MARKET RISK ESTIMATION USING GARCH MODELS, AI AND HYBRID METAMODELS: EVIDENCE FROM 30 FINANCIAL ASSETS

Paul Cristian Donoiu[1*], Bogdan Ionuț Anghel[2]
[1)2)] *The Bucharest University of Economic Studies, Bucharest, Romania*

Abstract
This article assesses the out-of-sample performance of classical volatility-based econometric models, modern AI models and meta-hybrid models in estimating financial risk for 30 international financial assets. Using daily returns computed from Yahoo Finance price data over the 2001–2024 period, we generated one-step-ahead forecasts of Value at Risk (VaR) and Expected Shortfall (ES). VaR accuracy was evaluated through unconditional and conditional coverage tests (Kupiec and Christoffersen), while ES was assessed using a complementary tail-oriented diagnostic. The empirical analysis includes models from the GARCH family (GARCH, EGARCH, GJR-GARCH and APARCH), AI-based models (GRU, TCN, Quantile Random Forest and Quantile Boosting), and four metamodelling frameworks that combine forecasts from different approaches and, in some cases, include exogenous information such as the VIX. The results show that no single model is uniformly superior across all assets. However, metamodelling, especially Meta 1 and Meta 2, provides the most stable overall performance across assets and the most consistent VaR-ES calibration. AI models can be competitive for some assets, but their performance is more heterogeneous, and some specifications tend to underestimate tail risk in the absence of additional calibration mechanisms. Overall, the findings suggest that metamodelling improves the robustness of financial risk estimation by reducing the limitations associated with relying on a single model and by increasing consistency across assets and market conditions.

Keywords
GARCH, artificial intelligence, market risk; value at risk; expected shortfall, backtesting, hybrid modelling.

JEL Classification
G11; G15; G17

[*] Corresponding author, **Paul Cristian Donoiu –** paul_donoiu@yahoo.com.

Introduction

The rigorous evaluation of market risk is a key component for both the stability of the financial system and the efficiency of risk management and regulatory processes. After the global financial crisis in 2008, the importance of the instruments used to calculate extreme losses, such as Value at Risk (VaR) and Expected Shortfall (ES), has increased significantly, with these two becoming international standards for market prudential supervision. Therefore, the scientific literature on risk estimation models has expanded significantly, from classical approaches to modern models often based on artificial intelligence.

The GARCH family models have been used widely in recent years due to their capacity to identify specific characteristics of financial time series, such as conditional heteroskedasticity and volatility clustering. These models have a solid theoretical background and a clear economic interpretation, which makes them very attractive for financial institutions and regulatory authorities. However, their parametric structure and restrictive hypotheses about the returns' distribution can affect their performance during periods of market turbulence or in the presence of exogenous shocks.

Additionally, technological advances and the rapid development of AI tools have led to the emergence of alternative models for risk estimation, capable of identifying nonlinear dependencies, complex patterns, and difficult-to-detect interactions in the data. Deep learning models can offer, at least theoretically, greater flexibility and a superior ability to adapt to different market regimes. However, exclusive reliance on these models also generates certain problems, such as the risk of overfitting, dependence on the volume and quality of available data, and difficulties in economic interpretation and justification.

In this context, the research questions of this paper have both practical and theoretical relevance: Could AI models enhance the performance of traditional GARCH models in estimating financial risk? Moreover, can hybrid solutions, resulting from combining individual models using a metamodelling technique, generate more accurate and robust results? To answer these questions, we performed a comparative analysis of representative GARCH models, modern AI tools, and a set of hybrid metamodels designed to efficiently exploit the complementarity of these approaches.

The empirical analysis used 30 financial assets with international relevance, covering a sufficiently long period to include both normal market regimes and intervals of high volatility and financial stress. To evaluate the performance of the models, we used tests commonly used in the literature, such as the Kupiec test for unconditional coverage, the Christoffersen test for conditional coverage, and a test designed to assess the accurate estimation of ES. This methodological approach allowed a rigorous validation of the results.

