

Computational Frontiers in Volatility Forecasting: A Systematic Review of Deep Learning Approaches in Financial Markets

Mariam Ait Al^{1*}, Said Achchab², Younes Lahrichi³

ENSIAS, Mohammed V University in Rabat, Rabat, Morocco^{1,2}

The Higher Institute of Commerce and Business Administration (Groupe ISCAE), Casablanca, Morocco³

Abstract—Deep learning is now routinely applied to financial volatility forecasting, but the evidence is dispersed and has not been mapped in a way that separates what the field has established from what it has merely asserted. This study reports a PRISMA-guided systematic review of 186 peer-reviewed studies published between 2002 and 2026. Every study is located on five dimensions by a published rule table, so each assignment can be audited rather than taken on trust, and the corpus is synthesised by structured vote counting, the method that fits evidence whose reported effect sizes are not commensurable. The field is architecturally concentrated: recurrent and feedforward networks account for 70.4% of the corpus, graph-based models, the natural representation for cross-asset spillovers, for only 4.8%. The evidence base is close to uniformly positive: of the 81 studies whose direction of effect is determinable, 76 (93.8%) report their model beating its baseline, yet only 10.8% report a test of forecast-difference significance and as many benchmark against the heterogeneous autoregressive family, the standard for realized volatility. A validation against 36 open-access full texts supports the classification where abstracts state a value and indicate how far the indicators understate practice: an interpretability gap narrows under correction but holds, at 12 of 36 full texts, while the robustness gap the abstract-based coding would have supported is withdrawn. The field has advanced faster in architecture than in evaluation discipline, which defines an agenda centred on volatility-appropriate loss functions, formal forecast comparison and deployment-aware validation.

Keywords—Volatility forecasting; deep learning; realized volatility; systematic literature review; graph neural networks; financial time series

I. INTRODUCTION

Volatility is the central input to risk measurement, derivative pricing, portfolio construction and prudential supervision. The econometric tradition running from the autoregressive conditional heteroskedasticity model [1] and its generalisation [2] through asymmetric extensions [3], [4] to the realized measures [5], [6] and the heterogeneous autoregressive (HAR) model [7] has supplied the reference specifications for four decades, and they are hard to beat: that a simple GARCH(1,1) is difficult to outperform [8] remains a standing benchmark for any competing method.

Deep learning entered with a different promise: to learn the mapping to future volatility without fixing a functional form, and to absorb heterogeneous inputs in one estimator. Recurrent architectures [9], convolutional filters and attention

mechanisms [10]–[12] have all been transplanted into the problem, and published work has grown steeply since 2020. That growth has outpaced the field's capacity to say what it collectively knows: studies differ in the object forecast, the horizon, the loss function, the benchmark, and the market, so claims of superiority are rarely comparable. A review is useful only if it locates each study on explicit dimensions and reports the evidence with the qualifications the primary studies impose.

This study reports a PRISMA-guided review [13] of 186 peer-reviewed studies published between 2002 and 2026 that apply neural architectures to a volatility object in a financial market. It contributes a search and screening protocol executed programmatically, with every decision recorded against a reason code (Fig. 1); a positioning of every included study on five dimensions against a published codebook (Tables III and IV); a synthesis by structured vote counting [14], with the baselines and loss functions the corpus actually uses (Tables V and VI); and a measurement of the operational gap validated against 36 open-access full texts, which leads to the withdrawal of one gap a naive reading would have claimed (Section III-G, Table II).

II. RELATED REVIEWS AND THE CONTRIBUTION OF THIS STUDY

Volatility forecasting with neural methods is now reviewed often enough that a new review has to justify itself against the existing ones, here separated into those that share the object of this review and those adjacent to it.

A. Reviews with the Same Object

Six reviews target volatility forecasting with machine or deep learning directly. Ge et al. [15] is the closest antecedent: a systematic review in ACM Computing Surveys covering 35 studies published after 2015, whose central contribution is the diagnosis that the field's results cannot be meaningfully compared and whose remedy is a shared evaluation task. That diagnosis is the starting point of the present review, and Section VI measures how far it still holds. Gunnarsson et al. [16] cover realized and implied volatility, but pool econometric, machine learning, and AI methods, so deep architectures are not isolated. Mansilla-Lopez et al. [17] is a PRISMA review of 40 articles from 105, organised around the factors driving volatility rather than the architectures forecasting it. Qiu et al. [18] trace the historical evolution and

*Corresponding author.

contribute a benchmark comparison of landmark models from GARCH through LSTM and the Transformer. Leushuis and Petkov [19] review 32 methodologies across linear, machine learning and deep learning classes, organised by method rather than by study. Behera et al. [20] survey the area as a book chapter.

B. Adjacent Reviews

Six further reviews are frequently cited in this space but do not share the object. Ajiga et al. [21] and Khattak [22] survey artificial intelligence for financial market forecasting across heterogeneous targets, so volatility-specific questions do not arise. Ashtiani and Raahemi [23] organise by input modality rather than by target. Casolaro et al. [24] and Kong [25] survey deep learning for time series in general. Sezer et al. [26] is volatility-adjacent but covers financial time series broadly and stops in 2019, before the attention and graph literature emerged.

C. What this Review Adds

Table I compares these reviews on the dimensions that determine what a review can establish, and four differences,

each checkable against the table, define the contribution of the present one. Coverage: the corpus contains 186 studies, where the largest volatility-specific review stating a size is 40 [17], with 35 in [15] and 32 methodologies in [19]; coverage is not itself a contribution but it is a precondition for the distributional claims of Sections V to VII. A reproducible search: the query set, filters and retrieval window fully determine the result, so the identification stage can be re-run rather than trusted, and none of the reviews above supplies an executable search. Per-study positioning against a published codebook: every study is located on five dimensions by an explicit rule table, whereas the comparable reviews organise by architecture or theme and none publishes an assignment a reader can contest rule by rule. And a synthesis bounded by its own validation: the coding is checked against 36 open-access full texts, which calibrates every reported share and led to the withdrawal of one claim (Sections III-G and VII-C).

The last two matter most. Ge et al. [15] argued that this literature cannot be meaningfully compared across studies; describing more studies would inherit that problem, whereas an inspectable extraction with a measured error addresses it.

