Une image contenant texte, capture d’écran, Police, graphisme

Le contenu généré par l’IA peut être incorrect.

Table des matières

[I. Introduction 3](#_Toc209998567)

[1) Contexte : la volatilité comme variable clé en finance 3](#_Toc209998568)

[2) Problématique : limites des modèles classiques 3](#_Toc209998569)

[3) Objectif du projet 4](#_Toc209998570)

[II. Données et prétraitement 4](#_Toc209998571)

[1) Source des données 4](#_Toc209998572)

[2) Présentation des séries utilisées 5](#_Toc209998573)

[3) Variables dérivées et prétraitement 5](#_Toc209998574)

[III. Méthodes 6](#_Toc209998575)

[1) Modèles GARCH et extensions 6](#_Toc209998576)

[2) Modèle HAR-RV 6](#_Toc209998577)

[3) Modèle LSTM 6](#_Toc209998578)

[4) Schéma expérimental et méthodologie de prévision 7](#_Toc209998579)

[5) Tests statistiques 7](#_Toc209998580)

[IV. Résultats empiriques 8](#_Toc209998581)

[1) Performances prévisionnelles 8](#_Toc209998582)

[2) Visualisation des résultats 8](#_Toc209998583)

[3) Analyse statistique des différences (tests DM et t-tests) 9](#_Toc209998584)

[V. Backtest applicatif 9](#_Toc209998585)

[1) Stratégie de straddles ATM avec couverture delta 9](#_Toc209998586)

[2) Résultats du backtest et interprétation 10](#_Toc209998587)

[VI. Discussion 11](#_Toc209998588)

[1) Horizon court : EWM et HAR-RV comme références solides 11](#_Toc209998589)

[2) Horizons lissés : apport clair des modèles LSTM 11](#_Toc209998590)

[3) Cohérence entre métriques et backtest 12](#_Toc209998591)

[4) Limites identifiées 12](#_Toc209998592)

[5) Perspectives d’amélioration 12](#_Toc209998593)

[VII. Conclusion du projet 13](#_Toc209998594)

## I. Introduction

### Contexte : la volatilité comme variable clé en finance

La volatilité occupe une place centrale en finance de marché. Elle mesure l’incertitude ou le risque associé à un actif, et influence directement le pricing des options, la gestion des portefeuilles, la Value-at-Risk et les stratégies de couverture. Pour les investisseurs comme pour les régulateurs, disposer d’estimations fiables de la volatilité est essentiel afin d’anticiper les fluctuations des marchés et de prendre des décisions rationnelles. Cependant, la volatilité n’est pas directement observable : elle doit être estimée ou prédite, ce qui justifie l’existence d’une vaste littérature en économétrie et en apprentissage automatique.

### Problématique : limites des modèles classiques

Historiquement, plusieurs approches ont été proposées pour modéliser la volatilité. Les modèles Exponentially Weighted Moving Average (EWM) offrent une méthode simple et robuste, mais leur paramétrisation avec un facteur de décroissance fixé peut limiter la capture de dynamiques complexes. Les modèles ARCH et GARCH introduisent une mémoire conditionnelle permettant de modéliser les clusters de volatilité, mais ils reposent sur des hypothèses paramétriques rigides, comme la normalité des résidus ou une persistance quasi unitaire, qui réduisent leur capacité à représenter des comportements non linéaires ou des chocs extrêmes. Des extensions comme l’EGARCH, le GJR-GARCH ou le HAR-RV ont été développées pour corriger certaines limites, mais elles restent basées sur des spécifications fonctionnelles imposées par le modélisateur. Ces contraintes ouvrent la voie à des méthodes issues de l’apprentissage profond. Les réseaux récurrents de type LSTM ont la capacité d’extraire automatiquement des dépendances complexes dans des séries temporelles, ce qui en fait des candidats naturels pour la prévision de volatilité. La question centrale est donc de savoir si ces approches peuvent réellement surpasser, en pratique, la robustesse des méthodes financières traditionnelles.

