SCIENTIFIC OASIS

Journal of Soft Computing and Decision Analytics

Journal homepage: www.jscda-journal.org
ISSN: 3009-3481

Mapping Hybrid and Ensemble Models for Financial Volatility Forecasting: A Bibliometric and LLM-Assisted Review

Rodrigo Baggi Prieto Alvarez[1,*], Jorge Miguel Bravo[1]

[1] NOVA Information Management School (NOVA IMS) and MagIC, NOVA University Lisbon, UNL, Portugal

ARTICLE INFO

Article history:
Received 2 May 2026
Received in revised form 24 July 2026
Accepted 9 September 2026
Available online 15 September 2026

Keywords:
Financial volatility forecasting; Hybrid models; Ensemble learning; Machine learning; Bibliometric analysis; Systematic review; Large language models

ABSTRACT

Financial volatility forecasting has undergone a rapid methodological transition from parametric econometric models towards machine and deep learning, hybrid architectures, and ensemble techniques. Despite this expansion, the literature lacks a synthesis combining bibliometric mapping with a granular methodological taxonomy of hybrid and ensemble models for volatility forecasting. This paper addresses that gap by retrieving 690 publications from *Scopus* and *Web of Science*, mapping the bibliometric landscape with `bibliometrix`, and constructing a 121-paper core corpus through multi-stage filtering on citation impact, recency, and Bradford Zone 1 sources. The corpus is classified using a ten-dimensional taxonomy covering model category, hybrid subtype, base models, combination strategy, forecast target, input features, data frequency, market and asset class, evaluation framework, and methodological novelty. To assess scalable annotation, we implement a three-model LLM-assisted pipeline using *Claude 4.6*, *Gemini 3*, and *GPT-5*, validated against a domain-expert human audit on a stratified subsample. Bibliometric results show a marked acceleration after 2020 and convergence between financial econometrics and computational predictive modelling. In the evidence reported by the primary studies, hybrid and ensemble architectures generally outperform single-model benchmarks, with sequential GARCH--DL cascades, stacking and shrinkage ensembles, and decomposition-based hybrids emerging as prominent designs; CEEMDAN and VMD provide the most transferable gains. LLM agreement is strongly dimension-specific: lexically observable dimensions (data frequency, model category, forecast target) achieve moderate agreement, whereas inferential dimensions (evaluation framework, methodological novelty) remain unreliable under abstract-only classification. These findings position LLMs as useful but bounded research assistants for systematic reviews in quantitative finance, capable of scaling first-pass classification when embedded in transparent, multi-model, and human-validated workflows.

1. Introduction

Forecasting financial volatility is one of the central problems in quantitative finance, with direct implications for risk management, derivatives pricing, portfolio allocation, and prudential regulation. Volatility is not directly observable, and this latent nature has shaped both the econometric modelling tradition and the subsequent development of realised and implied volatility measures. Since Engle [1] introduced the ARCH framework and Bollerslev [2] generalised it to GARCH, the literature

*Corresponding author.
E-mail address: rodrigo.alvarez@novaims.unl.pt

https://doi.org/10.31181/jscda41202694

has accumulated a substantial body of theoretical and empirical evidence on the dynamics of conditional variance. Subsequent contributions extended the methodological toolkit through realised volatility measures constructed from high-frequency data [3, 4] and through option-implied volatility indices that capture forward-looking market expectations [5]. Each generation of measurement paradigms, however, remained anchored in the parametric tradition of econometric specification and asymptotic inference, leaving the literature with sophisticated but increasingly constrained modelling instruments.

The past decade has witnessed a methodological transition away from this tradition. The growing availability of high-frequency and alternative data, the maturation of machine learning algorithms, and the deployment of deep learning architectures have together generated a new class of forecasting models that combines the structural insight of econometric specifications with the representational flexibility of data-driven learners. Hybrid models, which structurally combine components from different modelling families within a single forecasting pipeline, and ensemble models, which aggregate the outputs of independently trained learners, have become an increasingly central frontier of the volatility forecasting literature. Empirical evidence points to performance gains of these multi-model architectures over single-model benchmarks across asset classes and forecasting horizons [6–8]. The bibliometric evidence reported in this paper further shows that publication activity in this field has accelerated markedly since 2020.

Despite this expansion, the literature lacks a synthesis that combines bibliometric mapping with a granular methodological taxonomy explicitly targeting hybrid and ensemble architectures for financial volatility forecasting. Existing systematic reviews of volatility forecasting do not restrict their focus to this model family, while reviews of hybrid or ensemble approaches in financial prediction generally target stock prices or returns rather than volatility, which entails different model requirements, evaluation metrics, and economic implications. At the same time, no review in the financial econometrics domain has examined large language models as structured annotation instruments for methodological classification, despite emerging evidence from adjacent fields that LLMs can support large-scale taxonomic coding while also exhibiting model-specific biases and boundary sensitivities [9, 10].

This paper addresses these gaps through three interrelated contributions. Conceptually, it provides, to the best of our knowledge, the first review dedicated to hybrid and ensemble models for financial volatility forecasting, organised around a ten-dimensional taxonomy: Model Category, Hybrid Sub-Type, Base Models, Combination Strategy, Forecast Target, Input Features, Data Frequency, Market/Asset Class, Evaluation Framework, and Methodological Novelty. Methodologically, it advances a replicable review protocol that integrates bibliometric mapping of 690 publications with LLM-assisted classification of a 121-paper core corpus through a three-model pipeline (*Claude 4.6*, *Gemini 3*, and *GPT-5*), followed by a domain-expert human audit on a stratified subsample. Empirically, it provides a comparative assessment of LLM classification agreement, identifying where the three models converge and diverge in their interpretation of the taxonomy and where abstract-only classification imposes reliability limits that cannot be resolved by prompt refinement alone.

The study therefore makes a dual contribution. For the volatility forecasting literature, it maps the methodological evolution from single-model econometric specifications towards hybrid, ensemble, and decomposition-based architectures. For the methodology of systematic reviews in quantitative finance, it provides evidence on the conditions under which LLM-assisted annotation can be used reliably, showing that dimensions directly signalled by standardised abstract vocabulary are more amenable to automation than dimensions requiring full-paper interpretation, such as evaluation framework and methodological novelty.

The remainder of the paper is organised as follows. Section 2 reviews the literature on volatility forecasting and on LLM-assisted classification in systematic reviews. Section 3 sets out the theoretical foundations of volatility measurement and the taxonomy of hybrid and ensemble architectures. Section 4 describes the data sources, sampling procedure, classification pipeline, and human validation protocol. Section 5 reports the bibliometric findings. Section 6 presents the methodological classification results and the inter-annotator agreement analysis. Section 7 discusses the implications for future systematic reviews in quantitative finance, and Section 8 concludes.

2. Relevant Literature

2.1 Reviews on Volatility Prediction and Hybrid and Ensemble Methods

The literature on financial volatility forecasting has expanded substantially over the past decade, prompting a growing number of systematic and bibliometric reviews. These reviews provide important coverage of econometric, machine learning, and deep learning approaches, but they do not yet

offer a synthesis centred on hybrid and ensemble architectures as a distinct methodological family within volatility forecasting.

The most closely related volatility-specific reviews focus primarily on machine learning and deep learning methods. Ge et al. [11] provide one of the most methodologically rigorous reviews in this domain, examining 35 post-2015 studies on neural network-based volatility forecasting and identifying a gap between modern machine learning architectures and those empirically applied to volatility prediction. Their proposal for a shared benchmarking task is particularly relevant for improving comparability across studies. However, the scope of the review is restricted to neural networks and therefore excludes many of the hybrid econometric–ML and ensemble configurations that have become central to the field since 2020. A complementary perspective is offered by Gunnarsson et al. [6], who review AI and ML methods for realised and implied volatility prediction. Their synthesis finds that memory-based architectures, particularly LSTM and GRU networks, rank among the best-performing models, while traditional econometric models remain highly competitive benchmarks. The authors identify the combination of econometric and ML approaches as a promising but still underexplored research area, thereby motivating the present focus on hybrid and ensemble architectures. More recently, Mansilla-Lopez et al. [12] conducted a PRISMA-compliant systematic review of ML-based stock volatility forecasting, identifying 27 prediction models and 15 volatility-influencing factors. Although their review confirms the increasing prominence of hybrid models, it does not combine bibliometric mapping with a granular architectural taxonomy.

A second stream of reviews examines deep learning in financial forecasting more broadly. Sezer et al. [13] survey deep learning methods applied to financial time-series forecasting between 2005 and 2019, covering CNN, DBN, and LSTM architectures across index, foreign exchange, and commodity applications. Their review establishes an important DL-in-finance taxonomy, but it covers a pre-2020 period and does not focus specifically on volatility or hybrid architectures. Similarly, Hu et al. [14] review DL methods for stock and foreign exchange prediction, classifying papers by model type, including CNN, LSTM, DNN, RNN, reinforcement learning, and hybrid attention mechanisms. They show that models combining LSTM with other architectures have become an important trajectory of recent research. This taxonomy is useful for situating hybrid models, but the review concerns return and price prediction rather than volatility forecasting.

A third stream addresses hybrid and ensemble approaches in financial prediction, but generally outside the volatility forecasting domain. Shah et al. [15] review combinations of ARIMA, LSTM, CNN, and CNN–LSTM architectures for stock price prediction, discussing the accuracy and limitations of each configuration. However, their focus is on price-level forecasting rather than volatility, and the study does not include a bibliometric or taxonomic classification component. Nosratabadi et al. [7] provide a broader PRISMA-based synthesis of deep learning, hybrid deep learning, hybrid ML, and ensemble models across economic applications, finding that hybrid models generally outperform their single-model counterparts. This result directly supports the rationale for studying hybrid and ensemble architectures as a coherent model family. In a related contribution, Sonkavde et al. [16] provide a systematic review and performance analysis of ML and DL models for stock price forecasting, with particular attention to ensemble methods. Their practical demonstration illustrates the type of hybrid architecture that the present study systematises taxonomically, although again in the context of price forecasting rather than volatility.

Bibliometric reviews provide a fourth relevant stream. Ferro et al. [17] combine systematic review and bibliometric analysis to examine ensemble ML approaches for stock market index prediction. Their work is methodologically close to the present study, but it targets stock index prediction rather than volatility forecasting and does not address hybrid econometric–ML architectures such as the GARCH–DL family. Raimundo and J. M. Bravo [18] examine the methodological convergence of forecasting and portfolio theory through a bibliometric and thematic analysis. They employ a hybrid framework that integrates co-occurrence analysis, bibliographic coupling, fractional authorship attribution, and machine learning-based semantic clustering. The study concludes by outlining opportunities for cross-disciplinary integration and proposing targeted research directions to enhance robustness, adaptability, and interpretability in forecasting and financial decision-making. In a similar vein, Vuong et al. [19] and Ahmed et al. [20] provide bibliometric overviews of AI and ML in finance, mapping influential authors, journals, and research clusters. However, their focus on price rather than volatility forecasting, and their absence of a formal model taxonomy, distinguish them from the present review. Goyal and Soni [21] apply a comprehensive bibliometric framework, including Bradford's Law, Lotka's Law, keyword co-occurrence, thematic mapping, and bibliographic coupling, to analyse 1,283 articles on stock market volatility during crisis periods. Their methodological apparatus

closely parallels the bibliometric component of the present study, while their crisis-period focus provides useful context for interpreting the post-2020 publication surge documented here. Bashir [22] similarly use the `bibliometrix` R package to analyse 684 studies on oil price shocks, stock market returns, and volatility spillovers, identifying research themes, influential institutions, and impediments in the existing literature.

Asset-specific bibliometric studies further illustrate the value of combining quantitative mapping with qualitative synthesis. Almeida and Gonçalves [23] conduct both a bibliometric analysis and a qualitative literature review of volatility and risk management in cryptocurrency investment, covering the period from 2009 to 2021. Their dual-method design is conceptually analogous to the present study, although it is applied to a narrower asset class. A similarly structured study by Pečiulis et al. [24] analyses cryptocurrency forecasting trends using bibliometric techniques following PRISMA guidelines, identifying key research themes and trajectories. More broadly, Manogna and Anand [25] conduct a bibliometric analysis of deep learning applications in finance using 693 *Scopus* articles from 2000 to 2022, identifying four broad research areas and noting the acceleration of DL applications in finance since 2017.

Taken together, these reviews delineate an active frontier of systematic synthesis in volatility forecasting and financial prediction. However, they leave a clear gap. Volatility-specific reviews do not focus on hybrid and ensemble architectures as a coherent model family; hybrid and ensemble reviews generally focus on prices or returns rather than volatility; and bibliometric studies do not provide a granular methodological classification of hybrid and ensemble volatility forecasting models. This study addresses that missing intersection by combining bibliometric mapping with a structured architectural taxonomy of hybrid and ensemble models for financial volatility forecasting.

2.2 LLM-Assisted Classification in Methodological Reviews

The integration of large language models into the systematic review pipeline is an emerging methodological area. Although most applications have appeared in biomedical and social science contexts, the underlying principles are directly relevant to financial econometrics: large literatures must be screened, classified, and mapped against conceptual taxonomies, often under constraints of scale, ambiguity, and limited human annotation capacity.

The earliest and most developed use case concerns title and abstract screening. Dennstädt et al. [9] evaluate the use of LLMs for title and abstract screening in systematic reviews, finding sensitivity rates of up to 97.6% across different models on published biomedical datasets. Their study shows that LLMs can serve as efficient pre-screeners and reduce manual workload, while also demonstrating that performance varies substantially across models and datasets. This underscores the importance of model selection, prompt design, and domain-specific calibration. Extending this line of inquiry, Delgado-Chaves et al. [10] evaluate 18 LLMs for systematic review screening across three separate reviews and find that LLMs can reduce a single reviewer's workload by between 33 and 93 percent during title and abstract screening. Their results also show that performance is shaped by the precision and operationality of inclusion and exclusion criteria in the prompt.

A second use case moves beyond screening towards structured extraction and multi-dimensional classification. Ye et al. [26] propose a hybrid semi-automated workflow for systematic reviews that integrates *Gemini-Pro* for the structured extraction of identifier, verifier, and data-field categories from formatted documents. Their approach improves classification accuracy relative to human-only processes in a case study and provides a methodological template analogous to the multi-stage classification pipeline adopted here. However, the application domain differs from financial econometrics, and the classification task does not involve the same type of architectural and methodological boundary decisions that arise in hybrid and ensemble volatility forecasting.

A third stream evaluates the broader suitability of LLMs for taxonomic classification. Ziems et al. [27] provide a systematic evaluation of 13 LLMs on 25 computational social science benchmarks, finding that LLMs do not consistently outperform fine-tuned models on taxonomic labelling, although they achieve fair levels of agreement with human annotators and perform particularly well on free-form explanation tasks. This result supports the use of LLMs as complements to, rather than substitutes for, human classification. A broader framework for integrating machine annotation into qualitative social science research is offered by Karjus [28], who argues for workflows that combine human expertise with machine scalability and explicitly considers how annotator error rates can be incorporated into subsequent statistical inference. This perspective is directly relevant to the present study, which treats LLM outputs as structured annotations requiring validation rather than as definitive labels.

The existing LLM-assisted review literature therefore provides both an opportunity and a warning. On the one hand, LLMs can scale annotation, standardise coding rules, and reduce the burden of manual classification. On the other hand, their reliability is likely to vary across dimensions, especially when labels require inferential judgements that are not directly signalled in titles, abstracts, or keywords. This is particularly important in financial econometrics, where the distinction between hybrid and ensemble models, or between statistical and economic evaluation frameworks, may depend on methodological details reported outside the abstract. No existing review in this domain has evaluated LLMs as structured annotation instruments for methodological classification. The present study therefore extends the LLM-assisted review literature by applying a three-model annotation pipeline to a specialised financial econometrics corpus and validating its outputs against a domain-expert human audit.