TABLE I. THIS REVIEW COMPARED WITH PRIOR REVIEWS OF VOLATILITY FORECASTING WITH MACHINE AND DEEP LEARNING

Review	Object	Corpus size	Executable search	Per-study positioning on published codebook	Evidence synthesis	Extraction error measured
Ge et al. [15]	Neural volatility forecasting	35	No	No (organised by architecture)	Narrative + call for a shared task	No
Gunnarsson et al. [16]	Realized and implied volatility	Not reported	No	No (by method class)	Narrative	No
Mansilla-Lopez et al. [17]	Stock volatility, ML	40 (of 105)	No	No (by volatility factor)	Narrative	No
Qiu et al. [18]	Volatility forecasting, historical	Not reported	No	No	Own benchmark experiment	No
Leushuis and Petkov [19]	Realized volatility methodologies	32 methodologies	No	No (by method)	Narrative	No
Behera et al. [20]	Financial market volatility, ML	Not reported	No	No	Narrative	No
Sezer et al. [26]	Financial time series (all targets)	Not reported	No	No (by architecture and task)	Narrative	No
Ashtiani and Raahemi [23]	News-based prediction	Not reported	No	No (by input modality)	Narrative	No
Khattak [22]	AI for market forecasting	Not reported	No	Partly (taxonomy of techniques)	Narrative	No
Ajiga et al. [21]	AI for stock market prediction	Not reported	No	No	Narrative	No
Casolaro et al. [24]; Kong [25]	Time series, all domains	Not reported	No	No (by architecture)	Narrative	No
This review	Volatility objects only (Section III-A)	186	Yes	Yes, five dimensions, rule table published	Structured vote counting, 81 of 186 determinable	Yes, 36 full texts (Table II)

Note. Corpus size is reported only where the review states one in a verifiable form; "not reported" means that it could not be established from the published record and is not a claim that the review is small. "Executable search" means the search can be re-run from supplied artefacts rather than reconstructed from a description.

III. METHOD

The review follows the PRISMA 2020 statement [13]. Supplementary Tables S1 to S7 are openly archived at <https://doi.org/10.5281/zenodo.23377132>.

A. Definition of the Target Object

"Volatility proxy" is used loosely in this literature, so the review fixes the extension of its target once and applies it throughout. A study is in scope if and only if its prediction target is a volatility object, a quantity measuring the second

moment or the dispersion of a financial return process, comprising exactly six classes: 1) realized volatility or variance, including realized semivariance; 2) conditional variance, as produced by a GARCH-type or stochastic volatility specification; 3) implied volatility, including volatility indices and the surface; 4) spot or instantaneous volatility, including integrated variance over an intraday window; 5) multivariate objects, including covariance and correlation matrices and spillover measures; and 6) dispersion-based risk measures driven by a volatility model, namely value-at-risk and expected shortfall.

Two consequences are applied without exception. Studies whose target is a price, a return, or a direction are out of scope even when a volatility model appears as an input. Spot and instantaneous volatility are explicitly in scope, which brings the high-frequency and cross-asset literature in rather than leaving its status undetermined. The target must additionally be estimated with a neural architecture of at least one hidden layer; econometric and shallow machine learning models are admitted only as baselines or as hybrid components.

B. Search Strategy and Eligibility Criteria

The search was executed programmatically against the Crossref REST API, which indexes the metadata of the major publishers of this literature, including Elsevier, Springer Nature, Wiley, Taylor and Francis, IEEE, Oxford University Press, MDPI and ACM; a programmatic interface is what makes the identification stage reproducible. The review issued 62 structured queries in two waves, spanning the six target classes of Section III-A crossed with architecture families, input modalities, markets and methodological concerns; a representative query is “graph neural network realized volatility forecasting spillover” and the full list is supplied as supplementary Table S1. Retrieval was restricted to journal articles published from 2000-01-01 to the date of search, capped at 100 records per query, and was last executed on 24 September 2026; because Crossref returns a relevance-ranked window that is not stable across executions, the identification set is the union of the runs, 12,400 records reducing to 4,050 unique records after deduplication by DOI. A targeted hand search of the finance, econometrics, and AI-in-finance literature contributed 3 further studies.

A record was included if it was peer-reviewed, in English, published between 2000 and the date of search in an outlet indexed in Scopus or Web of Science, forecast one of the six volatility classes of Section III-A with a neural architecture of at least one hidden layer, and reported an empirical evaluation against at least one baseline. Each excluded record carries exactly one reason code, E1 to E8, the indexing criterion being E6 and preprints or working papers E5; the codes and their counts are given in Fig. 1 and the per-record decisions in supplementary Table S7. Studies confined to a single cryptocurrency market are screened against the same criteria rather than excluded on that ground, since their volatility literature uses the same targets and architectures as the rest of the corpus.

C. Study Selection

Screening proceeded in two stages. At the title and abstract stage, a record was retained only if its title or abstract evidenced a volatility object, a neural architecture and a forecasting task; 3,616 records were excluded here, leaving 434 for eligibility assessment. At that stage, each of these 434 records was assessed on its full bibliographic record and abstract, and received an explicit include or exclude decision with a reason code. Of these, 251 were excluded and 183 retained, to which the 3 hand-searched studies were added, giving a final corpus of 186 studies. Screening decisions were taken by the authors without independent duplicate screening, so no inter-rater statistic is reported; in its place, every decision is recorded against a record and a reason code. Coding, by

contrast, applies the published rule table automatically (Section III-D), so it is deterministic, and its error is measured against full texts in Section III-G. Fig. 1 reports the flow. Supplementary Tables S1 to S7 are archived as one workbook (Section III), with S3 being the complete list of the 186 included studies with DOIs and their full coding. The reference list cites the studies discussed individually; the corpus is 186 throughout.

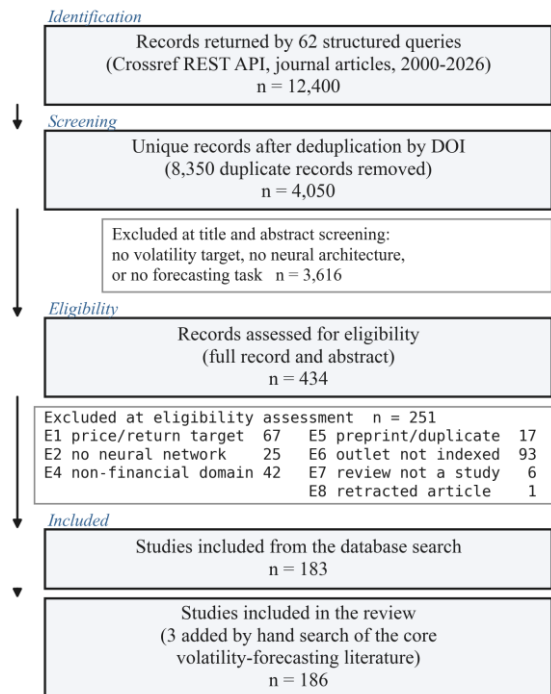


Fig. 1. PRISMA flow of the review. Counts are generated from the screening decision table, so each stage reconciles exactly with the next. Eligibility was assessed on the full bibliographic record and abstract of each candidate (Section III-C).

D. Data Extraction and the Multidimensional Codebook

Each study was positioned on five dimensions: model family, assigning one primary architecture under a priority ordering that resolves hybrids to their most specific component, with a separate flag for hybrid or ensemble designs; volatility target, the six classes of Section III-A; forecast resolution, distinguishing intraday, daily and lower-frequency settings, with multi-horizon evaluation recorded as a separate binary variable because the two are orthogonal; input modality, distinguishing price-only, engineered, order-book, macro-financial, textual and alternative web-derived inputs; and market context, distinguishing equity indices, individual equities, foreign exchange, commodity and energy markets, options and derivatives, and multi-asset settings including cryptocurrency.

