sans cancer métastatique et celles n'ayant pas connu de récidive ou de progression de leur maladie.

Les données incluent trois catégories principales de remboursement de soins de santé : médicaments, actes biologiques et médicaux, ainsi que la variable cible, la PRC. Cette dernière a été mesurée dans VICAN-5 à l'aide de la question suivante : « À quelle fréquence pensez-vous à la possibilité que la maladie récidive ? ». Les réponses ont été enregistrées sur une échelle de Likert à 5 points : jamais / quelques fois par mois / quelques fois par semaine / quelques fois par jour / plusieurs fois par jour. Ces réponses ont ensuite été regroupées en deux catégories : Non-PRC (jamais) et PRC (toutes les autres réponses).

Dans l'enquête VICAN-5, seules 930 participantes atteintes de cancer du sein ont été activement interrogées, avec des données collectées via des entretiens téléphoniques, incluant des informations sur la PRC et la consommation de soins de santé. En outre, nous disposons de données non labellisées pour 5110 patientes atteintes de cancer du sein, pour lesquelles seules les données de remboursements sont disponibles. En éliminant de notre étude les prescriptions de soins de santé émises au cours des deux premières années de traitement, nous avons exclu la phase directement influencée par le processus thérapeutique. Par ailleurs, nous avons retiré les prescriptions directement liées à l'hormonothérapie, en considérant qu'environ 70% des patientes poursuivent ce traitement dans le cas du cancer du sein. Cette décision repose sur des considérations cliniques et pratiques, en prenant en compte le fait que les deux premières années de traitement représentent une phase aiguë pendant laquelle l'hormonothérapie est généralement administrée. De plus, nous avons retiré les vaccins et d'autres données médicales jugées non pertinentes ou fréquemment utilisées dans la vie quotidienne, car elles ne sont pas congruentes avec l'investigation de la PRC. Après ces critères d'exclusion, le nombre de patients s'établit à 918.

Avant l'entraînement des modèles, nous avons effectué des étapes de prétraitement, telles que des tests statistiques et la sélection de variables, et avons traité le déséquilibre des données pour les algorithmes d'apprentissage automatique (ML) en construisant un pipeline complet. Ce pipeline intègre de manière fluide toutes les étapes de prétraitement, de rééquilibrage des données, de sélection de variable, d'entraînement des modèles et d'évaluation. Pour la sélection de variables, nous avons utilisé l'algorithme d'élimination récursive des variables (RFE), de meta-transformer (SFM) et de l'algorithme ReliefF. Afin de traiter le déséquilibre des données, nous avons employé les techniques Synthetic Minority Over-sampling Technique (SMOTE) et Adaptive Synthetic Sampling (ADASYN).

Pour les architectures de réseaux neuronaux (NN), nous avons construit des séquences médicales de patientes tout en préservant l'intégrité de la structure temporelle des séquences. Ensuite, pour chaque patiente, les données ont été organisées sous forme de séquence de visites médicales, où chaque visite est composée d'une séquence de codes médicaux. Dans ces cas, les codes médicaux ont été correctement encodés en entiers, et nous avons utilisé des pondérations par classe pour traiter le déséquilibre des classes, une technique qui consiste à attribuer des poids plus élevés aux classes sous-représentées et des poids plus faibles aux classes sur-représentées.

Pour analyser ces données, diverses techniques de Machine Learning et de réseaux de neurones ont été considérées, notamment : le Gradient Boosting Friedman (2001), le perceptron multicouche (MLP) Cybenko (1989), ainsi

que les réseaux de mémoire à long et court terme (LSTMs, pour Long Short-Term Memory) et leur variante bidirectionnelle, Bi-LSTM Hochreiter et Schmidhuber (1997); Graves et Schmidhuber (2005)

3 Expérimentations

Nous avons divisé l'ensemble de données en ensembles d'entraînement, de validation et de test en utilisant un ratio de 70%-15%-15%, en garantissant la continuité temporelle entre les différentes divisions. Les modèles GB et MLP ont été implémentés avec Scikit-learn, tandis que les modèles LSTM et Bi-LSTM ont été implémentés avec PyTorch et TensorFlow. Nous avons entrainé avec 100 epochs pour chaque modèle tout en surveillant la performance sur l'ensemble de validation pour éviter le surapprentissage. Nous avons évalué la performance des modèles à l'aide de mesures standards telles que la précision, le rappel, le score F1 et l'aire sous la courbe de ROC (AUC-ROC).

Le tableau 1 présente un résumé du profil de l'ensemble de données.

# Patients (P)	918	# Codes dans une visite	
# Codes médicaux (C)	443	Minimum	1
# Séquences de visites		Maximum	22
Visites totales (V)	56 501	Moyenne	2.59
Minimum	1	# Instances étiquetées	918
Maximum	217	PRC	582
Moyenne	83.92	Non-PRC	336
V/P	61.5	# Instances non étiquetées	5110

TAB. 1 – *Statistiques sur les données de consommation de soins de santé*

4 Résultats

Les résultats préliminaires de notre analyse en cours, résumés dans le tableau 2, fournissent des informations sur la performance comparative des différents modèles prédictifs. Il est à noter que les modèles GB, MLP, LSTM et Bi-LSTM présentent des métriques de performance prometteuses. Le tableau présente les principales métriques d'évaluation pour chaque modèle. Les modèles GB et MLP montrent des scores AUC comparables de 0,66, avec un rappel plus élevé pour GB à 0,70. Bien que les modèles LSTM et Bi-LSTM affichent des scores AUC inférieurs à ceux de GB et MLP, le modèle Bi-LSTM présente un rappel plus élevé, ce qui indique son potentiel pour identifier les vrais positifs. Ces résultats préliminaires et en cours servent de base pour de futures investigations et l'affinage des modèles prédictifs dans notre étude.

Modèles	AUC	Rappel	Précision	F1-score
GB	0.66	0.70	0.71	0.71
MLP	0.66	0.63	0.75	0.68
LSTM	0.56	0.65	0.67	0.66
Bi-LSTM	0.57	0.78	0.65	0.71

TAB. 2 – *Performance comparative des résultats de prédiction.*

5 Discussion

L'analyse en cours de divers modèles prédictifs pour la PRC promet d'améliorer notre compréhension des facteurs influençant la PRC chez les survivantes du cancer du sein. Les performances encourageantes des modèles de GB, MLP, LSTM et Bi-LSTM soulignent l'importance d'utiliser des techniques computationnelles avancées pour élaborer des plans de soins personnalisés pour les patientes atteintes de cancer du sein. Bien que le modèle GB soit actuellement le meilleur modèle en termes de performance, des recherches supplémentaires sur les modèles d'apprentissage profond, intégrant des dynamiques temporelles, pourraient conduire à des améliorations substantielles de la précision prédictive et de la performance globale.

Nos résultats expérimentaux ont plusieurs implications pour les soins de survie des patientes atteintes de cancer du sein. Le développement de modèles prédictifs précis pour la PRC peut aider les professionnels de santé à identifier les individus à risque élevé de PRC et à mettre en place des interventions ciblées pour atténuer leurs préoccupations. Cette avancée permet non seulement d'améliorer notre compréhension de la PRC à l'aide des données de remboursements, mais aussi de mettre en œuvre des programmes de dépistage à l'échelle de la population, dirigés par l'assurance maladie ou facilités par des initiatives d'autres instituts de santé publique. Cela sert aussi de preuve de concept que nous pourrons exploiter à l'avenir pour aborder des problèmes majeurs et multifactoriels tels que la récidive elle-même. En exploitant les données de remboursement de soins, les cliniciens peuvent personnaliser les plans de soins de survie et optimiser l'allocation des ressources pour répondre aux besoins uniques de chaque patiente. De plus, les techniques computationnelles avancées, telles que les modèles explorés dans cette étude, ouvrent de nouvelles perspectives pour des recherches futures sur les interactions complexes des codes médicaux influençant la PRC.

En outre, une évaluation supplémentaire en cours implique l'utilisation d'architectures de réseaux de neurones récurrents (RNN) intégrant un mécanisme d'attention et des techniques d'apprentissage semi-supervisé. Ces évaluations incluent des modèles tels que le REverse Time AttentIoN Mechanism (RETAIN), le Time-Aware Long-Short Term Memory (T-LSTM) et le Stacked Deep Generative Semi-Supervised Learning (SDGSSL). Ces méthodes utilisent à la fois des données medicoadministratives étiquetées et non étiquetées pour des interventions opportunes et des stratégies de soins personnalisées.

Références

Bai, T., B. L. Egleston, S. Zhang, et S. Vucetic (2018). Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 43–51.

Choi, E., M. T. Bahadori, et J. Sun (2015). Doctor AI : predicting clinical events via recurrent neural networks. *CoRR abs/1511.05942*.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems 2*(4), 303–314.

Finch, A., A. Crowell, Y. Chang, P. Parameshwarappa, J. Martinez, et M. Horberg (2021). A comparison of attentional neural network architectures for modeling with electronic medical records. *JAMIA Open 4*(3), ooab064.

Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *Annals of Statistics 29*(5), 1189–1232.

Graves, A. et J. Schmidhuber (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks 18*(5–6), 602–610.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

INCa (2013). *Plan Cancer 2009-2013*. France.

INCa (2014). *Plan Cancer 2014-2019*. France.

Koume, M., L. Seguin, A. Bouhnik, J. Mancini, M.-K. Bendiane, et R. Urena. Predicting fear of breast cancer recurrence in women five years after diagnosis using machine learning and healthcare reimbursement data ; the french nationwide vican survey. *Journal of Biomedical Informatics. Under review for publication (2024).*.

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, et D. I. Fotiadis (2014). Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J 13*, 8–17.

Magnani, C., A. Smith, D. Rey, A. Sarradon, M. Préau, M. Bendiane, A. Bouhnik, et J. Mancini (2022). Fear of cancer recurrence in young women five years after diagnosis with a good-prognosis cancer : The vican-5 national survey. *Journal of Cancer Survivorship*.

Miotto, R., L. Li, B. A. Kidd, et et al. (2016). Deep patient : An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports 6*, 26094.

Miotto, R., F. Wang, S. Wang, X. Jiang, et J. T. Dudley (2018). Deep learning for healthcare : Review, opportunities and challenges. *Briefings in Bioinformatics 19*, 1236–1246.

Otto, A. K., E. C. Soriano, S. D. Siegel, S. T. LoSavio, et J.-P. Laurenceau (2018). Assessing the relationship between fear of cancer recurrence and health care utilization in early-stage breast cancer survivors. *Journal of Cancer Survivorship 12*, 775–785.

Papanastasiou, G., G. Yang, D. I. Fotiadis, N. Dikaios, C. Wang, A. Huda, L. Sobolevsky, J. Raasch, E. Perez, G. Sidhu, et D. Palumbo (2023). Large-scale deep learning analysis to identify adult patients at risk for combined and common variable immunodeficiencies. *Communications in Medicine (London) 3*(1), 189.

Roorda, C., A. Berendsen, F. Groenhof, K. van der Meer, et G. de Bock (2013). Increased primary healthcare utilization among women with a history of breast cancer. *Supportive Care in Cancer 21*, 941–949.

Summary

La PRC est un problème majeur chez les survivantes du cancer du sein, affectant leur qualité de vie et nécessitant des stratégies efficaces. Les études précédentes ont surtout porté sur des interventions psychologiques, plutôt que sur la modélisation prédictive à partir des données de remboursement de soins. Ce travail propose une application innovante de l'apprentissage automatique (ML) pour prédire la PRC, offrant ainsi de nouvelles perspectives pour la survie à long terme et des interventions personnalisées.

Vers une interprétation robuste des anomalies contextuelles : Le rôle crucial de l'incertitude

Marwa Boulakbech

Aix Marseille Univ, CNRS, LIS, Marseille, France
marwa.boulakbech@univ-amu.fr,

Résumé. L'intégration de l'incertitude dans l'interprétabilté des anomalies représente un enjeu crucial pour l'intelligence artificielle centrée sur les données. Bien que l'estimation de l'incertitude soit reconnue pour améliorer la fiabilité des prédictions, son potentiel pour offrir des interprétations robustes et explicites reste largement inexploité. Ce papier de positionnement explore le rôle central de l'incertitude dans l'interprétabilté des anomalies contextuelles, en mettant l'accent sur les dépendances entre le contexte et l'incertitude. Nous proposons une architecture de haut niveau pour un système d'interprétation capable de détecter et d'interpréter les anomalies tout en intégrant l'incertitude de manière dynamique. Enfin, nous identifions des questions de recherche ouvertes pour guider le développement de ce système.

1 Introduction

L'intelligence artificielle centrée sur les données (IACD) a pour objectif de développer de nouvelles méthodes pour traiter de grandes quantités de données afin de calculer des prédictions ou prendre des décisions sur la base de ces données. Il s'agit de découvrir des relations et des connaissances dans les données, ce qui permet aux organisations de prendre des décisions fondées sur les données et d'obtenir de meilleurs résultats. Parmi les défis de l'IACD, nous retrouvons (1) la gestion de l'incertitude et (2) l'interprétabilité des résultats (Kumar et al. (2024)).

Dans le domaine de la détection d'anomalies, ces défis sont particulièrement critiques. L'incertitude joue un rôle essentiel en améliorant la robustesse des modèles et en fournissant des indications sur la fiabilité des prédictions (Ul Islam et al. (2018)). Au lieu de simplement classer une observation comme étant normale ou anormale, l'estimation de l'incertitude apporte une mesure de confiance associée à chaque prédiction. Par exemple, une prédiction avec une incertitude faible peut être considérée comme fiable, tandis qu'une incertitude élevée alerte les utilisateurs sur la nécessité de vérifier ou de valider les résultats avec des informations supplémentaires.

Bien que cette approche améliore la qualité des prédictions, elle présente des limites importantes en matière d'interprétabilité. Les travaux existants sur l'estimation de l'incertitude (Foldesi et Valdenegro-Toro, 2022; Vidmark, 2022; Wiessner et al., 2024) se concentrent principalement sur la détection d'anomalies, mais expliquent rarement pourquoi une observation est jugée anormale. Ce manque d'explication rend difficile la compréhension de l'origine de

l'anomalie, en particulier pour les utilisateurs non experts. Par exemple, dans le domaine des réseaux électriques, si un modèle signale une anomalie avec une incertitude élevée sans indiquer si elle est due à une fluctuation de tension ou à une panne d'équipement, il est difficile pour un opérateur de prendre des mesures correctives.

