

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ DE SOUSSE
INSTITUT SUPÉRIEUR D'INFORMATIQUE ET DES TECHNOLOGIES DE COMMUNICATION - SOUSSE
المعهد العالي للعلوم و تكنولوجيات الاتصال بسوسة

N° d'ordre

| | | | | | | | | | | | |
|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| <input type="checkbox"/> |
|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|

GESTION ET ANALYSE DES DONNÉES MASSIVES

*Rapport de synthèse des travaux de recherche, présenté en vue de l'obtention de l'
Habilitation Universitaire*

Discipline
Informatique

Présentée et soutenue publiquement par
Khaled JOUINI
Docteur de l'université Paris Dauphine-PSL (France)
Maître assistant à l'ISITCom

Janvier 2025

Table des matières

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Contexte et motivations | 2 |
| 1.2 | Contributions et organisation du rapport | 3 |
| 2 | Volume : Optimisation de la localité spatiale des bases NoSQL | 7 |
| 2.1 | Introduction | 8 |
| 2.2 | Modélisation des agrégats : défis et considérations | 10 |
| 2.2.1 | Normalisation vs. Dénormalisation | 10 |
| 2.2.2 | Racine d'agrégat et hiérarchisation des accès | 10 |
| 2.2.3 | Dénormalisation niveau entité vs. dénormalisation niveau attribut | 11 |
| 2.3 | Approche guidée par les coûts pour l'identification et la sélection des agrégats | 12 |
| 2.3.1 | Vue d'ensemble | 12 |
| 2.3.2 | Détermination des agrégats pertinents pour les requêtes individuelles | 14 |
| 2.3.3 | Mesure de l'intérêt des agrégats | 15 |
| 2.3.4 | Fusion d'agrégats pertinents | 16 |
| 2.3.5 | Sélection des agrégats : un problème de sac à dos multiple (MKP) | 17 |
| 2.4 | Résultats & discussion | 19 |
| 2.5 | Conclusion | 22 |
| 3 | Variété : Extraction de caractéristiques pour la recherche d'information et l'apprentissage automatique | 24 |
| 3.1 | Introduction | 25 |
| 3.2 | Extraction de caractéristiques pour la reconnaissance d'empreintes | 25 |

TABLE DES MATIÈRES

| | | |
|----------|---|-----------|
| 3.2.1 | Reconnaissance d'empreintes : concepts clés et travaux antérieurs | 27 |
| 3.2.2 | EDT-C : Comparaison basée sur la triangulation de Delaunay étendue | 29 |
| 3.2.3 | Consolidation des modèles de référence | 31 |
| 3.2.4 | Résultats & discussion | 33 |
| 3.3 | Fusion de caractéristiques pour l'amélioration de la classification LULC | 35 |
| 3.3.1 | Apprentissage par transfert et classification LULC | 36 |
| 3.3.2 | Fusion de caractéristiques conçues manuellement et extraites automatiquement . | 37 |
| 3.3.3 | Résultats & discussion | 41 |
| 3.4 | Conclusion | 43 |
| 4 | Vélocité : Apprentissage sur flux, adaptatif aux dérives de concept | 46 |
| 4.1 | Introduction | 47 |
| 4.2 | Apprentissage en ligne : modèles de référence | 48 |
| 4.2.1 | Arbre de Hoeffding [DH00] | 48 |
| 4.2.2 | Forêts aléatoires en ligne (<i>Online Random Forest</i>) [SLS ⁺ 09] | 49 |
| 4.2.3 | <i>Fast Incremental Model Trees with Drift Detection</i> [IGD11] | 49 |
| 4.3 | Classification en ligne adaptative aux drifts - cas de la détection d'intrusions | 50 |
| 4.3.1 | Classification en ligne appliquée à la détection d'intrusion | 51 |
| 4.3.2 | DDM-ORF : classification ensembliste, adaptative aux drifts | 51 |
| 4.3.3 | Résultats & discussion | 53 |
| 4.4 | Régression en ligne adaptative aux drifts - cas de l'évolution des pandémies | 54 |
| 4.4.1 | Régression incrémentale appliquée à la prédiction épidémiologique | 54 |
| 4.4.2 | <i>Extremely Fast Regression Tree with Drift Detection</i> (EFRT-DD) | 55 |
| 4.4.3 | <i>Collaborative Drift-Driven Regression</i> (CDR) | 56 |
| 4.4.4 | Résultats & discussion | 57 |
| 4.5 | Conclusion | 62 |
| 5 | Véritacité : Détection des Fake News dans les médias sociaux | 64 |
| 5.1 | Introduction | 65 |
| 5.2 | Contributions à la détection d'opinions | 65 |
| 5.2.1 | <i>Stance Detection vs. Fake News Detection</i> | 66 |
| 5.2.2 | Apprentissage ensembliste basé sur l'augmentation | 67 |
| 5.2.3 | Résultats & discussion | 69 |
| 5.3 | De la <i>Stance</i> à la <i>Fake News Detection</i> : la Blockchain pour combler le chaînon manquant | 72 |
| 5.3.1 | La Blockchain au service de la vérification des faits : principales approches . . | 73 |
| 5.3.2 | Vérification décentralisée des faits, basée sur la Blockchain, la Stance Detection et la borne de Hoeffding | 74 |
| 5.3.3 | Résultats & discussion | 79 |

TABLE DES MATIÈRES

| | |
|--|-----------|
| 5.4 Conclusion | 82 |
| 6 Conclusion & perspectives | 84 |
| Récapitulatif des publications et des encadrements | 88 |

Table des figures

| | | |
|------|--|----|
| 1.1 | Thématiques de recherche | 3 |
| 2.1 | Modèle orienté agrégat vs. modèle relationnel. | 8 |
| 2.2 | Modèles de données alternatifs de la base DBLP | 9 |
| 2.3 | Étapes clés de l'approche de sélection d'agrégats [Jou22]. | 13 |
| 2.4 | Algorithme de sélection des agrégats | 18 |
| 2.5 | Illustration de la sélection des agrégats par séparation/évaluation | 19 |
| 2.6 | Schéma dénormalisé vs. schéma obtenu avec l'approche BFS. | 19 |
| 2.7 | Temps d'exécution réalisés par BFS [KA15], CLDA [YLJ18] et <i>Distorted Replicas</i> | 20 |
| 2.8 | Amélioration du temps d'exécution en fonction du nombre de répliques | 20 |
| 3.1 | Types de minuties : (a) Terminaison ; (b) Bifurcation [MMJP09]. | 26 |
| 3.2 | Triangulation de Delaunay [Del34] | 28 |
| 3.3 | Triangulation de Delaunay étendue $EDT(P)$ | 29 |
| 3.4 | Descripteur utilisé dans EDT-C. | 30 |
| 3.5 | Détermination des zones compatibles [GJK17b]. | 32 |
| 3.6 | Correspondances trouvées entre deux impressions d'une même empreinte par M3gl [MPGBGRAR12] et EDT-C. | 34 |
| 3.7 | Images extraites du jeu de données EuroSAT [Lom98] | 36 |
| 3.8 | Baseline SIFT-NN | 38 |
| 3.9 | Baseline Shallow CNN | 38 |
| 3.10 | Fusion précoce | 38 |
| 3.11 | Fusion tardive : Attention-Enhanced Dual Learning (ADL) | 39 |

TABLE DES FIGURES

| | |
|--|----|
| 3.12 Exemple illustrant le mécanisme d'attention guidé par les points d'intérêt SIFT. | 39 |
| 3.13 Fusion à mi-niveau : <i>Fusion of Local Attended CNN Features and Global CNN Features with Gating Mechanism</i> | 40 |
| 3.14 Distribution des poids dans la <i>feature map</i> podérée par SIFT (LFGF-CNN) | 43 |
| 4.1 Vue d'ensemble de l'approche DDM-ORF [JJK25b]. | 52 |
| 4.2 CDR - Processus d'apprentissage. | 57 |
| 4.3 CDR - Processus d'inférence. | 57 |
| 4.4 Nombre de cas quotidien - MAE réalisé par EFRT-DD, CDR et les modèles incrémentaux (Tunisie). | 59 |
| 4.5 Nombre de décès quotidien - MAE réalisé par EFRT-DD, CDR et les modèles incrémentaux (Tunisie). | 59 |
| 5.1 Exemple de <i>Stance Detection</i> repris du banc FNC-1 [HPS ⁺¹⁸] | 67 |
| 5.2 Stratégie classique de l'utilisation de l'augmentation avec le Stacking. | 69 |
| 5.3 Stratégie proposée pour la combinaison de l'augmentation et du Stacking. | 69 |
| 5.4 F1-scores obtenus sur le banc FNC-1 avec et sans augmentation du texte | 70 |
| 5.5 F1-scores obtenus sur le banc de tests FNN avec et sans augmentation du texte | 70 |
| 5.6 Vue d'ensemble du processus d'évaluation de la crédibilité d'une information | 75 |
| 5.7 Exemple extrait du jeu de données RumourEval [GBD ⁺¹⁸] | 79 |

Liste des tableaux

| | | |
|-----|---|----|
| 2.1 | Algorithme d'identification et de sélection d'agrégats appliqué à la base TPC-H. | 21 |
| 2.2 | Nombre d'itérations pour la résolution du problème de sélection d'agrégats. | 21 |
| 2.3 | Temps d'exécution de l'approche de regroupement des agrégats en fonction du facteur de réPLICATION C ($N=21$ items). | 21 |
| 3.1 | Performances obtenues avec différentes topologies sur la base DB1_A du FVC2000. | 33 |
| 3.2 | Précision (en %) sur les empreintes de la base DB1_A | 34 |
| 3.3 | Précision (en %) et temps de calcul obtenus sur les bases du FVC. | 34 |
| 3.4 | Accuracy obtenue par les modèles étudiés | 41 |
| 3.5 | Comparaison des performances des approches existantes et proposées | 42 |
| 4.1 | DDM-ORF vs. approches existantes | 53 |
| 4.2 | Taux d'ingestion et de traitement de DDM-ORF sous Apache Spark Structured Streaming | 53 |
| 4.3 | MAE et RMSE réalisés par les modèles incrémentaux (tous pays confondus) | 58 |
| 4.4 | MAE et RMSE réalisés par les modèles incrémentaux (Tunisie) | 58 |
| 4.5 | Nombre de décès quotidien - MAE et RMSE réalisés par CDR et l'arbre Batch sous-jacent. | 60 |
| 4.6 | Nombre de cas quotidien - MAE et RMSE réalisés par CDR et l'arbre Batch sous-jacent. | 61 |
| 5.1 | Accuracy et F1-Score obtenus par le stacking conventionnel, [LMN ⁺ 19] et notre approche | 71 |
| 5.2 | Distribution des classes dans FNC-1 [HPS ⁺ 18] | 71 |
| 5.3 | F1-Score par classe avec et sans équilibrage des classes par augmentation (FNC-1) . . . | 71 |

LISTE DES TABLEAUX

| | | |
|-----|---|----|
| 5.4 | Score F1 macro-moyen obtenu par notre framework avec le modèle de référence BranchLSTM [KLZ18] | 80 |
| 5.5 | Score F1 macro-moyen obtenu par notre framework lorsque toutes les stances sont correctement identifiées (similaire à un système de vote) | 80 |
| 5.6 | Comparaison des scores F1 macro-moyen obtenu par les différentes approches. | 81 |

CHAPITRE 1

Introduction

Sommaire

| | | |
|-----|--|---|
| 1.1 | Contexte et motivations | 2 |
| 1.2 | Contributions et organisation du rapport | 3 |

1.1 Contexte et motivations

La célérité de l'évolution de l'informatique et la complexité accrue à laquelle doivent faire face les systèmes centrés sur les données, exigent des différents acteurs (enseignants, chercheurs et industrie) une veille informationnelle permanente et une capacité d'adaptation rapide à ce que l'on peut considérer comme des "dérives de concept"¹, aussi fréquentes qu'abruptes, de paradigmes, de besoins et de technologies. Ce rapport, présenté en vue de l'obtention de l'habilitation universitaire, est une rétrospective retraçant ma trajectoire de chercheur depuis mes travaux de thèse et son évolution façonnée par les avancées connues par le monde de la gestion et du traitement de ce qui est couramment appelé la *data*.

Les données sont un peu la matière première de tout système informatique. Comme toute matière première, les données existent sous différentes formes, parfois désorganisées, et nécessitent un traitement approprié pour en tirer de la valeur. Ceci englobe aussi bien les processus de transformation et de stockage que ceux d'interrogation et d'analyse. Depuis mes travaux de thèse qui ont porté sur l'organisation logique et physique des données dans les bases relationnelles, le monde du stockage et du traitement des données a connu d'importants développements, principalement en raison de la croissance exponentielle du volume de données produites au quotidien et des avancées technologiques qui s'en sont suivies. Cette prolifération a été alimentée par la numérisation croissante de nos vies et l'expansion des technologies de l'information qui en ont investi aussi bien les aspects les plus ludiques que les aspects les plus critiques.

Avec leurs fondements théoriques solides (transactions ACID, théorie des ensembles, etc.) ainsi que l'expérience et la maturité cumulées sur près d'un demi-siècle, les formidables systèmes que sont les SGBD relationnels ont résisté avec brio au poids de l'âge et continuent à être aujourd'hui le premier moyen pour gérer les données structurées. Cependant, l'émergence depuis près d'une décennie de la nécessité de gérer de gros volumes de données, distribuées à grande échelle et de différents types et formats, a fait naître de nouveaux besoins et de nouveaux défis auxquels les bases relationnelles ne peuvent répondre, en raison principalement des règles strictes qu'elles s'imposent (théorème CAP [GL02]). Ceci a eu comme effet de bord un ralentissement de la recherche sur les bases relationnelles au profit des systèmes NoSQL et plus globalement des systèmes destinés à la gestion et à l'analyse des données massives (*Big Data*).

Lorsque l'on évoque les données massives, il est courant de mettre l'accent sur leur *volume* considérable. Cependant, le volume n'est pas le seul défi que posent les données massives. Au côté du *Volume*, la *Variété*, la *Véracité* et la *Vélocité* sont les autres dimensions complétant les 4 Vs les caractérisant. La variété dénote l'hétérogénéité des formats et la diversité des sources qui peuvent être à l'origine des données (réseaux sociaux, télédétection, etc.). La véracité fait référence à l'exactitude, l'intégrité et la fiabilité des données et/ou des sources qui en sont à l'origine. Finalement, la vélocité désigne la vitesse

1. Le terme "dérive de concept" est utilisé dans ce contexte pour désigner un changement important résultant de l'évolution des connaissances ou des progrès technologiques et remettant en question les méthodes et les approches établies. La définition précise du terme dans le contexte de l'apprentissage incrémental continu est donnée plus tard.

1.2. Contributions et organisation du rapport

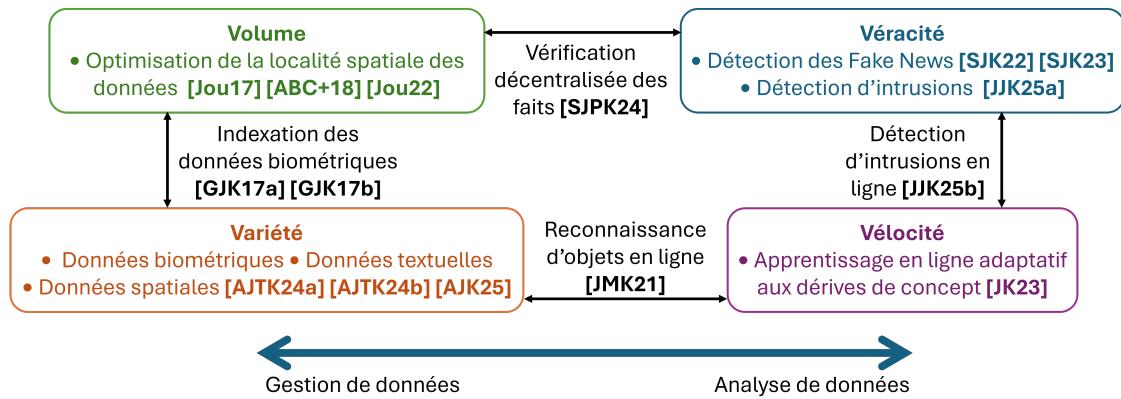


FIGURE 1.1 – Thématisques de recherche

à laquelle les données sont générées, collectées, transformées et traitées. La complexité inhérente aux données massives a conduit à une convergence des fonctionnalités, d'ingestion, de transformation, de stockage et d'analyse dans des plateformes unifiées et des écosystèmes intégrés (*e.g.* Apache Hadoop et Apache Spark). Ces plateformes ont pour mérite de couvrir l'ensemble du cycle de vie des données, de manière cohérente et relativement simple.

Aussi bien dans mes activités d'enseignement que de recherche, j'ai toujours été animé par le souci de ne pas me cantonner dans mes acquis et de rester en permanence en phase avec les évolutions que connaît le monde du stockage et du traitement des données. Le fléchissement de la recherche sur les bases relationnelles, conjugué à l'émergence des données massives et des écosystèmes dédiés à leur gestion, a donc déclenché un tournant dans ma carrière d'enseignant-chercheur. Les défis et les infinies opportunités qu'offrent les données massives m'ont ainsi incité à ouvrir le spectre de mes thématiques de recherche vers les bases NoSQL et les systèmes de fichiers distribués, dans un premier temps, et à initier une transition progressive de la gestion vers l'analyse des données, dans un second temps. Comme illustré dans la Figure 1.1, les thématiques de recherche que j'ai pu explorer et qui font l'objet de ce manuscrit, correspondent aux 4 Vs caractérisant les données massives. Il convient de noter que ces axes ne sont pas cloisonnés, dans le sens où un même travail de recherche peut être placé sous différentes bannières.

1.2 Contributions et organisation du rapport

Les travaux exposés dans ce manuscrit suivent une progression chronologique illustrant la transition progressive que j'ai entamé vers les thématiques ayant trait aux dimensions caractérisant le Big Data. Les contributions que j'ai pu apporter au regard de ces axes sont regroupées dans les chapitres ci-après.

Chapitre 2 – Volume : Optimisation de la localité spatiale des bases NoSQL.

Les thématiques de recherche abordées dans le second chapitre de ce manuscrit sont dans la lignée de mes travaux de thèse et portent sur l'optimisation de la localité spatiale des données. Le principe de *localité spatiale* stipule que les données susceptibles d'être lues ou écrites au même moment, doivent être regroupées ensemble. Dans le cas des bases relationnelles, ce principe se matérialise par le placement des données ayant une forte probabilité d'être manipulées ensemble de manière contiguë sur la mémoire persistante, afin de minimiser les accès disque et d'améliorer les performances des requêtes.

Les systèmes Big Data sont par essence des systèmes distribués. Le principe de localité spatiale appliqué aux données massives implique donc de regrouper sur un même nœud, les données auxquelles on accède au même moment, pour éviter les transactions et les jointures inter-nœuds *cross-nodes*. La flexibilité des modèles de données des bases NoSQL fait cependant que l'on soit confronté à une multitude de choix de regroupement. Dans nos travaux de recherche, nous exploitons la réPLICATION, ubiquitaire dans les systèmes NoSQL, pour empaqueter les données répliquées de différentes manières et améliorer ainsi la capacité de la base à prendre en charge des patterns d'accès hétérogènes. En sus de cette forme de réPLICATION dite déformée, nous proposons une approche guidée par les coûts dans laquelle nous transposons le problème de regroupement de données en un problème de sac à dos multiple et nous le résolvons avec la technique de séparation/évaluation (*Branch & Bound*).

Chapitre 3 – Variété : Extraction de caractéristiques pour la recherche d'information et l'apprentissage automatique

L'*extraction de caractéristiques* consiste à transformer des données brutes en représentations compactes et informatives, adaptées à l'*analyse* et à l'*indexation*. Les défis liés à la variété caractérisant le Big Data se matérialisent essentiellement dans l'extraction de caractéristiques : une fois les données brutes transformées en descripteurs numériques, un même algorithme opère globalement de la même manière, avec des variations selon le paramétrage et l'objectif. Les travaux exposés dans le troisième chapitre abordent les deux principaux cas d'usage de l'extraction de caractéristiques : la *recherche d'information* et l'*apprentissage automatique*. Ces travaux ont été menés dans le cadre de deux thèses que j'ai eu l'honneur de co-encadrer sous la direction du Pr. Ouajdi KORBA.

La première thèse a porté sur la reconnaissance d'empreintes et a été conduite dans le cadre du programme MOBIDOC en collaboration avec la société Sam's-Tech, au profit du ministère de l'intérieur. La plupart des algorithmes de reconnaissance d'empreintes n'ont aucune difficulté à reconnaître une empreinte lorsque la recherche se fait dans une petite base ou lorsque les images sont de bonne qualité. Ces algorithmes se révèlent cependant moins précis lors que la recherche se fait dans des bases volumineuses (à cause des concordances fortuites inévitables) ou bien lorsque les images sont de mauvaise qualité (e.g. empreintes relevées sur des scènes de crime ou obtenues par encrage des doigts). Dans nos travaux de recherche nous proposons notamment une approche novatrice basées sur la triangulation de Delaunay pour représenter les empreintes par des topologies à haut pouvoir discriminant et à grande robustesse à l'égard des anomalies pouvant les affecter.

La seconde thèse a donné lieu à une collaboration avec la *Northern Technical University* (Mossoul, Irak) et a porté sur la classification des images satellitaires. Les caractéristiques conçues manuellement (*Hand-crafted*), telles que celles générées par l'illustre SIFT (*Scale-Invariant Feature Transform* [Low04]), excellent dans la capture de détails locaux distinctifs. Cependant, de par leur nature locale, ces méthodes peinent à saisir le contexte global. À l'inverse, les réseaux de neurones convolutionnels (*Convolutional Neural Networks*, ou CNN) sont capables d'apprendre des représentations hiérarchiques riches, mais tendent souvent à lisser les détails subtiles à granularité fine. Dans nos travaux, nous explorons diverses stratégies basées sur des mécanismes d'attention pour tirer profit des avantages complémentaires des caractéristiques *Hand-Crafted* et *CNN-Learned*.

Chapitre 4 - Vélocité : Apprentissage sur flux, adaptatif aux dérives de concept

Le quatrième chapitre de ce manuscrit s'intéresse à la dimension temporelle du Big Data, la *Vélocité*. Celle-ci désigne la capacité d'un système à traiter les données de manière incrémentale, au fur et à mesure de leur génération, étant donné que les réponses tardives sont souvent inutiles. Ce chapitre explore l'apprentissage incrémental (également appelé *online learning* ou *lifelong learning*) adaptatif aux dérives de concept, à travers deux projets menés en collaboration avec ma collègue Dr. Farah JEMILI : la *détection d'intrusions* dans les systèmes informatiques et la *prédiction de l'évolution des pandémies*.

Bien que ces deux problématiques puissent sembler sans lien, elles partagent des défis communs liés à la nature dynamique des données. Tout comme les intrusions ajustent leurs tactiques pour contourner les mécanismes de sécurité, les virus biologiques adaptent leurs comportements pour échapper aux défenses immunitaires humaines. Par ailleurs, dans les deux cas, les données arrivent sous forme de flux potentiellement infinis, rendant impossible l'attente de leur disponibilité complète avant d'initier l'apprentissage. Enfin, dans ces environnements dynamiques où l'apprentissage se fait sur le long terme, la relation entre la variable cible et les variables explicatives évolue inévitablement avec le temps. La détection et l'adaptation à ces évolutions, appelées *dérives de concept*, sont le seul garant pour éviter la dégradation des performances prédictives.

Les travaux présentés dans ce chapitre portent sur la classification et la régression incrémentales. Nous proposons notamment un nouvel arbre de régression incrémental, conçu pour s'adapter aux dérives de concept grâce à une stratégie d'induction proactive (*Eager*), en contraste avec les approches conservatrices classiques. Nous introduisons également des stratégies collaboratives intégrant la détection des dérives, l'apprentissage incrémental et par lots, ayant pour objectif d'améliorer l'adaptabilité et d'accroître la précision.

Chapitre 5 - Vérité : Détection des Fake News dans les médias sociaux

Le dernier chapitre de ce manuscrit aborde une dimension qualitative du Big Data, la *Vérité*, en mettant l'accent sur la détection des Fake News dans les médias sociaux. Les travaux qui y sont présentés introduisent un nouveau paradigme de vérification des faits, combinant la détection d'opinions

1.2. Contributions et organisation du rapport

et la technologie blockchain. Ces recherches ont été menées dans le cadre de la thèse de Melle Ilhem Salah, en collaboration avec l'École Hexagone (Versailles, France) et l'École d'ingénieurs ESIGELEC (Rouen, France).

Les systèmes conventionnels de vérification des faits, souvent basés sur un processus de validation manuel et/ou centralisé, sont intrinsèquement limités en termes de scalabilité et exposés à des risques de biais, de manipulation et de censure. Parallèlement, bien qu'utiles, les approches classiques basées sur l'apprentissage automatique nécessitent souvent un ré-apprentissage périodique des modèles pour mettre à jour leurs connaissances statiques et potentiellement obsolètes du monde (comme le font, par exemple périodiquement, les grands modèles de langage). Grâce à sa nature décentralisée, transparente et sécurisée, la blockchain offre un potentiel unique pour une vérification des faits scalable et affranchie de toute autorité centrale.

Dans nos travaux de recherche, nous proposons un framework décentralisé de vérification des faits, construit autour de la blockchain et du mécanisme de consensus *Proof of Reputation*. Ce framework vise à évaluer la crédibilité d'une information en analysant les opinions exprimées par une communauté d'utilisateurs, ces opinions étant pondérées par la fiabilité (*i.e.* la *réputation*) respective de ces derniers. Au sein de ce framework, la détection d'opinions (*Stance Detection*) est utilisée pour inférer de manière automatisée les "votes" de différentes sources sur une information, même lorsque ceux-ci sont exprimés de manière implicite. En décentralisant et automatisant de bout en bout la vérification des faits, notre framework se démarque des approches existantes, qui se limitent pour la plupart à une simple traçabilité des informations sans exploiter pleinement le potentiel de la blockchain.

CHAPITRE 2

Volume : Optimisation de la localité spatiale des bases NoSQL

Sommaire

| | | |
|------------|---|-----------|
| 2.1 | Introduction | 8 |
| 2.2 | Modélisation des agrégats : défis et considérations | 10 |
| 2.2.1 | Normalisation vs. Dénormalisation | 10 |
| 2.2.2 | Racine d'agrégat et hiérarchisation des accès | 10 |
| 2.2.3 | Dénormalisation niveau entité vs. dénormalisation niveau attribut | 11 |
| 2.3 | Approche guidée par les coûts pour l'identification et la sélection des agrégats | 12 |
| 2.3.1 | Vue d'ensemble | 12 |
| 2.3.2 | Détermination des agrégats pertinents pour les requêtes individuelles | 14 |
| 2.3.3 | Mesure de l'intérêt des agrégats | 15 |
| 2.3.4 | Fusion d'agrégats pertinents | 16 |
| 2.3.5 | Sélection des agrégats : un problème de sac à dos multiple (MKP) | 17 |
| 2.4 | Résultats & discussion | 19 |
| 2.5 | Conclusion | 22 |

2.1 Introduction

LES SGBDs NoSQL ont été conçus et optimisés pour tourner sur de larges grappes de serveurs géographiquement distants et pour assurer la haute disponibilité, la résistance aux pannes, la scalabilité horizontale et de hautes performances pour les accès en lecture et en écriture. Les bases de données orientées document font partie de la mouvance NoSQL et en sont le type le plus diffusé et le plus populaire.

Les bases NoSQL orientées document s'appuient sur la flexibilité des documents structurés (*e.g.* JSON, XML, etc.) pour empaqueter les données qui ont de grandes chances d'être lues au même moment dans un seul document autonome. Ce faisant, les bases orientées document manipulent les données ainsi associées en une seule opération et réduisent les transactions et les jointures *cross-nodes*. Le concept de document est similaire au concept d'*agrégat* issu de la conception pilotée par le domaine (*Domain-Driven-Design*) [Eri03]. La Figure 2.1 illustre un document au format JSON provenant des archives de la bibliographie DBLP [Ley09] et illustre les principales différences entre le modèle de données agrégées et le modèle relationnel.

Un des défis majeurs dans les bases NoSQL est de structurer les documents de la manière qui sied le mieux à la fois à la distribution et aux patterns fréquents d'accès aux données. Plusieurs facteurs rendent ce problème non trivial : (i) Il n'existe pas de règles de passage claires d'un modèle conceptuel à un schéma agrégé, comme il en existe pour le schéma normalisé des bases relationnelles [RLRJ17] ; (ii) Un large éventail d'alternatives doit souvent être considéré (comme illustré dans la Figure 2.2, il est courant d'avoir une explosion combinatoire de schémas alternatifs) ; et (iii) Un choix de modélisation judicieux est entièrement tributaire de la façon dont nous tendons à manipuler les données [SF12], et



FIGURE 2.1 – Modèle orienté agrégat vs. modèle relationnel.

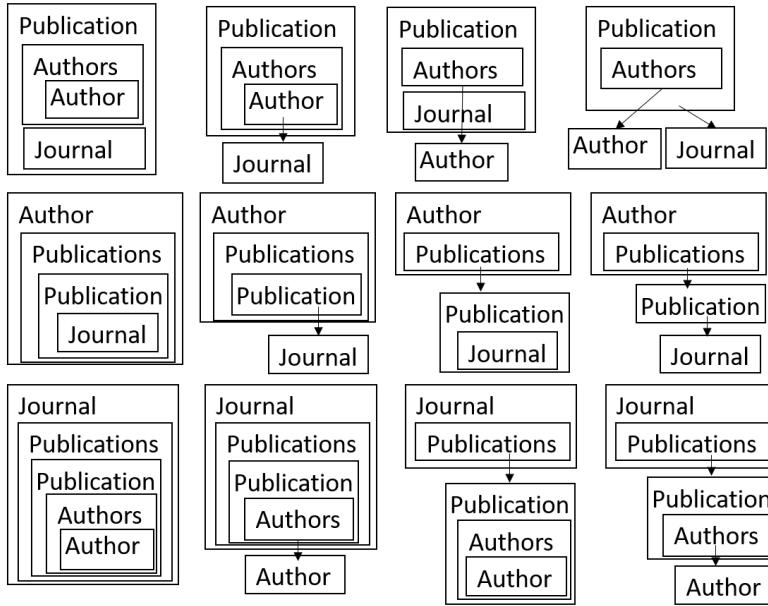


FIGURE 2.2 – Schémas alternatifs pour la base DBLP. Les rectangles représentent les documents (imbriqués ou non). Les flèches représentent des références entre agrégats. Le champ Authors (resp. Publications) représente un tableau de références ou de documents imbriqués.

doit donc être dicté par les coûts des requêtes fréquentes et influencé par la charge de travail subie par le système [dLdSM15].

La réPLICATION est ubiquitaire dans les systèmes NoSQL et est l'un des fondements de tout système destiné aux données massives. Dans nos travaux, nous proposons une nouvelle forme de réPLICATION, appelée *Distorted Replicas* (répliques déformées). L'idée derrière les répliques déformées est de restructurer les données répliquées de différentes manières, pour mieux faire face à l'hétérogénéité des patterns d'accès aux données. En complément, nous proposons une approche guidée par les coûts pour la sélection des agrégats. Un point saillant de cette approche est qu'elle transpose le problème de sélection en un problème de sac à dos multiple et le résout avec la technique *Branch & Bound*.

La suite de ce chapitre s'organise en quatre sections. La première présente les principaux défis liés à la modélisation des agrégats et les principales approches pour les surmonter. La deuxième décrit notre approche de répliques déformées et notre méthode de sélection d'agrégats. La troisième section présente une étude expérimentale menée sur le banc d'essai standard TPC-H. Suit la conclusion.

2.2 Modélisation des agrégats : défis et considérations

2.2.1 Normalisation vs. Dénormalisation

L'un des choix les plus cruciaux dans la modélisation des bases orientées agrégat est la matérialisation des associations entre entités : par référencement (*normalisation*) ou par imbrication de structures (*dénormalisation*). Avec le référencement, l'association entre deux entités se matérialise par l'imbrication d'un lien ou d'une référence d'une entité vers une autre, comme dans le modèle relationnel. En ne gardant qu'une seule copie des données, le modèle normalisé minimise la redondance et favorise la cohérence [RLRJ17]. Cependant, si les données qui ont de grandes chances d'être manipulées ensemble se trouvent dispersées entre plusieurs documents, voire serveurs, les lectures/écritures peuvent être prohibitivement lentes.

L'imbrication de structures matérialise l'association entre deux entités en les empaquetant dans une unité d'information autonome, *i.e.* le document ou l'*agrégat*. Ce faisant, l'imbrication favorise la localité spatiale des données, évite la jointure et présente une organisation des données qui sied à la distribution. Les avantages de l'imbrication ne sont pas sans contrepartie. L'imbrication n'est intéressante que pour les requêtes qui ont besoin d'accéder à plusieurs entités en même temps. Cependant, souvent, il existe des requêtes qui ont besoin d'accéder aux entités de manière individuelle, voire même, de n'accéder qu'à une faible proportion des attributs d'une même entité. Dans les bases orientées document l'unité de lecture/écriture est l'agrégat (*i.e.* le document). Lorsqu'une requête ne concerne qu'une faible proportion d'attributs, ce ne sont pas uniquement ces attributs qui sont lus, mais également tous les attributs de toutes les entités contenues dans le même agrégat. La lecture d'attributs inutiles à une requête conduit à : (i) une sous-utilisation de la bande passante du réseau; (ii) la pollution de la mémoire vive; et (iii) l'augmentation du temps CPU perdu en attente du chargement des données. Nous appelons dans la suite, le temps perdu dans le chargement de données inutiles à une requête le **sur-coût à la lecture** (*Read Overhead*).

2.2.2 Racine d'agrégat et hiérarchisation des accès

Un agrégat est plus qu'un simple ensemble d'entités. Il est en plus caractérisé par une racine, *i.e.* par l'entité selon laquelle le reste des entités de l'agrégat sont groupées. Deux agrégats empaquant les mêmes entités et ayant des racines différentes sont en effet deux agrégats différents avec des performances à la lecture/écriture différentes.

Considérons la base DBLP exemple et supposons que les entités Auteur et Publication soient regroupées dans un même agrégat. Les entités peuvent être regroupées alternativement soit par Auteur soit par Publication. Regrouper les auteurs par publication, permet de favoriser les accès par Publication et pénalise les accès par Auteur. Inversement, regrouper les publications par Auteur, favoriserait les accès par Auteur aux dépens des accès par Publication. Cette *hiérarchisation des accès*, inéluctable dans le modèle orienté agrégat, a un fort impact sur les performances des requêtes. Le plus souvent, il

existe plusieurs entités candidates à être racines. Le choix d'une racine parmi les alternatives possibles, sied à un certain nombre de requêtes, mais est forcément un obstacle pour beaucoup d'autres. Comme relevé dans [SF12] le bon fonctionnement de toute l'approche d'agrégation dépend du fait que les accès soient alignés ou non aux racines d'agrégat.

Dans la suite, nous représentons un agrégat a par une paire (r, E) , où $a.E$ est l'ensemble des entités apparaissant dans a et $a.r \in a.E$ est la racine de a . Lors qu'un agrégat est formé par une seule entité, il est dit *atomique*. Pour bénéficier de la rapidité des opérations binaires dans nos algorithmes, nous représentons $a.E$ par un bitmap de $|\mathbb{E}|$ bits, où \mathbb{E} est l'ensemble des entités modélisées. Le $i^{\text{ème}}$ bit du bitmap est mis à 1 si la $i^{\text{ème}}$ entité modélisée apparaît dans a , et à 0 sinon.

2.2.3 Dénormalisation niveau entité vs. dénormalisation niveau attribut

Sans surprise, la plupart des travaux existants sur la modélisation des bases NoSQL et la migration des bases relationnelles vers les bases NoSQL tentent de trouver le meilleur compromis entre référencement (dénormalisation) et imbrication de structures (dénormalisation).

