Introduire l'IA dans la lutte contre la fraude : Comment choisir et convaincre ?

F. Dama[1], R. Sleiman[1], S. Bellart[1]

[1] Centre de Recherche et d'Innovation, Talan France

fatoumata.dama@talan.com, rita.sleiman@talan.com, steve.bellart@talan.com

Résumé

La lutte contre le blanchiment constitue une priorité pour les institutions financières et les conduit à exercer une surveillance permanente. Face aux limites des outils traditionnels, les algorithmes d'Intelligence Artificielle (IA) représentent une opportunité en permettant la conception de dispositifs performants et flexibles. Ces travaux réalisent un benchmark des modèles d'IA (ML/DL) de détection de fraudes financières et discutent de leur intégration en pratique.

Mots-clés

Blanchiment d'argent, Machine Learning, Deep Learning, XAI (Intelligence Artificielle eXplicable), Données financières, Augmentation de données

Abstract

Combatting money laundering is a priority for financial institutions, leading them to implement continuous monitoring. Given the limitations of traditional tools, Artificial Intelligence (AI) algorithms represent an opportunity by enabling the design of effective and flexible systems. This work benchmarks AI (ML/DL) models for financial fraud detection, and discusses their practical integration.

Keywords

Money Laundering, Machine Learning, Deep Learning, XAI (Explainable Artificial Intelligence), Financial Data, Data Augmentation

1 Introduction

La lutte contre le blanchiment de capitaux et le financement du terrorisme (LCB-FT) représente un enjeu majeur pour la stabilité financière et la sécurité internationale. Il est estimé que 2 à 5% du PIB mondial est compromis par des activités de blanchiment d'argent chaque année, mettant en lumière l'urgence et la gravité de ce fléau, comme le montre certains modèle d'estimations [7]. Dans ce contexte, un cadre juridique international, incluant les recommandations du Groupe d'action financière (GAFI), a été établi pour contraindre les institutions financières à adopter des mesures de vigilance renforcée.

L'une de ces mesures essentielles est la norme *Know Your Customer* (KYC) [14], qui oblige les institutions financières à vérifier l'identité de leurs clients et à évaluer et surveiller en continu les risques associés à ces derniers. La conformité aux exigences KYC est cruciale pour prévenir le blanchiment d'argent, le financement du terrorisme et d'autres formes de fraude financière.

Toutefois, les systèmes LCB-FT actuels, principalement fondés sur des règles prédéfinies, se confrontent à des limites notables. Leur rigidité et leur incapacité à s'ajuster aux tactiques de blanchiment en constante évolution se traduisent par un taux élevé de faux positifs, nuisant à l'efficacité des processus de détection. Dans ce cadre, l'intelligence artificielle (IA) se présente comme une solution prometteuse, offrant des capacités de détection améliorées et adaptatives grâce à des algorithmes avancés d'apprentissage automatique et de *deep learning*. Néanmoins, leur adoption pour la détection de fraude se confronte à plusieurs défis, notamment la nécessité de données financières souvent restreintes et déséquilibrées, et la complexité de comprendre le fonctionnement souvent opaque des modèles d'IA les plus performants.

Notre étude vise à évaluer et comparer diverses méthodes d'IA dans le domaine du LCB-FT, en analysant leur performance sur des jeux de données réels et synthétiques. Nous évaluons la capacité de détection des modèles et leur aptitude à réduire les faux positifs qui représentent un coût opérationnel important pour les établissements financiers. En outre, l'explicabilité intrinsèque des modèles est examinée et discutée pour assurer la transparence et renforcer la confiance des utilisateurs, un aspect crucial dans l'adoption de systèmes d'IA dans des domaines aussi sensibles que la LCB-FT.

La structure de cet article comprend une revue des travaux antérieurs pertinents, la présentation de notre méthodologie et de nos résultats, suivie d'une discussion sur les implications et les perspectives futures de notre recherche.