3. Theoretical Foundations

3.1 Financial Volatility Measures

Volatility, understood as the time-varying dispersion of financial asset returns, is not directly observable. Its latent nature has shaped both the measurement problem and the architecture of the forecasting literature. Three broad families of volatility measures dominate academic and practitioner research: conditional volatility, realised volatility, and implied volatility. Each reflects a distinct solution to the problem of measuring an unobserved risk process, and each implies different benchmark models, evaluation metrics, and forecasting objectives.

The first and most established family is conditional volatility, rooted in the ARCH model of Engle [1] and its GARCH generalisation proposed by Bollerslev [2]. ARCH-class models treat conditional variance as a latent process evolving according to a specified dynamic equation. This framework has since been extended to accommodate leverage effects (EGARCH, GJR-GARCH), long memory (FIGARCH), and time-varying higher moments. As Bollerslev et al. [29] document, ARCH-class models remain the dominant econometric paradigm for volatility modelling and constitute the primary benchmark against which many subsequent approaches are evaluated. Their central limitation, formalised by Hansen and Lunde [30], is that model comparisons based on noisy proxies for a latent variable can generate systematically inconsistent rankings.

Realised volatility (RV) addresses this limitation by constructing an ex-post observable measure of daily variance from high-frequency intraday returns. Andersen et al. [3] provide the foundational theoretical framework, showing that, under standard continuous-time assumptions, RV converges to integrated variance as sampling frequency increases. In this sense, realised volatility transforms the latent volatility process into an observable empirical target. The most influential reduced-form model for RV forecasting is the Heterogeneous Autoregressive model of Corsi [4], which aggregates daily, weekly, and monthly realised volatility into a parsimonious additive specification designed to reproduce the long-memory dynamics observed in RV series. The HAR model has become the *de facto* econometric benchmark in the realised volatility literature, while recent evidence shows that machine learning algorithms, including regularised regression and neural networks, can compete with and frequently outperform HAR-type specifications [31].

The third family, implied volatility (IV), is forward-looking rather than backward-looking. It is the volatility level that, when inserted into an option pricing formula, equates the model-implied option price to the observed market price. The CBOE VIX index, constructed from S&P 500 options across a range of strikes and maturities, is the most widely cited implied volatility measure and is commonly interpreted as a market-based forecast of near-term equity risk [5, 32]. Unlike realised or conditional volatility, implied volatility incorporates market participants' expectations and risk premia. It is therefore informationally richer, but also more complex to model, because implied volatility surfaces display systematic smile, skew, and term-structure effects that vary across regimes.

These three measurement paradigms structure the forecasting literature. Conditional volatility is most closely associated with parametric econometric models; realised volatility has generated HAR-type benchmarks and high-frequency forecasting designs; and implied volatility links volatility forecasting to derivatives markets and risk-neutral expectations. The recent rise of machine learning and deep learning does not replace these paradigms. Rather, it increasingly combines them, producing hybrid and ensemble architectures that seek to retain the theoretical structure of econometric models while exploiting the representational flexibility of data-driven learners.

3.2 Hybrid and Ensemble Models: Taxonomy and Classification Structure

A clear terminological foundation is necessary before a taxonomy can be applied consistently. The distinction between hybrid and ensemble models is central to this study because it defines the primary architectural boundary used in the classification exercise. A hybrid model structurally integrates components from different modelling families within a single forecasting pipeline, such that the components interact during training or inference, share representations, pass residuals, or jointly optimise a loss function. An ensemble model, by contrast, aggregates the outputs of independently trained models through a fixed or learned combination rule, without altering the internal training dynamics of the constituent models [33].

The distinction is not merely semantic. Hybrid models modify the information flow between components and can exploit structural complementarities among modelling families. Ensemble models are more modular: they exploit diversity across independently trained learners and reduce the risk that the forecast is dominated by the failure of a single specification. Hybrids therefore operate through integration, whereas ensembles operate through aggregation. This distinction provides the main theoretical basis for the taxonomy used in this review.

The broader architecture of the taxonomy first separates Single Model from Multi-Model systems before subdividing the latter into Hybrid and Ensemble branches. This structure follows from the foundational forecast-combination literature. Bates and Granger [34] showed that a weighted combination of two forecasts with imperfectly correlated errors can be superior in mean-squared-error terms to either individual forecast, provided that neither forecast is already optimal. Subsequent forecast-combination research [35, 36] has confirmed that diversification across model families can reduce systematic forecast error. This principle applies to both ensembles and hybrids, although through different mechanisms: ensembles diversify at the output level, whereas hybrids embed complementarity within the forecasting architecture itself.

Within the Multi-Model branch, the Hybrid/Ensemble distinction maps onto two different theories of complementarity. Ensembles exploit output-level diversity: independently trained models capture different features of the data-generating process, and their outputs are combined to cancel idiosyncratic errors. Hybrids exploit structural complementarity: different modelling families are linked so that the strengths of one component compensate directly for the limitations of another. A GARCH model, for example, is well specified for conditional heteroskedasticity but has limited capacity to capture complex non-linear interactions among predictors. An LSTM can capture such interactions but lacks the explicit variance-dynamics structure of GARCH. A GARCH–LSTM hybrid can therefore exploit the strengths of both approaches by allowing the econometric component to transform or filter the input signal before the neural component models residual non-linearities.

Within hybrid models, five architectural configurations are distinguished (Figure 1).

Sequential hybrids apply models in a cascade, whereby the output, residuals, or filtered series generated by one model become the direct input to a subsequent model. The canonical volatility-forecasting example is a GARCH–LSTM architecture in which a GARCH model first filters conditional heteroskedasticity and the LSTM then models remaining non-linear structure. The theoretical motivation is residual learning: once the component that a parametric model explains efficiently has been removed, the second-stage learner faces a lower-dimensional and more regular prediction problem.

Parallel hybrids train multiple components on the same input and fuse their outputs within a common architecture. Unlike sequential hybrids, where information flows from one component to the next, parallel hybrids preserve the separate inductive biases of each component before combining them through a fusion layer. This configuration lies close to the ensemble paradigm, and the operational boundary depends on whether the components are trained jointly as part of one integrated architecture or independently before ex-post aggregation.

Decomposition-based hybrids introduce a signal-preprocessing stage before prediction. Methods such as empirical mode decomposition (EMD), complete ensemble EMD with adaptive noise (CEEMDAN), variational mode decomposition (VMD), and discrete wavelet transforms decompose the target series into components associated with different frequencies or scales. Predictive models are then applied to each component before the forecasts are reconstructed. The theoretical motivation is frequency-isolation: volatility series are non-stationary and often combine dynamics operating at different temporal scales, so modelling decomposed components may produce more parsimonious and stable forecasts than modelling the composite series directly [37].

Feature-based hybrids use one model primarily as a representation or feature extractor, with the learned features serving as inputs to a second predictive model. The canonical example is a CNN–LSTM architecture, where convolutional layers extract local temporal patterns and the LSTM captures

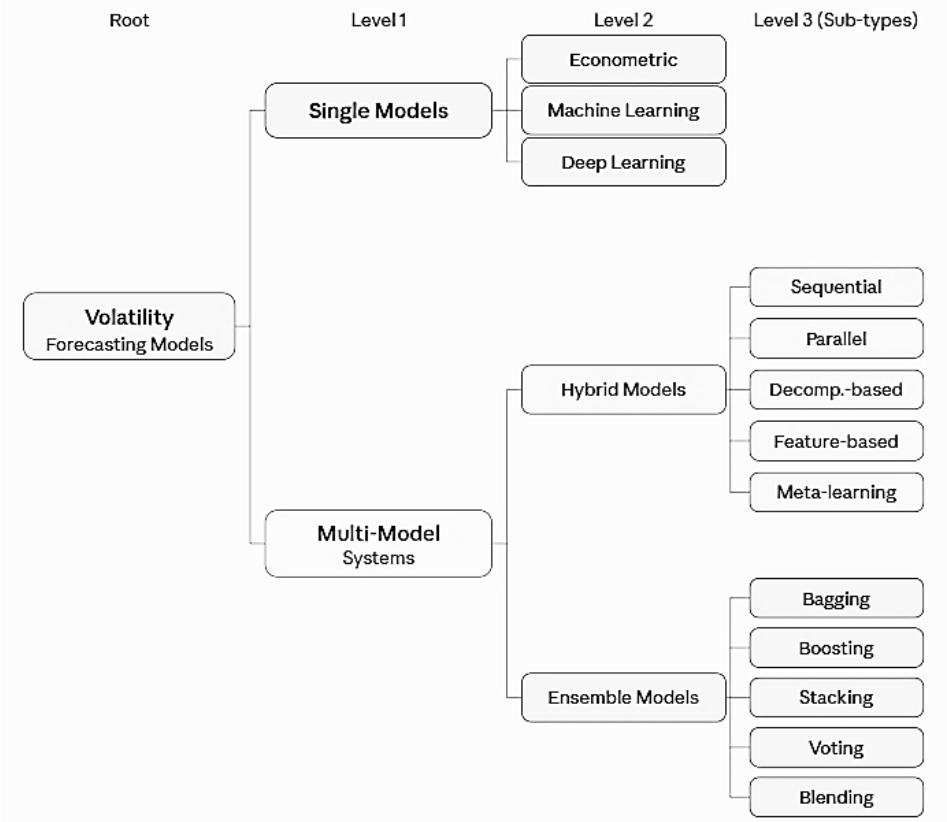

Fig. 1. Hybrid and Ensemble Models Taxonomy: Model Categories and Sub-Types

sequential dependencies. The theoretical basis is the transfer of representational capacity: the first model transforms the raw input space into a more informative feature space, reducing the burden placed on the second-stage predictor.

Meta-learning hybrids use a second-stage learner to estimate how base-model outputs should be combined, possibly in a way that varies across market regimes. This configuration extends stacking by allowing the weighting function itself to be learned from data. It is motivated by the instability of fixed combination weights in financial time series, where structural breaks and regime shifts can alter the relative informativeness of different forecasting models [35].

Within ensemble models, the taxonomy distinguishes five aggregation strategies. *Bagging* averages forecasts from models trained on bootstrapped samples, reducing variance through decorrelation. *Boosting* trains base learners sequentially, with each learner placing greater weight on the errors of its predecessors. *Stacking* trains a second-level meta-learner to combine base-learner outputs and is particularly relevant when the relative performance of constituent models varies across regimes. *Voting* assigns the final prediction by majority or plurality rule and is most applicable to classification or directional forecasting tasks. *Blending* is a held-out variant of stacking in which the meta-learner is trained on a validation partition rather than on out-of-fold predictions from the full training sample.

Although the Hybrid/Ensemble distinction is theoretically useful, the boundary between the two families is better understood as a continuum than as a perfectly discrete dichotomy. The forecast-combination literature, surveyed by Wang et al. [38], documents a proliferation of intermediate architectures, including stacking with non-linear meta-learners, decomposition-ensemble pipelines, and the hybridisation of pre-existing hybrid models [39]. Similarly, Cawood and Zyl [40] show that hybrid base learners can themselves be embedded within ensemble fusion frameworks, producing architectures that are hybrid in construction but ensemble-like in deployment. The taxonomy adopted in this study therefore preserves the Hybrid/Ensemble distinction as the primary classificatory axis, while recognising that some empirical papers occupy borderline positions. This is precisely why the classification protocol treats ambiguous cases as an empirical object of analysis rather than as simple coding noise.

4. Methodology

4.1 Data Sources and Search Strategy

Data were retrieved from two major bibliographic databases: *Scopus* and *Web of Science*, selected for their broad, high-quality, and complementary coverage of scientific literature [41]. The search was designed to identify studies at the intersection of financial volatility forecasting and hybrid or ensemble modelling. The final query was executed on 22 March 2026. The full Boolean search string is reported in Appendix A1. The search strategy combined three thematic pillars:

(i) *Volatility Forecasting or Prediction*
(ii) *Financial Markets or Financial Variables*
(iii) *Hybrid or Ensemble Models*

Records were screened on the basis of titles, abstracts, and keywords. The inclusion criteria were defined by topic relevance, publication type, and the selected time frame. All records were downloaded in *BibTeX* format and imported into *R* for processing. Duplicates across *Scopus* and *Web of Science* were removed using the `bibliometrix` package [42, 43]. The `biblioshiny` graphical interface was used for descriptive bibliometric analysis and science mapping, while all figures and maps were generated within *RStudio*.

This procedure yielded the full bibliometric sample used in the first stage of the analysis. The sample provides the basis for descriptive statistics, source analysis, author productivity, keyword co-occurrence, trend-topic analysis, and thematic mapping.

4.2 Core Sample Construction

Following the bibliometric analysis, a multi-stage core sampling procedure was implemented to derive a representative and analytically tractable subset for in-depth qualitative and methodological classification. The script used to construct the core sample is reported in Appendix A2. This approach is consistent with recent practice in hybrid systematic–bibliometric reviews, where the objective is to balance citation impact, recency, and source concentration [44, 45].

The first stage applied an *impact-based filter*, retaining the top 20% of documents according to total citations per year. Citation normalisation by publication age was used to reduce the temporal advantage of older papers in cumulative citation counts [46]. The 20% cut-off was selected to isolate the high-impact tail of the literature while retaining a candidate pool large enough to populate all ten taxonomic dimensions and small enough to permit classification of every retained abstract by three language models and, for the audited subsample, by a human expert. Stricter and more permissive thresholds were considered. A top 10% cut-off would have reduced the candidate pool below the level required to represent the full label space of the taxonomy, particularly for the less frequent hybrid subtypes and aggregation strategies, whereas a top 30% cut-off would have weakened the impact criterion without materially increasing architectural diversity, since the additional documents remained subject to the Bradford Zone 1 restriction applied in the second stage. To ensure that emerging methods were not excluded solely because they had not yet accumulated citations, all publications from 2025 were also retained irrespective of citation count. The impact and recency filters therefore operated as alternative inclusion routes: a document entered the candidate pool if it was either highly cited on an age-adjusted basis or published in 2025.

The second stage applied a *source-based refinement* using Bradford's Law of Scattering*. Documents were retained only if they appeared in Bradford Zone 1 sources. These included leading outlets such as *Expert Systems with Applications*, *Journal of Forecasting*, *Computational Economics*, *IEEE Access*, *International Journal of Forecasting*, and *Journal of Econometrics*, among others. Operationally, the final core sample can be expressed as:

$$\textit{Core Sample} = (\textit{Top 20\% by annual citations} \cup \textit{2025 publications}) \cap \textit{Bradford Zone 1 sources}.$$

This procedure produced a final core sample of 121 papers. The sample therefore privileges documents that are either influential or recent, while also ensuring that the retained studies are concentrated in the most productive sources in the field. The resulting corpus forms the basis for the qualitative synthesis and LLM-assisted methodological classification.

*Bradford's Law states that, for a given topic, the literature can be divided into zones according to source productivity. Zone 1 consists of a small number of sources with the highest number of publications, Zone 2 contains an intermediate set of sources, and Zone 3 consists of a larger number of sources with relatively few publications. It is commonly used to identify the core journals of a research field.

4.3 LLM-Assisted Classification

The increasing heterogeneity of volatility forecasting models poses substantial challenges for manual classification. Studies frequently combine econometric, machine learning, and deep learning techniques in non-standardised ways, making purely rule-based categorisation inefficient and potentially inconsistent. Following the construction of the core sample, this study therefore implements a hybrid human–AI classification framework to classify the 121 papers according to the taxonomy developed in Section 3. The full prompt is reported in Appendix A3.

The LLM-assisted classification was conducted using three frontier general-purpose models: *Claude 4.6*, *Gemini 3*, and *GPT-5*. The use of three models reflects three methodological considerations. First, the models are developed by different providers and therefore reduce the risk that the classification results are driven by a single model family or alignment protocol. Second, comparative evidence from scientific classification tasks suggests that proprietary frontier models continue to perform strongly on structured extraction and taxonomic labelling tasks [47–49]. Third, all three models support long-context reasoning at the scale required to process the full core corpus within a single classification workflow. The use of closed-source models nevertheless imposes a reproducibility cost, because provider-side updates may alter model behaviour over time. For this reason, the execution window and prompt specification are documented in Appendix A3.