Nous distinguons deux types de dénormalisation : la *dénormalisation au niveau entité* et la *dénormalisation au niveau attribut*. Pour la dénormalisation au niveau entité, [KA15] propose un algorithme de recherche en largeur d'abord (*Breadth-First Search*) pour trouver un chemin à partir d'une table e et les tables qui lui sont liées. Ce chemin sert à créer un agrégat de racine e , par l'imbrication récursive (de proche en proche) des entités ayant des liens avec e ou avec une des entités déjà imbriquées dans l'agrégat. Pour tenter de réduire la redondance, l'approche de [KA15], appelée BFS dans la suite, n'imbrique pas une même entité plus d'une fois dans un même agrégat. L'un des inconvénients majeurs de cette approche est qu'elle n'est pas guidée par les coûts : BFS empaquette ensemble des entités liées qui ne sont pas forcément interrogées fréquemment ensemble.

Dans [YLJ18], Yoo & al. ont proposé l'approche CLDA (*Column-Level Denormalization with Atomicty*) pour la modélisation des bases NoSQL. L'idée fondamentale dans CLDA est de dénormaliser au niveau des attributs et non au niveau des entités. Plus explicitement, CLDA commence par garder toutes les entités modélisées dans des agrégats atomiques. Ensuite, CLDA considère les requêtes de la charge de travail incluant des jointures. Au lieu d'éviter une jointure par l'imbrication de toute une entité dans une autre, CLDA n'imbrique que les colonnes (*i.e.* attributs) référencées dans la requête. Ce faisant, CLDA vise à éviter les opérations de jointure tout en minimisant le sur-coût à la lecture. Les résultats expérimentaux présentés dans [YLJ18] confirment le bien-fondé de l'approche et montre qu'elle permet une nette amélioration des performances. L'inconvénient majeur de cette approche est qu'elle tend à imbriquer systématiquement les colonnes sans tenir compte des cas où le gain obtenu par l'imbrication est inférieur à la perte causée par la sur-coût à la lecture.

2.3 Approche guidée par les coûts pour l'identification et la sélection des agrégats

Dans cette section nous présentons notre approche pour la structuration d'une base orientée agrégat en fonction de la charge de travail à laquelle elle répond. Notre travail est particulièrement utile lors de la migration d'une base relationnelle vers une base NoSQL ou lors qu'il est nécessaire de restructurer une base NoSQL pour en optimiser les performances.

2.3.1 Vue d'ensemble

L'idée directrice derrière notre approche *Distorted Replicas* (*répliques déformées*) [Jou17] est de restructurer les données répliquées de différentes manières pour mieux faire face à l'inévitable hétérogénéité des patterns d'accès aux données. Dans notre travail, la restructuration des données se matérialise par : (i) le remplacement d'un référencement par une imbrication et inversement ; et (ii) la réorganisation des données en fonction de racines d'agrégat différentes. Il convient de noter ici que la restructuration des agrégats selon de nouvelles racines n'est pas fastidieuse compte tenu du fait que la plupart des systèmes NoSQL orientés document fournissent des opérateurs optimisés permettant de promouvoir une entité d'"imbriquée" à "racine" (*e.g. replaceRoot* dans MongoDB). Nous devons également noter que dans notre travail, les données sont uniquement réorganisées localement, *i.e.* au sein d'un même membre d'un ensemble de réPLICATION. Par conséquent, même si des références sont utilisées, nous n'avons pas à effectuer de jointures distribuées pour reconstruire un agrégat.

Comme illustré dans la Figure 2.3, nous supposons que l'on dispose d'un ensemble de requêtes fréquentes W (charge de travail ou *workload*) et d'une base de données D ayant un facteur de réPLICATION $C \geq 1$ ($C = 1$ correspond au cas particulier où la base n'est pas répliquée). L'objectif est de trouver C sous-ensembles d'agrégats, complets et disjoints, permettant de réduire le temps d'exécution de W , sous deux contraintes :

- i *Contrainte de disjonction (non-chevauchement)* : les agrégats d'un même sous-ensemble doivent être disjoints entre-eux, dans le sens où chaque entité modélisée n'apparaît que dans un seul agrégat ;
- ii *Contrainte de restaurabilité* : bien qu'elles ne soient pas physiquement identiques, les C répliques doivent être logiquement identiques et constructibles les unes à partir des autres. Cela signifie que toutes les entités doivent apparaître dans chacun des C sous-ensembles.

En des termes plus explicites, ces deux contraintes imposent que toutes les entités modélisées apparaissent une et une seule fois dans chacun des C membres d'un Replica Set.

Comme indiqué précédemment, énumérer toutes les combinaisons de racines, d'imbrications et de référencements peut aisément conduire à une explosion combinatoire du nombre d'agrégats à considérer et à évaluer. Pour réduire ce nombre, nous nous limitons dans la première étape de notre approche "Détermination des agrégats pertinents à partir des requêtes individuelles" aux agrégats qui sont perti-

2.3. Approche guidée par les coûts pour l'identification et la sélection des agrégats

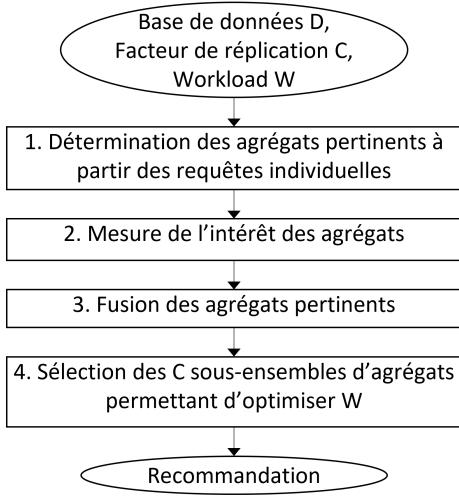


FIGURE 2.3 – Étapes clés de l'approche de sélection d'agrégats [Jou22].

nents pour au moins une requête de W . Les agrégats que nous devons retenir dans notre solution finale ont pour contrainte fondamentale d'être disjoints, *i.e.* de ne pas imbriquer les mêmes entités. Cette contrainte peut conduire à l'élimination d'un grand nombre d'agrégats pertinents. L'objectif de l'étape de "Fusion d'agrégats pertinents" est donc de trouver des agrégats supplémentaires, qui bien qu'ils ne soient pas optimaux pour une requête particulière, pourraient se révéler utiles pour différentes requêtes de W . Étant donné l'ensemble d'agrégats intéressants déterminés comme décrit ci-haut, l'objectif de l'étape ultime de notre approche est de sélectionner le sous-ensemble d'agrégats disjoints et complet permettant de maximiser l'intérêt total de la base, *i.e.* de réduire le temps d'exécution de W . Dans notre travail nous transposons la problématique de sélection d'agrégats intéressants à un problème de sac à dos multiple et nous le résolvons avec la technique de séparation/évaluation (*Branch & Bound*).

Comme dans [CN07, JQD11, dLdSM15], nous supposons l'existence d'une fonction $Cost(q, a)$ qui donne le coût estimé d'une requête q , lorsqu'elle est exécutée en utilisant l'agrégat a . Comme noté dans [CN07], plusieurs SGBDs disposent d'optimiseurs de requêtes permettant de fournir de telles estimations. En l'absence de cette fonctionnalité, il est possible de suivre la même approche que [JQD11, dLdSM15] et estimer $Cost(q, a)$ par le volume de données lues par q , *i.e.* par le nombre d'octets lus pour répondre à q en utilisant a . Le nombre d'octets lus correspond à :

$$Cost(q, a) = \text{Taille}(a) * \text{Cardinalité}(a) * \text{Selectivité}(q)$$

, où $\text{Taille}(a)$ est la taille estimée de a , $\text{Cardinalité}(a)$ le nombre d'instances de a et $\text{Selectivité}(q)$, le pourcentage d'instances de a que q sélectionne. Sans perte de généralité, il est possible d'utiliser d'autres estimations du coût d'une requête dans notre algorithme de sélection d'agrégats sans en modifier la démarche.

2.3.2 Détermination des agrégats pertinents pour les requêtes individuelles

La première étape dans notre approche consiste à identifier les agrégats pertinents pour les requêtes individuelles. Un agrégat est défini par l'ensemble des entités qu'il imbrique et par sa racine. Les racines potentielles désignent les entités référencées dans une requête qui pourraient être de "bonnes" racines d'agrégat. L'identification des racines potentielles est à la base de l'identification des agrégats pertinents.

Racines potentielles Soit $q.E$ l'ensemble des entités modélisées référencées dans une requête $q \in W$. Intuitivement, $e \in q.E$ est une racine candidate pour q , s'il est éventuellement utile de créer un agrégat ayant pour racine e et imbriquant toutes les entités référencées dans q . Formellement, nous considérons qu'une entité $e \in q.E$ est une racine potentielle pour q , s'il existe un champs $e.f$ de e tel que f apparaît dans la clause **Group By** ou **Order By** de q ou bien dans un des prédictats filtre de la clause **Where** de q . En l'absence d'au moins une entité satisfaisant les conditions susmentionnées, chaque entité $e \in q.E$ est considérée comme une racine potentielle.

Agrégats pertinents pour requêtes individuelles Intuitivement, un agrégat a est pertinent pour une requête q , s'il est possible de répondre à q en utilisant (uniquement) a , i.e. si $q.E \subseteq a.E$. Nous procédons à l'identification des agrégats pertinents en deux étapes. À partir de chaque requête $q_i \in W$, nous dérivons tout d'abord un ensemble préliminaire A_i d'agrégats pertinents pour q_i . Chaque agrégat de A_i : (i) contient exactement toutes les entités référencées dans q_i (et rien d'autre); et (ii) a pour racine, une racine potentielle pour q_i . Pour chaque $a_j \in A_i$, nous vérifions ensuite sa pertinence pour les autres requêtes de W . Le résultat de cette étape est une matrice de pertinence $R(q_i, a_j)$, indiquant si un agrégat a_j est pertinent pour une requête q_i , i.e. $R(q_i, a_j) = 1$ si a_j est pertinent pour q_i , $R(q_i, a_j) = 0$ sinon.

Soit a_1 et a_2 , deux agrégats pertinents ayant la même racine : $a_1.r = a_2.r$. Si $a_2.E \subseteq a_1.E$, alors il est évident que a_2 est pertinent pour chaque requête pour laquelle a_1 est pertinent, i.e. si q_i est une requête à laquelle on peut répondre en utilisant a_1 , il est également possible de répondre à q_i en utilisant a_2 . La matrice de pertinence est mise à jour en conséquence :

$$\forall q_i, a_j, a_k \quad | \quad R(q_i, a_j) = 1 \quad \& \quad a_k.r = a_j.r \quad \& \quad a_k.E \subseteq a_j.E, \quad R(q_i, a_k) \leftarrow 1.$$

Exemple Soit les requêtes auto-descriptives ci-après, exprimées en SQL pour des raisons de lisibilité.
 $q_1 : "SELECT author.mail FROM author WHERE author.name=.."$

$q_2 : "SELECT author.id, count(*) FROM author, publication WHERE publication.year = 2023 ... GROUP BY author.id".$

$a_1 = (author, \{author\})$ est un agrégat pertinent pour q_1 . $a_2 = (author, \{author, publication\})$ et $a_3 = (publication, \{author, publication\})$ sont des agrégats pertinents pour q_2 . Comme $a_1.r = a_2.r$ et

$a_1.E \subseteq a_2.E$ nous pouvons conclure que a_2 est également pertinent pour q_1 . Conséquemment, $R(q_1, a_2)$ est mis à 1. Contrairement à a_2 , a_3 n'est pas pertinent pour q_1 , puisque $a_3.r \neq a_1.r$ (ce qui signifie que "*publication*" n'est pas une racine potentielle pour q_1). $R(q_1, a_3)$ ne change pas et reste égale à 0. La matrice de pertinence finale est donc comme suit.

| | q_1 | q_2 |
|-------|-------|-------|
| a_1 | 1 | 0 |
| a_2 | 1 | 1 |
| a_3 | 0 | 1 |

2.3.3 Mesure de l'intérêt des agrégats

L'objectif principal de cette étape est de définir une métrique $Int(a)$ qui rend compte de l'intérêt relatif d'un agrégat pertinent a . Une telle mesure est essentielle pour classer les agrégats pertinents et sélectionner le sous-ensemble complet et disjoint permettant d'optimiser W . Intuitivement, un agrégat est intéressant pour une charge de travail W , s'il accélère significativement une fraction des requêtes de W . La matrice de pertinence est utile pour indiquer quels agrégats sont pertinents pour quelles requêtes. Cependant, elle ne peut pas être utilisée en l'état pour déterminer l'intérêt des agrégats, car elle ne prend pas en compte ni l'importance relative des requêtes, ni le sur-coût à la lecture introduit par la réponse à une requête q avec un agrégat incorporant des entités inutiles à q .

Soit A l'ensemble de tous les agrégats pertinents, $A_i \subseteq A$ est l'ensemble des agrégats pertinents pour une requête q_i ($\forall a_j \in A_i, R(q_i, a_j) = 1$), $Opt(q_i)$ le coût optimal de q_i et $OptW(W)$ le coût total optimal de W . En s'inspirant des travaux de [CN07, JQD11] $Opt(q_i)$, le coût le plus bas possible pour répondre à q_i , se calcule comme suit [CN07, JQD11] : $Opt(q_i) = Min_{a_j \in A_i}(Cost(q_i, a_j))$. De

même, $OptW(W)$, le coût le plus bas possible de W , se calcule comme suit : $Opt(W) = \sum_1^{|W|} Opt(q_i)$. Finalement, nous définissons $Int(a)$, la *mesure d'intérêt* d'un agrégat a , comme la fraction des coûts des requêtes de W pour lesquelles a est pertinent. Formellement $Int(a) \in [0, 1]$ se définit comme suit

[CN07, JQD11] : $Int(a) = \frac{\sum_1^{|W|} R(q_i, a) \times \frac{Opt(q_i)^2}{Cost(q_i, a)}}{Opt(W)}$. Bien que simple à calculer, $Int(a)$ prend en compte la fraction de requêtes pour laquelle un agrégat est pertinent, l'importance relative des requêtes et le sur-coût à la lecture introduit par un agrégat. D'un autre côté, $Int(a)$ est normalisée par le coût total de la charge du travail pour le rendre comparable. Ceci peut également être utilisé s'il est nécessaire d'élaguer des agrégats en écartant ceux qui ont un intérêt en dessous d'un seuil prédéfini. Dans le cas extrême où un agrégat est optimal pour toutes les requêtes $Int(a) = 1 : \forall q_i \in W, Cost(q_i, a) = Opt(q_i)$, $\frac{Opt(q_i)^2}{Cost(q_i, a)} = Opt(q_i)$ et donc $Int(a) = 1$. Dans l'autre cas extrême où un agrégat n'est pertinent pour aucune requête ($\forall q_i \in W, R(q_i, a) = 0$) et $Int(a)$ prend la valeur nulle.

Exemple Considérons le cas où W serait composé de 2 requêtes q_1 et q_2 et supposons qu'il existe 4 agrégats pertinents a_1, a_2, a_3 et a_4 . Chaque cellule du tableau ci-après donne $\text{Cost}(q_i, a_j)$, i.e. le coût de la requête q_i si on y répond en utilisant l'agrégat a_j (avec "-", signifiant que a_j n'est pas pertinent pour $q_i : R(q_i; a_j) = 0$).

| | q_1 | q_2 |
|-------|-----------|-----------|
| a_1 | 10 | - |
| a_2 | - | 20 |
| a_3 | 15 | 25 |
| a_4 | 20 | 40 |

Comme illustré dans le tableau ci-dessus, a_1 est optimal pour q_1 et a_2 est optimal pour la requête q_2 plus coûteuse (ayant un plus grand coût). a_1 (resp. a_2) ne permet pas de répondre à la requête q_2 (resp. q_1). a_3 et a_4 permettent de répondre à la fois aux requêtes q_1 et q_2 , mais introduisent un sur-coût à la lecture.

Selon la formule susmentionnée, $\text{Int}(a_1) = 0,33$, $\text{Int}(a_2) = 0,67$, $\text{Int}(a_3) = 0,76$ et $\text{Int}(a_4) = 0,5$. Intuitivement, a_2 est plus intéressant que a_1 , parce qu'optimal pour une requête plus coûteuse que celle pour laquelle a_1 est optimal. De même, a_3 est plus intéressant que a_4 , car il permet de répondre à q_1 et à q_2 avec moins de surcoût à la lecture. a_3 est plus intéressant que a_2 , parce qu'il permet de répondre à q_1 sans introduire un "grand" sur-coût à la lecture pour q_2 .

2.3.4 Fusion d'agrégats pertinents

L'objectif de l'étape de fusion d'agrégats intéressants est de dériver un ensemble supplémentaire d'agrégats intéressants, qui bien que non optimaux pour aucune requête individuelle, pourraient être intéressants globalement, i.e. pour plusieurs requêtes de W .

Deux agrégats sont dits disjoints s'ils n'ont aucune entité en commun. La contrainte de disjonction que nous nous imposons indique que les agrégats non-disjoints ne peuvent pas être sélectionnés conjointement dans un même membre de Replica Set. Ceci est essentiel pour éviter la redondance des données et donc le risque d'incohérence. En raison de la contrainte de disjonction, une grande partie des agrégats pertinents dérivés des requêtes individuelles est rejetée et nous pouvons aboutir à des recommandations sous-optimales pour la charge de travail. L'intuition derrière la "fusion d'agrégats" est que la fusion de deux agrégats mutuellement exclusifs dans un agrégat sous optimal est parfois préférable à la conservation d'un seul et au rejet de l'autre.

Étant donné l'ensemble I d'agrégats intéressants dérivés des requêtes individuelles, l'objectif de la fusion d'une paire d'agrégats (a_i, a_j) , appelés *agrégats parents*, est de générer un nouvel agrégat a_k , appelé *agrégat fusionné*, qui, même s'il n'est pas optimal pour aucune requête individuelle est utile pour plusieurs requêtes (l'union des requêtes pour lesquelles a_i et a_j sont intéressants). Plus explicitement, L'agrégat a_k obtenu par la fusion d'une paire d'agrégats (a_i, a_j) doit respecter les deux contraintes ci-après : (i) a_k doit être pertinent pour toute requête pour laquelle un de ses parents est pertinent (a_k doit pouvoir être utilisé pour répondre à toute requête pour laquelle un de ses parents est pertinent) ; (ii) le coût de l'utilisation de a_k pour répondre à une requête q , ne devrait pas être "beaucoup plus élevé" que celui de l'utilisation de l'un de ses parents pour la même requête.

À l'état actuel de nos travaux, nous ne fusionnons que les paires d'agrégats (a_i, a_j) répondants aux deux critères suivants : (i) a_i et a_j doivent avoir la même racine (*i.e.* $a_i.r = a_j.r$) ; et (ii) a_i et a_j doivent encapsuler au plus une entité non commune. La première condition est nécessaire pour garantir que l'agrégat a_k issu de la fusion est pertinent pour toutes les requêtes pour lesquelles ses parents sont pertinents. La deuxième condition est nécessaire pour garantir que le coût de l'utilisation de a_k pour répondre à une requête ne soit pas beaucoup plus élevé que celui de l'utilisation de l'un de ses parents.

2.3.5 Sélection des agrégats : un problème de sac à dos multiple (MKP)

Étant donné l'ensemble I d'agrégats intéressants et un Replica Set RS composé de C membres ($C \ll |I|$), notre objectif est de sélectionner C sous-ensembles de I , tels que : (i) l'intérêt global des agrégats sélectionnés est un maximum; et (ii) les contraintes de disjonction et de restaurabilité sont respectées. Le problème de sélection d'agrégats est analogue à un problème de sac à dos multiple, dans lequel C sac à dos de capacités différentes sont remplis par des objets apportant chacun un certain profit et ayant un certain poids. L'objectif étant de maximiser le profit global sans dépasser la capacité de chacun des sacs.

Dans notre travail [Jou22] nous transposons le problème de sélection d'agrégats à un problème de sac à dos multiple en variables binaires (*0-1 knapsack*) et nous le résolvons avec l'approche de séparation-évaluation (*Branch & Bound*). Plus explicitement, étant donné C sac à dos (*i.e.* C membres d'un Replica Set) et N éléments (*i.e.* agrégats intéressants), nous cherchons à trouver les variables binaires x_{ij} , $i \in \{1..C\}$, $j \in \{1..N\}$, ayant la signification suivante : $x_{ij} = 1$ si l'agrégat j est assigné au membre i , et 0 sinon.

Soit B un bitmap de cardinalité $|\mathbb{E}|$ bits, tous mis à 1 et w_j le bitmap de la même cardinalité associé à un agrégat a (*i.e.* le $k^{\text{ème}}$ bit de w_j est mis à 1 si la $k^{\text{ème}}$ entité modélisée apparaît dans a et à 0 sinon). Formellement, le problème est défini comme suit :

$$\begin{aligned} & \text{maximize} \sum_{i=1}^C \sum_{j=1}^N Int(a_j)x_{ij} \quad \text{subject to,} \\ & x_{ij} \in \{0, 1\}, \quad i \in \{1, \dots, C\}, j \in \{1, \dots, N\} \quad (1) \end{aligned}$$

$$\sum_{i=1}^N x_{ij} = 1, \quad j \in \{1, \dots, N\} \quad (2)$$

$$\sum_{j=1}^N w_j x_{ij} \leq B \quad i \in \{1, \dots, C\} \quad (3)$$

La contrainte (1) est explicite. La contrainte (2) indique qu'un agrégat ne peut pas être assigné à plus d'un membre de l'ensemble de réPLICATION. La contrainte (3), que nous appelons *contrainte de disjonction*, remplace la contrainte de capacité maximale classique du problème du sac à dos. Elle

```

Input : items, knapsacks, level, nodeID, bestNode, maxProfit
Output: Id of the decision-tree node that corresponds to the optimal solution
1 if level < items.length() – 1 then
2   | B&B(items, knapsacks, level+1, nodeID + ".0", bestNode, maxProfit)
3 for i = 0; i < knapsacks.length(); i ++ do
4   | if Hamming_weight(knapsacks[i].bitmap  $\wedge$  items[level].bitmap) == 0 then
5     |   | knapsacks[i].profit += items[level].profit
6     |   | knapsacks[i].bitmap = knapsacks[i].bitmap  $\vee$  items[level].bitmap
7     |   | if level < items.length() – 1 then
8       |   |   | B&B (items, knapsacks, level+1, nodeID + "." + str(i+1), bestNode, maxProfit)
9     |   | else if knapsacks[i].profit > maxProfit then
10    |   |   | maxProfit = knapsacks[i].profit
11    |   |   | bestNode = nodeID

```

FIGURE 2.4 – Algorithme de sélection des agrégats

indique que les agrégats assignés à un même membre doivent être disjoints entre-eux et donc que l’union des bitmaps des agrégats assignés à un même membre doit être inférieure ou égale à B .

Le pseudo-code de notre algorithme est présenté dans la Figure 2.4. Dans cet algorithme, la construction de l’arbre de recherche se fait par la considération tour à tour de l’ensemble des éléments (le $i^{\text{ème}}$ niveau de l’arbre correspond au $i^{\text{ème}}$ agrégat). Chaque agrégat est d’abord assigné à un sac à dos fictif (lignes 1 – 2), impliquant son exclusion de la solution correspondant au chemin de l’arbre en cours d’exploration. L’agrégat est ensuite ajouté à chaque sac qui satisfait la contrainte (3) (lignes 3 – 8) : l’agrégat n’est ajouté à un sac que s’il n’imbrique aucune entité déjà ajoutée au sac par le biais d’un autre agrégat. Lorsque la contrainte (3) est respectée, B&B est appelé récursivement pour continuer à explorer le sous-arbre correspondant. Dans le cas contraire, le sous-arbre est abandonné. Ceci est illustré dans l’exemple de la Figure 2.5, où les agrégats 1 et 2 ne sont pas disjoints (*i.e.* la conjonction de leurs bitmaps n’est pas nulle). Les chemins qui assignent ces agrégats au même sac (non fictif) sont abandonnés (nœuds 1.1 et 2.2).

Lorsque tous les éléments d’un chemin ont été considérés, l’algorithme vérifie si le chemin en cours d’exploration permet un profit supérieur à celui des chemins déjà explorés et met à jour la solution optimale en conséquence (lignes 9–11).

Une caractéristique saillante de notre travail est que nous identifions les noeuds de l’arbre de recherche de manière à ce que la connaissance de l’identifiant d’un noeud implique celle de la solution qui lui est associée (Algorithme 2.4 et Figure 2.5). L’identification des noeuds est inspirée de l’identification des versions dans [KU95] où le $i^{\text{ème}}$ fils d’un noeud n est identifié par $n.i$. Par exemple, un noeud identifié par "2.0.1" dans notre arbre de recherche signifie que le premier élément a été ajouté au second sac, le troisième au premier sac et que le second élément a été exclu (ajouté au sac fictif).

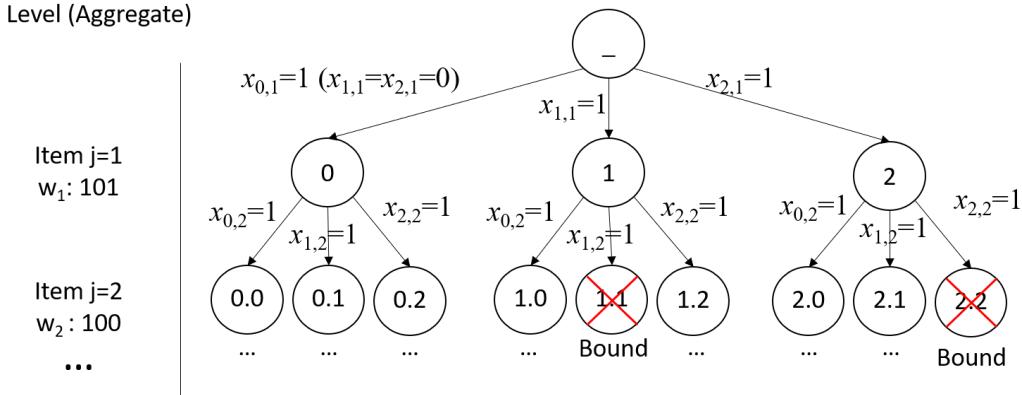


FIGURE 2.5 – Les agrégats 1 et 2 ne sont pas disjoints. Les chemins de l’arbre qui les sélectionnent ensemble sont abandonnés (*Branch & Bound*).

2.4 Résultats & discussion

Nous avons évalué et comparé les performances de BFS [KA15], CLDA [YLJ18] et les répliques déformées sur le banc de test standard TPC-H [tpc]. La base du TPC-H est composée de huit tables simulant l’activité d’une entreprise de livraison de produits [tpc]. La taille totale de la base dénormalisée (figure 2.6a) est de ≈ 114 Go et est d’environ ≈ 11 fois plus volumineuse que la base initiale du TPC-H.



(a) Schéma totalement dénormalisé de la base TPC-H. (b) Schéma obtenu avec l’approche BFS [KA15].

FIGURE 2.6 – Schéma dénormalisé vs. schéma obtenu avec l’approche BFS.

Nous avons considéré 10 requêtes du TPC-H : $q_1, q_3, q_5, q_6, q_{10}, q_{11}, q_{14}, q_{15}, q_{17}$ et q_{19} . Pour chaque requête, nous reportons le temps d’exécution moyen après trois exécutions successives (une exécution à froid et deux exécutions après l’initialisation du tampon de la base).

Comparaison de BFS [KA15], CLDA [YLJ18] et Distorted Replicas [Jou17, Jou22] La Figure 2.7 reporte les temps d’exécution réalisés par l’approche BFS [KA15], CLDA [YLJ18] et notre

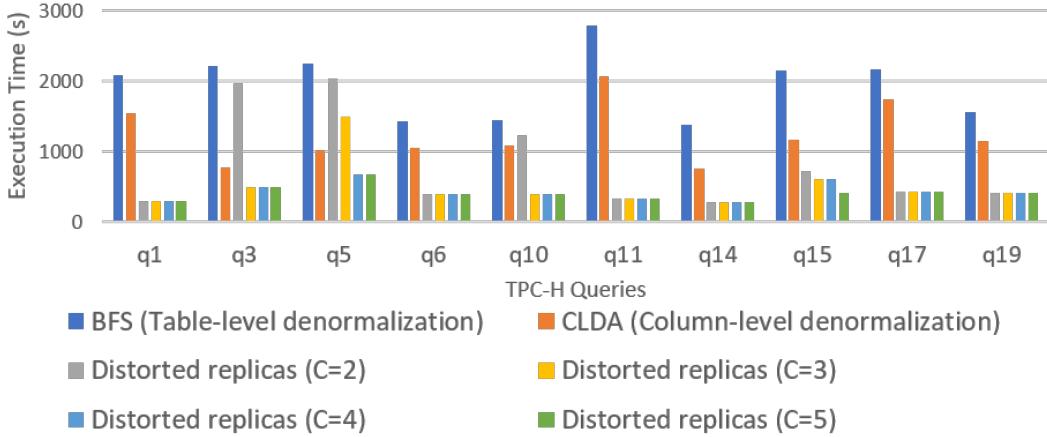
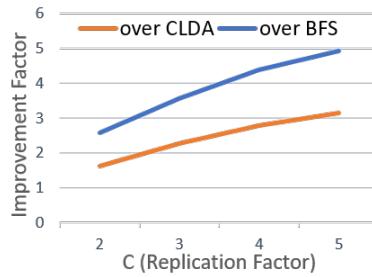

 FIGURE 2.7 – Temps d'exécution réalisés par BFS [KA15], CLDA [YLJ18] et *Distorted Replicas*.


FIGURE 2.8 – Amélioration du temps d'exécution global permise par Distorted Replicas, en comparaison avec BFS [KA15] et CLDA [YLJ18].

approche de répliques déformées [Jou17, Jou22] pour chacune des requêtes. La Figure 2.8 reporte l'accélération permise par notre approche sur l'ensemble des requêtes en fonction du facteur de réplication C . Comme on pouvait s'y attendre, CLDA et les répliques déformées surpassent largement BFS dans tous les cas de figure et permettent de réduire le temps d'exécution de la charge de travail de resp. ≈ 1.51 fois et jusqu'à ≈ 4.97 fois. Pour mieux comprendre le fonctionnement et les différences entre CLDA et les répliques déformées, considérons l'agrégat atomique $\{\text{Lineitem}, (\text{Lineitem})\}$ qui est optimal pour les requêtes q_1 et q_6 et contribue à la réduction du temps d'exécution de q_{15} and q_{17} . L'agrégat $\{\text{Lineitem}, (\text{Lineitem})\}$ est retenu par notre algorithme sélection d'agrégats $\forall C \geq 2$. Compte tenu des requêtes q_5 , q_{14} , et q_{19} , CLDA dénормalise partiellement les associations entre *Lineitem*, *Supplier*, *Part*, et *Nation* : i.e. certains attributs des entités *Supplier*, *Part* et *Nation* sont imbriquées dans *Lineitem*. CLDA imbrique en plus *Lineitem* dans *Order* pour mieux prendre en charge l'atomicité des transactions (i.e. *Order* imbrique un tableau contenant des sous-documents de type *Lineitem*). Répondre aux requêtes q_1 , q_6 , q_{15} et q_{17} avec l'agrégat construit par CLDA introduit un sur-coût à la

2.4. Résultats & discussion

| # de requêtes | # d'agrégats pertinents | # d'agrégats fusionnés | # d'agrégats retenus ($C=2$) | # d'agrégats retenus ($C=3$) | # d'agrégats retenus ($C=4$) | # d'agrégats retenus ($C=5$) |
|---------------|-------------------------|------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| 10 | 18 | 3 | 3 | 4 | 5 | 6 |

TABLE 2.1 – Algorithme d'identification et de sélection d'agrégats appliqué à la base TPC-H.

lecture important (du fait que l'agrégat imbrique des données inutiles à ces requêtes) et requiert en plus d'exécuter une opération *unwind* coûteuse, nécessaire pour "aplatir" le tableau de sous-documents *Lineitem* imbriqué dans *Order*. Ceci explique la raison pour laquelle les répliques déformées présentent de meilleures performances que CLDA pour ces requêtes.

La Figure 2.7 montre que lorsque le facteur de réPLICATION est petit, CLDA permet de meilleures performances que les répliques déformées pour les requêtes q_3 et q_5 . Ceci est dû à la contrainte de disjonction qui ne permet pas de sélectionner tous les agrégats optimaux. L'augmentation de C permet de sélectionner plus d'agrégats et de réduire le temps d'exécution de plus de requêtes.

| | $C=2$ | $C=3$ | $C=4$ | $C=5$ |
|-------------------------------------|----------------|-------------|-------------|---------------|
| # de noeuds visités | 47 600 | 2 252 212 | 84 579 458 | 2 584 233 662 |
| # de noeuds de l'arbre de recherche | 10 460 353 203 | 4.39805E+12 | 4.76837E+14 | 2.1937E+16 |

TABLE 2.2 – Nombre d'itérations pour la résolution du problème de sélection d'agrégats.

| $C=2$ | $C=3$ | $C=4$ | $C=5$ |
|--------|-------|-------|-------|
| 794 ms | 8 s | 302 s | 172 m |

TABLE 2.3 – Temps d'exécution de l'approche de regroupement des agrégats en fonction du facteur de réPLICATION ($N=21$ items).

Performances de l'algorithme de sélection d'agrégats Comme montré dans la Table 2.1, le nombre total d'agrégats à évaluer et à regrouper par notre algorithme de sélection d'agrégats s'élève à 21. La Table 2.2 donne le nombre d'itérations avec et sans la contrainte de disjonction en fonction du facteur de réPLICATION (*i.e.* en fonction du nombre de sac à dos). Avec la contrainte de disjonction la fraction de nœuds de l'arbre de recherche visités pour $C = 5$ est $\approx 1.17803\text{E}-07$.

La Table 2.3 reporte le temps nécessaire pour identifier les agrégats et les regrouper en des sous-ensembles disjoints. Ce temps est largement dominé par la résolution du problème du sac à dos. Pour 21 agrégats et 5 répliques, le temps nécessaire pour constituer les groupes est d'environ 172 minutes. Ce temps est acceptable du fait qu'il s'agisse d'un processus hors ligne et de l'accélération substantielle du temps d'exécution des requêtes que notre approche permet.

2.5 Conclusion

L'optimisation de la localité spatiale des données est un principe fondamental des bases de données. Appliquée aux bases NoSQL orientées agrégats, ce principe revient à regrouper dans une même unité autonome (l'agrégat), les données qui ont de grandes chances d'être manipulées ensemble. Ce principe est particulièrement crucial dans les environnements distribués à grande échelle où les transactions et les jointures *cross-nodes* sont prohibitives. Néanmoins, la définition des frontières d'un agrégat est un problème complexe, compte tenu de : (i) l'absence de règles formelles équivalentes à la normalisation relationnelle; (ii) l'explosion combinatoire des choix de modélisation possibles; et (iii) le fait qu'un choix de regroupement accélère certaines requêtes, mais en ralentit inévitablement d'autres.

Contrairement aux bases relationnelles où elle est considérée comme auxiliaire, la réplication est un élément central des bases NoSQL, et plus globalement, de tout système destiné au Big Data. Dans nos travaux de recherche nous proposons une approche de réplication innovante dite *Distorted Replicas* dans laquelle les données répliquées sont restructurées de différentes manières pour permettre à la base de prendre efficacement en charge des patterns d'accès antagonistes. Avec les *Distorted Replicas*, les répliques de la base sont physiquement différentes, mais logiquement identiques pour conserver la caractéristique fondamentale d'être constructibles les unes à partir des autres. En complément aux *Distorted Replicas*, nous avons proposé une approche guidée par les coûts pour la sélection des agrégats optimaux à inclure dans chaque réplique. Cette approche transpose le problème de sélection d'agrégats en un problème de sac-à-dos multiple (*Multiple Knapsack Problem*), où les répliques sont les sacs et les agrégats les éléments à regrouper dans les sacs de manière à en maximiser le profit (*i.e.* réduire le coût d'exécution des requêtes).