### Objectif du projet

Ce projet vise à comparer rigoureusement des approches classiques, représentées par l’EWM, le GARCH et le HAR-RV, et des approches modernes de deep learning, à travers des réseaux LSTM, pour la prédiction de la volatilité du S&P 500. L’évaluation est conduite à deux niveaux. Sur le plan statistique, il s’agit de comparer les erreurs de prévision à travers différentes métriques et de tester la significativité des écarts au moyen de procédures comme le test de Diebold-Mariano. Sur le plan pratique, la pertinence des prévisions est évaluée dans un backtest d’options consistant à simuler le pricing et la couverture delta d’un straddle at-the-money. Dans ce cadre, les modèles sont intégrés dans un protocole hors échantillon de type rolling avec une fenêtre de cinq années glissantes, recalibrée quotidiennement, produisant des prévisions de volatilité à un et cinq jours.

## II. Données et prétraitement

### Source des données

L’évaluation des modèles a été conduite sur deux horizons de prévision, un jour et cinq jours, afin de comparer leur capacité à estimer la volatilité réalisée. Les métriques retenues incluent le MAE et le RMSE, déjà standards dans la littérature, auxquels nous ajoutons la mesure QLIKE. Pour le MAE et le RMSE, des intervalles de confiance à 95 % ont été calculés à l’aide d’un bootstrap par blocs chevauchants de 21 jours, afin de tenir compte de la dépendance temporelle des données.

À l’horizon 1 jour, EGARCH et GJR présentent, dans notre échantillon, des MAE/RMSE numériquement inférieurs à ceux de la baseline EWM. Cependant, les tests de Diebold–Mariano (perte MSE, variance HAC de Newey–West, correction de Harvey) ne rejettent pas l’égalité de performance avec EWM (p ≥ 0,05). Nous concluons donc qu’aucune amélioration significative n’est établie à 1 jour ; EWM demeure une baseline robuste, le gain du HAR-RV restant modeste.  
À 5 jours, en revanche, le LSTM (et le Blend, le cas échéant) surperforment EWM/HAR de façon nette et significative (p < 0,05), avec des baisses concordantes de MAE/RMSE et un QLIKE plus faible.

### Présentation des séries utilisées

Trois séries principales ont été retenues afin d’assurer une analyse représentative et robuste. La première est le S&P 500 index (^GSPC), qui constitue le principal indice actions américain et la référence mondiale des marchés financiers. Il représente la série centrale du projet puisqu’il sert de base au calcul des rendements journaliers et de la volatilité réalisée. La seconde série est le SPY ETF (SPY), l’ETF le plus liquide répliquant le S&P 500. Son inclusion permet de rapprocher l’analyse du monde réel de la gestion de portefeuille, puisque le SPY est le support effectivement tradable par les investisseurs.

Enfin, la troisième série est le VIX index (^VIX), qui mesure la volatilité implicite à 30 jours extraite des options SPX. Le VIX offre une vision prospective du risque de marché et constitue un point de comparaison intéressant pour confronter volatilité réalisée, implicite et prédite. Le choix de ces trois actifs repose donc sur leur représentativité, leur complémentarité et leur accessibilité. Le S&P 500 et le SPY sont des benchmarks mondiaux, tandis que l’ajout du VIX permet de combiner une vision historique et une vision implicite de la volatilité. Par ailleurs, leur disponibilité gratuite et transparente via Yahoo Finance renforce la reproductibilité du projet.

### Variables dérivées et prétraitement

À partir de ces données brutes, plusieurs transformations ont été réalisées afin de préparer les séries pour la modélisation. Les rendements journaliers ont été calculés à partir des prix ajustés du S&P 500, corrigés des dividendes et des splits, afin d’obtenir une série cohérente et statistiquement exploitable. La volatilité réalisée a été estimée sur une fenêtre de vingt et un jours de bourse, ce qui correspond à environ un mois de marché et permet de disposer d’une mesure ex post du risque. En parallèle, une volatilité lissée a été calculée au moyen d’une moyenne exponentielle pondérée avec un facteur de décroissance fixé à 0,94, une valeur de référence largement utilisée dans le secteur financier, notamment dans le modèle RiskMetrics de J.P. Morgan.

Cette mesure sert de baseline robuste pour comparer les modèles. Afin de renforcer les capacités prédictives du LSTM, des variables supplémentaires ont été construites, notamment les retards des carrés de rendements et les retards de la volatilité EWM. Une transformation en log-volatilité a également été appliquée pour stabiliser la variance et limiter l’influence des valeurs extrêmes. Toutefois, cette transformation n’est utilisée qu’en phase d’entraînement, et l’ensemble des résultats finaux est systématiquement restitué en niveau journalier décimal afin de garantir la comparabilité entre les approches statistiques classiques et les approches fondées sur l’intelligence artificielle.