D'un autre côté, les approches d'interprétabilité d'anomalies proposées dans la littérature par (Jiang et al., 2023; Amarasinghe et al., 2018; Pang et Aggarwal, 2021) mettent l'accent sur l'interprétabilité, mais négligent souvent la gestion de l'incertitude (Pang et al. (2021)). Cette lacune limite l'applicabilité des modèles dans des environnements dynamiques où les données sont souvent bruitées ou incomplètes (Amarasinghe et al. (2018)).

Ainsi, dans ce papier, nous adressons la problématique suivante **comment exploiter l'incertitude pour améliorer l'interprétabilité des anomalies contextuelles ?**

Une anomalie contextuelle se définit comme une observation qui dévie des attentes normales dans un contexte spécifique, bien qu'elle puisse paraître normale dans un autre contexte (Chandola et al. (2009))

Nous pensons que l'intégration explicite de l'incertitude est essentielle pour interpréter ces anomalies de manière robuste. Cette intégration permet d'ajouter une couche de confiance et de transparence aux modèles de détection d'anomalies, tout en tenant compte des variations normales dans des contextes spécifiques. Cela devient particulièrement crucial dans des domaines complexes, tels que l'énergie, la finance ou la santé, où des erreurs d'interprétation peuvent entraîner des décisions coûteuses ou dangereuses. Pour les anomalies contextuelles, où les conditions externes influencent fortement les observations, combiner l'interprétabilité et la gestion de l'incertitude offre un cadre plus robuste et informatif pour guider les utilisateurs dans la prise de décision. **Il s'agit d'améliorer l'interprétabilité des prédictions en distinguant les sources fiables des sources incertaines**.

Ainsi, l'incertitude couplée à l'analyse des contributions, fournit des interprétations détaillées, permettant aux utilisateurs de comprendre pourquoi une anomalie est générée.

Supposons un système de gestion énergétique qui surveille la consommation électrique des conteneurs réfrigérés, l'objectif est de détecter des anomalies dans la consommation énergétique.

Variable	Valeur observée	Valeur normale
Température	40°C	25°C
Charge	2200 kg	1800 kg
Consommation	150 KW	120 KW

TAB. 1 – *Cas de détection d'anomalies contextuelles dans la consommation énergétique*

Le système observe une consommation de 150 kW, d'après le tableau 1, bien au-dessus de la consommation prédite (120 kW) alors il détecte une anomalie. L'utilisateur ne sait pas si cette anomalie est due à un dysfonctionnement ou à des conditions normales mais rares. Le système calcule une incertitude élevée pour les prédictions en raison de la rareté des données sur des températures aussi élevées avec un intervalle de confiance [115 kW,145 kW]. Ainsi, la consommation observée est légèrement au-dessus de l'intervalle de confiance et peut être interprétée par la température extérieure qui a conduit à 20% de surconsommation. Donc, l'utilisateur comprend que la température extérieure inhabituelle est le principal facteur de consommation élevée.

2 Système d'interprétation d'anomalies

Cette section décrit l'architecture d'un système d'interprétation des anomalies, conçu pour détecter et interpréter les anomalies tout en tenant compte du *contexte* et de *l'incertitude*. L'architecture est structurée en 4 composants comme illustré dans la Figure 1.

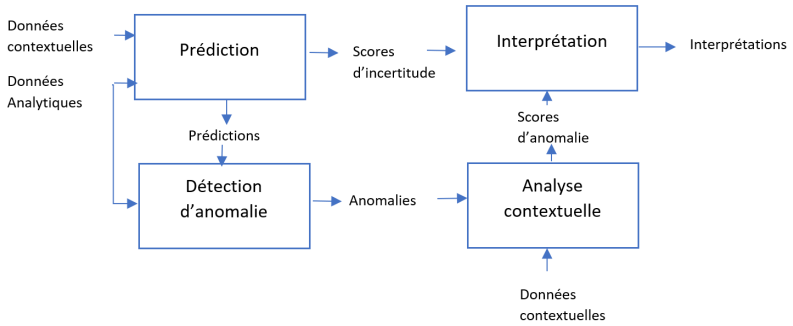


FIG. 1 – *Architecture fonctionnelle du système d'interprétation d'anomalies*

2.1 Module de prédiction

Le rôle principal de ce module est de prédire la variable cible $\widehat{y}_t$ et d'estimer l'incertitude associée σ_t. Les modèles bayésiens proposées par Heard et al. (2010); Forti et al. (2021); Yong et Brintrup (2022) peuvent être appliqués pour prédire la variable cible à partir des données analytiques et des données contextuelles.

2.2 Module de détection d'anomalies

Ce module identifie les anomalies en comparant les prédictions du modèle $\widehat{y}_t$ avec les observations réelles y_t , tout en prenant en compte l'incertitude.

— Si l'écart entre $\widehat{y}_t$ et y_t est important et l'incertitude σ_t est faible, l'anomalie est considérée comme *confirmée*.

— Si l'incertitude σ_t est élevée, l'anomalie est qualifiée de *potentielle* et nécessite une validation supplémentaire. Dans ce cas, un écart plus important est requis pour qu'une observation soit considérée comme réellement anormale.

Ce module améliore la robustesse du système en tenant compte des variations normales dans les contextes incertains, réduisant ainsi les faux positifs.

2.3 Module d'analyse contextuelle

Pour interpréter les anomalies détectées, ce module identifie les facteurs contextuels ayant influencé la détection d'anomalie. Il évalue la contribution de chaque variable contextuelle à la prédiction en utilisant des techniques d'interprétabilité tel que le mécanisme d'attention (Vaswani (2017)) ou la technique SHAP (Antwarg et al. (2019)). Il s'agit d'attribuer des poids

aux différentes variables contextuelles indiquant leur importance relative pour la prédiction et expliquer ainsi quelles parties des données ont influencé le plus la décision. Par exemple, lors d'une journée très chaude, la température peut recevoir un poids plus élevé que les autres variables (e.g., charge ou heure). Cette quantification permet d'isoler les variables ayant contribué significativement à l'anomalie et d'expliquer leur impact.

2.4 Module d'interprétation

Le dernier module permet une interprétation claire et exploitable des anomalies détectées, en combinant les résultats des modules précédents. Il utilise l'estimation de l'incertitude σ_t, et les scores d'anomalie s_t afin d'ajuster les contributions et générer les interprétations. Par exemple, le module peut calculer les contributions comme l'illustre le tableau 2. Ainsi, il génère les interprétations suivantes :

— La température ambiante est une cause probable et fiable de l'anomalie.
— La charge pourrait également avoir influencé l'anomalie, mais les données nécessitent une validation.
— l'heure est peu contributive et incertaine, probablement un bruit dans les données.

Variable	Score d'anomalie	Incertitude	Contribution
Température	0.7	0.1	0,63 Élevée
Charge	0.6	0.3	0,42 Modérée
Heure	0.5	0.5	0.25 Faible

TAB. 2 – *Contributions des variables contextuelles dans la détection d'anomalie*

3 Conclusion

L'intégration de l'incertitude dans l'interprétabilité des anomalies soulève plusieurs questions de recherche. Tout d'abord, les anomalies contextuelles, qui dépendent souvent de relations temporelles ou spatiales complexes, nécessitent des approches capables de modéliser l'incertitude de manière dynamique pour mieux refléter l'évolution des dépendances. Ensuite, l'incertitude peut également affecter la fiabilité des contributions des variables contextuelles à l'anomalie : une variable avec une forte contribution mais une incertitude élevée peut compromettre la confiance dans l'interprétation. Ainsi, un défi clé est de concevoir des méthodes pour relier efficacement les scores d'incertitude aux contributions des variables, afin de fournir des interprétations robustes et exploitables. Ces enjeux soulignent l'importance de développer des approches intégrant à la fois des mécanismes d'estimation d'incertitude et des outils d'interprétation contextuelle avancés.

Références

Amarasinghe, K., K. Kenney, et M. Manic (2018). Toward explainable deep neural network based anomaly detection. In *2018 11th international conference on human system interaction (HSI)*, pp. 311–317. IEEE.

Antwarg, L., R. M. Miller, B. Shapira, et L. Rokach (2019). Explaining anomalies detected by autoencoders using shap. *arXiv preprint arXiv :1903.02407*.

Chandola, V., A. Banerjee, et V. Kumar (2009). Anomaly detection : A survey. *ACM computing surveys (CSUR) 41*(3), 1–58.

Foldesi, L. et M. Valdenegro-Toro (2022). Comparison of uncertainty quantification with deep learning in time series regression. *arXiv preprint arXiv :2211.06233*.

Forti, N., L. M. Millefiori, P. Braca, et P. Willett (2021). Bayesian filtering for dynamic anomaly detection and tracking. *IEEE Transactions on Aerospace and Electronic Systems 58*(3).

Heard, N. A., D. J. Weston, K. Platanioti, et D. J. Hand (2010). Bayesian anomaly detection methods for social networks.

Jiang, R., Y. Xue, et D. Zou (2023). Interpretability-aware industrial anomaly detection using autoencoders. *IEEE Access 11*, 60490–60500.

Kumar, S., S. Datta, V. Singh, S. K. Singh, et R. Sharma (2024). Opportunities and challenges in data-centric ai. *IEEE Access*.

Pang, G. et C. Aggarwal (2021). Toward explainable deep anomaly detection. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Pang, G., C. Shen, L. Cao, et A. V. D. Hengel (2021). Deep learning for anomaly detection : A review. *ACM computing surveys (CSUR) 54*(2), 1–38.

Ul Islam, R., M. S. Hossain, et K. Andersson (2018). A novel anomaly detection algorithm for sensor data under uncertainty. *Soft Computing 22*(5), 1623–1639.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vidmark, A. (2022). Anomaly or not anomaly, that is the question of uncertainty : Investigating the relation between model uncertainty and anomalies using a recurrent autoencoder approach to market time series.

Wiessner, P., G. Bezirganyan, S. Sellami, R. Chbeir, et H.-J. Bungartz (2024). Uncertainty-aware time series anomaly detection. *Future Internet 16*(11), 403.

Yong, B. X. et A. Brintrup (2022). Bayesian autoencoders with uncertainty quantification : Towards trustworthy anomaly detection. *Expert Systems with Applications 209*, 118196.

Summary

Integrating uncertainty into anomaly interpretability represents a crucial challenge for data-centric artificial intelligence. Although uncertainty estimation is known to improve the reliability of predictions, its potential to offer robust and explicit interpretations remains largely untapped. This position paper explores the central role of uncertainty in the interpretability of contextual anomalies, focusing on the dependencies between context and uncertainty. We propose a high-level architecture for an interpretation system capable of detecting and interpreting anomalies while dynamically integrating uncertainty. Finally, we identify open research questions to guide the development of this system.

ESERAC: Explicabilité SEmantique des Recommandations basée sur l'Apprentissage profond pour la gestion des Crises

Firas Zouari*, Chirine Ghedira-Guegan*
Khouloud Boukadi** Nadia Kabachi***

*Univ Lyon, Université Jean-Moulin Lyon 3, LIRIS UMR5205, iaelyon School of Management, France
contact@firas-zouari.com
chirine.ghedira-guegan@univ-lyon3.fr
** University of Sfax, Sfax,Tunisia
khouloud.boukadi@fsegs.usf.tn
***Univ Lyon, Univ Lyon 1, UR ERIC and UR4129 P2S - Laboratory "Health, Systemic, Process", France.
nadia.kabachi@univ-lyon1.fr

Résumé. Au cours des dernières années, diverses crises complexes ont émergé, nécessitant des politiques adaptées pour anticiper, gérer et atténuer leurs impacts. Ces crises, qu'elles soient sanitaires, économiques ou sociales, mettent en évidence l'absence d'une solution universelle pour une gestion efficace. Les stratégies doivent être ajustées en fonction de multiples facteurs contextuels, telles que les données caractéristiques d'une région ou les priorités des parties prenantes. Dans cette optique, des modèles basés sur l'apprentissage profond ont été développés pour recommander des mesures adaptées à différents contextes décisionnels. Bien que performants, ces modèles présentent souvent un faible niveau d'explicabilité, limitant la confiance et la compréhension des utilisateurs. Pour y remédier, une approche sémantique a été proposée, exploitant des ontologies externes et les interactions des utilisateurs avec le système pour générer des explications personnalisées. L'évaluation de cette approche a démontré son efficacité grâce à des métriques de performance telles que la précision, le rappel et le F1-score, ainsi que des indicateurs d'ontologie, attestant de la pertinence des recommandations et de la clarté des explications fournies.

