Online ISSN: 3104-9338
Print ISSN: 3104-932X

A Review of Stock Index Forecasting Methods from ARIMA to Time-Series Foundation Models

Li Su*

* Chengdu University of Information Technology

Received: April 15, 2026

Revised: April 20, 2026

Accepted: April 25, 2026

Published online: April 26, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 5 (May 2026)

* Corresponding Author: Author Name (suli3607@foxmail.com)

Abstract. Stock index forecasting has evolved from linear statistical baselines to hybrid deep neural architectures and, more recently, to large-scale time-series foundation models. This review synthesizes the development path represented by the supplied literature, covering ARIMA, GARCH, and VAR models; classical machine learning methods such as random forests and boosting; recurrent, convolutional, and attention-based deep learning models; decomposition-driven hybrids; selective state space models; and emerging large-model approaches for time-series analysis. The review is organized around the inductive biases that different model families impose on financial data, with special attention to nonstationarity, volatility clustering, multimodal information fusion, and distribution shift. Compared with generic forecasting domains, stock index prediction places stronger demands on robustness, interpretability, and economic usefulness because signal-to-noise ratios are low and model errors can be magnified by trading decisions. Across the surveyed studies, no single architecture dominates all settings; instead, performance depends on how well a method aligns with data frequency, exogenous information, market regime, and evaluation objective. The review concludes that future progress is likely to come from financially informed hybrid systems, stronger benchmark design, and better integration between domain-specific supervision and foundation-model pretraining.

Keywords: *Stock Index Forecasting; Financial Time Series; Deep Learning; Transformers; State Space Models; Foundation Models*

1. Introduction

Forecasting stock indexes remains one of the most demanding tasks in applied time-series

analysis because market prices aggregate macroeconomic information, firm-level expectations, policy shocks, liquidity conditions, and investor behavior in a continuously changing environment. Unlike many engineering forecasting tasks, financial prediction must cope with weak explanatory signals, abrupt regime shifts, nonlinear dependence, volatility clustering, and the possibility that model-aware traders alter the very dynamics being modeled. For this reason, stock index forecasting has long served as both a practical problem and a stress test for forecasting methodology. The recent comparison of ARIMA models across the S&P 500, FTSE, and SSEC by Xu [33] illustrates that even straightforward autoregressive baselines can still provide useful diagnostic insight when the purpose is to understand persistence, trend structure, and market-specific differences rather than to claim universal predictive superiority.

The historical foundation of the field is statistical. Bollerslev's generalized autoregressive conditional heteroskedasticity model [2] formalized time-varying volatility in a way that remains central to risk-aware financial modeling, while Sims's vector autoregressive perspective [28] established a flexible multivariate framework for studying dynamic interactions without imposing overly restrictive structural assumptions at the outset. These models made explicit two enduring lessons for stock forecasting research: first, mean dynamics and variance dynamics should not be conflated; second, model usefulness depends on whether the analyst seeks point prediction, volatility estimation, structural interpretation, or policy-sensitive scenario analysis. Even when later machine learning methods outperform classical baselines on accuracy metrics, the interpretability and diagnostic clarity of statistical models remain indispensable.

The next methodological wave imported ideas from machine learning and sequence modeling. Random forests [3] and boosting [10] showed that nonlinear prediction could be improved by ensembling relatively simple learners, especially when inputs were expanded through technical indicators, lagged returns, or handcrafted macro features. Recurrent neural networks, especially long short-term memory networks [14] and gated encoder-decoder variants [6], then offered a more flexible approach to temporal dependence by learning representations directly from sequences rather than relying entirely on manually engineered summary statistics. At a broader methodological level, Bai et al. [1] challenged the assumption that recurrence is always the natural tool for sequence modeling, while the transformer architecture of Vaswani et al. [29] made attention-based sequence representation a dominant paradigm across machine learning.

Finance-specific studies reflect this progression. Fischer and Krauss [9] demonstrated that LSTM networks could extract useful predictive structure from cross-sectional financial data,

while CNNpred [15] broadened the input space by combining price, technical, and cross-market variables in a convolutional framework. Subsequent comparative and hybrid studies [11,12,13,19,21,23,24,27,34] explored richer combinations of recurrent units, convolutions, attention mechanisms, decomposition techniques, sentiment information, and cross-scale feature fusion. These works collectively suggest that financial forecasting accuracy often depends less on raw model depth than on whether the architecture matches the heterogeneous temporal scales, exogenous drivers, and nonstationary distributions of market data.

The architecture of a LSTM is illustrated in Figure 1, which shows the cell state pathway and the three gating mechanisms—forget, input, and output—that enable the network to selectively retain or discard information over long temporal horizons.

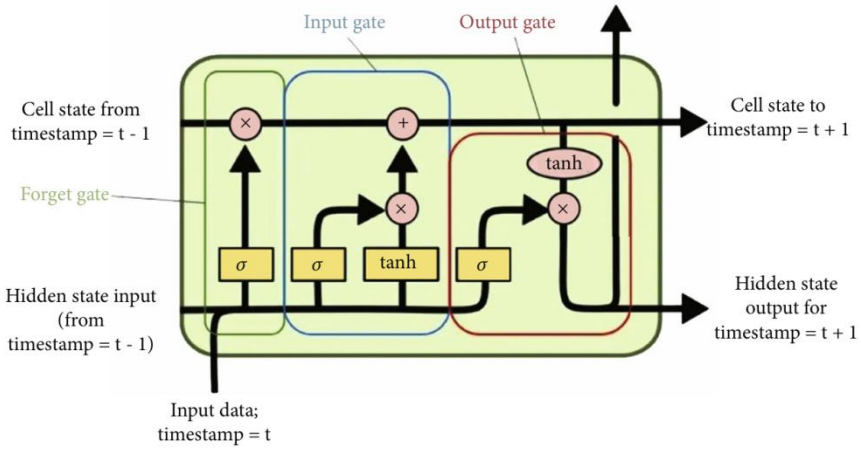

Figure 1. Architecture of a LSTM.

A further acceleration has occurred in recent years with the arrival of general-purpose time-series architectures and foundation-model thinking. Attention modules such as CBAM [31], normalization strategies such as RevIN [18], modern forecasting architectures including TimesNet [32], PatchTST [25], TiDE [7], TSMixer [5], iTransformer [20], TimeMixer [30], and SAMformer [16], and large-model approaches such as Time-LLM [17], TEMPO [4], Timer [22], decoder-only foundation models [8], and Lag-Llama [26] have expanded the design space well beyond the recurrent-versus-convolutional debate. This review examines that broader landscape from the standpoint of stock index forecasting, emphasizing inductive bias, data requirements, robustness under shift, and practical deployability rather than leaderboard-style comparison alone.