## III. Méthodes

### Modèles GARCH et extensions

Le modèle GARCH(1,1) est l’un des plus utilisés en économétrie financière pour modéliser la volatilité conditionnelle. Il repose sur l’idée que la volatilité d’aujourd’hui dépend à la fois des chocs passés (les rendements récents) et de la volatilité d’hier. Ce modèle capture bien le phénomène de clusters de volatilité, où les périodes calmes et agitées se succèdent.

Deux variantes ont également été étudiées. L’EGARCH permet de mieux représenter l’asymétrie, c’est-à-dire le fait que les chocs négatifs ont généralement un effet plus fort sur la volatilité que les chocs positifs. Le GJR-GARCH introduit un terme dit de “levier” qui vise à modéliser le même phénomène. Ces extensions corrigent certaines limites du GARCH de base, mais restent très paramétriques et imposent une structure fonctionnelle fixe à la dynamique de volatilité.

### Modèle HAR-RV

Le modèle HAR-RV (Heterogeneous AutoRegressive – Realized Volatility) a été introduit pour mieux prendre en compte les dynamiques de volatilité à différents horizons temporels. Plutôt que de ne regarder que le passé immédiat, ce modèle utilise des moyennes de volatilité quotidienne, hebdomadaire et mensuelle. Cela reflète l’idée que les acteurs de marché (day traders, gestionnaires de portefeuille, investisseurs institutionnels) agissent sur des horizons différents, et que la volatilité observée est le résultat de cette hétérogénéité. Le HAR-RV est simple, interprétable et souvent compétitif face à des modèles plus complexes, ce qui en fait une alternative pertinente au GARCH.

### Modèle LSTM

Les réseaux LSTM (Long Short-Term Memory) appartiennent à la famille des réseaux de neurones récurrents. Ils ont été conçus pour capter des dépendances complexes et de longue durée dans les séries temporelles, ce qui en fait un outil adapté à la prévision de volatilité.

Dans ce projet, l’architecture utilisée s’appuie sur des séquences de 60 à 90 jours de données en entrée. Les variables de base incluent les rendements journaliers du S&P 500, auxquelles ont été ajoutées des variables enrichies : les retards (lags) des carrés de rendements, les retards de la volatilité lissée (EWM), ainsi qu’une transformation en log-volatilité pour limiter l’influence des valeurs extrêmes. Le réseau comporte deux couches LSTM avec entre 64 et 128 neurones cachés, complétées par du dropout pour éviter le surapprentissage.

Pour l’entraînement, la fonction de perte choisie est la MAE ou le Huber loss appliquée sur la volatilité transformée en log, ce qui améliore la robustesse aux observations atypiques. Des techniques d’optimisation comme l’early stopping et la réduction adaptative du taux d’apprentissage ont également été utilisées pour stabiliser la convergence.

### Schéma expérimental et méthodologie de prévision

Les modèles présentés ont été évalués dans un cadre rigoureux de prévision hors-échantillon afin de reproduire des conditions proches de l’utilisation réelle. L’approche adoptée repose sur un protocole de type rolling, où une fenêtre d’apprentissage de cinq années glissantes est utilisée pour calibrer les modèles, puis génère des prévisions de volatilité à un jour et à cinq jours. Après chaque prévision, la fenêtre est décalée d’une journée afin de conserver une cohérence temporelle et d’éviter toute fuite d’information. Pour limiter le coût computationnel, les modèles paramétriques comme les GARCH sont recalibrés de manière mensuelle, ce qui permet de maintenir leur adaptabilité aux régimes de marché sans excès de recalcul.

Un soin particulier a été apporté à la définition des cibles afin d’assurer une parfaite correspondance entre les horizons de prévision et les métriques réalisées. Ainsi, les prévisions à un jour sont systématiquement comparées à la volatilité effectivement observée le jour suivant, tandis que les prévisions à cinq jours sont confrontées à la volatilité réalisée sur une période glissante de cinq jours. Ce calibrage évite tout décalage artificiel et garantit la validité des comparaisons entre approches.

Pour les modèles d’apprentissage profond, l’entraînement a été conduit avec des validations internes temporelles afin de choisir les hyperparamètres de manière robuste. Des techniques de régularisation comme le dropout et l’early stopping ont été employées afin de réduire le risque de surapprentissage, et le taux d’apprentissage a été ajusté de manière adaptative en fonction de la convergence. Ce protocole vise à assurer un compromis équilibré entre performance prédictive et stabilité statistique.