Each model received the same input fields for each paper: title, abstract, author keywords, source, and bibliographic metadata where available. The models were instructed to classify each paper according to ten dimensions:

1. Model Category;
2. Hybrid Sub-Type;
3. Base Models;
4. Combination Strategy;
5. Forecast Target;
6. Input Features;
7. Data Frequency;
8. Market/Asset Class;
9. Evaluation Framework;
10. Methodological Novelty.

This ten-dimensional structure captures not only the architectural typology of each model but also the forecasting object, information set, empirical context, and evaluation design. Dimensions 1 and 2 form the architectural core of the taxonomy. Dimension 3 records the constituent modelling components as free-text labels, while Dimension 4 captures the meta-level aggregation mechanism applied to model outputs. Dimensions 5 to 8 describe the empirical forecasting design: target, features, frequency, and market or asset class. Dimensions 9 and 10 capture the quality and contribution profile of each study through its evaluation framework and stated methodological novelty.

The classification output was standardised before analysis. Controlled vocabulary dimensions were normalised to remove minor formatting differences, while free-text dimensions were retained for qualitative audit rather than formal kappa computation. For each controlled dimension, the three LLM labels were compared pairwise using Cohen's kappa and jointly using Fleiss' kappa. A majority-vote rule was then used to derive the adjudicated label whenever at least two of the three models agreed. Cases in which all three models assigned different labels were flagged for manual adjudication. This design allows the pipeline to exploit the scalability of LLM annotation while preserving human oversight for dimensions and cases characterised by high classificatory ambiguity.

4.4 Human Validation Protocol

The inter-model agreement statistics establish the internal consistency of the LLM-assisted pipeline, but consistency among LLMs is not sufficient to establish classification quality. To evaluate whether the adjudicated classifications align with domain-expert interpretation, a structured human validation exercise was conducted. The human classifier acted as a reference annotator rather than as an infallible ground truth, and divergences between the human classifier and the LLMs were interpreted as evidence of possible taxonomic ambiguity, information constraints, or model-specific classification tendencies.

A stratified random sample of 25 papers, corresponding to 21% of the 121-paper core corpus, was drawn using a fixed random seed to ensure reproducibility. Stratification was performed by publication year, with three strata: pre-2022 (9 papers), 2022–2024 (7 papers), and 2025 (9 papers). This design ensures representation across the foundational period, the consolidation phase, and the most recent

wave of publications identified in the bibliometric analysis. The sampled papers span 14 journals and cover all Model Category labels present in the full corpus.

The human classifier was presented with the title, abstract, and author keywords of each paper, that is, the same informational input available to the LLMs. The full paper text was not consulted. This constraint preserves comparability between human and LLM classification conditions and isolates the abstract-level information signal used by the pipeline. Human classifications were recorded independently before the LLM outputs were consulted.

Each paper was classified on the eight computable taxonomic dimensions: Model Category, Hybrid Sub-Type, Combination Strategy, Forecast Target, Input Features, Data Frequency, Evaluation Framework, and Methodological Novelty. The two free-text dimensions, Base Models and Market/Asset Class, were recorded for audit purposes but excluded from the kappa analysis because their open-ended format makes exact string-matching unsuitable for chance-corrected agreement statistics.

Agreement between the human classifier and each LLM was quantified using Cohen's kappa [50]. A further pairwise kappa was computed between the human labels and the adjudicated majority-vote labels produced by the three-model pipeline. To assess global annotation consistency, Fleiss' kappa [51] was computed for each dimension across the four annotators jointly: *Human*, *Claude 4.6*, *Gemini 3*, and *GPT-5*. Kappa values are interpreted following the scale proposed by Landis and Koch [52].

For the Evaluation Framework dimension, agreement was computed on normalised multi-label strings. Where a paper used multiple evaluation approaches, the labels were sorted alphabetically and concatenated before comparison. This preserves the multi-label nature of the dimension while allowing consistent kappa computation under an exact-match convention.

Given the sample size of 25 papers, the validation statistics are reported as point estimates and interpreted primarily through their directional pattern across dimensions. The purpose of the human audit is not to establish definitive population-level reliability bounds, but to identify which dimensions are robust to abstract-only classification, which dimensions require manual adjudication, and which dimensions may need full-paper information to be classified reliably.

5. Results: Bibliometric Landscape

The bibliometric analysis is based on documents retrieved from *Scopus* (599 records) and *Web of Science* (420 records). Deduplication using the `bibliometrix` package in *R* identified and removed 329 overlapping records, yielding a final bibliometric sample of 690 unique publications (Figure 2). This sample constitutes the basis for the descriptive, source, authorship, citation, and keyword analyses reported in this section.

The reduction from the full bibliometric sample to the 121-paper core sample used for methodological classification is described in Section 4. In brief, the core corpus was constructed by retaining papers that were either in the top 20% by annual citations or published in 2025, and then restricting this set to Bradford Zone 1 sources. The present section therefore focuses on the structure and evolution of the broader 690-publication landscape.

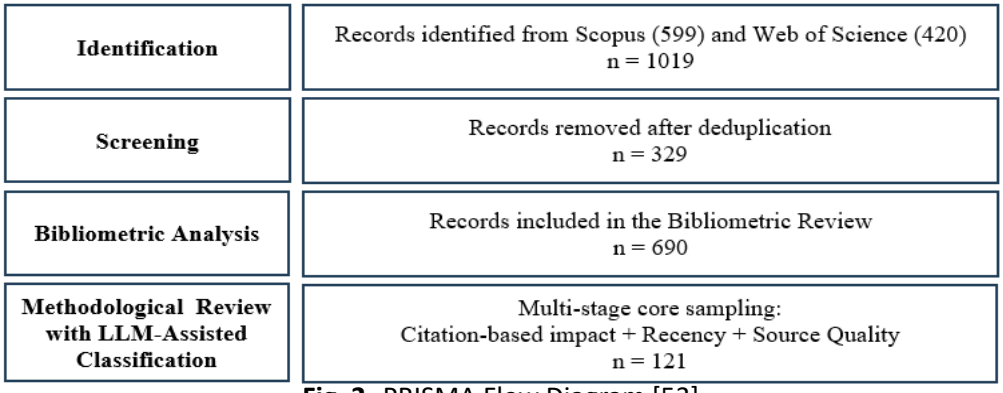

Fig. 2. PRISMA Flow Diagram [53]

Table 1 summarises the main descriptive statistics. The 690 documents span the period from 1997 to 2025 and are distributed across 413 sources, including journals, books, and conference proceedings. The annual growth rate of scientific production is 20.35%, and the average document age is 5.26 years, indicating that the literature is both recent and rapidly expanding. Journal articles dominate the sample, accounting for 494 documents, followed by conference papers and proceedings papers.

The dataset includes 1,708 authors and 2,184 author appearances. Only 90 documents are single-authored, while the average number of co-authors per document is 3.17 and the international co-authorship rate is 17.39%. These figures point to a collaborative but still moderately internationalised research field.

Table 1
Descriptive Statistics of the Bibliometric Analysis

Description	Results	Description	Results
MAIN INFORMATION ABOUT DATA		**DOCUMENT TYPES**	
Timespan	1997–2025	Article	494
Sources (Journals, Books, etc.)	413	Editorial material	1
Documents	690	Article; proceedings paper	5
Annual growth rate (%)	20.35	Article; retracted publication	2
Document average age	5.26	Book	1
Average citations per document	18.79	Book chapter	7
Average citations per year per document	2.46	Conference paper	114
DOCUMENT CONTENTS		Conference review	10
Keywords Plus (ID)	1654	Erratum	3
Author's keywords (DE)	1839	Proceedings paper	42
AUTHORS		Retracted	1
Authors	1708	Review	10
Author appearances	2184	**COUNTRIES – PUBLICATION RANKING**	
Authors of single-authored documents	71	China	183
AUTHORS COLLABORATION		India	75
Single-authored documents	90	USA	35
Documents per author	0.40	UK	34
Co-authors per document	3.17	Iran	18
International co-authorships (%)	17.39	Brazil	14

In `bibliometrix`, the "DE" field represents author-assigned keywords, while the "ID" field represents keywords indexed by the database, such as *WoS* or *Scopus*. *WoS* and *Scopus* use different indexing systems, which may lead to differences in the keywords assigned to the same article.

The temporal evolution of publications reveals a marked acceleration from 2021 onwards, culminating in 179 publications in 2025 (Figure 3). This surge is consistent with the rapid diffusion of machine learning, deep learning, and hybrid forecasting approaches in financial volatility research. From a citation perspective, the documents in the sample receive an average of 18.79 citations, corresponding to 2.46 citations per document per year. The peaks in average article citations observed in 2010 and 2017 appear to be associated with highly influential contributions, although this interpretation should be treated cautiously because annual citation averages are sensitive to a small number of highly cited papers.

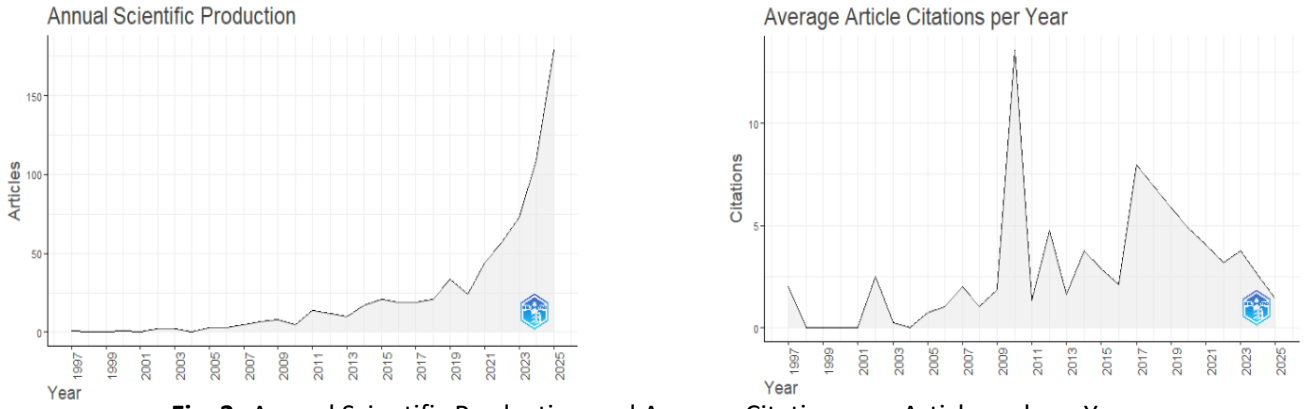

Fig. 3. Annual Scientific Production and Average Citations per Article and per Year

The source distribution is consistent with Bradford's Law of Scattering (Table 2). The 27 sources in Bradford's Zone 1 account for 230 documents (33.3% of the bibliometric sample). *Expert Systems with*

Applications is the leading source, with 27 articles, followed by *Journal of Forecasting* and *Computational Economics*. The presence of outlets such as *Energy Economics*, *International Journal of Forecasting*, and *Applied Soft Computing* indicates that the field is positioned at the intersection of financial econometrics, computational intelligence, and applied energy and commodity-market research.

Table 2
Core Journals from Bradford's Zone 1

	Source	Freq	CumFreq	CumPerc
1	EXPERT SYSTEMS WITH APPLICATIONS	27	27	3.9%
2	JOURNAL OF FORECASTING	22	49	7.1%
3	COMPUTATIONAL ECONOMICS	21	70	10.1%
4	ENERGY ECONOMICS	11	81	11.7%
5	IEEE ACCESS	11	92	13.3%
6	INTERNATIONAL REVIEW OF ECONOMICS & FINANCE	11	103	14.9%
7	ACM INTL CONFERENCE PROCEEDING SERIES	10	113	16.4%
8	INTERNATIONAL JOURNAL OF FORECASTING	9	122	17.7%
9	INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS	9	131	19.0%
10	APPLIED SOFT COMPUTING	8	139	20.1%
11	MATHEMATICS	8	147	21.3%
12	NORTH AMERICAN J OF ECONOMICS AND FINANCE	8	155	22.5%
13	RESOURCES POLICY	8	163	23.6%
14	FINANCE RESEARCH LETTERS	7	170	24.6%
15	JOURNAL OF EMPIRICAL FINANCE	6	176	25.5%
16	JOURNAL OF RISK AND FINANCIAL MANAGEMENT	6	182	26.4%
17	RISKS	6	188	27.2%
18	LECTURE NOTES IN COMPUTER SCIENCE	5	193	28.0%
19	PHYSICA A STAT MECHANICS AND ITS APPLICATIONS	5	198	28.7%
20	APPLIED SCIENCES-BASEL	4	202	29.3%
21	ECONOMIC MODELLING	4	206	29.9%
22	ENERGY	4	210	30.4%
23	ENG APPLICATIONS OF ARTIFICIAL INTELLIGENCE	4	214	31.0%
24	ENTROPY	4	218	31.6%
25	INTL J OF ADVANCED COMP SCIENCE AND APPLIC	4	222	32.2%
26	JOURNAL OF ECONOMETRICS	4	226	32.8%
27	J OF INTL FIN MARKETS INSTITUTIONS & MONEY	4	230	33.3%

The author-level indicators reveal a highly unequal distribution of scientific productivity. The most prolific contributors include *Zhang Y*, *Li Y*, *Wang Y*, and *Na N*, with Chinese- and US-affiliated authors occupying prominent positions in absolute publication counts (Table 3). Fractionalised authorship produces a similar ranking, suggesting that these authors' prominence is not driven solely by participation in large co-authored teams. Country-level patterns confirm the centrality of China in publication volume, followed by India, the USA, and the UK. Citation impact, however, is more concentrated among US-affiliated authors and papers, indicating a distinction between productivity leadership and citation leadership (Table 4).

At the journal level, *Expert Systems with Applications* is the most frequent outlet, contributing 27 articles, or 3.9% of the full sample (Table 5). Other important sources include *Journal of Forecasting*, *Computational Economics*, *Energy Economics*, and *IEEE Access*. The most cited papers also point to the importance of finance, econometrics, statistics, and computational intelligence outlets, including *Review of Financial Studies*, *Journal of Financial Economics*, *Journal of Applied Econometrics*, and *Computational Statistics & Data Analysis* (Table 6). This confirms that the literature is not confined to a single disciplinary domain but instead draws on financial economics, econometrics, statistical modelling, and AI-based forecasting.

5.1 Lotka's Law and Author Productivity

Lotka's Law [54] was applied to examine the distribution of author productivity within the corpus. The results show a highly unequal productivity structure. Approximately 86% of authors published only one paper, 9% published two papers, 3% published three papers, and fewer than 2% published four or more papers (Table 7).

Only five authors, corresponding to 0.3% of the 1,708 authors in the sample, contributed eight or more papers. This pattern indicates that the literature is characterised by a broad base of occasional contributors and a small core of highly productive researchers. In the context of hybrid and ensemble

Table 3
Most Productive Authors and Country of Affiliation – Publications

#	Authors	Country	Articles	#	Authors	Country	Fractionalised
1	ZHANG Y	China	18	1	NA N	China	12.00
2	LI Y	USA	14	2	ZHANG Y	China	6.00
3	WANG Y	USA	14	3	WANG Y	USA	4.57
4	NA N	China	12	4	NONEJAD N	Denmark	4.00
5	LIU J	Canada	9	5	LI Y	USA	3.91
6	MA F	China	9	6	KRISTJANPOLLER W	Chile	3.83
7	WANG H	China	9	7	WANG H	China	3.35
8	KRISTJANPOLLER W	Chile	7	8	LAHMIRI S	Canada	3.00
9	GUPTA R	South Africa	6	9	MA F	China	2.50
10	HU Y	New Zealand	6	10	OU P	China	2.50

Fractionalised authorship controls for co-authorship and quantifies an individual author's contribution to a published set of papers, under the assumption of equal contribution by all co-authors within each document.