The review deliberately places finance-specific papers alongside general-purpose forecasting models because the two literatures are now tightly connected. Many of the newest stock prediction studies are not built from wholly finance-native architectures; instead, they adapt

design ideas first validated on broader time-series benchmarks and then retrofit them to market data. This creates both an opportunity and a methodological risk. The opportunity is that finance can benefit from rapid progress in sequence modeling without rediscovering every architectural idea locally. The risk is that a model may inherit benchmark success from domains whose statistical properties differ sharply from those of financial returns. A useful survey must therefore examine not only what a model is, but also why its original design assumptions may or may not survive transfer into stock index forecasting.

2. From Statistical Benchmarks to Classical Machine Learning

2.1. Statistical Models: ARIMA, GARCH, and VAR

ARIMA remains the canonical entry point for financial forecasting because it imposes a transparent decomposition of a series into autoregressive, differencing, and moving-average components. Its main strength is not that it captures every market irregularity, but that it provides a disciplined baseline against which more complex models must justify their additional complexity. Xu [33] used ARIMA to compare three major equity indexes with different institutional and regional characteristics, reinforcing the practical value of model identification, stationarity testing, and residual diagnostics before moving to heavier architectures. In review terms, ARIMA serves as the benchmark for evaluating whether newer models genuinely learn nonlinear temporal structure or merely exploit sample-specific drift that can disappear out of sample.

However, stock indexes rarely behave like homoskedastic linear processes. Periods of calm and turbulence alternate, and the conditional variance often exhibits stronger persistence than the conditional mean. Bollerslev's GARCH model [2] addressed this issue by modeling volatility as its own dynamic process, thereby improving both risk estimation and the interpretation of forecast uncertainty. For stock index research, this distinction is crucial because a model that predicts returns poorly may still be useful for forecasting volatility, tail risk, or interval width. Many later neural architectures implicitly attempt to capture similar changing variance patterns, but they often do so without the explicit probabilistic structure that made GARCH attractive to financial econometrics.

VAR modeling introduced by Sims [28] is equally important when the objective is to forecast an index using interacting macroeconomic or cross-market variables rather than the target series alone. Its contribution lies in recognizing that financial prices do not evolve in isolation; exchange rates, interest rates, commodity prices, and foreign market indexes may all transmit

information through lagged interactions. Although modern deep learning systems can represent multivariate dependence more flexibly, the VAR framework still supplies the conceptual basis for exogenous-variable modeling, impulse-style reasoning, and scenario-aware forecasting. For review purposes, VAR is best understood as an early formalization of multivariate temporal coupling rather than merely a classical baseline.

Taken together, ARIMA, GARCH, and VAR establish three baseline questions that remain relevant in modern stock index forecasting. Is the predictive task primarily about linear mean reversion or trend continuation? Does volatility dynamics carry more stable information than level dynamics? And how much of the relevant signal is endogenous to the target series versus imported from correlated variables? Any new model family that cannot answer these questions more effectively than statistical benchmarks may offer sophistication without decision value. That is why the best contemporary studies still compare against econometric baselines even when the final winning architecture is nonlinear.

There is also a broader methodological virtue in preserving statistical baselines within modern experimental pipelines. Because ARIMA, GARCH, and VAR have relatively transparent failure modes, they help reveal whether a forecasting problem is intrinsically weak-signal or whether the issue lies in model misspecification. If a sophisticated deep architecture cannot outperform a carefully tuned linear or heteroskedastic benchmark, the result may indicate not that the deep model is poorly implemented, but that the available data do not support the complexity being imposed. In finance, where overfitting is easy and economic regimes are unstable, this diagnostic role is as important as the baseline score itself.

2.2. Classical Machine Learning and Feature-Centric Forecasting

Classical machine learning shifted the emphasis from explicit stochastic assumptions to flexible nonlinear decision boundaries. Random forests [3] are attractive in financial forecasting because they handle mixed feature types, nonlinear interactions, and moderate feature redundancy without heavy preprocessing, while still supporting variable-importance style diagnostics. Boosting methods such as AdaBoost [10] provide a different route to stronger prediction by repeatedly reweighting difficult examples and combining weak learners into a high-capacity ensemble. In stock index settings, these models are especially useful when the researcher constructs features from multiple horizons, technical indicators, macro releases, and sentiment summaries, because the feature space itself becomes the main source of predictive power.

The limitation of this feature-centric paradigm is that performance depends heavily on manual representation design. If the handcrafted features fail to expose the relevant temporal dependency, even strong nonlinear learners cannot recover the missing structure. Moreover, tree ensembles and boosted learners do not natively distinguish between stable long-run information and fast transient shocks unless those distinctions are already encoded in the features. For stock indexes, where the same raw price path may support multiple meaningful temporal views such as intraday momentum, weekly reversal, and crisis-regime persistence, this becomes a serious constraint. The transition to deep learning can therefore be seen not as a rejection of classical machine learning, but as an attempt to automate representation construction while retaining nonlinear predictive capacity.

Even so, classical machine learning continues to play a vital role in the landscape of modern literature on stock index forecasting. Firstly, it establishes competitive baselines that are often significantly stronger than naive linear models, providing a valuable point of reference for assessing more complex methodologies. Secondly, classical techniques remain particularly practical in low-data settings where deep learning models may be prone to overfitting due to insufficient training data. Thirdly, these traditional methods can be effectively integrated into hybrid workflows; for example, decomposition techniques can be employed to produce multi-scale components, allowing simpler learners to be assigned to different frequency bands for enhanced performance. For a comprehensive review of stock index forecasting, the main lesson learned is that model choice should thoughtfully reflect both the scale and representation of the available data: when domain knowledge leads to the development of highly informative features, classical ensemble methods can still prove challenging to outperform on a robustness-adjusted basis. This underlines the enduring relevance of classical approaches in contemporary predictive analytics and their potential to complement more advanced modeling techniques.

Classical machine learning methods highlight an important issue that continues to be relevant for contemporary deep learning models: the choice of target formulation. Some studies focus on predicting raw index levels, while others emphasize forecasting returns, and yet others simplify the task to direction classification. Techniques such as tree ensembles and boosting methods make this distinction explicit because they can be effectively trained as either regressors or classifiers with relatively minor architectural adjustments. The same degree of flexibility is indeed possible in the realm of deep learning; however, the existing literature does not always clearly state whether a given model is optimized specifically for error minimization, directional hit rate, or for facilitating downstream portfolio decisions. From an academic review

standpoint, this ambiguity is significant because architectures should only be compared when they are addressing the same forecasting problem under comparable loss functions. Consequently, researchers must exercise caution in their comparisons to ensure validity and relevance in performance evaluations.