Coding applies an ordered rule table to the title and abstract of each record. That rule table is the codebook and is published verbatim with the supplementary material, so any reader can re-derive every cell of Tables III to VI and contest a stated rule rather than an unobservable judgment. It trades fine-grained discretion for reproducibility, and the cost is stated in Section III-H.

E. Evidence Synthesis Strategy

No pooled effect size is reported, for a substantive reason. Volatility is latent, so every reported error metric is computed against a proxy whose measurement error contaminates model rankings unless the loss is robust to it [27]; and the corpus varies in target class, sampling frequency, market, sample period, and loss function, so a mean error reduction would aggregate quantities that are not the same quantity. The review therefore applies structured vote counting by direction of effect, the standard method when pooling is not defensible [14]. Each study is coded as reporting a gain, a conditional gain, no gain, or as not determinable; studies without a retrievable abstract are coded as not determinable and never counted as a win or a loss. The baseline classes and evaluation metrics for each study are recorded separately, because what a literature benchmarks against and measures is itself evidence about its evaluation discipline.

F. Practice Indicators and Bibliometric Mapping

The review reports five practice indicators as counts rather than as a composite score: whether interpretability is addressed, whether forecast uncertainty is quantified, whether performance is evaluated in a stress or regime-shift period, whether validation spans more than one market, and whether the study is openly accessible. A sixth records whether a formal test of forecast-difference significance [28], [29] is reported; results appear in Table VI. Separately, reference lists were retrieved for 170 of the 186 included studies and a co-citation network [30] was built over references cited by at least four of them, with communities extracted by greedy modularity maximisation [31] and interpreted in Section IV-B.

G. Validation of the Coding Against Full Texts

Coding from abstracts is defensible only if its error is measured rather than assumed. The open-access full text was retrieved for every included study for which one could be obtained, 36 of 186, and that sub-sample was re-coded; it is not random but selected by open-access retrievability, which is stated wherever its results are used. Three results follow, reported in Table II. Forecast resolution agrees with the full text in 7 of the 9 cases where both determine it, too few cases to support an interval. The abstract is silent far more often than it is wrong: resolution is determinable from 9 of 36 abstracts but from 34 of 36 full texts. And, most consequentially, the operational indicators understate full-text practice by factors running from 1.6 for significance testing and 2.0 for the HAR benchmark to 4.0 for stress-period evaluation, with multi-horizon evaluation understated by more than the table can quantify because only 1 abstract in the sub-sample flags it; Section VII uses these to separate the gaps that survive validation from the one that does not.

H. Limitations of the Protocol

Four limitations qualify what follows. Coding is derived from titles and abstracts, and Section III-G measures the consequence rather than assuming it: agreement rests on too few cases to establish a rate, and forecast resolution is undetermined for 75.8% of the corpus (Table III). All operational indicators are therefore floors, with an understatement that is large and uneven, from 1.6 times for significance testing to 4.0 for stress-period evaluation; Section

VII draws conclusions only where the floor is itself informative. The validation sub-sample is small and non-random, and its factors rest on counts as small as 7 against 11, so they indicate direction and order of magnitude rather than a precise correction. No risk-of-bias instrument was applied to the included studies: the practice indicators of Section III-F serve that function, and they measure reporting rather than conduct. Coverage is bounded by design: retrieval used a single index, restricted to journal articles and capped at 100 records per query, so preprints are excluded and conference papers, where much deep learning work first appears, are under-represented, while the indexing criterion of Section III-B trades coverage for a quality floor and may under-represent regional literatures.

TABLE II. VALIDATION OF THE CODING AGAINST OPEN-ACCESS FULL TEXTS

Quantity	From title and abstract	From full text	Factor
Forecast resolution determinable	9 / 36	34 / 36	-
Forecast resolution: codings agree where both determinable	7 / 9	-	-
Multi-horizon evaluation	1 / 36	13 / 36	n/a
Forecast-significance test reported	7 / 36	11 / 36	×1.6
HAR-family benchmark used	7 / 36	14 / 36	×2.0
Interpretability addressed	0 / 36	12 / 36	n/a
Stress or crisis-period evaluation	6 / 36	24 / 36	×4.0
QLIKE loss function used	1 / 36	6 / 36	n/a
Code publicly available	not measurable	4 / 36 (11%)	-
Data availability statement	not measurable	10 / 36 (28%)	-

Note. Sub-sample of 36 studies, selected by open-access retrievability and therefore not random. The factor is the within-sample understatement ratio, the full-text count to the abstract-based count for the same 36 studies, and is left undefined (n/a) where that count is zero or too small to carry precision; it calibrates the corpus-level shares in Tables III and VI rather than correcting them. The interpretability factor of 6.2 quoted in Section VII-A is not computed from this table: it is a cross-sample ratio, the full-text share here (33%) to the abstract-based share in the whole corpus of 186 studies (5.4%), and is indicative only.

IV. PROFILE OF THE CORPUS

A. Growth and Outlets

The corpus spans 2002 to 2026, and its growth is recent and steep (Fig. 2): 156 studies, 83.9% of the corpus, appeared from 2021 onwards; the median year of inclusion is 2024; and only 6 studies (3.2%) predate 2015, the earliest being a recurrent-network model of currency volatility [32]. The four leading outlets are Journal of Forecasting (23), Expert Systems with Applications (13), Journal of Risk and Financial Management (9) and IEEE Access (7); the corpus is dispersed across 88 outlets, 60 of which contribute a single study, together 32.3% of the corpus.

B. Intellectual Structure

Fig. 3 maps the co-citation structure, and the two communities it returns are not arbitrary. The first is organised around the conditional-variance tradition and the hybrid neural-GARCH studies extending it; the second around the realized-volatility tradition and the deep learning work targeting

realized measures. The second also contains the reference connecting them, the long short-term memory architecture itself.

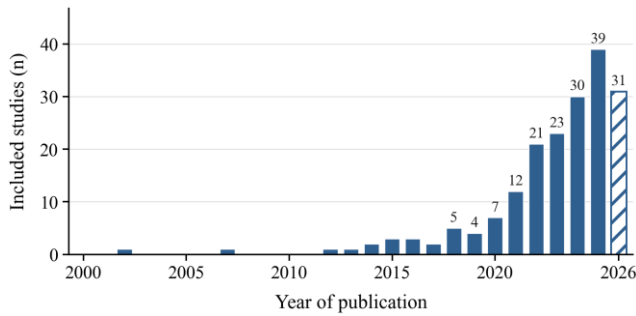


Fig. 2. Annual number of included studies, 2000-2026 ($n = 186$). The 2026 bar is hatched because that year is incomplete: the search closed in September, so it covers nine months.

