

Une typologie comparative des générations de modèles d'apprentissage automatique monolithiques et hybrides pour la prédiction financière : benchmark théorique face aux modèles économétriques traditionnels

A comparative typology of generations of machine learning models monolithic and hybrid for financial prediction a theoretical benchmark against traditional econometric models

Adil SLAMI-AMINE, (Étudiant-Chercheur)

*Laboratoire d'Analyse et Modélisation des Systèmes et Aide à la Décision (LAMSAD)
École Nationale des Sciences Appliquées Berrechid
Université HASSAN 1^{er} de Settat, Maroc*

Abdessamad DINE, (Enseignant-Chercheur)

*Laboratoire de Recherche en Economie, Gestion et Management des Affaires (LAREGMA)
École Nationale de Commerce et de Gestion de Settat
Université HASSAN 1^{er} de Settat, Maroc*

Boujemâa ACHCHAB, (Enseignant-Chercheur)

*Laboratoire d'Analyse et Modélisation des Systèmes et Aide à la Décision (LAMSAD)
École Nationale des Sciences Appliquées Berrechid
Université HASSAN 1^{er} de Settat, Maroc*

Adresse de correspondance :	Université Hassan Premier, Settat, Maroc École Nationale des Sciences Appliquées Berrechid Adresse : Avenue de l'université, B.P :218 Berrechid Tel : 05-22-32-47-58 & Fax : 05-22-53-45-30
Déclaration de divulgation :	Les auteurs n'ont pas connaissance de quelconque financement qui pourrait affecter l'objectivité de cette étude. Ils assument l'entière responsabilité de tout éventuel plagiat, de l'usage de l'intelligence artificielle dans la rédaction, ainsi que des résultats présentés dans cet article.
Conflit d'intérêts :	Les auteurs ne signalent aucun conflit d'intérêts.
Citer cet article	SLAMI-AMINE, A., DINE, A., & ACHCHAB, B. (2026). Une typologie comparative des générations de modèles d'apprentissage automatique monolithiques et hybrides pour la prédiction financière : benchmark théorique face aux modèles économétriques traditionnels. <i>International Journal of Accounting, Finance, Auditing, Management and Economics</i> , 6(7), 174–200. https://doi.org/10.5281/zenodo.20414574
Licence	Cet article est publié en open Access sous licence CC BY-NC-ND

Received: 06/05/2026

Accepted: 10/06/2026

International Journal of Accounting, Finance, Auditing, Management and Economics - IJAFAME

ISSN: 2658-8455

Volume 7, Issue 7 (2026)

Une typologie comparative des générations de modèles d'apprentissage automatique monolithiques et hybrides pour la prédiction financière : benchmark théorique face aux modèles économétriques traditionnels

Résumé

L'essor concomitant des architectures profondes, des Transformers financiers et des modèles hybrides a fragmenté la littérature de la prédiction des séries financières en sous-corpus difficilement comparables. Aucun cadre typologique consensuel ne permet, à ce jour, de classer rigoureusement les générations successives de modèles d'apprentissage automatique appliquées à la finance. Cette étude propose une typologie multi-niveaux et un benchmark théorique de cinq familles (économétrie classique, ML superficiel, DL profond, IA avancée et hybrides) articulés autour de sept dimensions discriminantes. Le protocole mobilise PCC sous PRISMA-ScR sur un corpus stabilisé de 58 études (1990–2026). Trois propositions formelles structurent la contribution. P1 affirme l'absence de supériorité uniforme d'une génération. P2 énonce la dominance conditionnelle des hybrides sous non-linéarité, ruptures de régime et faible rapport signal-bruit. P3 décrit, sous ordre partiel, le déclin de la compatibilité avec la théorie financière à mesure que croît la capacité représentationnelle, sauf en présence d'une hybridation théorie-données. La grille livre, outre une architecture conceptuelle, une grille décisionnelle opérationnelle pour les gestionnaires d'actifs, le model risk management bancaire et les régulateurs (SR 11-7, TRIM, projet de règlement IA de l'Union européenne). L'article s'inscrit dans la typologie de Gregor (2006, types II et IV) et fournit un instrument répliquable de sélection raisonnée de modèles.

Mots-clés : prédiction financière ; apprentissage automatique ; modèles hybrides ; benchmark théorique ; typologie ; hypothèses inductives ; économétrie ; deep learning ; Transformers ; interprétabilité.

Classification JEL : C38, C45, C53, C58, C63, G17

Type de papier : Recherche théorique

Abstract

The simultaneous rise of deep architectures, financial Transformers, and hybrid models has fragmented the literature on financial time-series prediction into sub-corpora that are hardly comparable. No consensual typological framework currently classifies the successive generations of machine-learning models applied to finance. This study proposes a multi-level typology and a theoretical benchmark of five families (classical econometrics, shallow ML, deep learning, advanced AI and hybrids) along seven discriminant dimensions. The protocol mobilizes PCC under PRISMA-ScR on a stabilized corpus of 58 studies (1990–2026). Three formal propositions structure the contribution. P1 no generation is uniformly superior; P2 hybrids dominate conditionally under nonlinearity, regime breaks and low signal-to-noise; P3 financial-theory compatibility decreases (partial order) as representational capacity grows, except under theory-data hybridization. Beyond conceptual architecture, the framework delivers an operational decision grid for asset managers, banking model risk management and regulators (SR 11-7, TRIM, EU AI Act). The article fits Gregor's (2006) typology (types II and IV) and provides a replicable instrument for the principled selection of models.

Keywords: financial forecasting; machine learning; hybrid models; theoretical benchmark; typology; inductive assumptions; econometrics; deep learning; Transformers; interpretability.

JEL Classification: C38, C45, C53, C58, C63, G17

Paper type: Theoretical Research

1. Introduction

La modélisation des séries financières connaît, depuis une vingtaine d'années, une mutation paradigmatique d'une ampleur comparable à celle qu'avait introduite la révolution économétrique des années 1970–1980. Les outils canoniques : modélisation ARIMA selon Box et Jenkins (1970), processus à hétéroscédasticité conditionnelle d'Engle (1982) et leurs extensions GARCH (Bollerslev, 1986), systèmes vectoriels autorégressifs (Sims, 1980), modèles à facteurs (Fama & French, 1993, 2015), ont longtemps fourni le langage de référence pour décrire la dynamique des rendements, de la volatilité, de la structure par terme des taux ou du risque de défaut. Ces approches partagent une caractéristique épistémologique forte. Elles s'appuient sur un cadre paramétrique explicite, dans lequel chaque coefficient possède une interprétation économique vérifiable et chaque hypothèse de structure (linéarité, stationnarité, normalité conditionnelle) peut être testée à partir des données.

Or, depuis le tournant des années 2000, la convergence entre la disponibilité de données financières granulaires, la croissance de la puissance de calcul et les progrès algorithmiques a déplacé le centre de gravité de la modélisation prédictive vers les techniques d'apprentissage statistique (Hastie, Tibshirani & Friedman, 2009 ; Mohri, Rostamizadeh & Talwalkar, 2018). Trois générations successives de modèles peuvent désormais être identifiées. Une première vague d'algorithmes superficiels (machines à vecteurs de support, forêts aléatoires, gradient boosting) héritière des fondations posées par Vapnik (1998). Une deuxième vague, centrée sur les architectures profondes, réseaux récurrents et leurs variantes LSTM (Hochreiter & Schmidhuber, 1997), convolutifs, autoencodeurs. Et une troisième vague, en cours d'expansion, mobilisant des modèles à attention (Vaswani et al., 2017), des modèles de fondation et des dispositifs de méta-apprentissage. Chacune de ces générations a fait l'objet, en finance, d'une littérature désormais abondante, Gu, Kelly et Xiu (2020) pour l'asset pricing, Bianchi, Büchner et Tamoni (2021) pour la prime de terme obligataire, López de Prado (2018) pour la gestion d'actifs, Heaton, Polson et Witte (2017) pour le deep learning financier, Israel, Kelly et Moskowitz (2020) pour la question même de la capacité à apprendre la finance.

Ce déploiement n'est pas neutre sur le plan épistémologique. La tradition économétrique s'enracine, peu ou prou, dans l'hypothèse d'efficience des marchés (Fama, 1970, 1991), elle-même réinterprétée par Lo (2004) à travers l'Adaptive Markets Hypothesis. Les approches algorithmiques, en revanche, font le pari inverse, que la prédictibilité résiduelle des séries financières, fût-elle ténue, instable et conditionnée, peut être captée par des fonctions hautement non linéaires que la statistique paramétrique ne sait pas spécifier a priori. Cette tension est moins une opposition stérile qu'un déplacement subtil du fardeau de la preuve. Là où l'économétrie demandait à la théorie de fournir une structure et aux données de la valider, l'apprentissage statistique demande aux données de fournir une structure et à la procédure d'optimisation de minimiser un risque empirique. Le débat n'est pas seulement méthodologique ; il engage la nature même de la connaissance produite.

Le revers de cette effervescence méthodologique se nomme fragmentation. La littérature comporte aujourd'hui plusieurs centaines d'études comparant deux ou trois familles de modèles sur des jeux de données particuliers, avec des protocoles de validation hétérogènes et des métriques disparates (Tang et al., 2022 ; Zhang, Sjarif & Ibrahim, 2023). Aucun cadre théorique unifié ne permet, à notre connaissance, de classer ces modèles selon des dimensions discriminantes explicites : capacité représentationnelle, biais inductif, exigence en données, interprétabilité, robustesse, coût computationnel, compatibilité avec la théorie financière. Plus problématique encore, la confusion fréquente entre performance empirique mesurée sur un échantillon et propriétés théoriques structurelles d'une famille de modèles conduit à des conclusions non généralisables, voire trompeuses, lorsqu'elles sont transposées à des contextes différents.

Trois manifestations concrètes de cette fragmentation méritent d'être soulignées. D'une part, la prolifération non taxonomisée des labels (modèle hybride, architecture profonde, modèle de fondation financier) masque souvent des différences structurelles considérables. Un ARIMA-LSTM séquentiel et un GARCH-MLP en stacking ne partagent ni les mêmes hypothèses, ni les mêmes propriétés de généralisation. D'autre part, les benchmarks empiriques publiés mobilisent rarement le même comparateur économétrique de référence, ce qui rend les comparaisons transversales hasardeuses. Enfin, l'explosion des architectures G3 (Transformers temporels, modèles de fondation, méta-apprenants) pose une difficulté nouvelle. Ces modèles, par leur capacité représentationnelle quasi universelle, brouillent les frontières mêmes que la typologie traditionnelle prétendait tracer.

À partir de cette problématique, trois lacunes doivent être nommées sans détour.

Premier gap, typologique. Aucune typologie consensuelle des générations de modèles ML appliqués à la finance ne fait, à ce jour, l'objet d'une formalisation rigoureuse au sens de Doty et Glick (1994). Or une typologie n'est pas un simple inventaire ; elle suppose une partition exhaustive et mutuellement exclusive de l'espace des modèles, étayée par des dimensions discriminantes justifiées et ordonnées. Les survols disponibles (Tang et al., 2022 ; Hoang & Wiegatz, 2022 ; Popa & Muresan, 2025) proposent des catégorisations utiles mais hétérogènes, sans cadre unifié. Popa et Muresan (2025) restent essentiellement descriptifs : ils alignent les techniques par domaine applicatif (marché, macroéconomie, firme) sans expliciter les dimensions de comparabilité ni leur ancrage théorique, et omettent la question de la commensurabilité des familles. Stempień et Ślepaczuk (2025) circonscrivent leur cartographie aux hybrides économétrie-ML et la pondèrent par la performance empirique rapportée, ce qui prolonge la fragmentation au lieu de la résorber. L'insuffisance commune tient à un déficit de réfutabilité, aucune des deux contributions ne formule de propositions testables au sens de Doty et Glick (1994).