Comme mentionné précédemment, dans mes activités de recherche j'ai été amené à entamer une transition progressive de la gestion des données vers le traitement des données. Cette transition a été dictée à la fois par le déclin de la recherche sur les bases relationnelles, la difficulté de disposer de matériel spécialisé pour les bases NoSQL et surtout par toutes les opportunités de recherche offertes par l'analyse de données. Les travaux discutés dans le chapitre suivant de ce manuscrit sont la première étape de cette transition et ont trait à une autre caractéristique fondamentale des données massives, la variété.

Publications en lien avec le chapitre

Revues internationales

[Jou22] **Aggregates Selection in Replicated Document-Oriented Databases**

Khaled JOUNI

Journal of Information Science and Engineering (Abbréviation JCR : J INF SCI ENG). 2022.

ISSN : 1016-2364, DOI : [10.6688/JISE.20220338\(2\).0012](https://doi.org/10.6688/JISE.20220338(2).0012). 2023.

SJR best quartile : Q3, SJR : 0.21, JCR IF (2022) : 1.1

Conférences internationales

[ABC⁺¹⁸] **The Database Version Approach: Overview and Future directions**

Talel ABDESSALEM, Claudia MEDEIROS BAUZER, Wojciech CELLARY, Stéphane GANÇARSKI,
Khaled JOUINI, Maude MANOVRIER, Marta RUKOZ, Michel ZAM

34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA).
Romania. 2018.

[Jou17] **Distorted Replicas: Intelligent Replication Schemes to Boost I/O Throughput in NoSQL Systems**

Khaled JOUINI

14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA).
DOI : <https://doi.org/10.1109/AICCSA.2017.8218110>. 2017.

CORE Rank : C.

Encadrement de masters de recherche

- Nada BEN LATIFA (Soutenu)
- Islem OTHMANI (Soutenu)
- Haythem SAOUDI (Soutenu)

CHAPITRE 3

Variété : Extraction de caractéristiques pour la recherche d'information et
l'apprentissage automatique

Sommaire

| | | |
|------------|---|-----------|
| 3.1 | Introduction | 25 |
| 3.2 | Extraction de caractéristiques pour la reconnaissance d'empreintes | 25 |
| 3.2.1 | Reconnaissance d'empreintes : concepts clés et travaux antérieurs | 27 |
| 3.2.2 | EDT-C : Comparaison basée sur la triangulation de Delaunay étendue | 29 |
| 3.2.3 | Consolidation des modèles de référence | 31 |
| 3.2.4 | Résultats & discussion | 33 |
| 3.3 | Fusion de caractéristiques pour l'amélioration de la classification LULC | 35 |
| 3.3.1 | Apprentissage par transfert et classification LULC | 36 |
| 3.3.2 | Fusion de caractéristiques conçues manuellement et extraites automatiquement | 37 |
| 3.3.3 | Résultats & discussion | 41 |
| 3.4 | Conclusion | 43 |

3.1 Introduction

Si le chapitre précédent s'est concentré sur l'aspect quantitatif des données massives, ce chapitre aborde un défi non moins important, la *variété*. La dernière décennie a été marquée par une explosion de données de différents formats : journaux d'évènements, enregistrements audio/vidéo, images, textes, etc. Bien que ces types de données présentent des besoins spécifiques, ils partagent néanmoins un défi commun : transformer les données brutes en représentations abstraites et informatives, *i.e.* *l'extraction de caractéristiques* ou *Feature Extraction*.

Dans les travaux de recherche présentés dans ce chapitre, nous explorons les deux principaux cas d'usage de l'extraction de caractéristiques : la *recherche d'information* (*Information Retrieval*) et l'*apprentissage automatique*. La première partie est ainsi consacrée à nos recherches sur l'extraction de caractéristiques appliquée à la reconnaissance d'empreintes digitales. Cette reconnaissance est triviale lorsqu'elle s'effectue sur de petites bases d'empreintes, mais devient beaucoup plus complexe lorsque les images sont de faible qualité et/ou lorsque la recherche s'effectue dans des bases volumineuses, comme c'est le cas des bases construites à l'échelle d'un pays. Dans nos travaux, nous proposons EDT-C et FZC, deux approches complémentaires permettant une reconnaissance robuste des empreintes présentant une forte variabilité intra-classe et une faible variabilité inter-classe.

Le rôle crucial de l'extraction de caractéristiques en apprentissage automatique est illustré par nos travaux sur la classification d'images satellitaires, présentés dans la deuxième partie de ce chapitre. Les caractéristiques conçues manuellement, telles que celles générées par l'illustre SIFT (*Scale-Invariant Feature Transform* [Low04]), excellent dans la capture de détails locaux distinctifs, mais peinent à saisir le contexte global et les relations spatiales complexes. À contrario, les CNN sont capables d'apprendre des représentations hiérarchiques riches, mais peuvent ne pas être en mesure de capturer les détails à granularité fine. Dans nos travaux de recherche, nous proposons et évaluons des stratégies de fusion, précoce, tardive et à mi-niveau pour la mise en synergie des caractéristiques manuelles et de celles apprises par les CNN, dans le but d'améliorer la classification des images satellitaires.

La suite de ce chapitre, présente d'abord nos travaux sur l'extraction de caractéristiques appliquée à la reconnaissance d'empreintes, puis à la classification d'images satellitaires.

3.2 Extraction de caractéristiques pour la reconnaissance d'empreintes

Une empreinte digitale se caractérise par un ensemble de crêtes et de points singuliers, dits *minuties* correspondant à la terminaison ou à la bifurcation de ses crêtes (voir Figure 3.1). Une minutie m est communément décrite par un quadruplet $m = (x, y, \theta, t)$, où (x, y) correspondent à sa localisation spatiale, θ à l'angle formé par l'axe horizontal et la tangente à la ligne de crête au point de minutie et t à son type. Bien qu'il existe différentes approches de comparaison d'empreintes, celle basée sur les points

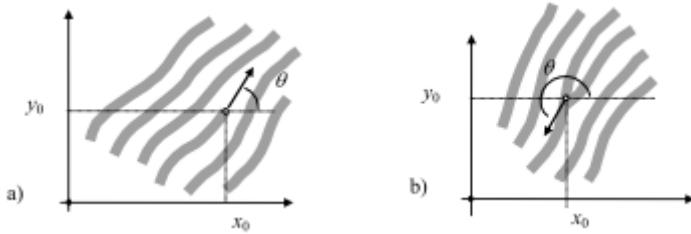


FIGURE 3.1 – Types de minuties : (a) Terminaison ; (b) Bifurcation [MMJP09].

de minuties demeure largement la plus répandue et la plus fiable [PGT⁺15]. Pour une même empreinte, il est extrêmement difficile que deux captures différentes aboutissent à une même représentation. Cette différence entre les représentations d'une même empreinte est communément appelée *variabilité intra-classe*. À contrario, la différence entre les représentations d'empreintes différentes est appelée *variabilité inter-classes*.

La variabilité inter-classes est grande pour les petites bases d'empreintes et faible dans les bases volumineuses, compte tenu de l'augmentation du risque de correspondances partielles fortuites. La variabilité intra-classe s'explique par des propriétés intrinsèques (*e.g.* élasticité de la peau) et *extrinsèques* (*e.g.* erreurs introduites lors de l'extraction de minuties) de l'empreinte. Dans le cas d'images d'empreintes de mauvaise qualité ou obtenues par encrage du doigt, les algorithmes de détection de minuties peuvent en effet introduire un grand nombre de fausses minuties (*a.k.a.* minuties parasites ou *spurious minutiae*) et ne pas être en mesure de détecter toutes les vraies minuties (*missing minutiae*). Les minuties déplacées, manquantes ou fausses expliquent la variabilité intra-classe et sont la principale source de faux rejets et d'erreurs de vérification [PGT⁺15].

Dans notre travail de recherche nous proposons deux approches complémentaires, intitulées EDT-C et FZC. EDT-C est un algorithme de comparaison d'empreintes qui se distingue par son utilisation d'une forme étendue de la triangulation de Delaunay (*i.e.* *Delaunay Triangulation*) [Del34]. FZC est une approche de fusion de différentes impressions d'une même empreinte qui vise à corriger la localisation spatiale des minuties déplacées, à restaurer celles manquantes et à supprimer celles parasites.

La suite de cette section est structurée en quatre sous-sections. La première introduit les concepts fondamentaux de la reconnaissance d'empreintes digitales et établit une taxonomie des approches antérieures à EDT-C et FZC. La deuxième (respectivement la troisième) sous-section détaille l'approche EDT-C (respectivement FZC). Enfin, la quatrième sous-section présente une étude expérimentale réalisée sur le banc d'essai FVC [MMC⁺04].

3.2.1 Reconnaissance d'empreintes : concepts clés et travaux antérieurs

3.2.1.1 Formulation du problème

Soit T (*Template*), l'ensemble de minuties d'une empreinte pré-stockée et I (*Input*) l'ensemble de minuties extraites d'une empreinte requête. La comparaison de T et I revient à superposer leurs minuties, *i.e.*, à trouver l'alignement ou la transformation géométrique (rotation et translation) $map_{(\delta_x, \delta_y, \delta_\theta)}$ qui maximise le nombre de minuties appariées. Le *score de similarité* se calcule alors par le ratio des minuties appariées par rapport au nombre total de minuties.

Les approches de comparaison d'empreintes peuvent être groupées en trois grandes familles : *comparaison globale*, *comparaison locale* et *comparaison locale avec consolidation*. Les approches procédant à un alignement global cherchent à apparier simultanément toutes les minuties de I et T . La recherche du bon alignement en une seule passe est fastidieux et consommateur de temps d'exécution [PGT+15]. Les approches procédant à des comparaisons locales, représentent une empreinte par un ensemble de structures constituées par des minuties mitoyennes. Les k plus proches voisins et les triangles formés par des triplets de minutie sont des exemples de structures locales. Les structures locales sont typiquement décrites par des mesures (angles, distances, etc.) invariantes aux transformations globales. La comparaison de structures locales ne requièrent pas de ce fait un pré-alignement et sont également moins enclines aux déformations causées par les distorsions que ne le sont les minuties prises à un niveau global. En dépit de ses avantages, la comparaison locale présente le défaut de la perte des relations spatiales globales entre les minuties. Pour bénéficier de la tolérance aux distorsions de la comparaison locale sans perdre la discrimination forte de la comparaison globale, plusieurs approches effectuent une comparaison locale, suivie par une étape de *consolidation*. La consolidation consiste à déduire de la comparaison des structures locales, la transformation permettant d'apparier le maximum de points de I et de T [MPGBRAR12].

3.2.1.2 Topologie des structures locales

Une structure locale est un sous-ensemble de minuties mitoyennes liées par une certaine relation spatiale. Dans la suite nous nous intéressons essentiellement aux *triplets* ou *triangles* de minuties. Les triangles ont deux grands avantages. Le premier est qu'ils permettent d'obtenir des descripteurs de longueur fixe. Le deuxième est que plusieurs caractéristiques tolérantes à la rotation et à la translation peuvent en être extraites. L'appariement de triangles de minuties issus de deux empreintes I et T a pour objectif de déterminer les paires de triangles $\{(\Delta^I, \Delta^T) \mid \Delta^I \in I \text{ et } \Delta^T \in T\}$ qui se "ressemblent" et qui ont donc de grandes chances d'être formés par des minuties qui correspondent les-unes aux-autres. Les principales approches de construction de triangles de minuties sont brièvement décrites ci-après.

Ensemble de tous les triangles possibles Les approches [GCC+97, CLLK03, BRAC08], construisent un triangle pour chaque combinaison de trois minuties. L'utilisation de tous les triangles possibles réduit le risque de rejet d'empreintes pertinentes, mais augmente le risque de correspondances fortuites

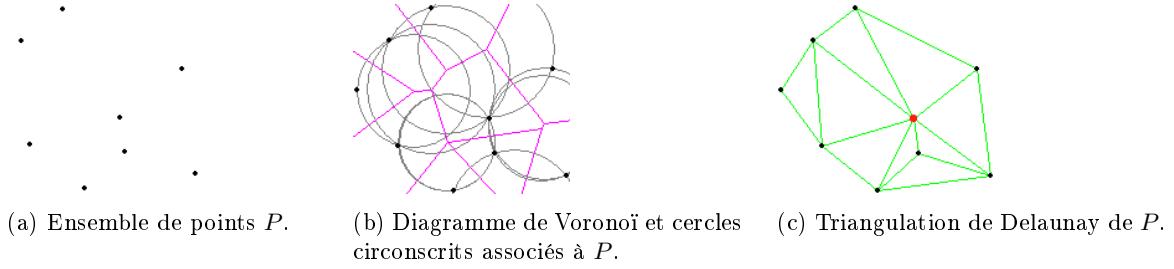


FIGURE 3.2 – Triangulation de Delaunay [Del34]

[UBEP07]. Cette méthode présente également l'inconvénient d'être plus lente que toutes les autres approches.

Plus proches voisins [MPGBGRAR12] construit pour chaque minutie p , tous les triangles possibles qu'elle peut former avec ses k plus proches voisines. L'inconvénient des structures locales basées sur les plus proches voisins est qu'elles sont considérablement affectées par les minuties manquantes ou fausses (*i.e.* parasites).

Triangulation de Delaunay Plusieurs approches de comparaison d'empreintes, telles que celles proposées dans [UBEP07] et [RM07], se basent sur la triangulation de Delaunay pour la construction des triangles de minuties. Une triangulation de Delaunay d'un ensemble de points P , notée $DT(P)$, est une triangulation telle qu'aucun point de P n'est à l'intérieur du cercle circonscrit d'un des triangles de $DT(P)$ (voir Figure 3.2c). $DT(P)$ permet de définir une structure topologique distinctive, mais se révèle peu robuste aux distorsions. Ceci est illustré dans les Figures 3.3b et 3.3c où le déplacement du point p modifie considérablement la topologie et la structure des triangles obtenus.

Combinaison des triangulations de Delaunay d'ordre 0 et d'ordre 1 Une triangulation de Delaunay d'ordre r d'un ensemble de points P , notée $DT_r(P)$, est une triangulation telle que, au plus, r points de P se trouvent à l'intérieur du cercle circonscrit de chaque triangle de $DT_r(P)$ [GHvK02]. Pour pallier au problème de déplacement de minuties, Liang & al. [LBA07] proposent de combiner les triangulations de Delaunay d'ordre 0 et d'ordre 1. Cette représentation, notée $LoD(P)$, permet dans le cas de déplacement de minuties de réduire le risque de rejet d'empreintes pertinentes. Cependant, l'absence de minuties réelles ou la présence de fausses minuties peut toujours modifier considérablement la structure des triangles construits.

3.2.2 EDT-C : Comparaison basée sur la triangulation de Delaunay étendue

3.2.2.1 Vue d'ensemble

Contrairement aux approches antérieures qui supposent que chaque minutie détectée p est authentique, EDT-C (*Extended Delaunay Triangulation based Comparison*) prend en plus en compte les cas où p pourrait être une minutie parasite, déplacée ou manquante (Figure 3.3). EDT-C s'appuie pour ce faire sur une variante de la triangulation de Delaunay, dite *triangulation de Delaunay étendue*, ou $EDT(P)$. Intuitivement, $EDT(P)$ contient, en plus des triangles de $DT(P)$, tous les triangles qui seraient obtenus si chacun des points de P était éliminé.

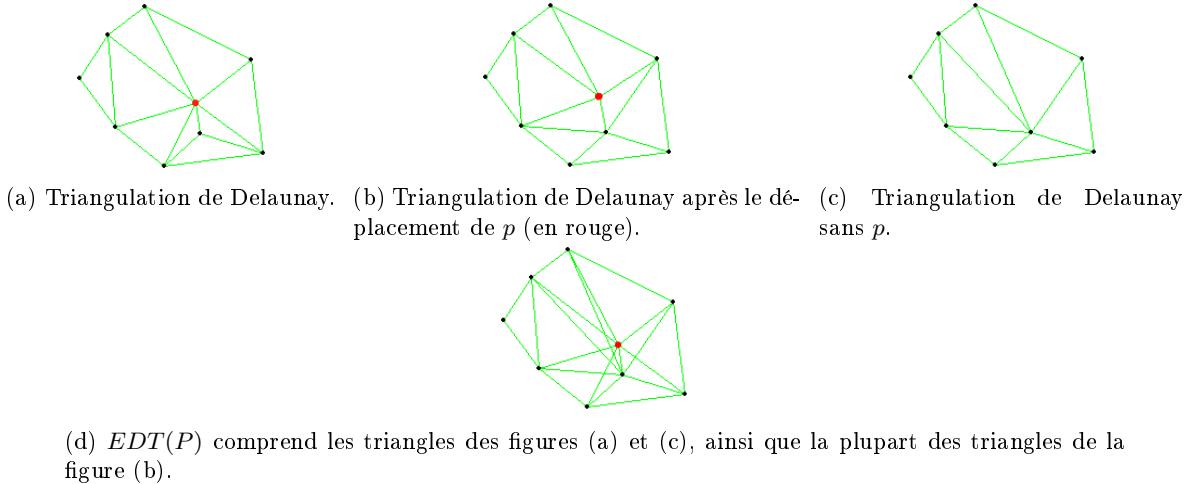


FIGURE 3.3 – Triangulation de Delaunay étendue $EDT(P)$.

Comme l'illustre la Figure 3.3.d, $EDT(P)$ permet de prendre en compte aussi bien les cas où p (en rouge sur les figures) est une minutie authentique (Figure 3.3.a) que les cas où p est déplacée (Figure 3.3.b), parasite ou manquante (Figure 3.3.c). Une fois les structures locales (triangles de minuties) déterminées, EDT-C suit les étapes classiques des algorithmes récents de comparaison d'empreintes : (i) caractérisation des structures locales, (ii) appariement des structures locales et (iii) consolidation et appariement global des minuties. EDT-C introduit cependant des améliorations à chacune de ces étapes, comme détaillé dans la suite.

3.2.2.2 Descripteurs associés aux triangles

Mesures géométriques Plusieurs mesures géométriques peuvent être extraites d'un triangle. Contrairement à la plupart des approches antérieures, EDT-C, ne décrit pas un triangle de minutie par les longueurs de ses côtés, mais par un ensemble de mesures qui en dérivent. Ces mesures ont été dé-

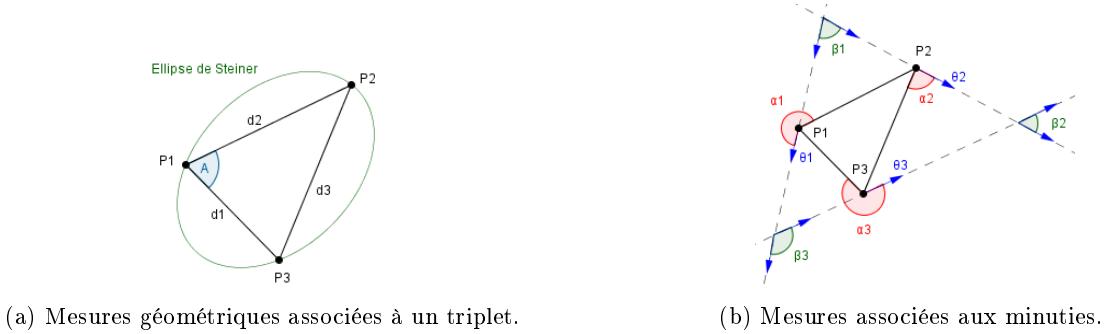


FIGURE 3.4 – Descripteur utilisé dans EDT-C.

terminées suite à l'étude expérimentale que nous avons menée sur les bases FVC [MMC⁺04]. Dans notre approche nous décrivons la forme d'un triangle de minuties Δ par : l'elongation de l'ellipse de *Steiner* circonscrite $\epsilon(\Delta)$, le cosinus du plus grand angle du triangle $\cos A(\Delta)$ et le périmètre $p(\Delta)$ (Figure 3.4.a). Intuitivement, l'elongation de l'ellipse de *Steiner* peut être interprétée comme le "degré d'étirement" d'un triangle (par rapport à un triangle équilatéral) [Far12].

Caractéristiques liées à l'empreinte Les mesures géométriques extraites d'un triangle se basent sur la disposition des minuties, mais ne prennent pas en compte leurs angles d'orientation.

Pour prendre en compte les angles d'orientation EDT-C inclut les mesures suivantes (Figure 3.4.b) :

- $\alpha_{i \in [1,3]}$: l'angle de rotation nécessaire pour superposer le vecteur d'orientation de la minutie p_i et le côté (p_i, p_{i+1}) du triangle.
- $\beta_{i \in [1,3]}$: l'angle de rotation nécessaire pour superposer le vecteur d'orientation de la minutie p_i avec celui de p_{i+1} .

3.2.2.3 Appariement local et consolidation

L'appariement local entre deux empreintes à comparer I et T se déroule en deux étapes : appariement des triangles de minuties, suivi de l'appariement des minuties elles-mêmes.

Appariement des triangles de minuties L'objectif de l'appariement des triangles est de construire une liste A contenant les paires de triangles de I et T qui se "ressemblent", sur la base des descripteurs qui leur sont associés, et selon une fonction de similarité locale SL ainsi qu'un seuil Th_{SL} :

$A = \{(\Delta^I, \Delta^T) \mid \Delta^I \in I, \Delta^T \in T \text{ et } SL(\Delta^I, \Delta^T) > Th_{SL}\}$. Si la similarité entre deux triangles Δ^I et Δ^T dépasse le seuil Th_{SL} , la paire (Δ^I, Δ^T) est ajoutée à la liste A .

Appariement local des minuties L'objectif de cette étape est de construire une liste M contenant les paires de minuties de I et T ayant une forte probabilité de correspondre. La liste M est générée à partir de A comme suit.

1. Pour chaque paire de triangles $(\Delta^I, \Delta^T) \in A$, on considère les triplets de minuties qui les forment : $(p_i, p_j, p_k) \in I$ et $(q_x, q_y, q_z) \in T$.
2. Pour chaque paire de minuties (p, q) , si p et q sont du même type (bifurcation ou terminaison) et que la paire (p, q) n'apparaît pas dans M , alors (p, q) est ajoutée à M avec une occurrence initiale égale à 1.
3. Si la paire (p, q) apparaît déjà dans M , son nombre d'occurrences est incrémenté de 1.

Le nombre d'occurrences associé à une paire (p, q) correspond au nombre de triangles similaires dans lesquels cette paire apparaît. Plus ce nombre est élevé, plus p a de chances de correspondre à q .

Appariement global et calcul de la similarité La consolidation constitue l'étape finale, où l'on cherche à déterminer la transformation (rotation et translation) permettant d'aligner au mieux I et T , ainsi qu'à calculer un score de similarité. Dans notre approche, nous considérons les r paires de M ayant les plus grands nombres d'occurrences. Les r transformations correspondantes sont testées, et celle permettant d'aligner le plus grand nombre de paires de M est retenue. Une fois la meilleure transformation identifiée, le score de similarité se calcule selon la formule [MMJP09] : $Score(I, T) = \frac{k}{(|I|+|T|)/2}$, où $|I|$ et $|T|$ désignent respectivement le nombre de minuties extraites de I et de T , et k représente le nombre de minuties qui se superposent après l'application de la transformation $(\delta_x, \delta_y, \delta_\theta)$.

3.2.3 Consolidation des modèles de référence

La qualité des modèles de référence est un facteur déterminant pour la fiabilité des systèmes de reconnaissance d'empreintes. Une méthode courante pour améliorer ces modèles consiste à les consolider en fusionnant plusieurs impressions d'une même empreinte. Dans nos travaux, nous avons proposé une approche de consolidation appelée Fusion de Zones Compatibles (FZC). FZC se distingue des approches antérieures en trois points : la détermination des zones concordantes, l'alignement localisé des impressions et l'estimation de l'authenticité des minuties.

Dans la suite, le modèle de référence et le modèle en entrée sont désignés respectivement par Ref et I . Soit $M = \{(p_i^R, q_j^I) \mid p_i^R \in Ref, q_j^I \in I\}$ l'ensemble des paires de minuties concordantes de Ref et I (déterminées par notre algorithme EDT-C). La première étape de FZC consiste à identifier les *zones compatibles*, *i.e.*, les zones concordantes de I et de Ref . Ensuite, pour chaque paire de zones compatibles, FZC effectue un alignement local, plus fin que l'alignement global, pour améliorer la zone de Ref par les minuties de la zone correspondante de I .

Pour détecter les zones compatibles, nous avons introduit l'algorithme 1, inspiré de l'algorithme K-means. Cet algorithme regroupe progressivement les minuties en clusters contenant au moins Th_n minuties situées à une distance inférieure à Th_c d'un centre ajustable. Les figures 3.5.a, 3.5.b et 3.5.c illustrent l'application de cet algorithme sur deux impressions réelles d'une même empreinte. Les ellipses de la figure 3.5.c montrent les zones compatibles obtenues.

Algorithm 1: Partitionnement des impressions en régions locales compatibles.

```

1 Input : Ensemble  $M$  de paires de minuties de  $Ref$  et  $I$ 
2 Output : Ensemble de régions compatibles  $Z$ 
3 foreach Paire  $(p_i, q_j)$  dans  $M$  do
4   if  $(p_i, q_j)$  n'a pas été traitée then
5     Créer une nouvelle région  $Z_k$  ;
6      $center_k^R = p_i$  ;  $center_k^I = q_j$  ;
7     Ajouter  $(p_i, q_j)$  à  $M_k$  ;
8     Marquer  $(p_i, q_j)$  comme traitée ;
9     repeat
10     $(p'_i, q'_j)$  : autre paire non traitée dans  $M$  ;
11    if  $distance(center_k^R, p'_i) < Th_c$  AND  $distance(center_k^I, q'_j) < Th_c$  then
12      Ajuster les centres ;
13      Ajouter  $(p'_i, q'_j)$  à  $M_k$  ;
14      Marquer  $(p'_i, q'_j)$  comme traitée ;
15    end
16    until Taille de  $M_k < Th_n$  ;
17    Ajouter  $Z_k$  à  $Z$  ;
18  end
19 end

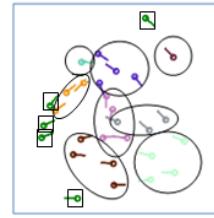
```



(a) Minuties extraits de l'impression 12_6 de la base DB1_A.



(b) Minuties extraits de l'impression 12_3 de la même empreinte.



(c) Zones compatibles et minuties sans correspondances.

FIGURE 3.5 – Détermination des zones compatibles [GJK17b].

Contrairement aux approches antérieures, FZC n'applique pas un alignement global unique entre Ref et I , mais un alignement local pour chaque paire de zones compatibles, dans le but de mieux gérer les distorsions spécifiques à chaque région. La transformation locale est calculée à partir de la minutie $p \in z_{Ref}$ ayant la plus haute fiabilité Q et de sa minutie correspondante $q \in z_I$. Une fois la transformation locale déterminée, les opérations suivantes sont effectuées : (i) *Remplacement* : Si $q \in I$ a une meilleure qualité que la minutie correspondante $p \in Ref$, ses attributs remplacent ceux de p dans Ref ; (ii) *Ajout* : Si q n'a pas de correspondant dans Ref et que sa qualité dépasse Th_Q , elle est ajoutée au modèle ; (iii) *Suppression* : Si p n'a pas de correspondant dans I et que sa qualité est inférieure à Th_Q , elle est supprimée. La majorité des approches antérieures à FZC se basent sur la fréquence d'une minutie pour juger de son authenticité. La fréquence d'une minutie peut se révéler insuffisante lorsqu'il n'est pas possible de collecter un nombre significatif d'impressions de bonne qualité. Dans

FZC, l'authenticité d'une minutie est estimée par la qualité de la capture de la zone dans laquelle se trouve la minutie : si la zone est affectée par une forte distorsion ou bien se trouve dans une région obscurcie de l'image, la minutie est jugée peu fiable.

3.2.4 Résultats & discussion

Les compétitions FVC (*Fingerprint Verification Competitions*) [MMC⁺02, MMC⁺04, CFFM07] sont des bancs de test, standards de fait, pour la comparaison des algorithmes de reconnaissance d'empreintes. FVC fournit quatre bases (DB1, DB2, DB3 et DB4), bien représentatives des variabilités intra-classe et inter-classe. Nous avons conduit trois séries de tests sur les différents bancs FVC pour valider nos approches. Par souci de concision, seuls quelques résultats sont présentés dans la suite.

La qualité de tout algorithme de comparaison d'empreintes dépend en premier lieu de son taux de faux positifs (*FMR* pour *False Match Rate*) et de son taux de faux négatifs (*FNMR* pour *False Non-Match Rate*). D'autres critères de qualité sont classiquement dérivés du FMR et du FNMR :

- Le taux *EER* (*Equal-Error Rate*), correspondant au seuil de décision pour lequel $FMR = FNMR$.
- Le taux *ZeroFMR*, qui correspond au plus bas FNMR pour lequel aucune fausse acceptation n'a lieu. *FMR100* (resp. *FMR1000*) est une variante de ce taux et correspond au plus bas FNMR lorsque le FMR est égal à 0,1 % (resp. 0,01 %).
- Le taux *ZeroFNMR*, qui correspond au plus bas FMR pour lequel aucun faux négatif n'a lieu.

Pour montrer l'apport de la triangulation de Delaunay étendue (EDT), nous l'avons remplacé dans notre approche par, tour à tour, les 3 plus proches voisins (NN), la triangulation de Delaunay classique (DT), et la combinaison des triangulations de Delaunay de premier et second ordre (LoD). Comme le montre la Table 3.1, la triangulation de Delaunay étendue consomme un temps de calcul légèrement supérieur aux autres topologies, mais permet également une précision nettement supérieure. La Table 3.2 compare notre approche de consolidation FZC à l'approche Hierarchical Matching [UBEP09]. Tel que le montre la Table 3.2, FZC présente une nette supériorité surtout lorsque le nombre d'empreintes fusionnées est relativement petit. La Table 3.3 compare notre approche EDT-C avec les algorithmes PN [PN04] et M3gl [MPGBGRAR12]. Ces deux algorithmes sont ceux qui se rapprochent le plus d'EDT-C. Comme le montre la Table 3.3, M3gl et EDT-C surpassent nettement PN. En général, M3gl est légèrement plus rapide que notre approche. Cependant, notre algorithme s'avère bien plus précis

| Triangulation | EER% | FMR100% | FMR1000% | zeroFMR% | Temps(ms) |
|---------------|--------------|--------------|--------------|--------------|--------------|
| NN | 2,500 | 3,179 | 4,893 | 6,750 | 4,059 |
| DT | 2,358 | 3,143 | 4,679 | 5,321 | 1,195 |
| LoD | 2,256 | 2,821 | 4,107 | 6,500 | 1,703 |
| EDT | 1,572 | 1,893 | 2,893 | 3,357 | 4,378 |

TABLE 3.1 – Performances obtenues avec différentes topologies sur la base DB1_A du FVC2000.

3.2. Extraction de caractéristiques pour la reconnaissance d'empreintes

| | | # d'échantillons fusionnés | 2 | 3 | 4 |
|--------------------------------|----------|----------------------------|-------|-------|---|
| | | | 2 | 3 | 4 |
| Hierarchical Matching [UBEP09] | FMR100% | 2,810 | 1,750 | 1,230 | |
| | FMR1000% | 4,680 | 3,100 | 2,030 | |
| | ZeroFMR% | 8,710 | 5,230 | 3,900 | |
| FZC | FMR100% | 0,833 | 0,800 | 1,000 | |
| | FMR1000% | 1,333 | 1,200 | 1,500 | |
| | ZeroFMR% | 2,167 | 1,400 | 4,250 | |

TABLE 3.2 – Précision (en %) sur les empreintes de la base DB1_A

sur l'ensemble des bases. La figure 3.6 illustre un cas réel d'empreintes de mauvaise qualité où notre algorithme détecte davantage de minuties concordantes que M3gl.

| Base | Algorithme | EER% | FMR100% | FMR1000% | ZeroFMR% | Temps (ms) |
|-------|------------|--------------|--------------|--------------|---------------|--------------|
| DB1_A | PN | 3,657 | 4,679 | 7,071 | 12,750 | 50,662 |
| | M3gl | 2,406 | 3,071 | 5,286 | 6,964 | 3,296 |
| | EDT-C | 1,572 | 1,893 | 2,893 | 3,357 | 4,378 |
| DB2_A | PN | 2,272 | 2,750 | 3,750 | 4,286 | 61,088 |
| | M3gl | 1,716 | 1,893 | 3,036 | 4,643 | 3,400 |
| | EDT-C | 1,122 | 1,250 | 1,964 | 2,500 | 4,341 |
| DB3_A | PN | 5,944 | 8,357 | 12,500 | 16,107 | 182,105 |
| | M3gl | 5,726 | 8,929 | 12,536 | 13,893 | 6,201 |
| | EDT-C | 4,228 | 6,036 | 7,750 | 13,464 | 8,656 |
| DB4_A | PN | 5,051 | 5,857 | 7,250 | 7,964 | 15,730 |
| | M3gl | 2,498 | 3,286 | 5,571 | 8,179 | 1,706 |
| | EDT-C | 2,074 | 2,714 | 4,893 | 9,250 | 1,403 |

TABLE 3.3 – Précision (en %) et temps de calcul obtenus sur les bases du FVC.



FIGURE 3.6 – Correspondances trouvées entre deux impressions d'une même empreinte par M3gl [MPGBRAR12] et EDT-C.

3.3 Fusion de caractéristiques pour l'amélioration de la classification LULC

Après avoir étudié l'extraction de caractéristiques dans le contexte de la recherche d'information, cette section explore son application à l'apprentissage automatique et la classification des usages du sol (*Land Use and Land Cover*, LULC).

L'imagerie satellitaire offre une perspective unique et exhaustive de la surface terrestre, cruciale pour des applications critiques comme la surveillance environnementale et la planification urbaine [AJTK24b]. Au cœur de ces applications se trouve la cartographie des usages du sol, qui consiste à attribuer des classes sémantiques aux images satellites. Les premières approches de classification LULC reposaient sur des descripteurs conçus manuellement, comme SIFT (*Scale-Invariant Feature Transform*) [Low04]. SIFT et les méthodes similaires excellent dans la capture des caractéristiques locales distinctives. Cependant, leur nature locale limite leur capacité à appréhender le contexte global et les relations spatiales complexes des images satellites.

L'avènement des réseaux convolutifs (CNN) profonds a transformé la classification d'images, surpassant les méthodes classiques dans diverses compétitions. La grande précision de ces modèles, couplée à leur voracité en données étiquetées, a orienté la majorité des approches existantes de classification LULC vers l'apprentissage par transfert. Ce type d'apprentissage consiste à ajuster des modèles profonds pré-appris sur de grands jeux de données généralistes, sur des jeux spécialisés et plus restreints. Les CNN apprennent des caractéristiques hiérarchiques : des éléments basiques dans les premières couches aux structures complexes dans les couches profondes. Cet apprentissage hiérarchique permet aux CNN d'exceller dans la capture des contextes globaux. En dépit de ces capacités, les CNN peuvent avoir du mal à préserver des détails très fins, comme les petites textures ou les variations subtiles de couleur, en raison des opérations de *pooling* qui suivent les couches convolutionnelles.

Dans nos travaux, nous explorons et quantifions les avantages de la fusion des descripteurs conçus manuellement (SIFT) avec les caractéristiques apprises automatiquement via les CNN. Cette fusion a pour but de tirer parti à la fois de la capacité de SIFT à capturer des détails locaux fins et de la capacité des CNN à saisir le contexte global et les relations spatiales complexes. En particulier, nous explorons trois stratégies de fusion : fusion précoce (*Early Fusion*), fusion tardive (*Late Fusion*) et fusion intermédiaire (*Mid-Level Fusion*). Nos résultats expérimentaux révèlent que ces approches varient en efficacité. Ils suggèrent également qu'au-delà de l'approche largement répandue de l'apprentissage par transfert, des alternatives telles que la fusion de caractéristiques méritent d'être explorées.