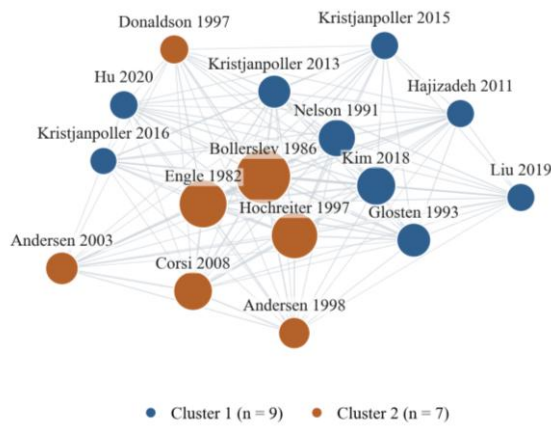


Fig. 3. Co-citation network of references cited by at least 4 included studies (16 nodes shown). Node area is proportional to citation frequency within the corpus, edge weight to co-citation count; colours mark communities from greedy modularity maximisation.

The reading is that the field has two intellectual bases that have not merged: papers benchmarking against GARCH rarely benchmark against HAR, and most papers benchmarking against HAR do not benchmark against GARCH. That separation reappears in Table V, where the GARCH family is a baseline in 44.1% of studies and the HAR family in only 10.8%, and it is the bibliometric counterpart of the evaluation problem documented in Section VI.

V. THE MULTIDIMENSIONAL MAP OF THE CORPUS

Every study occupies one cell on each of the five dimensions; the per-study assignment is supplied as Table S3, and the aggregate distributions are reported in Tables III and IV.

A. Architecture is Concentrated and Interacts with the Target

Recurrent and feedforward networks together account for 70.4% of the corpus, convolutional models for 9.7% and attention-based models for 10.8%, the last despite their prominence in the wider time-series literature [11], [12]. Graph neural networks account for 4.8% and deep reinforcement learning for 2.7%. Hybrid or ensemble designs, usually a neural network combined with a GARCH-type component or a

decomposition step, account for 66.1% and are the most characteristic design choice of the field.

Table IV, Panel A, crosses model family with volatility target. Conditional variance dominates (56.5%): it is the modal target of every family but one, often reached through GARCH-hybrid designs. Realized volatility (23.1%) is over-represented among graph-based studies, reflecting its origin in high-frequency econometrics, and implied volatility (8.6%) among convolutional ones, a consequence of its object: the surface is a two-dimensional grid, so convolutional inductive bias transfers naturally [33], [34]. Graph models are the only family whose modal target is realized as multivariate volatility, the expected signature of architectures built for cross-asset dependence.

TABLE III. DISTRIBUTION OF THE CORPUS ACROSS VOLATILITY TARGET, FORECAST RESOLUTION AND MARKET ($N = 186$)

Dimension	Category	Studies (n)	Share (%)
Volatility target	Realized volatility	43	23.1
	Conditional variance	105	56.5
	Implied volatility	16	8.6
	Spot / intraday volatility	5	2.7
	Multivariate/covariance	6	3.2
Forecast resolution	Downside risk (VaR / ES)	11	5.9
	Intraday/high-frequency	25	13.4
	Daily	20	10.8
Market context	Not specified in abstract	141	75.8
	Equity index	96	51.6
	Individual equities	3	1.6
	Commodity/energy	39	21.0
	Foreign exchange	20	10.8
	Options/derivatives	22	11.8
	Cryptocurrency (multi-asset)	6	3.2

Note. Percentages are of the full corpus. "Not specified in abstract" records that the value could not be determined from the abstract; it is reported rather than imputed. That row measures abstract reporting rather than practice: in the validation sub-sample the resolution is determinable from 9 of 36 abstracts but from 34 of 36 full texts (Table II, Section III-G).

B. Inputs Remain Overwhelmingly Price-Based

Panel B of Table IV crosses model family with input modality: price or return series alone feed 64.5% of the corpus, textual and sentiment inputs 9.1%, macro-financial inputs 7.0%, and high-frequency or order-book inputs 9.7%. The promise most often used to motivate deep learning here, its capacity to fuse heterogeneous inputs, is therefore exercised by a minority, and not by one family: textual inputs appear in graph-based, attention-based and recurrent designs alike [35]-[37].

C. Resolution, Horizon and Market

Intraday settings account for 13.4% of the corpus and daily settings for 10.8%, with the resolution undetermined from the abstract for 75.8%, a property of abstracts rather than of the field (Section III-G and the note to Table III). Multi-horizon evaluation is recorded separately because it is orthogonal to resolution; it is flagged from abstracts in 6.5% of the corpus

but appears in 13 of 36 full texts against 1 of 36 abstracts, so the review does not treat it as a gap.

Equity indices dominate the market dimension at 51.6%, followed by commodity and energy markets at 21.0%, options

and derivatives at 11.8%, and foreign exchange at 10.8%. Only 1.6% of the corpus works at the level of individual equities, which matters because cross-sectional dispersion and spillover are where multivariate and graph-based methods have an advantage to demonstrate.

TABLE IV. MODEL FAMILY BY VOLATILITY TARGET AND BY INPUT MODALITY (COUNTS)

Panel A. Model family by volatility target							
Model family	Realized volatility	Conditional variance	Implied volatility	Spot / intraday volatility	Multivariate/covariance	Downside risk (VaR / ES)	Total
Feedforward / MLP	18	41	6	3	1	1	70
Recurrent (LSTM / GRU)	14	34	3	1	2	7	61
Convolutional (CNN / TCN)	4	8	4	0	1	1	18
Transformer/attention	2	14	2	0	0	2	20
Graph neural network	4	2	0	1	2	0	9
Deep reinforcement learning	1	4	0	0	0	0	5
Generative representation /	0	2	1	0	0	0	3
Total	43	105	16	5	6	11	186
Panel B. Model family by input modality							
Model family	Price/return only	Technical / engineered	High-frequency / LOB	Macro-financial	Textual / sentiment	Alternative/web	Total
Feedforward / MLP	48	7	6	5	4	0	70
Recurrent (LSTM / GRU)	39	6	6	3	7	0	61
Convolutional (CNN / TCN)	12	2	2	1	1	0	18
Transformer/attention	11	2	2	2	2	1	20
Graph neural network	6	0	1	0	2	0	9
Deep reinforcement learning	2	0	1	1	1	0	5
Generative representation /	2	0	0	1	0	0	3
Total	120	17	18	13	17	1	186

VI. SYNTHESIS OF COMPARATIVE EVIDENCE

A. Direction of Effect

Of the 186 included studies, the direction of the comparative result is determinable for 81, and of those, 76 (93.8%) report the proposed neural specification outperforming its baseline, 2 report a gain conditional on regime, horizon, or market, and 3 report no gain or mixed evidence; Table V gives the breakdown by model family. The vote count therefore rests on fewer than half of the corpus: the other 105 studies state no direction in their abstract or have no retrievable abstract, and are excluded rather than imputed. A success rate of 93.8% is a finding about the literature rather than about deep learning: a published record in which essentially every comparison favours the proposed method cannot be read as an unbiased estimate of expected performance. Three mechanisms suffice to produce this without any individual study being at fault: tuning the proposed model on the evaluation sample while leaving the baseline at default settings; the lower probability that negative results are submitted and accepted; and the composition of the determinable subset itself, since an abstract

that states its direction of effect is more likely to do so when that direction is favourable. The third means the share is an upper bound on the proportion of the corpus reporting a gain rather than an estimate of it. Requiring a reported comparison (Section III-B), without which there is no direction to count, does not select on the result, but it confines the corpus to published comparative work, where the second mechanism operates. The appropriate inference is not that deep learning does not work, but that the magnitude of its advantage cannot be estimated from this corpus.