1 Introduction

Les récentes crises sanitaires, économiques, environnementales, etc transcendent les frontières et présentent des défis mondiaux majeurs. Les conséquences de ces crises, comme l'a illustré la pandémie de COVID-19, se manifestent simultanément sur les plans économique, social et géopolitique, nécessitant des réponses coordonnées et adaptées à la complexité de chaque situation. De ce fait, la gestion des crises exige non seulement une prédiction précise des risques, mais également des stratégies de réponse flexibles et contextuelles, capables de limiter les effets à court terme et d'anticiper les répercussions à long terme. Dans ce contexte, des systèmes de gestion de crises capables de surveiller les événements en temps réel et d'offrir des recommandations personnalisées deviennent essentiels. Ces systèmes doivent intégrer

des données globales et locales, tout en s'adaptant aux spécificités sociétales, économiques et culturelles des pays. Par exemple, face au COVID-19, les réponses variées de la Chine, de la Corée du Sud et du Japon ont démontré qu'aucune solution universelle ne peut répondre efficacement à une crise complexe. La Chine a misé sur des confinements stricts, la Corée du Sud sur le traçage et les tests massifs, et le Japon sur des mesures sociales modérées. Ces exemples soulignent l'importance d'une approche adaptable et basée sur les contextes spécifiques. Cependant, si ces systèmes de gestion offrent des recommandations pertinentes, leur utilité dépend fortement de leur capacité à fournir des explications claires et compréhensibles pour des utilisateurs aux profils variés. Les modèles de recommandation basés sur l'apprentissage profond, bien qu'efficaces pour analyser des données complexes, souffrent de leur nature opaque et de leur faible interprétabilité. Ce manque de transparence peut freiner leur adoption, particulièrement lorsqu'il s'agit de justifier des décisions critiques auprès d'un large éventail de parties prenantes. Pour répondre à ces défis, nous proposons un système novateur combinant intelligence artificielle et explicabilité pour la gestion des crises. Ce système repose sur deux contributions majeures : (i) un modèle de recommandation multi-sorties basé sur l'apprentissage profond conçu pour générer des recommandations adaptées aux spécificités locales des pays, tout en tenant compte des besoins divers des utilisateurs et des contextes décisionnels variés ; (ii) et une approche d'explicabilité, fondée sur les technologies du web sémantique, permettant de fournir des explications adaptées aux rôles et aux exigences des utilisateurs. Cette approche repose sur une ontologie d'explication dynamique exploitant des graphes de connaissances pour générer des explications contextualisées, telles que des contre-exemples, des comparaisons avec des pays voisins et/ou des analyses basées sur les données passées. Notre approche s'appuie sur les avancées en Intelligence Artificielle Explicable (XAI), garantissant une transparence post-hoc des recommandations, essentielle pour répondre aux attentes éthiques et pratiques. Le reste de cet article est structuré comme suit : la section 2 explore la revue de la littérature, la section 3 présente le modèle proposé, la section 4 détaille l'approche d'explication, et la section 5 expose les résultats expérimentaux. Enfin, la section 6 conclut et propose des pistes pour les travaux futurs.

2 Revue de la littérature

L'intelligence artificielle explicable (XAI) suscite un intérêt croissant parmi les chercheurs et les industriels, car elle cherche à concilier la performance des modèles d'apprentissage automatique avec leur interprétabilité. Bien que les réseaux neuronaux, en particulier ceux basés sur l'apprentissage profond, surpassent souvent des modèles plus transparents comme les arbres de décision, leur opacité demeure un frein à leur adoption dans des contextes nécessitant une prise de décision critique. Les approches XAI, et en particulier celles axées sur l'explicabilité post-hoc et les technologies sémantiques, offrent des pistes prometteuses pour surmonter ces défis. Les méthodes d'explication post-hoc visent à clarifier le fonctionnement des modèles sans en altérer les performances. Elles se concentrent notamment sur l'analyse des relations entre les entrées et les sorties du modèle pour générer des explications compréhensibles par les utilisateurs. En parallèle, les technologies sémantiques, telles que les ontologies et les graphes de connaissances, fournissent des cadres logiques structurés offrant de nombreux avantages tels que la cohérence des explications, leur adaptabilité à différents utilisateurs et leur capacité à raisonner sur des concepts complexes. Plusieurs travaux exploitent ces principes pour en-

richir l'explicabilité des modèles complexes. TREPAN Reloaded proposé par Confalonieri et Besold (2020), traduit les réseaux neuronaux en arbres de décision interprétables à l'aide d'ontologies. Panigutti et al. (2020) proposent des explications pour des données multi-étiquetées, reliant les caractéristiques des données à des concepts ontologiques. Sarker et al. (2017) utilisent des graphes de connaissances pour expliquer les relations entre les entrées et les sorties des modèles, offrant ainsi une perspective plus intuitive sur leur comportement. Dans une approche différente, Dragoni et al. (2020) ont développé un système exploitant les ontologies et l'intelligence artificielle explicable pour promouvoir les modes de vie sains. Leur modèle repose sur un raisonneur qui infère des informations et génère des explications contextuelles pratiques basées sur des règles définies dans l'ontologie de recommandation. Leur proposition illustre une utilisation innovante des technologies sémantiques pour fournir des explications directement intégrées pour la recommandation. Cependant, ces approches présentent plusieurs limites importantes. La plupart se focalisent sur un seul type d'explication ou visent un seul profil d'utilisateur, ce qui limite leur adaptabilité dans des contextes complexes impliquant des parties prenantes aux besoins variés. De plus, elles ne prennent pas toujours en charge la génération d'explications dynamiques qui pourraient évoluer en fonction du contexte ou des préférences voire exigences spécifiques des utilisateurs. Ces lacunes mettent en évidence le besoin de systèmes plus flexibles et robustes, capables de répondre aux attentes diversifiées des utilisateurs, tout en maintenant une performance élevée et une transparence adéquate.

3 Modèle de recommandation de mesures sanitaires

Nous proposons un modèle d'apprentissage profond conçu pour recommander des mesures sanitaires tout en fournissant des explications adaptées aux différents rôles impliqués dans la gestion des crises sanitaires. Ces rôles incluent des experts en santé publique, des économistes, et des analystes stratégiques. Ce modèle s'appuie sur les données collectées par l'Oxford Covid-19 Government Response Tracker (OxCGRT), un jeu de données de plus de 238 000 enregistrements et 56 variables recensant les politiques mises en place par les gouvernements pour lutter contre la pandémie de COVID-19. Les données couvrent une variété de mesures réparties en trois grandes catégories :

- **Politiques de fermeture et de confinement :** ciblent les analystes stratégiques et incluent, par exemple, la fermeture des écoles ou la limitation des rassemblements publics.
- **Politiques économiques :** intéressent particulièrement les économistes et incluent des mesures comme les subventions aux entreprises ou les allocations sociales.
- **Politiques du système de santé :** orientées vers les professionnels de santé, elles incluent des actions telles que le renforcement des capacités hospitalières ou la mise en œuvre de campagnes de vaccination.

Bien que conçues dans le contexte de la COVID-19, ces mesures peuvent également s'appliquer à d'autres crises sanitaires. Chaque mesure est identifiée par un code spécifique (par exemple, C1 pour la fermeture des écoles) et est caractérisée par une échelle de sévérité. Par exemple, pour la fermeture des écoles, l'échelle comprend trois niveaux : (1) recommandation de fermeture ou ajustement des horaires, (2) obligation de fermeture de certains niveaux scolaires (par exemple, les écoles primaires), et (3) obligation de fermeture de tous les niveaux scolaires. En adoptant ces données et les différentes mesures de sévérité associées, nous avons

conçu un modèle d'apprentissage profond multi-sorties capable de prédire le futur niveau de sévérité de chaque mesure. À notre connaissance, ce modèle original est le premier à recommander plusieurs mesures sanitaires, ainsi que leur degré de sévérité, en fonction des besoins de différents utilisateurs. Dans la section suivante, nous illustrons notre approche visant à expliquer les recommandations de notre modèle, en tenant compte des besoins spécifiques des différents utilisateurs.

4 Approche sémantique pour l'explication des recommandations sanitaires

Bien que les modèles d'apprentissage profond soient performants, leur opacité en tant que "boîtes noires" rend difficile leur interprétation. Pour pallier cette limitation, nous proposons une approche explicative basée sur des technologies sémantiques, visant à améliorer la transparence des résultats générés par le modèle de recommandation. Pour se faire, nous proposons la construction dynamique d'une ontologie d'explication, réalisée par le mappage avec des graphes sémantiques externes et des sources de données contextuelles. De plus, la qualité des données utilisées pour la construction de l'ontologie est rigoureusement évaluée grâce à notre approche Zouari et al. (2023), garantissant la pertinence et la fiabilité des explications générées. Une fois construite, cette ontologie permet d'extraire un sous-graphe spécifique au rôle et aux besoins de l'utilisateur, fournissant des explications claires, personnalisées et adaptées à chaque contexte décisionnel.

L'approche se déroule en deux étapes principales :

— **Construction de l'ontologie :** il s'agit de la création d'une ontologie d'explication en intégrant les informations contextuelles issues des graphes de connaissances et des ontologies de domaine.

— **Extraction du sous-graphe adapté :** consistant en la sélection des parties pertinentes de l'ontologie pour fournir des explications spécifiques aux besoins de chaque utilisateur.

4.1 Construction de l'ontologie d'explication

Les technologies sémantiques ont prouvé leur efficacité dans la structuration et la représentation des connaissances, ce qui les rend particulièrement adaptées à l'explicabilité des systèmes complexes. Dans ce cadre, notre approche repose sur une ontologie construite de manière dynamique, capable de s'adapter aux besoins et rôles variés des utilisateurs. L'ontologie intègre une variété de types d'explications, chacune répondant aux besoins spécifiques des utilisateurs. Par exemple, elle fournit des informations scientifiques destinées aux experts médicaux, détaillant les maladies, leurs symptômes, leurs traitements, et incluant des liens vers des ressources externes validées. Elle propose également des statistiques et des données contextuelles pour les analystes stratégiques, permettant des comparaisons des politiques sanitaires des pays voisins et des analyses des tendances globales ou régionales. Pour renforcer la pertinence des explications, notre approche s'appuie également sur des mesures historiques et des exemples concrets, permettant de contextualiser les recommandations dans des cas passés similaires. Une analyse de similarité identifie les pays ou contextes proches afin de fournir des contre-

exemples ou des références cohérentes, aidant ainsi à justifier les recommandations proposées. Enfin, l'explication par importance des caractéristiques utilise des méthodes comme SHAP (SHapley Additive exPlanations). Cette technique quantifie l'influence de chaque variable utilisée par le modèle, offrant ainsi une vision transparente des facteurs clés qui influencent les décisions. En combinant ces éléments, notre ontologie dynamique propose des explications claires, contextualisées et adaptées aux différents besoins, tout en exploitant efficacement les bases de connaissances internes et externes.

4.2 Extraction du sous-graphe d'explication

Après la construction de l'ontologie d'explication, l'étape suivante consiste à extraire un sous-graphe d'explication personnalisé, adapté aux besoins et préférences des utilisateurs. Les rôles variés des utilisateurs déterminent leurs attentes : un expert en santé peut privilégier des informations médicales détaillées, tandis qu'un analyste stratégique pourrait se concentrer sur des données comparatives, comme celles des pays voisins. A cette fin, nous utilisons une approche basée sur la factorisation de matrice, une méthode souvent employée dans les systèmes de recommandation tels que Netflix. En analysant les interactions des utilisateurs avec le système de recommandation, cette méthode nous permet de déduire implicitement leurs préférences. Ainsi, un sous-graphe d'explication adapté est extrait dynamiquement de l'ontologie d'explication, en tenant compte des préférences implicites des utilisateurs. Les interactions des utilisateurs sont également exploitées pour évaluer la satisfaction vis-à-vis des explications proposées. Par exemple, si un utilisateur choisit de cacher certaines explications ou explore activement des alternatives non suggérées, cela indique une insatisfaction probable. Ces retours sont utilisés pour affiner et adapter les explications futures, améliorant ainsi la pertinence des sous-graphes proposés. Nous avons implémenté cette approche via une matrice utilisateur-explication, dans laquelle les interactions sont enregistrées. En appliquant la factorisation de matrice, notamment l'algorithme SVD, nous avons projeté les utilisateurs et les explications dans un espace latent commun où leurs interactions sont modélisées comme des produits scalaires. Deux matrices principales, Q (utilisateurs) et P (explications), permettent de prédire l'intérêt d'un utilisateur pour une explication donnée en effectuant leur produit matriciel. Enfin, pour mieux refléter la diversité des contextes décisionnels, nous avons défini une matrice de préférences spécifique à chaque contexte. Cela nous permet de personnaliser davantage les sous-graphes d'explication en fonction des conditions. Cette capacité à s'adapter au contexte décisionnel garantit que notre système propose des explications dynamiques et pertinentes, alignées sur les besoins évolutifs des utilisateurs.

5 Résultats expérimentaux et discussion

Nous avons mené une batterie d'expériences pour évaluer les performances du modèle de recommandation basé sur l'apprentissage profond et de l'approche d'explication. Nous présentons dans cette section, les résultats obtenus en mettant particulièrement l'accent sur la robustesse de notre proposition et sur la transparence dudit modèle de recommandation, assurée par l'approche explicative.

5.1 Performance du modèle de recommandation

5.1.1 Jeu de données

Le modèle a été conçu en utilisant le jeu de données OxCGRT, qui compile les mesures stratégiques prises par les gouvernements face au COVID-19 entre le 1er janvier 2020 et le 31 janvier 2022. Les données couvrent plus de 180 pays et incluent des mesures comme la fermeture des écoles, les restrictions de déplacement, et les politiques économiques. Après un processus de prétraitement et de sélection des variables pertinentes, neuf caractéristiques d'entrée ont été retenues :

— Indices actuels (stringency, réponse gouvernementale, santé publique, et soutien économique),
— Taux de reproduction et taux de tests positifs,
— Taux de population, âge médian, espérance de vie, et indice de développement humain (IDH).

Pour la sortie, seules les mesures ayant une échelle de rigueur (closures, santé, et économie) ont été incluses. Les mesures de vaccination, sans échelle de rigueur, n'ont pas été considérées.

5.1.2 Évaluation du modèle

Le modèle de recommandation a été développé en Python à l'aide des bibliothèques Keras et TensorFlow, et son implémentation a été exécutée sous Google Colab. Formulé comme un problème de classification, il a été évalué à l'aide de plusieurs métriques. Les performances globales du modèle sont excellentes, avec un rappel de 95 %, une précision de 96 %, un F1-Score de 95 % et une précision globale également de 95 %. Ces résultats montrent la capacité du modèle à détecter efficacement les mesures tout en minimisant les erreurs de prédiction. Pour entraîner le modèle, les données issues de l'OxCGRT ont été divisées en trois ensembles : entraînement, test et validation. Le modèle a été optimisé avec l'algorithme ADAM sur 150 époques, garantissant une convergence rapide et stable. Une évaluation plus détaillée révèle que la précision par mesure spécifique varie entre 91 % et 99 %, confirmant la fiabilité du modèle pour une large gamme de prédictions. Enfin, l'analyse des performances d'entraînement et de validation a montré une convergence adéquate, sans signes de surapprentissage ou de sous-apprentissage. En conclusion, le modèle s'est révélé performant et robuste dans la prédiction des mesures sanitaires, démontrant sa pertinence pour une gestion efficace des crises.