3. Deep Neural Forecasting Architectures in Finance

3.1. Recurrent Sequence Models and Their Financial Adoption

Deep learning entered stock forecasting through recurrent sequence models because recurrence offers a direct mechanism for processing ordered observations and retaining temporal context. LSTM [14] addressed vanishing gradients through gated memory, making it possible to learn longer-range dependencies than standard recurrent networks. GRU-style encoder-decoder models [6] simplified the recurrent machinery while preserving gating behavior, and they encouraged the view that sequence-to-sequence learning could be adapted beyond language processing. In finance, Fischer and Krauss [9] provided an influential demonstration that LSTM networks could learn useful predictive regularities from stock-related inputs, helping establish recurrent neural networks as credible tools for market prediction rather than purely experimental imports from other domains.

The appeal of recurrent models in stock index forecasting is straightforward and well-founded: financial markets exhibit path dependence, meaning that the relevance of recent information often hinges on what transpired earlier in the same sequence. However, it is essential to note that recurrence is not inherently well matched to all financial tasks across various contexts. For instance, challenges such as training instability, the computational cost associated with sequential processing, and difficulties in effectively capturing very long contexts can become significant drawbacks when researchers transition from analyzing low-frequency end-of-day data to working with larger multivariate time windows. This limitation is one reason why the work by Bai et al. [1] became so pivotal in the broader literature: their empirical comparison demonstrated that generic convolutional sequence models could rival or even surpass recurrent networks on a wide range of sequence tasks. This finding served to weaken the prevailing belief that Recurrent Neural Networks (RNNs) were the inevitable default choice for handling time-dependent data, encouraging further exploration of alternative modeling approaches in financial forecasting.

Figure 2 depicts the architecture of a standard RNN, where the recurrent hidden-state connections allow previous time-step information to flow forward, providing the model with a

basic form of memory for sequential data.

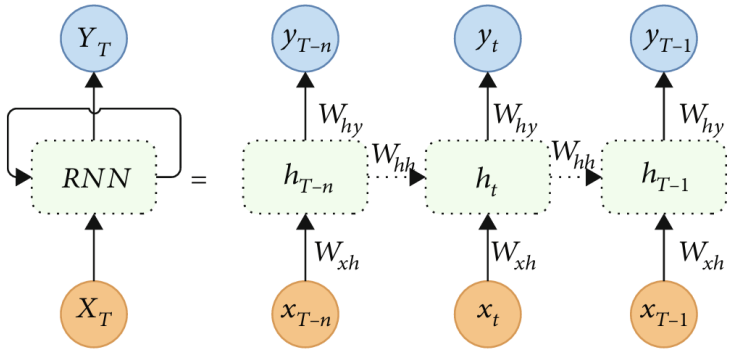

Figure 2. Architecture of RNN.

From a review perspective, recurrent models should be seen as a transitional breakthrough rather than the endpoint of deep financial forecasting. They established the feasibility of representation learning on raw temporal inputs, reduced dependence on manual lag design, and opened the door to richer hybrids with convolution, attention, and exogenous encoders. At the same time, their limitations made it natural for later work to search for architectures with more explicit multi-scale extraction, parallel computation, or stronger handling of distribution shift. The subsequent finance literature can therefore be interpreted as a sequence of attempts to preserve the temporal sensitivity of recurrence while addressing its computational and representational bottlenecks.

3.2. Convolutional and Attention-Augmented Financial Models

One of the earliest finance-specific steps beyond plain recurrence was to diversify the input representation. CNNpred [15] is significant because it used convolutional processing over a diverse set of variables, emphasizing that stock market prediction is rarely a univariate task in practice. By combining price-related features with broader market information, CNNpred implicitly treated forecasting as a structured representation problem rather than a simple autoregression problem. This insight remains central today: accuracy gains often arise not from replacing one sequence backbone with another in isolation, but from expanding the information channels the model can align across time.

Comparative studies help separate durable patterns from architecture-specific enthusiasm. The comparative review by He, Zhang, and Por [13] reflects an important stage in the literature, where different deep learning families are evaluated not simply by peak accuracy but by their sensitivity to data characteristics, parameterization, and training setup. Such comparative work matters because the stock forecasting literature is otherwise vulnerable to fragmented claims

across different markets, horizons, and preprocessing pipelines. A finance review that only reports the best result of each paper can easily exaggerate methodological progress; comparative studies instead remind us that many reported improvements are conditional on particular datasets and design choices.

As shown in Figure 3, a BiLSTM extends the unidirectional LSTM by processing the input sequence in both forward and backward directions, enabling the network to capture context from both past and future time steps simultaneously.

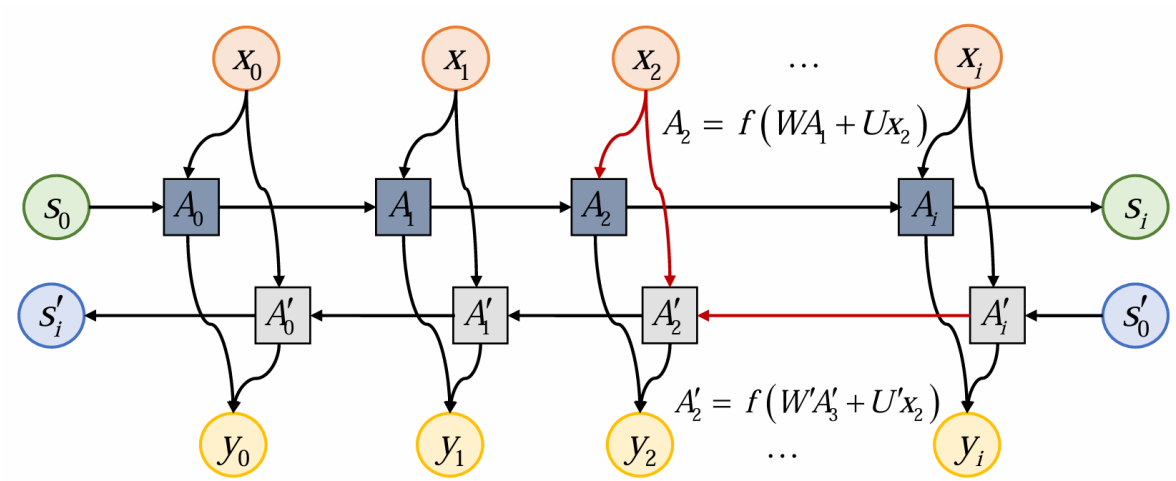

Figure 3. Architecture of a BiLSTM.

Attention-enhanced recurrent hybrids represent another significant theme in the evolution of forecasting models. In their recent work, Zhang, Ye, and Lai [34] innovatively combined Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory networks (BiLSTM), and attention mechanisms in an effort to capture local patterns, bidirectional temporal structures, and adaptive feature weighting within a single cohesive pipeline. Similarly, Mu et al. [23] proposed a spatiotemporal attention BiLSTM model specifically aimed at improving stock index prediction. This reflects a growing recognition among researchers that valuable signals may be distributed across both temporal positions and feature dimensions, necessitating more sophisticated modeling techniques. These advanced architectures are conceptually appealing because stock indexes are shaped by layered dependencies: short-term fluctuations, medium-term trend segments, and cross-variable interactions do not contribute equally at every time step. As a result, employing a uniform hidden representation can prove to be suboptimal for accurately capturing the complexities of financial data. By integrating these various elements, such models hold promise for enhancing predictive performance and providing deeper insights into market dynamics.