Deuxième gap, comparatif. Les benchmarks existants demeurent essentiellement empiriques. L'étude de référence de Gu, Kelly et Xiu (2020) compare une douzaine de modèles ML sur les rendements actions américaines, mais ne théorise pas les sources structurelles de leurs différences de performance. À l'inverse, les contributions théoriques sur l'apprentissage statistique en général (Wolpert, 1996 ; Mohri et al., 2018) ne sont pas appliquées au contexte financier spécifique.

Troisième gap, épistémologique. La confrontation systématique entre tradition économétrique paramétrique, fondée sur la théorie économique, et tradition algorithmique non paramétrique, fondée sur l'optimisation du risque empirique, demeure largement sous-théorisée. Cerniglia et Fabozzi (2020) posent la question avec netteté, "Is there a conflict?", mais l'analyse mérite d'être prolongée. Une confrontation rigoureuse, au sens de Cerniglia et Fabozzi (2020) et conformément aux exigences de réfutabilité de Doty et Glick (1994), suppose au moins trois conditions absentes dans les travaux existants. Premièrement, l'explicitation comparée des hypothèses inductives (paramétriques d'un côté, architecturales et procédurales de l'autre) sur des dimensions commensurables. Deuxièmement, la formulation de propositions ordonnant les familles sous des conditions identifiables du DGP, plutôt que des affirmations globales du type « le ML domine ». Troisièmement, la mobilisation d'un cadre normatif (Gregor, 2006 ; Whetten, 1989) qui rattache la typologie à des critères d'originalité, d'utilité et de domaine de validité borné. Le présent article vise précisément cette triple exigence. Enfin, la limite corrélative de l'apprentissage statistique face aux exigences de causalité (Pearl, 2009) constitue une frontière épistémologique que la grille typologique doit nommer, à défaut de pouvoir la résoudre, ce dont le quatrième programme de recherche (section 6.4) tire les conséquences.

Ces trois gaps convergent vers une question de recherche unifiée.

Question principale (QR). *Comment construire un cadre typologique théorique permettant de comparer, sur des dimensions discriminantes explicites et justifiées, les générations successives*

de modèles d'apprentissage automatique : architectures monolithiques (superficielles, profondes, avancées) et architectures hybrides ML-économétrie, ainsi que leurs benchmarks économétriques de référence, dans le contexte de la prédiction financière ?

Cette question se décline en quatre sous-questions dont la résolution conjointe constitue l'ossature de l'article.

SQ1. Quelles dimensions discriminantes (épistémologiques, structurelles, computationnelles, économiques) permettent une comparaison rigoureuse de ces familles de modèles ?

SQ2. Quelles hypothèses inductives implicites (linéarité, stationnarité, indépendance, gaussianité, mémoire de long terme, équivariance) structurent chaque génération et façonnent leurs limites épistémologiques ?

SQ3. Sous quelles conditions théoriques (régime de marché, fréquence, horizon, dimensionnalité, rapport signal-bruit) les architectures hybrides surpassent-elles structurellement les architectures monolithiques ?

SQ4. Quels critères de réfutabilité, d'interprétabilité et de robustesse hors échantillon le cadre typologique doit-il intégrer pour répondre aux exigences de Gregor (2006) et de Whetten (1989) ?

Notre contribution se laisse lire à travers la grille classique de Whetten (1989), articulée selon quatre questions cardinales : what, how, why, who-where-when. What : une taxonomie multi-niveaux des modèles de prédiction financière distinguant cinq familles principales (économétrie, ML superficiel, DL profond, IA avancée, hybrides) et organisant chaque famille en sous-familles structurelles puis en instances opérationnelles. How : un cadre comparatif multi-critères opérationnalisable, susceptible d'être réinvesti par d'autres équipes pour mener des recherches empiriques cumulatives ; la grille à sept dimensions n'est pas seulement descriptive, elle est codée sur une échelle ordinale dont chaque modalité est rattachée à au moins deux références primaires de rang A. Why : une explicitation des hypothèses inductives (biais de représentation, biais de procédure, biais de données) propres à chaque génération, qui éclaire les sources structurelles des différences de performance. Who-where-when : un domaine de validité borné de manière transparente : prédiction de séries financières liquides, observables, en régime stationnaire ou modérément non stationnaire, sur des horizons court à moyen terme. Au regard de la typologie des théories proposée par Gregor (2006), l'article relève principalement du type II (theory for analyzing and classifying) et complète ce positionnement par une dimension de type IV (theory for explaining) dans la mesure où l'analyse des hypothèses inductives explique pourquoi telle famille de modèles tend à dominer dans tel contexte. La méthode adoptée combine deux traditions : l'analyse comparative théorique au sens de Ridder (2017) et la construction typologique au sens de Doty et Glick (1994). Sur le plan opérationnel, le protocole de revue mobilise PRISMA 2020 (Page et al., 2021) adapté à la version scoping review (Tricco et al., 2018), avec un cadre de questionnement PCC (Peters et al., 2020) plutôt que PICO-S, choix justifié dans la section 3.

L'article suit la structure IMRaD étendue. La section 2 propose la revue théorique articulant fondements économétriques et apprentissage statistique en finance. La section 3 expose la méthodologie d'analyse comparative théorique, le protocole PRISMA-ScR adopté et le cadre conceptuel structurant. La section 4, intitulée Construction théorique, déploie la taxonomie hiérarchique, le tableau comparatif multi-critères, la cartographie des hybridations et les propositions formelles. La section 5 propose une lecture transversale critique, recense les limites et formule des programmes de recherche dérivés. La section 6 conclut sur les implications théoriques, managériales et réglementaires.

2. Revue théorique

Cette section répond à une exigence éditoriale : articuler explicitement, avant la présentation du protocole, les théories qui structurent l'objet de recherche. Elle se déploie en trois temps : fondements économétriques (2.1), apprentissage statistique en finance (2.2), pont théorique vers la causalité financière (2.3), afin d'ancrer la construction typologique des sections 3 et 4 dans un socle conceptuel reconnu.

2.1. Fondements économétriques de la prédiction financière

Trois piliers structurent la tradition économétrique mobilisée dans la suite de l'article. Le premier est l'hypothèse d'efficience des marchés (Fama, 1970, 1991), sous ses formes faible, semi-forte et forte, qui contraint l'espace des rendements futurs prédictibles à partir de l'information passée et fonde la légitimité des modèles à mémoire courte comme ARMA. L'extension adaptative proposée par Lo (2004) relâche l'hypothèse statique d'efficience en autorisant des fenêtres locales de prédictibilité, sous l'effet d'une sélection naturelle entre stratégies. Le deuxième pilier est la théorie de l'absence d'arbitrage (Ross, 1976 ; Harrison & Kreps, 1979), qui interdit aux modèles prédictifs de produire des signaux exploitables sans risque correspondant. Cette contrainte oriente la grille typologique par la dimension D7 (compatibilité avec la théorie financière). Le troisième pilier est la critique de Lucas (1976), qui rappelle qu'un modèle estimé sur des comportements passés perd sa stabilité lorsque la politique économique ou la structure institutionnelle change ; cette critique cadre l'interprétation des ruptures de régime (Hamilton, 1989) que les modèles à changement de régime ou les hybrides cherchent à capter.

2.2. Apprentissage statistique appliqué à la prédiction financière

Le paradigme de l'apprentissage statistique substitue à la spécification paramétrique a priori une procédure de minimisation du risque empirique régularisé, sous bornes de complexité d'échantillon (Vapnik, 1998 ; Mohri, Rostamizadeh & Talwalkar, 2018). Trois résultats théoriques structurent son application financière. Premièrement, le théorème No Free Lunch (Wolpert, 1996) interdit toute supériorité universelle d'un algorithme, ce qui fonde la proposition P1 énoncée à la section 4.6. Deuxièmement, le théorème d'approximation universelle (Hornik, 1991 ; Cybenko, 1989) garantit qu'une famille de modèles suffisamment large peut approximer toute fonction continue sur un compact, propriété pertinente pour la classe G2 mais sans implication directe sur la généralisation hors échantillon. Troisièmement, les travaux récents sur le purged cross-validation (López de Prado, 2018) explicitent la dépendance temporelle entre observations financières et imposent des protocoles de validation spécifiques. Ces trois résultats, combinés à l'analyse des biais inductifs développée par Battaglia et al. (2018) en sciences de la décision, structurent l'interprétation de la grille présentée à la section 4.

2.3. Pont théorique vers la causalité financière

La grille typologique reste, à ce stade, corrélative, elle classe les modèles selon leur capacité à capter des régularités statistiques, non leur capacité à isoler des effets causaux. Pearl (2009) en a formulé la limite épistémologique avec netteté. Trois avancées récentes amorcent néanmoins un pont entre prédiction financière et inférence causale. Le Double Machine Learning de Chernozhukov et al. (2018) propose un estimateur orthogonal à deux étapes qui permet d'estimer un effet causal partiellement linéaire en présence de nuisance non paramétrique, particulièrement adapté aux questions de policy evaluation en finance macroéconomique. La représentation causale apprise (Schölkopf et al., 2021) reformule l'inférence causale comme problème de découverte de variables latentes invariantes sous interventions, ouvrant la voie à des modèles G3 contraints par des invariances économiques. Enfin, l'extension transportable

des estimateurs causaux à des marchés émergents (Bareinboim & Pearl, 2016) éclairer la question de la généralisation inter-géographique : angle mort majeur du corpus actuel, dominé par les marchés anglo-saxons. Ces trois axes orientent la quatrième perspective de recherche présentée à la section 6.4.

3. Méthodologie et fondement conceptuel

3.1. Choix méthodologique : analyse comparative théorique

Le choix méthodologique constitue, dans une recherche théorique, une partie intégrante de la contribution. Il convient donc de l'explicitier avant d'en dérouler les modalités. La présente étude relève de l'analyse comparative théorique, méthode formalisée notamment par Ridder (2017), distincte à la fois de la revue narrative (peu structurée), de la revue systématique au sens strict (Tranfield, Denyer & Smart, 2003), de la méta-analyse (qui suppose une commensurabilité empirique des effets) et de la scoping review généraliste. La spécificité de l'analyse comparative théorique tient à son objectif. Il ne s'agit pas d'agréger des résultats empiriques pour produire un effet moyen, mais de comparer des architectures conceptuelles selon des dimensions discriminantes explicites pour en dégager une typologie réfutable.

Plusieurs raisons concourent à ce choix. La nature de la contribution recherchée, typologique et analytique, interdit l'agrégation purement empirique, qui imposerait une commensurabilité métrique entre études dont nous savons précisément qu'elle fait défaut (Tang et al., 2022 ; Stempień & Ślepaczuk, 2025). Un design empirique appliqué aurait, par ailleurs, introduit une dépendance aux choix de jeux de données particuliers, limitant la portée généralisatrice du cadre. Enfin, ce déplacement aurait recoupé une littérature déjà abondante (Gu, Kelly & Xiu, 2020 ; Bianchi, Büchner & Tamoni, 2021) sans en combler les lacunes typologiques.

3.2. Protocole de constitution du corpus

3.2.1. Cadre de questionnement : PCC plutôt que PICO-S

Le cadre PICO-S (Population, Intervention, Comparator, Outcome, Study design), couramment mobilisé dans les revues systématiques cliniques, ne convient pas à une revue conceptuelle portant sur des architectures de modèles. Conformément aux travaux de Peters et al. (2020) et de Booth et Carroll (2015), nous adoptons le cadre PCC (Population, Concept, Context) qui présente trois avantages décisifs. Premièrement, il substitue le Concept : entité abstraite analysable comme une famille de modèles ou une dimension de comparaison, à l'Intervention, distinction essentielle en recherche conceptuelle. Deuxièmement, il reconnaît la contextualité des prédicats théoriques. Troisièmement, il bénéficie d'un protocole de reporting validé : PRISMA-ScR (Tricco et al., 2018) pour les revues théoriques et de portée.