### Tests statistiques

Comparer des modèles ne se limite pas aux écarts de MAE/RMSE/QLIKE : nous vérifions systématiquement si ces écarts sont statistiquement significatifs. Pour les erreurs de prévision, nous appliquons un test de comparaison d’exactitude avec variance robuste et correction petit échantillon. Sauf mention contraire, la perte utilisée est la MSE sur la volatilité quotidienne, et la QLIKE (calculée sur la variance) est donnée en annexe. Les p-values sont évaluées bilatéralement avec un seuil de 5 %.  
Pour les backtests, nous comparons les retours sur les mêmes périodes à l’aide d’un test t apparié. Une méthode n’est dite supérieure que si l’écart est significatif (p < 0,05) et cohérent avec l’analyse économique.

## IV. Résultats empiriques

### Performances prévisionnelles

Nous discutons à la fois le niveau des erreurs (MAE/RMSE/QLIKE) et leur significativité via le test de Diebold–Mariano ; un écart n’est retenu que s’il est significatif (p < 0,05).

L’évaluation des modèles a été menée sur deux horizons, un jour et cinq jours, afin de comparer leur capacité à prédire la volatilité réalisée. Trois métriques sont utilisées : le MAE, le RMSE et le QLIKE, ce dernier étant couramment mobilisé dans la littérature car il pénalise particulièrement les prévisions sous-estimant les phases de forte volatilité. Pour le MAE et le RMSE, des intervalles de confiance à 95 % ont été obtenus par bootstrap en blocs de 21 jours, ce qui permet de tenir compte de la dépendance temporelle des séries financières.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Modèle | MAE (1j) | RMSE (1j) | MAE (5j) | RMSE (5j) |
| EWM | 9.84 | 13.41 | 5.55 | 8.33 |
| EGARCH(1,1) | 2.31 | 3.13 | N/A | N/A |
| GJR(1,1) | 2.88 | 4.28 | N/A | N/A |
| HAR-RV | 8.67 | 12.66 | 7.93 | 12.66 |
| LSTM | 7.78 | 11.89 | 4.54 | 6.91 |
| Blend | 7.73 | 11.64 | 4.54 | 6.91 |

Tableau 1 — Comparaison des erreurs de prévision (MAE, RMSE) aux horizons 1 jour et 5 jours.

Sauf mention contraire, toutes les erreurs sont exprimées en écart-type quotidien (σ\_d), valeurs décimales. Les comparaisons s’interprètent donc à unité identique.

Interprétation du Tableau 1 : Les valeurs brutes peuvent suggérer, à l’horizon 1 jour, des erreurs plus faibles pour EGARCH/GJR que pour EWM ; toutefois, ces écarts ne sont pas significatifs au test de Diebold–Mariano (p ≥ 0,05). EWM reste donc la référence opérationnelle à cet horizon. À 5 jours, les réductions d’erreur apportées par LSTM (et par le Blend) sont significatives et robustes, avec des p-values DM très faibles, ce qui valide leur supériorité sur EWM/HAR à cet horizon.

### Visualisation des résultats

Les barplots comparatifs des MAE, RMSE et QLIKE mettent en évidence la hiérarchie entre modèles. À un jour, les différences restent faibles et souvent couvertes par les intervalles de confiance. À cinq jours, en revanche, les LSTM et le Blend affichent des performances nettement supérieures, avec des intervalles bien séparés.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Figure 1 : Barplots des MAE avec intervalles de confiance à 95 % (bootstrap blocs 21 jours), horizon 1 jour. L’EWM reste robuste ; HAR apporte un léger gain. Les LSTM et Blend ne surclassent pas significativement à ce pas de temps.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Figure 2 : Barplots des MAE avec intervalles de confiance à 95 %, horizon 5 jours. Les LSTM et le Blend affichent des MAE significativement plus faibles, confirmant leur supériorité à moyen terme.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Figure 3 : Barplots des RMSE avec intervalles de confiance à 95 % (bootstrap en blocs de 21 jours) aux horizons 1 jour et 5 jours. À 1 jour, l’EWM/HAR restent difficiles à dépasser ; à 5 jours, les LSTM (et le Blend lorsqu’il est disponible) affichent des RMSE plus faibles.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Figure 4 : Barplots des QLIKE avec intervalles de confiance à 95 % aux horizons 1 jour et 5 jours. La métrique QLIKE, plus sévère sur les sous-estimations de volatilité, confirme l’avantage des modèles LSTM/Blend à 5 jours, avec des valeurs sensiblement plus bas.