Figure 4 presents the internal structure of a long short-term memory neural network, highlighting the cell state and the three gating mechanisms that together address the vanishing gradient problem inherent in vanilla recurrent architectures.

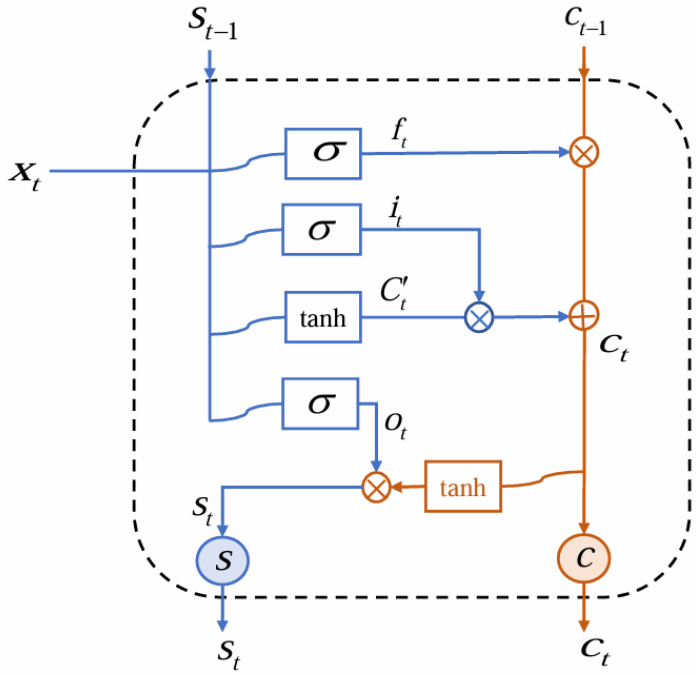

Figure 4. Long short-term memory neural network.

The incorporation of exogenous textual information is particularly notable in the innovative news-driven index prediction framework developed by Liu, Ge, and Gu [20]. By effectively combining a trellis-style temporal network with sentiment attention mechanisms, this model addresses a core weakness inherent in traditional price-only forecasting approaches: many market moves are influenced less by endogenous chart patterns and more by external narrative shocks, such as significant news events. Furthermore, news-based systems compel the field to confront complex alignment issues between the timing of textual information, the corresponding market response windows, and the challenges associated with noisy sentiment extraction. As a result, these advanced models are valuable not only for their potential gains in predictive accuracy but also for expanding the conceptual scope of stock index forecasting. They shift the focus from mere numerical extrapolation to a more nuanced form of multimodal event-aware reasoning that incorporates various types of information, thereby enhancing our understanding of market dynamics and investor behavior. This progression signifies an important step toward integrating qualitative insights into quantitative financial analysis.

The STBL architecture is visualized in Figure 5, where the spatial and temporal branches are combined through a fusion module to jointly model cross-sectional and temporal dependencies

in financial time-series data.

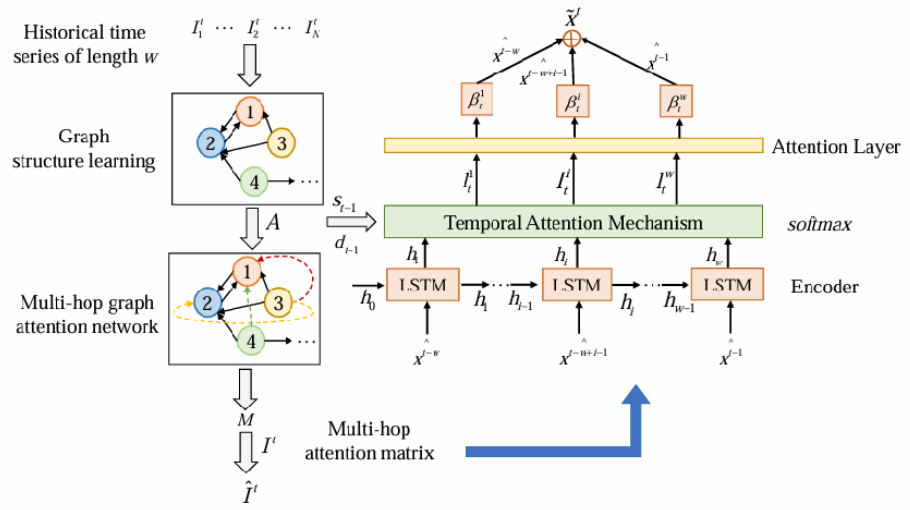

Figure 5. Visualization of the STBL architecture.

Attention-augmented financial models nonetheless require careful interpretation. In many papers, attention improves performance because it acts as a flexible weighting mechanism over already useful representations, not because it provides transparent economic explanations. A high attention score on a time step or feature channel should not automatically be read as causal market importance. This caveat is particularly relevant in stock index forecasting, where correlated signals, common shocks, and preprocessing artifacts can all produce seemingly meaningful weighting patterns. Future work should therefore distinguish more clearly between attention as an optimization device and attention as an interpretability claim.

3.3. Hybrid Decomposition Models and Selective State Spaces

A recurring conclusion in financial forecasting is that one architecture rarely handles every temporal scale equally well. This observation has motivated hybrid systems that explicitly decompose or route information before prediction. Ge [11], for example, proposed a hybrid model for forecasting S&P 500 and CSI 300 futures prices, reflecting the broader tendency to combine complementary modules rather than rely on a single backbone. The logic is economically sensible: low-frequency trend, medium-frequency cyclical movement, and high-frequency noise may each demand different inductive biases, and forcing one monolithic model to resolve all of them at once can reduce robustness.

Decomposition-based hybrids represent one of the most active branches of recent stock index research. Li et al. [19] combined CEEMDAN with a TCN-GRU-CBAM forecasting stack, while Mutinda and Geletu [24] used CEEMDAN with LSTM and BPNN components in an

ensemble decomposition model. Although the exact modules differ, the underlying hypothesis is shared: financial series become easier to model after they are separated into components with more homogeneous temporal properties. This is a practical response to nonstationarity because decomposition can isolate oscillatory behavior, slow trend, and residual noise before deep learning is applied. The price paid for this improvement is a more elaborate pipeline, more hyperparameters, and additional opportunities for data leakage if decomposition is not performed strictly within the training split.