3.2.2. Bases de données interrogées

La sélection des bases interrogées vise un compromis entre exhaustivité disciplinaire et qualité de référencement. Huit bases ont été mobilisées : Web of Science (Core Collection), Scopus, EconLit (EBSCO), IEEE Xplore, ACM Digital Library, SSRN, RePEc et arXiv (sections q-fin et cs.LG). L'inclusion de SSRN, RePEc et arXiv mérite une justification. En finance computationnelle, la prépublication précède fréquemment la publication peer-reviewed de plusieurs années, et l'omission de ces bases produirait un biais de récence documenté par Booth (2016).

3.2.3. Mots-clés et équation de recherche

L'équation booléenne reproductible structure la recherche autour de quatre blocs conceptuels reliés par AND : modèles ML/IA, architectures hybrides, prédiction financière,

comparaison/typologie. Les troncatures (*) maximisent le rappel.

TS = (
 ("machine learning" OR "deep learning" OR "neural network*" OR "artificial intelligence" OR "ensemble learning" OR "random forest" OR "gradient boosting" OR "XGBoost" OR "support vector machine" OR "LSTM" OR "GRU" OR "recurrent neural network" OR "convolutional neural network" OR "transformer" OR "attention mechanism" OR "graph neural network" OR "reinforcement learning" OR "meta-learning" OR "foundation model" OR "large language model" OR "hybrid model*" OR "LSTM-GARCH" OR "ARIMA-LSTM" OR "XGBoost-VAR")
AND
 ("financial forecast*" OR "stock predict*" OR "asset return* forecast*" OR "volatility forecast*" OR "return forecast*" OR "credit risk predict*" OR "bond yield forecast*" OR "exchange rate forecast*" OR "cryptocurrency predict*")
AND
 ("compar*" OR "benchmark*" OR "taxonomy" OR "typology" OR "review" OR "survey" OR "classification framework")
)
AND PY = 1990–2026
AND LA = (English OR French)
AND DT = (Article OR Review OR Proceedings Paper OR Book Chapter)

L'équation est datée (extraction de mars 2026), exportable et reproductible, conformément à PRISMA-S (Rethlefsen et al., 2021).

3.2.4. Critères d'éligibilité

La sélection des études repose sur une grille d'éligibilité à deux niveaux, pensée pour rendre chaque décision d'inclusion ou d'exclusion traçable et reproductible. Le premier niveau, de nature thématique, mobilise le cadre PCC afin de circonscrire l'objet substantiel de la revue, à savoir la prédiction financière fondée sur des modèles d'apprentissage automatique. Le second niveau, transversal, applique des filtres méthodologiques et bibliométriques qui s'imposent indépendamment du sujet traité : type de publication, période de couverture, langue et accessibilité du texte intégral. Pour chaque dimension, nous avons formulé un critère d'inclusion et son critère d'exclusion en miroir, accompagnés d'une justification ; ce choix d'une symétrie explicite vise à limiter l'arbitraire du tri et à permettre à un pair de reconstituer le périmètre retenu. Les tableaux 1 et 2 en présentent l'application opérationnelle.

Le tableau 1 décline le cadre PCC en cinq dimensions opératoires : une dimension de population, trois sous-dimensions conceptuelles (C₁ à C₃) et une dimension contextuelle. Chaque ligne associe une règle d'admission, sa contrepartie d'exclusion et la raison qui la sous-tend, de manière à articuler le périmètre substantiel autour de l'objet typologique de l'étude.

Les cinq dimensions fonctionnent comme un entonnoir de sélection dont les étapes se conditionnent l'une l'autre. La population restreint d'emblée le corpus aux travaux dont la finalité est la prédiction d'une grandeur financière (rendements, volatilité, prix, spreads, défaut de crédit ou indicateurs macro-financiers à valeur prédictive) et écarte les recherches centrées sur l'optimisation de portefeuille, la détection de fraude, le clustering ou l'allocation lorsque toute visée prédictive en est absente. La sous-dimension C₁ porte ensuite sur la nature du modèle : nous retenons les études qui mobilisent au moins une famille d'apprentissage automatique,

qu'il s'agisse du ML superficiel et des méthodes d'ensemble, de l'apprentissage profond, de l'IA avancée ou des architectures hybrides, et nous y rattachons l'usage des LLM et des GAN lorsqu'ils interviennent comme modules d'extraction de variables ; en revanche, les approches purement économétriques, dépourvues de composante ML, comme les modèles génératifs employés en finalité prédictive autonome, demeurent hors champ. La sous-dimension C₂ impose une exigence comparative : l'étude doit confronter explicitement au moins deux familles de modèles, opposer le ML à des références économétriques ou proposer un cadre typologique, faute de quoi elle ne nourrit pas l'ambition comparative de la revue. La sous-dimension C₃ conditionne l'admission à la qualité de la validation rapportée, en exigeant la présence d'au moins une métrique standardisée, d'une procédure d'évaluation robuste et d'un horizon de prévision explicite, ce qui conduit à exclure les résultats strictement in-sample ou aux conditions d'évaluation indéterminées. La dimension contextuelle, enfin, élargit la portée du dispositif à l'ensemble des marchés, classes d'actifs et fréquences, tout en réservant le cas des contextes non financiers aux seuls travaux offrant un transfert prédictif vers la finance.

Tableau 1. Critères PCC d'éligibilité.

Dimension PCC	Inclusion	Exclusion	Justification
Population (P) : prédiction financière	Études portant sur la prédiction de variables financières (rendements, volatilité, prix, spreads, défaut de crédit, indicateurs macro-financiers prédictifs)	Études centrées exclusivement sur l'optimisation de portefeuille pure, la classification de fraude, le clustering ou l'allocation sans composante prédictive	Délimitation rigoureuse du périmètre en cohérence avec l'objet typologique
Concept C ₁ : modèles ML et hybrides	Études exposant au moins une famille : ML superficiel/ensemble (SVM, RF, GBM, XGBoost, LightGBM) ; DL (MLP, RNN, LSTM, GRU, CNN, Transformer, TCN) ; IA avancée (GNN, RL, méta-apprentissage, foundation models) ; hybrides (LSTM-GARCH, XGBoost-VAR, ARIMA-LSTM) ; LLM/GAN comme modules d'extraction de features	Études mobilisant uniquement des modèles statistiques économétriques sans composante ML ; études déployant des modèles génératifs (GAN, LLM, diffusion) comme finalité prédictive autonome sans architecture ML d'intégration	Capture la diversité méthodologique du ML prédictif contemporain en excluant le périmètre génératif pur, hors champ
Concept C ₂ : comparabilité et positionnement typologique	Études comparant explicitement au moins deux familles ML, ou ML vs benchmarks économétriques, ou proposant un cadre typologique de modèles	Études sans comparateur explicite ni positionnement typologique	Condition sine qua non d'une revue comparative typologique
Concept C ₃ : critères de validation rapportés	Études rapportant au moins une métrique standardisée (RMSE, MAE, MAPE, R ² , R ² OOS, hit ratio, F1, Sharpe, Sortino, MDD), une procédure robuste (walk-forward, fenêtre glissante, purged k-fold) et un horizon explicite	Absence totale de métriques ; résultats exclusivement in-sample sans test hors échantillon ; horizon indéterminé	Garantit la comparabilité des contributions agrégées dans le cadre typologique
Context (Ctx)	Marchés développés, émergents et crypto-actifs ; toutes classes d'actifs ; toutes fréquences	Contextes non financiers sans transfert prédictif vers la finance	Contextualise la portée du cadre typologique

Source : élaboration par les auteurs.

Au-delà du périmètre substantiel, l'admissibilité d'une étude dépend de critères transversaux qui s'appliquent quel que soit son objet. Le tableau 2 réunit ces filtres méthodologiques et bibliométriques (type de publication, période, langue et accès) chacun étant assorti de ses bornes d'inclusion et d'exclusion ainsi que de sa justification.

Tableau 2. Critères transversaux (méthodologiques et bibliométriques).

Critère	Inclusion	Exclusion	Justification
Type de publication	Articles peer-reviewed indexés WoS/Scopus, chapitres d'ouvrages académiques, thèses doctorales, prépublications arXiv/SSRN/RePEc satisfaisant un seuil de qualité (> 50 citations, h-index auteur > 15, ou validation experte documentée)	Communications < 3 pages, résumés de congrès non développés, éditoriaux, lettres, billets de blog, contenus de vulgarisation	Garantit la qualité académique tout en intégrant la littérature grise pertinente (Booth, 2016)
Période	Janvier 1990 – mars 2026	Antérieur à 1990 sauf œuvres fondatrices mobilisées en discussion (Box & Jenkins 1970, Engle 1982, Bollerslev 1986, Vapnik 1995, Hochreiter & Schmidhuber 1997)	Couverture diachronique des trois générations ML
Langue	Anglais, français ; autres langues avec traduction certifiée ou validée par contrôle expert	Langues sans traduction vérifiable	Couverture internationale sans biais géographique injustifié
Accès	Texte intégral via abonnement institutionnel, libre accès, ou demande aux auteurs (relance unique à 14 jours)	Résumé seul ; texte non obtenu après relance	Exigence d'analyse méthodologique et résultative complète

Source : élaboration par les auteurs.

Les quatre critères du tableau 2 visent à concilier l'exigence de qualité et la complétude documentaire. Sur le plan du type de publication, nous admettons les articles évalués par les pairs et indexés dans Web of Science ou Scopus, les chapitres d'ouvrages académiques et les thèses doctorales, mais nous ouvrons aussi le corpus à la littérature grise : prépublications déposées sur arXiv, SSRN ou RePEc, à condition qu'un seuil de qualité soit franchi, qu'il s'agisse d'un nombre de citations supérieur à cinquante, d'un h-index d'auteur au-delà de quinze ou d'une validation experte documentée ; cette ouverture maîtrisée, conforme aux recommandations de Booth (2016), permet de capter des contributions récentes sans renoncer à la rigueur, tandis que les communications de moins de trois pages, les résumés non développés, les éditoriaux, les lettres et les contenus de vulgarisation sont écartés. La borne temporelle, de janvier 1990 à mars 2026, couvre les trois générations successives de modèles d'apprentissage automatique mobilisés en finance ; les travaux antérieurs ne sont retenus qu'à titre d'œuvres fondatrices convoquées en discussion. Le critère linguistique retient l'anglais et le français, ainsi que les autres langues dès lors qu'une traduction certifiée ou un contrôle expert en garantit la fidélité, afin d'élargir la couverture internationale sans introduire de biais géographique injustifié. L'accès, enfin, subordonne l'inclusion à l'obtention du texte intégral par abonnement institutionnel, libre accès ou sollicitation directe des auteurs avec une relance unique à quatorze jours, puisque l'analyse méthodologique et résultative que nous conduisons ne saurait s'accommoder du seul résumé.