Des timelines comparant volatilité réalisée et prévisions des modèles illustrent aussi les différences en période de crise. Les méthodes traditionnelles lissent les variations mais échouent à capter les pics, tandis que le LSTM enrichi suit mieux la dynamique moyenne et s’ajuste plus finement lors des phases turbulentes.

Une image contenant texte, capture d’écran, Tracé, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 5 : Timeline de la volatilité réalisée (Target) vs prévisions EWM, LSTM (et Blend le cas échéant) sur une fenêtre de crise. On observe que l’EWM lisse les pics tandis que le LSTM suit mieux la dynamique moyenne ; l’écart se réduit hors crise.

Une image contenant texte, ligne, Tracé, capture d’écran

Le contenu généré par l’IA peut être incorrect.

Figure 6 : Même comparaison à l’horizon 5 jours. La lissabilité accrue rend les prévisions LSTM/Blend plus proches de la volatilité réalisée, surtout lors du pic, ce qui corrobore les gains vus aux Figures 2–4.

### Analyse statistique des différences (tests DM et t-tests)

Pour vérifier la significativité des écarts, des tests de Diebold-Mariano ont été menés avec une variance HAC (Newey-West) et une correction de Harvey.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Horizon | Comparaison | Stat DM | p-value | Significatif ? | Meilleur modèle |
| 1 jour | HAR vs EWM | –3.533 | 0.00041 | Oui | HAR |
| 1 jour | LSTM vs EWM | –1.009 | 0.3128 | Non | NaN |
| 1 jour | LSTM vs HAR | 1.336 | 0.1815 | Non | NaN |
| 5 jours | HAR vs EWM | –0.123 | 0.9018 | Non | NaN |
| 5 jours | LSTM vs EWM | –4.248 | 2.15e–05 | Oui | LSTM |
| 5 jours | LSTM vs HAR | –4.289 | 1.80e–05 | Oui | LSTM |

Tableau 2 : Résultats des tests de Diebold–Mariano (statistique et p-value) pour HAR vs EWM, LSTM vs EWM et LSTM vs HAR, aux horizons 1 jour et 5 jours.

Les résultats montrent qu’à horizon d’un jour, seul le HAR-RV surpasse significativement l’EWM, tandis que le LSTM n’apporte pas de gain robuste. À horizon cinq jours, les LSTM et le Blend dépassent largement l’EWM et le HAR, avec des p-values très inférieures à 1 %.

## V. Backtest applicatif

### Stratégie de straddles ATM avec couverture delta

Pour évaluer la pertinence pratique des prévisions de volatilité, un backtest a été mis en place sur une stratégie d’options. Le choix s’est porté sur des straddles at-the-money, c’est-à-dire une combinaison simultanée d’un call et d’un put de même maturité et même prix d’exercice, centrée sur le prix courant du S&P 500. Cette stratégie est directement sensible au niveau de volatilité anticipé, ce qui en fait un terrain d’expérimentation idéal pour comparer les modèles.

Chaque straddle a été simulé avec une maturité fixe d’environ 21 jours de bourse, et une couverture dynamique par delta-hedging quotidien a été appliquée. Cette couverture permet de neutraliser l’exposition directionnelle au sous-jacent, de façon à isoler l’impact de la prévision de volatilité sur la performance de la stratégie. En pratique, la valeur initiale du straddle est calculée à l’aide de la formule de Black-Scholes en remplaçant la volatilité implicite par les prévisions issues des modèles (EWM, HAR-RV, LSTM et blend). Chaque jour, le portefeuille est rééquilibré pour maintenir une couverture delta-neutre.

Cette approche permet de tester non seulement la précision statistique des prévisions, mais surtout leur utilité économique : un modèle qui prédit mieux la volatilité doit logiquement générer une meilleure performance ajustée du risque dans ce type de backtest.

Unité des résultats. Sauf mention contraire, les retours de backtest sont exprimés en fraction (p. ex. 0,10 = 10 %), et non en pourcentage.