Table 4
Most Productive Authors and Country of Affiliation – Citations

#	Authors	Country	Total Citations	#	Authors	Country	Fractionalised Citations
1	RAPACH D	USA	1028	1	KRISTJANPOLLER W	Chile	410
2	STRAUSS J	USA	1028	2	PAYE B	USA	351
3	ZHOU G	USA	1028	3	RAPACH D	USA	343
4	KRISTJANPOLLER W	Chile	854	4	STRAUSS J	USA	343
5	MINUTOLO M	USA	604	5	ZHOU G	USA	343
6	ZHANG Y	China	568	6	MINUTOLO M	USA	280
7	KIM H	Korea	547	7	KIM H	Korea	274
8	WON C	Korea	513	8	WON C	Korea	257
9	WEI Y	China	437	9	ZHANG Y	China	175
10	LIU J	Canada	410	10	ROH T	Korea	145

Citations are calculated for documents included in this bibliometric analysis ($n = 690$).
Total citations correspond to the total number of citations (TC) recorded in the bibliographic database for documents included in this sample ($n = 690$). Citations are fully attributed to each co-author and are not fractionalised.
Fractionalised citations divide each document's total citations (TC) equally among its co-authors.

Table 5
Most Frequent Journals in the Bibliometric Review

#	Sources	SJR	H-Index	Articles	%
1	EXPERT SYSTEMS WITH APPLICATIONS	Q1	290	27	3.9%
2	JOURNAL OF FORECASTING	Q1	71	22	3.2%
3	COMPUTATIONAL ECONOMICS	Q2	51	21	3.0%
4	ENERGY ECONOMICS	Q1	230	11	1.6%
5	IEEE ACCESS	Q1	290	11	1.6%
6	INTERNATIONAL REVIEW OF ECONOMICS & FINANCE	Q1	87	11	1.6%
7	ACM INTL CONFERENCE PROCEEDING SERIES	–	164	10	1.4%
8	INTERNATIONAL JOURNAL OF FORECASTING	Q1	126	9	1.3%
9	INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS	Q1	110	9	1.3%
10	APPLIED SOFT COMPUTING	Q1	208	8	1.2%

Table 6
Most Cited Papers in the Bibliometric Review

#	Paper	SJR	TC	TC/Year	NTC	DOI
1	RAPACH D, 2010, **REV FINANC STUD**	Q1	1028	60.47	4.74	10.1093/rfs/hhp063
2	KIM H, 2018, **EXPERT SYST APPL**	Q1	513	57.00	9.26	10.1016/j.eswa.2018.03.002
3	PAYE B, 2012, **J FINANC ECON**	Q1	351	23.40	5.28	10.1016/j.jfineco.2012.06.005
4	BALCILAR M, 2017, **EMPIR ECON**	Q1	268	26.80	3.74	10.1007/s00181-016-1150-0
5	WEI Y, 2017, **ENERGY ECON**	Q1	230	23.00	3.21	10.1016/j.eneco.2017.09.016
6	CHRISTIANSEN C, 2012, **J APPL ECONOM**	Q1	206	13.73	3.10	10.1002/jae.2298
7	BASAK S, 2019, **N AM ECON FINANC**	Q1	204	25.50	4.97	10.1016/j.najef.2018.06.013
8	KASTNER G, 2014, **COMPUT STAT DATA ANL**	Q1	190	14.62	4.22	10.1016/j.csda.2013.01.002
9	VIDAL A, 2020, **EXPERT SYST APPL**	Q1	175	25.00	5.98	10.1016/j.eswa.2020.113481
10	PENG Y, 2018, **EXPERT SYST APPL**	Q1	165	18.33	2.98	10.1016/j.eswa.2017.12.004

Normalised Total Citations (NTC) are calculated by `bibliometrix` in *R* by dividing the total number of citations of an article by the average number of citations of all documents published in the same year. This measure accounts for differences in citation practices across disciplines and provides a more meaningful comparison of citation impact.

Table 7
Lotka's Law: Theoretical and Empirical Distribution in the Bibliometric Analysis

# Manuscripts Published	Authors	Empirical Share	Theoretical Share
1	1477	86.48%	67.34%
2	154	9.02%	7.99%
3	47	2.75%	2.30%
4	11	0.64%	0.95%
5	8	0.47%	0.48%
6	4	0.23%	0.27%
7	2	0.12%	0.17%
8	1	0.06%	0.11%
9	1	0.06%	0.08%
10	1	0.06%	0.06%
12	2	0.12%	0.03%
Displayed total	**1708**	**100%**	**80%**

The theoretical shares are reported for the observed productivity classes displayed in the table. Because the theoretical Lotka distribution extends beyond the displayed classes, the displayed theoretical shares need not sum to 100%.

volatility forecasting, such concentration may reflect the emergence of specialised research groups driving methodological innovation.

5.2 Keywords Co-occurrence, Trend Topics and Thematic Map

Keyword analysis using both database-assigned keywords (ID) and author-assigned keywords (DE) reveals the intellectual structure of the field. Figure 4 reports the ten most frequent keywords in each category. Among database-assigned keywords, terms such as "financial markets", "electronic trading", and "investments" reflect the broad financial domain of the corpus. Among author-assigned keywords, terms such as "machine learning", "deep learning", "LSTM", and "neural networks" indicate the increasing prominence of computational and AI-based forecasting methods.

Figure 5 provides a longitudinal view of the most relevant author keywords. Earlier periods are dominated by foundational econometric and time-series concepts, including *volatility*, *GARCH*, *ARIMA*, *time series*, and *forecasting*. These terms remain central to the field but exhibit relatively stable trajec-

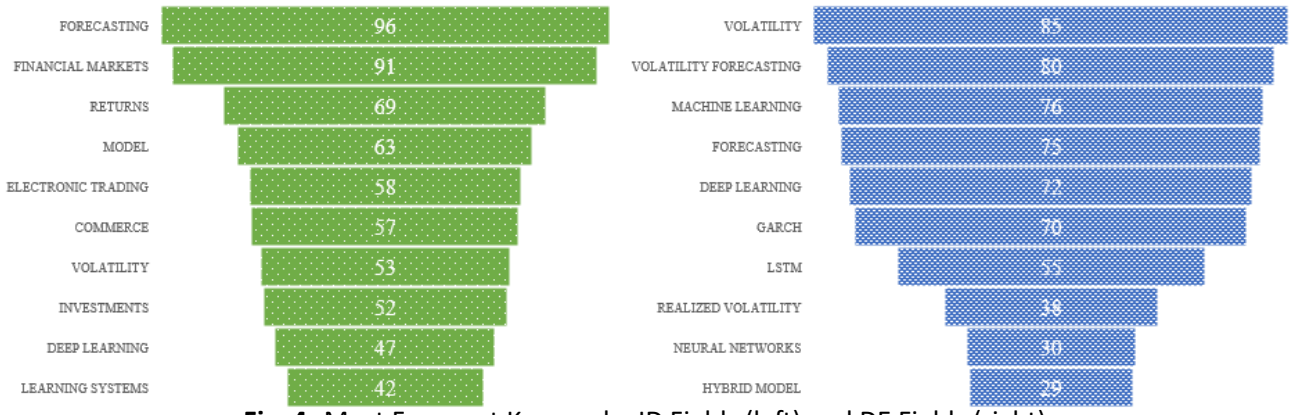

Fig. 4. Most Frequent Keywords: ID Fields (left) and DE Fields (right)

tories, suggesting their role as methodological baselines. From the mid-2010s onwards, terms such as *machine learning*, *deep learning*, *LSTM*, and *XGBoost* become increasingly prominent, reflecting the diffusion of non-linear and data-driven approaches. In the most recent period, terms such as *cryptocurrency* and *bitcoin* indicate diversification in the asset classes studied.

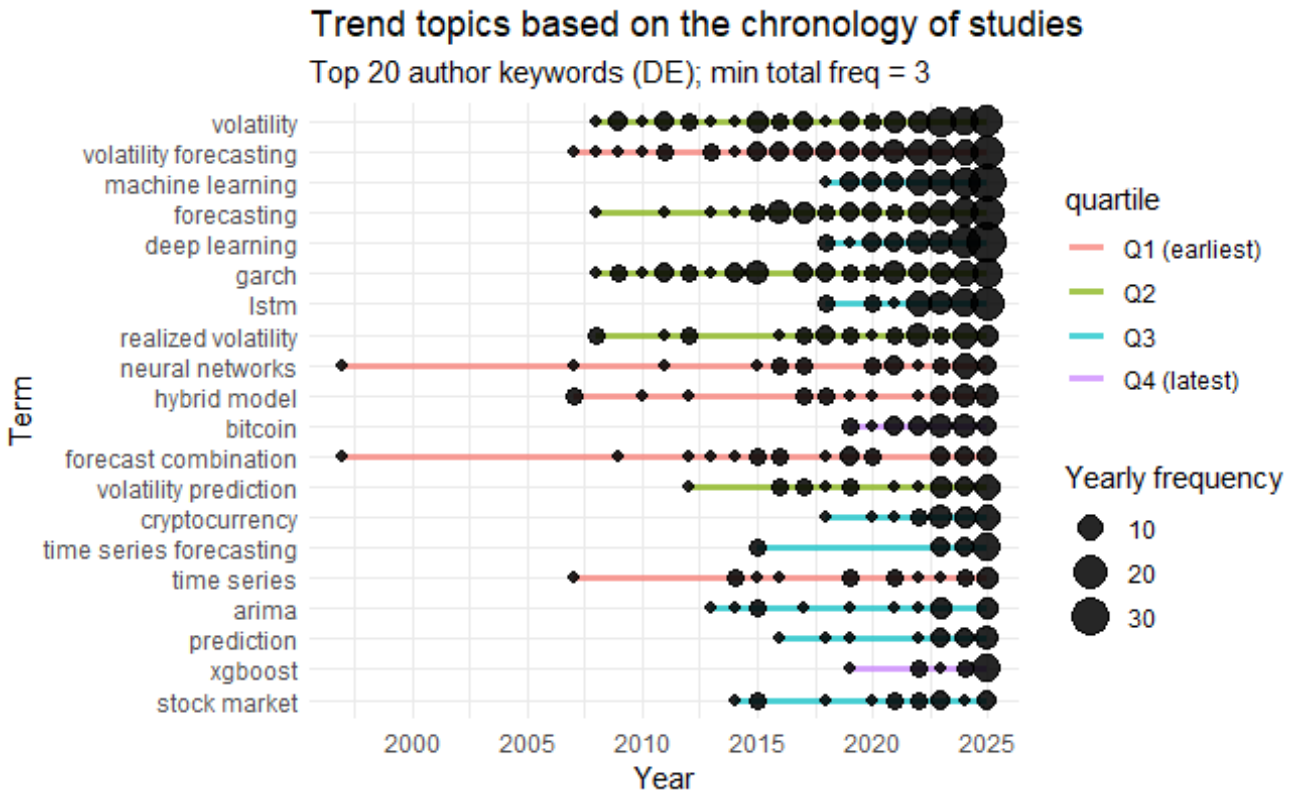

Fig. 5. Trend Topics Based on the Chronology of Studies

The thematic map in Figure 6 further clarifies the structure of the research field. Motor themes such as *market volatility* and *portfolio optimisation* are both central and well developed, indicating their role as organising themes in the literature. Niche themes such as *hybrid deep learning* and *EMD* are internally developed but less connected to the broader field, suggesting specialised methodological communities. Basic themes such as *machine learning*, *random forest*, and *forecast combinations* are highly central but less internally consolidated, indicating transversal concepts that continue to evolve. Emerging or declining themes, including *value-at-risk*, *high-frequency data*, and *wavelet transform*, occupy less central and less dense positions; their location may reflect either declining attention or early-stage development, depending on the specific theme.

Overall, the bibliometric evidence points to a clear methodological transition. Conventional econometric approaches remain foundational, but the research frontier is increasingly shaped by machine

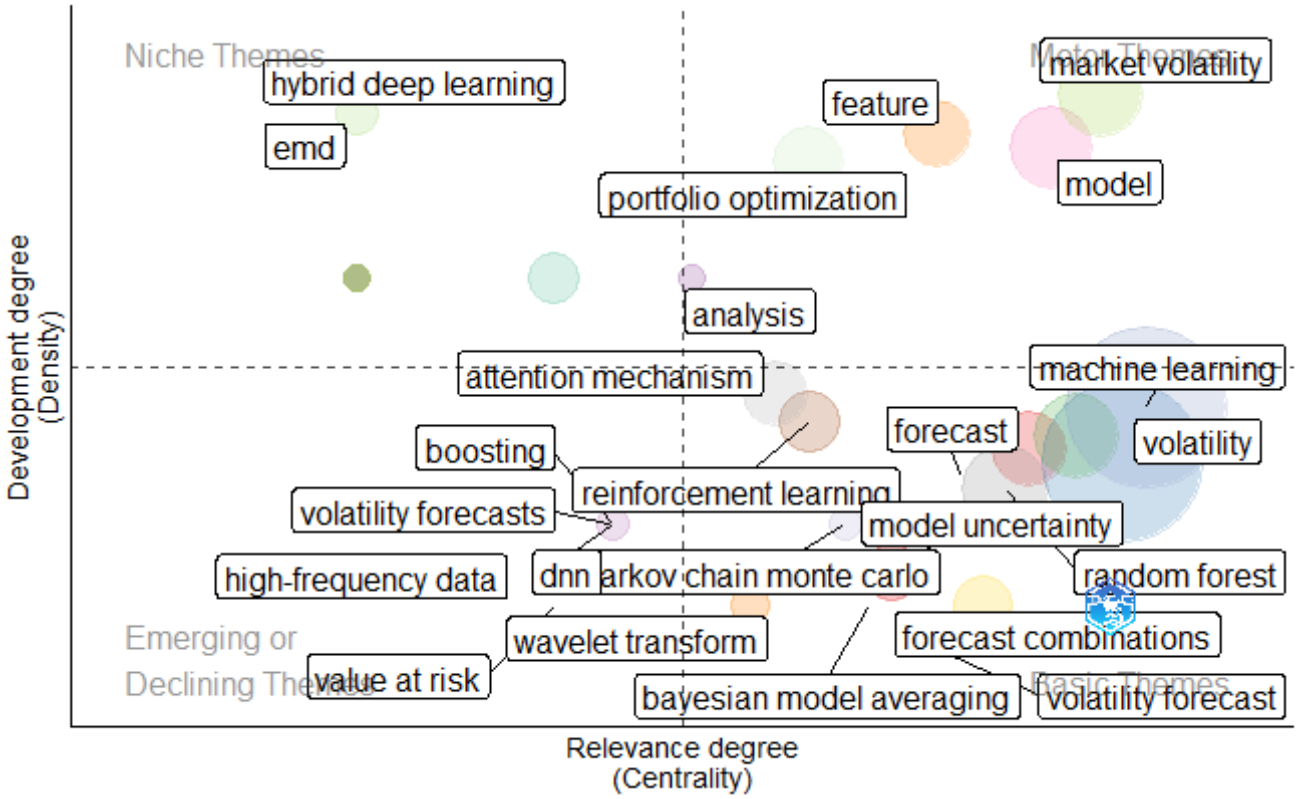

Fig. 6. Thematic Map Based on Article Keywords (DE)

learning, deep learning, hybrid architectures, ensemble methods, and signal decomposition techniques. The field is therefore best understood as an interdisciplinary convergence between financial econometrics and computational predictive modelling.

6. Results: LLM-Assisted Methodological Classification

This section reports the methodological classification of the 121-paper core corpus. The analysis proceeds in five steps. First, it describes the bibliometric profile of the core sample. Second, it evaluates inter-model agreement among *Claude 4.6*, *Gemini 3*, and *GPT-5*. Third, it analyses the main sources of classificatory divergence across models. Fourth, it reports the human audit of a stratified subsample, assessing whether the LLM-assisted classifications converge with domain-expert judgement under identical informational constraints. Finally, it uses the adjudicated classifications and qualitative audit to synthesise the substantive methodological evidence on hybrid and ensemble volatility forecasting architectures.

6.1 Profile of the Core Corpus

The core corpus consists of 121 studies selected after the multi-stage filtering procedure based on *citation-based impact*, *recency*, and *journal productivity*. The sample spans the period from 2007 to 2025 and is composed predominantly of journal articles, with three review papers and one proceedings contribution, distributed across 24 distinct sources (Table 8).