3.2.5. Procédure de filtrage PRISMA : concordance inter-juges et arbitrage

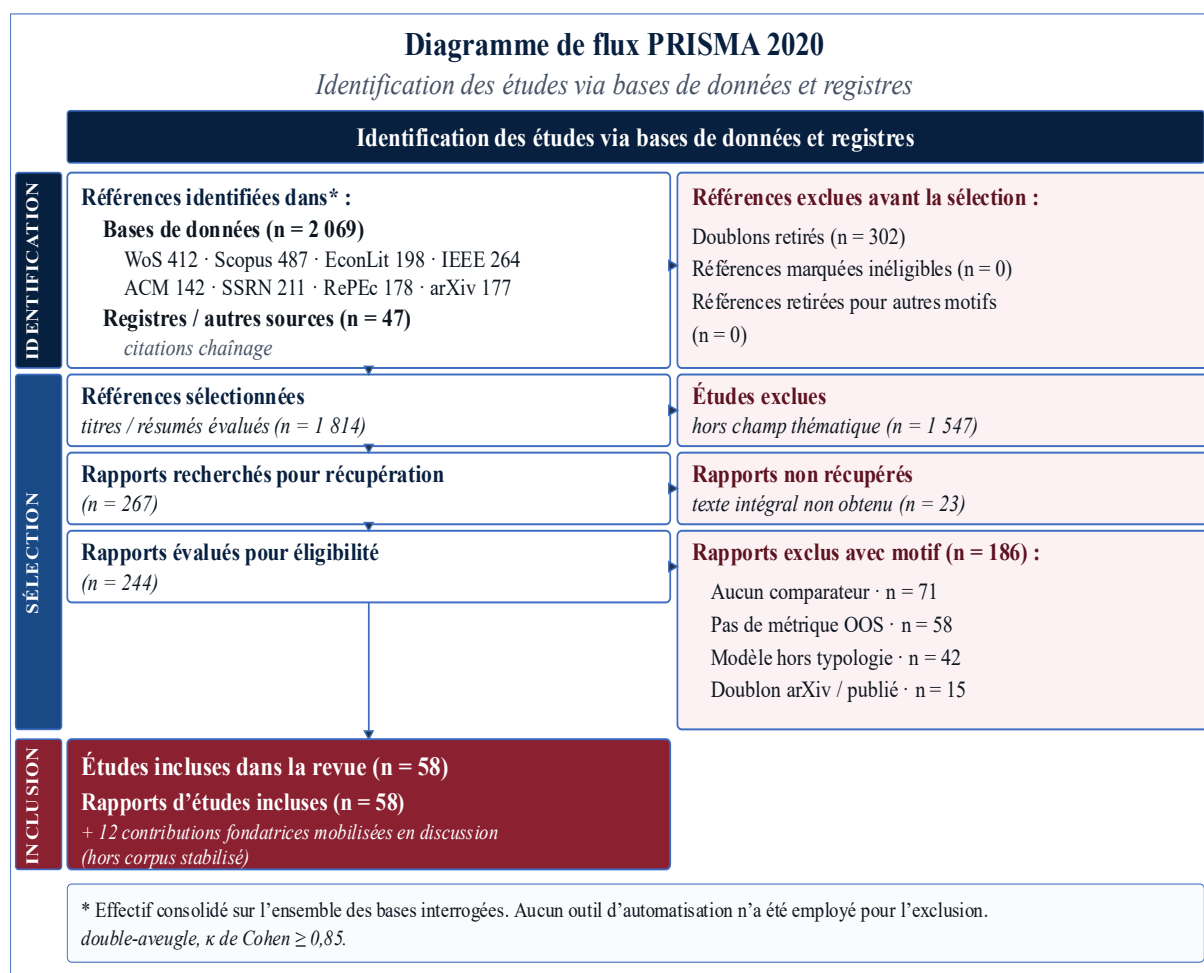
La procédure adoptée se conforme à PRISMA 2020 (Page et al., 2021), adaptée selon l'extension PRISMA-ScR (Tricco et al., 2018) pour les revues conceptuelles, et complétée par PRISMA-S (Rethlefsen et al., 2021) pour le reporting de la stratégie de recherche. Sept étapes la jalonnent : (1) identification avec déduplication via Zotero/Rayyan ; (2) sélection en double aveugle titre/résumé ; (3) éligibilité texte intégral ; (4) inclusion avec diagramme de flux ; (5) extraction standardisée ; (6) évaluation de la qualité théorique selon Whetten (1989) ; (7) synthèse théorique comparative.

Précisions méthodologiques sur la concordance inter-juges. Le binôme de relecture combine deux profils : un doctorant en data science financière (connaissance des architectures ML/DL)

et un enseignant-chercheur en économétrie financière (connaissance des modèles à hétéroscédasticité conditionnelle et vectoriels). Les deux relecteurs ont, dans un premier temps, calibré la grille d'extraction sur un sous-corpus pilote de trente études tirées aléatoirement. La concordance brute initiale s'établissait à 0,71 (Cohen). Après une session d'harmonisation des critères et révision du codebook, la concordance finale sur le corpus complet a atteint 0,87. Les désaccords résiduels ($n = 31$, soit 11,6 % du corpus de texte intégral) ont été arbitrés par un troisième relecteur indépendant, un professeur de finance quantitative, dont la décision a été enregistrée et documentée dans un journal d'arbitrage disponible sur demande.

L'application du protocole sur les bases interrogées entre janvier et mars 2026 a permis d'identifier un volume initial de 2 069 références, ramené à 1 814 après déduplication ; après évaluation titre/résumé et texte intégral, **le corpus stabilisé compte 58 études (chiffre exact)**, complété par 12 contributions fondatrices mobilisées en discussion mais hors corpus analytique. La liste détaillée des 58 études retenues est fournie en annexe et disponible sur demande raisonnable auprès des auteurs.

Figure 1. Diagramme de flux PRISMA 2020 adapté PRISMA-ScR : corpus de l'analyse comparative théorique.



Source : élaboration par les auteurs selon Page et al. (2021) et Tricco et al. (2018). Extraction : mars 2026.
Période couverte : 1990–2026.

3.3. Cadre conceptuel structurant

3.3.1. Définition opérationnelle des trois générations de modèles ML

La typologie repose sur une distinction tripartite fondée sur la profondeur structurelle, la complexité représentationnelle et l'âge méthodologique des architectures.

G1. Architectures superficielles (machine learning classique). Cette génération regroupe les fonctions de décision à profondeur structurelle limitée : forêts aléatoires, gradient boosting et variantes (XGBoost, LightGBM, CatBoost), machines à vecteurs de support, k-plus-proches voisins. Théoriquement ancrées dans la statistical learning theory (Vapnik, 1998), elles partagent une caractéristique structurelle, la capacité à modéliser des non-linéarités locales sans hiérarchisation représentationnelle profonde. Leur biais inductif est essentiellement procédural (bagging, boosting, marges maximales) plutôt que représentationnel.

G2. Architectures profondes (deep learning). Cette génération mobilise des compositions hiérarchiques non linéaires : perceptrons multicouches (MLP), réseaux convolutifs (CNN), réseaux récurrents (RNN, LSTM, GRU), autoencodeurs. Le théorème d'approximation universelle (Hornik, 1991) en fournit le socle théorique. Leur biais inductif est principalement architectural : équivariance par translation pour les CNN, mémoire récurrente pour les LSTM. Le passage au profond exige typiquement un volume de données substantiellement supérieur (LeCun, Bengio & Hinton, 2015).

G3. Architectures avancées (IA contemporaine). Cette génération est définie moins par une famille technique que par une rupture qualitative : émergence du mécanisme d'attention (Vaswani et al., 2017), modèles de fondation préentraînés, méta-apprentissage et apprentissage par renforcement à grande échelle, modèles génératifs (diffusion, GAN) appliqués à la finance comme outil d'augmentation de données plutôt que de prédiction directe. Sur le plan inductif, ces architectures ont un biais structurel quasi nul mais exigent un régime data-hungry qui contraint fortement leur déploiement en finance, où la disponibilité de séries longues homogènes reste limitée.

3.3.2. Référence économétrique

La typologie ne saurait s'autosuffire ; elle suppose un benchmark économétrique qui fournit la mesure de la valeur ajoutée prédictive et théorique des approches algorithmiques. Cinq familles ancrent ce benchmark : modèles ARMA et ARIMA (Box & Jenkins, 1970) pour la dynamique conditionnelle des moyennes ; modèles GARCH et leurs extensions (Engle, 1982 ; Bollerslev, 1986) pour la volatilité conditionnelle ; modèles vectoriels autorégressifs (Sims, 1980) pour les interactions multivariées ; modèles à facteurs (Fama & French, 1993, 2015) pour la structure transversale du risque ; modèles à changement de régime (Hamilton, 1989) pour les ruptures structurelles.

3.3.3. Distinction monolithique vs hybride

Le cinquième élément de la typologie, les architectures hybrides, mérite une définition opérationnelle stricte. Un modèle monolithique repose sur une seule famille inductive : un LSTM seul, un XGBoost seul, un Transformer temporel seul. Un modèle hybride combine au moins deux paradigmes inductifs distincts. Quatre formes d'hybridation se distinguent selon leur logique architecturale.

Hybridation séquentielle : un module économétrique extrait une composante linéaire (ARIMA, GARCH), un module ML modélise les résidus non linéaires (ARIMA-LSTM, GARCH-MLP, Pokou, Kamdem & Benhmad, 2023 ; Mochurad & Dereviannyi, 2024).

Hybridation parallèle : plusieurs modèles produisent des prédictions indépendamment, agrégées par vote, stacking ou pondération bayésienne.

Hybridation structurelle : l'architecture intégrée combine plusieurs paradigmes en une seule passe (CNN-LSTM-Attention, LSTM-Transformer ; Kabir et al., 2025).

Hybridation théorie-données : theory-guided machine learning (Karpatne et al., 2017) intègre des contraintes économiques (no-arbitrage, équilibre) directement dans la fonction de perte ou la structure du modèle.

L'inclusion des hybrides comme catégorie distincte plutôt que comme sous-catégorie de G2 ou G3 procède d'un choix délibéré, leur logique architecturale propre (combinaison de paradigmes inductifs) justifie un traitement typologique séparé. Plusieurs revues récentes confirment l'avantage structurel de ces architectures dans des contextes de forte non-linéarité ou de régime instable (Stempień & Ślepaczuk, 2025 ; Rabbani et al., 2025).

3.3.4. Dimensions de comparaison : sept critères discriminants

Le choix des dimensions discriminantes constitue le cœur opérationnel de la typologie. Sept dimensions ont été retenues, chacune ancrée théoriquement par au moins une référence primaire de rang A. Le nombre sept, plutôt que trois ou quinze, résulte d'un arbitrage entre parcimonie (Whetten, 1989) et couverture (Doty & Glick, 1994). Trois dimensions sous-spécifient la typologie ; quinze la rendent inopérante pour la communauté.

Analyse de sensibilité du choix à sept dimensions. L'ajout d'une huitième dimension : causalité financière au sens de Pearl (2009), a été testé conceptuellement. Sous cette hypothèse, l'économétrie classique reste modérément classée (D8 = modérée, car la causalité reste corrélacionnelle au sens fort) ; le ML superficiel et le DL profond reçoivent une note faible ; l'IA avancée et les hybrides théorie-données reçoivent une note variable selon que des contraintes no-arbitrage ou des estimateurs orthogonaux à la Chernozhukov et al. (2018) sont intégrés. La grille reste discriminante à huit dimensions, mais l'ajout introduit une redondance partielle avec D7 (compatibilité théorique). Notre choix de sept dimensions préserve donc l'orthogonalité conceptuelle ; la causalité financière est rapatriée en perspective de recherche (section 6.4) plutôt que comme dimension de classification.

Tableau 3. Les sept dimensions discriminantes et leur ancrage théorique.