La suite de cette section est organisée comme suit : la sous-section 3.3.1 donne une brève revue des travaux antérieurs, la sous-section 3.3.2 décrit nos stratégies de fusion, et la sous-section 3.3.3 en évalue expérimentalement les performances.

3.3.1 Apprentissage par transfert et classification LULC

EuroSAT [HBDB18], que nous utilisons dans notre étude expérimentale, est l'un des jeux de données les plus utilisés pour la validation des approches de classification LULC. Ce jeu contient 27 000 images réparties sur dix classes représentant un éventail diversifié d'occupations des sols (figure 3.7). Par souci de concision, la suite se concentre essentiellement sur les approches utilisant EuroSAT. La majorité de ces approches repose sur l'apprentissage par transfert.



FIGURE 3.7 – Images extraites du jeu de données EuroSAT [Lom98]

Dans [DA21], les auteurs utilisent différents modèles pré-entraînés et affinés sur EuroSAT pour l'extraction de caractéristiques (VGG19, ResNet50 et InceptionV3). Les caractéristiques extraites par ces modèles sont recalibrées et fournies comme entrée à un classifieur TSVM (Twin WSVM), qui atteint une précision de 94,39 %. Dans [HBDB19], diverses architectures CNN ont été comparées, notamment un CNN peu profond (*shallow CNN*), un modèle basé sur ResNet50 et un modèle basé sur GoogleNet. Les précisions obtenues sur EuroSAT sont respectivement de 89,03 %, 98,57 % et 98,18 %. [NPZH20] a exploré l'affinement des modèles pré-entraînés, d'abord sur des jeux d'images satellitaires (*In Domain Fine Tuning*), puis sur le jeu EuroSAT. [NPZH20] a démontré que les modèles affinés sur des jeux de données spécialisés surpassent de manière significative ceux affinés uniquement sur EuroSAT. Le modèle ResNet50v2 affiné sur des jeux de données spécialisés a notamment atteint une précision de 99,2 % sur EuroSAT.

Alors que l'apprentissage par transfert consiste à adapter un modèle pré-entraîné pour améliorer les performances sur un jeu de données spécialisé, le *transfert de connaissances* implique d'entraîner un seul modèle sur plusieurs tâches simultanément, en tirant parti des représentations et des connaissances partagées entre ces tâches. [GD22] a utilisé le transfert de connaissances en proposant un apprentissage

multitâche dans lequel le modèle apprend à partir de divers jeux d'images satellitaires simultanément. Le "réseau multitâche mutant" (μ 2Net), introduit par [GD22], atteint des performances remarquables sur EuroSAT avec une précision de 99,2 %. Le transfert de connaissances est également utilisé par [WZX⁺24], où les *Vision Transformers* (ViT) [SKZ⁺21] avec le mécanisme d'attention *Rotatable Variance Scaled Attention* (RVSA) sont exploités dans le cadre d'un *Multi-Task Pretraining* (MTP). Évalué sur EuroSAT, le modèle MTP atteint également une précision de 99,2 %.

L'étude [TZJ18] se distingue des approches existantes par la combinaison des caractéristiques conçues manuellement et de celles extraites via un CNN. [TZJ18] démontre les avantages de cette combinaison sur le jeu de données EuroSAT. Cependant, [TZJ18] n'exploré qu'une approche de fusion précoce simple, où les différentes caractéristiques sont combinées à l'entrée du modèle. Notre travail propose, en plus de la fusion précoce, des méthodes de fusion plus élaborées basées sur les mécanismes d'attention, permettant, comme nous allons le voir, d'obtenir de meilleurs résultats.

3.3.2 Fusion de caractéristiques conçues manuellement et extraites automatiquement

Dans nos travaux, nous proposons trois stratégies distinctes de fusion : la fusion précoce (simple) et des approches plus élaborées de fusion tardive et à mi-niveau, basées sur des mécanismes d'attention. Sommairement, la fusion précoce combine les différentes modalités des caractéristiques dès le début du processus d'apprentissage (fusion à l'entrée). À l'inverse, la fusion tardive utilise un mécanisme d'attention par modalité pour sélectionner les caractéristiques les plus pertinentes avant de les fusionner au dernier stade du processus d'apprentissage (fusion au niveau de la décision). La fusion à mi-niveau adopte une approche plus nuancée, en permettant aux caractéristiques d'interagir à un stade intermédiaire du réseau, favorisant ainsi un échange d'informations plus riche. La suite de cette section détaille chacune de ces stratégies de fusion que nous proposons.

3.3.2.1 Modèles de référence et fusion précoce

SIFT [Low04] est un algorithme de détection et de description de points d'intérêt (*keypoints*) dans les images. Dans le contexte des images satellitaires, les points d'intérêt correspondent souvent à des transitions entre différentes couvertures du sol. Le descripteur SIFT associé à un point d'intérêt encode les gradients ou les changements directionnels d'intensité dans son voisinage [Low04].

Les Figures 3.8 et 3.9 illustrent les modèles de référence utilisés dans notre travail. Le premier modèle est un réseau neuronal classique prenant des descripteurs SIFT en entrée et dont l'architecture alterne des couches denses et des mécanismes de régularisation. Les couches denses contiennent des activations ReLU pour introduire de la non-linéarité. La normalisation par lot (*Batch Normalization*) est utilisée pour stabiliser l'apprentissage. Enfin, la dernière couche *Softmax* est utilisée pour produire les probabilités d'appartenance à chaque classe. Le second modèle de référence est un CNN prenant en entrée des images RGB. Ce modèle adopte une architecture standard, comprenant des couches convo-

lutionnelles pour l'extraction des caractéristiques, des couches de pooling pour la réduction d'échelle, une couche d'aplatissement pour transformer les caractéristiques en vecteur, et des couches entièrement connectées pour la classification.

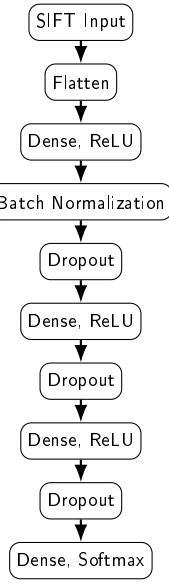


FIGURE 3.8 – Baseline
SIFT-NN

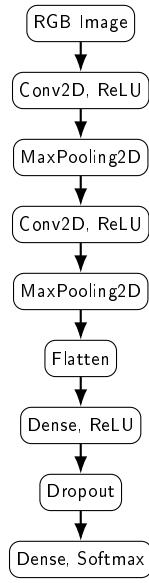


FIGURE 3.9 – Baseline
Shallow CNN

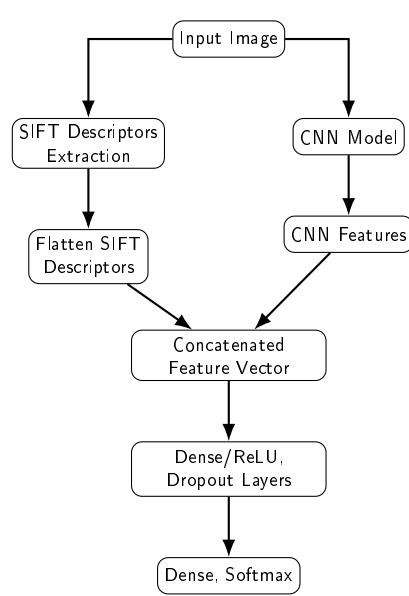


FIGURE 3.10 – Fusion précoce

La Figure 3.10 illustre le modèle de fusion précoce utilisé dans notre travail. Ce modèle comprend deux branches, traitant chacune une modalité différente. Après l'extraction des caractéristiques dans chaque branche, le vecteur de caractéristiques concaténé est transmis à des couches neuronales standards. Malgré sa simplicité, la fusion précoce permet d'obtenir de bons résultats en exploitant les informations provenant à la fois des images RGB et des descripteurs SIFT.

3.3.2.2 Fusion tardive : *Attention-Enhanced Dual Learning* (ADL)

Les *mécanismes d'attention*, inspirés de l'attention visuelle humaine, permettent aux réseaux neuronaux de se focaliser de manière dynamique sur les parties les plus pertinentes des données d'entrée. Le modèle de fusion tardive que nous proposons (Figure 3.11), repose sur une stratégie d'apprentissage dual intégrant des mécanismes d'attention. L'image d'entrée est traitée séparément par deux branches, chacune dédiée à une modalité spécifique. Dans la branche RGB, l'image passe par un CNN intégrant des blocs *Squeeze-and-Excitation* (SE), appliquent une attention par canal (*Channel-Wise Attention*), et donnent ainsi plus de poids aux canaux les plus pertinents. Parallèlement, dans la deuxième branche, les descripteurs SIFT sont extraits de l'image et traités par un réseau neuronal intégrant un mécanisme d'attention sur les caractéristiques (*Feature-Wise Attention*), qui attribue des poids aux descripteurs

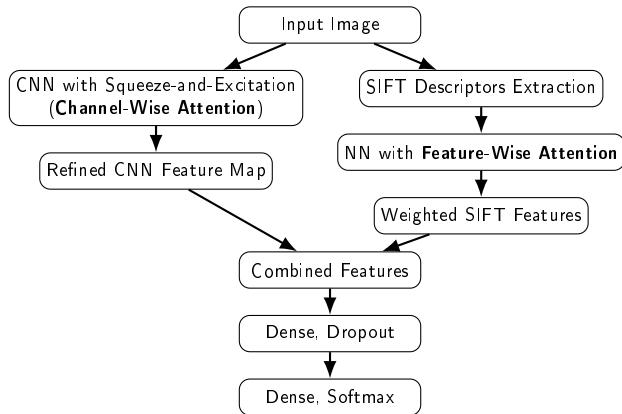


FIGURE 3.11 – Fusion tardive : Attention-Enhanced Dual Learning (ADL)

SIFT selon leur pouvoir informatif. Après le traitement indépendant dans chaque branche, les caractéristiques CNN affinées et les descripteurs SIFT pondérés sont fusionnés en un vecteur de caractéristiques unifié, transmis aux couches subséquentes pour la classification. En somme, la fusion tardive vise à retenir les caractéristiques SIFT et CNN les plus informatives, avant la fusion.

3.3.2.3 Fusion à mi-niveau : *Fusion of Local Attended CNN Features and Global CNN Features (LFGF)*

Les approches de fusion précoce et tardive considèrent les caractéristiques extraites d'un CNN et celles extraites avec SIFT comme étant d'égale importance. Dans nos travaux, nous proposons une nouvelle approche de fusion à mi-niveau visant à tirer parti des avantages combinés de SIFT et des CNN, en adoptant une logique de fusion différente. Le principe de cette approche est d'orienter l'attention du CNN vers les régions les plus informatives de l'image, en s'appuyant sur les points d'intérêt SIFT, sans pour autant exclure les caractéristiques obtenues sans l'application de l'attention basée sur SIFT. La Figure 3.12, issue de nos expériences, illustre la manière dont les points d'intérêt SIFT guident l'attention du réseau vers des zones spécifiques de l'image.

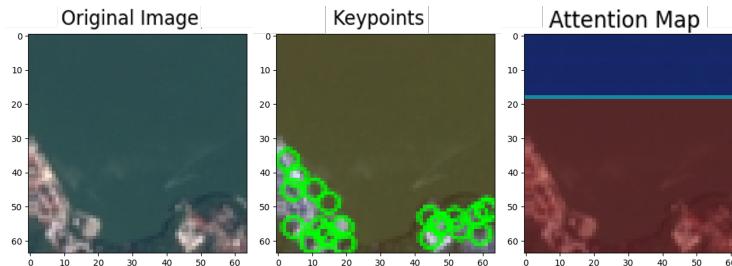


FIGURE 3.12 – Exemple illustrant le mécanisme d'attention guidé par les points d'intérêt SIFT.

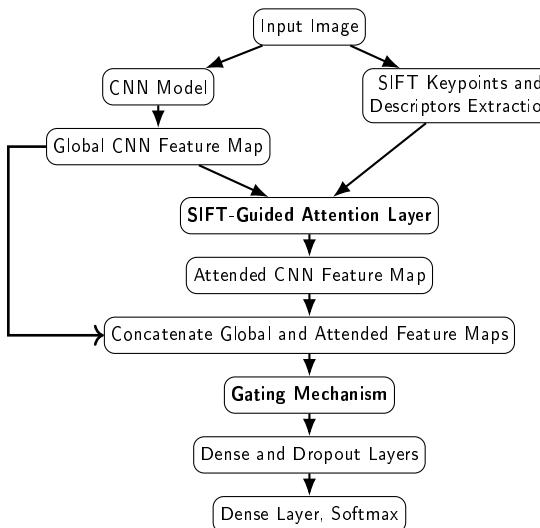


FIGURE 3.13 – Fusion à mi-niveau : *Fusion of Local Attended CNN Features and Global CNN Features with Gating Mechanism*

Comme montré dans la Figure 3.13, dans cette approche intitulée LFGF (*Fusion of Local Attended CNN Features and Global CNN Features*), la *feature map* globale issue de la branche CNN passe par une couche d’attention guidée par SIFT (*SIFT-Guided Attention Layer*). Cette couche affine la *feature map* originale du CNN en mettant l’accent sur les régions locales informatives identifiées par les descripteurs SIFT, générant une *Attended Feature Map*. La *feature map* originale, spécialisée dans la capture du contexte global, et la *feature map* affinée, qui se concentre sur les détails locaux, sont ensuite concaténées pour former un vecteur de caractéristiques unifié.

La couche d’attention guidée par SIFT que nous proposons repose sur trois composants : les *couches de projection*, l’*attention par produit scalaire normalisé*, et la *pondération et agrégation des caractéristiques*. Les caractéristiques CNN et les points-clés SIFT sont projetés dans un espace latent commun via des couches entièrement connectées, qui les transforment en vecteurs de même dimension pour permettre les calculs de similarité. Ensuite, les caractéristiques projetées sont soumises à un mécanisme d’attention par produit scalaire normalisé. Ce mécanisme calcule un score mesurant la similarité entre les caractéristiques issues de SIFT (descripteurs et points d’intérêt) et celles du CNN. Ces scores indiquent la pertinence de chaque région dans la *feature map* CNN par rapport aux points d’intérêt SIFT. Les poids d’attention sont utilisés pour moduler les caractéristiques CNN : la *feature map* originale du CNN est multipliée par les poids d’attention, produisant l’"Attended Feature Map".

Le vecteur de caractéristiques unifié passe par un *mécanisme de gating*, qui sélectionne les caractéristiques les plus pertinentes, réduisant ainsi la redondance entre les deux *feature maps*. Pour renforcer la généralisation et réduire les risques de sur-apprentissage, une régularisation L1 est appliquée aux couches denses projetant les descripteurs SIFT et les caractéristiques CNN, ainsi qu’au mécanisme de

gating. Cette régularisation favorise une sélection parcimonieuse des poids, maximisant l'utilisation des caractéristiques les plus pertinentes.

3.3.3 Résultats & discussion

Cette section évalue la performance des stratégies de fusion proposées sur le jeu de données EuroSAT. Les stratégies de fusion ont été implémentées avec un CNN simple peu profond (*shallow CNN*) et MobileNetV2 pré-entraîné [H2C+17] affiné sur EuroSAT. Bien que MobileNetV2 ne soit pas le modèle le plus performant, il offre un bon compromis entre précision et rapidité. Tous les modèles ont été implémentés avec Keras et TensorFlow [AAB+15]. Les descripteurs SIFT ont été extraits avec OpenCV [Ope]. Les techniques courantes d'augmentation d'images ont été appliquées pour améliorer la robustesse des modèles. Le jeu de données a été stratifié par classe et divisé en ensembles d'apprentissage, de validation et de test selon un ratio de 70/15/15. Chaque modèle a été entraîné sur 100 epochs avec les stratégies d'arrêt anticipé (*early stopping*) et de réduction du taux d'apprentissage sur plateau (*learning rate reduction on plateau*).

| Modèle | Accuracy | F1-Score |
|----------------------------------|----------|----------|
| <i>Modèles de référence</i> | | |
| SIFT-NN | 0,619 | 0,617 |
| CNN peu profond | 0,845 | 0,845 |
| Fine-tuned MobileNetV2 | 0,966 | 0,965 |
| <i>Fusion précoce</i> | | |
| Shallow CNN | 0,887 | 0,887 |
| Fine-tuned | 0,976 | 0,975 |
| <i>Fusion tardive (ADL)</i> | | |
| Shallow CNN | 0,911 | 0,910 |
| Fine-tuned MobileNetV2 | 0,984 | 0,984 |
| <i>Fusion à mi-niveau (LFGF)</i> | | |
| Shallow CNN | 0,924 | 0,923 |
| MobileNet V2 | 0,985 | 0,985 |

TABLE 3.4 – Accuracy obtenue par les modèles étudiés

Les résultats de la Table 3.4 montrent que les approches de fusion surpassent les modèles de référence SIFT-NN et CNN. La fusion tardive (ADL) améliore la précision de 7,68 % par rapport au CNN de base et de 47,17 % par rapport à SIFT-NN, grâce à l'intégration dynamique des caractéristiques saillantes des branches CNN et SIFT avant la classification finale. La fusion à mi-niveau (LFGF) apporte des améliorations encore plus significatives, avec une précision accrue de 9,22 % par rapport au CNN de base et de 49,27 % par rapport à SIFT-NN. Ces résultats démontrent l'efficacité des mécanismes d'attention pour combiner de manière optimale les caractéristiques locales et globales. Par ailleurs, les gains obtenus par les approches de fusion avec le fine-tuned MobileNetV2 sont moindres qu'avec le shallow CNN. Cela s'explique par le fait que les caractéristiques apprises par MobileNetV2 atteignent déjà une performance élevée, laissant moins de marge d'amélioration par l'incorporation de

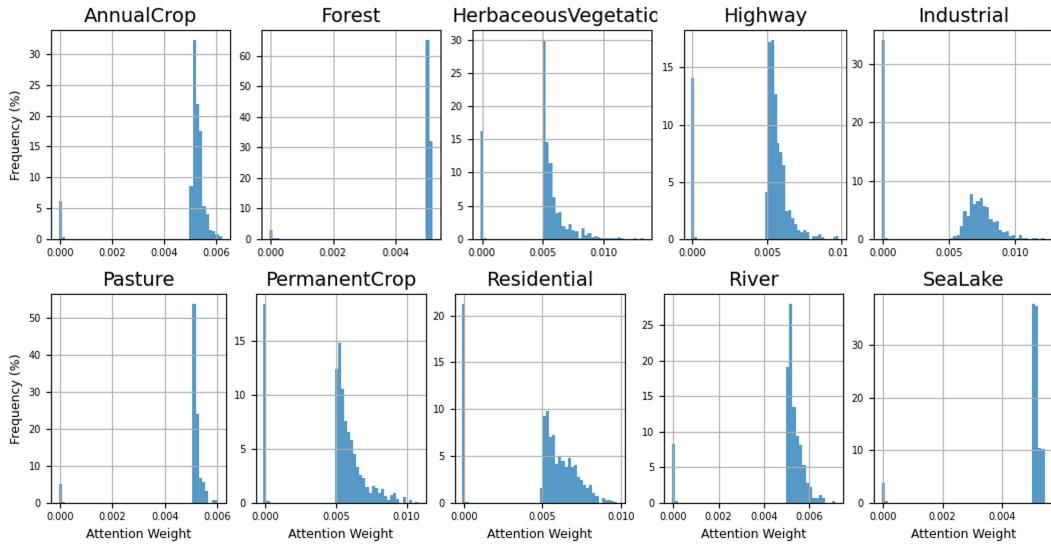
| Modèle | Accuracy | # Paramètres (Approx.) |
|---|----------|------------------------|
| Random Forest [TP22] | 0,567 | n.c. |
| VGG19 pré-entraîné avec TWSVM [DA21] | 0,946 | 143,7 M |
| SHAP [TTK ⁺ 23] | 0,947 | 1,9 M |
| Fusion ResNet50 et D-CNN [BB24] | 0,952 | >25,6 M |
| Fusion SIBNet et Tiny ViT [RKH ⁺ 24] | 0,978 | 13,4 M |
| Fine-tuned ResNet50 [HBDB18] | 0,986 | 25,6 M |
| ResNet50 fine-tuned sur des jeux spécialisés [NPZH20] | 0,992 | 25,6 M |
| μ 2Net [GD22] | 0,992 | 301 M |
| LFGF - CNN peu profond | 0,924 | 1,37 M |
| LFGF - MobileNetV2 | 0,985 | 3,55 M |

TABLE 3.5 – Comparaison des performances des approches existantes et proposées

caractéristiques supplémentaires. Cependant, les gains observés sur les deux modèles démontrent la généralisabilité des approches de fusion proposées.

La Table 3.5 met en lumière la compétitivité des approches proposées face aux méthodes existantes. Bien que des modèles pré-entraînés tels que ResNet50 ou μ 2Net atteignent des précisions très élevées (jusqu'à 99,2 %), ils sont extrêmement coûteux en termes de complexité computationnelle et de taille (jusqu'à 301 millions de paramètres pour μ 2Net). À l'inverse, avec un CNN peu profond notre approche de fusion à mi-niveau atteint une précision compétitive de 92,4 % avec seulement 1,37 million de paramètres. Avec MobileNetV2, notre méthode atteint 98,5 % de précision, approchant les meilleures performances tout en restant nettement plus légère. Ces résultats montrent que les stratégies proposées fournissent un bon compromis entre précision et efficacité, offrant des solutions adaptées aux applications opérant dans des environnements contraints.

La Figure 3.14 illustre la distribution des poids d'attention dans la *feature map* pondérée par SIFT dans l'approche de fusion à mi-niveau. Ces poids modulent l'importance des régions en fonction des informations locales. Les classes homogènes (Forest, SeaLake, Pasture) présentent des distributions étroites avec des pics autour de zéro, traduisant une prédominance des caractéristiques globales, cohérente avec leurs textures uniformes. À l'inverse, les classes complexes (Industrial, Residential, River) affichent des distributions plus larges, révélant une dépendance accrue aux détails locaux pour capturer la diversité structurelle. Les classes agricoles (AnnualCrop, PermanentCrop) combinent pics proches de zéro et poids non nuls, reflétant la nécessité d'équilibrer motifs globaux et variations locales. Enfin, pour des infrastructures linéaires comme Highway, la distribution suggère une attention modérée aux détails locaux, en phase avec les motifs directionnels des routes. La présence de poids non nuls, particulièrement marquée dans les classes complexes, confirme la capacité du modèle à capturer des détails fins tout en maintenant le contexte global.


 FIGURE 3.14 – Distribution des poids dans la *feature map* podérée par SIFT (LFGF-CNN)

3.4 Conclusion

Les travaux présentés dans ce chapitre ont pour objectif commun de développer des approches novatrices d'extraction de caractéristiques capables de s'adapter aux défis spécifiques posés par des données complexes et variées. Dans la première partie de ce chapitre nous avons présenté nos contributions relatives à l'extraction de caractéristiques dans le contexte de la recherche d'information et plus spécifiquement dans celui de la reconnaissance d'empreintes digitales. Nous y avons notamment présenté notre approche EDT-C, basée sur la triangulation de Delaunay étendue. Comme l'a montré nos études expérimentales, cette approche est particulièrement utile pour les bases volumineuses d'empreintes ou lors que les empreintes sont de faible qualité. Ces travaux de recherche ont été menés dans le cadre d'une thèse MOBIDOC en partenariat avec l'industrie et le ministère de l'intérieur.

Dans la deuxième partie de ce chapitre, nous avons présenté nos contributions relatives à l'extraction de caractéristiques dans le contexte de l'apprentissage automatique, et plus spécifiquement dans celui de la classification des images satellitaires. Nous y avons notamment proposé de fusionner les caractéristiques locales extraites par SIFT avec celles globales extraites par des CNN. Trois stratégies de fusion ont été proposées. La première, dite *précoce*, combine les deux modalités dès le début du processus d'apprentissage. La seconde, dite *tardive*, applique un mécanisme d'attention à chaque modalité afin de ne retenir que le sous-ensemble de caractéristiques le plus informatif, avant de les combiner au stade de décision. La troisième, dite *à mi-niveau*, adopte une logique de fusion différente où un mécanisme d'attention guidée par SIFT oriente l'attention du CNN vers les régions les plus informatives de l'image. Les résultats expérimentaux ont confirmé le bien-fondé des approches proposées et ont mis en

évidence l'apport des mécanismes d'attention. Ces travaux de recherche se poursuivent dans le cadre de la thèse de Mme Vian Abdulmajeed Ahmad et de notre collaboration avec la *Northern Technical University* (Mossoul, Irak).

Les travaux présentés dans ce chapitre illustrent la diversité des défis liés à l'extraction de caractéristiques dans des domaines variés du Big Data. Ils montrent également l'importance d'adapter les méthodes d'extraction à la nature des données et aux objectifs spécifiques. Ce chapitre illustre également l'importance accordée aux partenariats industriels et académiques dans mon parcours de chercheur.

Après le Volume et la Variété, le chapitre suivant explicite nos contributions relatives à la *Vélocité*, un autre aspect important caractérisant le Big Data.

Publications en lien avec le chapitre

Conférences internationales

[AJK25] Enhancing LULC Classification with Attention-Based Fusion of Handcrafted and Deep Features

Vian ABDULMAJEED AHMAD, Khaled JOUINI & Ouajdi KORBA

17th International Conference on Agents and Artificial Intelligence (ICAART). Feb 2025. (*Accepted, to appear*)

CORE Rank : B.

[AJTK24a] Integrating Deep and handcrafted Features for Enhanced Remote Sensing Image Classification

Vian ABDULMAJEED AHMAD, Khaled JOUINI, Amel TUAMA & Ouajdi KORBA

21st ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). 2024. (*Accepted, To appear*)

CORE Rank : C.

[AJTK24b] A Fusion Approach for Enhanced Remote Sensing Image Classification

Vian ABDULMAJEED AHMAD, Khaled JOUINI, Amel TUAMA & Ouajdi KORBA

19th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP. DOI : <https://doi.org/10.5220/0012376600003660>. 2024.

CORE Rank : B (lors de la soumission).

[GJK17a] Fast and Accurate Fingerprint Matching using Expanded Delaunay Triangulation

Mohamed Hédi GHADDAB, Khaled JOUINI & Ouajdi KORBA

14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA).

DOI : <https://doi.org/10.1109/AICCSA.2017.8239101>. 2017.

CORE Rank : C

[GJK17b] Fusion de minuties pour une reconnaissance efficiente des empreintes digitales

Mohamed Hédi GHADDAB, Khaled JOUINI & Ouajdi KORBAA

Colloque sur l'Optimisation et les Systèmes d'Information (COSI). Algérie. 2017.

Encadrement doctoral

- Mohamed Hédi GHADDAB (Soutenue le 07-12-2018, Directeur de thèse : Prof. Ouajdi KORBAA)
- Vian ABDELMAJEED AHMAD (En cours, Directeur de thèse : Prof. Ouajdi KORBAA)

CHAPITRE 4

Vélocité : Apprentissage sur flux, adaptatif aux dérives de concept

Sommaire

| | | |
|------------|---|-----------|
| 4.1 | Introduction | 47 |
| 4.2 | Apprentissage en ligne : modèles de référence | 48 |
| 4.2.1 | Arbre de Hoeffding [DH00] | 48 |
| 4.2.2 | Forêts aléatoires en ligne (<i>Online Random Forest</i>) [SLS ⁺ 09] | 49 |
| 4.2.3 | <i>Fast Incremental Model Trees with Drift Detection</i> [IGD11] | 49 |
| 4.3 | Classification en ligne adaptative aux drifts - cas de la détection d'intrusions | 50 |
| 4.3.1 | Classification en ligne appliquée à la détection d'intrusion | 51 |
| 4.3.2 | DDM-ORF : classification ensembliste, adaptative aux drifts | 51 |
| 4.3.3 | Résultats & discussion | 53 |
| 4.4 | Régression en ligne adaptative aux drifts - cas de l'évolution des pandémies | 54 |
| 4.4.1 | Régression incrémentale appliquée à la prédiction épidémiologique | 54 |
| 4.4.2 | <i>Extremely Fast Regression Tree with Drift Detection</i> (EFRT-DD) | 55 |
| 4.4.3 | <i>Collaborative Drift-Driven Regression</i> (CDR) | 56 |
| 4.4.4 | Résultats & discussion | 57 |
| 4.5 | Conclusion | 62 |

4.1 Introduction

En septembre 2018, j'ai eu l'opportunité d'effectuer un séjour de recherche au sein du Laboratoire de Traitement et Communication de l'Information (LTCI) de Télécom -pariTech (France). Lors de ce séjour, j'ai eu notamment le privilège d'échanger longuement avec Pr. A. BIFET, référence mondiale dans l'apprentissage incrémental, à l'origine des systèmes précurseurs *Massive OnLine Analysis* [BGHP18], Apache SAMOA [KMB18] et Scikit-Multiflow [MRBA18]. Ces échanges m'ont ouvert la voie vers les sujets ayant trait à l'apprentissage en ligne adaptatif aux dérives de concept.

L'apprentissage automatique a pour but de trouver la relation entre les variables explicatives en entrée d'un modèle et la variable cible que le modèle tente de prédire. Dans certaines situations, il peut arriver que cette relation change au fil du temps, de manière à ce qu'un modèle appris sur les données du passé, ne reflète plus fidèlement la relation actuelle entre les variables explicatives et la variable cible. Dans l'apprentissage automatique et les domaines connexes, les changements au fil du temps dans la relation entre ces variables sont connus sous le nom de *dérives conceptuelles (Concept Drift)*. Les dérives sont communes dans les environnements dynamiques, comme la cybersécurité ou l'épidémiologie. En cybersécurité, elles se traduisent par l'évolution constante des menaces et des tactiques adoptées pour contourner les défenses existantes. En épidémiologie, elles se matérialisent par les changements dans la façon dont le virus se propage ou mute, les évolutions dans la réponse immunitaire de la population, la saisonnalité, etc..

La majorité des algorithmes d'apprentissage conventionnels (*batch* ou par lots), prennent l'hypothèse forte que toutes les données d'apprentissage sont disponibles avant la formation du modèle et que la distribution des données est stationnaire. En présence de dérives et sans une intervention appropriée, les modèles batch deviennent inéluctablement moins précis au fil du temps. Une intervention courante consiste à *ré-apprendre* à nouveau le modèle. Outre la charge de calcul, le ré-apprentissage soulève trois défis importants : (i) déterminer quand un modèle n'est plus assez précis et nécessite un ré-apprentissage (*i.e. dilemme stabilité-plasticité*) ; (ii) déterminer les données sur lesquelles le modèle doit être ré-appris ; et (iii) décider de la quantité de nouvelles données à collecter avant le ré-apprentissage, du fait que collecter plus de données augmente les chances de générer un modèle précis, mais retarder le remplacement de l'ancien modèle devenu imprécis.

L'*apprentissage incrémental* est une forme d'apprentissage particulièrement adapté aux flux de données, les données massives et plus globalement aux situations où il est impossible de traiter en même temps l'intégralité des données [MKS23]. Avec l'apprentissage incrémental, le modèle se construit *à la volée*, de manière *continue* et *progressive*, au fur et à mesure que de nouvelles instances d'apprentissage deviennent disponibles. Du fait qu'ils prennent en permanence en compte les données récentes, les modèles incrémentaux résistent mieux aux dérives de concept que les modèles batch. Une autre caractéristique fort intéressante de ces modèles est qu'ils commencent à fournir des prédictions après les toutes premières instances d'apprentissage (*i.e. prédiction anytime* [BGHP18]).

Dans les travaux de recherche présentés dans ce chapitre, nous nous intéressons à la classification et à la régression incrémentales adaptatives aux dérives de concept. En collaboration avec Dr. Farah Jemili, spécialiste en détection d'intrusions, nous avons notamment développé une approche de classification en ligne spécialement conçue pour s'adapter aux évolutions constantes des menaces cybernétiques. Parallèlement, nous avons proposé une approche de régression en ligne pour la prédiction de l'évolution des épidémies.

Dans la suite, la section 4.2 approfondit les concepts clés de l'apprentissage incrémental et des dérives conceptuelles. Les sections 4.3 et 4.4 présentent respectivement les contributions relatives à la classification incrémentale et à la régression incrémentale. Enfin, une synthèse des résultats est présentée. Suit de la conclusion.

4.2 Apprentissage en ligne : modèles de référence

Dans les domaines sensibles, tels que la cybersécurité et la santé publique, l'interprétabilité d'un modèle est au moins aussi importante que la précision de ses prédictions et les modèles *white-box* sont souvent préférés aux modèles *black-box*. Les arbres de décision font partie des modèles white-box les plus répandus. Cette section présente les concepts clés liés aux arbres incrémentaux, et notamment l'arbre de Hoeffding [DH00], qui est à la base des approches proposées.

4.2.1 Arbre de Hoeffding [DH00]

Un arbre de décision (*Decision Tree*) se construit de haut en bas (division par les feuilles), par le remplacement récursif des feuilles par des nœuds de test. La récursion se termine lorsqu'une feuille est considérée comme suffisamment homogène, ou lorsque la division ne permet plus d'améliorer l'homogénéité. La décision cruciale lors de l'induction est de déterminer quand diviser un nœud et selon quel attribut. Les arbres de décision batch supposent que toutes les données d'apprentissage sont disponibles avant l'induction et parcourront l'ensemble des données pour découvrir le meilleur attribut de division. Cette méthode d'induction ne peut pas être adoptée en l'état dans les contextes où seule une petite fraction des données est accessible pendant l'apprentissage (cas typique avec des flux de données).

L'arbre de Hoeffding (*Hoeffding Tree* ou HT) [DH00] est le standard de facto pour l'apprentissage incrémental sur les flux et est à la base de la majorité des approches récentes [IGD11, IGZD11, IGD14, MWS18, GBFB18, GMM⁺20]. L'idée principale de HT est qu'une petite fraction de données peut souvent suffire pour choisir un bon attribut de division. Cette idée est soutenue par la *Borne de Hoeffding (Hoeffding Bound)* qui stipule qu'après n observations indépendantes et avec une probabilité $1-\delta$, la moyenne réelle d'une variable aléatoire d'amplitude R , ne différera pas de la moyenne empirique de plus de [Hoe63] : $\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$.

Soit X_a et X_b les deux attributs les plus discriminants au niveau d'une feuille (X_a en premier, suivi par X_b). Par "les plus discriminants", nous entendons les deux attributs permettant le meilleur et le

second meilleur gain d'information. Dans HT, la variable aléatoire à estimer est la différence entre le gain d'information que permet X_a et celui que permet X_b , $G(X_a) - G(X_b)$. La borne de Hoeffding est alors utilisée pour garantir avec un certain degré de confiance $1-\delta$ que, si $G(X_a) - G(X_b) > \epsilon$, alors X_a demeurera toujours un meilleur attribut de segmentation que X_b .

Le *Hoeffding Adaptive Tree* [BG07] le *Fast Incremental Model Tree with Drift Detection* (FIMT-DD) [IGD11], et l'*Online Random Forest* (ORF) [SLS⁺09] sont des extensions de l'arbre de Hoeffding intégrant des mécanismes spéciaux permettant une adaptabilité accrue aux dérives de concept.

4.2.2 Forêts aléatoires en ligne (*Online Random Forest*) [SLS⁺09]

La forêt aléatoire (*Random Forest*) est un modèle ensembliste composé de plusieurs arbres de décision, chacun construit à partir d'un échantillon aléatoire de données et de caractéristiques. Chaque arbre produit des décisions indépendantes et le résultat final est obtenu par un vote majoritaire. Ce mécanisme vise à améliorer la robustesse et la précision globales du modèle, les erreurs des arbres individuels ayant tendance à se compenser. ORF (*Online Random Forest*) [SLS⁺09] adapte ce principe à l'apprentissage en ligne en utilisant des arbres de Hoeffding comme modèles sous-jacents.

Pour garantir la diversité, chaque nouvelle instance de données est soumise à chacun des arbres avec un nombre aléatoire k . L'instance est alors soit ignorée ($k = 0$) ou utilisée une ou plusieurs fois ($k \geq 1$) pour mettre à jour l'arbre, simulant ainsi l'échantillonnage avec remise. En parallèle, ORF effectue une sous-sélection aléatoire des caractéristiques disponibles à chaque nœud, permettant à chaque arbre de se concentrer sur aspects différents des données.