2 Travaux antérieurs

Au cours de la dernière décennie, plusieurs travaux de recherche se sont intéressés à l'utilisation de l'IA pour concevoir de nouveaux outils de détection de fraudes financières dont le blanchiment de capitaux [11][4]. Une analyse de la littérature permet d'identifier deux grands groupes de modèles : les modèles supervisés entraînés sur des données labélisées (normale/suspecte) et les modèles non-supervisés

qui cherchent à séparer les données en différents groupes homogènes [4].

Dans cette étude nous nous intéressons aux modèles supervisés qui sont entraînés à détecter les *patterns* indicatifs de blanchiment spécifiés dans la base d'apprentissage. L'application de ces modèles dans la LCB-FT se divise principalement en deux grandes classes, à savoir les modèles de *Machine Learning* traditionnels (ML) et les modèles de *Deep Learning* (DL). Les modèles classiques de ML comprennent des algorithmes comme les arbres de décision [15], les modèles ensemble par Bagging (Random Forest, Extra Trees) [17], les arbres boostés (XGB, LGBM, CatBoost) [1], et les modèles probabilistes (Logistic Regression, Naive Bayes) [12]. De l'autre côté, les modèles de DL incluent des modèles comme les Perceptrons multicouches (MLP), les Auto-Encodeurs, et les graphes de réseaux de neurones (GNN, GCN)[16, 3].

D'après les études recensées, les modèles reposant sur des arbres sont largement utilisés en comparaison avec les modèles de DL. Ceci s'explique par plusieurs facteurs : d'une part, ces modèles génèrent des règles explicites utilisées pour effectuer les prédictions (normale/suspecte) ; d'autre part, les modèles de DL sont perçus comme des boîtes noires, ce qui complique l'interprétation de leurs prédictions, un défi significatif dans un domaine aussi critique.

Malgré l'existence de plusieurs études abordant l'apport de l'IA à la LCB-FT, ces dernières présentent plusieurs limites qui nécessitent d'être prises en considération dans des prochaines études. En effet, la majorité des études se concentrent sur l'utilisation de types de modèles bien déterminés, ce qui limite la comparaison et l'évaluation des performances entre les modèles cités. En plus, les conditions expérimentales souvent hétérogènes d'une étude à l'autre compliquent la généralisation des résultats et l'évaluation fiable des différentes approches. Ceci souligne l'importance de recherches plus approfondies qui explorent une plus grande diversité d'algorithmes avec une standardisation des protocoles expérimentaux.

Ainsi, bien que l'étude de l'usage de modèles d'apprentissage automatique dans la lutte contre la fraude a déjà été traité dans la littérature [9] en y montrant notamment les défis a relever pour avoir des modèles efficaces, notre étude vise à compléter ces travaux, en proposant une comparaison entre 17 modèles d'IA, tout en ayant un regard sur les difficultés vis-à-vis de leur adoption. Nos travaux intègrent donc également l'aspect explicabilité peu discuté dans la littérature, mais fondamental pour des raisons de conformité, de transparence et pour l'approbation de ces approches dans un secteur aussi critique.

3 Expérimentations

3.1 Modèles et Datasets

Dans nos expérimentations, nous avons considéré **17 modèles** de détection de fraude dont 13 modèles de *Machine Learning* et 4 modèles *Deep Learning*.

Modèles ML. Arbre de décision, les modèles ensemble par Bagging (Random Forest, BaggingClassifier et Extra Trees), les arbres boostés (AdaBoost, Logiboost, XGBoost, LightGBM et CatBoost), les modèles probabilistes (Naive Bayes et Logistic Regression), le modèle Support Vector Machine (SVM) et le modèle des K plus proches voisins (KNN).

Modèles DL. Les modèles Perceptron et Perceptron multicouches (MLP) avec 1, 2 ou 3 couches cachées.

Les modèles ont été évalués sur 5 jeux de données réelles et synthétiques, présentés dans le tableau 1. Les données sont des transactions bancaires (transferts, payements, dépôts, ...) labélisées normales (la classe 0) ou suspectes (la classe 1). Remarquons la faible proportions de transactions suspectes dans les 5 jeux de données. Il est bien connu que le déséquilibre entre les classes peut être source de biais dans les modèles. Afin d'éviter un tel biais, la méthode d'échantillonnage SMOTE (*Synthetic Minority Oversampling Technique*) [6] a été utilisée pour équilibrer les données en générant des transactions suspectes supplémentaires grâce à une technique d'interpolation linéaire.