Table 8
Bibliometric Profile of the Core Sample (n = 121)

Description	Results	Description	Results
MAIN INFORMATION ABOUT DATA		DOCUMENT TYPES	
Timespan	2007:2025	article	117
Sources (Journals, Books, etc.)	24	article; proceedings paper	1
Documents	121	review	3
Annual Growth Rate %	23.7	COUNTRIES – PUBLICATION RANKING	
Document Average Age	4.12	China	49
Average citations per doc	46.69	India	10

Table 8
Continued

Description	Results	Description	Results
Average citations per year per doc	6.62	Chile	8
DOCUMENT CONTENTS*		UK	6
Keywords Plus (ID)	310	Iran	6
Author's Keywords (DE)	485	Brazil	5
AUTHORS		USA	5
Authors	331	COUNTRIES – CITATION RANKING	
Author Appearances	402	China	1499
Authors of single-authored docs	9	Chile	926
AUTHORS COLLABORATION		Korea	683
Single-authored docs	9	India	568
Documents per Author	0.37	Brazil	388
Co-Authors per Doc	3.32	Iran	255
International co-authorships %	23.97	UK	219

The average document age of 4.12 years confirms the recency emphasis embedded in the selection procedure, while the average citation count of 46.69 per document, corresponding to 6.62 citations per year, reflects the citation-based impact threshold applied at the filtering stage. The corpus is internationally distributed but geographically concentrated: China leads both publication volume (49 documents) and total citations received (1,499), followed by Chile in citations (926) and India in publications (10). The 331 authors produced an average of 3.32 co-authors per document, and the international co-authorship rate is 23.97%, indicating meaningful cross-border collaboration alongside a predominance of single-country studies.

The journal distribution reveals a concentration structure consistent with Bradford's Law. The leading source, *Expert Systems with Applications*, accounts for 21 articles, corresponding to 17.4% of the core corpus (Table 9). The first five journals alone account for 42.1% of all documents, while the remaining 19 sources contribute the other 57.9%, with individual journal shares ranging from 0.8% to 5.0%. The majority of sources hold Q1 SJR rankings, with the remainder classified as Q2; no Q3 or Q4 journals are present, reflecting the source-based refinement applied in the sampling procedure.

Table 9
Core Sources: Journal Distribution, SJR Ranking, and Share (n = 121)

#	Sources	SJR	H-Index	Articles	%
1	EXPERT SYSTEMS WITH APPLICATIONS	Q1	290	21	17.4%
2	COMPUTATIONAL ECONOMICS	Q2	51	17	14.0%
3	ENERGY ECONOMICS	Q1	230	9	7.4%
4	IEEE ACCESS	Q1	290	7	5.8%
5	JOURNAL OF FORECASTING	Q1	71	7	5.8%
6	MATHEMATICS	Q2	84	6	5.0%
7	APPLIED SOFT COMPUTING	Q1	208	5	4.1%
8	INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS	Q1	110	5	4.1%
9	NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE	Q1	65	5	4.1%
10	INTERNATIONAL JOURNAL OF FORECASTING	Q1	126	4	3.3%
11	INTERNATIONAL REVIEW OF ECONOMICS & FINANCE	Q1	87	4	3.3%
12	JOURNAL OF RISK AND FINANCIAL MANAGEMENT	Q2	54	4	3.3%
13	RESOURCES POLICY	Q1	138	4	3.3%
14	APPLIED SCIENCES-BASEL	Q2	162	3	2.5%
15	ECONOMIC MODELLING	Q1	126	3	2.5%
16	ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE	Q1	149	3	2.5%
17	ENERGY	Q1	274	2	1.7%
18	ENTROPY	Q2	112	2	1.7%
19	FINANCE RESEARCH LETTERS	Q1	123	2	1.7%
20	JOURNAL OF EMPIRICAL FINANCE	Q1	98	2	1.7%
21	PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS	Q2	195	2	1.7%
22	RISKS	Q2	38	2	1.7%
23	JOURNAL OF ECONOMETRICS	Q1	198	1	0.8%
24	JOURNAL OF INTERNATIONAL FINANCIAL MARKETS INSTITUTIONS & MONEY	Q1	89	1	0.8%
				121	**100%**

6.2 Inter-Model Agreement

The inter-model agreement analysis shows that the reliability of LLM-assisted classification is strongly dimension-specific. Dimensions whose labels are directly signalled in titles and abstracts, such as Data Frequency, Model Category, and Forecast Target, achieve moderate agreement. By contrast, dimensions requiring inference about research design or contribution type, such as Evaluation Framework and Methodological Novelty, show only slight agreement. This pattern indicates that the limiting factor is not simply model capability, but the interaction between taxonomy design and the information available in abstracts.

Agreement among *Claude 4.6*, *Gemini 3*, and *GPT-5* was quantified using two complementary statistics. For each pair of models, Cohen's kappa [50] was computed for each taxonomic dimension, yielding a chance-corrected measure of pairwise concordance. For the three-model system as a whole, Fleiss' kappa [51] was computed for each dimension, extending the pairwise measure to the multi-annotator case. Both statistics are interpreted according to the scale proposed by Landis and Koch [52]: values below 0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and values above 0.80 almost perfect agreement.

Agreement was computed on normalised label strings to abstract away from formatting conventions. The Market/Asset Class dimension was excluded from formal kappa computation because the three models produced free-text annotations using incompatible geographic and typological conventions, rendering exact string-matching-based agreement systematically underestimated even where substantive concordance was present.

Three dimensions reach moderate agreement. Data Frequency achieves the highest Fleiss' κ (0.531) and unanimous three-way agreement on 96 of 121 papers (79.3%), reflecting the direct observability of data granularity in abstract descriptions. Model Category (0.449) and Forecast Target (0.455) also attain moderate agreement, with the *Claude–GPT* pairing achieving the strongest pairwise kappa for Model Category (0.602) and the *Gemini–GPT* pairing the strongest for Forecast Target (0.501). Hybrid Sub-Type (0.342), Combination Strategy (0.208), and Input Features (0.319) fall within the fair agreement range. Evaluation Framework (0.089) and Methodological Novelty (0.107) remain in the slight agreement range, indicating that these dimensions are structurally more difficult to classify from abstract-level information alone.

Table 10
Cohen's Kappa (Pairwise) and Fleiss' Kappa by Dimension ($n = 121$)

Dimension	κ Claude–Gemini	κ Claude–GPT	κ Gemini–GPT	Fleiss' κ	Interpretation
Model Category	0.368	**0.602**	0.382	**0.449**	Moderate
Hybrid Sub-Type	0.248	**0.553**	0.309	**0.342**	Fair
Combination Strategy	0.272	0.287	0.155	**0.208**	Fair
Forecast Target	0.490	0.407	**0.501**	**0.455**	Moderate
Input Features	0.250	**0.543**	0.194	**0.319**	Fair
Data Frequency	0.390	**0.509**	**0.691**	**0.531**	Moderate
Evaluation Framework	0.366	0.118	0.160	0.089	Slight
Methodological Novelty	-0.047	0.353	0.096	0.107	Slight

Kappa is computed on normalised label strings across 121 papers. Market/Asset Class is excluded because of free-text format incompatibility. Bold values indicate the highest pairwise kappa within each row. Interpretation of Fleiss' kappa follows Landis and Koch [52]: < 0.20 Slight; 0.21–0.40 Fair; 0.41–0.60 Moderate; 0.61–0.80 Substantial; > 0.80 Almost perfect. Negative kappa indicates observed agreement below the level expected by chance.

The majority-vote analysis confirms that the three-model pipeline resolves a high proportion of cases across most dimensions (Table 11). Data Frequency achieves a 100% majority-resolution rate, with no cases requiring manual adjudication. Model Category (97.5%), Forecast Target (93.4%), and Hybrid Sub-Type (90.9%) are also efficiently resolved. The two dimensions with the highest residual adjudication burden are Evaluation Framework and Methodological Novelty, where 18.2% and 14.0% of papers, respectively, present three-way splits. The relatively high adjudication rate for Evaluation Framework is consistent with its low Fleiss' kappa: the models disagree not merely pairwise, but in

mutually incompatible ways for a non-trivial share of the corpus.

Table 11
Majority Vote: Agreement Tiers by Taxonomic Dimension ($n = 121$)

Dimension	All 3 Agree	Majority ($\geq$ 2/3)	Three-way Split
Model Category	68 (56.2%)	118 (97.5%)	**3 (2.5%)**
Hybrid Sub-Type	34 (28.1%)	110 (90.9%)	**11 (9.1%)**
Combination Strategy	34 (28.1%)	107 (88.4%)	**14 (11.6%)**
Forecast Target	56 (46.3%)	113 (93.4%)	**8 (6.6%)**
Input Features	51 (42.1%)	107 (88.4%)	**14 (11.6%)**
Data Frequency	96 (79.3%)	121 (100.0%)	**0 (0.0%)**
Evaluation Framework	22 (18.2%)	99 (81.8%)	**22 (18.2%)**
Methodological Novelty	39 (32.2%)	104 (86.0%)	**17 (14.0%)**

All 3 Agree is a subset of *Majority* ($\geq 2/3$).
Three-way Split denotes cases in which all three models assign different labels, requiring manual adjudication. Percentages are computed over $n = 121$ papers per dimension.

6.3 Divergence Analysis

The disagreement cases are not treated as random annotation noise. Rather, they provide a diagnostic test of the taxonomy by identifying the boundaries where architectural categories are least stable under abstract-only classification. A structured inspection of the disagreement cases reveals four recurring patterns of classificatory divergence.

The first pattern is boundary divergence at the Hybrid–Ensemble frontier. The most frequent source of pairwise disagreement in the Model Category dimension concerns papers that train multiple independent models and combine their outputs. *Claude* and *Gemini* tend to classify these configurations as Ensemble, whereas *GPT-5* more frequently assigns them to Hybrid, particularly when the combination involves learned weights or an attention mechanism. This pattern reflects a genuine boundary problem: under the taxonomy used here, Ensemble requires that component models are trained independently without interaction, whereas Hybrid requires structural integration during training or inference. Many papers, however, describe architectures that sit precisely on this boundary, combining independently trained forecasters through a learned aggregation layer that introduces some degree of structural coupling.

The second pattern is sub-type conflation in the Hybrid Sub-Type dimension. *Gemini* assigns the Sequential label to 63 papers (52.1%), compared with 16 for *Claude* and 24 for *GPT-5*, while classifying only three papers as Feature-based, compared with 35 for *Claude* and 24 for *GPT-5*. This asymmetry suggests that *Gemini* interprets Sequential as encompassing any multi-stage pipeline, regardless of whether outputs are structurally propagated as inputs to the next stage. The taxonomy definition restricts Sequential to architectures in which the output or residuals of one model become the direct input to the next; simply training models in a comparative sequence does not qualify. *Claude* and *GPT-5*, by contrast, reserve Sequential for this narrower meaning and distribute a larger share of papers to Feature-based, indicating that they more consistently apply the distinction between output-propagation and feature-transformation architectures.

The third pattern is category saturation in Forecast Target. *Gemini* assigns the label Other to 21 papers (17.4%) and *GPT-5* to 33 papers (27.3%), compared with only three for *Claude*. Inspection of the affected papers reveals that the majority forecast general price levels or directional movement as a secondary objective alongside volatility, a configuration that *Claude* classifies as Return, while *Gemini* and *GPT-5* prefer Other. This divergence inflates disagreement rates on Forecast Target and underscores the need for an explicit scope rule within each categorical definition specifying how adjacent forecasting objectives are to be treated.

The fourth pattern is dimension-specific vocabulary divergence in Combination Strategy. *GPT-5* assigns the Cascade label to 39 papers (32.2%), compared with three for both *Claude* and *Gemini*. Inspection confirms that *GPT-5* applies Cascade to any architecture in which a model's output is passed forward as input to a subsequent component, a broad interpretation that includes what *Claude* and *Gemini* typically classify as None, meaning a single pipeline with no separate ensemble step, or Weighted Average. *Claude* and *Gemini* restrict Cascade to explicit multi-stage pipelines with a structured hand-off between distinct forecasting stages. This is the single largest source of pairwise

disagreement between *GPT-5* and the other two models on this dimension, and its resolution through adjudication accounts for most of the 14 cases referred to manual review.

6.4 Human Audit Results

The human validation exercise extends the inter-model agreement analysis by introducing a fourth annotator: the researcher, acting as a domain-expert reference classifier. The human classifier serves as a reference annotator rather than as an infallible ground truth. Divergences between the human classifier and any LLM are therefore interpreted as evidence of possible taxonomic ambiguity, information constraints, or model-specific classification tendencies, rather than as a unidirectional measure of LLM error. Cohen's kappa is a symmetric measure, and all pairwise statistics reported below are interpreted accordingly.

The audited subsample comprises 25 papers drawn through stratified random sampling, with strata defined by publication period to ensure representativeness across the pre-2022 (n = 9), 2022–2024 (n = 7), and 2025 (n = 9) cohorts. Each paper was classified using only the title, abstract, and author keywords, that is, under the same informational constraints applied to the three LLMs. Classification was completed independently before the LLM outputs were consulted, preserving the integrity of the human annotation.

Table 12 reports pairwise Cohen's kappa between the human classifier and each LLM individually, as well as Fleiss' kappa across all four annotators jointly. Table 13 reports the corresponding percentage agreement and agreement-tier distributions under the four-annotator scheme.

Table 12
Cohen's Kappa and Fleiss' Kappa (Four Annotators) by Dimension ($n = 25$)

Dimension	κ(H, Claude)	κ(H, Gemini)	κ(H, GPT)	Fleiss' κ	Interpretation
Model Category	0.385	**0.876**	0.648	**0.587**	Moderate
Hybrid Sub-Type	0.468	0.407	**0.582**	**0.437**	Moderate
Combination Strategy	**0.324**	0.301	0.111	0.258	Fair
Forecast Target	**0.450**	0.299	0.292	0.394	Fair
Input Features	**0.354**	0.083	0.167	0.208	Fair
Data Frequency	0.609	0.224	**0.711**	**0.516**	Moderate
Evaluation Framework	−0.012	−0.098	0.021	−0.037	Poor
Methodological Novelty	**0.102**	−0.014	−0.101	0.055	Slight

Cohen's kappa compares the human classifier with each LLM, while Fleiss' kappa is computed across all four annotators: *Human*, *Claude 4.6*, *Gemini 3*, and *GPT-5*. The human classifier is treated as a reference annotator in parity with the three LLMs rather than as an infallible ground truth.

Table 13
Agreement-Tier Distribution: Four-Annotator Scheme by Dimension ($n = 25$)

Dimension	All 4 Agree	All 4 %	Majority $\geq$ 3/4	Majority %	4-way Split	4-way Split %
Model Category	14	56%	**21**	**84%**	4	16%
Hybrid Sub-Type	9	36%	**15**	**60%**	10	40%
Combination Strategy	8	32%	**17**	**68%**	8	32%
Forecast Target	9	36%	**19**	**76%**	6	24%
Input Features	7	28%	**13**	**52%**	12	48%
Data Frequency	19	76%	**25**	**100%**	0	0%
Evaluation Framework	2	8%	**16**	**64%**	9	36%
Methodological Novelty	4	16%	**14**	**56%**	11	44%

All 4 Agree is a subset of *Majority* ($\geq 3/4$).
Four-way Split denotes cases in which all four annotators assign different labels. Percentages are computed over $n = 25$ papers per dimension (seed = 42).

Several structural patterns emerge from the four-annotator analysis. The first is high concordance on dimensions that are well specified and lexically observable. Data Frequency achieves moderate

Fleiss' kappa across all four annotators (κ = 0.516), with all 25 papers reaching a three-annotator majority. This result is consistent with the three-model finding in Table 10 (Fleiss' κ = 0.531) and confirms that data granularity is reliably signalled by standardised vocabulary in abstracts. Model Category attains the highest Fleiss' kappa across the four annotators (κ = 0.587), above the three-model estimate of 0.449, with four-way splits confined to four papers lying on the Hybrid–Ensemble or Machine Learning–Deep Learning boundary. Hybrid Sub-Type also reaches moderate four-annotator agreement (κ = 0.437), despite its five-category structure, with divergence concentrated in the Sequential–Feature-based and Parallel–Decomposition-based boundary pairs identified in the three-model error analysis.