5.2 Performance de l'approche d'explication

Des expérimentations ont été menées pour évaluer l'efficacité de l'approche d'explication proposée. Trois aspects principaux ont été étudiés : la qualité de l'ontologie d'explication, l'adéquation des sous-graphes extraits, et la performance du mécanisme de recommandation. La qualité de l'ontologie a été mesurée en utilisant des métriques qui prennent en compte sa structure et les connaissances qu'elle contient, couvrant des dimensions telles que la fonction, la sémantique et la pratique. Les résultats montrent que l'ontologie générée est riche, bien équilibrée et facile à réutiliser, avec des valeurs de profondeur et d'étendue indiquant une faible complexité tout en offrant une couverture sémantique élevée (richesse des classes de 75%) Hallinan et Striphas (2014), Wang et Zhang (2013). En ce qui concerne les sous-graphes d'explication, ils ont également été évalués à l'aide des mêmes métriques, avec des résultats similaires.

Les sous-graphes sont sémantiquement riches et bien équilibrés, en termes de structure et de contenu. Bien qu'ils soient limités par le nombre de relations d'héritage pour certaines catégories, leur richesse en attributs et leur capacité à fournir des explications claires sont probantes Zalewski et al. (2019), Lemire et Maclachlan (2007). Le mécanisme de recommandation a été testé à l'aide de divers algorithmes, y compris la Factorisation de Matrice (SVD, SVD++, NMF) Hallinan et Striphas (2014), Wang et Zhang (2013), KNN (baseline, moyennes, ZScore) Zalewski et al. (2019), et des méthodes comme Slope One Lemire et Maclachlan (2007), Co-Clustering Govaert et Nadif (2013), et des recommandations aléatoires. Les résultats montrent que les algorithmes SVD et SVD++ offrent les meilleures performances selon les métriques de Précision, Rappel, MAE et RMSE Wang et al. (2018), Chai et Draxler (2014). Les méthodes basées sur KNN et la Factorisation de Matrice Non-Négative ont également montré de bons résultats, bien que Co-Clustering ait été le moins performant. Ces résultats confirment l'efficacité de la Factorisation de Matrice pour recommander des explications adaptées dans différents contextes décisionnels.

6 Conclusion

Nous avons présenté une nouvelle approche explicable pour recommander des mesures sanitaires dans le cadre de la gestion des crises sanitaires. Cette approche s'appuie sur les mesures sanitaires proposées par le suivi OxCGRT COVID-19 pour concevoir un nouveau modèle basé sur l'apprentissage profond. Le modèle génère plusieurs sorties, chaque couche de sortie prédisant le niveau de rigueur des mesures sanitaires. Ces mesures concernent les politiques de confinement et de fermeture, les politiques économiques et celles relatives au système de santé.

Étant donné que les algorithmes d'apprentissage profond génèrent des modèles en « boîte noire », nous avons développé une approche basée sur la sémantique pour expliquer les recommandations produites. L'originalité de ce travail repose sur plusieurs points :

1. L'intégration d'un modèle d'apprentissage profond pour la recommandation de mesures sanitaires ;
2. La diversification des types d'explications fournies par le modèle, incluant des exemples et contre-exemples, l'analyse de l'importance des caractéristiques, ainsi que des comparaisons avec les pays voisins, ce qui permet une meilleure interprétabilité et transparence des résultats ;
3. L'adaptation dynamique des explications aux besoins des utilisateurs, via la conception d'une ontologie d'explications, assurant une personnalisation des recommandations et une amélioration continue des processus de traitement des données.

Dans les perspectives, nous envisageons plusieurs axes d'amélioration. Nous projetons d'intégrer la dimension temporelle dans la recommandation des mesures sanitaires en exploitant des architectures adaptées aux séries temporelles telles que LSTM, GRU ou CNN, tout en veillant à la qualité et à la représentativité des données utilisées pour assurer la robustesse des prédictions tout en veillant à préserver les données sensibles lors de la génération des recommandations et des explications.

Références

Chai, T. et R. Draxler (2014). Root mean square error (rmse) or mean absolute error (mae) ?–arguments against avoiding rmse in the literature. *Geoscientific Model Development 7*, 1247–1250.

Confalonieri, R. et T. Besold (2020). Trepan reloaded : A knowledge-driven approach to explaining black-box models. *ECAI*.

Dragoni, M., I. Donadello, et C. Eccher (2020). Explainable ai meets persuasiveness : Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine 105*, 101840.

Govaert, G. et M. Nadif (2013). Co-clustering. *Encyclopedia of Database Systems*.

Hallinan, B. et T. Striphas (2014). Recommended for you : The netflix prize and the production of algorithmic culture. *New Media & Society 18*.

Lemire, D. et A. Maclachlan (2007). Slope one predictors for online rating-based collaborative filtering. *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005 5*.

Panigutti, C., A. Perotti, et D. Pedreschi (2020). Doctor xai : An ontology-based approach to black-box sequential data classification explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 629–639.

Sarker, M. K., N. Xie, D. Doran, M. Raymer, et P. Hitzler (2017). Explaining trained neural networks with semantic web technologies : First steps. *Twelveth International Workshop on Neural-Symbolic Learning and Reasoning, NeSy*.

Wang, L., Y. Wang, F. Shen, M. Rastegar-Mojarad, et H. Liu (2018). Predicting practice setting using topic modeling. *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, 62–63.

Wang, Y.-X. et Y.-J. Zhang (2013). Nonnegative matrix factorization : A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering 25*, 1336–1353.

Zalewski, J., M. Ganzha, et M. Paprzycki (2019). Recommender system for board games. *2019 23rd International Conference on System Theory, Control and Computing*, 249–254.

Zouari, F., C. Ghedira-Guegan, K. Boukadi, et N. Kabachi (2023). A semantic and service-based approach for adaptive mutli-structured data curation in data lakehouses. *World Wide Web*.

Summary

Over the past decade, the world has faced multiple crises requiring effective management strategies to minimize impacts on populations and economies. These crises, whether related to health, the environment, or social challenges, highlight the need for adaptable decision-making frameworks tailored to the unique characteristics of each situation and country. To address this, we propose a deep learning-based recommendation model designed to suggest suitable actions and policies for different user roles. While these models demonstrate high accuracy, their lack of transparency often hinders trust and understanding among decision-makers. To overcome

this, we introduce a semantic explanation approach that utilizes ontologies to enhance the interpretability of the recommendations. By adapting explanations to user preferences and presenting them as intuitive sub-graphs, this approach ensures relevance and clarity. The proposed solution has been evaluated using performance and ontology-based metrics, demonstrating its effectiveness in generating actionable recommendations and providing tailored explanations across various decision-making contexts.

Un système de Récupération d'Information bilingue espagnol-nahuatl: un jouet d'application réaliste

Juan-José Guzmán-Landa*, Ligia Quintana-Torres*,** Juan-Manuel Torres-Moreno*

*Laboratoire Informatique d'Avignon
{juan-jose.guzman-landa,juan-manuel.torres}@univ-avignon.fr,
http://lia.univ-avignon.fr/equipe-cornet-2/
**Universidad Veracruzana
liquintana@uv.mx https://www.uv.mx

Résumé. Nous présentons un système de récupération d'informations à partir de documents PDF (thèses et mémoires master de l'Université Veracruzana UV, Mexique) via des requêtes des utilisateurs. Le système RIEN (Récupération d'Information Espagnol–Nahuatl) est centré sur les données et il cherche à résoudre une problématique actuelle de l'entrepôt de documents académiques de l'UV. RIEN utilise un approche IA/RI et des représentations statiques et dynamiques des embeddings (plongements de mots) pour établir des correspondances avec la requête et construire un rang des documents. Les premiers tests sur des ensembles réduits de données réelles sont très prometteurs.

1 Introduction

Actuellement l'Universidad Veracruzana (Mexique) compte plus de 160 programmes d'enseignement supérieur. Les étudiants, les enseignants-chercheurs et les chercheurs maintiennent une production académique continue que a besoin d'être triée. L'idée de ce projet est de classer leurs produits académiques selon plusieurs critères, qui seront exprimés sous forme de requêtes. La plupart des documents sont en espagnol, mais un sous-ensemble nouveau contient des documents en nahuatl, la langue autochtone mexicaine. Cette langue n'appartient pas à aucune famille des langues indo-européennes mais à la famille uto-aztèque (Torres-Moreno et al., 2024). Le problème est doublement complexe de par la nature interdisciplinaire des publications et par la nature bilingue de l'entrepôt.

Afin de tester le système dans un contexte réaliste, nous allons établir des correspondances entre un ensemble de requêtes et un ensemble de documents. L'objectif consiste donc à analyser l'ensemble de cette production -y compris les thèses, les mémoires master, les articles de recherche, les chapitres de livres, etc-. afin de obtenir des classements pertinents des documents recherchés via des requêtes des utilisateurs.

2 Corpus de données et approche retenue

Pour le problème de jouet, nous avons utilisé un ensemble pilot de $P = 56$ fichiers type .pdf venant de plusieurs catégories disponibles dans la base de données de l'UV [1]. Nous avons donc établi 160 catégories pour constituer notre corpus d'étude. Or, une problématique étant que ces catégories ne reflètent pas forcément l'interdisciplinarité présente dans les documents. Elles plutôt les cloisonnent. D'un autre côté, les requêtes peuvent aussi être interdisciplinaires, par exemple : **Requête 1 :** { **informatique, gastronomie, Veracruz** } ; **Requête 2 :** { **mathématiques, musique** } ; **Requête 3 : (nahuatl) :** { **tepostototl, informatikah** } (avion, informatique). Pour effectuer nos tests en espagnol, nous avons défini les $R_i, i = 1...10$ requêtes présentées dans la Table (1).

R_i	**Requête**	
0	Musica jazz	*musique jazz*
1	eeg cerebro computadora	*eeg cerveau ordinateur*
2	Enriquecidos fortificados	*enrichies fortifiés*
3	Violencia género	*violence genre*
4	Obesogénico neuroconductuales	*obésogenique neuroconducte*
5	Tripanosomiasis tropismo infección	*tripanosomiasis tropisme infection*
6	Conservación ecológica	*conservation écologique*
7	Política gobierno	*politique gouvernement*
8	Aguacate artificial microondas	*avocat artificiel microondes*
9	Antioxidante fisicoquímicas bioactivos	*antioxidant physicochimie bioactives*

TAB. 1 – *Requêtes utilisées lors de nos expériences*

2.1 Pré-traitement

Pour nos expériences, nous avons utilisé les P documents. Nous avons extrait les entêtes, c'est-à-dire, les résumés ou les introductions (dans le cas d'absence du résumé). Ensuite nous avons utilisé le pré-traitement suivant pour construir les entêtes :

1. Les $P = 56$ documents .pdf ont été transformés automatiquement en format .txt utf8.
2. Nous avons gardé les premiers 200 caractères de chaque document afin d'extraire soit le résumé soit l'introduction.

Le titre n'est pas utilisé pour le moment, car il demande d'heuristiques plus poussée d'analyse du PDF. Ainsi a été constitué notre jeu de données d'entêtes.

2.2 Modèles

Les $M_j; j = 1...9$ modèles à utiliser dans nos expériences sont les suivants :
— Sac-de-mots (BoW)
— TF-IDF

1. `https://cdigital.uv.mx/home` : Entrepôt de données institutionnel de l'Université Veracruzana.

— Word2Vec (CBOW et Skipgram) (Bojanowski et al., 2017; Mikolov et al., 2013)
— Word2Vec (pré-entrainé)
— FastText (pré-entrainé) (Bojanowski et al., 2016)
— Glove (pré-entrainé) (Pennington et al., 2014)
— BETO (CLS, Moyen ; BERT [2] en espagnol, pré-entrainés) (Cañete et al., 2020)

Nous avons utilisé des modèles pre-entrainées (Word2Vec, FastText, Glove et BETO) et nous avons entraîné le modèle Word2Vec (CBOW et skipgram) à partir des textes disponibles. BoW et TF-IDF utilisent une vecteur de termes par entête. BETO, Word2Vec, FastText et Glove utilisent leur propre tokeniseur, où chaque token produit un vecteur de plongements. Nous avons calculé un vecteur moyen normalisé pour les autres modèles. BETO utilise un token spécial (CLS), qui contient l'information de toute l'entête. Nous avons utilisé également ce vecteur dans nos expériences. Nous avons retenue comme méthode de tri des documents par rapport à la requête une approche de Recherche d'Information classique (Amini et Gaussier, 2013) avec des représentations classiques et denses à base de plongements de mots. Nous avons calculé le plongement E_{ij} des mots de la requête $R_i; i = 1, ..., 10$, selon le modèle $j = 1...9$.

2.3 Algorithme

L'algorithme d'analyse d'une requête i et un modèle j est le suivant :

1. Lire une requête R_i
2. Calculer son plongement E_{ij} selon un modèle M_j et la requête R_i.
3. Pour $k = 1...P$ documents faire
4. Calculer le plongement e_{kj} selon un modèle M_j et l'entête e_k.
5. $d = \min(sim = \cos(E_{ij}, e_{kj}))$
6. Fin

A la fin de l'algorithme la distance minimale d correspond au document k le plus proche à la requête i. La même méthode doit être repété pour tous les modèles et toutes les requêtes.

3 Résultats

Nous avons mesuré la précision p des systèmes sur l'ensemble des requêtes. Nous présentons nos résultats des 9 systèmes employées sur la table (2). La colonne VOTE montre le numéro de document les plus voté par les systèmes par rapport à chaque requête. Ceci a été vérifié manuellement et corresponde donc à une précision de 100%.

Les temps d'exécution T sont bien plus grandes dans les modèles classiques du au grand nombre de dimensions (4 459) par rapport à BETO (768), Word2Vec, FastText et Glove (300) et Word2Vec entraîné (100).

On peut constater l'excellente performance des modèles classiques et statiques par rapport aux modèles dynamiques. Nous constatons que les résultats obtenus avec la méthode RIEN sont assez encourageants.