Figure 6 illustrates the construction framework of the CEEMDAN-TCN-GRU-CBAM model, in which the original price series is first decomposed by CEEMDAN, then each sub-series is independently modeled by a TCN-GRU encoder with CBAM attention before final reconstruction.

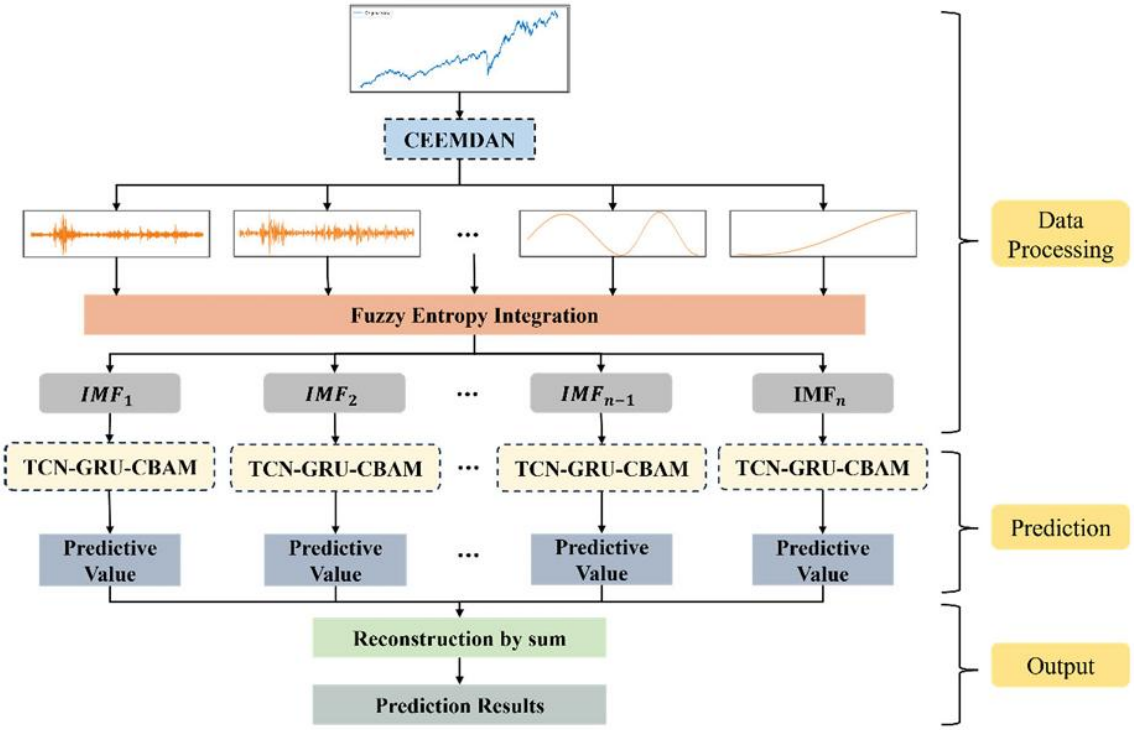

Figure 6. Construction framework of the CEEMDAN-TCN-GRU-CBAM model.

Attention modules also became increasingly modular in this hybrid literature. CBAM [31], though originally proposed in computer vision, offers a lightweight way to recalibrate channel-wise and spatial feature emphasis, and its adoption in financial hybrids signals a methodological pattern that extends beyond one paper: researchers are willing to import compact attention mechanisms when they can improve feature prioritization without the full computational burden of transformer-style global attention. In finance, such modules are especially attractive when multivariate inputs are noisy and only a subset of channels is informative under a given regime.

The specific steps of CEEMDAN decomposition are shown in Figure 7, where white Gaussian noise is added and iteratively averaged across ensemble trials to extract intrinsic mode functions from the original signal.

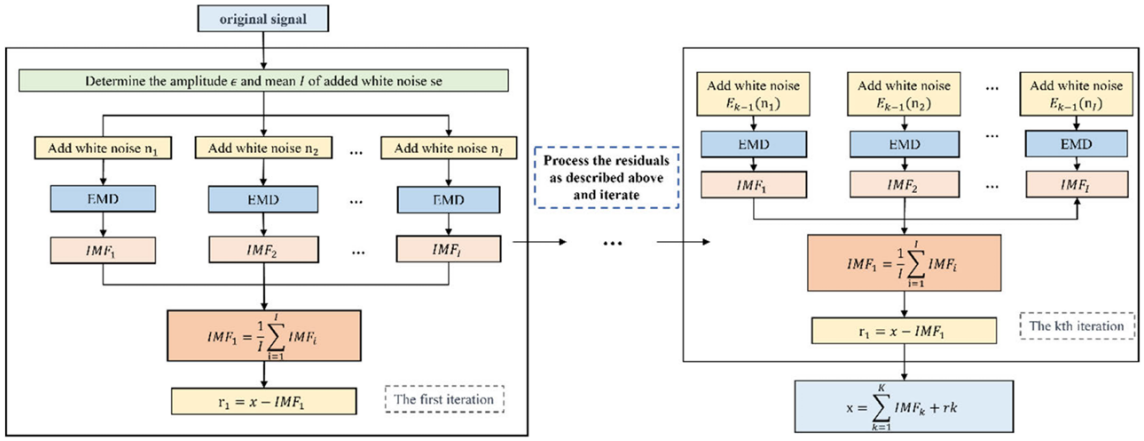

Figure 7. Specific steps of CEEMDAN decomposition.

Selective state space modeling introduces a different response to the limitations of both recurrence and full attention. Mamba [12] proposed linear-time sequence modeling through selective state spaces, offering a mechanism for content-aware state updates without quadratic attention cost. Its significance for stock forecasting lies less in immediate benchmark dominance than in architectural principle: if long-horizon dependence can be modeled through efficient selective transitions, then financial forecasting may scale to longer contexts and richer multivariate streams without the memory burden of standard transformers. Shi's MambaStock [27] is therefore noteworthy as an early domain-specific adaptation, showing how general state space advances can be specialized for stock prediction tasks.

Across these hybrid and state-space studies, the main trend is clear. Financial forecasting research is moving away from single-mechanism architectures toward modular systems that separate representation by frequency, modality, or state-update logic. The stronger these systems become, however, the more important benchmark discipline becomes as well. A complicated hybrid can outperform a plain LSTM or ARIMA in a fixed experiment and still fail to generalize under a different horizon, market, or crisis regime. For that reason, architecture design should be judged jointly with evaluation design, an issue returned to in later sections.

The decomposition literature also reminds researchers that reported gains should be interpreted in relation to task difficulty and data geography. For example, the CEEMDAN-LSTM-BPNN study of Mutinda and Geletu [24] focuses on the DAX index, while other hybrid studies consider different markets and targets [11,19]. An architecture that benefits a European

benchmark under one sampling frequency may not translate directly to U.S. or Chinese indexes under another. Consequently, the most defensible conclusion is not that one decomposition hybrid has solved stock forecasting, but that decomposition is a useful strategy for handling nonstationary mixtures when the evaluation pipeline is transparent and the forecast horizon is clearly defined.