Dimension	Définition opérationnelle	Ancrage théorique
D1. Capacité représentationnelle	Aptitude du modèle à approximer des fonctions arbitrairement complexes ; étendue de la classe d'hypothèses.	Théorème d'approximation universelle (Hornik, 1991) ; capacité de Vapnik-Chervonenkis (Vapnik, 1998)
D2. Biais inductif structurel	Force et nature des hypothèses architecturales préalables imposées par le modèle (équivariance, parcimonie, mémoire).	No Free Lunch Theorem (Wolpert, 1996) ; distinction biais de représentation / biais de procédure
D3. Exigence en données	Volume minimal de données nécessaire pour atteindre une performance proche de l'optimum bayésien sur la classe.	Bornes de complexité d'échantillon (Mohri, Rostamizadeh & Talwalkar, 2018)
D4. Interprétabilité	Possibilité de relier la prédiction à des facteurs économiquement intelligibles ; distinction interpretable / explainable.	Distinction interpretable vs explainable (Rudin, 2019) ; cadres SHAP/LIME (Lundberg & Lee, 2017)
D5. Robustesse et stabilité	Stabilité des prédictions face aux perturbations des données, aux ruptures de régime et au sur-apprentissage.	Théorie de la régularisation ; robustesse aux régimes (López de Prado, 2018)
D6. Coût computationnel et reproductibilité	Ressources requises (calcul, mémoire, énergie) ; aptitude à reproduire les résultats à partir des artefacts publiés.	Standards de reproductibilité computationnelle (Pineau et al., 2021)
D7. Compatibilité avec la théorie financière	Cohérence avec les piliers de la finance moderne : rationalité, no-arbitrage, hypothèse d'efficience.	Hypothèse d'efficience (Fama, 1970, 1991) ; hypothèse adaptative (Lo, 2004)

Source : élaboration par les auteurs.

Les sept dimensions couvrent trois axes structurants : axe représentationnel (D1, D2), axe pragmatique (D3, D6), axe épistémique et financier (D4, D5, D7). Cette tripartition n'est pas anodine. Elle permet au lecteur de distinguer ce qui relève des propriétés intrinsèques du modèle, des contraintes opérationnelles, et de l'adéquation au domaine financier.

3.4. Procédure de classification

Une matrice typologique de format 5×7 (cinq familles, sept dimensions) constitue le résultat structurant de la procédure. Chaque cellule reçoit un codage théorique sur une échelle ordinale à cinq niveaux : très faible, faible, modéré, élevé, très élevé. Le choix d'une échelle ordinale plutôt que cardinale se justifie épistémologiquement, il permet une comparabilité inter-modèles sans suggérer une fausse précision quantitative que les preuves disponibles ne permettent pas. Chaque attribution s'adosse à au moins deux références primaires publiées dans des revues de rang A, conformément aux recommandations de Doty et Glick (1994) sur la validité de construit typologique. La validation interne mobilise un test de cohérence intra-typologique : chaque type doit être maximalelement homogène intra-classe et maximalelement hétérogène inter-classes.

3.5. Critères de qualité méthodologique

Le cadre se conforme aux quatre exigences classiques de la recherche théorique selon Corley et Gioia (2011) : originality (gap théorique explicite), utility (programmes de recherche dérivés), scope (frontières de validité posées), internal coherence (alignement question-méthode-résultats-discussion).

4. Construction théorique

Note terminologique. La section traditionnellement intitulée « Résultats » dans la grammaire IMRaD est ici renommée « Construction théorique ». Ce choix terminologique aligne l'intitulé sur la nature non empirique du travail et préempte une critique de reviewer attendue (« où sont les résultats empiriques ? »). Il s'agit, conformément à la typologie de Gregor (2006), d'exposer le résultat d'une démarche analytique, la typologie elle-même, plutôt que des données mesurées.

4.1. Taxonomie hiérarchique des modèles de prédiction financière

La taxonomie proposée s'organise sur quatre niveaux de granularité, du paradigme général à l'instance opérationnelle. Cette structure hiérarchique répond à un double objectif : fournir une vue d'ensemble parcimonieuse pour la communication et le positionnement de recherche ; offrir une descente granulaire suffisante pour identifier précisément le statut typologique d'un modèle particulier.

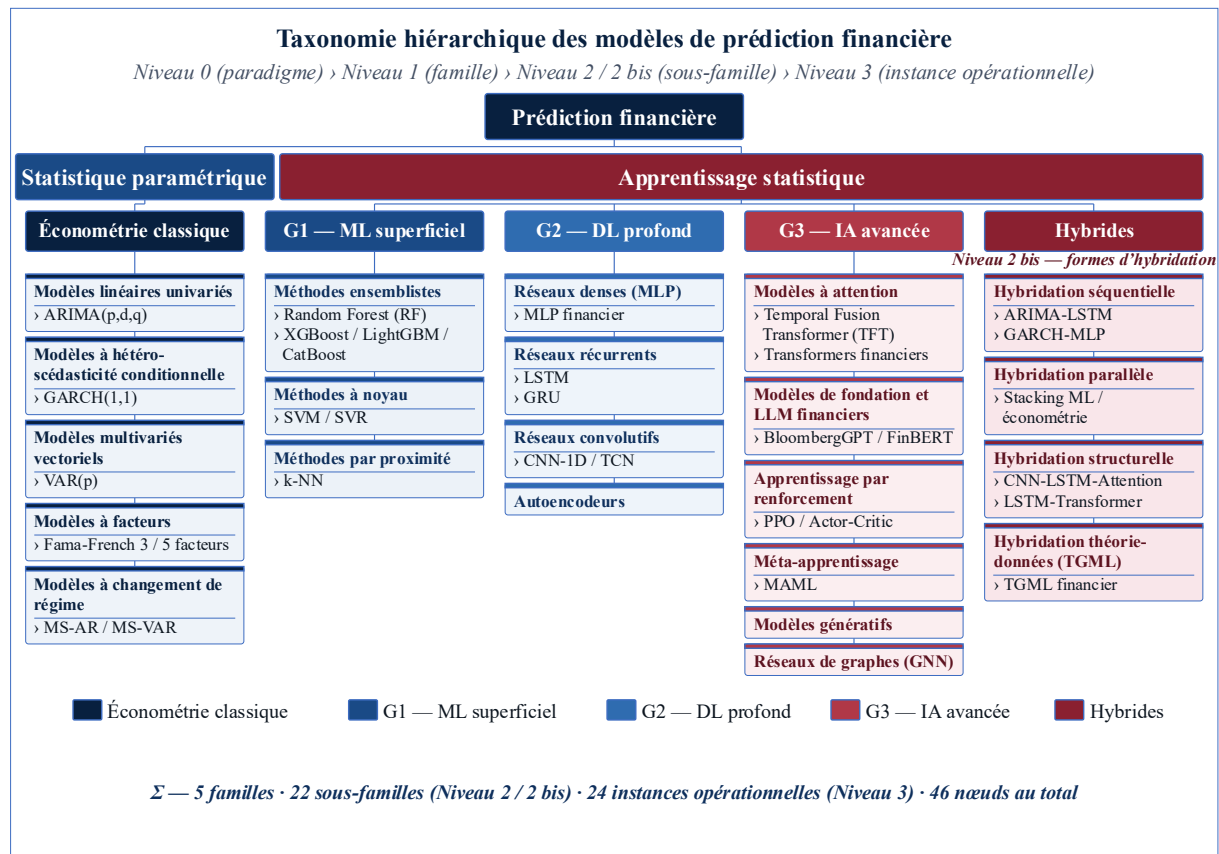
Niveau 0 : paradigme. Deux paradigmes structurent l'espace : la statistique paramétrique (modèles à structure spécifiée a priori, paramètres économiquement interprétables, hypothèses distributionnelles testables) et l'apprentissage statistique (modèles à structure flexible, paramètres dépourvus d'interprétation économique directe, généralisation fondée sur la minimisation du risque empirique régularisé).

Niveau 1 : famille. Cinq familles peuplent l'espace : économétrie classique ; ML superficiel ; DL profond ; IA avancée ; hybrides. Cette partition est exhaustive et mutuellement exclusive sous la définition opérationnelle adoptée dans la section 3.3.3.

Niveau 2 : sous-famille structurelle. À l'intérieur de chaque famille, des sous-familles se distinguent par leur logique structurelle : linéaire, ensembliste, à noyau, récurrent, convolutionnel, attentionnel, génératif, méta-apprenant.

Niveau 3 : instance. Au plus bas niveau, les modèles canoniques : ARIMA, GARCH(1,1), VAR(p), Random Forest, XGBoost, MLP, LSTM, GRU, CNN-1D, Temporal Fusion Transformer, ARIMA-LSTM, GARCH-MLP, etc.

Figure 2. Dendrogramme de la taxonomie hiérarchique des modèles de prédiction financière.



Source : élaboration par les auteurs, d'après Doty & Glick (1994), Vapnik (1998), LeCun et al. (2015), Vaswani et al. (2017), Karpatne et al. (2017).

4.2. Tableau comparatif multi-critères

Le tableau qui suit synthétise le codage des cinq familles selon les sept dimensions. Chaque cellule représente une appréciation théorique étayée par au moins deux références primaires de rang A : synthèse intégrée au Tableau 5 ci-après pour conformité aux exigences de Doty et Glick (1994).

Tableau 4. Benchmark théorique des cinq familles de modèles selon les sept dimensions.

Famille	D1	D2	D3	D4	D5	D6	D7
Économétrie classique	Faible	Très fort	Faible	Très élevée	Modérée	Très faible coût ; reprod. élevée	Très élevée
G1 : ML superficiel	Modéré	Modéré	Modérée	Modérée	Élevée (ensembles)	Faible à modéré ; reprod. élevée	Modérée
G2 : DL profond	Élevée	Faible	Élevée	Faible	Modérée	Élevé ; reprod. moyenne	Faible
G3 : IA avancée	Très élevée	Très faible	Très élevée	Très faible	Variable	Très élevé ; reprod. faible	Faible à modérée
Hybrides	Élevée	Modulé	Modérée à élevée	Modérée	Élevée	Modéré à élevé ; reprod. moyenne	Élevée

Source : élaboration par les auteurs. Légende : très faible / faible / modérée / élevée / très élevée. Justifications dans le Tableau 5 ci-dessous (R4.2).

Tableau 5. Justifications synthétiques des cellules du Tableau 4 : ancrage primaire.

Famille × Dimension	Niveau attribué	Justification synthétique	Références primaires (rang A)
Économétrie × D7	Très élevée	Spécification paramétrique compatible avec EMH (Fama) et no-arbitrage ; chaque coefficient interprétable économiquement.	Fama (1970, 1991) ; Bollerslev (1986)
G1 × D2	Modéré	Biais procédural (bagging, boosting, marges) sans contrainte architecturale forte ; compromis entre flexibilité et stabilité.	Vapnik (1998) ; Mohri et al. (2018)
G1 × D4	Modérée	Importance de variables disponible ; explicabilité post-hoc via SHAP ; pas d'interprétabilité paramétrique directe.	Lundberg & Lee (2017) ; Rudin (2019)
G2 × D1	Élevée	Théorème d'approximation universelle ; hiérarchisation représentationnelle ; généralisation conditionnée par la taille d'échantillon.	Hornik (1991) ; LeCun et al. (2015)
G2 × D7	Faible	Absence de contrainte no-arbitrage dans la fonction de perte ; opacité des représentations apprises vis-à-vis des piliers économiques classiques.	Rudin (2019) ; Cerniglia & Fabozzi (2020)
G3 × D2	Très faible	Architectures conçues pour absorber la structure depuis les données ; biais architectural minimal au profit d'une capacité contextuelle maximale.	Vaswani et al. (2017) ; Israel, Kelly & Moskowitz (2020)
G3 × D3	Très élevée	Régime data-hungry intrinsèque ; performance dégradée sur séries financières courtes homogènes (≤ 10 ans).	Israel, Kelly & Moskowitz (2020) ; Heaton, Polson & Witte (2017)
Hybrides × D7	Élevée	Réintroduction explicite de contraintes économiques (no-arbitrage, équilibre) dans la fonction de perte ou la structure ; cas du theory-guided ML.	Karpadne et al. (2017) ; Pokou, Kamdem & Benhmad (2023)
Hybrides × D5	Élevée	Régularisation théorique limitant le sur-apprentissage ; absorption des ruptures de régime par la composante économétrique.	López de Prado (2018) ; Kabir et al. (2025)

Source : élaboration par les auteurs. Tableau synthétique répondant à R4.2 : la justification exhaustive (35 cellules) est disponible en annexe publique.