Dataset	n_{feat}	n_{obs}	p_{fraude}
CreditCard	30	284 807	0.2%
Ethereum	48	9 840	22%
Bitcoin	167	203 796	10%
IBM ALMSim	8	250 000	0.7%
Mobile Money	11	250 000	3%

TABLE 1 – Description des datasets. De gauche à droite : le nombre de features, d'observations et la ratio de fraudes (https ://www.kaggle.com/datasets). Les 3 premiers datasets sont composés de données réelles et les 2 derniers sont composés de données synthétiques.

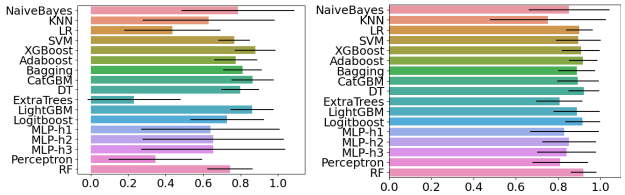
3.2 Protocole expérimental

Les 17 modèles considérés dans nos expérimentations ont été évalués sur les 5 datasets précédemment présentés. Chaque dataset a été découpé en deux parties : un jeu d'entraînement (80%) et un jeu de test (20%). Les hyperparamètres des différents modèles (la profondeur des arbres, le taux d'apprentissage, les termes de régularisation, le nombre K de voisins proches, ...) ont été calibrés par la méthode de la validation croisée à 5 champs. Les modèles ensemble sont composés de 200 arbres de décision.

Les métriques d'évaluation considérées sont la **fiabilité** et **l'efficacité opérationnelle**, définies ci-dessous. La fiabilité décrit la capacité du modèle à identifier les transactions suspectes. Tandis que l'efficacité opérationnelle décrit sa capacité à réduire le volume des fausses alarmes (faux positifs) et réduire par la même occasion le coût opérationnel lié au traitement manuel des alarmes.

$$\text{Fiabilité } (\textit{Recall}) = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs + Faux Négatifs}} \tag{1}$$

$$\text{Efficacité } (\textit{Precision}) = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs + Faux Positifs}} \tag{2}$$



(a) Données brutes (b) SMOTE

FIGURE 1 – Moyennes et écart-types des scores de fiabilité (calculées sur 5 datasets) obtenus par les modèles entraînés sur : (a) les données brutes ; (b) les données équilibrées par la méthode SMOTE.

3.3 Résultats et Analyse

3.3.1 Données brutes versus SMOTE

La figure 1 présente les scores de fiabilité obtenus par les modèles testés avec les deux modalités : (a) modèles entraînés sur les données brutes (déséquilibrées) ; et (b) modèles entraînés sur les données équilibrées par la méthode SMOTE pour obtenir 50% de cas de fraudes.

Les résultats montrent une amélioration notable des capacités de détection des modèles (entre $3 - 57\%$) lorsque les données d'apprentissage sont équilibrées. Par ailleurs, les arbres de décision boostés (XGBoost, CatGBM et LightGBM) montrent une certaine robustesse au déséquilibre entre les classes avec une amélioration de 3%.

3.3.2 Cartographie des modèles

La figure 2 présente la moyenne des performances obtenues par les 17 modèles testés sur les 5 datasets considérés. Les jeux d'apprentissage ont été préalablement équilibrés par la méthode SMOTE.

Les résultats montrent une bonne capacité de détection des transactions suspectes pour l'ensemble des modèles testés avec un score de fiabilité supérieur à 75%. Par ailleurs, chaque modèle testé surpasse certains systèmes experts (basés sur des règles prédéfinies) avec un score d'efficacité opérationnelle au moins égal à 15% contre seulement 5% pour ces derniers. Les arbres de décision boostés obtiennent les meilleures performances (jusqu'à 90% de fiabilité et d'efficacité opérationnelle) et surpassent significativement les modèles de *Deep Learning*, corroborant l'avantage de ces modèles sur des données tabulaires [8]. Il reste à comparer ces résultats avec les systèmes experts actuels, mais cela est difficile, les informations à leur propos n'étant pas accessibles en raison de la sensibilité du secteur.