Whether the share is an artefact of corpus composition can be checked directly, because every record carries the query that retrieved it: the sub-corpus reachable by the first search wave is recoverable exactly, 157 studies against 186, 16% smaller and differently composed. On it, the share reporting a gain is 93.0%, against 93.8% on the full corpus, a shift of 0.8 points, small beside the distance from either figure to an even split. The check bears on the breadth of the search, not on selection: both corpora apply the same inclusion criteria, so neither can show how often unpublished work fails to beat its baseline.

TABLE V. DIRECTION-OF-EFFECT VOTE COUNTING BY MODEL FAMILY

Model family	Studies (n)	Direction of effect determinable (n)	Gain over baseline reported (n, % of determinable)	Conditional or no gain (n)	GARCH-family baseline (n, %)	HAR-family baseline (n, %)	Forecast-significance test reported (n, %)
Feedforward / MLP	70	24	23 (95.8%)	1	29 (41%)	5 (7%)	7 (10%)
Recurrent (LSTM / GRU)	61	28	25 (89.3%)	3	38 (62%)	9 (15%)	7 (11%)
Convolutional (CNN / TCN)	18	10	10 (100.0%)	0	7 (39%)	1 (6%)	0 (0%)
Transformer/attention	20	11	11 (100.0%)	0	6 (30%)	3 (15%)	5 (25%)
Graph neural network	9	4	3	1	1 (11%)	1 (11%)	1 (11%)
Deep reinforcement learning	5	2	2	0	1 (20%)	0 (0%)	0 (0%)
Generative representation /	3	2	2	0	0	1	0
ALL	186	81	76 (93.8%)	5	82 (44%)	20 (11%)	20 (11%)

Note. Percentages in the gain column are of the determinable subset, not of the family total. Studies whose direction of effect cannot be established are excluded from that subset rather than assigned a value. Shares are given only where the base reaches five studies; below that, the count stands alone.

B. Benchmarks and Loss Functions

The GARCH family appears as a baseline in 44.1% of the corpus. The HAR family, which is the standard benchmark for realized volatility [7] and the model any realized-volatility forecast must beat to be interesting, appears in only 10.8%. Restricting attention to the 43 studies whose target is realized volatility raises the rate more than threefold but leaves it a minority: the HAR family is named by 16 of them. A substantial part of the corpus is therefore demonstrating superiority over a benchmark that the target literature does not regard as the relevant one.

The loss functions are equally revealing. Mean squared error, mean absolute error and root mean squared error dominate the metrics named across the corpus, at 15.1%, 12.9% and 13.4% of studies respectively. QLIKE, which together with MSE is one of the two loss functions robust to the fact that volatility is measured with error [27], is named by 5 studies. Formal tests of forecast-difference significance [28], [29] are reported by 10.8%. The combination—a proxy-sensitive loss function, no significance test, and a benchmark drawn from a different tradition—means that a large share of the reported improvements in this corpus is not established at conventional standards of forecast evaluation.

These three shares are abstract-based and therefore floors. The validation of Section III-G puts numbers on the gap: in the full-text sub-sample the HAR benchmark is used by 14 of 36 studies rather than 7, forecast-significance testing by 11 of 36 rather than 7, and QLIKE by 6 of 36 rather than 1. The conclusion survives, at the corrected magnitude: even on full texts, a volatility-appropriate loss function is used by one study in six, a formal forecast comparison by fewer than one in three, and the standard realized-volatility benchmark by fewer than half. Whether that discipline moderates the reported advantage cannot be tested on direction of effect: among the determinable studies, those using the HAR benchmark, QLIKE or a significance test report a gain in all or all but one case, as do nearly all of the others. The test needs effect sizes on a common loss, which the corpus does not report commensurably (Section III-E), and is left to the agenda of Section IX.

VII. OPERATIONAL DIMENSIONS

Table VI reports the practice indicators by model family. Every value is a lower bound whose looseness Section III-G measures, so each gap below is reported at its validated magnitude where one exists, and withdrawn where full-text practice proves common.

TABLE VI. METHODOLOGICAL AND OPERATIONAL PRACTICE INDICATORS BY MODEL FAMILY

Model family	Studies (n)	Interpretability (n, %)	Uncertainty or density (n, %)	Stress/crisis evaluation (n, %)	Cross-market validation (n, %)	Multi-horizon evaluation (n, %)	Forecast-significance test (n, %)	Open access (n, %)
Feedforward / MLP	70	2 (3%)	3 (4%)	12 (17%)	9 (13%)	3 (4%)	7 (10%)	35 (50%)
Recurrent (LSTM / GRU)	61	2 (3%)	3 (5%)	10 (16%)	5 (8%)	3 (5%)	7 (11%)	37 (61%)
Convolutional (CNN / TCN)	18	1 (6%)	1 (6%)	1 (6%)	5 (28%)	2 (11%)	0 (0%)	13 (72%)
Transformer/attention	20	2 (10%)	0 (0%)	7 (35%)	4 (20%)	3 (15%)	5 (25%)	15 (75%)
Graph neural network	9	3 (33%)	0 (0%)	4 (44%)	3 (33%)	1 (11%)	1 (11%)	4 (44%)
Deep reinforcement learning	5	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (60%)
Generative / representation	3	0	0	1	2	0	0	3
ALL	186	10 (5%)	7 (4%)	35 (19%)	28 (15%)	12 (6%)	20 (11%)	110 (59%)

Note. All indicators are coded from titles and abstracts and are lower bounds on true practice; Table II reports the measured understatement factors. Shares are given only where the family reaches five studies. Open access is a dissemination indicator and is not equivalent to code or data availability, which is reported separately in Section VII-D.

A. Interpretability

Interpretability is addressed in 5.4% of the corpus. Post-hoc attribution [38], [39] is only one of the senses in which the ten studies [40]-[49] use the term: others explain graph models through their learned structure or design architectures that keep an econometric reading. Graph-based studies report it most often, at 3 of 9, which is plausible: an attention coefficient over an explicit asset graph has a direct economic reading as a spillover intensity, whereas the same coefficient over an unstructured input does not. Supervisory guidance on advanced analytics in regulated institutions treats explainability as a requirement rather than a preference [50], which makes this an operational constraint.

No indicator is more completely absent from abstracts, so the claim must be stated at the corrected magnitude: interpretability is addressed in 12 of the 36 validated full texts but flagged in none of their abstracts, which leaves its within-sample factor undefined (Table II). Set against the 5.4% read from abstracts across the corpus, the full-text share of 33% gives a factor of 6.2, a cross-sample ratio that is indicative only. A gap remains: two studies in three in a domain whose forecasts feed regulated processes, but it is far narrower than the abstract-based coding suggests.