2. https://github.com/dccuchile/beto

R_i	**BETO cls**	**BETO moyn**	**Word2Vec cbow**	**Word2Vec skipgram**	**FastText**	**Glove**	**Word2Vec**	**BoW**	**TF.IDF**	**VOTE**
0	32	32	26	26	20	7	26	26	26	26
1	32	40	12	12	12	12	3	12	12	12
2	32	32	11	30	27	27	27	27	27	27
3	32	32	35	35	8	8	8	11	11	8
4	32	4	4	4	1	32	34	4	4	4
5	32	54	41	41	41	23	41	41	41	41
6	32	32	30	2	49	49	49	51	51	49
7	12	32	23	37	56	56	56	43	46	56
8	32	32	43	43	43	43	43	43	43	43
9	32	27	19	19	2	2	2	19	19	19
p	0	10%	60%	60%	70%	60%	70%	70%	70%	**100%**
T	0.456	0.426	0.066	0.073	0.207	0.204	0.194	0.92	1.244	

TAB. 2 – *Précision p et temps d'exécution T (s) des systèmes selon les requêtes.*

4 Conclusion et perspectives

L'utilisation de TF-IDF et BoW se sont révélés intéressantes pour cette tâche. Ceci peut s'expliquer en partie parce que nous l'avons posé comme un problème de RI classique. Word-2Vec et FastText produisent également des bons résultats montrant ainsi que les plongements statiques ont encore un intérêt réelle dans des tâches de RI. Glove montre des perfomances moyennes mais supérieures à celle de BETO. Ce dernier modèle (CLS ou moyenne) montre des résultats décevants. Ils peuvent être expliqués de par la nature même de nos requêtes : elles sont interdisciplinaires voire exotiques et donc leurs représentations sont très orthogonales. Cela montre que BETO a du mal à trouver la proximité sémantique avec les documents. D'un autre côté, l'entraînement direct des plongements sur les données du corpus a pris quelques secondes sur une machine i7 sous GNU/Linux. En plus, nous pensons que l'ajout du titre et l'entraînement des plongements dynamiques BETO (Cañete et al., 2020)) pourrait mieux focaliser le rang de documents, à condition de re-faire un entraînement sur nos propres données, mais cela sera fait dans une prochaine étape. La version préliminaire du système RIEN est fonctionnelle sur des documents en espagnol et nous pensons pouvoir l'adapter pour des requêtes et documents nahuatl. En effet, pour la récupération de ce genre de documents, nous utiliserons d'abord les embeddings statiques venant du projet NAHU² (Torres-Moreno et al., 2024) : Word2vec et FastText, et puis ceux dynamiques de BERTL dès qu'ils seront disponibles. Enfin, l'ajout d'un module de résumé automatique (Torres-Moreno, 2014) pourra aider à générer un résumé exploitable où cette information est absente.

Références

Amini, M.-R. et E. Gaussier (2013). *Recherche d'information. Application, modèles et algorithmes*. Eyrolles, Paris.

D_i	**File name**	D_i	**File name**
1	Efectos de la Exposición	30	Efecto de la Frecuencia de Inundación
2	Bioaccesibilidad	32	Sistema de Sensado Remoto
3	Umbral de Respuesta	34	Efecto del Stress
4	Conducta Alimentaria	35	Estudio Teórico del Daño Oxidativo
7	La construcción del Imaginario	37	Análisis Experimental
8	Acoso y Hostigamiento	40	Diseño de una Etapa de Control Digital
11	Relación entre Impulsividad y Género	41	Evaluación de la Susceptibilidad
12	Sistema Clasificador para Ondas	43	Analisis Digital de Imagenes
19	Efecto de Diferentes Métodos de Cocción	46	Habitabilidad Urbana
20	Algunos Aspectos Transformacionales	49	Monitoreo de Poblaciones de Tortugas
23	Evaluación de Factores	51	Desarrollo Comunitario
26	La Educación del Jazz en México	54	Identificador de Posible Detector de Células
27	Fortificación de Tierras	56	La Apropiación del Espacio Público

TAB. 3 – *Documents .pdf et leur id récupérés par les modèles*

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2016). Enriching word vectors with subword information. *CoRR abs/1607.04606*.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2017). Enriching word vectors with subword information. *Transactions of the ACL 5*, 135–146.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, et J. Pérez (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *26th NIPS Volume 2*, NIPS'13, Red Hook, NY, USA, pp. 3111–3119. Curran Associates Inc.

Pennington, J., R. Socher, et C. D. Manning (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Torres-Moreno, J.-M. (2014). *Automatic Text Summarization*. Wiley, London.

Torres-Moreno, J.-M., M.-L. Avendaño-Garrido, M. Figueroa-Saavedra, G. Ranger, C. González-Gallardo, E. L. Pontes, P. V. Morales, L. Q. Torres, et J.-J. Guzmán-Landa (2024). piyalli : Un nouveau corpus pour le Nahuatl. *ArXiv arXiv :2412.15821v1*.

Summary

We present a retrieval information system from PDF documents (theses and master's theses from the Universidad Veracruzana UV, Mexico) via user queries. The RIEN system (from *Récupération d'Informations Espagnol–Nahuatl*) is data-centred and it seeks to solve a current problem in the UV's thesis repository. RIEN uses an AI/IR approach and static representations to establish correspondences with the query and build a document rank. Initial tests on reduced sets of real data are very promising.

Méthode automatique pour décider un réentrainement en apprentissage supervisé : application au phénotypage de populations de plantes par imagerie

Lakamy Thiam, Mathis Cordier, Félix Mercier,
Angélina El Ghaziri, Nizar Bouhlel, David Rousseau

ImHorPhen, Université d'Angers, Institut Agro Rennes-Angers,
Institut de Rechetche en Horticulture et Semences (IRHS), UMR INRAe, Angers, France
https://irhs.angers-nantes.hub.inrae.fr/recherche/imagerie-pour-l-horticulture-et-le-phenotypage
david.rousseau@unviv-angers.fr

Résumé. Nous proposons une méthode non supervisée opérant dans l'espace latent d'un réseau de neurones pour décider du réentraînement ou non d'un modèle quand de nouvelles données sont apportées. Nous montrons qu'en détectant les outliers par rapport aux données exisantes il est possible de limiter la quantité de données à annoter pour un réentraînement. Ceci est illustré dans le domaine du phénotypage des plantes par imagerie via une tâche de segmentation de symptômes foliaires sur une étude pluri-annuelle.

1 Introduction

À l'ère du traitement du phénotypage à haut-débit des végétaux, les mesures sont désormais réalisées automatiquement par des systèmes d'imagerie qui remplacent et augmentent les capacités humaines d'observations Costa et al. (2019). Le goulot d'étranglement du phénotypage des végétaux n'est donc plus au niveau des capteurs ni du développement de codes Minervini et al. (2015), il se déplace vers la gestion des données et notamment de la labellisation des données Douarre et al. (2019); Samiei et al. (2020); Sapoukhina et al. (2019). Actuellement les données en phénotypage des plantes sont principalement traitées au repos une fois que les acquisitions sont réalisées. Ce n'est pas compatible avec un suivi sur des objets en évolution constante comme les plantes. Cette situation appelle à une approche dite "Data centric", i.e. centrée sur les données plus que sur les codes, pour le domaine du phénotypage de population de plantes par imagerie.

Dans cette communication nous abordons la question de l'approche "Data centric" par le prisme de l'apprentissage actif. Le paradigme le plus courant en apprentissage actif suppose qu'on dispose d'un premier modèle et que lorsque de nouvelles données arrivent, un autre algorithme (a priori non supervisé) doit décider si ces nouvelles données peuvent être traitées par le modèle ou bien si elles doivent être labellisées pour donner lieu à un réentraînement. Un gain de temps est attendu par rapport à la situation où toutes les données seraient systématiquement ajoutées au jeu de données d'entraînement.

L'apprentissage actif a été seulement récemment considéré en imagerie des plantes Nagasubramanian et al. (2021); Rawat et al. (2022); Chandra et al. (2020). L'apport de l'apprentissage actif a été montré pour de la détection de maladie dans Nagasubramanian et al. (2021) et pour de la segmentation et détection d'organes dans Rawat et al. (2022); Chandra et al. (2020). Nous abordons ici une problématique de segmentation de symptômes de maladies. Les méthodes d'apprentissage actif fonctionnent principalement via des métriques appliquées à des instances seules ou par petits lots Ren et al. (2021). Dans le cas de l'imagerie des plantes, la récolte des données est souvent synchronisée (pour ce qui est des données au champ) avec les saisons. Nous proposons une méthode adaptée à cette spécificité.

2 Matériel et Méthode

2.1 Images, labels et réseau de neurones

Les données de cette étude sont des images couleur rouge vert bleu (RVB). La Fig. 1 montre des exemples rassemblés lors des années 2020, 2021, 2022 en partenariat avec l'Institut Technique de la Betterave (ITB) pour le suivi de la cercosporiose sur des feuilles de cultures de betterave sucrière par analyse d'images.

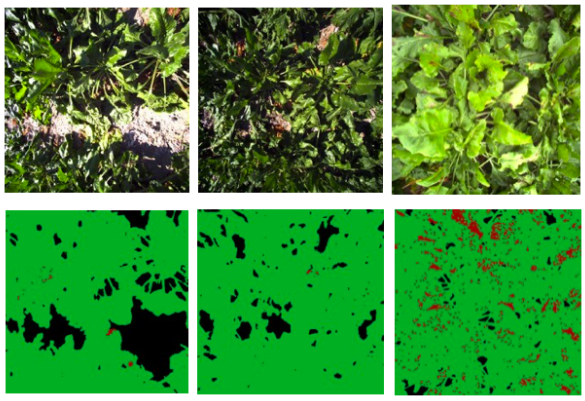


FIG. 1 – *Exemple d'images RGB et annotations associées. En vert le feuillage, en noir le sol en rouge la cercosporiose. Gauche image prise en 2020, milieu image de 2021, droite image de 2022.*

Les pixels des images sont labelisés en trois classes : sol, feuillage, cercosporiose comme montré sur la Fig. 1. Pour la segmentation des images, un classique réseau de neurones de type U-Net est utilisé Ronneberger et al. (2015). À titre d'illustration, l'annotation d'images comme celles de la Fig. 1 peut nécessiter plusieurs minutes. La Fig. 1 montre que les images présentent une grande diversité. D'une année à l'autre, les conditions météorologiques, la qualité du sol ou les techniques de culture subissent des modifications. L'année d'acquisition est un facteur important de variabilité de la base de données. Nous avons récolté 58 images en 2020, 122

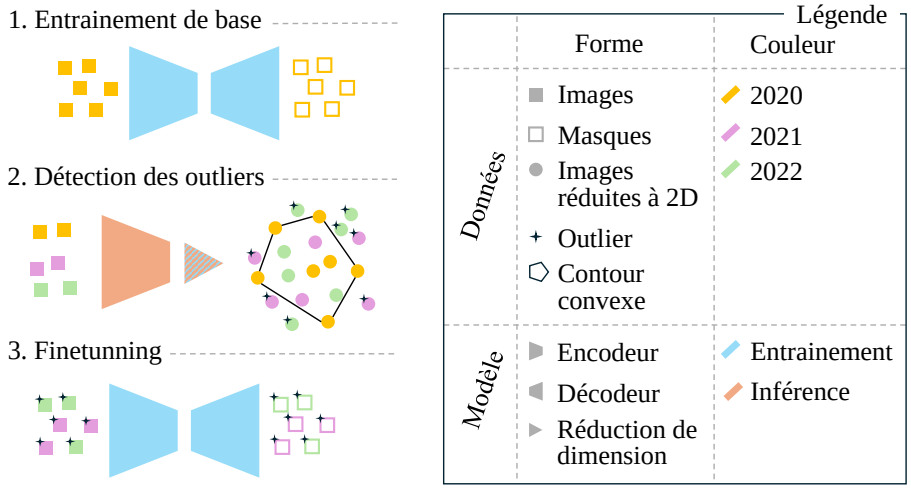


FIG. 2 – *Algorithme d'apprentissage actif proposé.*

en 2021 et 200 en 2022. Notre objectif est d'arriver à optimiser la quantité d'image à annoter chaque année pour améliorer le modèle. Pour ce faire nous proposons l'algorithme non supervisé de la section suivante. Nous mesurons la performance du modèle de segmentation au moyen de la métrique classique du coefficient de Dice-Sørensen qui mesure le rapport de l'intersection de l'union entre le masque de vérité terrain et la segmentation divisé par l'union entre ce masque et la vérité terrain.

2.2 Algorithme d'apprentissage actif proposé

Nous introduisons l'algorithme non supervisé de la Fig. 2. Dans cet algorithme, la première étape consiste en un entraînement de base sur les données d'entraînement de la première année. Puis, dans l'étape 2 de la Fig.2, lorsque de nouvelles données sont acquises (sous la forme d'une nouvelle année ici), elles sont passées dans l'encodeur du U-Net qui avait été entraîné sur les données précédemment accumulées. Une réduction en 2 dimension est ensuite réalisée sur cet espace latent de l'encodeur. Nous procédons à une détection de point aberrants (Outlier). Ces Outliers sont annotés puis, ce qui correspond à l'étape 3 de la Fig. 2, un réentraînement est réalisé avec les Outliers sous forme de fine tuning. Dans cette communication, nous montrons les résultats avec une méthode de réduction de dimension classiques de type ISOMAP Tenenbaum et al. (2000) (mais les résultats sont similaires pour d'autres méthodes classiques). Pour ce qui est de la détection d'Outlier, nous avons calculé l'enveloppe convexe des données. Une image de la nouvelle année est considérée comme un Outlier si elle n'est pas positionnée dans l'enveloppe convexe des données ayant servies à entraîner la précédente version du U-Net.

3 Résultats

Le Tableau 3 montre les performances obtenues en comparant l'apprentissage de base si on n'annote que les donnés de 2020, l'approche qui consiste à annoter toutes les données d'entraînement sur les 3 années et notre approche où seuls les Outliers détectés sont réannotés. On observe tout d'abord que les performances du modèle entraîné sur les données de 2020 généralise mal aux données de 2021 et 2022. Ceci justifie quantitativement le besoin d'ajouter davantage de données. On observe une amélioration significative des performances (tant sur le Dice Global que sur le Dice lié à la seule classe de la maladie) avec notre méthode qui se montre très proche du cas où nous aurions dû tout réannoter. Le nombre d'Outlier est de 30% en 2021 et 25% en 2022. On note également un gain significatif sur le temps d'entraînement et l'énergie consommée avec notre approche.