The second pattern concerns dimensions where pairwise concordance with the human classifier is heterogeneous across models, revealing possible model-specific interpretive tendencies rather than uniform pipeline limitations. In Model Category, *Gemini 3* achieves the highest pairwise concordance with the human classifier (0.876), while *Claude 4.6* achieves the lowest (0.385), indicating that the two models may apply substantively different decision boundaries for the Hybrid–Ensemble distinction. In Data Frequency, *GPT-5* and the human classifier agree most closely (0.711), whereas *Gemini 3* shows lower concordance (0.224), suggesting that the model applies a finer-grained frequency vocabulary that produces label fragmentation. In Input Features (κ = 0.208), the human classifier systematically assigns Multimodal where the LLMs assign Price or Technical, reflecting a tendency for LLMs to anchor on the most prominent feature category mentioned in the abstract rather than infer the full composition of the input set from contextual cues.

The third pattern is structural unreliability on interpretively demanding dimensions that cannot be resolved from abstract-level information alone. Evaluation Framework achieves Fleiss' $\kappa = -0.037$ across all four annotators, a value below the chance-corrected baseline, despite raw agreement with *Claude* and *Gemini* ranging from 60% to 64%. This apparent inconsistency reflects the kappa paradox: when one label dominates the distribution, in this case statistical metrics only, high raw agreement can coexist with near-zero or negative kappa because the expected agreement under chance is itself high. Methodological Novelty exhibits a related but distinct problem. The human classifier disagrees with the three-LLM majority in 10 of 25 cases on this dimension, most frequently assigning Benchmark Replication where the LLM majority assigns Architecture.

Taken together, the human audit confirms that the three-model pipeline provides a reliable approximation of the broader inter-annotator agreement structure. The addition of a human domain expert raises agreement on the more tractable dimensions and leaves the structurally unreliable dimensions largely unchanged. This pattern is consistent with the interpretation that the primary limiting factor is information availability in the abstract rather than LLM capability alone.

6.5 Substantive Evidence from the Methodological Classification

Having established the reliability profile of the classification pipeline, the remainder of this section uses the adjudicated labels and the accompanying qualitative audit to synthesise the substantive methodological evidence in the core corpus. The objective is not to produce a formal meta-analysis of forecasting accuracy, since the studies differ in assets, horizons, targets, and evaluation metrics, but to identify recurring methodological regularities across the principal classification dimensions.

The dominant empirical regularity is that hybrid and ensemble architectures outperform their constituent standalone components in the majority of reported comparisons across asset classes, data frequencies, forecast targets, and evaluation metrics. Papers reporting mixed results or only marginal improvements represent a minority. This regularity should be read as a synthesis of the evidence reported by the primary studies rather than as the outcome of a formal meta-analysis: the studies differ substantially in assets, horizons, forecast targets, benchmark specifications, and evaluation metrics, so the direction of the reported effects can be summarised consistently while their magnitudes cannot be pooled. The question facing the literature is therefore no longer simply *whether* combining approaches adds value, but *which* combination, *under what market conditions*, and *with what efficiency*.

Architecture and Combination Strategy

Among hybrid architectures, the *sequential cascade*—in which GARCH-family outputs serve as engineered features for a neural network—is the most studied and most consistently successful design. H. Y. Kim and Won [55] report a GEW-LSTM model achieving a 37% reduction in MAE and a 57% reduction in MSE relative to the best existing hybrid model for the KOSPI 200. Kristjanpoller and Minutolo [56] document a 30% improvement in HMSE for oil price volatility using a GARCH–ANN cascade with fi-

nancial variables, while Kristjanpoller and Minutolo [57] confirm a 25% reduction in MAPE for gold volatility using an ANN–GARCH framework. The cascade advantage is most pronounced when GARCH estimates serve as non-linear signals rather than merely as pre-processed inputs.

For *ensemble-by-combination*, the forecast combination puzzle extends to machine learning: equally weighted or shrinkage-based combinations of individual models frequently dominate any single architecture [58, 59]. Shrinkage methods, including Elastic Net and LASSO, outperform both individual HAR extensions and standard combinations for oil volatility, while also generating sizeable economic gains for mean–variance investors [60]. Stacking ensembles add a further layer: Aras [61] show that stacking with LASSO meta-learning and higher-order GARCH processes outperforms both standard hybrids and plain stacking for Bitcoin volatility.

Decomposition Methods

Signal decomposition prior to modelling is the clearest and most transferable empirical regularity in the corpus. When a series is first decomposed into intrinsic mode functions and a residual, and each component is modelled separately before reconstruction, forecasting errors fall substantially across virtually all base learners. Lin et al. [62] find CEEMDAN–LSTM to be uniformly the best performer across the CSI 300, S&P 500, and STOXX 50, with all single models inferior to their CEEMDAN-enhanced counterparts. Zhang et al. [63] quantify this gain precisely: introducing CEEMDAN into a Random Forest model reduces MSE by 52% for the Shanghai Stock Exchange. Koo and G. Kim [64] report a 21% RMSE gain through volume-upped distribution transformation prior to GARCH–LSTM fitting on S&P 500 realised volatility. A methodological consensus emerging from this strand is that *the choice of decomposition method matters more than the choice of the subsequent learner*: CEEMDAN and VMD consistently outperform standard EMD in recent papers, and the performance gap between alternative base learners, such as LSTM, GRU, and Random Forest, tends to narrow once decomposition is applied.

Forecast Target and Risk Measures

For *conditional volatility* targets, which constitute the largest group at approximately 55% of the corpus, DL-enhanced GARCH models improve upon pure GARCH specifications, particularly at longer horizons and during periods of elevated volatility. Transformer-based hybrids are especially effective: Mishra et al. [65] document that MT-GARCH and MTL-GARCH outperform all individual benchmarks, with evidence of robustness during the COVID-19 pandemic.

For *realised volatility*, the HAR model serves as the dominant benchmark, and the evidence consistently supports ML augmentation. Ribeiro et al. [66] achieve an average R^2 of 0.635 at the one-day horizon with HAR–PSO–ESN across three NASDAQ stocks, with statistically significant results under the Friedman–Nemenyi test. Ngwaba [67] confirm the outperformance of HAR-RV-CARMA over standalone models across five ETFs using DM tests.

For *Value-at-Risk*, hybrid models show particularly strong gains during crisis episodes. Kakade et al. [68] evaluate GARCH–LSTM and GARCH–BiLSTM ensembles for crude oil VaR during the 2007–2009 and 2020–2021 crises, finding that standard GARCH fails systematically in high-volatility regimes, while the FHS-BiLSTM-Hybrid achieves the lowest loss values across all coverage tests. Léber and Egyed [69] confirm that sentiment-augmented GARCH–LSTM significantly outperforms conventional GARCH for S&P 500 individual stocks in VaR estimation over the period 2019–2024.

Input Features: Incremental Value of Alternative Information

Three categories of alternative features yield well-documented incremental content. First, **macroeconomic and policy uncertainty**: Global EPU and US EPU dominate traditional oil market predictors in a GARCH–MIDAS–DMA framework [70], while military build-ups and war escalation are the most impactful geopolitical risk components for US stock volatility [71]. Second, **climate variables**: ENSO decomposed via STL significantly improves GARCH–MIDAS forecasts for US grain futures [72]; the SOI trend is the most economically valuable climate predictor for WTI crude oil [73]; and climate risk indices provide incremental predictive accuracy for energy futures [74]. Third, **textual sentiment**: Twitter sentiment improves return and volatility forecasts for renewable energy stocks, with gains concentrated in deep learning specifications [75], while Google Search Volume provides incremental predictive power over GARCH models for energy commodities [76].

A recurring caveat is that these gains are often context-specific and horizon-dependent. Importantly, Fu et al. [77] provide a cautionary counterpoint: for daily INE crude oil futures, low-frequency macro variables have only weak incremental effects, and adding more predictors does not systemati-

cally improve accuracy. Feature selection, rather than feature breadth, is therefore the critical design decision.

Asset Class, Market Regime, and Evaluation

Market-specific patterns are consistent. For *equity* markets, CEEMDAN/VMD–DL ensembles dominate the post-2020 Chinese market cohort, while Transformer hybrids with macro variables represent the US frontier. For *commodity* markets, model combination and shrinkage methods within the HAR-RV framework consistently dominate individual models; GARCH–MIDAS models with climate and weather predictors represent the leading approach for agricultural and energy futures. For *cryptocurrency* markets, a notable finding challenges the complexity premium: García-Medina and Aguayo-Moreno [78] show that MLP is not statistically distinguishable from LSTM–GARCH under the DM test at hourly frequency, suggesting diminishing returns to architectural complexity in highly non-linear series.

With respect to market regimes, hybrid models with regime-switching maintain performance during crises, while standard GARCH models fail. Lu et al. [79] document that MS–LASSO improves US equity volatility forecasts both statistically and economically across business cycles. Liu and Pan [80] identify a business-cycle asymmetry: technical indicators outperform macro variables during expansions, while the reverse holds during recessions, with the optimal strategy being to combine both. Transformer hybrids demonstrate robustness to COVID-19 structural changes in both equity and multi-asset contexts [65, 81].

A cross-cutting methodological concern is evaluation heterogeneity. Approximately 75% of papers rely solely on statistical loss functions (MSE, MAE, RMSE, and MAPE), while a subset employs the Model Confidence Set [62, inter alia] or the Diebold–Mariano test [67, 71, inter alia] for formal predictive ability testing. Economic evaluation, employed in approximately 20% of papers, consistently confirms that statistically significant improvements translate into meaningful gains: shrinkage-based volatility forecasts generate sizeable portfolio utility gains [60, 71], forecast combination for cryptocurrencies yields an average 3.46% wealth-equivalent utility gain [82], and decomposition-based ensemble strategies for VIX implied volatility improve cumulative trading returns across 1–22-day horizons [83].

Synthesis

Table 14 consolidates the evidence across the principal classification dimensions. Three overarching conclusions emerge from the cross-dimensional reading. First, *no single model dominates across all contexts*: the winning architecture for equity realised volatility differs from those optimal for crude oil conditional volatility or cryptocurrency VaR. Second, *decomposition is the most universally transferable performance lever*: CEEMDAN or VMD prior to modelling consistently reduces forecasting error by substantial margins, regardless of base learner or market. Third, *regime-awareness and alternative information* (EPU, climate, and sentiment) define the frontier where statistically and economically significant gains remain to be more systematically explored, particularly for assets and markets currently underrepresented in the corpus.

Table 14
Meta-Summary of Empirical Evidence by Dimension (121-Paper Core Sample)

Dimension	Dominant Finding	Key Evidence	Caveats
Architecture or Combination	Sequential cascade (GARCH–DL) and stacking ensembles outperform parallel and simple-average designs	37% MAE reduction [55]; 30% HMSE gain [56]; stacking + LASSO outperforms plain hybrid [61]	Cascade gains are largest for daily equity; evidence is less clear for high-frequency crypto
Decomposition Method	CEEMDAN/VMD decomposition consistently reduces errors across base learners; the gain from decomposition exceeds the gain from learner choice	52% MSE reduction [63]; CEEMDAN–LSTM best across CSI 300, S&P 500, and STOXX 50 [62]; 21% RMSE gain [64]	Abstract-based results; PDF-level quantification is needed for precise meta-statistics

Table 14
Continued

Dimension	Dominant Finding	Key Evidence	Caveats
Forecast Target	VaR forecasting is most sensitive to crisis regimes; RV is most tractable for HAR–ML hybrids; conditional volatility is the largest and most heterogeneous group	FHS-BiLSTM-Hybrid best for VaR in crisis periods [68]; HAR–PSO–ESN achieves R^2=0.635 at the one-day horizon [66]	Evaluation metric heterogeneity limits cross-study comparison
Input Features	EPU, GPR, climate indices, and sentiment provide consistent incremental content; feature selection is crucial	GEPU dominates oil predictors [70]; ENSO improves grain GARCH–MIDAS forecasts [72]; Twitter sentiment improves DL-based forecasts [75]	Low-frequency macro variables have weak daily effects [77]; gains are often horizon- and context-specific
Asset Class	Crypto is often well modelled by simpler ML; oil and agricultural commodities benefit most from macro/climate predictors; the Chinese equity market is dominated by CEEMDAN–DL in the post-2020 cohort	MLP $\approx$ LSTM–GARCH for crypto [78]; HAR–LASSO superior for oil [60]; VMD–DL dominant for A-shares [84]	Corpus is skewed towards equity and commodities; FX and fixed income are underrepresented
Market Regime	Hybrid models with regime-switching maintain performance during crises; standard GARCH fails in high-volatility regimes; technical versus macro feature informativeness is cycle-dependent	MS–LASSO improves US equity forecasts across cycles [79]; Transformer hybrids are robust to COVID-19 [65, 81]; feature informativeness is cycle-dependent [80]	Regime analysis is based mainly on subperiod comparisons; formal structural break tests are infrequent

7. Discussion: LLMs as Research Assistants in Financial Econometrics

The deployment of *Claude 4.6*, *Gemini 3*, and *GPT-5* as structured annotation instruments across a specialised financial econometrics corpus provides evidence on both the promise and the limits of LLM-assisted methodological review. The results indicate that LLMs can support scalable and reproducible classification when the taxonomy is well specified and the relevant information is visible in titles, abstracts, and keywords. At the same time, they show that LLMs cannot reliably infer methodological details that are absent from the abstract, even when multiple frontier models are combined through a majority-vote pipeline.

The discussion proceeds in four steps. First, it considers what inter-model disagreement reveals about the taxonomy itself. Second, it examines the limits of abstract-level LLM classification. Third, it draws implications for future systematic and bibliometric reviews in quantitative finance. Fourth, it discusses the substantive implications of the review for volatility forecasting research and practice.

7.1 What Inter-Model Disagreement Reveals About the Taxonomy

A central finding of the study is that inter-model disagreement is not merely a technical artefact of LLM variability. Rather, it provides a diagnostic test of the taxonomy. The dimensions with the highest agreement are those whose labels are directly signalled by standardised vocabulary in abstracts. Data Frequency, Model Category, and Forecast Target fall into this group: terms such as *daily returns*, *high-frequency data*, *GARCH*, *realised volatility*, and *implied volatility* typically appear explicitly in the abstract or keywords. These dimensions are therefore comparatively robust to automated annotation.

By contrast, the dimensions with weaker agreement require interpretive judgement about the structure of the model or the nature of the paper's contribution. Hybrid Sub-Type, Combination Strategy, Evaluation Framework, and Methodological Novelty are more dependent on how architectural relationships are described. For example, the distinction between a Sequential hybrid and a Feature-based hybrid depends not only on the list of models used, but also on the direction of information flow between components. Similarly, distinguishing an Ensemble from a Hybrid architecture requires knowing whether component models are trained independently or structurally integrated

during training or inference. These details are not always available in abstracts.

The disagreement patterns therefore identify where the taxonomy is empirically stable and where it requires sharper boundary rules. The Hybrid–Ensemble boundary is theoretically defensible, but empirically porous. Many papers combine independently trained learners through learned aggregation layers, producing architectures that are ensemble-like in construction but hybrid-like in deployment. The classification difficulty observed across the LLMs is therefore not simply annotation error; it reflects a genuine ambiguity in the literature's use of architectural terminology.

7.2 The Limits of Abstract-Level LLM Classification

The human audit confirms that the weakest dimensions are constrained primarily by information availability rather than by LLM capability alone. Evaluation Framework and Methodological Novelty remain unreliable even when a domain-expert human classifier is added as a fourth annotator under the same abstract-only conditions. This is an important methodological result. It suggests that prompt refinement can reduce some forms of disagreement, but cannot compensate for information that is not present in the input.