Du point de vue de l'explicabilité, les modèles à ensembles d'arbres offrent une meilleure interprétabilité que les modèles basés sur les réseaux de neurones, malgré la supériorité habituelle de ces derniers en termes de précision. Nos expériences démontrent une exception à cette règle générale, avec une performance supérieure des modèles ensemblistes, surtout ceux issus du *boosting*. Bien que les arbres de décision uniques soient intuitivement compréhensibles, la complexité des modèles ensemblistes en diminue la transparence. Néanmoins, l'avantage en précision des ensembles sur les arbres individuels est significatif. Nous sommes donc encouragés à continuer d'explorer l'explicabilité des modèles ensemblistes dans le cadre de la LCB-FT, en utilisant des outils d'explicabilité avancés pour faciliter leur acceptation par les experts.

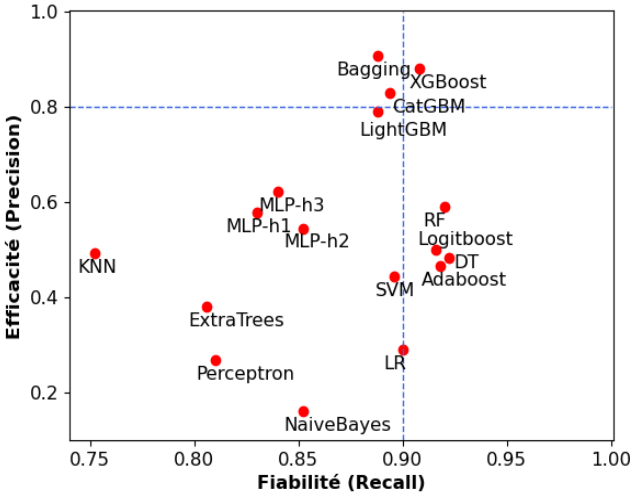


FIGURE 2 – Moyennes des performances (fiabilité et efficacité) obtenues par les 17 modèles testés sur les 5 datasets considérés.

4 Conclusion

Dans cette étude, nous avons évalué la performance de 17 modèles d'apprentissage automatique, incluant 13 modèles traditionnels de *Machine Learning* et 4 modèles de *Deep Learning*, testés sur 5 ensembles de données financières.

Les résultats obtenus montrent que les modèles testés possèdent une bonne capacité de détection des transactions suspectes (correspondant à du blanchiment d'argent). De plus, ces modèles se sont révélés plus efficaces que certains systèmes experts. Notamment, les arbres de décision boostés se sont distingués en atteignant jusqu'à 90% de fiabilité et d'efficacité opérationnelle. Ces modèles se révèlent être des outils précieux pour les institutions financières cherchant à améliorer leurs outils de détection des activités frauduleuses et optimiser le coût opérationnel de leurs systèmes de surveillance.

5 Perspectives et suite du projet

Le travail précédent s'est concentré sur l'évaluation de différents modèles d'apprentissage automatique dans la LCB-FT. D'après nos résultats, il apparaît que les modèles à ensembles d'arbres constituent une catégorie particulièrement prometteuse pour des analyses approfondies. Bien que nos recherches actuelles démontrent le potentiel de ces IA, l'intégration de ces technologies dans un domaine aussi sensible que la finance pose plusieurs défis. Il est crucial que ces modèles gagnent la confiance des experts du secteur et fassent l'objet d'une vérification rigoureuse avant leur déploiement. L'usage d'outils d'explicabilité pour l'IA peut jouer un rôle clé dans ce processus.

La recherche en explicabilité des modèles d'apprentissage automatique est récente et dynamique ayant déjà généré une multitude d'approches, chacune présentant ses avantages et ses inconvénients [10] [2]. Nous prévoyons pour la suite

d'exploiter certains de ces outils pour analyser les connaissances extraites par ces modèles. En particulier, nous envisageons d'utiliser des méthodes agnostiques au modèle (LIME, SHAP et Anchors), ainsi que des outils spécifiquement conçus pour les modèles à ensembles d'arbres (TreeSHAP et PyXAI).