B. Uncertainty Quantification

Forecast uncertainty is quantified in 3.8% of the corpus, through quantile regression [51], [52] and probabilistic forecasting [41]. The omission is substantive rather than stylistic: a volatility forecast is an input to a risk calculation, and a point forecast without a predictive distribution cannot support the interval and tail statements such calculations require [53]. The value-at-risk and expected shortfall studies of target class (vi) are no exception: none of the 11 carries the indicator, and where their abstracts state the mechanism, the risk measure is derived from a point volatility forecast, under an assumed conditional distribution [49], [54] or by filtered historical simulation [54], not read from a quantile or distributional output of the network. The few studies that do address uncertainty show the machinery is available and simply not used.

C. Robustness Under Stress: A Claim the Validation Withdraws

Stress-period evaluation is the one indicator on which the review reverses a conclusion. Coded from abstracts, evaluation in a stress, crisis, or regime-shift period appears in 18.8% of the corpus, which would suggest that the setting where volatility forecasts matter most is tested by roughly one study in five. The validation of Section III-G does not support that reading: the same indicator appears in 24 of 36 full texts, an understatement factor of 4.0, the largest within-sample factor in Table II. Crisis sub-samples are a routine part of the empirical design here but rarely a selling point, so they reach the results section and not the abstract. This review therefore cannot establish a robustness gap and makes no such claim. What the corpus does show is a difference in emphasis: attention-based and graph-based families are the most likely to foreground stress evaluation in their abstracts (7 of 20 and 4 of 9),

consistent with architectures introduced to handle regime-dependent and spillover-driven dynamics. Cross-market validation (15.1% from abstracts) was not validated and is reported as a floor only.

D. Reproducibility

Code availability cannot be established from bibliographic records, so no corpus-level figure is reported. Among the 36 studies whose full text could be retrieved, 4 (11%) point to publicly available code and 10 (28%) carry a data availability statement; because that sub-sample is selected by open-access retrievability, it is biased towards openly disseminated work, so 11% should be interpreted as an upper bound rather than as an estimate. Open access itself, at 59.1%, is a dissemination indicator, not a reproducibility one. The structural point holds independently: the hybrid and ensemble designs that are 66.1% of the literature involve decomposition steps, tuned components, and multi-stage pipelines whose replication depends on detail that published methods sections rarely carry. They also carry a validity risk this review can name but not measure: a wavelet, empirical-mode or variational-mode decomposition, or a GARCH filter, fitted to the whole series before the train-test split lets the evaluation period shape the inputs, a look-ahead that inflates out-of-sample accuracy. The codebook records neither the step nor its position relative to the split, so its prevalence is left to Section IX.

VIII. WHAT THE CORPUS SHOWS, BY MODEL FAMILY

A. Feedforward, Recurrent, Convolutional and Attention Models

Feedforward and recurrent networks are the backbone of the corpus at 70.4% of studies, and their trajectory is one of progressive hybridisation rather than architectural advance: early work pairs a feedforward network with a GARCH-type variance equation [55], [56] and the design recurs with a recurrent core [57], [58], remaining dominant in the most recent years. Their reported gains are rarely tested, about one study in ten reporting a significance test against one in four among attention models; their most credible results come from the studies benchmarking against HAR on realized measures rather than against GARCH [59], [60].

Convolutional models (9.7%) serve two distinct purposes: dilated causal convolution over a one-dimensional volatility series, as in DeepVol [61] and the temporal convolutional approaches to volatility and value-at-risk [62], and genuine two-dimensional structure in the implied volatility surface [34]. The second is the stronger case, because there the inductive bias matches the object. Attention models (10.8%) arrive late, 17 of 20 from 2024 onwards, and are distinguished by the settings they make tractable: multi-horizon evaluation [63], multimodal designs combining text with market data [36], and an explicit attempt to keep econometric structure inside an attention model [64]. Deep reinforcement learning (2.7%) and generative approaches (1.6%) remain small, the former optimising the pipeline [65] rather than forecasting directly, the latter using adversarial and autoencoder structures for augmentation and denoising [66].

B. Graph Neural Networks

Graph-based models account for 4.8% of the corpus, which understates their relevance: volatility is a multivariate object with time-varying cross-asset spillover, a structure a graph encodes explicitly. The 9 studies follow four main lines. The first forecasts multivariate realized volatility from limit order book data over a relational market graph [67]. The second models spillover directly, through a spillover index embedded in a spatio-temporal graph [68] or by building the graph from estimated nonlinear spillovers [69]. The third targets spot volatility, with SpotV2Net using Fourier estimates of spot volatility and of volatility of volatility as node and edge features [43], and the HAR-type graph model retaining the HAR structure while letting the graph carry cross-sectional information [44]. The fourth fuses graph structure with sentiment and attention data [35]. How the graph is built matters most under stress, when the dependence it encodes is itself what changes. The explicitly time-varying designs all date from 2025 onwards: edges carrying spot vol-of-vol estimates [43], evolving graph-structure learners [46], a dynamic spillover-index graph [70] and a time-varying transfer-entropy network [71]. The codebook does not record graph construction, so the share of static graphs is not quantified.

This family also leads the corpus on two indicators that matter: interpretability and stress-period evaluation (Table VI). With nine studies coded from abstracts, that profile is suggestive rather than established, and the case for the family is structural rather than empirical: an architecture matched to the multivariate structure of the object, whose attention coefficients carry an economic reading as spillover intensities, is what a well-posed deep learning study of volatility calls for. Whether it also delivers larger or more robust gains is a question for the effect-size synthesis proposed in Section IX.

IX. RESEARCH AGENDA

Each item below names the measurement it responds to. Benchmark against the right model: only 10.8% of the corpus, and 16 of the 43 realized-volatility studies, use the HAR family, yet a realized-volatility forecast that does not beat HAR is not useful. Use a proxy-robust loss and test the difference: QLIKE is named by 5 studies and a formal forecast comparison by 10.8%, so MSE or QLIKE with a Diebold-Mariano or model confidence set result should be the minimum standard. Report predictive distributions: uncertainty is quantified in 3.8% of studies, while the downstream use of a volatility forecast is almost always an interval or a tail statement.

Exploit cross-sectional structure: only 1.6% of the corpus works at individual equities, and graph-based models remain 4.8% of studies despite leading the corpus on interpretability and stress-period evaluation. Make interpretability structural rather than post-hoc, as the graph literature does, by building models whose learned quantities carry an economic reading; this is reported in 5.4% of abstracts. Close the reproducibility gap: even in the open-access sub-sample, biased towards openly disseminated work, only 4 of 36 studies point to available code. Finally, widen and deepen the evidence base: a search of Scopus, Web of Science, IEEE Xplore and arXiv

with duplicate screening would test the coverage limits of Section III-H, and a full-text re-coding of the corpus would settle what Section III-G can only bound. That re-coding should extract effect sizes on a common loss, to test whether evaluation discipline moderates the reported advantage, and three variables abstracts rarely carry: how a risk measure is derived from the volatility forecast, whether the graph is static or time-varying, and where any decomposition step sits relative to the train-test split.