Evaluation Framework illustrates this limitation clearly. Abstracts often report that a model outperforms benchmarks, but they do not always specify whether the evaluation relies only on statistical loss functions, includes formal predictive-accuracy tests, or incorporates economic value measures. These distinctions are usually documented in the empirical results section of the full paper. Methodological Novelty is similarly difficult to classify from abstracts alone because a paper's contribution claim may concern architecture, data, evaluation, or benchmark replication, and this claim is often articulated in the introduction, conclusion, or methodological sections rather than in the abstract.

The implication is that LLM-assisted classification should be designed around the information architecture of scientific papers. Some dimensions are suitable for abstract-level annotation; others should be treated as full-paper dimensions from the outset. In this sense, the study supports a bounded view of LLMs as research assistants. They are valuable for scaling well-specified classification tasks, identifying ambiguous cases, and producing reproducible first-pass annotations. They are not substitutes for domain expertise where the classification requires full-paper interpretation or theoretical judgement.

The model-specific divergence patterns further support this bounded view. *Gemini 3* aligns closely with the human classifier on Model Category but performs less consistently on Input Features and Data Frequency. *GPT-5* aligns strongly on Data Frequency but applies a broader interpretation of Cascade in Combination Strategy. *Claude 4.6* tends to classify structurally integrated models as Hybrid more readily than the other models. These differences imply that frontier LLMs should not be treated as interchangeable annotators. A multi-model pipeline is valuable precisely because it exposes such model-specific tendencies and prevents a single model's boundary interpretation from silently determining the final taxonomy.

7.3 Implications for Future Reviews in Quantitative Finance

The findings have several implications for the design of future systematic and bibliometric reviews in quantitative finance. First, taxonomy construction should distinguish between lexically observable dimensions and inferential dimensions. Lexically observable dimensions, such as Data Frequency and Forecast Target, are suitable candidates for near-automated annotation with limited adjudication. Inferential dimensions, such as Evaluation Framework and Methodological Novelty, should either be redesigned using coarser categories or classified from full-text inputs.

Second, future reviews should report dimension-specific reliability rather than a single aggregate agreement statistic. A taxonomy may be highly reliable for some dimensions and unreliable for others. Reporting only an overall agreement score would obscure this structure. The combination of pairwise Cohen's kappa, Fleiss' kappa, majority-vote resolution rates, and human-audit agreement tiers provides a more informative validation framework.

Third, adjudication rates should be treated as substantive evidence about the taxonomy. A high three-way split rate indicates that the classification boundary is either under-specified or insufficiently observable in the available text. In this study, the low adjudication burden for Data Frequency and Model Category suggests that these dimensions are stable, while the higher burden for Evaluation Framework and Methodological Novelty indicates that they require either fuller input information or redesigned coding rules.

Fourth, provider diversity should be regarded as a methodological safeguard. A single-model pipeline would not reveal whether a given classification reflects a stable taxonomic assignment or

a model-specific interpretive tendency. The three-model design used here makes disagreement observable and therefore auditable. Future reviews could extend this logic by assigning model weights by dimension, using human-audit concordance to determine which model receives greater authority for each classification task.

More broadly, the study suggests that the appropriate role of LLMs in financial econometrics reviews is not autonomous classification, but assisted evidence synthesis. LLMs can reduce manual workload, impose consistent coding structures, and identify ambiguous cases for expert review. Their value is highest when they are embedded in a transparent workflow that includes prompt documentation, output normalisation, inter-model agreement analysis, and human validation.

7.4 Implications for Volatility Forecasting Research and Practice

The substantive findings of the methodological classification also have implications for volatility forecasting research. The evidence from the core corpus indicates that hybrid and ensemble architectures now form a central methodological frontier. Across asset classes, forecast targets, and evaluation designs, these architectures generally outperform single-model benchmarks in the results reported by the primary studies. The relevant research question has therefore shifted from whether model combination adds value to which form of combination is most appropriate for a given target, market, horizon, and regime.

One of the strongest empirical regularities is the value of decomposition-based modelling. CEEMDAN, VMD, and related methods consistently improve performance by separating volatility series into components associated with different temporal scales before prediction. This suggests that, for non-stationary and multi-scale financial series, preprocessing architecture may matter as much as, or more than, the choice of the final learner.

A second implication concerns regime sensitivity. Hybrid and ensemble models appear particularly valuable during periods of elevated volatility, when standard GARCH-type specifications may fail to capture non-linear dynamics, structural breaks, or tail-risk behaviour. This is especially relevant for VaR forecasting and internal model validation, where underestimating risk during crisis periods has direct prudential consequences.

A third implication concerns the use of alternative information. Economic Policy Uncertainty indices, geopolitical risk measures, climate variables, search intensity, and textual sentiment all provide incremental predictive content in specific contexts. However, the evidence also shows that these gains are horizon- and asset-dependent. Adding more predictors does not automatically improve forecasting accuracy. Feature selection and economic interpretability remain central design considerations.

For practitioners, the results suggest that hybrid and ensemble forecasting systems should not be adopted as generic black-box solutions. Their value depends on matching architecture to forecasting task. Decomposition-based hybrids are especially promising for multi-scale volatility series; GARCH–DL cascades are well suited to settings where econometric structure and non-linear learning are both valuable; and shrinkage or stacking ensembles are appropriate when the relative informativeness of individual models varies across regimes. The methodological challenge is therefore no longer simply to increase model complexity, but to design architectures that are aligned with the statistical structure of the volatility measure, the available information set, and the economic purpose of the forecast.

8. Conclusion

This paper addressed a gap at the intersection of three literatures: financial volatility forecasting, hybrid and ensemble modelling, and LLM-assisted scientific classification. While existing reviews have examined volatility forecasting, machine learning in finance, and hybrid forecasting methods separately, none has combined bibliometric mapping with a granular methodological taxonomy focused specifically on hybrid and ensemble architectures for financial volatility forecasting. Nor has prior work in financial econometrics evaluated large language models as structured annotation instruments for methodological review.

The study makes three contributions. Conceptually, it develops a taxonomy that distinguishes single-model, hybrid, and ensemble forecasting architectures and further classifies multi-model systems by architectural sub-type, combination strategy, forecast target, input features, data frequency, market or asset class, evaluation framework, and methodological novelty. Methodologically, it proposes a replicable review protocol that combines bibliometric analysis of 690 publications with LLM-assisted classification of a 121-paper core corpus using *Claude 4.6*, *Gemini 3*, and *GPT-5*. Empirically, it evaluates the reliability of this classification pipeline through inter-model agreement statistics and a domain-expert human audit conducted under identical abstract-level information constraints.

The bibliometric results show that the literature has expanded rapidly, with a marked acceleration after 2020. The field remains grounded in financial econometrics but is increasingly shaped by machine learning, deep learning, hybrid architectures, ensemble methods, and signal decomposition techniques. Source and keyword analyses reveal an interdisciplinary research space located at the intersection of forecasting, computational intelligence, financial economics, and applied risk modelling.

The methodological classification confirms that hybrid and ensemble architectures have become a central frontier in volatility forecasting. Across the core corpus, multi-model approaches generally outperform single-model benchmarks in the comparisons reported by the primary studies, although their gains are context-dependent and the heterogeneity of assets, horizons, targets, and evaluation metrics precludes a formal pooling of effect sizes. Sequential GARCH–DL cascades, stacking and shrinkage ensembles, and decomposition-based hybrids emerge as particularly important architectural families. Among these, decomposition methods such as CEEMDAN and VMD appear to provide the most transferable performance lever, often reducing forecasting error across different base learners, asset classes, and volatility targets. At the same time, the evidence shows that no single architecture dominates across all contexts: forecast target, data frequency, market regime, asset class, and input feature set all condition model performance.

The LLM-assisted classification results provide a second set of findings. Agreement among the three models is strongly dimension-specific. Dimensions whose labels are directly signalled in abstracts, such as Data Frequency, Model Category, and Forecast Target, achieve moderate agreement and are suitable candidates for near-automated annotation. Dimensions requiring fuller methodological interpretation, such as Evaluation Framework and Methodological Novelty, remain unreliable under abstract-only classification, even when a human domain expert is added as a fourth annotator. This finding indicates that the main constraint is not simply LLM capability, but the information available in the input text. Prompt refinement cannot recover methodological details that are absent from abstracts.

For systematic reviews in quantitative finance, the implication is that LLMs are best understood as bounded research assistants rather than autonomous reviewers. They can scale first-pass classification, impose consistent coding structures, and identify ambiguous cases for human adjudication. Their use should nevertheless be embedded in transparent workflows that include prompt documentation, controlled vocabularies, output normalisation, inter-model agreement analysis, and human validation. Provider diversity is also important: the divergence patterns observed across *Claude 4.6*, *Gemini 3*, and *GPT-5* reveal model-specific boundary interpretations that would remain invisible in a single-model pipeline.

For volatility forecasting research and practice, the results suggest that the relevant question is no longer whether hybrid and ensemble models improve upon traditional benchmarks, but which architecture is appropriate for a given forecasting task. GARCH–DL cascades are well suited to settings where econometric structure and non-linear learning are both valuable; decomposition-based hybrids are especially appropriate for non-stationary and multi-scale volatility series; and stacking or shrinkage ensembles are useful when model informativeness varies across regimes. Alternative information sets, including Economic Policy Uncertainty indices, geopolitical risk, climate variables, search intensity, and textual sentiment, provide incremental value in specific contexts, but their contribution remains horizon- and asset-dependent.

The study has limitations that define a clear research agenda. First, the abstract-only classification protocol imposes a reliability ceiling on dimensions that require detailed methodological interpretation. Future work should implement a two-stage pipeline in which abstract-level classification is used for lexically observable dimensions and full-text classification is used for Evaluation Framework, Methodological Novelty, and other inferential dimensions. Second, the use of closed-source frontier models introduces a reproducibility cost, since provider-side updates may alter model outputs over time. Future studies should compare proprietary and open-weight models under identical prompts and corpora. Third, the core corpus is geographically and empirically concentrated, with strong representation of Chinese equity and commodity-market applications and weaker coverage of foreign exchange, fixed income, ESG assets, carbon markets, and emerging-market risk systems. Extending the search strategy to these underrepresented domains would test the portability of the proposed taxonomy and may reveal additional architectural configurations. Fourth, the performance regularities synthesised in Section 6.5 are subject to publication and selection bias. The corpus is composed predominantly of studies that propose or extend hybrid and ensemble architectures, and such studies are more likely to report favourable comparisons against the benchmarks they select. The citation-based and source-based filters applied during core sampling may reinforce this tendency, since highly

cited papers in leading outlets are themselves more likely to report positive findings. The reported superiority of multi-model architectures should therefore be interpreted as a robust regularity in the published record rather than as an unbiased estimate of relative forecasting performance. Registered benchmarking exercises with common datasets, horizons, and loss functions, of the kind proposed by Ge et al. [11], would be required to establish the latter.

Financial volatility forecasting has entered a methodological phase in which hybrid and ensemble models are no longer peripheral alternatives to econometric benchmarks, but central instruments in the forecasting toolkit. The frontier now lies in matching architectures to volatility measures, information sets, market regimes, and economic objectives. This review provides a bibliometric map, a methodological taxonomy, and an empirically validated LLM-assisted review protocol from which that research agenda can proceed with greater precision.

Appendix

A1. Boolean Search String: Query for *Scopus* and *Web of Science*

```
TITLE-ABS-KEY (
  (
    (volatil* W/3 (forecast* OR predict* OR estimat*))
    OR
    ((forecast* OR predict* OR estimat*) W/3 volatil*)
  )
  AND
  (
    financ* OR "stock market*" OR "financial market*" OR equit*
    OR "exchange rate*" OR forex OR currenc* OR crypto*
    OR commodit* OR bond* OR "interest rate*"
  )
  AND
  (
    hybrid* OR "hybrid model*" OR "hybrid approach*" OR ensemble*
    OR "ensemble model*" OR "ensemble learning"
    OR "forecast combination" OR "model combination"
    OR stacking OR bagging OR boosting OR blending
    OR (
      (LSTM OR GRU OR RNN OR CNN OR SVM OR "random forest"
      OR XGBoost OR LightGBM OR CatBoost OR "k-nearest neighbor"
      OR GARCH OR EGARCH OR TGARCH OR APARCH OR FIGARCH
      OR HAR OR HAR-RV OR ARIMA)
      W/3
      (hybrid OR ensemble OR combin*)
    )
    OR (
      ("empirical mode decomposition" OR EMD OR CEEMDAN
      OR ICEEMDAN OR "variational mode decomposition" OR wavelet*)
      W/5
      (hybrid OR ensemble OR combin*)
    )
  )
)
AND PUBYEAR < 2026
```

A2. *R* Script for Core Sample Construction

```r
# STEP 1 --- Data Preparation
M$TC <- as.numeric(M$TC)
M$PY <- as.numeric(M$PY)
current_year <- 2025
M <- M %>% mutate(TC_per_year = TC / (current_year - PY + 1))

# STEP 2 --- Impact Filter (Top 20%)
threshold <- quantile(M$TC_per_year, 0.80, na.rm = TRUE)
M_top <- M %>% filter(TC_per_year >= threshold)

# STEP 3 --- Recency Filter
M_recent <- M %>%
  filter(PY >= 2025) %>%
  slice_max(order_by = TC_per_year, n = 150)

# STEP 4 --- Merge Impact + Recency
M_core1 <- bind_rows(M_top, M_recent) %>%
  distinct(DI, .keep_all = TRUE)

# STEP 5 --- Core Sources (Bradford Zone 1, ~33.3%)
source_freq <- M %>%
  group_by(SO) %>%
  summarise(n = n(), .groups = "drop") %>%
  arrange(desc(n)) %>%
  mutate(cum_articles = cumsum(n),
         total_articles = sum(n),
         cum_prop = cum_articles / total_articles)
core_sources <- source_freq %>%
  filter(cum_prop <= 0.3334) %>% pull(SO)

# STEP 6 --- Apply Core Source Filter (STRICT)
M_final <- M_core1 %>% filter(SO %in% core_sources)

# STEP 7 --- Export Final Dataset
write.csv(M_final,
          file.path(output_dir, "M_final_clean_dataset.csv"),
          row.names = FALSE)
```

A3. LLM Classification Prompt: Taxonomy and Output Schema

```
You are an expert in financial econometrics and machine learning.
Read the full dataset in the attached file. Your task is to classify
this full dataset of academic papers based on each paper's title,
abstract and keywords. The classification must follow the detailed
taxonomy of hybrid and ensemble models. Return your answer strictly
as a JSON file, with one object per document, ordered from 1 to 121.

CLASSIFICATION STRUCTURE:

1. Model Category
   (Econometric / ML / DL / Hybrid / Ensemble)

2. Hybrid Sub-Type
   -- If Hybrid: Architecture type (Sequential / Parallel /
      Decomposition-based / Feature-based / Meta-learning)
      Decomposition Method, if applicable
      (EMD / CEEMDAN / ICEEMDAN / VMD / Wavelet / None)
   -- If Ensemble: Aggregation strategy
      (Bagging / Boosting / Stacking / Voting / Blending)
   -- If Econometric / ML / DL: N/A

3. Base Models
   (e.g. GARCH, EGARCH, HAR-RV, LSTM, CNN, GRU,
    Transformer, XGBoost, Random Forest, SVM, etc.)

4. Combination Strategy  [meta-level aggregation only]
   (Weighted Average / Simple Average / Stacking /
    Cascade / Boosting / Bagging / None)
   Note: internal architectural mechanisms such as attention fusion
   are NOT combination strategies and should not be reported here.

5. Forecast Target
   (Realised Volatility (RV) / Conditional Volatility (GARCH-class) /
    Implied Volatility / VaR / Return / Other)

6. Input Features
   (Price only / Technical / Macro / Sentiment / Multimodal)

7. Data Frequency
   (High-frequency (tick/min) / Daily / Weekly / Monthly)

8. Market / Asset Class
   (Equity / FX / Commodity / Crypto / Fixed Income / Multi-asset
    -- specify country or region if identifiable)

9. Evaluation Framework  [multi-label: select all that apply]
   (Statistical metrics (MSE / MAE / QLIKE) /
    Statistical tests (DM / MCS) /
    Economic metrics (VaR backtesting / Sharpe Ratio))

10. Methodological Novelty  [primary contribution category]
    (Architecture / Data / Evaluation / Benchmark-Replication)

OUTPUT FORMAT:
{
  "Paper Order": "",  "Title": "",  "Authors": "",
  "Journal": "",  "Year": "",  "DOI": "",
  "Model Category": "",
  "Hybrid Sub-Type": {
    "Architecture Type": "",  "Decomposition Method": "" },
  "Base Models": [],  "Combination Strategy": "",
  "Forecast Target": "",  "Input Features": "",
  "Data Frequency": "",  "Market / Asset Class": "",
  "Evaluation Framework": [],  "Methodological Novelty": ""
}
```

Acknowledgement
This work was supported by national science funds through FCT I.P. (Fundação para a Ciência e a Tecnologia, I.P.), under the projects UID/04152/2025 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS https://doi.org/10.54499/UID/04152/2025, and UID/00315/2025 – Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL) https://doi.org/10.54499/UID/00315/2025, and by the European Union – NextGenerationEU under the project UID/PRR/04152/2025 https://doi.org/10.54499/UID/PRR/04152/2025. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest
The authors declare no conflicts of interest.