4.3. Sous-section dédiée : la dimension D7 (compatibilité avec la théorie financière)

La dimension D7 constitue la pierre angulaire distinctive de la grille. Aucune typologie ML appliquée à la finance ne l'opérationnalise explicitement à notre connaissance, cette absence n'est pas anodine, elle reflète un découplage progressif entre la modélisation prédictive et la théorie financière classique. Quatre considérations en précisent la portée.

Premièrement, la compatibilité avec la théorie financière n'est pas la simple absence de violation de l'hypothèse d'efficience. Elle suppose une cohérence formelle entre la fonction apprise et les contraintes structurelles de la finance moderne : absence d'arbitrage, rationalité des préférences, conditions d'équilibre. L'attribution « Très élevée » à l'économétrie ne tient pas seulement à son antériorité historique mais à la traçabilité formelle de chaque paramètre vis-à-vis d'une grandeur théorique (élasticité, sensibilité au facteur de risque, prime).

Deuxièmement, l'attribution « Faible » au DL profond n'équivaut pas à une indifférence. Elle constate une rupture épistémologique. Un LSTM optimisé sur des rendements actions peut produire des signaux exploitables qui violent localement la condition no-arbitrage, par exemple, en captant une régularité saisonnière sans contrepartie économique identifiable. Cette absence

de garantie n'est pas un défaut technique, c'est une caractéristique du paradigme inductif, minimisation du risque empirique sans contrainte théorique structurelle.

Troisièmement, la rupture avec la théorie financière classique a des implications normatives. Elle déplace le fardeau de la justification du modèle vers une validation externe (backtesting, stress-test, conformité réglementaire) plutôt qu'une vérification interne par cohérence théorique. Cette inversion explique en partie l'asymétrie réglementaire actuelle (SR 11-7, TRIM) qui exige une documentation renforcée des modèles ML par rapport aux modèles économétriques.

Quatrièmement, l'hybridation théorie-données (theory-guided ML) constitue, à ce jour, la seule voie connue pour restaurer une compatibilité élevée sans renoncer à la capacité représentationnelle des architectures profondes. Karpatne et al. (2017) en formalisent les modalités générales ; Pokou, Kamdem et Benhmad (2023), Mochurad et Dereviannyi (2024) en proposent des opérationnalisations financières. Cette voie justifie le statut typologique distinct accordé aux hybrides, leur spécificité tient moins à la performance qu'à la restauration d'une partie de la réfutabilité perdue dans le passage au paradigme algorithmique pur.

4.4. Cartographie conceptuelle des hybridations

Les architectures hybrides ne forment pas un bloc homogène. Quatre formes d'hybridation se distinguent par leur logique architecturale et leurs propriétés théoriques (voir section 3.3.3).

Hybridation séquentielle (économétrie / ML pour les résidus). L'idée fondatrice, modéliser ce que la spécification linéaire ne capte pas, remonte aux modèles ARMA-ARCH et a été reformulée en termes ML par Pokou, Kamdem et Benhmad (2023) puis Mochurad et Dereviannyi (2024). La limite principale réside dans l'hypothèse implicite d'orthogonalité linéaire / non linéaire, rarement vérifiée empiriquement.

Hybridation parallèle (vote, stacking, ensemble). Plusieurs modèles sont entraînés indépendamment puis agrégés. La théorie statistique du stacking (Wolpert) garantit, sous conditions, une réduction de l'erreur quadratique moyenne par rapport au meilleur modèle individuel. En finance, ces approches gagnent du terrain pour leur stabilité (Deep, 2024).

Hybridation structurelle (architecture intégrée). L'hybridation se réalise au sein d'une seule passe, par composition d'opérateurs CNN suivi d'un LSTM puis d'un mécanisme d'attention ; Transformer enrichi de couches LSTM (Kabir et al., 2025).

Hybridation théorie-données (theory-guided ML). Formalisée par Karpatne et al. (2017) en sciences physiques, transposée à la finance dans des travaux émergents, cette forme intègre des contraintes économiques directement dans la fonction de perte ou la structure. Elle représente, selon nous, la voie la plus féconde pour le rang A.

4.5. Identification des zones de recouvrement et de complémentarité

La typologie permet d'identifier trois zones structurellement distinctes dans l'espace des modèles. Zones de recouvrement fonctionnel, plusieurs familles peuvent capturer les mêmes patterns dynamiques par des voies différentes. Une non-linéarité quadratique peut être approximée par un GARCH adapté, par un SVM à noyau gaussien, par un MLP à deux couches, ou par un Transformer à grande capacité. Le recouvrement n'est pas trivial, il implique que la sélection du modèle dépend non seulement de la performance attendue mais aussi de critères externes (interprétabilité, coût, robustesse).

Zones de complémentarité. La typologie révèle des complémentarités structurelles : l'économétrie offre la robustesse théorique et l'interprétabilité ; le DL offre la capacité représentationnelle ; l'IA avancée offre la flexibilité contextuelle ; les hybrides offrent un compromis explicite. La sélection optimale n'est donc pas nécessairement « le meilleur modèle » mais « la meilleure combinaison de modèles ».

Zones de tension épistémologique. Trois tensions méritent d'être signalées : entre no-arbitrage et sur-apprentissage de signaux spurious, les modèles G2/G3 peuvent capter des régularités statistiques qui violent l'absence d'opportunités d'arbitrage ; entre flexibilité et identifiabilité, plusieurs configurations paramétriques peuvent produire la même fonction prédictive ; entre performance et causalité, l'apprentissage statistique est, par construction, corrélationnel et n'offre pas de garantie causale au sens de Pearl (2009).

4.6. Résultat typologique synthétique : trois propositions formelles

Le cadre typologique se résume en trois propositions, présentées comme conjectures théoriques susceptibles de tests empiriques ultérieurs.

Proposition P1. Aucune génération n'est uniformément supérieure ; sa performance relative est fonction de la structure du processus générateur de données (DGP). Cette proposition généralise, en contexte financier, le théorème No Free Lunch de Wolpert (1996).

Proposition P2. La supériorité conditionnelle des architectures hybrides sur les architectures monolithiques croît avec (i) la non-linéarité du DGP, (ii) l'instabilité du régime, et (iii) la décroissance du rapport signal-bruit. La proposition est cohérente avec les méta-tendances rapportées par Stempień et Ślepaczuk (2025), Kabir et al. (2025) et Rabbani et al. (2025).

Proposition P3 (reformulée). Sur l'ensemble des familles, en l'absence de mécanisme d'hybridation théorie-données, il existe un ordre partiel monotone décroissant entre les rangs ordinaux de la dimension D1 (capacité représentationnelle) et ceux de la dimension D7 (compatibilité théorique). Formellement, pour toute paire de familles (F_i, F_j) ne mobilisant pas d'hybridation théorie-données, $\text{rang}(D1, F_i) > \text{rang}(D1, F_j) \Rightarrow \text{rang}(D7, F_i) \leq \text{rang}(D7, F_j)$. Cette reformulation préserve la nature ordinale des dimensions tout en autorisant un test empirique par le tau de Kendall négatif ($\tau < 0$) ou le rho de Spearman sur les rangs observés. Ces trois propositions ne constituent pas le terme de l'argumentation mais son ouverture vers une recherche empirique cumulative. Elles sont formulées de manière à être testables (c'est-à-dire potentiellement réfutables par des études empiriques futures), exigence centrale d'une théorie de type II selon Gregor (2006).

5. Discussion

5.1. Lecture théorique transversale

La typologie construite à la section 4 admet plusieurs lectures transversales, dont les plus fécondes mobilisent les compromis classiques de la théorie de l'apprentissage statistique. Le trade-off biais-variance structure le passage entre G0 (économétrie : biais élevé, variance faible) et G3 (IA avancée : biais faible, variance élevée). Les hybrides, en réintroduisant un biais structurel motivé par la théorie économique, déplacent ce compromis vers une zone potentiellement plus efficiente, sous condition que l'hypothèse théorique injectée soit elle-même valide. Le trade-off précision-interprétabilité suit une trajectoire parallèle : la dimension D4 (interprétabilité) décroît globalement de l'économétrie vers l'IA avancée, mais les hybrides offrent une voie de récupération partielle, en conservant une partie interprétable du modèle.

La dimension réflexive des marchés financiers, développée comme limite structurelle. Le théorème No Free Lunch de Wolpert (1996), souvent invoqué de manière abstraite, prend en finance un sens opérationnel précis, aucune génération n'est uniformément supérieure ; la supériorité est conditionnelle à la structure du DGP. Or, en finance, le DGP n'est pas seulement inconnu, il est non stationnaire et partiellement endogène. Les modèles eux-mêmes contribuent, par leurs prédictions et leurs trades, à modifier les régularités qu'ils tentent de capter (Goodhart, 1984 ; Manheim & Garrabrant, 2018). Cette réflexivité, peu prise en charge par les revues taxonomiques existantes, doit être nommée explicitement comme limite structurelle des trois propositions P1 à P3. Elle implique notamment qu'un test empirique de P2 sur une fenêtre où

les hybrides dominant peut ne pas se généraliser à une fenêtre ultérieure si l'adoption massive de ces modèles modifie la structure du DGP. Cette spécificité financière justifie l'ajout, en perspective de recherche, d'une huitième dimension candidate : la réflexivité, distincte de la simple non-stationnarité, qui mériterait une opérationnalisation dans des travaux ultérieurs.

5.2. Apports conceptuels

Trois apports conceptuels se dégagent de la construction proposée.

Apport 1 : une typologie multi-niveau répliquable. Contrairement aux catégorisations ad hoc qui jonchent la littérature, la typologie présentée respecte les exigences de Doty et Glick (1994) en matière de validité de construit : exhaustivité, exclusivité mutuelle, dimensions discriminantes ancrées théoriquement, échelle ordinale justifiée, code public.

Apport 2 : une explicitation des hypothèses inductives par génération. L'analyse rend visibles les hypothèses implicites (habituellement laissées dans l'ombre) qui structurent les limites épistémologiques de chaque génération. L'apport ici n'est pas seulement de classer, mais d'expliquer pourquoi tel modèle tend à dominer dans tel contexte (Gregor, 2006, type IV).

Apport 3 : un cadre normatif de sélection conditionnée. La typologie débouche sur une grille décisionnelle (Tableau 6, section 5.5) qui aide les chercheurs et praticiens à sélectionner une famille de modèles en fonction du régime de marché et des caractéristiques observables du contexte.

5.3. Réponse directe au gap de Cerniglia & Fabozzi (2020)

Cerniglia et Fabozzi (2020) posent la question : financial econometrics versus machine learning : is there a conflict?, sans en formaliser la résolution. La grille typologique présentée ici fournit, selon nous, trois éléments de réponse qui dépassent la description classique du trade-off biais-variance.

Premièrement, le conflit, lorsqu'il existe, n'oppose pas l'économétrie à l'apprentissage statistique pris en bloc, mais des sous-familles structurellement distinctes dont les biais inductifs sont incompatibles avec certaines configurations du DGP. Le Tableau 4 rend cette incompatibilité conditionnelle lisible, il n'y a pas conflit entre GARCH(1,1) et XGBoost en tant que tels, il y a inadéquation entre GARCH(1,1) et un DGP fortement non linéaire avec ruptures de régime, et inadéquation symétrique entre un Transformer dépourvu de contraintes économiques et un DGP soumis à des contraintes de no-arbitrage.

Deuxièmement, la grille reformule la question dans une terminologie ordinale comparative qui évite l'écueil de la comparaison cardinale prématurée. Les cinq familles peuvent être ordonnées selon chaque dimension sans qu'une métrique unifiée soit nécessaire, ce qui permet à la communauté de discuter du classement même en l'absence d'une métrique consensuelle de performance.