X. CONCLUSION

This review mapped 186 peer-reviewed studies of deep learning for financial volatility forecasting onto five explicit dimensions and synthesised their comparative evidence under the constraints that evidence imposes. Three conclusions follow. The field is architecturally concentrated and input-conservative: recurrent and feedforward hybrids dominate, 64.5% of studies use price data alone, and the architectures matched to the multivariate structure of volatility remain a small minority. Its published evidence is close to uniformly positive, with 76 of 81 determinable comparisons favouring the proposed model, while the apparatus that would make such claims credible- the HAR benchmark, a proxy-robust loss, and a significance test- is present in a minority of cases, even on full texts. And where operational practice can be validated, the picture is uneven rather than uniformly thin: correction leaves interpretability addressed by a third of full texts and available code by 4 of 36, whereas robustness testing is common practice that abstracts do not report.

The implication is not that deep learning has failed here, but that the field's methodological development has outpaced its evaluation discipline: the binding constraint on progress is no longer architectural novelty but the willingness to test new models against the benchmarks, loss functions and market conditions that determine whether a volatility forecast is worth deploying.

DECLARATION ON GENERATIVE AI

During the preparation of this manuscript, the authors used generative AI tools to assist with drafting, editing, and language refinement of the text. No generative AI tool was used to generate, simulate, or analyse any data, or to produce any numerical result. The authors reviewed and edited all content and take full responsibility for the content of this publication.

REFERENCES

- [1] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987-1007, 1982.
- [2] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307-327, 1986.
- [3] D. B. Nelson, "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica*, vol. 59, no. 2, pp. 347-370, 1991.
- [4] L. R. Glosten, R. Jagannathan, and D. E. Runkle, "On the relation between the expected value and the volatility of the nominal excess return on stocks," *J. Finance*, vol. 48, no. 5, pp. 1779-1801, 1993.
- [5] T. G. Andersen and T. Bollerslev, "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *Int. Econ. Rev.*, vol. 39, no. 4, pp. 885-905, 1998.

- [6] T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys, "Modeling and forecasting realized volatility," *Econometrica*, vol. 71, no. 2, pp. 579-625, 2003.
- [7] F. Corsi, "A simple approximate long-memory model of realized volatility," *J. Financial Econometrics*, vol. 7, no. 2, pp. 174-196, 2009.
- [8] P. R. Hansen and A. Lunde, "A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?," *J. Appl. Econometrics*, vol. 20, no. 7, pp. 873-889, 2005.
- [9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [10] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000-6010.
- [11] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748-1764, 2021.
- [12] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 11106-11115.
- [13] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
- [14] J. E. McKenzie and S. E. Brennan, "Synthesizing and presenting findings using other methods," in *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed., Wiley, 2019, pp. 321-347.
- [15] W. Ge, P. Lalbakhsh, L. Isai, A. Lenskiy, and H. Suominen, "Neural network-based financial volatility forecasting: A systematic review," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1-30, 2022, doi: 10.1145/3483596.
- [16] E. S. Gunnarsson, H. R. Isern, A. Kaloudis, M. Ristad, B. Vigdel, and S. Westgaard, "Prediction of realized volatility and implied volatility indices using AI and machine learning: A review," *Int. Rev. Financial Anal.*, vol. 93, p. 103221, 2024.
- [17] J. Mansilla-Lopez, D. Mauricio, and A. Narvaez, "Factors, forecasts, and simulations of volatility in the stock market using machine learning," *J. Risk Financial Manag.*, vol. 18, no. 5, p. 227, 2025, doi: 10.3390/jrfm18050227.
- [18] Z. Qiu, C. Kownatzki, F. Scalzo, and E. S. Cha, "Historical perspectives in volatility forecasting methods with machine learning," *Risks*, vol. 13, no. 5, p. 98, 2025, doi: 10.3390/risks13050098.
- [19] R. M. Leushuis and N. Petkov, "Advances in forecasting realized volatility: A review of methodologies," *Financial Innovation*, vol. 12, no. 14, pp. 1-29, 2026, doi: 10.1186/s40854-025-00809-5.
- [20] I. Behera, P. Nanda, S. Mitra, and S. Kumari, "Machine learning approaches for forecasting financial market volatility," in *Machine Learning Approaches in Financial Analytics, Intelligent Systems Reference Library*, Springer, 2024, pp. 431-451, doi: 10.1007/978-3-031-61037-0_20.
- [21] D. Ajiga, R. A. Adeleye, O. F. Asuzu, O. R. Owolabi, B. G. Bello, and N. L. Ndubuisi, "Review of AI techniques in financial forecasting: Applications in stock market analysis," *Finance Account. Res. J.*, vol. 6, no. 2, pp. 125-145, 2024.
- [22] B. H. A. Khattak, "A systematic survey of AI models in financial market forecasting for profitability analysis," *IEEE Access*, vol. 11, pp. 125359-125380, 2023.
- [23] M. N. Ashtiani and B. Raahemi, "News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review," *Expert Syst. Appl.*, vol. 217, p. 119509, 2023.
- [24] A. Casolaro, V. Capone, G. Iannuzzo, and F. Camastra, "Deep learning for time series forecasting: Advances and open problems," *Information*, vol. 14, no. 11, p. 598, 2023.
- [25] X. Kong, "Deep learning for time series forecasting: A survey," *Int. J. Mach. Learn. Cybern.*, 2025, doi: 10.1007/s13042-025-02560-w.
- [26] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005-2019," *Appl. Soft Comput.*, vol. 90, p. 106181, 2020.
- [27] A. J. Patton, "Volatility forecast comparison using imperfect volatility proxies," *J. Econometrics*, vol. 160, no. 1, pp. 246-256, 2011.
- [28] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Stat.*, vol. 13, no. 3, pp. 253-263, 1995.
- [29] P. R. Hansen, A. Lunde, and J. M. Nason, "The model confidence set," *Econometrica*, vol. 79, no. 2, pp. 453-497, 2011.
- [30] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *J. Amer. Soc. Inf. Sci.*, vol. 24, no. 4, pp. 265-269, 1973.
- [31] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, 2004.
- [32] C. L. Dunis and X. Huang, "Forecasting and trading currency volatility: an application of recurrent neural regression and model combination," *J. Forecasting*, vol. 21, no. 5, pp. 317-354, 2002.
- [33] J. Choi and D. W. Shin, "Parallel architecture of CNN-bidirectional LSTMs for implied volatility forecast," *J. Forecasting*, vol. 41, no. 6, pp. 1087-1098, 2022.
- [34] H. Shao, B. Zhou, and S. Gong, "Prediction of the implied volatility surface—An empirical analysis of the SSE 50ETF option based on CNNs," *Finance Res. Lett.*, vol. 77, p. 107119, 2025.
- [35] B. Lei and Y. Song, "Volatility forecasting for stock market incorporating media reports, investors' sentiment, and attention based on MTGNN model," *J. Forecasting*, vol. 43, no. 5, pp. 1706-1730, 2024.
- [36] Y. Song et al., "Volatility and value-at-risk forecasting using BERT and transformer models incorporating investors' textual sentiments," *Finance Res. Lett.*, vol. 85, p. 108210, 2025.
- [37] Y. Li et al., "The role of news sentiment in oil futures returns and volatility forecasting: Data-decomposition based deep learning approach," *Energy Econ.*, vol. 95, p. 105140, 2021.
- [38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4768-4777.
- [39] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135-1144.
- [40] J. Amerić, T. Poklepović, and Z. Aljinović, "GARCH-based artificial neural networks in forecasting conditional variance of stock returns," *Croat. Oper. Res. Rev.*, vol. 5, no. 2, pp. 329-343, 2014.
- [41] H. Xia et al., "A new probability forecasting model for cotton yarn futures price volatility with explainable AI and big data," *J. Forecasting*, vol. 44, no. 1, pp. 112-135, 2024.
- [42] W. Zhou and D. Luo, "Decomposing, Learning, and Predicting Realized Volatilities: A Comparison Analysis From the Global Stock Markets," *J. Forecasting*, vol. 45, no. 1, pp. 135-155, 2025.
- [43] A. Brini and G. Toscano, "SpotV2Net: Multivariate intraday spot volatility forecasting via vol-of-vol-informed graph attention networks," *Int. J. Forecasting*, vol. 41, no. 3, pp. 1093-1111, 2025.
- [44] N. Hu, X. Yin, and Y. Yao, "A novel HAR-type realized volatility forecasting model using graph neural network," *Int. Rev. Financial Anal.*, vol. 98, p. 103881, 2025.
- [45] D. Herremans and K. W. Low, "Forecasting Bitcoin Volatility Spikes From Whale Transactions and CryptoQuant Data Using Synthesizer Transformer Models," *IEEE Access*, vol. 13, pp. 117788-117807, 2025.
- [46] Y. Zhou et al., "Forecasting cryptocurrency volatility: a novel framework based on the evolving multiscale graph neural network," *Financial Innov.*, vol. 11, no. 1, 2025.
- [47] S. Liu et al., "Volatility prediction of high-frequency data based on deep feature extraction and fusion: An interpretable deep learning architecture," *Appl. Soft Comput.*, vol. 189, p. 114535, 2026.
- [48] S. Agal, "A hybrid deep learning framework for volatility prediction in financial markets," *Scientific Reports*, 2026.
- [49] N. Acevedo-Prins and J. D. Velásquez, "ARCH-LSTM Structural Equivalence with Hybrid Student-t Likelihood Loss for Financial Volatility Forecasting," *J. Risk Financ. Manag.*, vol. 19, no. 9, p. 688, 2026.
- [50] European Banking Authority, "EBA report on big data and advanced analytics," EBA/REP/2020/01, Jan. 2020.