### Résultats du backtest et interprétation

Lecture des backtests (horizon straddle ≈ 21 j). Les résultats corrigés ne présentent plus d’artefacts (séries identiques/écart-types séquentiels). À 1 jour, les retours (en fraction) sont en moyenne d’environ 0,059 pour EWM, 0,380 pour HAR et 0,934 pour LSTM, avec des écarts-types ~0,362, 0,641 et 0,808 respectivement. Les t-tests appariés confirment HAR > EWM et LSTM > HAR/EWM (p-values ≪ 1 %). À 5 jours, la hiérarchie est similaire et reste significative (cf. Tableau 3).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Horizon | Modèle | Ret moyen | Ret écart-type | Sharpe |
| 1 jour | EWM | 0.058709 | 0.362495 | 0.162 |
| 1 jour | HAR | 0.379937 | 0.641432 | 0.592 |
| 1 jour | LSTM | 0.933809 | 0.808385 | 1.155 |
| 5 jours | EWM | 0.058709 | |  | | --- | | 0.362495 |  |  | | --- | |  | | 0.162 |
| 5 jours | HAR | 0.127632 | 0.448827 | 0.284 |
| 5 jours | LSTM | 0.218025 | 0.384721 | 0.567 |

Tableau 3 : Présente le retour moyen, l’écart-type des retours et un ratio de type Sharpe (ret\_moyen / ret\_écart-type) pour chaque modèle.  
Note. Les valeurs sont des retours fractionnels (0,10 = 10 %). Les colonnes PnL sont retirées pour éviter un biais d’échelle ; les comparaisons s’appuient sur Ret et les t-tests appariés.

En revanche, les modèles basés sur le deep learning, et en particulier le LSTM enrichi, se distinguent. Leurs prévisions se traduisent par des PnL moyens nettement supérieurs, avec des ratios de type Sharpe plus élevés et des distributions de résultats plus favorables. Le modèle blend (pondération entre LSTM et EWM) présente également de bonnes performances, proches de celles du LSTM seul, avec une stabilité accrue. Sur l’horizon de cinq jours, les résultats sont encore plus marqués : les LSTM dominent systématiquement les approches classiques, aussi bien en moyenne qu’en termes de régularité des retours.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Figure 7 : Boxplots des PnL journaliers pour les modèles à l’horizon 1 jour. LSTM et Blend affichent des distributions décalées vers le haut, mais avec une variabilité plus forte, tandis que l’EWM reste concentré autour de zéro.

Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 8 : Boxplots des PnL journaliers à l’horizon 5 jours. LSTM et Blend présentent des rendements journaliers plus stables et supérieurs, confirmant leur avantage économique par rapport aux modèles traditionnels.

Une image contenant texte, Tracé, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Figure 9 : Courbes cumulées des equity à l’horizon 1 jour. Les performances sont globalement modestes, avec une légère surperformance du LSTM et du Blend.

Une image contenant texte, Tracé, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Figure 10 : Courbes cumulées des equity à l’horizon 5 jours. Les stratégies LSTM et Blend dominent nettement en termes de croissance cumulée et de résilience, confirmant les résultats du Tableau 3.

Les graphiques issus du backtest illustrent clairement cette hiérarchie. Les boxplots révèlent une distribution plus favorable pour les modèles neuronaux, et les courbes cumulées montrent une progression régulière pour le LSTM, tandis que celles de l’EWM et du HAR restent quasi plates.

Afin d’analyser plus finement la nature des retours, des histogrammes des PnL journaliers ont également été produits. Ces distributions mettent en évidence la différence de profil de risque entre modèles : les approches traditionnelles génèrent des résultats concentrés mais souvent asymétriques, tandis que les LSTM produisent des distributions plus équilibrées et moins exposées aux queues extrêmes.

Une image contenant Tracé, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 11 : Histogrammes des PnL journaliers par modèle à l’horizon 1 jour. Les distributions présentent de fortes queues à droite ; les modèles LSTM et Blend offrent des gains plus fréquents et plus élevés, tandis que l’EWM reste concentré autour de zéro.

Une image contenant capture d’écran, Tracé, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 12 : Histogrammes des PnL journaliers à l’horizon 5 jours. Les distributions sont plus resserrées, mais LSTM et Blend se distinguent encore par une masse décalée vers des valeurs positives, confirmant les résultats du Tableau 3.