Declaration of generative AI and AI-assisted technologies
During the preparation of this work the authors used *Claude (Anthropic)* to assist with language editing and prose refinement. The authors reviewed and edited the content and take full responsibility for the content of the published article. The authors note that the three large language models examined in this study, namely *Claude 4.6 (Anthropic)*, *Gemini 3 (Google DeepMind)*, and *GPT-5 (OpenAI)*, are employed as research instruments within the LLM-assisted classification pipeline described in Section 4. Their role as objects of methodological investigation is fully documented in the Methods section and is conceptually distinct from the manuscript-preparation use declared above.

References
[1] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*. https://doi.org/10.2307/1912773
[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*. https://doi.org/10.1016/0304-4076(86)90063-1
[3] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*. https://doi.org/10.1016/s0304-405x(01)00055-1
[4] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*. https://doi.org/10.1093/jjfinec/nbp001
[5] Britten-Jones, M., & Neuberger, A. (2000). Option prices, implied price processes, and stochastic volatility. *Journal of Finance*. https://doi.org/10.1111/0022-1082.00228
[6] Gunnarsson, E. S., Isern, H. R., Kaloudis, A., Risstad, M., Vigdel, B., & Westgaard, S. (2024). Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *International Review of Financial Analysis*. https://doi.org/10.1016/j.irfa.2024.103221
[7] Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., Reuter, U., Gama, J., & Gandomi, A. H. (2020). Data science in economics: Comprehensive review of advanced machine learning and deep learning methods. *Mathematics*. https://doi.org/10.3390/math8101799
[8] Alvarez, R. B. P., & Bravo, J. M. V. (2026). Forecast combination for asset classes ETFs: Insights on market efficiency and arbitrage. In Proceedings of 20th Iberian Conference on Information Systems and Technologies (CISTI 2025) (pp. 274–286). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-10721-3_25
[9] Dennstädt, F., Zink, J., Putora, P. M., Hastings, J., & Cihoric, N. (2024). Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain. *Systematic Reviews*. https://doi.org/10.1186/s13643-024-02575-4
[10] Delgado-Chaves, F. M., Jennings, M. J., Atalaia, A., Wolff, J., Horvath, R., Mamdouh, Z. M., Baumbach, J., & Baumbach, L. (2025). Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.2411962122
[11] Ge, W., Lalbakhsh, P., Isai, L., Lenskiy, A., & Suominen, H. (2022). Neural network-based financial volatility forecasting: A systematic review. *ACM Computing Surveys*. https://doi.org/10.1145/3483596
[12] Mansilla-Lopez, J., Mauricio, D., & Narváez, A. (2025). Factors, forecasts, and simulations of volatility in the stock market using machine learning. *Journal of Risk and Financial Management*. https://doi.org/10.3390/jrfm18050227
[13] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2020.106181
[14] Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*. https://doi.org/10.3390/asi4010009
[15] Shah, J., Vaidya, D., & Shah, M. (2022). A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intelligent Systems with Applications*. https://doi.org/10.1016/j.iswa.2022.200111
[16] Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*. https://doi.org/10.3390/ijfs11030094
[17] Ferro, J. V. R., Santos, R. J. R. D., de Barros Costa, E., & da Silva Brito, J. R. (2024). Machine learning techniques via ensemble approaches in stock exchange index prediction: Systematic review and bibliometric analysis. *Applied Soft*

Computing. https://doi.org/10.1016/j.asoc.2024.112359
[18] Raimundo, B., & Bravo, J. M. (2026). Forecasting meets portfolio theory: A bibliometric approach to decision-making under uncertainty [Epub ahead of print]. Humanities and Social Sciences Communications. https://doi.org/10.1057/s41599-026-07252-6
[19] Vuong, P. H., Phu, L. H., Nguyen, T. H. V., Duy, L. N., Bao, P. T., & Trinh, T. D. (2024). A bibliometric literature review of stock price forecasting: From statistical model to deep learning approach. *Science Progress*. https://doi.org/10.1177/00368504241236557
[20] Ahmed, S., Alshater, M. M., Ammari, A. E., & Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*. https://doi.org/10.1016/j.ribaf.2022.101646
[21] Goyal, P., & Soni, P. (2025). Stock markets volatility during crises periods: A bibliometric analysis. *Qualitative Research in Financial Markets*. https://doi.org/10.1108/qrfm-06-2023-0143
[22] Bashir, M. F. (2022). Oil price shocks, stock market returns, and volatility spillovers: A bibliometric analysis and its implications. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-021-18314-4
[23] Almeida, J., & Gonçalves, T. C. (2022). A systematic literature review of volatility and risk management on cryptocurrency investment: A methodological point of view. *Risks*. https://doi.org/10.3390/risks10050107
[24] Pečiulis, T., Ahmad, N., Menegaki, A. N., & Bibi, A. (2024). Forecasting of cryptocurrencies: Mapping trends, influential sources, and research themes. *Journal of Forecasting*. https://doi.org/10.1002/for.3114
[25] Manogna, R. L., & Anand, A. (2024). A bibliometric analysis on the application of deep learning in finance: Status, development and future directions. *Kybernetes*. https://doi.org/10.1108/k-04-2023-0637
[26] Ye, A., Maiti, A., Schmidt, M., & Pedersen, S. J. (2024). A hybrid semi-automated workflow for systematic and literature review processes with large language model analysis. *Future Internet*. https://doi.org/10.3390/fi16050167
[27] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? Computational Linguistics, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502
[28] Karjus, A. (2025). Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*. https://doi.org/10.1057/s41599-025-04503-w
[29] Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*. https://doi.org/10.1016/0304-4076(92)90064-x
[30] Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2005.01.005
[31] Christensen, K., Siggaard, M., & Veliyev, B. (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*. https://doi.org/10.1093/jjfinec/nbac020
[32] Poon, S.-H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*. https://doi.org/10.1257/002205103765762743
[33] Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*. https://doi.org/10.1007/3-540-45014-9_1
[34] Bates, J. M., & Granger, C. W. J. (1969). Journal of the Operational Research Society, 20(4), 451–468. https://doi.org/10.1057/jors.1969.103
[35] Timmermann, A. (2018). Forecasting methods in finance. *Annual Review of Financial Economics*. https://doi.org/10.1146/annurev-financial-110217-022713
[36] Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting, 29(1), 108–121. https://doi.org/10.1016/j.ijforecast.2012.06.004
[37] Cao, J., Li, Z., & Li, J. (2019). Financial time series forecasting model based on CEEMDAN and LSTM. *Physica A: Statistical Mechanics and its Applications*. https://doi.org/10.1016/j.physa.2018.11.061
[38] Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2022.11.005
[39] Hajirahimi, Z., & Khashei, M. (2022). Hybridization of hybrid structures for time series forecasting: A review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10199-0
[40] Cawood, P., & Zyl, T. V. (2022). Evaluating state-of-the-art, forecasting ensembles and meta-learning strategies for model fusion. *Forecasting*. https://doi.org/10.3390/forecast4030040
[41] Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*. https://doi.org/10.3390/publications9010012
[42] Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2017.08.007
[43] Martins, J. N., Bravo, J. M., & Martins, J. M. (2024). Mapping the scientific landscape of knowledge management in IT SMEs: A bibliometric analysis. *International Journal of Innovation and Technology Management*. https://doi.org/10.1142/s0219877024300064
[44] Marzi, G., Balzano, M., Caputo, A., & Pellegrini, M. M. (2025). Guidelines for bibliometric-systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. *International Journal of Management Reviews*. https://doi.org/10.1111/ijmr.12381
[45] Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2021.04.070
[46] Tóth, I., Lázár, Z. I., Varga, L., Járai-Szabó, F., Papp, I., Florian, R. V., & Ercsey-Ravasz, M. (2021). Mitigating ageing bias in article level metrics using citation network analysis. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2020.101105
[47] Bernasconi, E., Redavid, D., & Ferilli, S. (2025). Integrated survey classification and trend analysis via LLMs: An ensemble approach for robust literature synthesis. *Electronics*. https://doi.org/10.3390/electronics14173404
[48] Sanghera, R., Thirunavukarasu, A. J., Khoury, M. E., O'logbon, J., Chen, Y., Watt, A., Mahmood, M., Butt, H., Nishimura,

G., & Soltan, A. A. (2025). High-performance automated abstract screening with large language model ensembles. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocaf050
[49] Borovič, M., Tomovski, E., Dobnik, T. L., & Majninger, S. (2025). Evaluating proprietary and open-weight large language models as universal decimal classification recommender systems. *Applied Sciences*. https://doi.org/10.3390/app15147666
[50] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. https://doi.org/10.1177/001316446002000104
[51] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. https://doi.org/10.1037/h0031619
[52] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. https://doi.org/10.2307/2529310
[53] Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., \ Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*. https://doi.org/10.1136/bmj.n160
[54] Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*.
[55] Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.03.002
[56] Kristjanpoller, W., & Minutolo, M. C. (2016). Forecasting volatility of oil price using an artificial neural network-GARCH model. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2016.08.045
[57] Kristjanpoller, W., & Minutolo, M. C. (2015). Gold price volatility: A forecasting approach using the artificial neural network–GARCH model. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2015.04.058
[58] Patton, A. J., & Sheppard, K. (2009). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2009.01.011
[59] Lu, X., Ma, F., Xu, J., & Zhang, Z. (2022). Oil futures volatility predictability: New evidence based on machine learning models. *International Review of Financial Analysis*. https://doi.org/10.1016/j.irfa.2022.102299
[60] Zhang, Y., Ma, F., & Wei, Y. (2019). Out-of-sample prediction of the oil futures market volatility: A comparison of new and traditional combination approaches. *Energy Economics*. https://doi.org/10.1016/j.eneco.2019.05.018
[61] Aras, S. (2021). Stacking hybrid GARCH models for forecasting Bitcoin volatility. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2021.114747
[62] Lin, Y., Lin, Z., Liao, Y., Li, Y., Xu, J., & Yan, Y. (2022). Forecasting the realized volatility of stock price index: A hybrid model integrating CEEMDAN and LSTM. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2022.117736
[63] Zhang, Y., Zhang, T., & Hu, J. (2025). Forecasting stock market volatility using CNN-BiLSTM-attention model with mixed-frequency data. *Mathematics*. https://doi.org/10.3390/math13111889
[64] Koo, E., & Kim, G. (2022). A hybrid prediction model integrating GARCH models with a distribution manipulation strategy based on LSTM networks for stock market volatility. *IEEE Access*. https://doi.org/10.1109/access.2022.3163723
[65] Mishra, A. K., Renganathan, J., & Gupta, A. (2024). Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture. *Engineering Applications of Artificial Intelligence*. https://doi.org/10.1016/j.engappai.2024.108223
[66] Ribeiro, G. T., Santos, A. A. P., Mariani, V. C., & dos Santos Coelho, L. (2021). Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2021.115490
[67] Ngwaba, C. A. (2025). HAR-RV-CARMA: A Kalman filter-weighted hybrid model for enhanced volatility forecasting. *Risks*. https://doi.org/10.3390/risks13110223
[68] Kakade, K., Jain, I., & Mishra, A. K. (2022). Value-at-Risk forecasting: A hybrid ensemble learning GARCH-LSTM based approach. *Resources Policy*. https://doi.org/10.1016/j.resourpol.2022.102903
[69] Léber, D., & Egyed, B. (2025). The sentiment augmented GARCH-LSTM hybrid model for value-at-risk forecasting. *Computational Economics*. https://doi.org/10.1007/s10614-025-11042-8
[70] Wei, Y., Liu, J., Lai, X., & Hu, Y. (2017). Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? Energy Economics, 68, 141–150. https://doi.org/10.1016/j.eneco.2017.09.016
[71] Niu, Z., Wang, C., & Zhang, H. (2023). Forecasting stock market volatility with various geopolitical risks categories: New evidence from machine learning models. *International Review of Financial Analysis*. https://doi.org/10.1016/j.irfa.2023.102738
[72] Su, Y., Liang, C., Zhang, L., & Zeng, Q. (2022). Uncover the response of the U.S. grain commodity market on El Niño–Southern Oscillation. *International Review of Economics & Finance*. https://doi.org/10.1016/j.iref.2022.05.003
[73] Hong, Y., Yu, J., Su, Y., & Wang, L. (2023). Southern oscillation: Great value of its trends for forecasting crude oil spot price volatility. *International Review of Economics & Finance*. https://doi.org/10.1016/j.iref.2022.11.023
[74] Gong, X., Lai, P., He, M., & Wen, D. (2024). Climate risk and energy futures high frequency volatility prediction. *Energy*. https://doi.org/10.1016/j.energy.2024.132466
[75] Herrera, G. P., Constantino, M., Su, J. J., & Naranpanawa, A. (2022). Renewable energy stocks forecast using Twitter investor sentiment and deep learning. *Energy Economics*. https://doi.org/10.1016/j.eneco.2022.106285
[76] Afkhami, M., Cormack, L., & Ghoddusi, H. (2017). Google search keywords that best predict energy price volatility. *Energy Economics*. https://doi.org/10.1016/j.eneco.2017.07.014

[77] Fu, T., Huang, D., Feng, L., & Tang, X. (2024). More is better? The impact of predictor choice on the INE oil futures volatility forecasting. *Energy Economics*. https://doi.org/10.1016/j.eneco.2024.107540
[78] García-Medina, A., & Aguayo-Moreno, E. (2024). LSTM–GARCH hybrid model for the prediction of volatility in cryptocurrency portfolios. *Computational Economics*. https://doi.org/10.1007/s10614-023-10373-8
[79] Lu, X., Ma, F., Wang, J., & Zhu, B. (2021). Oil shocks and stock market volatility: New evidence. *Energy Economics*. https://doi.org/10.1016/j.eneco.2021.105567
[80] Liu, L., & Pan, Z. (2020). Forecasting stock market volatility: The role of technical variables. *Economic Modelling*. https://doi.org/10.1016/j.econmod.2019.03.007
[81] Zeng, H., Wu, R., Abedin, M. Z., & Ahmed, A. D. (2025). Forecasting volatility of Australian stock market applying WTC-DCA-informer framework. *Journal of Forecasting*. https://doi.org/10.1002/for.3264
[82] Qiu, Y., Qu, S., Shi, Z., & Xie, T. (2025). Predicting cryptocurrency volatility: The power of model clustering. *Economic Modelling*. https://doi.org/10.1016/j.econmod.2024.106986
[83] Degiannakis, S., & Kafousaki, E. (2025). International Journal of Forecasting, 41(4), 1559–1588. https://doi.org/10.1016/j.ijforecast.2025.01.007
[84] Mutinda, J. K., & Yong, L. (2025). Decomposition-Ensemble approach for realized volatility prediction. Computational Economics. https://doi.org/10.1007/s10614-025-11020-0