The contribution of this paper lies in providing a unified comparative evaluation of classical GARCH models, AI-based models and metamodelling frameworks for VaR and ES estimation across a broad cross-section of international financial assets. In particular, the study highlights not only differences in predictive performance, but also the relative robustness of these approaches across heterogeneous assets and market conditions.

All in all, the paper argues that the evolution of financial risk modelling does not imply replacing classical econometric models, but integrating them with modern AI tools to develop more robust, flexible, and efficient risk estimation frameworks that can better adapt to contemporary market features.

1. Review of the scientific literature

Financial risk represents one of the most important concepts in modern finance, highlighting the uncertainties associated with the future evolution and the deviations from the estimated trend.

One of the benchmark scientific works on defining financial risk is Christoffersen (2003), which provides a perspective on the concept grounded in econometric principles. According to the author, financial risk results from the variance of returns and from the exposure of companies and investors to adverse market conditions. In the introductory part of his paper, Christoffersen links risk with the uncertain dynamics of financial time series and concludes that risk should be defined as an inherent characteristic of all financial activities rather than an anomaly to be avoided.

Unlike Christoffersen, Jorion (2001) uses a more operational approach, defining financial risk as the probability of financial losses generated by adverse movements in market conditions. Thus, Jorion focuses on risk dimensions that can be measured and managed, resulting from the exposure of assets and institutions to market fluctuations. In this context, the author does not treat risk merely as a theoretical manifestation of uncertainty but defines it as a measurable variable with a fundamental role in the decision-making process in financial institutions.

Alexander (2005) adds new dimensions to financial risk and, implicitly, to its management activity, defining it as a systemic and integrative discipline rather than as a set of isolated volatility-control functions. Moreover, Hull (2018) extends the perspective and the level of understanding of the concept of financial risk by placing it at the intersection between investment decisions and institutional stability. The author defines financial risk not from the perspective of exposure to uncertainty in financial markets, but rather as a condition/state inherent to corporate and financial survival.

Financial risk modelling has undergone an important transformation in recent decades, through the transition from deterministic models, which assumed the existence of stable and predictable relationships between variables, to more complex models, which identified and highlighted the need to evaluate the conditional relationships among the analysed variables.

A relevant milestone in the scientific literature on financial risk modelling was Robert Engle's study in 1982, which introduced the concept of conditional variance in the analysis of financial time series and developed the ARCH model. Bollerslev (1986), Engle's PhD student, strengthened the ARCH methodology by developing the GARCH model, one of the most widely used models for analysing volatility even today, both by researchers and practitioners. The most important methodological evolution is represented by this extension of ARCH: in Bollerslev's view, variance is influenced not only by past shocks, but also by autoregressive components.

Diebold and Mariano's (1995) study marked a paradigm shift in the scientific literature, as the quality of models began to be evaluated not only by their econometric properties,

but also by their ability to forecast the evolution of financial time series. In this study, the authors introduced a framework for comparing the predictive power of models, known as the Diebold–Mariano test. This test focuses on evaluating forecast performance, unlike earlier instruments that primarily assess estimation accuracy. The flexibility of the Diebold–Mariano test has made it one of the most widely used instruments in financial risk management, as it is applied extensively in forecasting and validating indicators such as VaR.

Over the past 10–20 years, a conceptual evolution has taken place in financial risk management, through a transition from linear and deterministic models to nonlinear models, a necessary shift given the increasing complexity and uncertainty of financial systems. One of the studies highlighting the need for conceptual adaptation to the new realities of financial markets is Sirignano and Cont (2019). The authors used deep learning models to analyse large volumes of data to assess the relationship between market order dynamics and price movements. According to them, there is a universal and stationary relationship between the two elements, thereby highlighting the importance of investors' orders in the formation of financial asset prices. Moreover, the study shows that the model developed by Sirignano and Cont delivers better results than models specific to a single asset or an asset class, which indicates its broad applicability.