4. General Time-Series Modeling Advances Relevant to Financial Forecasting

4.1. Attention, Normalization, and Representation Beyond Plain Recurrence

The transformer era changed stock forecasting indirectly before it changed it directly. Vaswani et al. [29] showed that attention could replace recurrence in sequence representation, enabling parallel training and long-range dependency modeling through adaptive token-to-token interaction. Although vanilla transformers are not automatically ideal for financial data, the shift in design philosophy was profound: sequence modeling no longer had to be organized around recurrent hidden-state propagation. This encouraged a wave of forecasting architectures that treated time windows as token sets, patch collections, or channel-specific sequences rather than as strictly recurrent streams.

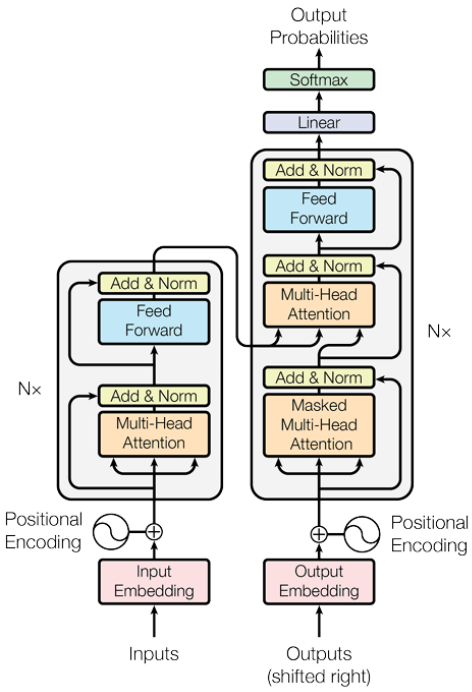

Figure 8. The Transformer model architecture.

Figure 8 shows the Transformer model architecture proposed by Vaswani et al. [29], where

multi-head self-attention replaces recurrence entirely, allowing every position in the sequence to attend to every other position in parallel.

At the same time, two practical concerns became impossible to ignore in time-series forecasting: selective feature emphasis and distribution shift. CBAM [31] offered a compact attention mechanism that many applied models could embed inside larger stacks, while RevIN [18] directly targeted distribution shift by normalizing each instance and reversing the normalization after prediction. RevIN is especially relevant to stock indexes because financial series frequently undergo level, variance, and seasonal changes across regimes. A model that performs well under one volatility regime may degrade sharply when the marginal distribution shifts, so normalization should be treated as a substantive design choice rather than a preprocessing afterthought.

4.2. Modern Forecasting Backbones: TimesNet, PatchTST, TiDE, TSMixer, iTransformer, TimeMixer, and SAMformer

A major branch of recent research focuses on designing backbones that encode temporal variation more efficiently and with stronger structural bias than generic sequence models. TimesNet [32] treats temporal patterns through two-dimensional variation modeling, reflecting the idea that multi-periodicity and local pattern repetition can be captured more naturally when time is reorganized into richer structural views. For stock indexes, where daily, weekly, and event-driven periodicities interact, this kind of representation is attractive because it avoids assuming that all useful information lies in a single one-dimensional dependency chain.

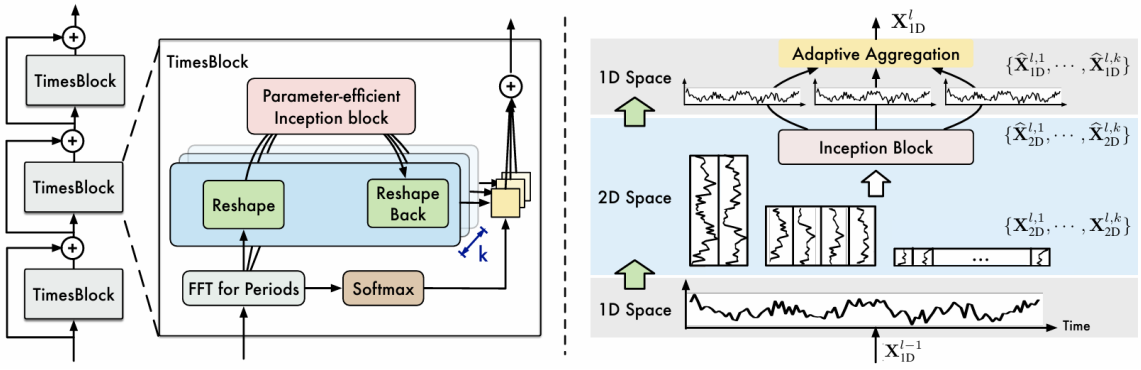

Figure 9. Overall architecture of TimesNet.

As depicted in Figure 9, TimesNet transforms one-dimensional time-series data into two-dimensional tensors by exploiting the detected period length, then applies 2D convolutional kernels to capture both intra-period and inter-period temporal patterns.

PatchTST, introduced in the work titled A Time Series Is Worth 64 Words [25], applies a patching strategy that converts contiguous time segments into token-like units. This approach is important for finance because local motifs such as breakout formation, volatility expansion, or post-news consolidation often occupy short contiguous windows rather than isolated timestamps. By working with patches and channel-wise treatment, the model reduces sequence length while preserving local temporal context. The underlying insight is that for many forecasting tasks, including stock indexes, the right unit of representation may be a short segment rather than a single scalar observation.

Other modern architectures challenge transformer dominance from different directions. TiDE [7] uses a dense encoder-decoder structure for long-term forecasting, while TSMixer [5] demonstrates that all-MLP architectures can be surprisingly competitive when temporal and feature mixing are well designed. These models matter because they show that good forecasting does not necessarily require attention at every layer. For stock prediction, where model efficiency and reproducibility are often as important as raw accuracy, simpler architectures can be attractive if they preserve enough capacity to model cross-horizon interactions and regime-sensitive features.

The transformer family itself has also diversified. iTransformer [20] rethinks the assignment of tokens and channels, arguing that inverted treatment of feature dimensions can improve forecasting effectiveness. TimeMixer [30] emphasizes decomposable multiscale mixing, explicitly acknowledging that temporal dynamics unfold across nested scales rather than a single homogeneous horizon. SAMformer [16] adds sharpness-aware optimization and channel-wise attention, highlighting that successful forecasting depends not only on representation structure but also on how the model is trained and regularized. For stock indexes, these contributions collectively strengthen the case that architecture search should be guided by scale decomposition, channel relevance, and generalization stability rather than by adopting the latest generic transformer variant uncritically.