Enfin, des t-tests pairés sur les retours des stratégies de straddle delta-hedgées confirment quantitativement ces résultats.

|  |  |  |  |
| --- | --- | --- | --- |
| Horizon | Comparaison | t\_stat | p\_value |
| 1 jour | LSTM vs EWM | 46.6252 | 2.023e-266 |
| 1 jour | LSTM vs HAR | 33.2069 | 1.056e-169 |
| 1 jour | Blend vs LSTM | -36.7180 | 1.933e-195 |
| 1 jour | Blend vs EWM | 49.2585 | 1.322e-284 |
| 1 jour | HAR vs EWM | 25.3373 | 1.036e-112 |
| 5 jours | LSTM vs EWM | 24.4477 | 1.804e-106 |
| 5 jours | LSTM vs HAR | 10.3903 | 3.182e-24 |
| 5 jours | Blend vs LSTM | 1.5522 | 0.1209 |
| 5 jours | Blend vs EWM | 24.4477 | 1.804e-106 |
| 5 jours | HAR vs EWM | 8.6144 | 2.309e-17 |

Tableau 4 : Résultats des t-tests pairés (p-values en notation scientifique).  
Les comparaisons montrent que les trois modèles (HAR, LSTM, Blend) battent très significativement EWM, tant à 1 jour qu’à 5 jours.

À 5 jours, les différences entre LSTM et Blend ne sont pas significatives, suggérant des performances équivalentes, tandis que tous deux surpassent HAR et EWM.

En conclusion, le backtest confirme que l’intérêt des modèles IA ne se limite pas aux métriques statistiques classiques mais s’étend à une meilleure performance économique dans un cadre de trading simulé. Cette dimension pratique valide l’apport concret du LSTM enrichi et de ses variantes face aux méthodes paramétriques traditionnelles.

## VI. Discussion

### Horizon court : EWM et HAR-RV comme références solides

Les résultats montrent que sur l’horizon très court, à savoir la prévision de la volatilité à un jour, les méthodes classiques demeurent difficiles à battre. L’EWM avec λ = 0,94, déjà largement utilisée dans l’industrie (notamment via le modèle RiskMetrics), fournit une estimation robuste et réactive. Le HAR-RV confirme un léger avantage statistique, mis en évidence par les tests de Diebold-Mariano. Toutefois, les intervalles de confiance obtenus par bootstrap révèlent que les différences entre modèles à cet horizon restent limitées et souvent recouvertes, ce qui explique pourquoi les gains du HAR-RV ou du LSTM ne sont pas systématiquement significatifs. Dans ce contexte, les modèles de machine learning, et en particulier le LSTM, n’apportent pas d’amélioration décisive et peuvent même souffrir du bruit élevé présent dans les séries quotidiennes.

### Horizons lissés : apport clair des modèles LSTM

À des horizons plus longs, comme la prévision de la volatilité sur cinq jours, la situation change radicalement. Les LSTM enrichis par des variables supplémentaires (lags, log-vol, volatilité lissée) exploitent des dynamiques que les approches paramétriques classiques n’intègrent pas. Les tests de Diebold-Mariano, corrigés selon Harvey et avec variance HAC, confirment cette supériorité, et les backtests applicatifs traduisent cette précision accrue en performances économiques mesurables. Les PnL moyens des straddles simulés sous couverture delta sont significativement plus élevés avec les prévisions issues du LSTM qu’avec celles des méthodes traditionnelles, et les ratios de type Sharpe en témoignent également. Ainsi, l’intelligence artificielle démontre un intérêt réel dès lors que l’on cherche à lisser le signal et à capturer des structures complexes au-delà de l’horizon journalier.

### Cohérence entre métriques et backtest

Une différence notable apparaît entre les résultats statistiques et applicatifs à horizon d’un jour : les tests de Diebold-Mariano ne montrent pas de supériorité robuste du LSTM face au HAR-RV, mais les backtests indiquent pourtant une amélioration des performances économiques. Cette apparente contradiction s’explique par la nature des erreurs de prévision. Les erreurs du LSTM, bien que proches en moyenne de celles du HAR-RV, interagissent différemment avec la structure de payoff d’un straddle delta-hedgé. En pratique, une meilleure gestion des phases de forte volatilité permet au LSTM de générer un PnL supérieur malgré une absence de domination statistique claire sur le plan du MSE.