After the methodological shift represented by the use of nonlinear and deep learning models, the next relevant conceptual step in financial scientific research was the development of AI instruments, used especially in fields such as credit risk, solvency evaluation, or financial volatility estimation. A comprehensive analysis of the impact of AI on financial risk modelling was conducted by David et al. (2024). The authors evaluate the impact of AI on analysing several dimensions of financial risk, such as market risk, portfolio risk, or systemic risk, showing that models using this technology perform better than classical econometric models in terms of both robustness and accuracy. For example, the authors show that classical models such as VaR fail to identify nonlinear relationships between financial time series, which highlights the need to use modern approaches such as neural networks, which have this capability. Moreover, the authors state that AI-based models allow the development of a more precise and stable strategy for managing financial risk.

2. Research methodology

The purpose of this study is to evaluate the efficiency of some categories of models, both individual and resulting from a metamodelling process, in estimating financial risk. The empirical analysis uses 30 internationally relevant assets, selected to ensure global coverage and sectoral diversification. The dataset consists of major global indices (S&P 500, NASDAQ, FTSE 100, DAX, NIKKEI 225, HANG SENG INDEX) and blue-chip companies from the technology, financial, energy, and consumer sectors. Daily data were obtained from the Yahoo Finance database for the 2001–2024 period, which allows us to assess model efficiency across multiple phases of the economic cycle and different market regimes, including both stable periods and intervals with high volatility. Using the price data series, we computed logarithmic returns, which were used in all modelling stages.

The first methodological layer involves classical econometric models from the GARCH family used to estimate financial risk. These models are well known in the financial literature for their ability to identify volatility clustering and conditional risk dynamics. We used several models from the same family (GARCH, EGARCH, GJR-GARCH, and APARCH) to identify asymmetrical effects, different reactions to volatility, and risk persistence. For each model, we obtained out-of-sample estimates of VaR and ES, which are the benchmarks for the analysis.

The second methodological approach consists of using AI-based models. We selected them because of their capacity to identify nonlinear patterns in financial time series and structural changes in market dynamics. We used four widely studied AI models: GRU (a recurrent neural network for time series), LightGBM (quantile-calibrated), Quantile Random Forest (QRF) and Temporal Convolutional Network (TCN).

The central methodological element is represented by metamodelling, which implies developing models that combine the strengths of classical and AI approaches.

Meta 1 is the first model family, which consists of different combination schemes of VaR and ES estimates provided by the individual models. We applied five different models: QRA-Simplex (a weighted average of the baseline VaR estimates),

Weighted Quantile Combination (also a weighted average, but with regularization on the weights, to reduce instability), SQRA-CP (a weighted combination that also includes a coverage penalty, to keep the exception rate close to the theoretical level - 5%), EG-Hedge (an online mechanism that shifts weight toward the models that have performed better recently in terms of quantile loss) and Trimmed Quantile Pool (a robust combination that computes a trimmed mean of the model quantiles).

The Meta 2 model uses a more conservative and more pragmatic approach. An ensemble of GARCH models is used as an "anchor", and information from the AI models is incorporated subject to constraints: AI is used only when there is sufficient evidence that the estimation can be improved (for example, during periods of high volatility, when AI models are known to deliver superior performance). Practically, Meta 2 was designed to use the potential of AI without affecting the stability and robustness of the GARCH family models.

The Meta 3 model is based on the idea that model efficiency is influenced by the market regime. Specifically, the weights assigned to the GARCH and AI inputs adapt depending on volatility dynamics and the prevailing market regime.

The Meta 4 model extends the methodological framework by introducing an exogenous variable, namely the VIX, which is a widely used global index of uncertainty and financial stress. The purpose is to examine whether including a systemic risk indicator can improve VaR and ES estimates—specifically, whether the models become more cautious when the VIX signals high financial stress and less conservative during stable periods.

The entire analysis is conducted within a rigorous methodological framework, with no look-ahead bias and no use of future information. The models are recalibrated using a rolling-window scheme, and performance is evaluated with tests widely used in the risk literature: the Kupiec test for VaR coverage, the Christoffersen test for checking dependence and clustering of exceptions, and Expected Shortfall evaluation to capture the severity of extreme losses.

Overall, the methodological approach is designed to answer relevant questions for both researchers and practitioners: can combining classical econometric models (GARCH) with AI models improve the estimation of financial risk, providing more robust, stable, and informative estimates than using individual models, both in terms of coverage (VaR) and the severity of extreme losses (ES).