What distinguishes these modern backbones is not simply accuracy, but the way each one defines the fundamental forecasting object. TimesNet [32] highlights temporal variation patterns, PatchTST [25] highlights local segments, TiDE [7] highlights dense temporal compression, TSMixer [5] highlights separable mixing, iTransformer [20] highlights variable-centric representation, and TimeMixer [30] highlights explicit multi-scale decomposition. This variety is useful for finance because stock indexes are simultaneously path-dependent, feature-dependent, and scale-dependent. A researcher choosing among these architectures should

therefore ask which representation lens best matches the available data and the intended decision horizon, rather than treating all transformer-era models as interchangeable upgrades.

4.3. Foundation Models and Large-Model Thinking for Time Series

The latest stage of the literature extends beyond task-specific architectures, venturing into the realm of foundation-model style forecasting. A notable example is Time-LLM [17], which reprograms large language models specifically for time-series forecasting. This innovative approach effectively poses the question of whether pretrained language representations can be adapted for temporal reasoning through appropriate interfaces and techniques. This is a conceptually bold move as it treats time series not merely as numeric sequences to be extrapolated, but rather as structured signals that may significantly benefit from the application of broad pretrained priors derived from extensive datasets. The appeal for stock forecasting is particularly strong: if large pretrained systems can successfully transfer robust sequence abstractions and insights, they may help alleviate the severe data scarcity challenges that are often encountered by domain-specific models trained solely on a single market or specific forecasting horizon. By leveraging the strengths of foundational models, researchers could unlock new avenues for enhancing predictive performance and improving decision-making in financial markets.

Related efforts deepen this emerging trend in various innovative ways. For instance, TEMPO [4] formulates a prompt-based generative pretraining approach specifically tailored for time-series forecasting, while Timer [22] advocates for the use of generative pre-trained transformers as large-scale models designed for time-series data. Additionally, the decoder-only foundation model introduced by Das et al. [7] advances the intriguing idea that large autoregressive forecasting systems can be constructed directly for time-series domains without significant modifications. Furthermore, Lag-Llama [26] introduces a probabilistic forecasting perspective, which is particularly valuable in finance because the calibration of uncertainty is just as crucial as producing mean predictions. Collectively, these studies represent a significant shift in the overarching research question from simply asking, "Which architecture performs best on a given benchmark?" to a more nuanced inquiry: "What kind of pretraining and transfer learning regime can effectively produce adaptable forecasting behavior across diverse tasks and datasets?" This evolution not only reflects a growing sophistication in modeling approaches but also emphasizes the importance of flexibility and generalization in predictive analytics within financial contexts.

Figure 10 presents the model framework of TIME-LLM, which reprograms a frozen large

language model for time-series forecasting by converting temporal data into token embeddings through a prompt-based interface.

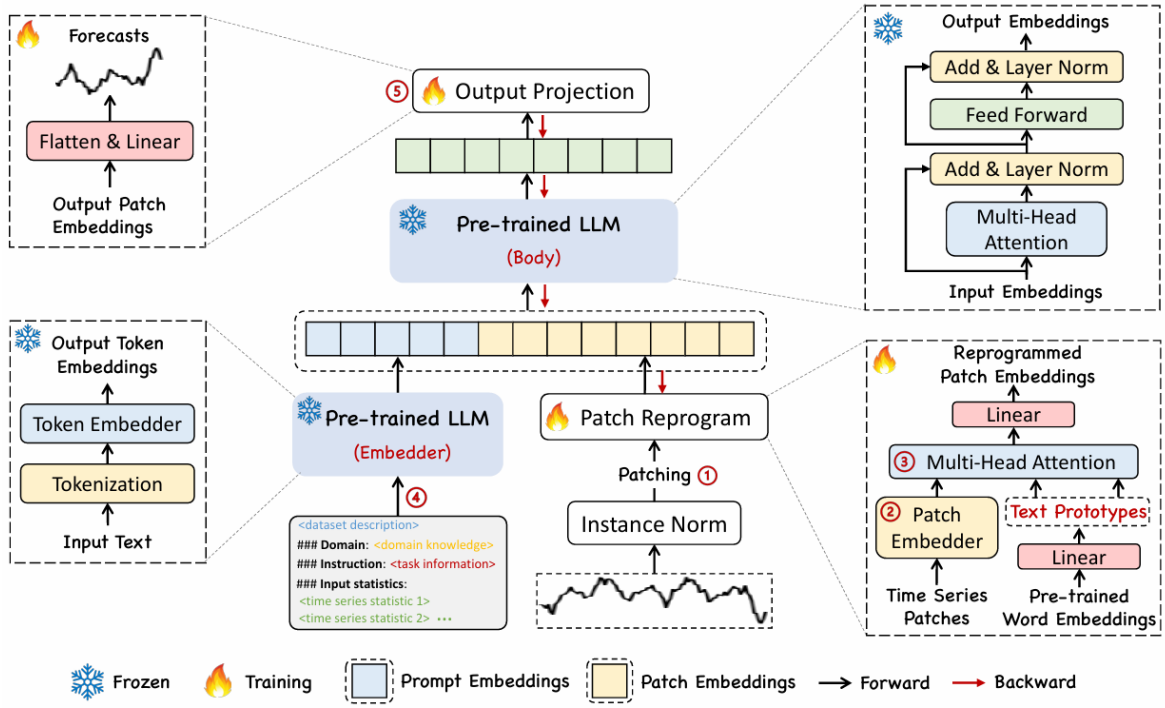

Figure 10. The model framework of TIME-LLM.

Yet stock index forecasting also exposes the limits of foundation-model optimism. Financial data are sparse relative to web-scale corpora, strongly regime dependent, and heavily influenced by institutional changes that may not repeat in the training history. A large model pretrained on generic time series or even on language may capture broad sequential regularities while still missing the economic semantics of market microstructure, policy surprise, or risk transmission. Therefore, the main value of foundation models in finance may lie in transfer-efficient representation learning, probabilistic calibration, and multimodal integration rather than zero-shot market prediction. The most promising direction is likely not replacement of domain models, but controlled fusion between pretrained sequence priors and finance-specific supervision.

Probabilistic output is especially important in this context. Large forecasting systems that report only point predictions are often less useful to financial decision makers than models that provide calibrated uncertainty, because trading, hedging, and risk budgeting all depend on the distribution around the forecast. Lag-Llama [26] is therefore conceptually important even beyond its empirical results, since it represents the move from deterministic large-model extrapolation toward probabilistic large-model forecasting. For stock index applications, the next generation of foundation models should ideally support scenario ranges, tail-aware

intervals, and confidence measures that remain stable under regime change.

Another unresolved question is how prompting and adaptation should work in financial settings. In language tasks, prompting can externalize task instructions, contextual hints, and examples. For stock forecasting, however, the useful context may consist of recent market regimes, macro calendars, event descriptions, or retrieved analog periods rather than natural-language task statements alone. This suggests that prompt-based time-series models such as TEMPO [4] and reprogramming approaches such as Time-LLM [17] may achieve their greatest financial value when combined with retrieval, regime annotation, and economically meaningful metadata rather than raw numeric windows in isolation.