ORF présente un compromis entre la précision, améliorée par l'apprentissage ensembliste, et l'interprétabilité partiellement préservée par les arbres de Hoeffding.

4.2.3 *Fast Incremental Model Trees with Drift Detection* [IGD11]

Plusieurs travaux relativement récents sur la régression incrémentale, tels qu'*OBag* (*Online Bagging of FIMT-DD trees*) [IGD14] et *ARF-FIMT-DD* (*Adaptive Random Forest*) [GBFB18, GMM⁺20], se sont intéressés à l'apprentissage ensembliste. Le point commun entre ces approches est qu'elles s'appuient sur le FIMT-DD (*Fast Incremental Model Trees with Drift Detection*) [IGD11], considéré à ce jour comme l'un des arbres de régression incrémentale les plus efficaces [BBG⁺21]. Les points saillants du FIMT-DD sont synthétisés ci-après.

— *Critère de division* : FIMT-DD utilise la réduction de l'écart-type comme critère de division.

Étant donné une feuille l où un échantillon de cardinalité N a été observé, une segmentation binaire selon un attribut X divise les échantillons de l en deux sous-ensembles l_l et l_r , de cardinalités respectives N_l et N_r . La réduction de l'écart-type $SDR(X)$ se calcule comme suit : $SDR(X) = sd(l) - \left(\frac{N_l}{N} sd(l_l) + \frac{N_r}{N} sd(l_r) \right)$, où $sd(node)$ est l'écart-type de la variable cible dans $node$. Soit X_a et X_b les deux attributs permettant resp. la plus grande et la seconde plus grande réduction de l'écart-type. De manière similaire à HT, FIMT-DD utilise la borne de Hoeffding

pour contrôler le risque que, à mesure que les données arrivent, la division selon X_b devienne meilleure que celle selon X_a . En pratique, FIMT-DD tarde la division d'un noeud jusqu'à ce que la condition ci-après soit satisfaite : $\frac{SDR(X_b)}{SDR(X_a)} < 1 - \epsilon$

- *Perceptrons dans les feuilles* : FIMT-DD utilise un perceptron dans chacune des feuilles pour produire les prédictions finales. Les poids de ces perceptrons sont continuellement mis à jour à mesure que de nouvelles données arrivent. FIMT-DD utilise la descente de gradient stochastique avec comme objectif de réduire l'erreur quadratique moyenne. En plus de leur efficacité avérée, les perceptrons ont l'avantage crucial de s'adapter naturellement aux dérives [IGD11].
- *Adaptation aux dérives* : L'intuition derrière l'adaptation au changement dans FIMT-DD est de construire des sous-arbres alternatifs dans un noeud, puis de ne garder que le sous-arbre le plus prometteur. Cette stratégie initialement introduite dans le *Very Fast Decision Tree* [DH00] a été également reprise par le *Hoeffding Adaptive Tree* (HAT) [Bif10]. FIMT-DD utilise le test Page-Hinkley (PH) [MMMS04] au niveau des noeuds internes pour détecter les changements dans le taux d'erreur. Lorsqu'un changement est détecté dans un noeud interne *inner*, la construction d'un nouveau sous-arbre alternatif à celui existant et enraciné à *inner* est initiée : chaque nouvelle instance atteignant *inner* fait croître à la fois l'ancien et le nouveau sous-arbre. Le nouveau sous-arbre remplace le sous-arbre d'origine si les prédictions qu'il produit sont meilleures au regard de la fonction de perte.

Les approches existantes en apprentissage incrémental, telles que les arbres de Hoeffding, leurs variantes (HAT, FIMT-DD), et les forêts aléatoires en ligne (ORF), ont démontré leur efficacité dans divers contextes dynamiques. Cependant, elles présentent des limitations importantes. Premièrement, la majorité de ces modèles repose sur des décisions de division irrévocables, ce qui peut affecter leur capacité à s'adapter à des changements imprévus dans les données. Deuxièmement, bien que des mécanismes de détection de dérives soient intégrés dans certaines variantes, leur sensibilité aux dérives subtiles ou graduelles reste limitée.

4.3 Classification en ligne adaptative aux drifts - cas de la détection d'intrusions

Les systèmes de détection d'intrusions (*Intrusion Detection Systems* ou IDS) sont essentiels pour protéger les réseaux informatiques contre diverses menaces [MFMT22]. Face à l'évolution inéluctable des tactiques de contournement des mesures de sécurité, les IDS doivent pouvoir s'adapter rapidement aux changements des schémas d'attaque. Dans [JJK25b], nous proposons l'utilisation conjointe de l'apprentissage en ligne et de la détection automatique des dérives pour maintenir l'efficacité des IDS dans les environnements dynamiques. Plus précisément, nous introduisons une nouvelle méthode, dite DDM-ORF utilisant ORF comme méthode d'apprentissage et DDM (*Drift Detection Method*)

comme détecteur de dérives de concept. Sommairement, DDM surveille en permanence les erreurs de classification d'ORF pour repérer les changements significatifs. Lorsqu'un changement est détecté, ORF ajuste automatiquement ses critères de classification pour s'adapter aux nouveaux schémas d'attaque. Dans les sections suivantes, nous passons en revue les principales approches utilisant l'apprentissage en ligne dans le contexte de la détection d'intrusions. Puis, nous détaillons l'architecture de DDM-ORF, ses mécanismes d'apprentissage, ainsi que les résultats expérimentaux.

4.3.1 Classification en ligne appliquée à la détection d'intrusion

Plusieurs approches ont été proposées pour gérer la dérive conceptuelle dans les IDS, dont notamment [YWZ⁺18], [SRJ22] et [LZZ⁺23]. Dans [YWZ⁺18] le flux de données est d'abord segmenté, chaque segment étant analysé et comparé aux précédents pour identifier des dérives de concept. La détection d'une dérive déclenche la mise à jour de l'ensemble de classificateurs. Bien que cette approche permette un certaine adaptabilité, son efficacité est fortement dépendante de la granularité des segments : des segments trop grands retardent la détection de dérives, tandis que des segments trop petits augmentent le risque de fausses alertes.

[LZZ⁺23] propose une approche de détection d'intrusions dans les environnements IoT dynamiques. L'approche proposée utilise un encodeur automatique épars empilé (SSAE) pour extraire les caractéristiques des données de trafic réseau, ainsi qu'un réseau de neurones incrémental auto-organisé (SOINN), capable d'apprendre de nouveaux schémas en ajustant sa topologie par l'ajout de nouveaux nœuds. Afin d'éviter l'oubli des schémas d'attaque antérieurs, des contraintes sont appliquées à la topologie. Un inconvénient majeur de l'approche SSAE-SOINN est qu'elle n'intègre pas un mécanisme automatique de détection des dérives, essentiel pour garantir des mises à jour du modèle en temps opportun.

L'approche PCA-based Deep Neural Networks (PCA-DNN) proposée dans [SRJ22] utilise l'analyse en composantes principales (PCA) pour détecter les dérives par l'évaluation des changements entre les données courantes et celles antérieures. Pour s'adapter aux dérives, l'approche utilise un réseau neuronal profond en ligne, capable d'ajuster dynamiquement la taille de ses couches cachées grâce à un mécanisme de pondération basé sur l'algorithme *Hedge*. La performance de chaque couche du réseau est évaluée à l'aide d'une fonction de perte adaptative, qui met à jour les pondérations en fonction des résultats. Contrairement à PCA, qui se limite à analyser les variations des caractéristiques, la méthode DDM offre une détection plus efficace des changements dans les données.

4.3.2 DDM-ORF : classification ensembliste, adaptative aux drifts

Les approches telles que SSAE-SOINN et PCA-DNN ont exploré diverses stratégies pour adapter les modèles aux évolutions des schémas d'attaque. Cependant, ces méthodes souffrent d'un manque d'intégration explicite de mécanismes automatiques de détection de dérives, limitant leur efficacité dans des environnements hautement dynamiques. En réponse à ces limitations, notre approche DDM-ORF combine l'apprentissage en ligne ensembliste avec un détecteur de dérives (DDM). Contrairement à ces

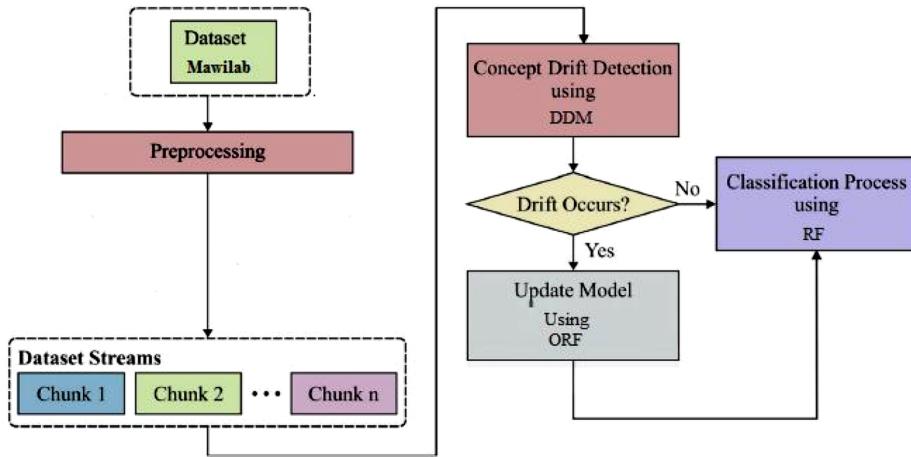


FIGURE 4.1 – Vue d'ensemble de l'approche DDM-ORF [JJK25b].

approches, DDM-ORF détecte les changements significatifs dans le flux de données en temps réel, tout en mettant à jour le modèle de manière incrémentale pour préserver à la fois la précision et la réactivité. La Figure 4.1 donne une vue d'ensemble de notre approche. Similairement à un apprentissage batch, le pipeline comprend trois étapes : collecte et pré-traitement des données, apprentissage et inférence. La différence fondamentale est que ces étapes se déroulent en parallèle et non en séquence comme il est le cas avec l'apprentissage batch.

La collecte de données implique l'utilisation d'un outil de web scraping (*beautifulSoup*) pour extraire les données en temps réel. Le processus de nettoyage et de transformation des données brutes extraites, inclut notamment la normalisation et la sélection des caractéristiques pertinentes avec le *Chi-Square Test* et la matrice de corrélation. Une fois prétraitées, les données servent à mettre à jour de manière incrémentale ORF au fur et à mesure que DDM détecte des dérives.

DDM détecte les dérives de concept par le suivi au fil du temps du taux d'erreur de classification ϵ_t du modèle ORF courant. DDM surveille l'évolution de ϵ_t à l'aide de l'erreur moyenne cumulée m_t et de l'écart-type s_t . Ces deux mesures sont mises à jour de manière incrémentale à chaque nouvelle instance comme suit : $m_t = m_{t-1} + \frac{\epsilon_t - m_{t-1}}{t+1}$, $s_t = \sqrt{s_{t-1}^2 + (\epsilon_t - m_{t-1})(\epsilon_t - m_t)}$. Ici, m_t est la moyenne du taux d'erreur jusqu'à l'instant t , et s_t est l'écart-type de ces erreurs. Ces mesures permettent de quantifier la variabilité de l'erreur au fil du temps. Le mécanisme de détection de dérives de DDM repose sur un seuil, dit *Seuil de Dérive* défini comme suit : $m_t + 3 \times s_t$. Si le taux d'erreur dépasse ce seuil, cela indique une dérive de concept dans le flux de données.

Lorsqu'une dérive est détectée, DDM réinitialise ses compteurs se préparant ainsi à la détection de la prochaine dérive. Parallèlement, notre approche déclenche une mise à jour incrémentale d'ORF. Cette mise à jour se matérialise par les trois actions suivantes. (i) *Mise à Jour des Arbres*. La mise à jour des arbres de base se fait en incorporant les nouvelles instances de données. Pour chaque arbre, le

nœud feuille auquel appartient la nouvelle instance de données est identifié et les statistiques associées à ce nœud (*e.g.* nombre d'observations) sont mises à jour; (*ii*) *Pruning*. Les performances de chaque arbre de base sont réévaluées. Les arbres qui ne contribuent pas significativement à la précision globale du modèle sont élagués. (*iii*) *Ajout de Nouveaux Arbres*. Si le nombre total d'arbres dans la forêt est insuffisant pour capturer les nouvelles tendances dans les données, de nouveaux arbres sont ajoutés à l'ensemble. Chaque nouvel arbre est appris en utilisant les nouvelles données et des caractéristiques sélectionnées aléatoirement.

4.3.3 Résultats & discussion

Nous avons comparé notre modèle DDM-ORF aux principales approches de détection d'intrusions sur le banc de tests MAWILab [FBAF10]. Par souci de concision, seuls les principaux résultats sont présentés. Comme illustré dans la Table 4.1, DDM-ORF surpassé significativement les autres méthodes en termes de précision, de rappel et de F1-score. Pour évaluer l'efficacité de notre approche dans des conditions d'utilisation réelles, nous l'avons également testée avec Apache Spark Structured Streaming, une plateforme de traitement des flux massifs de données en temps réel. Deux métriques ont été utilisées : le taux d'ingestion (nombre de lignes chargées par seconde) et le taux de traitement (nombre de lignes traitées par seconde). Comme indiqué dans la Table 4.2, le système peut charger et traiter les données à des taux élevés, atteignant respectivement ≈ 560 k et ≈ 55 k événements par seconde.

| Méthode | Accuracy | Precision | Recall | F1-Score |
|---|----------|-----------|--------|----------|
| CNN-based Model [DJ21] | 98.7% | 98.7% | 97.0% | 97.8% |
| Ensemble Incremental Learning [YWZ ⁺ 18] | 98.2% | 98.5% | 97.0% | 97.7% |
| Adaptive Class Incremental Learning [LZZ ⁺ 23] | 98.8% | 98.0% | 97.5% | 97.3% |
| Active Incremental Learning [SRJ22] | 98.6% | 98.8% | 97.2% | 97.0% |
| DDM-ORF | 99.96% | 99.93% | 99.95% | 99.94% |

TABLE 4.1 – DDM-ORF vs. approches existantes

| | Résultats |
|-----------------|----------------------|
| Input Rate | 560 808 événements/s |
| Processing Rate | 55 175 événements/s |

TABLE 4.2 – Taux d'ingestion et de traitement de DDM-ORF sous Apache Spark Structured Streaming

Dans l'ensemble, ces résultats démontrent l'efficacité du modèle DDM-ORF et sa capacité à traiter en ligne de grands volumes de données avec une haute précision. Cependant, comme toute approche, DDM-ORF présente certaines limitations. La première concerne le détecteur de dérives DDM, qui peut rencontrer des difficultés à identifier les dérives subtiles ou graduelles non présentes dans MAWILab. Un autre défi important réside dans les besoins croissants en mémoire vive d'ORF, qui augmentent à mesure que le modèle s'adapte aux nouvelles données tout en essayant de conserver ses connaissances

antérieures. Cela peut, à terme, conduire à un modèle volumineux et gourmand en ressources. Les travaux futurs pourraient explorer des détecteurs de dérives plus sophistiqués, tels qu'ADWIN, afin d'améliorer la sensibilité aux dérives subtiles. De plus, l'utilisation de stratégies d'élagage plus agressives et de pondérations basées sur des facteurs d'oubli pourrait contribuer à une meilleure gestion de la mémoire par ORF.

4.4 Régression en ligne adaptative aux drifts - cas de l'évolution des pandémies

Les pandémies ne cesseront jamais d'émerger et de menacer à la fois la santé publique et l'économie mondiale. Il est donc crucial de tirer des leçons des pandémies passées et de développer des outils pour prévenir et contrôler celles futures. Cela implique non seulement la recherche de nouvelles thérapies, mais également la mise en place de systèmes de surveillance épidémiologique efficents. La prédiction précise de l'évolution d'une pandémie est cependant une tâche particulièrement difficile, car elle implique de faire face simultanément à des dérives soudaines, des dérives graduelles et des dérives récurrentes qui peuvent interagir et s'influencer mutuellement.

Comme indiqué précédemment, les modèles incrémentaux s'adaptent mieux aux dérives de concept que les modèles batch. Le revers de l'apprentissage incrémental, cependant, est qu'il construit ses modèles en faisant des hypothèses sur les données à venir (un modèle incrémental est en effet une approximation du modèle batch correspondant). Les algorithmes batch ont besoin de temps pour collecter suffisamment de données, mais une fois le modèle construit, il est souvent plus précis que le modèle incrémental correspondant. L'apprentissage incrémental et l'apprentissage par lots sont souvent considérés comme des approches mutuellement exclusives [Jac19].

Dans notre travail, nous introduisons un nouvel arbre de régression appelé *The Extremely Fast Regression Tree with Drift Detection* ou EFRT-DD qui se distingue des arbres incrémentaux existants par sa stratégie d'induction "optimiste" ou *Eager*. Nos travaux introduisent également une nouvelle approche de régression collaborative dirigée par les dérives de concept, que nous intitulons *Collaborative Drift-Driven Regression* (CDR). L'idée derrière CDR est d'utiliser conjointement l'apprentissage incrémental et l'apprentissage batch pour bénéficier des avantages des deux types d'apprentissage.

Dans la suite de cette section, nous commençons par passer brièvement en revue les principales approches utilisant la régression incrémentale pour la prédiction de l'évolution des pandémies. Ensuite, nous détaillons nos approches EFRT-DD et CDR. Nous terminons cette section par la discussion des principaux résultats expérimentaux.

4.4.1 Régression incrémentale appliquée à la prédiction épidémiologique

La littérature sur la modélisation et la prédiction de l'évolution de la pandémie de COVID-19 propose diverses techniques, allant des méthodes élémentaires, telles qu'ARIMA, aux techniques d'ap-

apprentissage profond les plus avancées [MKP⁺22]. Cependant, à l'exception de [CAQ⁺22] et [MKS23], peu d'études ont porté sur l'apprentissage incrémental.

L'approche de [CAQ⁺22] comporte deux étapes. La première utilise un algorithme génétique en conjonction avec des modèles ARIMA pour sélectionner les prédicteurs optimaux pour les variables SEIRD (*Susceptible, Exposed, Infected, Recovered, Dead*). Dans la deuxième étape chaque variable SEIRD est prédite par un ensemble de modèles. Le modèle utilisé à un instant t est celui qui a été le plus précis à $t - 1$. Lorsque la précision est jugée insuffisante, de nouveaux modèles sont créés par ré-apprentissage complet sur l'intégralité des données disponibles à t . Un inconvénient majeur de l'approche [CAQ⁺22] est l'absence d'un mécanisme de détection des dérives, indiquant de manière claire et automatisée, les points de rupture où la précision d'un modèle est considérée comme insuffisante et donc où un ré-apprentissage est nécessaire.

Dans [MKS23], une étude comparative entre algorithmes de régression incrémentaux et batch sur les données de 50 pays a montré que les méthodes incrémentales s'adaptent mieux aux dérives tout en réduisant la charge computationnelle. [MKS23] teste et compare trois stratégies de prédiction : la première forme un modèle par pays, la deuxième utilise un modèle global appris sur les données de l'ensemble des pays, et la troisième combinant les données des pays où l'évolution de la pandémie est similaire (les évolutions similaires sont identifiées via *Dynamic Time Warping*). Les résultats obtenus montrent que cette troisième approche est la plus efficace. L'étude a inclus l'arbre de Hoeffding (HT), l'arbre de Hoeffding adaptatif (HAT), et la forêt aléatoire (ORF), mais a omis le FIMT-DD. Comme dans [CAQ⁺22], aucun mécanisme de détection de dérive n'a été inclus, et l'apprentissage batch et incrémental sont considérés comme mutuellement exclusifs.

4.4.2 *Extremely Fast Regression Tree with Drift Detection* (EFRT-DD)

Les arbres incrémentaux de régression sont un domaine de recherche relativement peu exploité comparativement aux arbres incrémentaux de classification [GMM⁺20]. L'*Extremely Fast Regression Tree with Drift Detection* (EFRT-DD) que nous proposons est une adaptation de l'*Extremely Fast Tree* [MWS18, MGS⁺22] à la régression et aux dérives de concept. L'EFRT-DD peut également être perçu comme une amélioration du FIMT-DD [IGD11]. Comme le FIMT-DD, EFRT-DD forme un perceptron à chaque feuille et utilise le test PH dans les nœuds internes. Contrairement à FIMT-DD, EFRT-DD divise un nœud dès qu'il est suffisamment confiant que la division est utile, et reconsidère ensuite cette décision s'il s'avère, à mesure que les données arrivent, qu'il existe une meilleure option de division.

Soit X_a et X_b , respectivement, l'actuel meilleur et second meilleur attribut de division au niveau d'une feuille l . HT et ses variantes (y compris le FIMT-DD) retardent la division de l jusqu'à ce qu'ils soient suffisamment confiants que X_a restera toujours meilleur que X_b , et ce, indépendamment de l'amélioration en homogénéité que X_a permet. Cette stratégie d'induction, que nous qualifions de "*Lazy*", présente deux inconvénients majeurs. Premièrement, différer la division des feuilles peut altérer les performances prédictives, du fait que l'arbre en cours de construction est également utilisé pour

l'inférence. Deuxièmement, la borne de Hoeffding est utilisée pour contrôler le risque que X_b devienne ultérieurement une meilleure option de segmentation que X_a . Cependant, à mesure de nouvelles instances arrivent, rien n'empêche qu'un troisième attribut se révèle être une meilleure option de division que X_a . Dans de telles situations, il n'y a pas de recours, puisque dans HT et ses variantes les divisions sont irrévocables.

Algorithm 2: AttemptToSplit

```

Input :  $l$ , a leaf node
         $n_{min}$ , grace period
1 Let  $N$  be the number of samples seen in  $l$ 
2 if  $N \bmod n_{min} = 0$  then
3   Compute  $SD(l)$ 
4   Compute  $\text{SumSDchildren}(X)$  for each attribute  $X_i$ 
5   Let  $X_a$  be the attribute with the highest  $SD_{ratio}$ 
6   Compute the Hoeffding Bound  $\epsilon$ 
7   if  $\frac{\text{SumSDchildren}(X_a)}{SD(l)} < 1 - \epsilon$  then
8     Split  $l$  on  $X_a$ 
9     forall each branch do
10    | Initialize a new leaf;
11    end
12  end
13 end

```

L'arbre EFRT-DD que nous proposons adopte une stratégie d'induction que nous qualifions de "*Eager*" ou optimiste. La stratégie de division de l'EFRT-DD est présentée dans l'algorithme 2. EFRT-DD divise une feuille l dès qu'il est assez confiant que la division est une meilleure option que la non-division (*i.e.* la décision de division dépend uniquement du gain en homogénéité que X_a permet). EFRT-DD utilise donc la borne de Hoeffding pour contrôler le risque que la non-division devienne, à mesure que les données arrivent, une meilleure option que la division. Avec EFRT-DD les décisions de division ne sont pas irrévocables. Ainsi, s'il s'avère ultérieurement que X_a n'est plus le meilleur attribut de division, EFRT-DD supprime le sous-arbre enraciné à l et le remplace par le sous-arbre engendré par la division selon le nouveau meilleur attribut. De même, s'il s'avère ultérieurement que la non-division de l est une meilleure option que sa division, EFRT-DD supprime le sous-arbre enraciné à l et transforme l en un noeud feuille.

4.4.3 Collaborative Drift-Driven Regression (CDR)

L'approche CDR que nous proposons utilise conjointement l'apprentissage incrémental et batch pour bénéficier à la fois de : (i) la précision des modèles batch ; et (ii) l'adaptabilité aux dérives et la propriété *anytime prediction* des modèles incrémentaux. Comme illustré dans la Figure 4.2, dans CDR l'apprentissage est un processus continu où l'apprentissage incrémental et l'apprentissage batch

coexistent. L'apprentissage incrémental est utilisé pour former et affiner en continu un modèle I , à mesure que de nouvelles données deviennent disponibles. L'apprentissage batch est utilisé pour former une séquence de modèles batch B_1, B_2, \dots, B_n . Comme le montre la Figure 4.2, chaque fois qu'une dérive est détectée, le modèle batch courant B_i est invalidé et est remplacé par un nouveau modèle B_{i+1} (Figure 4.2). Dans notre travail nous utilisons l'algorithme ADWIN (*ADaptive WINdowing*) [BG07, Bif10] pour la détection des dérives de concept.

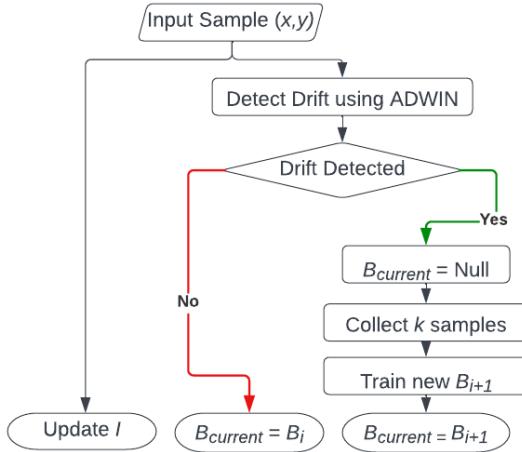


FIGURE 4.2 – CDR - Processus d'apprentissage.

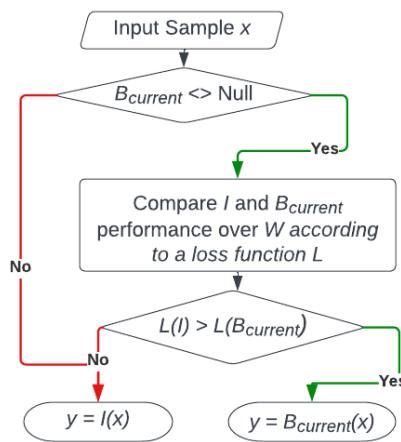


FIGURE 4.3 – CDR - Processus d'inférence.

Le processus d'inférence de CDR est présenté dans la Figure 4.3. Comme montré dans ce diagramme, lorsqu'un modèle batch B_i est invalidé et jusqu'à ce que k instances d'apprentissage soient collectées pour former un nouveau modèle batch B_{i+1} , seul le modèle incrémental I est utilisé pour l'inférence. Lorsqu'un modèle batch $B_{current}$ est disponible (soit parce que B_i n'est pas invalidé ou bien parce que B_i a été invalidé, mais l'apprentissage de B_{i+1} a pris fin), CDR compare les performances prédictives de $B_{current}$ et I pour savoir lequel utiliser pour l'inférence. Pour ce faire, CDR maintient une fenêtre temporelle glissante contenant les instances les plus récentes, *i.e.* les instances arrivées durant les W dernières unités de temps. Le modèle utilisé pour l'inférence à t est celui dont les prédictions sont les plus précises sur la période $[t - W, t]$ au regard de la fonction de perte.

4.4.4 Résultats & discussion

Pour évaluer CDR et EFRT-DD, nous avons considéré le jeu de données *Coronavirus Pandemic*, publiquement accessible sur le site d'*Our World in Data* (OWID) [CGL22]. La tâche de régression que nous considérons est la prédiction du nombre quotidien de nouveaux cas et de nouveaux décès par million d'habitants (désignés resp. par *new_cases* et *new_deaths* dans la suite), en fonction du nombre quotidien de cas par million d'habitants. Pour prédire le nombre de cas/décès à j , nous considérons le

nombre de cas signalés à 8 points dans le temps : $j - 1$ semaine, $j - 2$ semaines, ..., $j - 8$ semaines. Le jeu de données obtenu couvre la période allant du 28 mars 2020 au 30 novembre 2022 et contient ≈ 177 k instances. Lorsqu'il est restreint aux données de la Tunisie, le jeu contient ≈ 910 instances.

EFRT-DD et CDR ont été implémentés avec MOA [BGHP18], Scikit-Multiflow [MRBA18] et Scikit-Learn [PVG+11a]. CDR a été configuré pour utiliser EFRT-DD pour l'apprentissage incrémental et l'arbre de décision [PVG+11b] pour l'apprentissage batch. Les performances des modèles batch et des modèles incrémentaux ont été comparées sur une fenêtre temporelle glissante d'une semaine. La précision des modèles a été évaluée avec l'erreur absolue moyenne (*Mean Absolut Error* ou MAE) et l'erreur quadratique moyenne (*Root Mean Squared Error* ou RMSE).

4.4.4.1 Résultats

CDR/EFRT-DD vs. modèles incrémentaux Dans cette première série d'expérimentations, nous comparons les performances de CDR et de l'EFRT-DD à celles des arbres incrémentaux existants. Les modèles incrémentaux s'évaluent généralement selon une stratégie dite *Psequential* ou *Interleaved test-then-train*. Cette stratégie traite les données séquentiellement selon leur ordre d'arrivée. Chaque nouvelle instance est d'abord utilisée pour évaluer les performances du modèle, puis la même instance est utilisée pour l'apprentissage. La précision du modèle est ainsi évaluée de manière incrémentale qui sied aux situations où l'ordre d'arrivée des données est important et aux situations où il est impossible de disposer de toutes les données préalablement à l'apprentissage.

| | Daily New Confirmed Deaths | | | Daily New Confirmed Cases | | |
|---------|----------------------------|-----------------|----------------|---------------------------|----------------|---------------|
| | Time | $RMSE_{Deaths}$ | MAE_{Deaths} | Time | $RMSE_{Cases}$ | MAE_{Cases} |
| HT | 12m31s | 2.01 | 1.21 | 5m28s | 174.1 | 91.51 |
| HAT | 11m36s | 2.03 | 1.23 | 6m19s | 150.12 | 80.67 |
| FIMT-DD | 8m27s | 1.72 | 1.01 | 6m8s | 137.75 | 74.53 |
| EFRT-DD | 44m27s | 1.68 | 0.92 | 39m56s | 132.27 | 71.54 |
| CDR | - | 1.66 | 0.88 | - | 126.86 | 68.11 |

TABLE 4.3 – MAE et RMSE réalisés par les modèles incrémentaux (tous pays confondus)

| | Daily New Confirmed Deaths | | | Daily New Confirmed Cases | | |
|---------|----------------------------|-----------------|----------------|---------------------------|----------------|---------------|
| | Time | $RMSE_{Deaths}$ | MAE_{Deaths} | Time | $RMSE_{Cases}$ | MAE_{Cases} |
| HT | 4.89s | 2.47 | 2.44 | 2.14s | 61.56 | 58.71 |
| HAT | 4.45s | 2.46 | 2.43 | 1.42s | 59.35 | 56.68 |
| FIMT-DD | 4.28s | 2.27 | 2.23 | 1.56s | 56.50 | 52.29 |
| EFRT-DD | 22.23s | 1.02 | 0.83 | 9.38s | 49.64 | 41.06 |
| CDR | - | 0.98 | 0.74 | - | 44.37 | 36.83 |

TABLE 4.4 – MAE et RMSE réalisés par les modèles incrémentaux (Tunisie)

4.4. Régression en ligne adaptative aux drifts - cas de l'évolution des pandémies

Les résultats de cette première série d'expérimentations sont synthétisés dans les Tables 4.3 et 4.4 et partiellement illustrés dans les Figures 4.4 et 4.5. Comme le montre la Table 4.3, EFRT-DD et FIMT-DD surclassent de loin HAT (*Hoeffding Adaptive Tree*) et HT (*Hoeffding Tree*). En comparaison avec FIMT-DD, EFRT-DD permet une amélioration de 2,38%, 9,78%, 4,14% et 4,18% (*resp.* 122,55%, 168,67%, 13,82% et 27,35%) au regard des métriques $RMSE_{Deaths}$, MAE_{Deaths} , $RMSE_{Cases}$ et MAE_{Cases} sur les données confondues de tous les pays (*resp.* les données d'un seul pays). De même, par rapport au FIMT-DD, CDR permet une amélioration de 3,61%, 14,77%, 8,58% et 9,43% (*resp.* 131,63%, 149%, 27,34% et 41,98%) au regard des métriques $RMSE_{Deaths}$, MAE_{Deaths} , $RMSE_{Cases}$ et MAE_{Cases} sur les données confondues de tous les pays (*resp.* les données d'un seul pays).

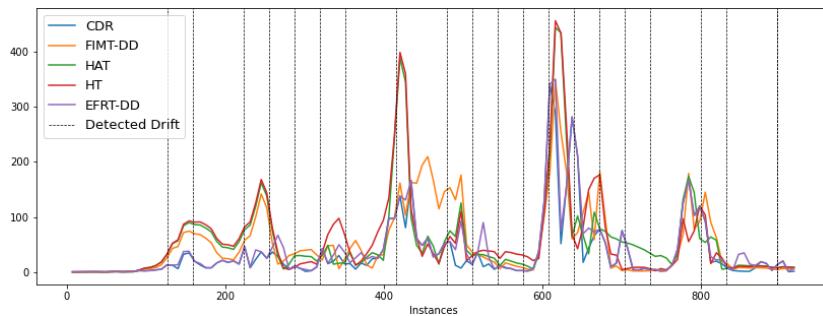


FIGURE 4.4 – Nombre de cas quotidien - MAE réalisé par EFRT-DD, CDR et les modèles incrémentaux (Tunisie).

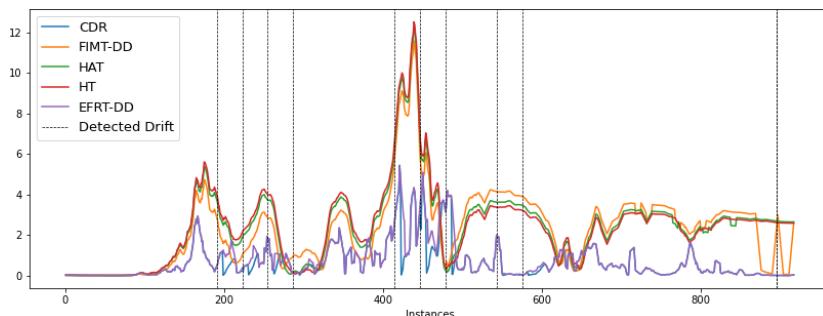


FIGURE 4.5 – Nombre de décès quotidien - MAE réalisé par EFRT-DD, CDR et les modèles incrémentaux (Tunisie).

La contrepartie des bonnes performances prédictives de l'EFRT-DD est sa relative lenteur. En moyenne, EFRT-DD est de 5 à 6 fois plus lent que FIMT-DD. Ceci s'explique par le fait que EFRT-DD révise en continu ses décisions de division pour ajuster le modèle et remplacer les divisions s'avérant non optimales. Ces révisions ralentissent le processus d'apprentissage mais sont essentielles à l'amélioration des performances prédictives du modèle.

CDR vs. Arbre de Régression Batch Cette deuxième série d'expériences compare les performances de CDR à celles d'un arbre de décision batch. L'évaluation *Prequential*, typiquement utilisée en apprentissage incrémental, peut s'appliquer à l'apprentissage batch en répétant les cycles d'apprentissage/évaluation [MKS23]. À chaque cycle, les données sont divisées en un ensemble d'apprentissage et un ensemble de test, tout en respectant l'ordre chronologique. Les instances testées avec le modèle batch B_i sont ensuite ajoutées à l'ensemble d'apprentissage du modèle suivant, B_{i+1} . Ce processus continue jusqu'à épuisement des données. Nous reprenons ici la terminologie de [MKS23], où chaque cycle est appelé *milestone*. Nous adoptons également un scénario réaliste où un nouveau modèle batch est ré-appris tous les ≈ 3 mois, ce qui résulte en 9 modèles batch (*i.e.* 9 *milestones* : M_1, M_2, \dots, M_9).