Expected Shortfall was evaluated as a complementary measure of tail-risk adequacy, with the purpose of assessing whether the models provide not only acceptable VaR coverage, but also a sufficiently reliable description of the severity of losses beyond the VaR threshold. This additional perspective is important because, in market risk applications, a model may perform reasonably well in terms of exception counting while still providing weak information about the magnitude of extreme losses. Consequently, the joint use of VaR and ES diagnostics allows for a broader assessment of model performance and supports a more informative comparison between classical, AI-based and hybrid approaches.

At the implementation level, all models were estimated and evaluated under the same forecasting protocol to ensure a fair comparison across specifications and asset classes. The empirical exercise was conducted in an out-of-sample framework based on sequential rolling-window re-estimation and one-step-ahead forecasting, so that the information set available at each prediction date remained strictly limited to past observations. This common setup was applied to both the individual models and the metamodels, ensuring that differences in performance reflect the properties of the modelling approaches rather than differences in estimation design. In this sense, the methodological objective was not only to compare forecast accuracy but also to assess the robustness and consistency of VaR and ES estimation across heterogeneous financial assets and market conditions.

3. Results and discussion

Across the 30 assets, (Table no.1) points to three robust patterns. First, the strongest overall performers are concentrated among the better-calibrated AI models (most notably GRU Quantile-VaR) and the leading meta-combinations, especially Meta 2 (Selective Overlay) and the best Meta 1 variants (the more stable convex/regularized combination rules). In practical terms, these approaches tend to be more consistently acceptable across assets: they more frequently deliver VaR exception behaviour that is statistically defensible and, at the same time, show comparatively stronger validation on the ES side, which matters when the objective is not only to "count exceedances" but also to capture the severity of tail losses when exceedances occur. Second, the GARCH-family benchmarks remain competitive and often stable—particularly on VaR coverage and clustering diagnostics—yet the results suggest a tendency toward conservatism (hit ratios often below the nominal 5%), and they do not dominate on ES validation in a uniform way, indicating that modelling conditional volatility well does not automatically translate into the best performance for tail expectation. Third, the weakest outcomes are concentrated among some AI quantile learners in this implementation—especially QRF and certain LightGBM quantile specifications—which exhibit systematically higher exception rates (frequently above the nominal level) and therefore fail VaR coverage criteria for a sizeable share of assets; in a risk-management setting,

such behaviour is difficult to defend without additional calibration, regularisation, or a more conservative post-processing layer. Overall, the table supports the interpretation that metamodeling improves robustness primarily by stabilizing tail forecasts across heterogeneous assets, while "pure" AI can be excellent when well-calibrated (GRU) but can also be fragile when the quantile-learning setup underestimates risk.

Table no. 1. Overall Model Performance

Model	Mean HitRatio	PassRate VaR_UC	PassRate VaR_CC	PassRate ES	Mean Kupiec_p	Mean CC_p	Mean ES_p	Mean Score
AI-GRU-Quantile-VaR	0.0534	1.0000	0.8333	1.0000	0.4782	0.3361	0.4765	2846.1711
META-2-SelectiveOverlay	0.0497	0.9667	0.9667	0.8333	0.4987	0.3871	0.4255	2779.7075
META 1-WQC-Convex	0.0735	0.9667	0.9000	1.0000	0.2778	0.3067	0.3403	2875.7887
META 1-SQRA-CP	0.0730	0.9333	0.9000	1.0000	0.3116	0.3109	0.3682	2843.1180
META 1-EG-Hedge	0.0486	0.9333	0.8667	0.9333	0.4287	0.3740	0.4861	2746.1461
AI-TCN-Quantile-VaR	0.0581	0.9333	0.7333	0.9667	0.4773	0.2819	0.5304	2646.1536
GARCH	0.0385	0.9000	0.8667	0.5333	0.3072	0.2980	0.2376	2308.3309
META 1-TQP-θ(0.10)	0.0532	0.9000	0.8333	0.9000	0.4473	0.3806	0.4743	2646.2747
APARCH	0.0379	0.8333	0.8000	0.5333	0.3130	0.2772	0.2343	2174.8124
GJR-GARCH	0.0383	0.8000	0.8667	0.5333	0.3414	0.2975	0.2384	2208.6786
EGARCH	0.0383	0.8000	0.8333	0.5333	0.2930	0.2741	0.2256	2174.4917
META 1-QRA-Simplex	0.0831	0.7667	0.7333	1.0000	0.2119	0.2455	0.2809	2507.2158
META-4-ExoVIXMix	0.0934	0.5333	0.6000	0.8333	0.1055	0.1569	0.2011	1971.0844
META-3-RegimeMix(VOL)	0.1066	0.3000	0.3667	0.6000	0.0734	0.0805	0.1148	1269.0700
AI-QRF-VaR	0.1268	0.1000	0.0667	0.0333	0.0109	0.0091	0.0096	199.9116
AI-LGBM-Quantile-VaR	0.1486	0.0000	0.0000	0.0333	0.0006	0.0006	0.0045	32.8971