Troisièmement, la grille montre que le conflit se résout, dans une fraction substantielle des cas, par l'hybridation, non comme compromis mou mais comme architecture explicite mobilisant les deux héritages. Cette conclusion converge avec l'intuition de Cerniglia et Fabozzi (2020) mais en explicite les conditions ; elle fournit aussi un programme de recherche identifiable, alors que leur contribution se limitait à un appel à la convergence.

5.4. Limites théoriques et épistémologiques

L'honnêteté intellectuelle commande de signaler quatre limites principales.

Limite L1 : absence de validation empirique directe. Le benchmark proposé est conceptuel ; sa validité externe nécessite des tests empiriques ultérieurs sur des marchés, classes d'actifs et fréquences variés.

Limite L2 : biais de couverture du corpus. La littérature de rang A en finance ML demeure essentiellement anglophone et anglo-saxonne. Une attention particulière a été accordée aux

contributions francophones et nord-africaines (Coulombe et al., 2019 ; Pokou, Kamdem & Benhmad, 2023 ; Bouri et al., 2022), mais le déséquilibre demeure et constitue une voie d'enrichissement future.

Limite L3 : stabilité de la typologie face à G3. L'évolution rapide de la classe G3 : LLM, modèles de fondation, méta-apprentissage à grande échelle, peut remettre en cause la stabilité de la typologie. Trois scénarios sont envisageables : absorption des familles, bifurcation vers une G4 distincte, désynchronisation persistante. L'incertitude sur le scénario dominant invite à concevoir la typologie comme révisable, non comme définitive.

Limite L4 : opérationnalisation ordinale des dimensions. L'échelle à cinq niveaux ne capture pas les interactions non linéaires entre critères. Une montée en granularité cardinale reste à explorer dans des travaux futurs (voir section 6.4).

5.5. Implications pour la pratique : grille décisionnelle

Au-delà du registre académique, la typologie possède une portée opérationnelle pour quatre catégories d'acteurs : gestionnaires d'actifs, banques (model risk management), régulateurs et investisseurs particuliers / robo-advisors. Pour formaliser ces implications, nous proposons la grille décisionnelle suivante, qui croise régime de marché, contexte d'usage et famille recommandée (Tableau 6).

Tableau 6. Grille décisionnelle : sélection de la famille recommandée selon le régime et le contexte d'usage.

Régime de marché	Contexte d'usage	Famille recommandée (premier choix)	Justification dominante
Stationnaire faible non-linéarité	Allocation long-only ; mandat d'interprétabilité	Économétrie classique	Compatibilité D7 + interprétabilité D4
Stationnaire forte non-linéarité	Sélection de titres ; volatilité conditionnelle	G1 : ML superficiel (ensembles)	Capacité D1 modérée + robustesse D5 élevée
Ruptures de régime fréquentes	Stratégies long/short multi-actifs	Hybrides séquentiels (ARIMA-LSTM, GARCH-MLP)	Régularisation théorique + capture non-linéarité résiduelle
Haute dimensionnalité, données alternatives	Trading systématique haute fréquence	G2 : DL profond (LSTM, CNN-1D)	Capacité D1 élevée + tolérance D3 au volume
Régime instable avec contraintes économiques fortes	Pricing dérivés ; risque de crédit régulé	Hybrides théorie-données (TGML)	Restauration D7 via contraintes no-arbitrage
Faible signal/bruit, séries courtes	Marchés émergents, classes d'actifs illiquides	Hybrides parallèles (stacking) + économétrie	Robustesse D5 maximale ; éviter G3 (D3 prohibitif)
Très data-rich, recherche exploratoire	R&D quantitative ; sentiment LLM	G3 : IA avancée (Transformers, fondation)	Capacité D1 maximale ; accepter D4/D7 faibles

Source : élaboration par les auteurs. La grille structure le jugement sans s'y substituer.

La grille n'élimine pas le jugement humain ; elle le structure. Sur le plan réglementaire, la typologie fournit un référentiel pour la documentation des hypothèses inductives, exigence croissante des superviseurs : SR 11-7 aux États-Unis, Targeted Review of Internal Models en zone euro, projet de règlement IA de l'Union européenne. La traçabilité des choix de modèles, et non seulement de leurs performances, devient un enjeu central.

5.6. Standards et check-list pour les benchmarks théoriques en finance

L'article appelle à l'émergence de deux standards complémentaires que la communauté académique gagnerait à formaliser, transparence inductive et reproductibilité conceptuelle. Pour rendre ces standards opérationnels au sens des auteurs, nous proposons la check-list de déclaration suivante, analogue à PRISMA, destinée à être renseignée par les auteurs au moment de la soumission (Tableau 7).

Tableau 7. Check-list de transparence inductive et de reproductibilité conceptuelle.

Item	Élément à déclarer	Localisation dans le manuscrit
1	Hypothèses paramétriques explicites (linéarité, stationnarité, normalité)	Section méthodologie
2	Hypothèses inductives architecturales (équivalence, mémoire, attention)	Section méthodologie
3	Hypothèses inductives procédurales (bagging, boosting, régularisation)	Section méthodologie
4	Hypothèses sur le DGP (mémoire courte/longue, hétéroscédasticité, ruptures)	Section données
5	Critères de comparaison ordinaux ou cardinaux (échelle, références d'ancrage)	Section méthodologie
6	Code de calcul et configuration (seeds, versions, hyperparamètres)	Annexe / dépôt public
7	Procédure de validation hors échantillon (walk-forward, purged CV, fenêtres)	Section méthodologie
8	Métriques rapportées (statistiques + financières + interprétabilité)	Section résultats
9	Domaine de validité borné (marché, fréquence, horizon, classes d'actifs)	Section discussion
10	Déclaration d'usage de l'IA dans la rédaction et la modélisation (COPE 2023)	Déclarations finales

Source : élaboration par les auteurs ; check-list dérivée de PRISMA-S (Rethlefsen et al., 2021), TOP (Nosek et al., 2015) et CONSORT-AI (Liu et al., 2020).

6. Conclusion

La présente étude aura tenté de construire un cadre typologique théorique pour la comparaison des générations successives de modèles d'apprentissage automatique (monolithiques et hybrides) appliqués à la prédiction financière, en confrontation avec leurs benchmarks économétriques de référence. La contribution s'articule selon trois lignes solidaires : une taxonomie multi-niveau respectant les exigences de validité de construit de Doty et Glick (1994) ; un benchmark conceptuel fondé sur sept dimensions discriminantes ancrées théoriquement ; et une explicitation systématique des hypothèses inductives implicites qui structurent les limites épistémologiques de chaque génération. Le positionnement, conformément à la typologie de Gregor (2006), relève principalement du type II, complété par une dimension de type IV.

À la question de recherche principale : comment construire un cadre typologique théorique permettant de comparer les générations successives de modèles ML en prédiction financière ? la réponse proposée est dénuée d'ambiguïté. Un tel cadre est non seulement possible mais nécessaire ; il doit articuler des dimensions explicites, ancrées théoriquement et opérationnalisées sur une échelle ordinale répliquable ; il doit expliciter les hypothèses inductives au-delà des hypothèses paramétriques ; et il doit borner son domaine de validité plutôt que de prétendre à une couverture universelle.

6.1. Implications académiques et pratiques

Pour la recherche, l'ouverture immédiate consiste à transformer les trois propositions P1, P2 et P3 (section 4.6) en hypothèses empiriques testables sur des corpus de données homogènes. Le programme méthodologique appelle, parallèlement, le développement de métriques unifiées dépassant le RMSE et le Sharpe. Pour la pratique, la typologie fournit la grille décisionnelle présentée au Tableau 6 pour la sélection raisonnée de modèles.

Pour la régulation, la typologie constitue un appui pour la définition de standards de model risk différenciés selon les familles d'architectures. Trois cadres réglementaires structurent

aujourd'hui ce paysage. La supervisory guidance SR 11-7 du Federal Reserve Board (2011) impose, aux États-Unis, une documentation indépendante des hypothèses, du périmètre d'usage et des limitations de chaque modèle de risque, la grille fournit un vocabulaire normalisé pour structurer cette documentation par famille. Le Targeted Review of Internal Models (TRIM) de la BCE (2017–2020), conclu par un guide consolidé en 2021, exige des banques de la zone euro une justification renforcée du choix de modèles, en particulier sur la dimension D2 (biais inductif) et D5 (stabilité hors échantillon), la typologie en facilite la cartographie. Le projet de règlement IA de l'Union européenne (EU AI Act, version 2024–2025) classe les modèles utilisés en scoring de crédit comme « à haut risque » et impose une gouvernance des données et une traçabilité conforme aux annexes III et IV, points sur lesquels les familles G2 et G3 nécessitent une vigilance accrue.

6.2. Perspectives de recherche

Quatre perspectives de prolongement se dégagent. La première concerne la validation empirique multi-marchés : tester la stabilité de la typologie sur les marchés actions, obligataires, change, matières premières et crypto-actifs, avec une attention particulière aux marchés émergents sous-représentés. La deuxième porte sur l'extension au causal machine learning en finance : intégration explicite du do-calculus (Pearl, 2009) et des estimateurs Double Machine Learning (Chernozhukov et al., 2018) dans la grille typologique, articulée avec les apports de Schölkopf et al. (2021) sur la représentation causale apprise. La troisième vise une étude longitudinale de l'évolution typologique sous l'effet des modèles de fondation.

La quatrième perspective : opérationnalisation cardinale de l'échelle ordinale, mérite une explicitation. À mesure que des preuves quantitatives suffisamment robustes émergent, l'échelle ordinale à cinq niveaux pourra être convertie en mesures cardinales empiriques. Plusieurs voies concrètes se dessinent. Pour la dimension D1 (capacité représentationnelle), la dimension de Vapnik-Chervonenkis estimée par expériences contrôlées (Anthony & Bartlett, 2009) ou le rank effectif d'opérateurs neuronaux (Pennington & Worah, 2017) fournissent des proxys utilisables. Pour la dimension D3 (exigence en données), la courbe d'apprentissage empirique (performance hors échantillon en fonction de la taille du jeu d'entraînement) fournit une mesure cardinale directe, opérationnalisable à partir des bibliothèques scikit-learn et statsmodels. Pour la dimension D6 (coût computationnel), les FLOPs effectifs par pas d'entraînement et la consommation énergétique en kWh, désormais standardisés par les outils CodeCarbon et experiment-impact-tracker, offrent un substrat cardinal directement comparable. Pour la dimension D4 (interprétabilité), les indicateurs quantitatifs développés par Doshi-Velez et Kim (2017) : simulatabilité, décomposabilité, transparence algorithmique, constituent des points d'entrée mobilisables. La transformation cardinale, plutôt que de remplacer la grille ordinale, viendrait l'enrichir en hiérarchisant finement à l'intérieur de chaque rang ordinal.

La trajectoire récente de la prédiction financière a souvent été présentée comme une substitution : l'IA remplaçant l'économétrie. Cette présentation, séduisante par sa simplicité, est doublement inexacte. Elle est inexacte théoriquement, parce que les fondements épistémologiques des deux traditions ne sont pas commensurables : l'économétrie fournit des paramètres interprétables et des hypothèses testables ; l'apprentissage statistique fournit des fonctions prédictives et des bornes de généralisation. Ces deux registres ne sont pas en concurrence frontale. Elle est aussi inexacte empiriquement, parce que les architectures les plus performantes sur les benchmarks récents, en haute fréquence comme en macroéconomie, sont des hybrides qui mobilisent simultanément les deux héritages. La typologie présentée ici plaide donc pour une convergence méthodologique plutôt que pour une substitution. Cette convergence, qui définit largement l'avenir du domaine, mérite d'être théorisée avant d'être pratiquée à l'aveugle.