		Données d'entraînement		Données de finetunning
Métrique	Données Test	2020	2020-2021-2022	Outlier 2021-2022
Dice	2020	0.83 ± 0.01	0.79 ± 0.02	0.75 ± 0.01
	2021	0.86 ± 0.01	0.92 ± 0.01	0.93 ± 0.00
	2022	0.72 ± 0.03	0.97 ± 0.01	0.98 ± 0.00
Dice Maladie	2020	0.37 ± 0.02	0.35 ± 0.03	0.33 ± 0.05
	2021	0.30 ± 0.01	0.44 ± 0.04	0.40 ± 0.05
	2022	0.03 ± 0.00	0.28 ± 0.10	0.25 ± 0.09
Consommation d'énergie (dWh)		2.0 ± 1.0	13.7 ± 2.3	3.8 ± 1.0
Duré d'entraînement		$16\ min$	$1\ h\ 37\ min$	$27\ min$

TAB. 1 – *Trois modèles testés : 1- U-Net avec le jeu d'entraînement de 2020. 2- U-Net avec les jeux d'entraînement des trois années. 3- Détection des Outliers de 2021-2022 obtenue avec le jeu d'entraînement de 2020 appliqué à l'apprentissage actif de notre algorithme de la Fig. 2. Les métriques sont les coefficients Dice sur les jeux de données tests, par année et pour la classe maladie. Les calculs sont estimés sur une station de travail Dell Precision 7960.*

4 Conclusions

L'algorithme d'apprentissage actif non supervisé introduit montre de bonnes performances avec un gain de temps d'annotation de 75% et un gain de temps de réentraînement de 27% pour le jeu de données pris ici pour illustration. Pour étendre la portée de ces travaux préliminaires, il serait important de l'appliquer par la suite à d'autres jeux de données, à d'autres modalités image que l'année (comme la qualité des images), d'autres méthodes de détection d'Outliers, d'autres méthodes de réduction de dimension et de le comparer avec des algorithmes classiques d'apprentissage actif Wu et al. (2022). Autant de pistes que nous explorons actuellement.

5 Remerciements

Les auteurs remercient le fond CASDAR et l'ITB pour le financement de ces travaux dans la cadre du projet CERCOCAP.

Références

Chandra, A. L., S. V. Desai, V. N. Balasubramanian, S. Ninomiya, et W. Guo (2020). Active learning with point supervision for cost-effective panicle detection in cereal crops. *Plant Methods 16*, 1–16.

Costa, C., U. Schurr, F. Loreto, P. Menesatti, et S. Carpentier (2019). Plant phenotyping research trends, a science mapping approach. *Frontiers in plant science 9*, 1933.

Douarre, C., C. F. Crispim-Junior, A. Gelibert, L. Tougne, et D. Rousseau (2019). Novel data augmentation strategies to boost supervised segmentation of plant disease. *Computers and electronics in agriculture 165*, 104967.

Minervini, M., H. Scharr, et S. A. Tsaftaris (2015). Image analysis : the new bottleneck in plant phenotyping [applications corner]. *IEEE signal processing magazine 32*(4), 126–131.

Nagasubramanian, K., T. Jubery, F. Fotouhi Ardakani, S. V. Mirnezami, A. K. Singh, A. Singh, S. Sarkar, et B. Ganapathysubramanian (2021). How useful is active learning for image-based plant phenotyping ? *The Plant Phenome Journal 4*(1), e20020.

Rawat, S., A. L. Chandra, S. V. Desai, V. N. Balasubramanian, S. Ninomiya, et W. Guo (2022). How useful is image-based active learning for plant organ segmentation ? *Plant Phenomics*.

Ren, P., Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, et X. Wang (2021). A survey of deep active learning. *ACM computing surveys (CSUR) 54*(9), 1–40.

Ronneberger, O., P. Fischer, et T. Brox (2015). U-net : Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015 : 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer.

Samiei, S., P. Rasti, P. Richard, G. Galopin, et D. Rousseau (2020). Toward joint acquisition-annotation of images with egocentric devices for a lower-cost machine learning application to apple detection. *Sensors 20*(15), 4173.

Sapoukhina, N., S. Samiei, P. Rasti, et D. Rousseau (2019). Data augmentation from rgb to chlorophyll fluorescence imaging application to leaf segmentation of arabidopsis thaliana from top view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0.

Tenenbaum, J. B., V. d. Silva, et J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *science 290*(5500), 2319–2323.

Wu, M., C. Li, et Z. Yao (2022). Deep active learning for computer vision tasks : methodologies, applications, and challenges. *Applied Sciences 12*(16), 8103.

Un pipeline automatisé pour le développement des modèles d'IA prédictifs des niveaux d'eau souterraine au service d'un jumeau numérique environnemental

Vy-Thuy-Lynh HOANG*, Abdallah EL MAAZOUZI**, Yann DANTAL***

*,**,***BRGM - 3 av Claude Guillemin, 45100 Orléans
*vtl.hoang@brgm.fr, **a.elmaazouzi@brgm.fr, **y.dantal@brgm.fr

Résumé. Cet article présente une méthodologie exploitant l'automatisation des pipelines pour les jumeaux numériques environnementaux. En intégrant des sources de données diversifiées adaptées aux modèles d'IA, il offre des capacités d'optimisation des prédictions des niveaux d'eau souterraine tout en améliorant la scalabilité. Cette synergie entre données, IA et workflows automatisés renforce l'efficacité et soutient les objectifs des jumeaux numériques environnementaux.

1 Introduction

Un jumeau numérique (JN) est une réplique virtuelle d'une entité réelle, vise à mieux comprendre, surveiller et prédire l'évolution cette réelle, tout en permettant de tester des modifications sans impact réel (de Koning et al., 2023). Bien que prometteurs pour l'optimisation des ressources naturelles, les JNs en environnement sont encore émergents, avec peu de méthodologies établies (Tzachor et al., 2022). Le programme JUNON [1], financé par la région Centre-Val de Loire, vise à créer des JNs des ressources naturelles de la Beauce, en intégrant les dimensions de l'AIR, du SOL et de l'EAU. Son premier prototype établit une base conceptuelle pour l'intégration, l'interopérabilité et l'extension du système à divers cas d'utilisation. À partir de la source de données, le JN génère automatiquement (ou adapte) différents modèles. Ces modèles présentent des capacités d'*interpolation* et d'*extrapolation* (des points de mesure aux zones) et de *prédiction* (des valeurs passées aux valeurs futures), et sont capables de construire des cartes qui peuvent être comparées à des références, des seuils externes, et combinées à des bases de connaissances [(Iglesias et al., 2024), (Iglesias Vázquez, 2024)].

Le volet *EAU* du JUNON a pour objectif de simplifier la représentation du fonctionnement de la nappe de Beauce. Cela vise à automatiser l'actualisation des données et des informations dans une optique de prévision, afin de démontrer les avantages d'un JN environnemental. Pour atteindre cet objectif, il est essentiel de développer des modèles d'IA qui viennent compléter les modèles physiques existants, tels que GARDÉNIA [2] - un outil de modélisation hydrologique des bassins versants. Cette approche repose sur l'automatisation des pipelines de science des données via un workflow structuré, permettant de centraliser et de traiter efficacement des données d'entrée riches et complexes (Section 2). Les nombreuses variables disponibles offrent

1. https ://www.brgm.fr/en/programme/junon-digital-twins-working-natural-resources
2. https ://www.brgm.fr/fr/logiciel/gardenia-logiciel-modelisation-hydrologique-globale-bassin-versant

diverses options pour développer des modèles d'IA adaptés à différentes configurations de données (Section 3). En utilisant des modèles métiers comme référence ou pour le *labeling*, l'efficacité des prédictions des niveaux d'eau est significativement renforcée (Section 4).

2 Sources et types de données nécessaires

Boo et al. (2024) offre une analyse complète des modèles d'apprentissage automatique (ML) utilisés pour la prévision du niveau des eaux souterraines. Les auteurs discutent également de concepts clés comme la taille des ensembles de données, la sélection des variables d'entrée, les indicateurs de performance,... Selon plusieurs études [(Sun et al., 2024), (Rajaee et al., 2019)], les variables clés pour prévoir les niveaux d'eau souterraine (GWL-*groundwater level*) incluent les précipitations (P), les niveaux antérieurs (GWL antécédent) et la température (T), en raison de leur impact direct. L'évapotranspiration potentielle (ETP) complète ces données pour mieux saisir les interactions complexes entre l'atmosphère et les aquifères.

Les données historiques des niveaux d'eau souterraine proviennent de la plateforme ADES [3], qui fournit des mesures piézométriques à l'échelle nationale via les API REST de Hub'Eau [4]. Ces données, disponibles en formats CSV, JSON et GeoJSON, sont collectées à travers le Code BSS [5] et permettent de suivre l'évolution des niveaux d'eau au fil du temps. Météo-France fournit des données météorologiques via API [6], téléchargeables en formats CSV/JSON, et propose le modèle SAFRAN [7], qui intègre observations et modèles numériques pour créer des données atmosphériques cohérentes à haute résolution spatiale et temporelle - idéales pour les études hydrologiques. L'INRAE offre une API [8] dédiée aux données SAFRAN. ERA5, du service Copernicus, fournit des données climatiques globales (précipitations, ETP) sur une grille de 25 km, accessibles via l'API Climate Data Store [9].

Les données de "*pluie efficace*" - calculées par le BRGM, sont obtenues à partir de plusieurs sources. Tout d'abord, des variables météorologiques (précipitations, neige, température, et ETP) issues des modèles climatiques SAFRAN sont utilisées. Ensuite, la capacité de stockage des sols (CSS) et les données sur le couvert végétal (CLC 2018) sont intégrées pour mieux représenter les processus d'infiltration et d'évapotranspiration. À l'échelle de chaque maille SAFRAN (8x8 km²), le bilan hydrique journalier du sol est modélisé en utilisant plusieurs approches hydrologiques (*Thornthwaite, Dingman, Edijatno&Michel*), prenant en compte les précipitations, la fonte des neiges, l'ETR (évapotranspiration réelle) et la CSS. Les résultats des trois modèles sont ensuite moyennés, puis des moyennes interannuelles sont calculées pour produire une carte nationale de la pluie efficace [10].

3 Automatisation des pipelines

Traitements nécessaires. L'élaboration des modèles de prévision des niveaux d'eau souterraine reposent sur plusieurs étapes essentielles : (1) *Récupération et Prétraitement des don-*

3. https ://ades.eaufrance.fr/
4. https ://hubeau.eaufrance.fr/page/api-piezometrie
5. https ://infoterre.brgm.fr/page/nouveau-code-bss
6. https ://portail-api.meteofrance.fr/web/fr/
7. https ://meteo.data.gouv.fr/datasets/donnees-changement-climatique-sim-quotidienne/
8. https ://geosas.fr/web/ ?page_id=6345
9. https ://cds.climate.copernicus.eu/datasets/reanalysis-era5-land ?tab=download
10. Journée "Eau & Connaissance", Lyon 11 déccembre 2023

nées : Les données hydrologiques et météorologiques sont extraites via des APIs. Ce processus nécessite le nettoyage des valeurs anormales, l'interpolation des données manquantes et la gestion des incohérences (les différences entre les codes BSS anciens et nouveaux, ou la présence de mesures multiples par jour). Les moyennes quotidiennes de température et les totaux journaliers des précipitations/neige sont ensuite calculés pour normaliser les données. (2) *Spatialisation et Temporalisation* : Les données météorologiques sont associées aux points de mesure hydrologiques en fonction des périodes de mesure et des coordonnées géographiques (par une moyenne pondérée des 4 mailles voisines ou par l'utilisation de la maille unique la plus proche). (3) *Calcul de la Pluie efficace* : Ce calcul (réalisé en MATLAB) doit pouvoir être appliqué à toute période demandée par les modèles.

Modèles exploités. Le modèle Gardenia, basé sur des équations hydrologiques classiques, sert de référence pour valider les prédictions des modèles d'IA. Il peut également être utilisé comme label pour entraîner les modèles d'IA à certains points où il est disponible, bien qu'il ne couvre pas tous les points piézométriques. Les modèles d'IA exploitent des techniques d'apprentissage pour prédire les niveaux d'eau souterraine en combinant données historiques et variables météorologiques. Actuellement en phase de développement, ces modèles présentent encore des incertitudes sur le choix des données d'entrée et des algorithmes.

Un workflow personnalisable permettant de traiter efficacement des données d'entrée complexes et riches, destinées à alimenter des modèles, afin de sélectionner l'option optimale :

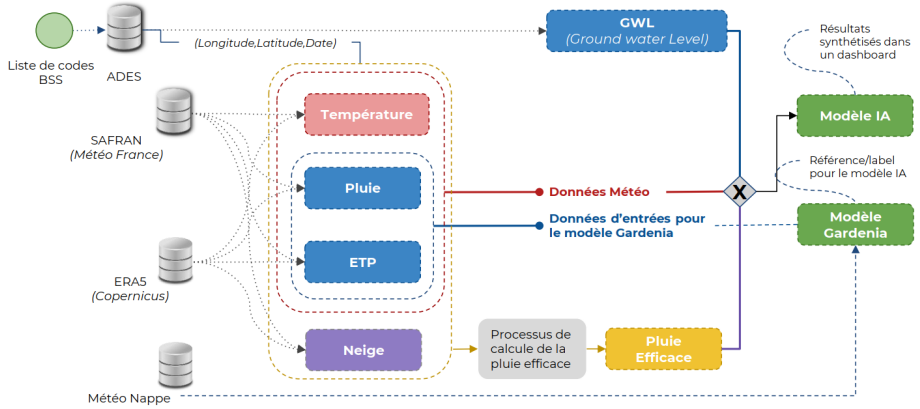


FIG. 1 – *Pipeline Global.*

Technologies d'automatisation des workflows. *Apache Airflow* est flexible pour les développeurs grâce à Python et une gestion avancée des dépendances, mais complexe pour les non-techniciens et moins adapté aux workflows évolutifs. *Galaxy* offre une interface intuitive pour les utilisateurs non techniques, idéale pour les tâches reproductibles, mais limité en personnalisation. La *dockerisation* des étapes garantit portabilité et compatibilité des workflows.