- [51] D. Pradeepkumar and V. Ravi, "Forecasting financial time series volatility using Particle Swarm Optimization trained Quantile Regression Neural Network," *Appl. Soft Comput.*, vol. 58, pp. 35-52, 2017.
- [52] J. Song et al., "Forecasting realized volatility using deep learning quantile function," *Appl. Soft Comput.*, vol. 175, p. 113016, 2025.
- [53] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359-378, 2007.
- [54] K. Kakade, I. Jain, and A. K. Mishra, "Value-at-Risk forecasting: A hybrid ensemble learning GARCH-LSTM based approach," *Resour. Policy*, vol. 78, p. 102903, 2022.
- [55] W. Kristjanpoller and M. C. Minutolo, "Gold price volatility: A forecasting approach using the Artificial Neural Network-GARCH model," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7245-7251, 2015.
- [56] W. Kristjanpoller and M. C. Minutolo, "A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis," *Expert Syst. Appl.*, vol. 109, pp. 1-11, 2018.
- [57] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models," *Expert Syst. Appl.*, vol. 103, pp. 25-37, 2018.
- [58] Y. Liu, "Novel volatility forecasting using deep learning-Long Short-Term Memory Recurrent Neural Networks," *Expert Syst. Appl.*, vol. 132, pp. 99-109, 2019.
- [59] A. Bucci, "Realized Volatility Forecasting with Neural Networks," *J. Financial Econometrics*, vol. 18, no. 3, pp. 502-531, 2020.
- [60] K. Christensen, M. Siggaard, and B. Veliyev, "A Machine Learning Approach to Volatility Forecasting," *J. Financial Econometrics*, vol. 21, no. 5, pp. 1680-1727, 2022.
- [61] F. Moreno-Pino and S. Zohren, "DeepVol: volatility forecasting from high-frequency data with dilated causal convolutions," *Quant. Finance*, vol. 24, no. 8, pp. 1105-1127, 2024.
- [62] C. X. Zhang et al., "Forecasting stock volatility and value-at-risk based on temporal convolutional networks," *Expert Syst. Appl.*, vol. 207, p. 117951, 2022.
- [63] A. K. Mishra, J. Renganathan, and A. Gupta, "Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108223, 2024.
- [64] L. Petrosino et al., "A GARCH-temporal fusion transformer model for the volatility prediction of exchange-traded funds," *Neural Comput. Appl.*, vol. 37, no. 26, pp. 21435-21458, 2025.
- [65] Y. Yu et al., "Novel optimization approach for realized volatility forecast of stock price index based on deep reinforcement learning model," *Expert Syst. Appl.*, vol. 233, p. 120880, 2023.
- [66] A. H. A. Gadhi, S. Peiris, and D. E. Allen, "Improving Volatility Forecasting: A Study through Hybrid Deep Learning Methods with WGAN," *J. Risk Financ. Manag.*, vol. 17, no. 9, p. 380, 2024.
- [67] Q. Chen and C. Y. Robert, "Multivariate Realized Volatility Forecasting with Graph Neural Network," *Proc. ACM ICAIF*, pp. 156-164, 2022.
- [68] B. Son et al., "Forecasting global stock market volatility: The impact of volatility spillover index in spatial-temporal graph-based model," *J. Forecasting*, vol. 42, no. 7, pp. 1539-1559, 2023.
- [69] C. Zhang et al., "Forecasting realized volatility with spillover effects: Perspectives from graph neural networks," *Int. J. Forecasting*, vol. 41, no. 1, pp. 377-397, 2025.
- [70] Z. Chi, J. Gao, and C. Wang, "Global Stock Market Volatility Forecasting Incorporating Dynamic Graphs and All Trading Days," *J. Forecasting*, vol. 45, no. 3, pp. 1311-1324, 2025.
- [71] J. Yu, L. Wang, and X. Sun, "An Entropy-Regularised AI Framework for Multi-Asset Volatility Spillover Forecasting and CVaR-Constrained Portfolio Allocation in Financial Markets," *Entropy*, vol. 28, no. 7, p. 756, 2026.