L'objectif principal est de déterminer et d'examiner les règles ou *patterns* dans les données qui conduisent à la classification d'une action comme frauduleuse d'après les modèles d'IA. Nous proposons d'analyser ces règles à l'aide de méthodes statistiques, en examinant la longueur (le nombre d'assertions à respecter pour obéir à la règle) et la couverture (le nombre d'instances qui y obéissent) des règles, ainsi que leur contenu, en les comparant avec des connaissances établies et en sollicitant l'avis d'experts pour obtenir des évaluations constructives. Cette démarche vise à déterminer quel est la meilleure façon de les générer et si les règles calculées sont cohérentes ou si elles révèlent de nouveaux patterns de fraude qui sont plausibles mais auparavant non identifiés.

Finalement, nous soutenons que l'adoption généralisée des méthodes d'apprentissage automatique dépendra non seulement de leur performance statistique (bien que leur optimisation reste un défi important), mais aussi de la capacité à être comprises et contrôlées par les utilisateurs [13]. Les outils d'XAI améliorent la confiance en ces systèmes et nous pensons que développer des approches permettant aux utilisateurs d'interagir avec les systèmes d'IA et de les corriger en cas d'erreurs ou d'imprécisions est un enjeu important à leur mise en œuvre pratique. Nos travaux futurs se concentreront également sur l'intégration d'outils d'ajustement interactif [5], afin de renforcer l'efficacité et la fiabilité envers ces approches dans la LCB-FT.

Références

[1] A Ahmed. Anti-money laundering recognition through the gradient boosting classifier. *Academy of Accounting and Financial Studies Journal*, 25(5), 2021.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai) : What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99 :101805, 2023.

[3] Zhiyuan Chen, Waleed Mahmoud Soliman, Amril Nazir, and Mohammad Shorfuzzaman. Variational autoencoders and wasserstein generative adversarial networks for improving the anti-money laundering process. *IEEE Access*, 9 :83762–83785, 2021.

[4] Zhiyuan Chen, Le Dinh Van Khoa, Ee Na Teoh, Amril Nazir, Ettikan Kandasamy Karuppiah, and Kim Sim Lam. Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection : a review. *Knowledge and Information Systems*, 57 :245–285, 2018.

[5] Sylvie Coste-Marquis and Pierre Marquis. Rectifying binary classifiers. In *The 26th European Conference on Artificial Intelligence (ECAI'23)*. IOS Press, 2023.

[6] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data : progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61 :863–905, 2018.

[7] Joras Ferwerda, Alexander van Saase, Brigitte Unger, and Michael Getzner. Estimating money laundering flows with a gravity model-based simulation. *Scientific Reports*, 10(1) :18552, 2020.

[8] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data ? *Advances in neural information processing systems*, 35 :507–520, 2022.

[9] Prince Grover, Julia Xu, Justin Tittelfitz, Anqi Cheng, Zheng Li, Jakub Zablocki, Jianbo Liu, and Hao Zhou. Fraud dataset benchmark and applications, 2023.

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5) :1–42, 2018.

[11] Dattatray Vishnu Kute, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE access*, 9 :82300–82317, 2021.

[12] Mark E Lokanan. Predicting money laundering using machine learning and artificial neural networks algorithms in banks. *Journal of Applied Security Research*, 19(1) :20–44, 2024.

[13] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning : a state of the art. *Artificial Intelligence Review*, 56(4) :3005–3054, 2023.

[14] Plaid. What is kyc ? financial regulations to reduce fraud. https ://plaid.com, 2022.

[15] Omri Raiter. Applying supervised machine learning algorithms for fraud detection in anti-money laundering. *Journal of Modern Issues in Business Research*, 1(1) :14–26, 2021.

[16] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money laundering in bitcoin : Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv :1908.02591*, 2019.

[17] Wai Weng Lo, Gayan K Kulatilleke, Mohanad Sarhan, Siamak Layeghy, and Marius Portmann. Inspection-l : Self-supervised gnn node embeddings for money laundering detection in bitcoin. *arXiv e-prints*, pages arXiv–2203, 2022.