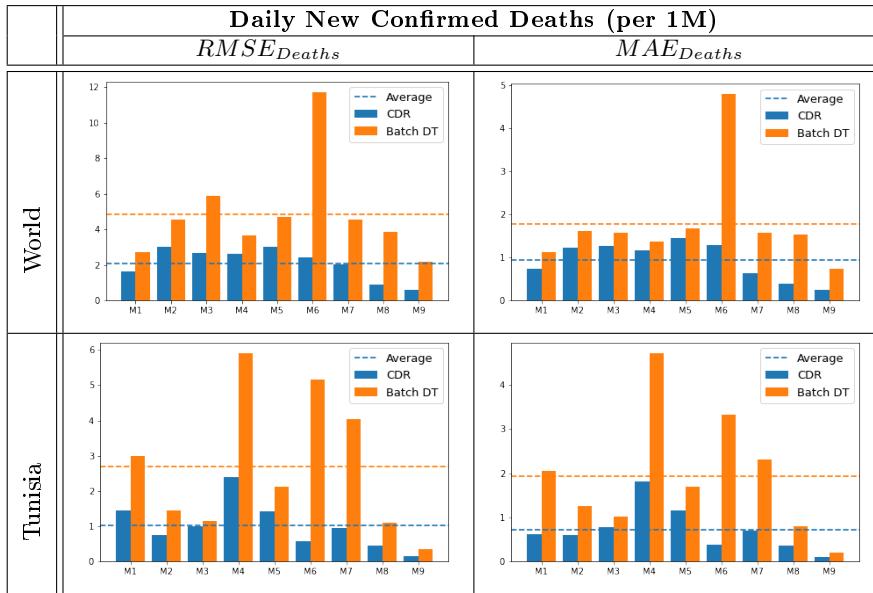


TABLE 4.5 – Nombre de décès quotidien - MAE et RMSE réalisés par CDR et l'arbre Batch sous-jacent.

Les résultats de cette deuxième série d'expérimentations sont présentés dans les Tables 4.5 et 4.6. Comme illustré dans ces tables, CDR surpasse largement le modèle batch et réalise une amélioration moyenne de 131,61%, 89,54%, 63,23% et 21,85% (*resp.* 165,78%, 167,08%, 137,88% et 93,06%) au regard des métriques $RMSE_{Deaths}$, MAE_{Deaths} , $RMSE_{Cases}$ et MAE_{Cases} sur les données confondues de tous les pays (*resp.* les données d'un seul pays). Les résultats expérimentaux de la première et de la seconde série d'expérimentations montrent que notre approche collaborative dirigée par les dérives de concept, permet de meilleurs résultats que ceux obtenus par chacun des modèles sous-jacents pris séparément. Les résultats expérimentaux confirment également que, dans le cas particulier de la prédition de l'évolution d'une pandémie, une approche de ré-apprentissage dirigée par les dérives permet de meilleures performances prédictives que le ré-apprentissage à intervalles fixes.

4.4. Régression en ligne adaptative aux drifts - cas de l'évolution des pandémies

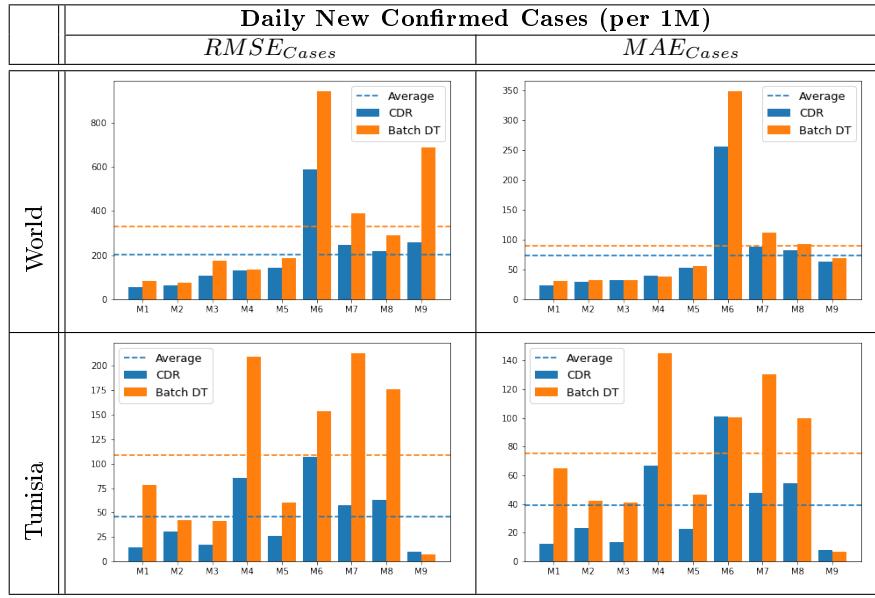


TABLE 4.6 – Nombre de cas quotidien - MAE et RMSE réalisés par CDR et l'arbre Batch sous-jacent.

4.4.4.2 Discussion

Les résultats de notre étude expérimentale permettent de tirer les conclusions importantes ci-après.

- Induction Eager (active) vs. Induction Lazy (conservatrice).** La majorité des arbres incrémentaux de régression adoptent une stratégie d'induction conservatrice, où la division d'un nœud est retardée jusqu'à ce que l'algorithme soit suffisamment confiant que l'actuel meilleur attribut de division, restera toujours une meilleure option que l'actuel second meilleur attribut. Dans EFRT-DD nous adoptons une stratégie active, dans laquelle un nœud est divisé aussitôt que l'on est suffisamment confiant que la division améliore l'homogénéité des feuilles. Comme le montre la table 4.3, dans le cas particulier de la prédiction de l'évolution d'une pandémie, la stratégie *Eager* s'avère être meilleure que la stratégie *Lazy*. Ceci s'explique par la nature abrupte des dérives de concept affectant les courbes de progression de la COVID-19 qui requiert une adaptation rapide.
- Ré-apprentissage sur les données récentes vs. ré-apprentissage sur toutes les données.** Ré-apprendre un modèle batch sur l'ensemble de toutes les données disponibles est utile dans les cas où il est nécessaire d'apprendre des concepts stables ou pour s'adapter à des dérives récurrentes. Dans le cas particulier de la prédiction de l'évolution d'une pandémie, nous constatons que le ré-apprentissage d'un modèle uniquement sur les données récentes s'avère être une meilleure stratégie. Ceci peut être observé dans les Tables 4.5 et 4.6, où au fil du temps, CDR creuse de plus en plus l'écart avec le modèle batch. Nous expliquons ceci par le fait que le

modèle batch est (ré)-appris sur des ensembles de données avec une proportion décroissante de données récentes, ce qui le rend, au fil du temps, de moins en moins sensible aux changements.

3. **Apprentissage sur les données d'un seul pays vs. apprentissage sur les données de plusieurs pays.** Comme on peut l'observer dans les Tables 4.3, 4.5 et 4.6, les améliorations de CDR par rapport aux modèles batch et aux modèles incrémentaux, sont plus conséquentes lorsque l'apprentissage se fait sur les données provenant d'un seul pays. Cela est principalement dû à l'occurrence des dérives à des moments différents d'un pays à un autre. Utiliser des données provenant de différents pays, résulte en conséquence à un modèle qui ne représente de manière précise aucun de ces pays.

4.5 Conclusion

Les travaux de recherche présentés dans ce chapitre ont exploré à la fois la classification et la régression incrémentales adaptatives aux dérives de concept dans les environnements dynamiques que sont la cybersécurité et l'épidémiologie. Tout comme les intrusions ajustent leurs tactiques pour contourner les mécanismes de sécurité, les virus biologiques adaptent leurs stratégies pour échapper aux défenses immunitaires. Nous avons proposé les approches DDM-ORF pour la détection en ligne des intrusions et EFRT-DD/CDR pour la prédition de l'évolution des pandémies. Ces approches se distinguent par leur capacité à s'adapter de manière dynamique et automatisée aux dérives de concept, principale source de dégradation des performances dans les environnements dynamiques.

Nos travaux sur l'apprentissage en ligne se poursuivent vers l'exploration de nouvelles stratégies permettant de s'adapter aux dérives subtiles, difficiles à détecter avec les approches basées sur le fenêtrage. Nous nous intéressons notamment aux facteurs d'oubli (*Fading Factors*) qui permettent d'accorder une pondération décroissante aux observations selon leur ancienneté, favorisant ainsi l'adaptation aux nouvelles tendances. Les facteurs d'oubli peuvent également être appliqués aux méthodes ensemblistes comme ORF, en accordant une importance croissante aux prédictions des nouveaux arbres et en élargissant progressivement les anciens. Cette approche permet de s'adapter aux dérives tout en optimisant l'utilisation des ressources matérielles.

Publications en lien avec le chapitre

Revues internationales

[JJK25b] Intrusion Detection based on Concept Drift Detection & Online Incremental Learning

Farah JEMILI, Khaled JOUINI & Ouajdi KORBAA

International Journal of Pervasive Computing and Communications. ISSN : 1742-7371, Emerald group publishing ltd. DOI : <https://doi.org/10.1108/IJPCC-12-2023-0358>. 2025.

SJR best quartile : Q2, SJR : 0.36, JCR IF (2023) : 0.6.

[JK25a] Detecting Unknown Intrusions from Large Heterogeneous Data through Ensemble Learning

Farah JEMILI, Khaled JOUINI & Ouajdi KORBA

Intelligent Systems with Applications (ISWA 200465). ISSN : 26673053, Elsevier B.V., DOI : <https://doi.org/10.1016/j.iswa.2024.200465>. 2025.

SJR best quartile : Q1, SJR : 0.96

Conférences internationales

[JK23] Drift-Driven Regression for Predicting the Evolution of Pandemics

Khaled JOUINI & Ouajdi KORBA

IADIS International Conference Applied Computing (IADIS AC). pp. 77-84, ISBN : 978-989-8704-53-5. DOI : https://doi.org/10.33965/ICWI_AC_2023_202307L009. 2023.

CORE Rank : C.

[JMK21] Real-Time, CNN-Based Assistive Device for Visually Impaired People

Khaled JOUINI, Mohamed Hédi MAALOUL & Ouajdi KORBA

International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (IEEE CISP-BMEI). pp. 1-6, DOI : <https://doi.org/10.1109/CISP-BMEI53629.2021.9624387>. 2021.

Encadrement de masters de recherche

- Mouna MZIOU (Soutenu)
- Marwa GHEZAIEL (Soutenu)

CHAPITRE 5

Vérité : Détection des Fake News dans les médias sociaux

Sommaire

| | | |
|------------|---|-----------|
| 5.1 | Introduction | 65 |
| 5.2 | Contributions à la détection d'opinions | 65 |
| 5.2.1 | <i>Stance Detection vs. Fake News Detection</i> | 66 |
| 5.2.2 | Apprentissage ensembliste basé sur l'augmentation | 67 |
| 5.2.3 | Résultats & discussion | 69 |
| 5.3 | De la <i>Stance</i> à la <i>Fake News Detection</i> : la Blockchain pour combler le chaînon manquant | 72 |
| 5.3.1 | La Blockchain au service de la vérification des faits : principales approches | 73 |
| 5.3.2 | Vérification décentralisée des faits, basée sur la Blockchain, la Stance Detection et la borne de Hoeffding | 74 |
| 5.3.3 | Résultats & discussion | 79 |
| 5.4 | Conclusion | 82 |

5.1 Introduction

Après avoir exploré la dimension temporelle du Big Data à travers la vélocité, nous nous tournons vers une dimension qualitative en examinant la véracité des données.

À l'ère de l'Internet ubiquitaire et des plateformes de médias sociaux, où des informations de divers types sont accessibles partout, à tout moment et à tous et où tout point de vue peut trouver un auditoire, le défi n'est plus la démocratisation de l'information, mais désormais la véracité et l'authenticité de l'information. Évaluer la véracité d'une information en temps opportun est toutefois une tâche fastidieuse et extrêmement complexe, non pas uniquement pour les algorithmes d'apprentissage automatique, mais également parfois pour les journalistes les plus expérimentés. Consciente de la difficulté de la tâche, la communauté scientifique l'aborde sous différents angles, dépassant les solutions technologiques singulières et décomposant le processus en sous-tâches indépendantes [HPS⁺18]. Une première étape pratique vers la vérification automatisée des faits consiste à inférer l'opinion ou le point de vue (*i.e. Stance*) de différentes sources sur une même information. Dans nos travaux, la *détection d'opinions (Stance Detection)* est un composant clé d'un pipeline automatisé de vérification des faits, construit autour de la technologie blockchain.

Grâce à sa scalabilité et à sa nature décentralisée lui permettant de s'affranchir de toute autorité centrale, la technologie blockchain est de plus en plus utilisée pour lutter contre la désinformation [KJKU23]. Les approches existantes basées sur la blockchain se sont principalement concentrées sur la traçabilité et le suivi des modifications apportées aux informations [ans, pro, AR22, WXJ⁺23, CDPG22, SRLB⁺21, YU23]. Bien que ces efforts aient apporté des avancées précieuses, nous estimons que la simple traçabilité n'exploite pas pleinement le potentiel de la blockchain. Dans notre travail, nous franchissons une étape supplémentaire en nous plaçant dans le contexte d'un réseau social décentralisé et en proposant un framework où la blockchain comble le chaînon manquant entre la détection d'opinions et la détection des fausses informations. L'objectif ultime de ce framework est de dériver, de manière décentralisée, dynamique et automatisée de bout en bout, un score de crédibilité pour les informations, basé sur les opinions (*e.g.* accord/désaccord) de diverses sources, les opinions étant pondérées par la fiabilité (*i.e. réputation*) respective de ces sources.

La suite de ce chapitre commence par présenter nos contributions relatives à la détection d'opinions. Ensuite, nous présentons notre approche mettant en synergie la détection d'opinions et la technologie blockchain pour la lutte à grande échelle contre la désinformation.

5.2 Contributions à la détection d'opinions

Plusieurs facteurs rendent la détection d'opinions et plus globalement de fausses informations extrêmement complexe, dont notamment :

- *Ambiguïté du langage naturel.* La polysémie, les homographes, le contexte etc. sont autant de facteurs qui rendent difficile la classification des données textuelles.

- *Difficultés liées à l’apprentissage.* Les fausses nouvelles ont souvent pour particularité d’avoir été créées pour paraître les plus ressemblantes possible aux informations vraies.
- *Manque de données.* La qualité et la diversité des ensembles de données sont souvent insuffisantes, limitant l’efficacité et la généralisation des modèles d’apprentissage supervisé.
- *Textes courts.* Les textes sur les réseaux sociaux sont souvent brefs, ce qui complique l’identification de régularités statistiques.
- *Déséquilibre des classes.* Les fausses informations sont moins nombreuses que les vraies, ce qui biaise les modèles vers les classes majoritaires et dégrade les performances prédictives.

L’*augmentation* est une technique consistant à synthétiser de nouvelles instances d’apprentissage à partir de celles disponibles, pour enrichir et diversifier les données. Dans nos travaux [SJK22, SJK23] nous explorons différents usages de l’*augmentation des données textuelles* pour résoudre (une partie) des problématiques susmentionnées. En particulier, nous explorons et quantifions l’effet de différentes techniques d’augmentation du texte sur les performances des algorithmes de classification usuels. Nous tirons parti des résultats de cette évaluation pour proposer une approche d’apprentissage ensembliste basée sur l’augmentation. Finalement, nous étudions expérimentalement l’utilisation de l’augmentation pour résoudre le problème déséquilibre des classes, un problème récurrent dans la détection d’opinions et des fausses nouvelles.

Dans la suite, la section 5.2.1 revoit brièvement les concepts liés à la détection d’opinions et son utilisation dans le contexte de la détection des fausses nouvelles. Puis la section 5.2.2 présente notre approche de détection d’opinions basée sur l’augmentation. La section 5.2.3 étudie expérimentalement les approches proposées et discute les principaux résultats.

5.2.1 Stance Detection vs. Fake News Detection

L’*opinion* ou la *position* est définie ici comme l’expression du point de vue ou du jugement d’une entité sur un fait, une information ou une proposition qui lui est présentée. Comme illustré dans l’exemple de la Figure 5.1, extrait du banc de tests FNC-1 [HPS+18], étant donné un titre et un corps (provenant d’un même article ou d’articles différents), la variable cible prend l’une des quatre valeurs suivantes :

- *Agree* : le corps confirme le fait annoncé dans le titre ;
- *Disagree* : le corps contredit le fait annoncé dans le titre ;
- *Discuss* : le corps discute le fait annoncé dans le titre, sans prendre position ;
- *Unrelated* : le corps aborde un sujet différent de celui du titre.

Les auteurs du challenge FNC-1 ont réalisé un modèle de base (*baseline classifier*) donnant une valeur de référence des performances attendues. Le modèle utilise les n-grammes pour représenter le texte, *Gradient Boosting* [CG16] pour l’apprentissage et réalise un F1-Score de 79,53%. Les modèles "SOLAT in the SWEN" [STVS+21], "Team Athene" [HPS+18] et "UCL Machine Reading" (UCLMR) [RASR17], ont occupé les trois premières place du challenge. "SOLAT in the SWEN" utilise une

5.2. Contributions à la détection d'opinions

| Headline: Hundreds of Palestinians flee floods in Gaza as Israel opens dams | |
|--|---|
| Agree (AGR) | GAZA CITY (Ma'an) – Hundreds of Palestinians were evacuated from their homes Sunday morning after Israeli authorities opened a number of dams near the border, flooding the Gaza Valley in the wake of a recent severe winter storm. The Gaza Ministry of Interior said in a statement that civil defense services and teams from the Ministry of Public Works had evacuated more than 80 families from both sides of the Gaza Valley (Wadi Gaza) after their homes flooded as water levels reached more than three meters [...] |
| Discuss (DSC) | Palestinian officials say hundreds of Gazans were forced to evacuate after Israel opened the gates of several dams on the border with the Gaza Strip, and flooded at least 80 households. Israel has denied the claim as "entirely false". [...] |
| Disagree (DSG) | Israel has rejected allegations by government officials in the Gaza strip that authorities were responsible for released storm waters flooding parts of the besieged area. "The claim is entirely false, and southern Israel does not have any dams," said a statement from the Coordinator of Government Activities in the Territories (COGAT). "Due to the recent rain, streams were flooded throughout the region with no connection to actions taken by the State of Israel." At least 80 Palestinian families have been evacuated after water levels in the Gaza Valley (Wadi Gaza) rose to almost three meters. [...] |
| Unrelated (UNR) | Apple is continuing to experience 'Hairgate' problems but they may just be a publicity stunt [...] |

FIGURE 5.1 – Exemple de *Stance Detection* repris du banc FNC-1 [HPS+18]

approche ensembliste basée sur un CNN et des *Gradient-Boosted Decision Trees* (GBDT). Le modèle "Team Athene" [HPS+18] qui s'est classé deuxième, utilise un ensemble composé de 5 perceptrons multicouches (*MultiLayer Perceptron* ou MLP), où les prédictions sont obtenues par vote majoritaire. Le modèle de UCLMR [RASR17], classé troisième, utilise également un classifieur MLP avec une couche softmax pour la classification. Plus récemment, d'autres approches ont utilisé FNC-1 pour valider leurs modèles [DDDW19, SA20, LMN+19, STVS+21]. L'approche proposée dans [LMN+19] est celle qui s'approche le plus de l'esprit de notre travail. [LMN+19] utilise LSA pour la réduction de dimensionnalité et un stacking comprenant cinq modèles de base : GradBoost, Random Forest (RF), XGBoost, Bagging et Light Gradient Boosting Machine (Lightgbm). Une comparaison expérimentale entre [LMN+19] et notre travail est donnée dans la sous-section 5.2.3.

Une des conclusions importantes que l'on peut tirer des travaux susmentionnés est que les modèles ensemblistes obtiennent les meilleurs résultats. Par ailleurs, en dépit du potentiel d'amélioration que peut apporter l'augmentation, à notre connaissance, il n'existe aucune étude antérieure à la nôtre qui compare les techniques d'augmentation et/ou l'utilise en conjonction avec l'apprentissage ensembliste pour la détection d'opinions.

5.2.2 Apprentissage ensembliste basé sur l'augmentation

5.2.2.1 Augmentation du texte

L'augmentation des données permet de : (i) pallier le manque d'instances d'apprentissage en offrant un moyen relativement simple et peu coûteux de collecter et d'annoter des données; (ii) améliorer la capacité de généralisation des modèles; et (iii) aider à résoudre le problème de déséquilibre des classes.

L'augmentation est très répandue en vision par ordinateur, mais reste encore peu exploitée pour les données textuelles [SKF21]. Cette différence s'explique par les propriétés intrinsèques des données textuelles (*e.g.*, polysémie, homonymie, etc.), qui rendent plus difficile la définition de transformations préservant le sens ("déformer" un texte peut en altérer le sens). Dans nos travaux, nous avons utilisé les techniques suivantes, qui permettent (dans le cas général) de préserver le sens.

- Plongement lexical (*Word Embedding*) : La technique du *plongement lexical* repose sur l'hypothèse que les mots apparaissant dans des contextes similaires ont des significations apparentées [Cou18]. À partir de cette hypothèse, le plongement associe des vecteurs aux mots, de sorte que ceux apparaissant dans des contextes similaires aient des vecteurs proches dans l'espace vectoriel. La substitution basée sur le plongement consiste à identifier les plus proches voisins d'un mot à l'aide de la similarité cosinus, puis à le remplacer par l'un de ces voisins. Dans notre travail, nous avons choisi BERT [DCLT18] comme représentant de cette technique.
- Rétro-traduction (*Back Translation*) : La rétro-traduction consiste à traduire un texte, puis à retraduire le texte obtenu vers sa langue d'origine. Si le texte original et sa version doublement traduite ne sont pas identiques, le nouveau texte est ajouté aux données d'apprentissage.
- Synonymes : Cette technique, également appelée *substitution lexicale avec dictionnaire*, consiste à remplacer des mots choisis aléatoirement par des synonymes tirés d'une base de données lexicale (*i.e.* dictionnaire de synonymes).
- Augmentation basée sur le vecteur TF-IDF : Deux techniques d'augmentation basées sur le TF-IDF ont été utilisées :
 - Substitution : Proposée dans [XDH⁺19], cette technique consiste à remplacer les mots ayant de faibles scores TF-IDF par d'autres mots ayant des scores similaires.
 - Injection : Cette technique ajoute au texte des mots ayant de faibles scores TF-IDF. L'idée derrière ces techniques est que les mots ayant de faibles scores TF-IDF sont peu informatifs et peuvent être substitués ou injectés sans altérer la classification.

5.2.2.2 Apprentissage ensembliste

Les méthodes d'ensemble (*i.e.* apprentissage ensembliste) reposent sur le phénomène de la "sagesse de la foule" (*i.e.* *Wisdom of the Crowd*) [Sur05] qui voudrait qu'"*un grand nombre d'amateurs peut mieux répondre à une question qu'un seul expert*". Pour que le principe de la "sagesse de la foule" puisse s'appliquer, il est primordial que les membres de la foule aient un minimum de compétence et des opinions diversifiées. Dans nos travaux de recherche, nous proposons une nouvelle approche d'apprentissage ensembliste basée sur l'augmentation et à mi-chemin entre le Bagging et le Stacking. La méthodologie globale que nous suivons est comme suit. Nous menons d'abord une étude expérimentale approfondie pour identifier les paires (algorithme d'apprentissage, technique d'augmentation) les plus "compétentes", *i.e.* celles permettant les meilleures performances prédictives. Pour les besoins de cette étude nous avons considéré dix des méthodes d'apprentissage les plus courantes, à savoir : Decision

Tree, SVM, Random Forest, lightGBM, Logistic Regression, Gradient Boosting, eXtreme Gradient Boost, Adaptative Boosting, Bootstrap Aggregation et Naive Bayes. Ensuite, nous distinguons deux stratégies possibles pour combiner augmentation des données et Stacking, une classique et une autre que nous proposons.

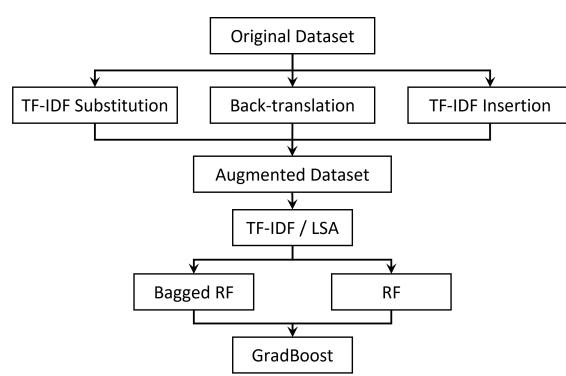


FIGURE 5.2 – Stratégie classique : le même ensemble de données est utilisé pour générer les modèles de base.

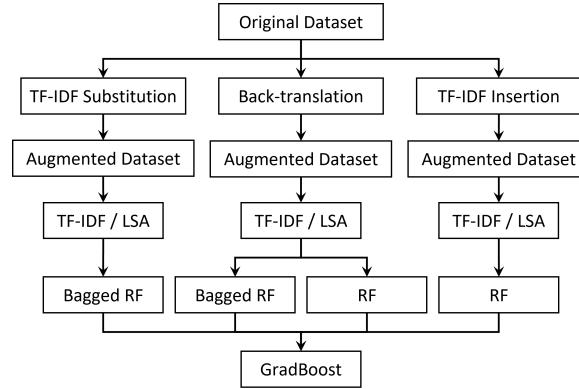


FIGURE 5.3 – Stratégie proposée : différents ensembles d'apprentissage sont utilisés pour générer les modèles de base.

Dans la stratégie classique (Figure 5.2), les données originales et celles obtenues par augmentation, sont d'abord fusionnées dans un même ensemble d'apprentissage. Puis ces données sont fournies à des algorithmes d'apprentissage hétérogènes pour former les modèles de base. Dans la stratégie que nous proposons (Figure 5.3), des algorithmes hétérogènes apprennent sur des ensembles de données différents pour promouvoir davantage la diversité. Chacun de ces ensembles est obtenu par l'application d'une technique différente d'augmentation (en plus des données originales). Cette stratégie se situe entre le Bagging et le Stacking. Comme le Stacking, elle n'utilise pas l'échantillonnage et construit des modèles de base avec des algorithmes d'apprentissage hétérogènes. Comme le Bagging et contrairement au Stacking, chaque modèle de base est appris sur un ensemble de données différent. Notre approche hérite ainsi des avantages du bagging et du stacking en diversifiant les apprentissages à la fois sur les données et sur les algorithmes.

5.2.3 Résultats & discussion

5.2.3.1 Résultats

Les différentes techniques et approches comparées ont été implémentées avec le langage Python et les bibliothèques NLTK (*Natural Language ToolKit*) [NLT], SciKit-Learn (version 0.24.2) [PVG⁺11a] et NLPAug [Ma19]. Les expérimentations ont été faites sur les bancs de tests FNC [PR17] et FNN [Shu19]. Par souci de concision, seuls les principaux résultats sont rapportés.

5.2. Contributions à la détection d'opinions

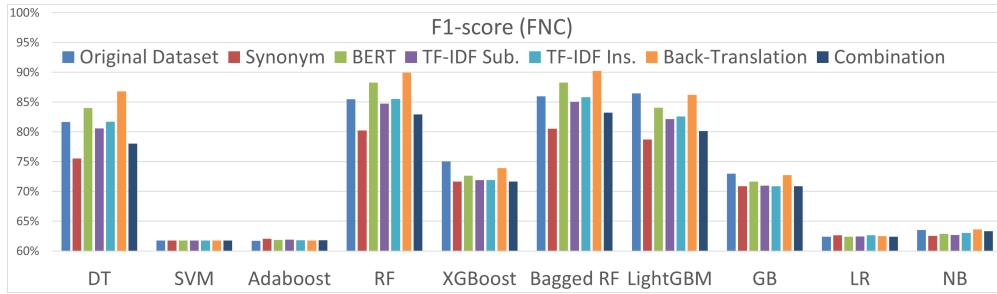


FIGURE 5.4 – F1-scores obtenus sur le banc FNC-1 avec et sans augmentation du texte

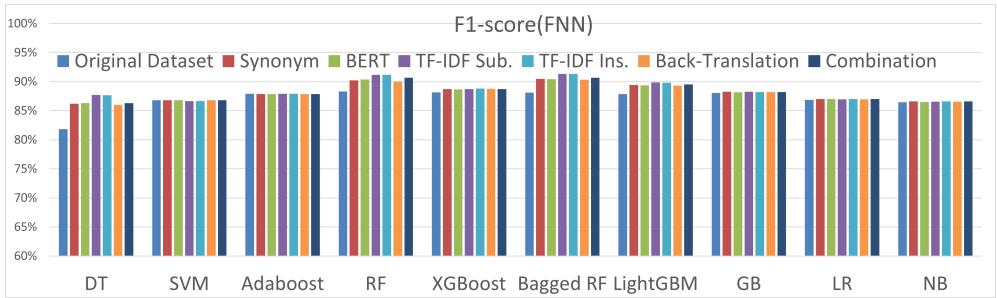


FIGURE 5.5 – F1-scores obtenus sur le banc de tests FNN avec et sans augmentation du texte

Paires compétentes Dans cette première série d’expérimentations nous évaluons les gains en précision permis par l’augmentation du texte sur 10 algorithmes d’apprentissage courants. Comme le montre la Figure 5.4, les associations (*BootStrap Aggregation, Back-translation*) et (*Random Forest, Back-translation*) permettent les meilleures performances prédictives sur le dataset FNC-1. L’amélioration permise par ces associations par rapport aux meilleures valeurs que l’on peut atteindre sans augmentation est de $\approx 4.16\%$ sur le F1-score. La Figure 5.5 montre que les associations (*Random Forest, Substitution TF-IDF*) et (*BootStrap Aggregation, Insertion TF-IDF*) permettent les meilleures performances sur le dataset FNN. L’amélioration permise par ces associations par rapport aux meilleures valeurs que l’on peut atteindre sans augmentation est de l’ordre de $\approx 5.87\%$ sur le F1-score.

Apprentissage ensembliste L’idée principale derrière notre approche ensembliste est de tirer profit de l’augmentation pour améliorer à la fois la diversité et la compétence des modèles de base. Comme illustré dans la Figure 5.3 nous utilisons les algorithmes *Bagged RF* et *Random Forest* pour former nos modèles de base et *GradBoost* pour le méta-modèle. Nous comparons notre modèle à une approche de stacking plus classique, où tous les modèles de base sont appris sur le même ensemble de données, composé des données d’origine et de celles obtenues par augmentation (Figure 5.2). Nous comparons également notre modèle à l’approche ensembliste proposée dans [LMN⁺19]. De même que notre approche, [LMN⁺19] utilise LSA pour la réduction de la dimensionnalité et le stacking pour l’apprentissage ensembliste. Comme montré dans la Table 5.1, à l’exception de la méthode Back-translation

5.2. Contributions à la détection d'opinions

| Modèle | F1-Score | | Accuracy | |
|---------------------------------|---------------|---------------|---------------|---------------|
| | FNC | FNN | FNC | FNN |
| (TF-IDF Insertion, Stacking) | 85,58% | 90,92% | 86,62% | 85,67% |
| (TF-IDF Substitution, Stacking) | 84,57% | 90,43% | 85,78% | 84,72% |
| (Back-Translation, Stacking) | 90,31% | 89,80% | 89,70% | 83,47% |
| (BERT, Stacking) | 87,93% | 90,26% | 88,62% | 84,23% |
| (Synonym, Stacking) | 80,71% | 90,28% | 82,71% | 84,17% |
| (Combinaison, Stacking) | 83,11% | 90,73% | 84,98% | 85,28% |
| [LMN ⁺ 19] | 83,72% | 88,45% | 83,67% | 79,31% |
| Notre approche [SJK22, SJK23] | 90,15% | 91,07% | 90,67% | 85,78% |

TABLE 5.1 – Accuracy et F1-Score obtenus par le stacking conventionnel, [LMN⁺19] et notre approche

sur le jeu de données FNC-1, l'approche que nous proposons surclasse l'approche de stacking classique dans tous les cas de figure. Par rapport à [LMN⁺19], notre approche permet une amélioration du F1-Score et de l'Accuracy de 7,72 % et 7,13 % (resp. 7,54 % et 2,88 %) sur FNC-1 (resp. FNN).

| Dataset | # Titres | # Articles | # Instances | # Agree | # Disagree | # Discuss | # Unrelated |
|---------|----------|------------|-------------|---------|------------|-----------|-------------|
| FNC-1 | 2,587 | 2,587 | 75,385 | 7,4% | 2,0% | 17,7% | 72,8% |

TABLE 5.2 – Distribution des classes dans FNC-1 [HPS⁺18]

| Model | Original dataset | | | | Balanced FNC | | | |
|----------------|------------------|----------|---------|-----------|---------------|---------------|---------|-----------|
| | Agree | Disagree | Discuss | Unrelated | Agree | Disagree | Discuss | Unrelated |
| DT | 40,70% | 28,49% | 68,80% | 89,95% | 69,47% | 84,52% | 59,06% | 81,13% |
| SVM | 0% | 0% | 33,51% | 81,36% | 36% | 38,89% | 21,48% | 63,27% |
| AdaBoost | 0% | 0% | 0,77% | 84,30% | 31,26% | 42,30% | 6,08% | 59,09% |
| RF | 45,99% | 27% | 74,74% | 92,53% | 78,11% | 89,53% | 68,08% | 87,11% |
| XGBoost | 1,85% | 0% | 41,08% | 86,98% | 43,34% | 45,28% | 34,10% | 66,14% |
| Bagged RF | 45,68% | 28,87% | 72,61% | 91,94% | 74,56% | 87,25% | 66,09% | 84,77% |
| LightGBM | 14,65% | 8,56% | 57,81% | 88,99% | 54,40% | 74,54% | 47,30% | 75,13% |
| GradBoost | 6,17% | 3,45% | 46,96% | 87,56% | 44,87% | 54,85% | 39,81% | 68,12% |
| LR | 0% | 0% | 2,42% | 84,50% | 1,01% | 28,20% | 2,86% | 59,04% |
| NB | 0% | 0% | 24,70% | 80,92% | 15,27% | 28,87% | 18,83% | 57,75% |
| [SJK22, SJK23] | 47,78% | 26,34% | 75,35% | 92,64% | 77,23% | 89,26% | 65,85% | 86,82% |

TABLE 5.3 – F1-Score par classe avec et sans équilibrage des classes par augmentation (FNC-1)

Déséquilibre des classes Le problème de déséquilibre des classes survient lorsqu'une modalité de la variable cible présente une faible proportion dans les données par rapport aux autres modalités. Comme noté dans [FGG⁺18] : (i) *Les classes minoritaires sont souvent plus difficiles à prédire* : Ceci s'explique par le fait que les algorithmes d'apprentissage sont biaisés vers les classes majoritaires, puisqu'ils essaient généralement de maximiser le taux global de bon classement ; et (ii) *Les classes*

minoritaires sont généralement les plus importantes à prédire et à identifier correctement : Ceci signifie que la capacité d'un modèle à prédire correctement une classe minoritaire est souvent plus importante que celle à prédire correctement une classe majoritaire.

Le biais vers les classes majoritaires peut être observé dans la Table 5.3, où tous les algorithmes ont de mauvaises performances sur les classes minoritaires "Agree" et "Disagree". Pour équilibrer le jeu de données FNC-1, nous avons généré 18 390 instances supplémentaires pour la classe "Agree" et 4 200 pour la classe "Disagree". Comme le montre la Table 5.3, l'augmentation permet une amélioration substantielle des F1-scores obtenus sur les classes minoritaires : une amélioration moyenne de 189,14 % pour la classe "Agree" et de 586,39 % pour la classe "Disagree".

5.2.3.2 Discussion

Les résultats de notre étude mettent en évidence plusieurs aspects importants de l'augmentation des données dans la détection d'opinions et de fausses nouvelles :

- *L'augmentation des données n'améliore pas toujours les performances prédictives.* Certains algorithmes (SVM, XGBoost, GradBoost sur FNC-1 et AdaBoost sur FNN) montrent de meilleures performances sans augmentation.
- *Pas de solution universelle (one-size-fits-all).* Les techniques d'augmentation se comportent différemment selon les algorithmes. Par exemple, la substitution par synonymes est efficace avec AdaBoost, mais se révèle être la moins performante avec Naive Bayes.
- *Le motto “the more, the better” n'est pas la bonne approche au regard de l'augmentation.* Combiner toutes les techniques d'augmentation ne garantit pas de meilleures performances. Des associations ciblées donnent souvent de meilleurs résultats.
- *L'augmentation est un levier essentiel pour améliorer la diversité et la précision des modèles ensemblistes.* Ceci est confirmé par les excellents résultats de notre approche ensembliste.
- *L'augmentation s'avère particulièrement efficace pour remédier au déséquilibre des classes.* Le déséquilibre de classe est un problème fréquent en détection d'opinions. Notre étude a permis de quantifier et de montrer dans quelle mesure l'augmentation peut atténuer ses effets et améliorer les performances prédictives sur les classes minoritaires.