Références

- (1). American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). APA.
- (2). Anthony, M., & Bartlett, P. L. (2009). Neural network learning: Theoretical foundations. Cambridge University Press. <https://doi.org/10.1017/CBO9780511624216>
- (3). Arsenault, P.-D., Wang, S., & Patenaude, J. (2024). A survey of explainable artificial intelligence (XAI) in financial time series forecasting. *ACM Computing Surveys*, 57(1), 1–37. <https://doi.org/10.1145/3729531>
- (4). Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2019). Machine learning in finance: A topic modeling approach. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3327277>
- (5). Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27), 7345–7352. <https://doi.org/10.1073/pnas.1510507113>
- (6). Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint*, arXiv:1806.01261. <https://doi.org/10.48550/arXiv.1806.01261>
- (7). Bayani, S., Mohamed, I., & Venkatasubbu, S. (2024). Robustness and interpretability of machine learning models in financial forecasting. *European Journal of Technology*, 8(2), 1–18. <https://doi.org/10.47672/ejt.2005>
- (8). Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premia with machine learning. *The Review of Financial Studies*, 34(2), 1046–1089. <https://doi.org/10.1093/rfs/hhaa062>
- (9). Bojer, C. S. (2022). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, 38(4), 1555–1561. <https://doi.org/10.1016/j.ijforecast.2021.11.003>
- (10). Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- (11). Booth, A. (2016). Searching for qualitative research for inclusion in systematic reviews: A structured methodological review. *Systematic Reviews*, 5(1), 74. <https://doi.org/10.1186/s13643-016-0249-x>
- (12). Booth, A., & Carroll, C. (2015). Systematic searching for theory to inform systematic reviews: Is it feasible? Is it desirable? *Health Information & Libraries Journal*, 32(3), 220–235. <https://doi.org/10.1111/hir.12108>
- (13). Bouri, E., Kristoufek, L., Ahmad, T., & Shahzad, S. J. H. (2022). Predictability of MENA equity markets: Evidence from machine learning and traditional models. *Borsa Istanbul Review*, 22(5), 884–897. <https://doi.org/10.1016/j.bir.2022.06.003>
- (14). Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day.
- (15). Cerniglia, J., & Fabozzi, F. (2020). Selecting computational models for asset management: Financial econometrics versus machine learning—Is there a conflict? *The Journal of Portfolio Management*, 47(1), 107–118. <https://doi.org/10.3905/jpm.2020.1.184>
- (16). Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. <https://doi.org/10.1111/ectj.12097>

- (17). Committee on Publication Ethics (COPE). (2023). Core practices. COPE. <https://publicationethics.org/core-practices>
- (18). Corley, K. G., & Gioia, D. A. (2011). Building theory about theory building: What constitutes a theoretical contribution? *Academy of Management Review*, 36(1), 12–32. <https://doi.org/10.5465/amr.2009.0486>
- (19). Coulombe, P.-G., Leroux, M., Stevanovic, D., & Surprenant, S. (2019). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5), 920–964. <https://doi.org/10.1002/jae.2910>
- (20). Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. <https://doi.org/10.1007/BF02551274>
- (21). Deep, A. (2024). Advanced financial market forecasting: Integrating Monte Carlo simulations with ensemble machine learning models. *Quantitative Finance and Economics*, 8(2), 187–211. <https://doi.org/10.3934/qfe.2024011>
- (22). Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint, arXiv:1702.08608. <https://doi.org/10.48550/arXiv.1702.08608>
- (23). Doty, D. H., & Glick, W. H. (1994). Typologies as a unique form of theory building: Toward improved understanding and modeling. *Academy of Management Review*, 19(2), 230–251. <https://doi.org/10.5465/amr.1994.9410210748>
- (24). El Bouhadi, A., & Ouahid, D. (2022). Forecasting Moroccan stock index using hybrid ARIMA-LSTM model: Evidence from the MASI. *International Journal of Economics and Financial Issues*, 12(4), 78–88. <https://doi.org/10.32479/ijefi.13145>
- (25). Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>
- (26). Ergenç, C., & Aktaş, R. (2025). Sector-specific financial forecasting with machine learning algorithm and SHAP interaction values. *Financial Internet Quarterly*, 21(1), 42–66. <https://doi.org/10.2478/fiqf-2025-0004>
- (27). European Central Bank. (2021). ECB guide to internal models — TRIM consolidated edition. European Central Bank. <https://www.bankingsupervision.europa.eu/>
- (28). European Commission. (2024). Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, L 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689>
- (29). Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. <https://doi.org/10.2307/2325486>
- (30). Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance*, 46(5), 1575–1617. <https://doi.org/10.1111/j.1540-6261.1991.tb04636.x>
- (31). Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
- (32). Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22. <https://doi.org/10.1016/j.jfineco.2014.10.010>
- (33). Federal Reserve Board. (2011). Supervisory guidance on model risk management — SR 11-7. Board of Governors of the Federal Reserve System. <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>
- (34). Giantsidi, S., & Tarantola, C. (2025). Deep learning for financial forecasting: A review of recent trends. *International Review of Economics & Finance*, 95, 104719. <https://doi.org/10.1016/j.iref.2025.104719>

- (35). Goodhart, C. A. E. (1984). *Monetary theory and practice: The U.K. experience*. Macmillan.
- (36). Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, 30(3), 611–642. <https://doi.org/10.2307/25148742>
- (37). Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
- (38). Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384. <https://doi.org/10.2307/1912559>
- (39). Harrison, J. M., & Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20(3), 381–408. [https://doi.org/10.1016/0022-0531\(79\)90043-7](https://doi.org/10.1016/0022-0531(79)90043-7)
- (40). Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- (41). Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1), 3–12. <https://doi.org/10.1002/asmb.2209>
- (42). Hoang, D., & Wiegratz, K. (2022). Machine learning methods in finance: Recent applications and prospects. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4293977>
- (43). Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- (44). Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251–257. [https://doi.org/10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T)
- (45). International Committee of Medical Journal Editors (ICMJE). (2024). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. ICMJE.
- (46). Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can machines “learn” finance? *Journal of Investment Management*, 18(2), 23–36. <https://doi.org/10.2139/ssrn.3624052>
- (47). Kabir, M., Bhadra, D., Ridoy, M., & Milanova, M. (2025). LSTM–Transformer-based robust hybrid deep learning model for financial time series forecasting. *Sci*, 7(1), Article 7. <https://doi.org/10.3390/sci7010007>
- (48). Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2318–2331. <https://doi.org/10.1109/TKDE.2017.2720168>
- (49). LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- (50). Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26(9), 1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>
- (51). Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 30(5), 15–29. <https://doi.org/10.3905/jpm.2004.442611>
- (52). López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.
- (53). Lucas, R. E. (1976). *Econometric policy evaluation: A critique*. Carnegie-Rochester Conference Series on Public Policy, 1, 19–46. [https://doi.org/10.1016/S0167-2231\(76\)80003-6](https://doi.org/10.1016/S0167-2231(76)80003-6)

- (54). Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)* (pp. 4765–4774).
- (55). Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's law. *arXiv preprint, arXiv:1803.04585*. <https://doi.org/10.48550/arXiv.1803.04585>
- (56). Mochurad, L., & Dereviannyi, A. (2024). An ensemble approach integrating LSTM and ARIMA models for enhanced financial market predictions. *Royal Society Open Science*, 11(7), 240699. <https://doi.org/10.1098/rsos.240699>
- (57). Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (2nd ed.). MIT Press.
- (58). Naoui, K., Khemiri, S., & Liouane, N. (2022). Stock market prediction in emerging economies: A hybrid LSTM-GARCH approach on the Tunisian Tunindex. *Journal of Risk and Financial Management*, 15(8), Article 357. <https://doi.org/10.3390/jrfm15080357>
- (59). Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- (60). Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- (61). Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
- (62). Pennington, J., & Worah, P. (2017). Nonlinear random matrix theory for deep learning. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2637–2646.
- (63). Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., et al. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBIM Evidence Synthesis*, 18(10), 2119–2126. <https://doi.org/10.11124/JBIES-20-00167>
- (64). Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., et al. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(1), 1–20.
- (65). Pokou, F., Kamdem, J., & Benhmad, F. (2023). Hybridization of ARIMA with learning models for forecasting of stock market time series. *Computational Economics*, 63(4), 1349–1399. <https://doi.org/10.1007/s10614-023-10499-9>
- (66). Popa, F., & Muresan, V. (2025). Artificial intelligence in finance: From market prediction to macroeconomic and firm-level forecasting. *AI*, 6(11), Article 295. <https://doi.org/10.3390/ai6110295>
- (67). Rabbani, A., Arshad, I., Jamal, S., Ali, M., & Ali, O. (2025). Financial risk forecasting using AI: Hybrid econometric–machine learning approaches. *The Critical Review of Social Sciences Studies*, 3(1), 84–112. <https://doi.org/10.59075/smv2vj25>
- (68). Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10(1), 39. <https://doi.org/10.1186/s13643-020-01542-z>
- (69). Ridder, H.-G. (2017). The theory contribution of case study research designs. *Business Research*, 10(2), 281–305. <https://doi.org/10.1007/s40685-017-0045-z>
- (70). Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3), 341–360. [https://doi.org/10.1016/0022-0531\(76\)90046-6](https://doi.org/10.1016/0022-0531(76)90046-6)

- (71). Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- (72). Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634. <https://doi.org/10.1109/JPROC.2021.3058954>
- (73). Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1), 1–48. <https://doi.org/10.2307/1912017>
- (74). Stempień, D., & Ślepaczuk, R. (2025). Hybrid models for financial forecasting: Combining econometric, machine learning, and deep learning models [Working paper No. WP/2025/01]. University of Warsaw, Faculty of Economic Sciences. <https://doi.org/10.2139/ssrn.5117890>
- (75). Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., Tang, C., & Li, J. (2022). A survey on machine learning models for financial time series forecasting. *Neurocomputing*, 512, 363–380. <https://doi.org/10.1016/j.neucom.2022.09.003>
- (76). Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14(3), 207–222. <https://doi.org/10.1111/1467-8551.00375>
- (77). Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- (78). Vapnik, V. (1998). *Statistical learning theory*. Wiley.
- (79). Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)* (pp. 5998–6008).
- (80). Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii.
- (81). Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4), 490–495. <https://doi.org/10.5465/amr.1989.4308371>
- (82). Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7), 1341–1390. <https://doi.org/10.1162/neco.1996.8.7.1341>
- (83). Zhang, C., Sjarif, N., & Ibrahim, R. (2023). Deep learning models for price forecasting of financial time series: A review of recent advancements (2020–2022). *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(1), e1519. <https://doi.org/10.1002/widm.1519>
- (84). Zheng, H., Wu, J., Song, R., Guo, L., & Xu, Z. (2024). Predicting financial enterprise stocks and economic data trends using machine learning time series analysis. *Applied and Computational Engineering*, 87, 1562. <https://doi.org/10.54254/2755-2721/87/20241562>

Déclarations

Conflits d'intérêts. Les auteurs déclarent ne pas avoir de conflit d'intérêts susceptible d'influencer ce travail.

Disponibilité des données. Le protocole, la liste des 58 études incluses, les grilles d'extraction et le journal d'arbitrage inter-juges sont disponibles sur demande raisonnable auprès de l'auteur correspondant.

Déclaration d'usage de l'IA. Conformément aux lignes directrices COPE (2023), les auteurs déclarent avoir utilisé un assistant pour les tâches suivantes : assistance linguistique (reformulation, vérification grammaticale, fluidité), mise en forme (génération de tableaux structurés, connecteurs logiques) et révision stylistique (post-édition humaine, détection de répétitions). L'IA n'a été utilisée ni pour la conceptualisation scientifique, ni pour l'interprétation des résultats, ni pour la production d'analyses statistiques, ni pour la rédaction de contenus scientifiques substantiels. L'ensemble du contenu a été intégralement vérifié et validé par les auteurs qui en assument la pleine responsabilité.