Source: Author's own calculations.

The score reported in the **table no. 2)** was constructed as a synthetic ranking indicator designed to combine the main dimensions of model adequacy considered in this study. More specifically, it is computed as 1000 × the number of tests passed among UC, CC and ES, plus 10 × the sum of the corresponding p-values, and minus a calibration penalty of 5 × |HitRatio − α|, to discourage specifications whose exception frequency departs from the nominal level. The logic of this composite measure is to give primary

importance to the formal acceptance of the relevant backtesting procedures, while also preserving additional information about the statistical strength of the results and the degree of calibration of the model. From (Table no. 2), it appears that the best models differ substantially across assets: for many individual assets, the best models are AI quantile-based models (AI-GRU-Quantile-VaR and AI-TCN-Quantile-VaR), while for indices the best-performing models are metamodels (for example, Meta-2-SelectiveOverlay and META-EG-Hedge). On the whole, while some GARCH variations remain competitive (for example, GJR-GARCH or even GARCH), metamodelling and AI can improve estimation accuracy, depending on the asset analysed and the market regime.

Table no.2. Best Performing Model by Asset

Ticker	Model	Score
AAPL	META-2-SelectiveOverlay	3022.9021
BAC	META-SQRA-CP	3014.9798
BP	META-QRA-Simplex	3011.8765
CAT	AI-TCN-Quantile-VaR	3025.4337
CVX	GJR-GARCH	3024.1304
DIS	AI-GRU-Quantile-VaR	3021.1558
GE	AI-GRU-Quantile-VaR	3016.5178
GS	META-EG-Hedge	3018.3910
HSBC	AI-GRU-Quantile-VaR	3020.0281
IBM	META-QRA-Simplex	3025.2952
INTC	META-EG-Hedge	3019.6795
JNJ	AI-GRU-Quantile-VaR	3020.1171
JPM	AI-GRU-Quantile-VaR	3012.9596
KO	META-TQP-θ(0.10)	3025.2545
MCD	META-TQP-θ(0.10)	3024.0916
MSFT	AI-GRU-Quantile-VaR	3019.5210
NVDA	AI-TCN-Quantile-VaR	3024.3431
NVS	META-TQP-θ(0.10)	3019.5115
PFE	AI-TCN-Quantile-VaR	3021.7710
PG	AI-GRU-Quantile-VaR	3016.1176
TSM	META-SQRA-CP	3019.2053
UBS	META-2-SelectiveOverlay	3018.9654
XOM	META-EG-Hedge	3025.1923
^DJI	META-TQP-θ(0.10)	3022.4707

^FTSE	GJR-GARCH	3021.5675
^GDAXI	META-2-SelectiveOverlay	3016.1283
^GSPC	META-EG-Hedge	3023.4779
^HSI	META-2-SelectiveOverlay	3019.9310
^IXIC	GJR-GARCH	3018.6466
^N225	GARCH	3008.6485

Source: Author's own calculations.

In Table no. 3, we used the same classification algorithm as in Table no. 2. Based on this criterion, Table no. 3 shows the frequency with which each model appears in the top 3 for every analysed asset. The results indicate that metamodelling approaches offer superior performance: Meta-Top-θ (14,4%), Meta-2-SelectiveOverlay (13,3%), while AI-GRU-QUANTILE-VAR is the best among AI models (11,1%). By contrast, GARCH models appear less often in the top 3 (GJR-GARCH-6, EGARCH-5, APARCH-4, GARCH-2), which suggests that, in terms of VaR-ES combined robustness, top performance is achieved more often by metamodels and some of the AI models.