5.3 De la *Stance* à la *Fake News Detection* : la Blockchain pour combler le chaînon manquant

Nos travaux sur la détection d'opinions nous ont révélé un besoin crucial : développer des méthodologies permettant de relier de manière systématique la détection d'opinions à la vérification de la véracité des informations. La plupart des travaux existants se concentrent en effet sur l'une ou l'autre des tâches, sans établir de lien pratique entre elles. Dans les travaux présentés dans cette section, nous proposons d'utiliser la technologie blockchain pour combler ce chaînon manquant.

Dans son article introduisant la blockchain, le premier argument avancé par son inventeur Satoshi Nakamoto¹ [Nak09] pour promouvoir cette technologie, était le souhait de créer un système de paiement électronique décentralisé, opérant *sans intermédiaire de confiance* et *sans autorité centrale* régulatrice :

"What is needed is an electronic payment system ... allowing any two willing parties to transact directly with each other **without the need for a trusted third party**"

Satoshi Nakamoto [Nak09]

L'idée de recourir à la blockchain est née dans notre esprit du constat qu'il est impossible dans la pratique de mettre en place une autorité centrale qui attesterait de la véracité de toutes les informations partagées sur Internet ou sur un réseau social ; et quand bien même si elle existait, le pouvoir qu'elle détiendrait augmenterait les risques de biais, de manipulation et de censure. Face à cette impossibilité, la blockchain apparaît comme une solution pertinente pour décentraliser le processus de vérification et permettre une validation de l'information, collective, sécurisée et affranchie de toute autorité centrale.

Plus explicitement, dans nos travaux nous proposons un framework décentralisé de vérification des faits, dont l'objectif est d'attribuer des *scores de réputation* aux participants et des *scores de crédibilité* aux informations. La crédibilité d'une information est estimée selon les opinions prises par les participants interagissant avec elle, les opinions étant pondérées par la réputation des sources. La réputation d'un participant est dynamique et évolue en fonction de la pertinence de ses actions passées : elle augmente lorsqu'il valide des informations véridiques ou réfute de fausses informations, et diminue dans le cas contraire. Afin de se prémunir des acteurs malveillants, notre framework adapte le mécanisme de consensus *Proof of Reputation* à la détection de fausses informations. Ainsi, seules les opinions des participants fiables sont prises en compte. Pour permettre une détection en temps opportun et gérer l'incertitude liée à l'arrivée des opinions, notre framework utilise la *borne de Hoeffding* [Hoe63] comme garantie statistique pour labelliser les informations.

Dans la suite, la première section expose brièvement les limites des approches utilisant la blockchain pour la lutte contre la désinformation. La deuxième section présente notre approche décentralisée de vérification des faits. La troisième section étudie expérimentalement le framework proposé.

5.3.1 La Blockchain au service de la vérification des faits : principales approches

Grâce à sa nature décentralisée et sécurisée, la technologie blockchain est de plus en plus utilisée pour lutter contre la désinformation [KJKU23]. Les premiers efforts se sont principalement concentrés sur la traçabilité de l'information et l'incitation à la création de contenus fiables. Parmi les systèmes notables de cette catégorie figurent BlockProof [AR22], AnsaCheck [ans] et le projet de provenance du NY Times [pro]. S'appuyant sur plusieurs études démontrant l'efficacité de la sagesse collective

1. Nom d'emprunt de l'inventeur du Bitcoin, inconnu à ce jour.

pour limiter les biais [DMG20, SRLB⁺21], des approches plus récentes ont intégré le crowdsourcing et les modèles d'apprentissage automatique [MC23]. Ceci inclut notamment les approches [CDPG22] et [YU23].

L'approche proposée dans [CDPG22] introduit un réseau social décentralisé utilisant un mécanisme d'incitation pour encourager les utilisateurs à promouvoir la véracité des informations partagées via un mécanisme de vote. L'algorithme de sélection des évaluateurs repose sur l'entropie et privilégie ceux dont les votes sont les moins prévisibles, maximisant ainsi la diversité des opinions et la robustesse. Si une majorité d'évaluateurs juge qu'une nouvelle est fausse, celle-ci est signalée comme telle et sa visibilité est réduite. Les utilisateurs ont la possibilité d'évaluer les évaluateurs en fonction de la pertinence et de l'utilité de leurs contributions, influençant ainsi leur score de fiabilité. La fiabilité d'un évaluateur est déterminée par plusieurs critères : la précision de ses votes, la variance de ses votes et sa réputation, elle-même calculée à partir des évaluations des autres utilisateurs. L'inconvénient majeur de l'approche [CDPG22] est qu'elle repose sur des votes manuels d'un groupe restreint d'évaluateurs jugés experts. Ces évaluateurs constituent une forme d'autorité centrale, ce qui va à l'encontre de l'esprit et de la philosophie de la blockchain et limite le système en termes de scalabilité et de neutralité.

Dans [YU23], les auteurs proposent un environnement décentralisé où les utilisateurs évaluent collectivement la véracité des informations. L'atteinte du consensus au sein de la communauté de votants est déterminée à l'aide des exposants de Lyapunov et de l'entropie de Shannon. La véracité d'une information est évaluée à l'aide d'un modèle neuronal à deux niveaux, composé d'un classificateur d'actions et d'un classificateur de nouvelles. Le classificateur d'actions, implémenté à l'aide d'un CNN, est conçu pour identifier les utilisateurs malveillants. Le classificateur de nouvelles utilise un réseau LSTM pour prédire le label d'une information en se basant sur le comportement collectif de la communauté des votants. Bien que cette approche présente des aspects intéressants, elle souffre de deux limitations majeures. Premièrement, le recours à des modèles de type « boîte noire » limite la transparence du système, ce qui peut affecter la confiance des utilisateurs. Deuxièmement, les labels attribués aux informations sont statiques et immuables. Par conséquent, si l'évolution des votes fait qu'une information initialement classée comme fausse s'avère vraie, il n'existe aucun recours pour rectifier l'erreur. Cette rigidité, combinée à l'absence de garantie statistique, constitue une limitation majeure dans un contexte où l'information et les interactions sont dynamiques.

À notre connaissance, notre framework est la seule approche adoptant des scores de réputation et de crédibilité dynamiques et fournissant une garantie statistique lors de la classification des informations.

5.3.2 Vérification décentralisée des faits, basée sur la Blockchain, la Stance Detection et la borne de Hoeffding

Cette section présente notre framework qui combine la détection d'opinions et la technologie blockchain afin d'établir un système décentralisé et automatisé de bout en bout pour la vérification des

faits. La suite donne une vue d'ensemble des principales composantes du système avant d'expliquer les concepts de réputation et de crédibilité.

5.3.2.1 Vue d'ensemble

Le framework que nous proposons a pour objectif d'évaluer la crédibilité d'une information en se basant sur l'opinion collective d'une communauté d'utilisateurs, tout en tenant compte de la fiabilité (*réputation*) de chaque participant. Ce cadre peut être vu comme une généralisation des systèmes de vote, où les votes sont inférés des opinions exprimées plutôt que fournis explicitement.

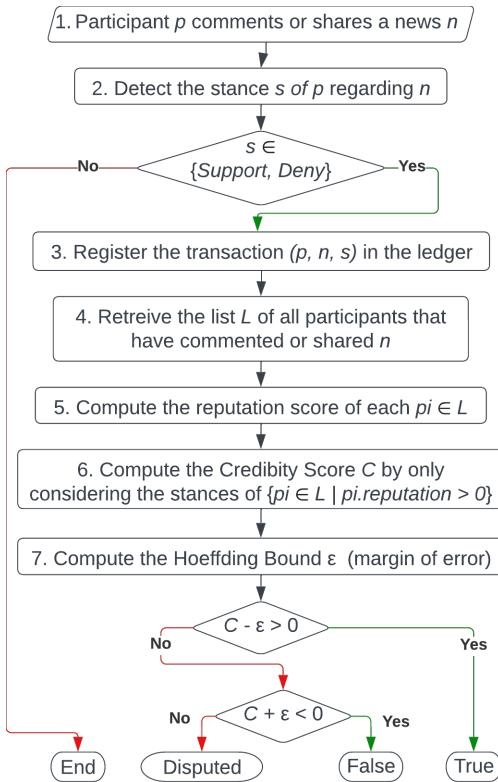


FIGURE 5.6 – Vue d'ensemble du processus d'évaluation de la crédibilité d'une information

Comme le montre la Figure 5.6, les participants interagissent avec le système en commentant ou en partageant des informations (Étape 1). Le système utilise ensuite la détection d'opinions pour déterminer si le participant soutient ("Support") ou réfute ("Deny") l'information (Étape 2). La position du participant est alors enregistrée dans le registre distribué (Étape 3). Le registre distribué garantit que l'historique des contributions (*i.e.* positions) ne peut pas être modifié, renforçant ainsi l'intégrité globale du système.

Le score de crédibilité d'une information dépend des opinions et des scores de réputation des participants qui l'ont commentée ou partagée. Le score de réputation s'inspire du solde conventionnel de cryptomonnaies, augmentant dynamiquement avec les contributions positives et diminuant avec celles négatives. Ce "solde" reflète le comportement global d'un participant au sein du système et est calculé en analysant l'ensemble de son historique d'actions (Étape 5). Cela implique de sommer les contributions positives (similaires aux transactions entrantes dans une blockchain classique), de déduire les contributions négatives (similaires aux transactions sortantes), puis de normaliser le résultat.

La réputation agit comme un indicateur de confiance, orientant le consensus sur les évaluations de véracité des nouvelles (Étape 6). Le mécanisme de consensus adopté dans notre cadre présente des similitudes avec les systèmes de Preuve de Réputation (PoR) et de Preuve d'Enjeu (PoS). Cela signifie que les contributeurs engagent leur réputation lors qu'ils interagissent avec le système et que seules les opinions des participants ayant une réputation positive sont prises en compte lors de la détermination du score de crédibilité d'une information. Ce score de crédibilité est recalculé dynamiquement à mesure que de nouvelles opinions deviennent disponibles.

5.3.2.2 Estimation dynamique de la réputation et de la crédibilité

Pour favoriser la transparence et donc la confiance et l'engagement, nous avons opté pour des formules intuitives et simples pour les scores de réputation et de crédibilité. Ces scores ne sont pas explicitement stockés, mais (re-)calculés dynamiquement à la volée,似ilairement à la blockchain conventionnelle, où le solde d'un compte n'est pas stocké, mais déterminé à la demande à partir de l'historique des transactions. Ce choix délibéré a été motivé par, d'un côté, la nature dynamique des interactions et donc de la réputation et de la crédibilité et, de l'autre, le caractère immuable de la blockchain.

Dans notre approche, la seule information stockée dans la blockchain est la position (*i.e.* vote) d'un participant à l'égard d'une information. Chaque fois qu'un utilisateur commente ou partage une information, une transaction est créée. La transaction contient l'adresse du participant, l'identifiant de la nouvelle et le vote du participant vis-à-vis de celle-ci. C'est l'ensemble de ces transactions qui permet de calculer, à tout moment, le score de réputation des participants et le score de crédibilité des nouvelles.

Score de réputation Comme mentionné précédemment, le score de réputation d'un participant est déterminé en fonction de son historique de contributions positives et négatives. Une contribution positive (*resp.* négative) correspond à la confirmation/partage d'une information vraie (*resp.* fausse) ou la réfutation d'une fausse (*resp.* vraie) information.

L'algorithme 3 décrit le processus de calcul du score normalisé de la réputation. L'opinion d'un participant concernant une nouvelle peut être soit 1 pour "Support" ou -1 pour "Deny", tandis que le statut de la nouvelle peut être soit 1 pour "True", -1 pour "False" ou 0 pour "Disputed". Le score de

Algorithm 3: Reputation Score Calculation

```

Input: participantHash
Result: Participant's reputation score
1 for each news in participants[participantHash].commentedAndSharedNews do
2     status ← newsMap[newsID].status;
3     stance ← participants[participantHash].stances[newsID];
4     if (status × stance > 0) then
5         positiveContributions++;
6     else
7         if status ≠ 0 then
8             negativeContributions++;
9     totalContributions ← positiveContributions + negativeContributions;
10    if totalContributions > 0 then
11        return (positiveContributions – negativeContributions)/totalContributions;
12    return 0;

```

réputation résultant varie de -1 (réputation totalement négative) à 1 (réputation totalement positive), la magnitude indiquant la fiabilité du participant.

Puisque la réputation est liée au comportement passé et aux contributions, les participants ont un intérêt à maintenir une réputation positive.

Score de crédibilité Le score de crédibilité évalue la fiabilité d'une information en fonction des interactions et du consensus au sein de la communauté des participants. Ce score est calculé dynamiquement, en considérant uniquement les opinions des participants dont la réputation est positive. Ces opinions sont pondérées par les scores de réputation respectifs des participants, accordant ainsi une influence plus importante aux contributeurs considérés comme plus fiables. Cette approche participative, fondée sur la réputation, renforce la capacité du système à identifier et à contrer la diffusion de fausses informations en exploitant la sagesse collective, tout en discriminant les contributions potentiellement malveillantes ou peu fiables. Le score de crédibilité d'une information, noté C , est calculé selon la formule suivante :

$$C = \frac{\sum_{i=1}^N (Stance_i \times Reputation_i)}{N} \quad (5.1)$$

où N représente le nombre total de participants ayant une réputation positive et interagissant avec l'information, $Stance_i \in [-1, 1]$ et $Reputation_i \in [-1, 1]$ désignant respectivement l'opinion et le score de réputation du participant i . Le score normalisé $C \in [-1, 1]$ indique la fiabilité générale de l'information. Un score positif suggère un consensus en faveur de l'information, tandis qu'un score négatif reflète un consensus général contre celle-ci. L'amplitude du score exprime la force de ce consensus.

5.3.2.3 Évaluation agile et dynamique de la véracité avec la Borne de Hoeffding

Les opinions exprimées sur une information dans un réseau social sont typiquement espacées dans le temps. Attendre de recueillir toutes les opinions avant d'évaluer la véracité d'une information pourrait rendre cette évaluation obsolète et sans intérêt : la fausse nouvelle se serait déjà propagée à un point tel qu'il deviendrait très difficile d'en limiter les conséquences. Dans [SJK24], nous utilisons la borne de Hoeffding comme garantie statistique pour déterminer la véracité d'une information lorsque "suffisamment" de "preuves" ont été collectées (*i.e.*, sans attendre l'intégralité des opinions). Il convient de souligner que cette évaluation n'est pas figée. Le fait que dans notre approche les scores de crédibilité et les statuts des informations ne soient pas stockés, mais (re-)calculés à la volée implique que ces derniers peuvent être modifiés si de nouvelles preuves viennent contredire les évaluations précédentes. Cette caractéristique confère à notre système une grande flexibilité et lui permet de s'adapter en continu à l'évolution des interactions.

Comme indiqué dans le chapitre traitant de la vélocité, la borne de Hoeffding [Hoe63] stipule qu'avec une probabilité de $1 - \delta$, la moyenne réelle d'une variable aléatoire d'amplitude R ne différera pas de la moyenne empirique après N observations de plus de : $\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2N}}$. Avec R et δ fixes, la seule variable pouvant modifier la borne de Hoeffding ϵ est le nombre d'observations N . À mesure que N augmente, ϵ diminue, conformément au fait que la moyenne empirique se rapproche de plus en plus de sa valeur réelle. D'un point de vue pratique, le niveau de confiance $\delta \in [0, 1]$ permet d'ajuster la marge d'erreur "acceptable" ϵ , ou en ce qui nous concerne le compromis souhaité entre la précision de l'estimation et le temps de réponse.

Pour décider du statut d'une nouvelle, la variable aléatoire estimée est $p = Stance \times Reputation$, c'est-à-dire l'opinion d'un participant sur une information, pondérée par son score de réputation. La moyenne de p correspond à $C = \frac{\sum_{i=1}^N (Stance_i \times Reputation_i)}{N}$, le score de crédibilité de l'information. Étant donné que la position (*Stance*) est soit 1, soit -1, et que la réputation (*Reputation*) varie entre -1 et 1, C est compris entre -1 et 1, ce qui signifie que $R = 2$.

Lorsque $C - \epsilon > 0$ (ϵ étant la marge d'erreur), cela signifie que C est significativement positive, et l'information peut être considérée comme "True" : *i.e.* nous pouvons affirmer avec un niveau de confiance $1 - \delta$ que C restera toujours positif lors que d'autres opinions deviennent disponibles, et donc que l'information restera toujours vraie. Inversement, lorsque $C + \epsilon < 0$, cela signifie que C est significativement négative, et nous pouvons affirmer, avec un niveau de confiance de $1 - \delta$, que l'information est fausse et le restera toujours. Comme illustré dans le diagramme de flux de la figure 5.6, lorsque les conditions pour déclarer une information "True" ou "False" ne sont pas réunies, celle-ci reste à l'état "Disputed".

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada PICTURE [support]
u2: @u1 Apparently a hoax. Best to take Tweet down. [deny]
u3: @u1 This photo was taken this morning, before the shooting. [deny]
u4: @u1 I dont believe there are soldiers guarding this area right now. [deny]
u5: @u4 wondered as well. Ive reached out to someone who would know just to confirm that. Hopefully get response soon. [comment]
u4: @u5 ok, thanks. [comment]

FIGURE 5.7 – Exemple extrait du jeu de données RumourEval [GBD⁺18]

5.3.3 Résultats & discussion

5.3.3.1 Outils et jeu de données

La difficulté majeure rencontrée lors de notre étude expérimentale a été l'absence de jeux de données couvrant tous les aspects de la détection de fausses nouvelles. Les jeux de données existants traitent séparément la détection d'opinions, la crédibilité des sources et la classification des nouvelles. Ceci explique en partie la raison pour laquelle la plupart des approches basées sur la blockchain se concentrent davantage sur les scénarios d'attaques que sur l'identification efficace des fausses nouvelles. Bien que le jeu de données RumourEval [GBD⁺18] ne lie pas les interactions aux participants, nous l'avons adopté dans notre étude expérimentale du fait qu'il soit le seul annotant les données à la fois pour les opinions et la véracité. RumourEval contient 3 342 conversations issues de Twitter sur divers sujets et inclut deux tâches : (A) détection d'opinion (SDQC : *Support*, *Deny*, *Query*, *Comment*) et (B) détection de la véracité (Vrai, Faux, Indéterminé). Un exemple extrait de RumourEval est donné dans la Figure 5.7. Les concepteurs de RumourEval ont fourni BranchLSTM [KLZ18] comme modèle de référence pour la détection d'opinions (Task A). Afin d'isoler au mieux l'impact de notre framework de tout biais pouvant être introduit par des modèles sophistiqués, nous l'avons implémenté avec BranchLSTM. La logique applicative a été implémentée à l'aide d'un contrat intelligent en Solidity [sol]. Pour évaluer l'impact de la réputation, nous avons simulé 42 utilisateurs², générant un même nombre de stances. L'attribution aléatoire des stances peut entraîner des réputations faibles, ne reflétant pas la diversité réelle des réseaux sociaux, où coexistent des participants fiables, des acteurs malveillants et des utilisateurs intermédiaires. Afin de mieux capturer cette hétérogénéité, nous avons classé les participants en trois groupes de réputation : inférieure à -0,5 (malveillants), supérieure à 0,5 (réputés) et comprise entre -0,5 et 0,5 (intermédiaires), et exploré différentes proportions de ces groupes dans nos expérimentations.

5.3.3.2 Résultats

Nos expérimentations ont été menées avec trois objectifs : (i) quantifier l'impact de l'efficacité du détecteur d'opinions sur les performances du système; (ii) examiner l'influence des proportions de

2. Le nombre maximum de commentaires sur un tweet dans le jeu de données est de 41

5.3. De la *Stance* à la *Fake News Detection* : la Blockchain pour combler le chaînon manquant

| δ | $Dist_1 : [0.05, 0.9, 0.05]$ | $Dist_2 : [0.1, 0.8, 0.1]$ | $Dist_3 : [0.15, 0.7, 0.15]$ | $Dist_4 : [0.2, 0.6, 0.2]$ |
|----------|------------------------------|----------------------------|------------------------------|----------------------------|
| 0.2 | 15.63% | 27.75% | 30.59% | 34.07% |
| 0.3 | 19.74% | 31.47% | 37.88% | 38.55% |
| 0.4 | 24.37% | 34.59% | 40.77% | 49.83% |
| 0.5 | 28.72% | 37.57% | 42.58% | 50.22% |
| 0.6 | 28.72% | 37.90% | 49.29% | 50.22% |

TABLE 5.4 – Score F1 macro-moyen obtenu par notre framework avec le modèle de référence BranchLSTM [KLZ18]

| δ | $Dist_1 : [0.05, 0.9, 0.05]$ | $Dist_2 : [0.1, 0.8, 0.1]$ | $Dist_3 : [0.15, 0.7, 0.15]$ | $Dist_4 : [0.2, 0.6, 0.2]$ |
|----------|------------------------------|----------------------------|------------------------------|----------------------------|
| 0.2 | 21.64% | 30.40% | 33.28% | 41.49% |
| 0.3 | 30.50% | 37.15% | 39.90% | 51.97% |
| 0.4 | 39.20% | 41.03% | 49.29% | 78.20% |
| 0.5 | 40% | 49.29% | 54.83% | 78.20% |
| 0.6 | 40% | 49.29% | 67.04% | 78.20% |

TABLE 5.5 – Score F1 macro-moyen obtenu par notre framework lorsque toutes les stances sont correctement identifiées (similaire à un système de vote)

participants malveillants et réputés ; et (iii) analyser la sensibilité du système aux variations de δ , le niveau de confiance de la borne de Hoeffding. Pour illustrer l’impact du détecteur d’opinions, nous avons comparé les résultats obtenus avec BranchLSTM [KLZ18] (Table 5.4) à ceux d’un scénario où toutes les opinions seraient correctement déterminées (Table 5.5). Ce dernier scénario correspond à un système de vote explicite, que notre framework prend en charge en permettant aux participants de corriger les opinions détectées. Les résultats montrent qu’améliorer la détection d’opinions pourrait améliorer le F1-score de 12,2% en moyenne.

Pour évaluer l’influence des différentes distributions de participants sur les performances du framework, nous avons examiné quatre scénarios représentés par les distributions $Dist_1$ à $Dist_4$: $Dist_1 : [0,05 \text{ Réputés}, 0,9 \text{ Intermédiaires}, 0,05 \text{ Malveillants}]$, $Dist_2 : [0,1 \text{ Réputés}, 0,8 \text{ Intermédiaires}, 0,1 \text{ Malveillants}]$, $Dist_3 : [0,15 \text{ Réputés}, 0,7 \text{ Intermédiaires}, 0,15 \text{ Malveillants}]$, et $Dist_4 : [0,2 \text{ Réputés}, 0,6 \text{ Intermédiaires}, 0,2 \text{ Malveillants}]$. Comme l’indiquent les tables 5.4 et 5.5, l’efficacité du framework n’est pas affectée par la proportion de participants malveillants, mais fortement influencée par la fraction de participants réputés. En moyenne, une augmentation de 5% de la proportion des participants réputés, améliore le F1-score de 8,75%.

Comme mentionné précédemment, la borne de Hoeffding est connue pour son caractère conservateur. Cette caractéristique est confirmée dans notre étude expérimentale, où de faibles valeurs du niveau de confiance ($1 - \delta$) se traduisent par de bonnes performances. Cela s’explique par le fait que lorsque δ est faible, la borne de Hoeffding évite de prendre des décisions concernant le statut des nouvelles en attendant que d’autres opinions soient disponibles et donc que la marge d’erreur se réduise.

| Approche | Score |
|---|--------------|
| Baseline Majority Class [GBD ⁺ 18] | 22.41% |
| BranchLSTM [KLZ18] | 49.29% |
| eventAI [LZS19] | 57.65% |
| Fine-tuned Longformer [Kha21] | 58.68% |
| Shared Multi-channel Interactions (MTL-SMI) [LYZ ⁺ 24] | 68.5% |
| Approche proposée avec BranchLSTM ($\delta = 0.5$) | up to 50.22% |
| Approche proposée avec vote ($\delta = 0.5$) | up to 78.20% |

TABLE 5.6 – Comparaison des scores F1 macro-moyen obtenu par les différentes approches.

En conséquence, une partie importante des nouvelles reste catégorisée comme "Undertermined" plutôt que comme "True" ou "False".

5.3.3.3 Discussion

Les résultats de notre étude expérimentale mettent en lumière les facteurs influençant l'efficacité de notre framework. En particulier, les résultats montrent que, bien que les participants ayant une faible réputation aient une influence négligeable sur les performances du système, l'efficacité globale dépend fortement de la proportion de participants réputés. Cette dépendance peut s'expliquer par deux facteurs : le mécanisme de *Proof of Reputation*, qui filtre les opinions des participants de faible réputation, et la borne de Hoeffding, qui exige un minimum de preuves corroborantes avant de rendre un verdict sur la véracité d'une information. Cela met en évidence une faiblesse potentielle dans les scénarios où les contributeurs fiables sont rares ou pendant les premières phases du système, car celui-ci pourrait avoir du mal à atteindre un consensus.

Les résultats ont également montré dans quelle mesure l'efficacité du détecteur d'opinions, représenté par le modèle BranchLSTM, impacte les performances globales du système, avec une amélioration potentielle moyenne d'environ 12,2 % de la détection des fausses nouvelles. Cela souligne la nécessité d'explorer davantage des détecteurs d'opinions plus sophistiqués.

En ce qui concerne la borne de Hoeffding, son caractère conservateur introduit un compromis entre les niveaux de confiance et la précision. Comme l'a montré notre étude expérimentale, abaisser le niveau de confiance incite la borne de Hoeffding à prendre des décisions plus rapidement, ce qui se traduit par des jugements plus décisifs et de meilleurs scores F1 (car moins de nouvelles sont classées comme "Disputed"). Cependant, cela augmente également le risque de classifications erronées, notamment dans les situations avec peu de données ou une grande incertitude. Bien que cette adaptabilité permette d'ajuster le framework en fonction des besoins spécifiques, elle soulève des questions quant à l'équilibre entre la rapidité et la précision.

Bien que ses performances soient comparables à celles des approches d'apprentissage automatique les plus avancées, notre framework se distingue par sa nature entièrement décentralisée et dynamique. Cette approche dynamique offre une évaluation plus nuancée de la véracité des nouvelles par rapport

aux modèles statiques d'apprentissage automatique et favorise un environnement où l'influence des participants est alignée avec leur fiabilité établie. Ces caractéristiques sont obtenues sans dépendre d'une autorité centrale ou nécessiter des ré-apprentissage réguliers pour mettre à jour des connaissances figées du monde. Ceci constitue un avantage de taille par rapport aux approches traditionnelles d'apprentissage automatique et positionne notre framework comme une solution prometteuse dans le paysage complexe de la détection des fausses nouvelles.

Enfin, il convient de noter que notre framework et les approches conventionnelles basées sur l'apprentissage automatique ne sont pas mutuellement exclusifs. Les scores de réputation et de crédibilité dérivés de notre framework peuvent servir d'entrées précieuses pour les modèles d'apprentissage automatique, fournissant des informations contextuelles supplémentaires qui ont le potentiel d'améliorer la détection des fausses nouvelles.

5.4 Conclusion

Ce chapitre a traité de la véracité des informations dans les médias sociaux. Dans un premier temps, nos travaux ont porté sur l'amélioration de la détection d'opinions à travers l'augmentation de texte. Ensuite, nous avons exploré l'utilisation de la blockchain pour combler le chaînon manquant entre la détection d'opinions et la détection des fausses informations. Le framework que nous proposons vise à décentraliser la vérification des faits et à tirer parti du mécanisme de consensus *Proof of Reputation* pour réduire les risques de biais et de censure tout en améliorant la scalabilité et la résilience contre les attaques malveillantes. En intégrant la borne de Hoeffding, notre framework offre une garantie statistique pour l'évaluation de la véracité des informations en temps opportun (*i.e* sans attendre que l'intégralité des opinions). L'un des avantages saillants de notre approche par rapport aux approches traditionnelles basées sur l'apprentissage automatique réside dans sa capacité à s'adapter à l'évolution perpétuelle des informations. Contrairement à ces modèles qui requièrent un ré-apprentissage régulier pour mettre à jour des connaissances statiques, notre framework est moins sujet aux fluctuations.

Les travaux présentés dans ce chapitre ont été menés dans le cadre du master de recherche et de la thèse Melle. Ilhem Salah, en collaboration avec l'École Hexagone (Versailles, France) et l'École d'ingénieurs ESIGELEC (Rouen, France). Notre collaboration avec l'ESIGELEC se poursuit pour explorer l'usage de la logique floue pour affiner l'évaluation des opinions, des réputations et des informations, en créant des labels plus nuancés tels que "partiellement vrai" ou "partiellement faux". Parallèlement, dans le cadre du master de recherche de M. Kamel nous explorons d'autres implémentations de notre framework, avec notamment la similarité cosinus pour la détermination de la réputation et de l'entropie pondérée des opinions comme alternative à la borne de Hoeffding pour évaluer la convergence du consensus sur la véracité des informations.

Publications en lien avec le chapitre

Revues internationales

[SJPK24] Connecting the Dots between Stance and Fake News Detection with Block-chain, Proof of Reputation, and the Hoeffding Bound

Ilhem SALAH, Khaled JOUINI, Cyril-Alexandre Pachon & Ouajdi KORBA

Cluster Computing. ISSN : 1386-7857, Springer Nature, DOI : <https://doi.org/10.1007/s10586-024-04637-7>. 2024.

SJR best quartile : Q1, SJR : 1.07, JCR IF (2023) : 3.6.

[SJK23] On the Use of Text Augmentation For Stance and Fake News Detection

Ilhem SALAH, Khaled JOUINI & Ouajdi KORBA

Journal of Information and Telecommunication. Vol. 7, No. 3, pp 359-375, ISSN : 24751839, Taylor and Francis Ltd., DOI : <https://doi.org/10.1080/24751839.2023.2198820>. 2023.

SJR best quartile : Q2, SJR : 0.67, JCR IF (2023) : 2.7

Conférences internationales

[SJK22] Augmentation-Based Ensemble Learning for Stance and Fake News Detection

Ilhem SALAH, Khaled JOUINI & Ouajdi KORBA

14th International Conference on Computational Collective Intelligence (ICCCI). DOI : https://doi.org/10.1007/978-3-031-16210-7_3. 2022.

CORE Rank : B

Encadrement doctoral

— Ilhem SALAH (En cours, Directeur de thèse : Prof. Ouajdi KORBA)

Encadrement de masters de recherche

— Ilhem SALAH (Soutenu)

— Abdelhamid KAMEL (Soutenu)

CHAPITRE 6

Conclusion & perspectives

L'INFORMATIQUE a ça de particulier qu'elle évolue à un rythme extraordinairement rapide, entraînant souvent l'obsolescence, en l'espace de quelques années, de paradigmes et de thématiques de recherche considérés à leurs débuts comme les plus prometteurs. Cette évolution rapide, couplée à une préoccupation constante de rester en phase avec les avancées dans le domaine du stockage et du traitement des données, m'a conduit à m'investir pleinement dans le vaste et multidisciplinaire domaine que représentent les données massives et à ouvrir le spectre de mes activités d'enseignement et de recherche à leurs différentes facettes. Mes recherches se sont ainsi articulées autour de trois types d'efforts.

- Poursuite des travaux de thèse : J'ai étendu mes travaux de thèse, portant sur l'optimisation de la localité spatiale dans les bases relationnelles, aux bases NoSQL.
- Lancement de nouvelles pistes : la multidisciplinarité des données massives m'a amené à explorer d'autres dimensions que celui du volume et à lancer de nouveaux axes ayant trait à la Véracité (lutte contre la désinformation), la Variété (classification LULC) et la Vélocité (apprentissage en ligne adaptatif aux dérives de concepts).
- Applications industrielles : l'informatique est aussi bien une science qu'une technologie mettant en pratique cette science. La thèse de M. H. Ghaddab [Gha17], menée dans le cadre du programme MOBIDOC, est à ce titre une parfaite illustration des les bénéfices qui peuvent être tirés de la mise des fondements théoriques au service de la résolution de problématiques concrètes.

Dans le cadre de ces travaux, j'ai encadré six sujets de mastère de recherche et eu l'honneur de co-encadrer avec Pr. Ouajdi KORBA, une thèse soutenue et deux thèses en cours. Ces encadrements

ont donné lieu à quinze publications scientifiques. Au-delà des contributions scientifiques, ces travaux ont permis de développer des collaborations enrichissantes au niveau national et international. En plus de la collaboration avec les membres de mon laboratoire de recherche, j'ai ainsi pu, avec le concours du Pr. O. KORBA, prendre part à des collaborations internationales, notamment avec l'École Hexagone (Versailles, France), l'École d'ingénieurs ESIGELEC (Rouen, France), Telecom paristech (Paris, France) et la *Northern Technical University* (Mossoul, Irak). Les travaux que j'ai menés couvrent ainsi diverses formes de partenariat et de collaboration. En dépit de la diversité des thématiques abordées, je pense qu'elles sont complémentaires et cohérentes à la fois avec mes activités d'enseignement, mes travaux de thèse et les sujets traités par mon équipe de recherche.

Perspectives

Les travaux de recherche présentés dans ce manuscrit se poursuivent activement dans trois grands axes : (i) apprentissage en ligne adaptatif aux dérives de concept ; (ii) détails locaux fins et contexte global dans la classification d'images ; et (iii) détection d'opinions et lutte contre la désinformation.

Apprentissage incrémental adaptatif aux dérives de concept

Les perspectives liées à l'apprentissage incrémental adaptatif aux dérives de concept se concentrent sur le développement de modèles capables de s'adapter en temps réel aux changements progressifs ou abrupts dans les données, tout en préservant leurs connaissances sur les concepts stables qui demeurent pertinents au fil du temps. En particulier, nous visons à concevoir des modèles capables d'anticiper de manière **proactive** les dérives de concept. Cette capacité repose sur l'identification des signes avant-coureurs des dérives, obtenue par une analyse approfondie des dépendances temporelles et des interactions entre différents types de dérives.

Le mécanisme d'attention, qui constitue le cœur des Transformers [VSP⁺¹⁷], offre une capacité unique à traiter efficacement des séquences complexes. En identifiant les éléments les plus pertinents et en modélisant leurs relations, l'attention permet de capturer des dépendances locales et globales au sein des données. Les Transformers exploitent ce mécanisme pour modéliser de manière contextuelle des dépendances à différentes échelles. Cette combinaison d'attention et de traitement hiérarchique confère aux Transformers leur capacité unique à comprendre les interactions complexes dans des séquences dynamiques. Dans le contexte spécifique de nos travaux, nous nous appuierons sur des extensions de ces mécanismes classiques, notamment sur l'**attention temporelle** [WCW20] et les Transformers temporels [OV24]. L'attention temporelle permet de modéliser les relations entre les éléments d'une séquence en prenant en compte leur position dans le temps, tandis que les Transformers temporels sont conçus pour analyser les séries chronologiques en capturant les dynamiques évolutives des données. Nous envisageons d'utiliser ces outils pour modéliser les dynamiques temporelles des dérives et identifier des **patterns récurrents** susceptibles d'annoncer des **changements imminents**. Par ailleurs, nous

approfondirons l’analyse des interactions entre différents types de dérives afin d’enrichir les prédictions en exploitant les données historiques. Par exemple, en étudiant comment une dérive progressive peut précéder une dérive abrupte, il sera possible d’estimer avec plus de précision la probabilité et la nature des dérives futures. Cette approche proactive, reposant sur l’attention temporelle et les Transformers temporels, a le potentiel d’améliorer significativement la rapidité et l’efficacité de l’adaptation aux dérives, surpassant ainsi les méthodes d’adaptation rétroactives classiques.