4 Conclusion et Perspectives

Ce travail initie l'automatisation des pipelines scientifiques pour les JNs environnementaux, en se concentrant sur le volet EAU/JUNON avec des extensions prévues pour AIR et

SOL. Une première version opérationnelle du workflow, intégrant les données hydrologiques (ADES) et météorologiques (SAFRAN), utilisant des modèles d'IA (TRANSFORMERS), et permettant de visualiser les résultats via un tableau de bord interactif, a été déployée avec succès. Ces travaux ouvrent des perspectives prometteuses, notamment en formalisant davantage l'impact de l'automatisation sur les JNs et en développant une intégration transversale des thématiques EAU, AIR et SOL. Par ailleurs, l'étude des choix algorithmiques et des incertitudes des données sera essentielle pour évaluer les performances des modèles, tout comme l'harmonisation des données et l'intégration directe de modules pour le calcul de la pluie efficace et le modèle Gardenia, afin d'améliorer la robustesse et l'efficacité du workflow. L'innovation repose sur la conception d'un workflow extensible et reproductible, réduisant les délais de traitement, élargissant les données disponibles pour les modèles d'IA, et posant les bases d'une validation à grande échelle, d'une évaluation comparative des modèles et d'une intégration élargie pour renforcer la contribution scientifique aux JNs environnementaux.

Références

Boo, K. B. W., A. El-Shafie, F. Othman, M. M. H. Khan, A. H. Birima, et A. N. Ahmed (2024). Groundwater level forecasting with machine learning models : A review. *Water Research 252*, 121249.

de Koning, K., J. Broekhuijsen, I. Kühn, O. Ovaskainen, F. Taubert, D. Endresen, D. Schigel, et V. Grimm (2023). Digital twins : dynamic model-data fusion for ecology. *Trends in Ecology and Evolution 38*(10), 916–926.

Iglesias, F., V.-T.-L. Hoang, L. Gourcy, A. Grossel, A. D. Filippis, J.-S. Moquet, et F. Ros (2024). Digital twins of natural resources in the loire valley. *To be published in the Proceedings of the 27th International Conference on Discovery Science*. Presented at the 27th International Conference on Discovery Science 2024.

Iglesias Vázquez, F. (2024). Proof of concept : Junon digital twin (v1.3).

Rajaee, T., H. Ebrahimi, et V. Nourani (2019). A review of the artificial intelligence methods in groundwater level modeling. *Journal of Hydrology 572*, 336–351.

Sun, W., L.-C. Chang, et F.-J. Chang (2024). Deep dive into predictive excellence : Transformer's impact on groundwater level prediction. *Journal of Hydrology 636*, 131250.

Tzachor, A., S. Sabri, C. E. Richards, A. Rajabifard, et M. Acuto (2022). Potential and limitations of digital twins to achieve the sustainable development goals. *Nature Sustainability 5*(10), 822–829.

Summary

This article outlines a methodology leveraging pipeline automation for environmental digital twins. By integrating diverse data sources tailored to AI models, it enhances scalability and optimizes groundwater prediction, showcasing the synergy between data, AI, and automation in advancing digital twin objectives.

MixMAS: Recherche automatisée d'architectures de fusion des données et d'apprentissage multimodal

Abdelmadjid Chergui*, Grigor Bezirganyan**

*Higher School of Computer Science, 8 Mai 1945,
SBA, Algeria
a.chergui@esi-sba.dz
**Aix-Marseille University, LIS, CNRS,
Marseille, France
grigor.bezirganyan@univ-amu.fr

Résumé. Choisir une architecture d'apprentissage profond adaptée à la fusion de données multimodales est une tâche complexe, car cela nécessite une intégration et un traitement efficaces de types de données divers, chacun ayant des structures et des caractéristiques distinctes. Dans cet article, nous présentons MixMAS, un nouveau *framework* pour la recherche d'architectures de fusion basée sur l'échantillonnage, spécialement conçu pour l'apprentissage multimodal. Notre approche sélectionne automatiquement l'architecture optimale basée sur les MLP pour une tâche donnée d'apprentissage machine multimodal (MML). Plus précisément, MixMAS utilise une stratégie de micro-benchmarking basée sur l'échantillonnage pour explorer diverses combinaisons d'encodeurs spécifiques à chaque modalité, de fonctions de fusion et de réseaux de fusion, identifiant systématiquement l'architecture qui répond le mieux aux métriques de performance de la tâche. Les expérimentations menées sur les jeux de données MM-IMDB et AV-MNIST démontrent que le pipeline proposé génère des architectures qui surpassent les modèles de référence M2-Mixer, avec des améliorations moyennes de +2,92% en F1-score et de +2,59% en précision, respectivement.

1 Introduction

La complexité croissante et la diversité des données dans divers domaines nécessitent l'utilisation de l'apprentissage multimodal, qui peut tirer parti et intégrer des informations provenant de différentes modalités, y compris le texte, l'image, l'audio, la vidéo, les séries temporelles, etc. (Baltrušaitis et al., 2018). L'application de l'apprentissage multimodal s'étend à un large éventail de domaines, notamment, mais sans s'y limiter, la génération de texte à partir d'images, la synthèse de texte à partir de vidéos, la robotique et la conduite autonome (Liang et al., 2022). L'essence de l'apprentissage multimodal réside dans sa capacité à fournir une compréhension plus globale des données issues de différents canaux, en exploitant la nature complémentaire des différents types de données. Cependant, la fusion des données multimodales présente des défis significatifs sur le plan computationnel et théorique (Liang et al.,

2022), en raison de l'hétérogénéité inhérente des sources de données, rendant plus difficile l'apprentissage des relations et représentations inter-modales. Chaque modalité présente souvent des propriétés, des structures et une pertinence diverses pour la tâche concernée. Sur le plan computationnel, le traitement et la fusion de ces divers types de données à grande échelle nécessitent des ressources matérielles considérables. Cela souligne l'importance croissante de concevoir des architectures spécialisées capables de traiter efficacement les données multimodales.

Face à ces défis, les architectures basées sur des perceptrons multicouches (MLP) ont émergé comme une solution prometteuse (Tolstikhin et al., 2021; Mai et al., 2023; Fu et al., 2023). Ces architectures offrent une alternative intéressante aux modèles de type *Transformer* en établissant un compromis avantageux entre performance et complexité de calcul (Tolstikhin et al., 2021). Leur efficacité, leur conception simplifiée en terme d'architecture (Liu et al., 2022), et leur robustesse dans la gestion de divers types de données et de tâches (Liu et al., 2022) en font des solutions adaptées à de nombreux contextes. L'intégration de ces composants dans une chaîne automatisée permet de tirer parti de leurs avantages, de sélectionner et de concevoir des architectures répondant aux exigences et contraintes spécifiques pour une tâche donnée, aboutissant à des modèles plus efficaces et spécialisés.

Dans cet article, nous introduisons MixMAS, un *framework* automatisé permettant de sélectionner les architectures MLP les plus performantes pour l'apprentissage multimodal (MML).

Nos contributions principales sont les suivantes :

1. Nous proposons un *pipeline* qui identifie l'architecture MLP optimale pour les tâches d'apprentissage et les données multimodales. Ce processus évalue diverses architectures MLP pour chaque modalité, sélectionne la fonction de fusion la plus efficace et détermine le réseau de fusion adapté. Bien que centré sur les MLP pour leur simplicité, la méthode proposée reste flexible et peut intégrer d'autres modèles, comme les *transformers* ou les réseaux de type CNN.
2. Nous intégrons une approche d'échantillonnage, permettant d'évaluer les modules sur un sous-échantillon des données, réduisant ainsi les coûts de calcul par rapport à une analyse complète.
3. Nous validons expérimentalement l'efficacité de notre approche, montrant une amélioration des métriques de performance telles que la précision par rapport aux réseaux multimodaux standards basés sur les MLP. Le code source est disponible ici : https://anonymous.4open.science/r/MixMAS.

Cet article est structuré comme suit. La section 2 passe en revue la littérature. La section 3 décrit le *framework* proposé. La section 4 présente les résultats expérimentaux. Enfin, la section 5 conclut cet article et propose des perspectives pour nos recherches futures.

2 Etat de l'art

Cette section passe en revue les travaux sur les modèles basés sur les MLP et les méthodes de recherche d'architectures multimodales.

MLP-Mixers. Les auteurs de (Tolstikhin et al., 2021) ont introduit une approche innovante en apprentissage profond, en obtenant des résultats compétitifs sur des *benchmarks* de

classification d'images, tout en utilisant des ressources computationnelles similaires. Leur architecture repose exclusivement sur des MLP, organisés en deux types de couches : des MLP appliqués indépendamment sur des patches d'images et des MLP appliqués entre ces patches. De nombreux travaux ultérieurs améliorent l'architecture MLP-Mixer, tels que : 1) *Region-aware MLP* (**RaMLP**) (Lai et al., 2023), qui résout une limitation des modèles MLP précédents imposant de fixer la taille des données en entrée et qui capture à la fois des indices visuels locaux et globaux de manière sensible aux régions ; 2) **HyperMixer** (Mai et al., 2023), qui introduit un mécanisme de mélange de *tokens* appelé *HyperMixing*. Ce mécanisme utilise des hyperréseaux pour générer dynamiquement les poids du MLP pour le mélange de *tokens* en fonction de l'entrée, permettant ainsi à HyperMixer de gérer des longueurs d'entrée variables et de systématiser la modélisation des interactions entre les *tokens* avec des poids partagés à travers les positions ; et 3) **Monarch-Mixer** (Fu et al., 2023), qui utilise des matrices *Monarch* pour améliorer les performances sur GPUs, montrant des résultats comparables voire supérieurs pour des tâches telles que la modélisation du langage ou la classification d'images, en employant moins de paramètres. Ces modèles sont conceptuellement plus simples par rapport à d'autres architectures comme les CNNs et les *Transformers*. Ils représentent également un bon compromis entre performance et efficacité.

L'architecture M2-Mixer (Bezirganyan et al., 2023) a été proposée pour la classification multimodale, tirant parti de la simplicité et de l'efficacité des MLP-Mixers. Elle utilise une fonction de perte multi-têtes pour résoudre le déséquilibre dans l'optimisation, garantissant qu'aucune des modalité ne domine le processus d'apprentissage. Cela aboutit à un modèle conceptuellement et computationnellement simple, surpassant les modèles de référence sur des ensembles de données multimodales utilisés dans le *benchmark*, avec une précision accrue et un temps d'entraînement significativement réduit. Cependant, les MLP-Mixers ne garantissent pas toujours des performances optimales pour toutes les modalités, et le choix de l'architecture MLP la plus efficace pour chaque tâche peut s'avérer complexe. Pour résoudre ce problème, le pipeline que nous proposons identifie systématiquement l'architecture MLP la mieux adaptée à chaque jeu de données, assurant ainsi des performances maximales tout en prenant en compte les spécificités des données.

NAS Multimodal : Les méthodes de recherche d'architecture neuronale (NAS) tentent d'automatiser le processus de recherche de l'architecture optimale pour des tâches et des ensembles de données choisis. Plusieurs approches de NAS multimodal ont été proposées dans la littérature en utilisant différents algorithmes et espaces de recherche (Pérez-Rúa et al., 2019; Yu et al., 2020; Yin et al., 2022; Xu et al., 2021). Ces approches sont souvent complexes à implémenter et à entraîner, et l'ajustement de certaines parties de l'architecture nécessite généralement de relancer l'ensemble du processus NAS. En revanche, le cadre que nous proposons est simple à mettre en œuvre et permet de conserver les micro-benchmarks existants, ce qui facilite les mises à jour rapides de l'architecture lorsque de nouveaux modules ou composants sont ajoutés à l'espace de recherche. De plus, la conception modulaire du pipeline permet d'appliquer les micro-benchmarks de manière sélective à des parties spécifiques de l'architecture, rationalisant ainsi le processus.

3 MixMAS

Nous proposons **MixMAS**, un *framework* pour la recherche automatisée d'architectures de fusion et d'apprentissage multimodal. MixMAS permet de sélectionner de manière efficace les architectures optimales basées sur les MLP pour diverses tâches et données multimodales, en exploitant la modularité et l'extensibilité afin de garantir une grande flexibilité. Le *pipeline* proposé se concentre sur des architectures MLP simples et efficaces, constituées uniquement de multiplications matricielles, de transformations linéaires et de fonctions d'activation, facilitant ainsi la recherche de configurations adaptées pour l'apprentissage multimodal.

Comme illustré dans la Figure 1, le processus est divisé en quatre étapes principales : 1) l'échantillonnage, 2) la sélection des encodeurs, 3) le choix de la fonction de fusion et 4) la sélection du réseau de fusion.

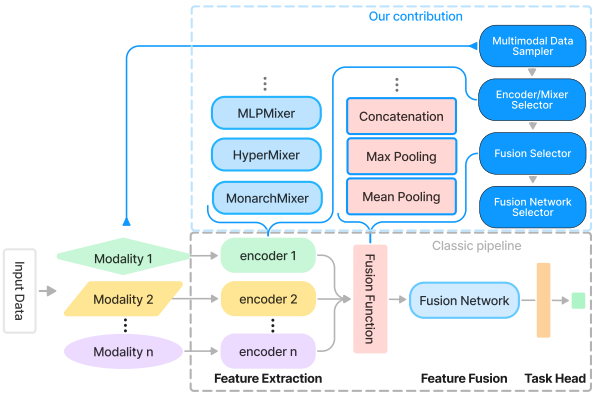


FIG. 1 – *Architecture de MixMAS*

Échantillonnage : Le module d'échantillonnage sélectionne un sous-ensemble du jeu de données pour comparer les performances des différents modules à chaque étape. Il garantit que l'échantillon sélectionné est représentatif de l'ensemble du jeu de données, ce qui est essentiel pour obtenir des métriques de performance précises et fiables afin de guider le processus de sélection. Le nombre d'échantillons est calculé selon l'équation (1), en utilisant la formule de détermination de la taille de l'échantillon Hogg et al. (2013) :

$$N' = \frac{n}{1 + \frac{z^2 \times \hat{p}(1-\hat{p})}{\varepsilon^2 N}}, \quad n = \frac{z^2 \times \hat{p}(1-\hat{p})}{\varepsilon^2}, \tag{1}$$

où N est la taille du jeu de données, z est le score-z, $\hat{p}$ est la proportion estimée de la population possédant l'attribut d'intérêt, et ε est la marge d'erreur (1%). Nous utilisons un échantillonnage aléatoire pour approximer la distribution des classes du jeu de données d'origine. Pour valider cette approche, nous calculons la distance entre les proportions des classes dans les jeux de données d'origine et échantillonné, en veillant à ce qu'elle reste inférieure à 0,05. Dans les travaux futurs, nous prévoyons de mettre en œuvre un échantillonnage basé sur l'incertitude pour obtenir des échantillons plus informatifs.