Table no.3. Frequency of Top-3 Appearances

Model	Top 3 Count	Top 3 Share
META-TQP-θ(0.10)	13	0.1444
META-2-SelectiveOverlay	12	0.1333
AI-GRU-Quantile-VaR	10	0.1111
META-EG-Hedge	9	0.1
META-SQRA-CP	9	0.1
META-WQC-Convex	8	0.0889
AI-TCN-Quantile-VaR	7	0.0778
GJR-GARCH	6	0.0667
EGARCH	5	0.0556
APARCH	4	0.04444
META-QRA-Simplex	3	0.0333
GARCH	2	0.0222
META-3-RegimeMix(VOL)	1	0.0111
META-4-ExoVIXMix	1	0.0111

Source: Author's own calculations.

Table no.4 compares each model family (AI, META 1, META 2, META 3, META 4) with a benchmark – the best GARCH model for each ticker. For each family, we showed the number of times that the best-performing model obtained better results than the best GARCH model. Results indicate that Meta-1 has the best performance, winning in 24 out of 30 cases (80%), followed by Meta-2 (76,7%) and AI (63,3%). By contrast, META-3 and META-4 rarely win against GARCH (20%, respectively 26,7%), which suggests that, in this configuration, these approaches do not provide robustness and consistency in financial risk estimation.

Table no.4. Win-Loss Comparison vs. Best Garch

ComparedGroup	Wins_vs_BestGARCH	Losses_vs_BestGARCH	WinRate
AI	19	11	0.6333
META-1	24	6	0.8
META-2	23	7	0.7666
META-3	6	24	0.2
META-4	8	22	0.2666

Source: Author's own calculations.

Table no.5 summarises the performance across model families (GARCH, AI, META) through the average exception rate (Avg_HitRatio) and the proportion of assets for which the model passes the tests at the chosen threshold (PassRate_Kupiec for unconditional VaR coverage, PassRate_CC for the independence/clustering of exceptions, and PassRate_ES for ES validation), complemented by the corresponding average p-values (Avg_Kupiec_p, Avg_CC_p, Avg_ES_p). The results suggest a clear advantage for the META family: Avg_HitRatio ≈ 0.0553 (close to the nominal level α = 5%) and pass rates of 1.0000 for Kupiec, CC, and ES, together with the highest average p-values (≈ 0.6784 for Kupiec, ≈ 0.5171 for CC, and ≈ 0.5979 for ES). AI also shows solid performance (Avg_HitRatio ≈ 0.0546), with pass rates of 1.0000 for Kupiec and ES and 0.9333 for CC, consistently outperforming the GARCH family. By contrast, GARCH appears more conservative in terms of the exception rate (Avg HitRatio ≈ 0.0391, below α) and shows the weakest ES validation (PassRate_ES ≈ 0.5333, Avg_ES_p ≈ 0.2450), which indicates that, although it controls VaR exceptions relatively well, it captures the severity of extreme losses less robustly than AI models and, especially, metamodels.

Table no.5. Performance Summary by Model Family

Family	PassRate_Kupiec	PassRate_CC	PassRate_ES	Avg_Kupiec_p	Avg_CC_p	Avg_ES_p
GARCH	0.9000	0.9000	0.5333	0.4160	0.3521	0.2450
AI	1.0000	0.9333	1.0000	0.5580	0.4146	0.5607
META	1.0000	1.0000	1.0000	0.6284	0.5171	0.5979

Source: Author's own calculations.

These results also have an important economic interpretation. The relatively stronger performance of the metamodels suggests that, in market risk estimation, robustness is often obtained not from relying exclusively on a single class of models, but from combining specifications that capture different features of return dynamics. In practical terms, this means that classical volatility models remain useful because they provide stability and discipline in risk estimation, while AI-based models may add value when market conditions become more complex or nonlinear. At the same time, the weaker results of some standalone AI quantile models indicate that flexibility alone is not sufficient for reliable tail-risk measurement unless it is accompanied by adequate calibration. From a risk-management perspective, these findings support a pragmatic approach in which econometric structure and data-driven adaptability are treated as complementary rather than competing tools.