Détails locaux fins et contexte global dans la classification d’images

Les perspectives pour cet axe portent sur les domaines où les détails à granularité fine sont aussi importants que le contexte global, à l’instar des images **satellitaires**, **médicales** ou des **images manipulées**. Les récents développements des **Visual Transformers** ont permis des avancées significatives, surpassant souvent les CNN dans la capture du contexte global [RKH⁺24]. Cependant, ces modèles peinent à restituer les détails fins, pourtant essentiels dans de nombreux cas critiques [LZC⁺21]. Ces limitations nous confortent dans l’idée que les descripteurs manuels, tels que SIFT ou SURF pourraient complémer les Visual Transformers. Ces descripteurs excellent en effet dans la capture de détails à granularité fine que l’attention globale des Transformers tend à négliger.

L’intégration de ces deux modalités pourrait ouvrir la voie à de nouvelles architectures hybrides combinant les capacités contextuelles des Transformers et la précision des descripteurs locaux. Cette stratégie serait particulièrement prometteuse pour la détection des **images manipulées**, un enjeu crucial dans divers contextes. La prolifération des *deepfakes* illustre notamment les défis croissants non seulement pour la société mais également pour les grands modèles de l’IA générative [BF22]. Entraînés sur des volumes massifs de données issues du web, ces modèles restent vulnérables à l’ingestion d’images manipulées pouvant entraîner des biais d’apprentissage. Dans ce contexte, la synergie entre Visual Transformers et descripteurs locaux permettrait de détecter des manipulations indécelables avec une seule modalité (comme des artefacts aux contours ou des incohérences globales), et aussi de renforcer la fiabilité des grands modèles en amont, en assainissant les données d’entraînement.

Détection d’opinions et lutte contre la désinformation

L’idée centrale de nos travaux sur la détection des fake news est d’inférer les opinions exprimées sur un fait (*stance detection*), puis d’estimer sa véracité en pondérant ces opinions par la crédibilité de la source. Nos perspectives explorent plusieurs axes pour améliorer la détection d’opinions et la lutte contre la désinformation. Un premier axe porte sur l’utilisation de l'**apprentissage contrastif** et de l'**apprentissage guidé par les prompts** (*prompt-based learning* [JJH⁺22]) pour affiner la détection des opinions concordantes et contradictoires. L’apprentissage contrastif optimise les représentations sémantiques en rapprochant les opinions similaires et en éloignant celles divergentes. L’apprentissage guidé par les prompts, quant à lui, guide un modèle de langage pré-entraîné via des prompts spécifiques afin de mieux saisir le contexte et les nuances sémantiques. Cette approche s’avère particulièrement

efficace pour désambiguïser le langage naturel. Les résultats préliminaires obtenus sur le banc d'essai RumourEval confirment le potentiel de cette approche.

L'analyse des conversations en ligne pour la détection d'opinions et la lutte contre la désinformation nécessite de modéliser les interactions complexes et l'évolution temporelle des échanges. Nous entendons exploiter la puissance des réseaux de neurones pour les graphes, notamment les **Temporal Graph Convolutional Networks** (T-GCN) [PDC⁺19]. En représentant les arborescences de conversation comme des graphes dynamiques, où les noeuds représentent les utilisateurs et leurs messages, et les arêtes les interactions, les T-GCN peuvent capturer la propagation des opinions et l'influence des échanges passés. Des travaux récents [AB24] ont démontré l'efficacité des GNN pour la détection de fake news, mais l'intégration fine de la temporalité et la modélisation des changements de stance restent des défis ouverts. Nos recherches exploreront des architectures T-GCN innovantes intégrant des mécanismes d'attention temporelle pour identifier les moments clés et les changements d'opinion significatifs dans les conversations.

Sur un plus long terme, nous envisageons de poursuivre l'exploration de l'intégration de la technologie blockchain, notamment pour sa capacité à garantir la transparence et la décentralisation du processus de vérification des faits. Nous envisageons à cet effet d'introduire la **logique floue** dans les mécanismes de consensus pour ajouter des degrés de certitude aux opinions exprimées par les participants, à leur fiabilité et à la véracité des informations elles-mêmes. Cela offrirait une gestion plus flexible des zones d'incertitude, tout en permettant une prise de décision plus nuancée.

Récapitulatif des publications et des encadrements

Publications

Revues internationales

[JJK25a] Detecting Unknown Intrusions from Large Heterogeneous Data through Ensemble Learning

Farah JEMILI, Khaled JOUINI & Ouajdi KORBA

Intelligent Systems with Applications (ISWA 200465). ISSN : 26673053, Elsevier B.V., DOI : <https://doi.org/10.1016/j.iswa.2024.200465>. 2025.

SJR best quartile : Q1, SJR : 0.96

[JJK25b] Intrusion Detection based on Concept Drift Detection & Online Incremental Learning

Farah JEMILI, Khaled JOUINI & Ouajdi KORBA

International Journal of Pervasive Computing and Communications. ISSN : 1742-7371, Emerald group publishing ltd. DOI : <https://doi.org/10.1108/IJPCC-12-2023-0358>. 2025.

SJR best quartile : Q2, SJR : 0.36, JCR IF (2023) : 0.6.

[SJPK24] Connecting the Dots between Stance and Fake News Detection with Blockchain, Proof of Reputation, and the Hoeffding Bound

Ilhem SALAH, Khaled JOUINI, Cyril-Alexandre Pachon & Ouajdi KORBA

Cluster Computing. ISSN : 1386-7857, Springer Nature, DOI : <https://doi.org/10.1007/s10586-024-04637-7>. 2024.

SJR best quartile : Q1, SJR : 1.07, JCR IF (2023) : 3.6.

[SJK23] On the Use of Text Augmentation For Stance and Fake News Detection

Ilhem SALAH, Khaled JOUINI & Ouajdi KORBA

Journal of Information and Telecommunication. Vol. 7, No. 3, pp 359-375, ISSN : 24751839, Taylor and Francis Ltd., DOI : <https://doi.org/10.1080/24751839.2023.2198820>. 2023.

SJR best quartile : Q2, SJR : 0.67, JCR IF (2023) : 2.7

[Jou22] Aggregates Selection in Replicated Document-Oriented Databases

Khaled JOUINI

Journal of Information Science and Engineering (Abbréviation JCR : J INF SCI ENG). 2022. ISSN : 1016-2364, DOI : [10.6688/JISE.20220338\(2\).0012](https://doi.org/10.6688/JISE.20220338(2).0012). 2023.

SJR best quartile : Q3, SJR : 0.21, JCR IF (2022) : 1.1

Conférences internationales

[AJK25] Enhancing LULC Classification with Attention-Based Fusion of Handcrafted and Deep Features

Vian ABDULMAJEED AHMAD, Khaled JOUINI & Ouajdi KORBA

17th International Conference on Agents and Artificial Intelligence (ICAART). Feb 2025. (*Accepted, to appear*)

CORE Rank : B.

[AJTK24a] Integrating Deep and handcrafted Features for Enhanced Remote Sensing Image Classification

Vian ABDULMAJEED AHMAD, Khaled JOUINI, Amel TUAMA & Ouajdi KORBA

21st ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). 2024. (*Accepted, To appear*)

CORE Rank : C.

[AJTK24b] A Fusion Approach for Enhanced Remote Sensing Image Classification

Vian ABDULMAJEED AHMAD, Khaled JOUINI, Amel TUAMA & Ouajdi KORBA

19th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP. DOI : <https://doi.org/10.5220/0012376600003660>. 2024.

CORE Rank : B (lors de la soumission).

[JK23] Drift-Driven Regression for Predicting the Evolution of Pandemics

Khaled JOUINI & Ouajdi KORBA

IADIS International Conference Applied Computing (IADIS AC). pp. 77-84, ISBN : 978-989-8704-53-5. DOI : https://doi.org/10.33965/ICWI_AC_2023_202307L009. 2023.

CORE Rank : C.

[SJK22] Augmentation-Based Ensemble Learning for Stance and Fake News Detection

Ilhem SALAH, Khaled JOUINI & Ouajdi KORBA

14th International Conference on Computational Collective Intelligence (ICCCI). DOI : <https://doi.org/10.1186/s13675-022-01237-0>.

//doi.org/10.1007/978-3-031-16210-7_3. 2022.

CORE Rank : B

[JMK21] Real-Time, CNN-Based Assistive Device for Visually Impaired People

Khaled JOUINI, Mohamed Hédi MAALOUL & Ouajdi KORBAA

International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (IEEE CIS-BMEI). pp. 1-6, DOI : <https://doi.org/10.1109/CISP-BMEI53629.2021.9624387>. 2021.

[ABC⁺18] The Database Version Approach: Overview and Future directions

Talel ABDESSALEM, Claudia MEDEIROS BAUZER, Wojciech CELLARY, Stéphane GANÇARSKI, Khaled JOUINI, Maude MANOUVRIER, Marta RUKOZ, Michel ZAM

34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA). Romania. 2018.

[Jou17] Distorted Replicas: Intelligent Replication Schemes to Boost I/O Throughput in NoSQL Systems

Khaled JOUINI

14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). DOI : <https://doi.org/10.1109/AICCSA.2017.8270014>. 2017.

CORE Rank : C.

[GJK17a] Fast and Accurate Fingerprint Matching using Expanded Delaunay Triangulation

Mohamed Hédi GHADDAB, Khaled JOUINI & Ouajdi KORBAA

14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). DOI : <https://doi.org/10.1109/AICCSA.2017.8270014>. 2017.

CORE Rank : C

[GJK17b] Fusion de minuties pour une reconnaissance efficiente des empreintes digitales

Mohamed Hédi GHADDAB, Khaled JOUINI & Ouajdi KORBAA

Colloque sur l'Optimisation et les Systèmes d'Information (COSI). Algérie. 2017.

Encadrement doctoral

- Mohamed Hédi GHADDAB (Thèse soutenue le 07-12-2018 École Nationale des Sciences de l'Informatique, Directeur de thèse : Prof. Ouajdi KORBAA)
- Vian ABDELMajeed AHMAD (En cours, Directeur de thèse : Prof. Ouajdi KORBAA)
- Ilhem SALAH (En cours, Directeur de thèse : Prof. Ouajdi KORBAA)

Encadrement de masters de recherche

- Nada BEN LATIFA (Soutenu)
- Islem OTHMANI (Soutenu)
- Haythem SAOUDI (Soutenu)
- Mouna MZIOU (Soutenu)
- Marwa GHEZAIEL (Soutenu)
- Ilhem SALAH (Soutenu)
- Abdelhamid KAMEL (Soutenu)

Bibliographie

- [AAB⁺15] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citron, G. Corrado, A. Davis, J. Dean, M. Devin, and et. al. Tensorflow : Large-scale machine learning on heterogeneous systems. In *Google Research*, 2015. Software available from tensorflow.org.
- [AB24] Mohammad Q. Alnabhan and Paula Branco. Fake news detection using deep learning : A systematic literature review. *IEEE Access*, 12 :114435–114459, 2024.
- [ABC⁺18] Talel Abdessalem, Claudia Medeiros Bauzer, Wojciech Cellary, Stéphane Gançarski, Khaled Jouini, Maude Manouvrier, Marta Rukoz, and Michel Zam. The database version approach : Overview and future directions. In *34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2018)*, Bucarest, Romania, 2018.
- [AJK25] Vian Abdulmajeed Ahmad, Khaled Jouini, and Ouajdi Korbaa. Enhancing lulc classification with attention-based fusion of handcrafted and deep features. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART)*, Porto, Portugal, February 2025. Accepted, to Appear.
- [AJTK24a] Vian Abdulmajeed Ahmad, Khaled Jouini, Amal Tuama, and Ouajdi Korbaa. Integrating deep and handcrafted features for enhanced remote sensing image classification. In *Proceedings of the 21st ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 2024.
- [AJTK24b] Vian Abdulmajeed Ahmad, Khaled Jouini, Amel Tuama, and Ouajdi Korbaa. A fusion approach for enhanced remote sensing image classification. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Italy, January 2024.
- [ans] Ansacheck. https://www.ansa.it/sito/static/ansa_check.html. [Accessed on 31 October 2024].
- [AR22] Meirylene Avelino and Antonio A. de A. Rocha. Blockproof : A framework for verifying authenticity and integrity of web content. *Sensors*, 22(3), 2022.

BIBLIOGRAPHIE

- [BB24] Abhishek Bhatt and Vandana Thakur Bhatt. Dcrff-lhrf : an improvised methodology for efficient land-cover classification on eurosat dataset. *Multimedia Tools and Applications*, 83(18) :54001–54025, May 2024.
- [BBG⁺21] Maroua Bahri, Albert Bifet, João Gama, Heitor Murilo Gomes, and Silviu Maniu. Data stream analysis : Foundations, major tasks and tools. *WIREs Data Mining and Knowledge Discovery*, 11, 03 2021.
- [BF22] Matyas Bohacek and Hany Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48) :e2216035119, 2022.
- [BG07] Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the 7th SIAM International Conference on Data Mining*, pages 443–448, 04 2007.
- [BGHP18] Albert Bifet, Ricard Gavaldà, Geoff Holmes, and Bernhard Pfahringer. *Machine learning for data streams : with practical examples in MOA*. MIT Press, 2018.
- [Bif10] Albert Bifet. Adaptive stream mining : Pattern learning and mining from evolving data streams. *Frontiers in Artificial Intelligence and Applications*, 207 :1–212, 01 2010.
- [BRAC08] Soma Biswas, Nalini K Ratha, Gaurav Aggarwal, and Jonathan Connell. Exploring ridge curvature for fingerprint indexing. In *Biometrics : Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–6. IEEE, 2008.
- [CAQ⁺22] E. Camargo, J. Aguilar, Y. Quintero, F. Rivas, and D. Ardila. An incremental learning approach to prediction models of seird variables in the context of the covid-19 pandemic. *Health and Technology*, 12(4) :2190–7196, 2022.
- [CDPG22] Chien-Chih Chen, Yuxuan Du, Richards Peter, and Wojciech M. Golab. An implementation of fake news prevention by blockchain and entropy-based incentive mechanism. *Soc. Netw. Anal. Min.*, 12(1) :114, 2022.
- [CFFM07] Raffaele Cappelli, Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Fingerprint verification competition 2006. *Biometric Technology Today*, 15(7) :7–9, 2007.
- [CG16] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.
- [CGL22] Angelo Capodici, Davide Gori, and Jacopo Lenzi. Deaths, countermeasures, and obedience : How countries' non-pharmaceutical measures have quelled the covid-19 death toll. *Frontiers in Public Health*, 10, 2022.
- [CLLK03] Kyoungtaek Choi, Dongjae Lee, Sanghoon Lee, and Jaihie Kim. An improved fingerprint indexing algorithm based on the triplet approach. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 584–591. Springer, 2003.

BIBLIOGRAPHIE

- [CN07] Surajit Chaudhuri and Vivek R. Narasayya. Self-tuning database systems : A decade of progress. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 3–14. ACM, 2007.
- [Cou18] Claude Coulombe. Text data augmentation made simple by leveraging NLP cloud apis. *CoRR*, abs/1812.04718, 2018.
- [DA21] H. I. Dewangkoro and A. M. Arymurthy. Land use and land cover classification using cnn, svm, and channel squeeze & spatial excitation block. *IOP Conf. Ser. Earth Environ. Sci.*, 704(1), 2021.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [DDDW19] Chris Duhanty, Jason L. Deglint, Ibrahim Ben Daya, and Alexander Wong. Taking a stance on fake news : Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. *CoRR*, abs/1911.11951, 2019.
- [Del34] Boris Delaunay. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800) :1–2, 1934.
- [DH00] Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 71–80, 2000.
- [DJ21] R. Dahabi and F. Jemili. A deep learning approach for intrusion detection. In *2021 IEEE 23rd International Conference on High Performance Computing & Communications (HPCC)*, pages 1–8, 2021.
- [dLdSM15] Claudio de Lima and Ronaldo dos Santos Mello. A workload-driven logical design approach for nosql document databases. In *Proceedings of the 17th International Conference on Information Integration and Web-Based Apps & Services*, pages 1–10. ACM, 2015.
- [DMG20] Ronald Denaux, Flavio Merenda, and José Manuél Gómez-Pérez. Towards crowdsourcing tasks for accurate misinformation detection. In *Joint Proceedings of Workshops AI4LEGAL2020, NLIWOD, PROFILES 2020, QuWeDa 2020 and SEMIFORM2020 Co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference, November, 2020*, volume 2722 of *CEUR Workshop Proceedings*, pages 159–167. CEUR-WS.org, 2020.
- [Eri03] Evans Eric. *Domain-Driven Design : Tacking Complexity In the Heart of Software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2003.
- [Far12] Gerald Farin. Shape measures for triangles. *IEEE transactions on visualization and computer graphics*, 18(1) :43–46, 2012.
- [FBAF10] Romain Fontugne, Pierre Borgnat, Patrice Abry, and Kensuke Fukuda. MAWILab : Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking. In *ACM CoNEXT '10*, Philadelphia, PA, December 2010.
- [FGG⁺18] Alberto Fernández, Salvador Garca, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer Publishing Company, Incorporated, 1st edition, 2018.

BIBLIOGRAPHIE

- [GBD⁺18] Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. Rumoureval 2019 : Determining rumour veracity and support for rumours. *CoRR*, abs/1809.06683, 2018.
- [GBFB18] Heitor Murilo Gomes, Jean Paul Barddal, Luis Eduardo Boiko Ferreira, and Albert Bifet. Adaptive random forests for data stream regression. In *26th European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium, April 25-27, 2018*, 2018.
- [GCC⁺97] Robert S Germain, Andrea Califano, Scott Colville, et al. Fingerprint matching using transformation parameter clustering. *IEEE Computational Science and Engineering*, 4(4) :42–49, 1997.
- [GD22] Andrea Gesmundo and Jeff Dean. An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems. *CoRR*, abs/2205.12755, 2022.
- [Gha17] Mohamed Hédi Ghaddab. *Comparaison d'empreintes à grande variabilité intra-classe et à faible variabilité inter-classes*. PhD thesis, École Nationale des Sciences de l'Informatique - Université de la Manouba, Décembre 2017.
- [GHvK02] Joachim Gudmundsson, Mikael Hammar, and Marc van Kreveld. Higher order delaunay triangulations. *Computational Geometry*, 23(1) :85–98, 2002.
- [GJK17a] Mohamed Hedi Ghaddab, Khaled Jouini, and Ouajdi Korbaa. Fast and accurate fingerprint matching using expanded delaunay triangulation. In *14th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2017, Hammamet, Tunisia, October 30 - Nov. 3, 2017*, pages 751–758. IEEE Computer Society, 2017.
- [GJK17b] Mohamed Hedi Ghaddab, Khaled Jouini, and Ouajdi Korbaa. Fusion de minuties pour une reconnaissance efficiente des empreintes digitales. In *Colloque sur l'Optimisation et les Systèmes d'Information*, 2017.
- [GL02] Seth Gilbert and Nancy Lynch. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News*, 33(2) :51–59, jun 2002.
- [GMM⁺20] Heitor Murilo Gomes, Jacob Montiel, Saulo Martiello Mastelini, Bernhard Pfahringer, and Albert Bifet. On ensemble techniques for data stream regression. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.
- [HBDB18] P. Helber, B. Bischke, A. Dengel, and D. Borth. Introducing eurosat : A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS*, page 204–207, 2018.
- [HBDB19] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat : A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 12(7) :2217–2226, 2019.
- [Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963.
- [HPS⁺18] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on*

BIBLIOGRAPHIE

- Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
- [HZC⁺17] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets : Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.
- [IGD11] Elena Ikonomovska, João Gama, and Sašo Džeroski. Learning model trees from evolving data streams. *Data mining and knowledge discovery*, 23(1) :128–168, 2011.
- [IGD14] Elena Ikonomovska, João Gama, and Sašo Džeroski. Online tree-based ensembles and option trees for regression on evolving data streams. *Neurocomputing*, 150, 11 2014.
- [IGZD11] Elena Ikonomovska, João Gama, Bernard Zenko, and Saso Dzeroski. Speeding-up hoeffding-based regression trees with options. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 537–544. Omnipress, 2011.
- [Jac19] Montiel Lopez Jacob. *Fast and slow machine learning*. PhD thesis, Télécom ParisTech, France, 2019.
- [JJH⁺22] Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furui Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. PromptBERT : Improving BERT sentence embeddings with prompts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [JKK25a] Farah Jemili, Khaled Jouini, and Ouajdi Korbaa. Detecting unknown intrusions from large heterogeneous data through ensemble learning. *Intelligent Systems with Applications (ISWA)*, 200465, 2025.
- [JKK25b] Farah Jemili, Khaled Jouini, and Ouajdi Korbaa. Intrusion detection based on concept drift detection and online incremental learning. *International Journal of Pervasive Computing and Communications*, 21(1) :81–115, 2025.
- [JK23] Khaled Jouini and Ouajdi Korbaa. Drift-driven regression for predicting the evolution of pandemics. In *Proceedings of the 20th International Conference on Applied Computing*, pages 77–84, Lisbon, Portugal, 2023. IADIS Press.
- [JMK21] Khaled Jouini, Mohamed Hédi Maaloul, and Ouajdi Korbaa. Real-time, cnn-based assistive device for visually impaired people. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6, 2021.
- [Jou17] Khaled Jouini. Distorted replicas : Intelligent replication schemes to boost I/O throughput in document-stores. In *IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 25–32, 2017.
- [Jou22] Khaled Jouini. Aggregates selection in replicated document-oriented databases. *J. Inf. Sci. Eng.*, 38(2) :479–496, 2022.

BIBLIOGRAPHIE

- [JQD11] Alekh Jindal, Jorge-Arnulfo Quiané-Ruiz, and Jens Dittrich. Trojan data layouts : right shoes for a running elephant. In *ACM Symposium on Cloud Computing*, page 21. ACM, 2011.
- [KA15] G. Karnitis and G. Arnicans. Migration of relational database to document-oriented database : Structure denormalization and data transformation. In *2015 7th International Conference on Computational Intelligence, Communication Systems and Networks*, pages 113–118, 2015.
- [Kha21] Anant Khandelwal. Fine-tune longformer for jointly predicting rumor stance and veracity. In *CODS-COMAD ’21*, page 10–19, New York, NY, USA, 2021. Association for Computing Machinery.
- [KJKU23] Asad Ullah Khan, Nadeem Javaid, Muhammad Asghar Khan, and Insaf Ullah. A blockchain scheme for authentication, data sharing and nonrepudiation to secure internet of wireless sensor things. *Cluster Computing*, 26(2) :945–960, 2023.
- [KLZ18] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one : Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA, aug 2018. Association for Computational Linguistics.
- [KMB18] Nicolas Kourtellis, Gianmarco De Francisci Morales, and Albert Bifet. Large-scale learning from data streams with apache SAMOA. *CoRR*, abs/1805.11477, 2018.
- [KU95] Arthur M. Keller and Jeffrey D. Ullman. A version numbering scheme with a useful lexicographical order. In *ICDE*, pages 240–248. IEEE Computer Society, 1995.
- [LBA07] Xuefeng Liang, Arijit Bishnu, and Tetsuo Asano. A robust fingerprint indexing scheme using minutia neighborhood structure and low-order delaunay triangles. *IEEE Transactions on Information Forensics and Security*, 2(4) :721–733, 2007.
- [Ley09] Michael Ley. DBLP - some lessons learned. *Proc. VLDB Endow.*, 2(2) :1493–1500, 2009.
- [LMN⁺19] Songqian Li, Kun Ma, Xuewei Niu, Yufeng Wang, Ke Ji, Ziqiang Yu, and Zhenxiang Chen. Stacking-based ensemble learning on low dimensional features for fake news detection. In *2019 IEEE 21st International Conference on High Performance Computing and Communications*, 2019.
- [Lom98] David B. Lomet. B-tree page size when caching is considered. *SIGMOD Rec.*, 27(3) :28–32, 1998.
- [Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.
- [LYZ⁺24] Yudong Liu, Xiaoyu Yang, Xi Zhang, Zhihao Tang, Zongyi Chen, and Liwen Zheng. Predicting rumor veracity on social media with cross-channel interaction of multi-task. *Neural Computing and Applications*, pages 1–12, 02 2024.
- [LZC⁺21] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit : Bringing locality to vision transformers. *CoRR*, abs/2104.05707, 2021.

BIBLIOGRAPHIE

- [LZS19] Quanzhi Li, Qiong Zhang, and Luo Si. eventAI at SemEval-2019 task 7 : Rumor detection on social media by exploiting content, user credibility and propagation information. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859, Minneapolis, Minnesota, USA, jun 2019. Association for Computational Linguistics.
- [LZZ⁺23] Q. Liu, Y. Zhang, W. Zhou, X. Jiang, W. Zhou, and M. Zhou. Adaptive class incremental learning-based iot intrusion detection system. *Computer Engineering*, 49(2) :169–174, 2023.
- [Ma19] Edward Ma. NLP Augmentation. <https://github.com/makcedward/nlpaug>, 2019. [Accessed : 2024-05-15].
- [MC23] Julian Marx and Marc Cheong. Decentralised social media : Scoping review and future research directions. In *Australasian Conference on Information Systems*, 12 2023.
- [MFMT22] E. Mahdavi, A. Fanian, A. Mirzaei, and Z. Taghiyarrenani. Itl-ids : Incremental transfer learning for intrusion detection systems. *Knowledge-Based Systems*, 253 :109542, 2022.
- [MGS⁺22] Chaitanya Manapragada, Heitor M. Gomes, Mahsa Salehi, Albert Bifet, and Geoffrey I. Webb. An eager splitting strategy for online decision trees in ensembles. *Data Mining and Knowledge Discovery*, 36(2) :566–619, March 2022.
- [MKP⁺22] Rathnamma V. Mydukuri, Suresh Kallam, Rizwan Patan, Fadi Al-Turjman, and Manikandan Ramachandran. Deming least square regressed feature selection and gaussian neuro-fuzzy multi-layered data classifier for early COVID prediction. *Expert Syst. J. Knowl. Eng.*, 39(4), 2022.
- [MKS23] Luis Miralles-Pechuán, Ankit Kumar, and Andrés L. Suárez-Cetrulo. Forecasting COVID-19 cases using dynamic time warping and incremental machine learning methods. *Expert Syst. J. Knowl. Eng.*, 40(6), 2023.
- [MMC⁺02] Dario Maio, Davide Maltoni, Raffaele Cappelli, James L. Wayman, and Anil K. Jain. Fvc2000 : Fingerprint verification competition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3) :402–412, 2002.
- [MMC⁺04] Dario Maio, Davide Maltoni, Raffaele Cappelli, Jim L Wayman, and Anil K Jain. Fvc2004 : third fingerprint verification competition. In *Biometric Authentication*, pages 1–7. Springer, 2004.
- [MMJP09] Davide Maltoni, Dario Maio, Anil Jain, and Salil Prabhakar. *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.
- [MMMS04] Hayet Mouss, D Mouss, N Mouss, and L Sefouhi. Test of page-hinckley, an approach for fault detection in an agro-alimentary production system. In *2004 5th Asian Control Conference (IEEE Cat. No. 04EX904)*, volume 2, pages 815–818. IEEE, 2004.
- [MPGBGRAR12] Miguel Angel Medina-Pérez, Milton García-Borroto, Andres Eduardo Gutierrez-Rodríguez, and Leopoldo Altamirano-Robles. Improving fingerprint verification using minutiae triplets. *Sensors*, 12(3) :3418–3437, 2012.

BIBLIOGRAPHIE

- [MRBA18] Jacob Montiel, Jesse Read, Albert Bifet, and Talel Abdessalem. Scikit-multiflow : A multi-output streaming framework. *Journal of Machine Learning Research*, 19(72) :1–5, 2018.
- [MWS18] Chaitanya Manapragada, Geoffrey I. Webb, and Mahsa Salehi. Extremely fast decision tree. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1953–1962. ACM, 2018.
- [Nak09] Satoshi Nakamoto. Bitcoin : A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260, 2009.
- [NLT] NLTK.org. Natural Language Toolkit. <https://github.com/nltk/nltk>. [Accessed : 2021-05-15].
- [NPZH20] Maxim Neumann, André Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *CoRR*, abs/1911.06721, 2020.
- [Ope] OpenCV. <https://opencv.org/>. [Accessed : 2024-12-01].
- [OV24] Kenniy Olorunnimbe and Herna Viktor. Ensemble of temporal transformers for financial time series. *Journal of Intelligent Information Systems*, 62(4) :1087–1111, August 2024.
- [PDC⁺19] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanazashi, Tim Kaler, and Charles E. Leiserson. Evolvegen : Evolving graph convolutional networks for dynamic graphs. *ArXiv*, abs/1902.10191, 2019.
- [PGT⁺15] Daniel Peralta, Mikel Galar, Isaac Triguero, Daniel Paternain, Salvador García, Edurne Barrenechea, José M Benítez, Humberto Bustince, and Francisco Herrera. A survey on fingerprint minutiae-based local matching for verification and identification : Taxonomy and experimental evaluation. *Information Sciences*, 315 :67–87, 2015.
- [PN04] Giuseppe Parziale and Albert Niel. A fingerprint matching using minutiae triangulation. In *Biometric Authentication*, pages 241–248. Springer, 2004.
- [PR17] Dean Pomerleau and Delip Rao. The fake news challenge : Exploring how artificial intelligence technologies could be leveraged to combat fake news. <http://www.fakenewschallenge.org/>, 2017. [Accessed : 2024-04-01].
- [pro] New york times news provenance project. <https://newsprovenanceproject.com>. [Accessed : 2024-10-31].
- [PVG⁺11a] F. Pedregosa, 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, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.
- [PVG⁺11b] F. Pedregosa, 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, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

BIBLIOGRAPHIE

- [RASR17] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264, 2017.
- [RKH⁺24] Saddaf Rubab, Muhammad Attique Khan, Ameer Hamza, Hussain Mobarak Albarakati, Oumaima Saidani, Amal Alshardan, Areej Alasiry, Mehrez Marzougui, and Yunyoung Nam. A novel network-level fusion architecture of proposed self-attention and vision transformer models for land use and land cover classification from remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17 :13135–13148, 2024.
- [RLRJ17] Vincent Reniers, Dimitri Van Landuyt, Ansar Rafique, and Wouter Joosen. Schema design support for semi-structured data : Finding the sweet spot between NF and De-NF. In *2017 IEEE International Conference on Big Data*, pages 2921–2930, 2017.
- [RM07] Arun Ross and Rajiv Mukherjee. Augmenting ridge curves with minutiae triplets for finger-print indexing. In *Defense and Security Symposium*, pages 65390C–65390C. International Society for Optics and Photonics, 2007.
- [SA20] Valeriya Slovikovskaya and Giuseppe Attardi. Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1211–1218, Marseille, France, May 2020. European Language Resources Association.
- [SF12] Pramod J. Sadalage and Martin Fowler. *NoSQL Distilled : A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley Professional, Upper Saddle River, NJ, 1st edition, 2012. ISBN : 978-0-321-82662-6.
- [Shu19] Kai Shu. FakeNewsNet. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UEMMHS>, 2019. [Accessed : 2024-05-15].
- [SJK22] Ilhem Salah, Khaled Jouini, and Ouajdi Korbaa. Augmentation-based ensemble learning for stance and fake news detection. In *Advances in Computational Collective Intelligence - 14th International Conference, ICCC 2022, Hammamet, Tunisia, September 28-30, 2022, Proceedings*, volume 1653 of *Communications in Computer and Information Science*, pages 29–41. Springer, 2022.
- [SJK23] Ilhem Salah, Khaled Jouini, and Ouajdi Korbaa. On the use of text augmentation for stance and fake news detection. *Journal of Information and Telecommunication*, 0(0) :1–17, 2023.
- [SJPK24] Ilhem Salah, Khaled Jouini, Cyril-Alexandre Pachon, and Ouajdi Korbaa. Connecting the dots between stance and fake news detection with blockchain, proof of reputation, and the hoeffding bound. *Clust. Comput.*, 27(9) :13395–13405, 2024.
- [SKF21] Ciarán Shorten, Taghi M. Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of Big Data*, 8(101), 2021.
- [SKZ⁺21] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit ? data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270, 2021.

BIBLIOGRAPHIE

- [SLS⁺09] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. Online random forests. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1393–1400, 2009.
- [sol] Solidity. <https://soliditylang.org/>. [Accessed : 2024-10-31].
- [SRJ22] Z. Sun, G. Ran, and Z. Jin. Intrusion detection method based on active incremental learning in industrial internet of things environment. *Journal on Internet of Things*, 4(2) :99–111, 2022.
- [SRLB⁺21] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. The many dimensions of truthfulness : Crowdsourcing misinformation assessments on a multidimensional scale. *Inf. Process. Manage.*, 58(6), nov 2021.
- [STVS⁺21] Robiert Sepúlveda Torres, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Sanz. Headlinestancechecker : Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, 71 :100660, 09 2021.
- [Sur05] James Surowiecki. *The Wisdom of Crowds*. Anchor Books, 1st edition, 2005.
- [TP22] R. Thakur and P. Panse. Classification performance of land use from multispectral remote sensing images using decision tree, k-nearest neighbor, random forest and support vector machine using eurosat da. *Orig. Res. Pap. Int. J. Intell. Syst. Appl. Eng. IJISAE*, 2022(1s) :67–77, 2022.
- [tpc] The TPC-H benchmark. <http://www.tpc.org/tpch>. [Accessed : 2020-01-15].
- [TTK⁺23] A. Temenos, N. Temenos, M. Kaselimi, A. Doulamis, and N. Doulamis. Interpretable deep learning framework for land use and land cover classification in remote sensing using shap. *IEEE Geosci. Remote Sens. Lett.*, 20 :1–5, 2023.
- [TZJ18] Zhou Tianyu, Miao Zhenjiang, and Zhang Jianhu. Combining cnn with hand-crafted features for image classification. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 554–557, 2018.
- [UBEP07] Tamer Uz, George Bebis, Ali Erol, and Salil Prabhakar. Minutiae-based template synthesis and matching using hierarchical delaunay triangulations. In *Biometrics : Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1–8. IEEE, 2007.
- [UBEP09] Tamer Uz, George Bebis, Ali Erol, and Salil Prabhakar. Minutiae-based template synthesis and matching for fingerprint authentication. *Computer Vision and Image Understanding*, 113(9) :979–992, 2009.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [WCW20] Jibang Wu, Renqin Cai, and Hongning Wang. Déjà vu : A contextualized temporal attention mechanism for sequential recommendation. *WWW '20*, New York, NY, USA, 2020. Association for Computing Machinery.

BIBLIOGRAPHIE

- [WXJ⁺23] Xia Wang, Hao Xie, Shouling Ji, Li Liu, and Debin Huang. Blockchain-based fake news traceability and verification mechanism. *Helijon*, 9(7) :e17084, 2023.
- [WZX⁺24] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. Mtp : Advancing remote sensing foundation model via multi-task pretraining. *arXiv preprint arXiv :2403.13430*, 2024.
- [XDH⁺19] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation. *CoRR*, abs/1904.12848, 2019.
- [YLJ18] Jaejun Yoo, Ki-Hoon Lee, and Young-Ho Jeon. Migration from RDBMS to NoSQL using column-level denormalization and atomic aggregates. *Journal of Information Science and Engineering*, 34(1) :243–259, 2018.
- [YU23] Tolga Yilmaz and Ozgur Ulusoy. Modeling and mitigating online misinformation : A suggested blockchain approach. *arXiv preprint arXiv :2303.10765*, 2023.
- [YWZ⁺18] X. Yuan, R. Wang, Y. Zhuang, K. Zhu, and J. Hao. A concept drift based ensemble incremental learning approach for intrusion detection. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 350–357, Halifax, NS, Canada, 2018.