Sélection d'encodeurs : Cette étape consiste à évaluer les performances de divers encodeurs basés sur les MLP sur le jeu de données échantillonné pour chaque modalité. Le choix

TAB. 1 – *Résumé des jeux de données*

Jeu de données	Modalités	Taille	Tâche
MM-IMDB	Texte, Image	25 959	Classification : Genre de film
AV-MINST	Image, Audio	70 000	Classification : Chiffre de 0 à 9
MIMIC-III	Séries temporelles, Tabulaire	36 212	Classification : Mortalité

des encodeurs dépend des modalités du jeu de données, en veillant à ce que toutes les options soient basées sur des MLP. Les utilisateurs peuvent personnaliser les métriques d'évaluation dans le cadre pour les adapter à la nature spécifique du problème. Par exemple, pour un problème de classification binaire équilibré, l'*accuracy* pourrait être privilégiée, tandis que pour des classes déséquilibrées, des métriques comme la précision, le rappel ou le score F1 pourraient être plus appropriées. Bien que des encodeurs basés sur les MLP aient été utilisés dans ce travail, notre méthode permet d'incorporer d'autres encodeurs, tels que les Transformers, les CNNs, les RNNs, etc. L'encodeur optimal pour chaque modalité est ensuite sélectionné sur la base des résultats des évaluations pour être utilisé à l'étape suivante.

Sélection de la fonction de fusion : Cette étape consiste à choisir la meilleure fonction de fusion pour combiner les caractéristiques de chaque modalité. Nous utilisons une fusion intermédiaire, car la fusion brute est souvent impraticable en raison des structures de données différentes, et la fusion tardive peut être sous-optimale pour des modalités fortement corrélées Stahlschmidt et al. (2022). La fonction de fusion est évaluée sur la base des performances de classification.

Sélection du réseau de fusion : Cette étape est responsable de sélectionner le réseau approprié pour encoder les informations inter-modales et préparer l'*embedding* final pour la tête de tâche. De manière similaire à la sélection des encodeurs, plusieurs réseaux basés sur des MLP sont évalués. Les encodeurs pour chaque modalité et la fonction de fusion sont déterminés et fixés à partir des étapes précédentes.

Notre approche consiste à conserver les scores des micro-benchmarks après identification de l'architecture finale. En évitant de réévaluer les architectures déjà testées, cette stratégie simplifie le processus lors de l'incorporation de nouveaux composants ou de modifications.

4 Expérimentations

4.1 Configuration

Les expérimentations ont été menées avec les jeux de données suivants 4) :

— **MM-IMDB** [1] est un jeu de données multimodal avec des images (affiches de films) et du texte (résumés) pour la classification des genres. Nous avons utilisé BERT (Devlin et al., 2019) pour les embeddings textuels.

1. https ://github.com/johnarevalo/gmu-mmimdb

TAB. 2 – *Résultats du micro-benchmark pour chaque étape et jeu de données. Le module avec le score le plus élevé sera sélectionné.*

MM-IMDB		**AV-MNIST**		**MIMIC-III**	
Échantillonnage (%)	23%		12%		21%
Module	Score F1-p(%)	Module	Score Acc(%)	Module	Score Acc(%)
Sélection d'encodeur Image		**Sélection d'encodeur Image**		**Sélection d'encodeur Séries Temporelles**	
MLPMixer	**24.02**	MLPMixer	44.27	MLPMixer	40.77
HyperMixer	16.89	**HyperMixer**	**56.15**	**HyperMixer**	**45.36**
RaMLP	14.44	RaMLP	47.52	MonarchMixer	44.38
Sélection d'encodeur Texte		**Sélection d'encodeur Audio**		**Sélection d'encodeur Tabulaire**	
MLPMixer	9.20	MLPMixer	27.40	—	—
HyperMixer	15.07	**HyperMixer**	**29.16**	—	—
MonarchMixer	**28.55**	MonarchMixer	28.49	—	—
Sélection de la Fonction de Fusion		**Sélection de la Fonction de Fusion**		**Sélection de la Fonction de Fusion**	
ConcatFusion	**19.56**	**ConcatFusion**	**18.38**	**ConcatFusion**	**28.55**
MeanFusion	10.20	MeanFusion	9.61	MeanFusion	4.28
MaxFusion	9.07	MaxFusion	6.20	MaxFusion	6.73
Sélection du Réseau de Fusion		**Sélection du Réseau de Fusion**		**Sélection du Réseau de Fusion**	
HyperMixer	**29.0**	**HyperMixer**	**53.47**	**HyperMixer**	**38.15**
MLPMixer	25.97	MLPMixer	42.17	MLPMixer	34.14

— **AV-MINST** [2] combine les images MNIST [3] avec les données FSDD [4] (prononciations de chiffres).
— **MIMIC-III** [5] est un jeu de données cliniques contenant des séries temporelles (12 mesures médicales par heure sur 24 heures) et des données tabulaires.

Pour les micro-benchmarks, nous utilisons un taux d'apprentissage de 0,001, en entraînant sur les données échantillonnées pendant 10 epochs. Pour l'entraînement complet, nous commençons avec un taux d'apprentissage de 0,001 sur MM-IMDB, en utilisant un planificateur qui le réduit d'un facteur de 10 si la perte de validation ne montre aucune amélioration pendant 2 epochs. Pour AV-MINST et MIMIC-III, nous suivons la configuration d'entraînement de M2-Mixer décrite dans (Bezirganyan et al., 2023). Pour le jeu de données MM-IMDB, nous calculons un score F1 pondéré en raison du déséquilibre des classes. Les autres jeux de données sont évalués avec la métrique *accuracy*.

Nous utilisons des MLP-Mixers, RaMLP (Lai et al., 2023), HyperMixer (Mai et al., 2023) et MonarchMixer (Mai et al., 2023) comme candidats pour la fonction d'encodeur, et HyperMixer et MLPMixer comme candidats pour le réseau de fusion. Dans MIMIC-III, nous avons

2. https ://github.com/slyviacassell/_MFAS/tree/master
3. https ://yann.lecun.com/exdb/mnist/
4. https ://github.com/Jakobovski/free-spoken-digit-dataset
5. https ://physionet.org/content/mimiciii/1.4/

TAB. 3 – *Résultats sur les jeux de données **MM-IMDB**, **AV-MNIST** et **MIMIC-III**.*

	MM-IMDB		**AV-MNIST**		**MIMIC-III**	
Architecture	F1-p. (%) (moy.)	Paramètres	Acc. (%) (moy.)	Paramètres	Acc. (%) (moy.)	Paramètres
M2-Mixer	46.66 ± 0.44	16.7	73.20 ± 0.2	8.3	78.32 ± 0.3	0.029
MixMAS	$\mathbf{49.58 \pm 0.5}$	10.37	$\mathbf{75.79 \pm 0.3}$	9.33	78.3 ± 0.73	0.033

TAB. 4 – *Description des jeux de données*

Jeu de données	Modalités	Taille	Tâche
MM-IMDB	Texte, Image	25 959	genre de film
AV-MINST	Image, Audio	70 000	chiffre
MIMIC-III	Séries temporelles, Tabulaire	36 212	mortalité

opté pour un simple MLP fixe comme encodeur pour la modalité tabulaire. Nous comparons les performances de MixMAS avec celles de M2-Mixer.

Nous avons utilisé deux clusters internes avec des GPU NVIDIA GeForce RTX 3090, GeForce RTX 2080, A40 et V100. Le temps total d'exécution de toutes les expériences a été de 144 heures.

4.2 Résultats

Le tableau 2 résume les résultats des micro-benchmarks, où notre méthode sélectionne les modules ayant les scores les plus élevés pour construire l'architecture finale. Les résultats montrent qu'il n'existe pas de solution universelle pour les encodeurs de modalités ou les réseaux de fusion, car certains modules sont plus adaptés à des jeux de données et modalités particuliers. Cela met en évidence l'efficacité de notre approche, car MixMAS recherche l'architecture et la fonction de fusion adaptées à la tâche. De plus, nous remarquons que *Concat-Fusion* est systématiquement sélectionné à l'étape de la fonction de fusion, validant notre hypothèse selon laquelle la concaténation préserve davantage d'informations issues des modalités par rapport à l'agrégation par moyenne ou par maximum, améliorant ainsi les performances globales du modèle.

Le tableau 3 présente les résultats de l'entraînement du modèle final sur les jeux de données complets. Sur MM-IMDB, l'architecture sélectionnée par MixMAS dépasse M2-Mixer, atteignant un score moyen F1 pondéré de 49,58% contre 42,3% pour M2-Mixer, avec moins de paramètres (10,37 millions contre 16,7 millions). Pour AV-MNIST, l'architecture trouvée par MixMAS surpasse également M2-Mixer, atteignant une précision moyenne de 75,79% contre 73,2%. Les résultats pour MIMIC-III montrent des performances similaires entre l'architecture sélectionnée par MixMAS et M2-Mixer. Nous émettons l'hypothèse que l'incompatibilité de la modalité tabulaire avec les MLP-Mixers, combinée à l'utilisation d'un simple MLP fixe pour cette modalité, réduit l'espace de recherche.

5 Conclusion

Dans cet article, nous introduisons MixMAS, un *framework* efficace pour sélectionner des architectures optimales basées sur des MLP à l'aide d'échantillonnage et de micro-benchmarking. Notre approche repose sur la simplicité et l'efficacité des MLP-Mixers, en les adaptant à l'apprentissage multimodal. Les expériences menées confirment l'efficacité de notre proposition sur des ensembles de données ayant deux modalités.

Pour les travaux futurs, nous prévoyons de tester cette approche sur des ensembles de données avec davantage de modalités. Nous envisageons également d'explorer des méthodes d'échantillonnage alternatives, comme l'échantillonnage basé sur l'incertitude et la diversité, tout en élargissant la recherche pour inclure une gamme plus étendue de modules à sélectionner.

6 Remerciements

Ce travail a été publié dans le Workshop Multimodal AI (MMAI 2024) de la conférence IEEE BIG DATA 2024. Nous remercions le Centre de Calcul Intensif d'Aix-Marseille pour l'accès à ses ressources de calcul haute performance.

Références

Baltrušaitis, T., C. Ahuja, et L.-P. Morency (2018). Multimodal machine learning : A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence 41*(2), 423–443.

Bezirganyan, G., S. Sellami, L. Berti-ÉQuille, et S. Fournier (2023). M2-mixer : A multimodal mixer with multi-head loss for classification from multimodal data. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 1052–1058. IEEE.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Volume 1, pp. 2. Minneapolis, Minnesota.

Fu, D. Y., S. Arora, J. Grogan, I. Johnson, E. S. Eyuboglu, A. W. Thomas, B. Spector, M. Poli, A. Rudra, et C. Ré (2023). Monarch mixer : A simple sub-quadratic gemm-based architecture. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et S. Levine (Eds.), *Advances in Neural Information Processing Systems 36*.

Hogg, R., E. Tanis, et D. Zimmerman (2013). *Probability and Statistical Inference*. Pearson.

Lai, S., X. Du, J. Guo, et K. Zhang (2023). Ramlp : Vision mlp via region-aware mixing. In *IJCAI*, pp. 999–1007.

Liang, P. P., A. Zadeh, et L.-P. Morency (2022). Foundations and recent trends in multimodal machine learning : Principles, challenges, and open questions. *CoRR abs/2209.03430*.

Liu, R., Y. Li, L. Tao, D. Liang, et H. Zheng (2022). Are we ready for a new paradigm shift ? a survey on visual deep mlp. *Patterns (N Y) 3*(7), 100520.

Mai, F., A. Pannatier, F. Fehr, H. Chen, F. Marelli, F. Fleuret, et J. Henderson (2023). Hypermixer : An mlp-based low cost alternative to transformers.

Pérez-Rúa, J.-M., V. Vielzeuf, S. Pateux, M. Baccouche, et F. Jurie (2019). Mfas : Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6966–6975.

Stahlschmidt, S. R., B. Ulfenborg, et J. Synnergren (2022). Multimodal deep learning for biomedical data fusion : a review. *Briefings Bioinform. 23*(2).

Tolstikhin, I. O., N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, et A. Dosovitskiy (2021). Mlp-mixer : An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems 34*, pp. 24261–24272.

Xu, Z., D. R. So, et A. M. Dai (2021). Mufasa : Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 35, pp. 10532–10540.

Yin, Y., S. Huang, et X. Zhang (2022). Bm-nas : Bilevel multimodal neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 36, pp. 8901–8909.

Yu, Z., Y. Cui, J. Yu, M. Wang, D. Tao, et Q. Tian (2020). Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3743–3752.

Summary

Choosing a suitable deep learning architecture for multimodal data fusion is a challenging task, as it requires the effective integration and processing of diverse data types, each with distinct structures and characteristics. In this paper, we introduce MixMAS, a novel framework for sampling-based network architecture search tailored to multimodal learning. Our approach automatically selects the optimal MLP-based architecture for a given multimodal machine learning (MML) task. Specifically, MixMAS utilizes a sampling-based micro-benchmarking strategy to explore various combinations of modality-specific encoders, fusion functions, and fusion networks, systematically identifying the architecture that best meets the task's performance metrics. Empirical evaluations on the MM-IMDB and AV-MNIST datasets demonstrate that the proposed pipeline generates architectures that outperform baseline M2-Mixer models, achieving average improvements of +2.92% in F1-score and +2.59% in accuracy, respectively.