### Limites identifiées

Malgré ces résultats encourageants, plusieurs limites doivent être soulignées. Le risque de surapprentissage reste important dans les modèles LSTM, qui exigent un réglage fin des hyperparamètres (taille des fenêtres, profondeur des couches, régularisation). Le choix des variables d’entrée influe fortement sur la qualité des prévisions : un jeu trop restreint limite la capacité du modèle, tandis qu’un jeu trop riche accroît le risque de bruit. Enfin, il faut rappeler que la baseline EWM est particulièrement difficile à battre, en raison de sa simplicité, de son efficacité et de son ancrage solide dans la pratique de gestion du risque. Ces contraintes rappellent que l’usage du deep learning en finance doit être accompagné d’une validation rigoureuse et d’un contrôle strict des biais.

### Perspectives d’amélioration

Ce projet ouvre néanmoins la voie à des perspectives intéressantes. Une première piste consiste à explorer des architectures plus récentes comme les Transformers et les modèles à mécanismes d’attention, qui ont montré leur efficacité dans d’autres domaines de séries temporelles. L’élargissement des données utilisées pourrait également renforcer la robustesse : intégration de volatilités implicites détaillées (smiles et surfaces), d’indicateurs macroéconomiques ou de données textuelles issues des marchés (actualités, discours des banques centrales). Enfin, un backtest plus réaliste incluant les coûts de transaction et des scénarios de volatilité implicite permettrait de rapprocher encore davantage les résultats des conditions de marché réelles.

## VII. Conclusion du projet

Ce projet avait pour objectif de comparer des approches classiques de modélisation de la volatilité (EWM, GARCH, HAR-RV) à des méthodes modernes de deep learning (LSTM) appliquées à l’indice S&P 500. Les résultats mettent en évidence la robustesse remarquable de l’EWM, qui reste difficile à surpasser à l’horizon d’un jour. Le HAR-RV obtient un léger avantage statistique sur ce pas de temps, tandis que les modèles LSTM n’apportent pas de gain significatif, en raison du bruit élevé des séries quotidiennes. En revanche, à l’horizon de cinq jours, les LSTM enrichis surpassent nettement les méthodes traditionnelles, capturant des dynamiques que les modèles paramétriques ne parviennent pas à exploiter.

La validation a été double. Sur le plan statistique, les métriques d’erreur (MAE, RMSE, QLIKE) et les tests de Diebold-Mariano confirment la supériorité des LSTM à moyen terme, avec des résultats hautement significatifs (p ≪ 1 %). Sur le plan économique, les backtests d’options montrent que les stratégies de straddles delta-hedgées construites à partir des prévisions LSTM génèrent des PnL moyens supérieurs, avec de meilleures distributions de retours et des ratios de type Sharpe plus élevés. Cette convergence entre validation statistique et validation économique constitue un apport majeur du projet.

Certaines limites doivent toutefois être rappelées. La prédiction de la volatilité quotidienne reste très bruitée, et les LSTM nécessitent un calibrage attentif pour éviter le surapprentissage. Par ailleurs, le backtest repose sur des hypothèses simplificatrices, notamment l’usage du modèle de Black-Scholes et l’absence de coûts de transaction, ce qui limite la portée immédiate des résultats pour une mise en œuvre réelle.

Enfin, ce travail illustre aussi l’importance de la reproductibilité scientifique. L’intégralité des analyses repose sur des données accessibles publiquement (Yahoo Finance) et sur des notebooks Python documentés, garantissant que l’ensemble des résultats peut être répliqué et prolongé. Cette ouverture constitue une base solide pour de futures extensions, qu’il s’agisse d’intégrer la volatilité implicite, des facteurs macroéconomiques ou de tester de nouvelles architectures de deep learning comme les Transformers.

### Références bibliographiques

* Diebold, F.X., & Mariano, R.S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*: [(LIEN)](https://scispace.com/pdf/comparing-predictive-accuracy-2tm17nazab.pdf)
* Corsi, F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility (HAR-RV). *Journal of Financial Econometrics*: [(LIEN)](https://statmath.wu.ac.at/~hauser/LVs/FinEtricsQF/References/Corsi2009JFinEtrics_LMmodelRealizedVola.pdf)
* J.P. Morgan & Reuters (1996). RiskMetrics: [(LIEN)](https://www.msci.com/documents/10199/5915b101-4206-4ba0-aee2-3449d5c7e95a)