Conclusions

The empirical evidence suggests that metamodelling is the most robust direction for market risk estimation in a cross-asset framework. GARCH models remain a solid and relatively stable benchmark, but they do not dominate uniformly in either VaR or ES, and in some cases, they may be excessively conservative (with exception rates below the nominal level). AI models directly calibrated on quantiles, especially the GRU and TCN variants, can match or improve the results for some assets, but overall, AI performance is more volatile and more dependent on model specification; some implementations lead to VaR exceedances that are too frequent, which is difficult to accept in a risk management application without an additional calibration control layer. The most consistent results are obtained from combination schemes that exploit the complementarity between models, rather than from the complete replacement of classical econometrics. Meta-1 (robust combinations/weighting schemes) and Meta-2 (selective overlay, with AI intervention only when there are repeated signals of improvement) frequently appear among the best options, suggesting that the stabilisation of tail forecasts through combination reduces sensitivity to model-specific errors and increases robustness during periods of stress. By contrast, the more "structural" extensions tested here—regime-based mixing (Meta-3) and the version including the exogenous VIX variable (Meta-4)—do not produce, in the current configuration, robust improvements relative to Meta-1 and Meta-2. This indicates that additional complexity does not guarantee better performance and probably requires richer state variables, a more careful definition of regimes, or stronger regularisation.

From a practical perspective, the pragmatic recommendation is that GARCH models could be used as a stable benchmark, while AI should be used selectively based on their specific characteristics. The best results, however, are obtained using metamodelling, when the purpose is to ensure consistency and stability over time. As future research directions, the study should be extended by including more exogenous variables, different market regimes and supplementary tests for ES in order to ensure a comprehensive evaluation of the efficiency of the models in predicting financial risk.

References

[1] Alexander, C. (2005) 'The Present and Future of Financial Risk Management', Journal of Financial Econometrics, 3(1), pp. 3–25. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=478802
[2] Bollerslev, T. (1986) 'Generalized Autoregressive Conditional Heteroskedasticity', Journal of Econometrics, 31(3), pp. 307–327. Available at: https://doi.org/10.1016/0304-4076(86)90063-1
[3] Christoffersen, P.F. (2003) Elements of Financial Risk Management. Elsevier Academic Press.

[4] David, L.K., Wang, J., Cisse, I.I. and Angel, V. (2024) 'Machine learning algorithms for financial risk prediction: A performance comparison', International Journal of Accounting Research, 9(2), pp. 49–55.
[5] Diebold, F.X. and Mariano, R.S. (1995) 'Comparing Predictive Accuracy', Journal of Business & Economic Statistics, 13(3), pp. 253–263. Available at: https://doi.org/10.1080/07350015.1995.10524599
[6] Engle, R.F. (1982) 'Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation', Econometrica, 50(4), pp. 987–1007. Available at: https://doi.org/10.2307/1912773
[7] Hull, J.C. (2018) Risk Management and Financial Institutions. 5th edn. Wiley.
[8] Jorion, P. (2001) Value at Risk: The New Benchmark for Managing Financial Risk. 3rd edn. McGraw-Hill.
[9] Schwaab, B., Zhang, Y. and Lucas, A. (2026) 'Joint extreme value-at-risk and expected shortfall dynamics: a score-driven peaks-over-threshold approach', ECB Working Paper Series, No. 3166.
[10] Sirignano, J. and Cont, R. (2019) 'Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning', Quantitative Finance, 19(9), pp. 1449–1459. Available at: https://doi.org/10.1080/14697688.2019.1622295
[11] Wang, Q., Wang, R. and Ziegel, J. (2025) 'E-backtesting'. Available at: University of Waterloo working paper.
[12] Wang, S., Qi, Y. and Huang, Z. (2025) 'Learning about tail risk: Machine learning and combination forecasting for value at risk and expected shortfall', Omega, 133, 103221.