5. Comparative Discussion

5.1. Inductive Bias and Data Regime

The surveyed literature clearly illustrates that model performance is fundamentally influenced by inductive bias. Statistical models tend to excel when the target process exhibits short-memory characteristics, is interpretable, and remains sufficiently stable over time for parameter estimates to retain their meaning across different periods. Classical machine learning approaches perform particularly well when domain expertise can contribute informative features that enhance predictive accuracy. Recurrent networks prove useful in scenarios where medium-range path dependence plays a significant role, while convolutional, patch-based, and multiscale architectures are better suited for capturing repeated local motifs and facilitating cross-horizon composition of data patterns. Additionally, hybrid decomposition models become increasingly attractive when the same raw time series contains multiple incompatible temporal regimes, allowing for more nuanced analysis. Meanwhile, foundation-model approaches gain plausibility when the ability to transfer knowledge across tasks can effectively compensate for limited labeled data in certain contexts. Thus, there is no architecture-independent answer to the question of what works best for stock index forecasting; instead, the choice of model should be informed by the specific characteristics of the data and the underlying processes being modeled. This highlights the importance of tailoring approaches to the unique challenges presented by financial time series data.

Dataset scale and market regime are equally important. Low-frequency index series with a few thousand observations do not justify the same parameterization as large multivariate panels or dense intraday sequences. Likewise, models tuned during tranquil periods can degrade badly during crises because the mapping from predictors to returns changes faster than the model can

adapt. This is why seemingly older ideas such as GARCH-style variance awareness [2] and RevIN-style shift handling [18] continue to matter even in otherwise modern pipelines. For practitioners, model selection should begin with the structure of the available data and the forecast use case, not with the popularity of the architecture.

5.2. Exogenous Information, Multimodality, and Representation Scope

Another important pattern is the widening scope of input representation. Price-only models remain useful benchmarks, but the stronger finance studies increasingly combine multiple information channels such as technical indicators, macro variables, cross-market covariates, and textual sentiment [15,21]. This shift reflects an economic reality: many index movements are driven by information that is not recoverable from recent prices alone. At the same time, adding modalities increases alignment risk, because timestamps, release lags, missingness, and varying update frequencies can introduce subtle leakage or dilution of signal. A robust review therefore cannot treat multimodality as automatically beneficial; its value depends on disciplined temporal synchronization and clear causal availability at forecast time.

Modern backbones also differ in how broadly they define representation scope. Patch-based and mixer-style models [5,25] widen the local receptive unit; multiscale models [30,32] widen temporal granularity; foundation models [4,8,17,22,26] widen transfer scope across tasks and datasets. These are all forms of representation expansion, but they respond to different bottlenecks. For stock indexes, the key challenge is to expand scope without drowning the weak predictive signal in irrelevant context. This is where financially grounded regularization, sparsity, and relevance weighting may become more valuable than simply increasing model size.

5.3. Evaluation, Robustness, and Economic Relevance

A persistent weakness of the stock forecasting literature is evaluation fragmentation. Studies differ in market selection, forecast horizon, target variable, split protocol, preprocessing pipeline, and reported metrics, making direct comparison difficult. Error metrics such as RMSE, MAE, or MAPE remain necessary, but they do not reveal whether a forecast is economically actionable after transaction costs, turnover, slippage, and risk constraints. This is especially relevant when comparing sophisticated hybrids or large models against simpler baselines: a marginal gain in point accuracy may not translate into a better trading or risk-management decision. Future stock index forecasting research should therefore report both statistical accuracy and decision-oriented utility whenever the intended application is financial action

rather than descriptive extrapolation.

Robustness is a second evaluation dimension that deserves more weight. A model should be tested across rolling windows, multiple market states, and realistic out-of-sample regimes rather than a single favorable split. Distribution-shift handling [18], decomposition robustness [19,24], and transfer behavior of foundation models [4,8,17,22,26] all need evaluation protocols that stress adaptation rather than just average fit. Without such testing, architectural novelty can easily be mistaken for durable forecasting ability. The literature surveyed here suggests that reproducibility, robustness, and economic interpretation are now at least as important as adding another modeling block to the pipeline.

Metric selection should also be broadened. Directional accuracy, hit rate, interval coverage, calibration error, and portfolio-level statistics such as drawdown or risk-adjusted return can reveal properties that RMSE alone conceals. A model with slightly worse point error may still be preferable if it identifies turning points more consistently or produces uncertainty estimates that prevent overconfident trading. Conversely, a model with excellent average error may fail exactly when risk management matters most, such as during crisis transitions. The next phase of stock index forecasting research should therefore evaluate models as forecasting systems embedded in decisions, not merely as function approximators optimized for one scalar score.

5.4. Practical Guidance for Model Selection

The reviewed evidence supports a pragmatic selection logic. When interpretability, data scarcity, or regulatory transparency is paramount, ARIMA, GARCH, VAR, and classical machine learning remain strong starting points [2,3,10,28,33]. When the dataset is rich enough and the objective is to exploit nonlinear sequential dependence, recurrent and convolutional financial models [1,6,9,11,12,13,14,15,22,23,27,29,34] are appropriate, especially if exogenous variables are available. When nonstationarity is severe and multi-scale behavior dominates, decomposition-driven hybrids and state-space methods [12,19,24,27] are promising. When the task requires transfer across datasets, probabilistic forecasting, or multimodal integration at scale, the more recent time-series and foundation-model literature [4,5,7,8,16,17,20,25,26,30,32] becomes relevant. The strongest research strategy is not to assume that one category supersedes all others, but to treat each family as a tool matched to a specific data regime and decision requirement.

6. Open Challenges and Research Directions

The first open challenge is benchmark quality. Stock forecasting studies still suffer from

inconsistent train-test splits, insufficient leakage control, and limited reporting of preprocessing choices. Standardized rolling-window benchmarks across multiple geographic markets and volatility regimes would make it much easier to judge whether a new model improves generalization or simply benefits from a favorable setup. Such benchmarks should include both statistical targets and decision-oriented targets, because a useful stock index model may be valuable for direction classification, volatility-aware allocation, or interval forecasting even when point error is not state of the art.

Benchmark design should also be horizon aware. A model that is competitive for one-step-ahead daily prediction may fail for weekly trend forecasting or monthly allocation support, and the reverse can also be true. Many architecture debates in the literature are implicitly debates about horizon mismatch rather than absolute model quality. Future evaluations should therefore report results across short, medium, and longer horizons whenever possible, making it clear whether a method is best interpreted as a micro-pattern detector, a regime tracker, or a long-horizon planner.

The second challenge is multimodal and multi-resolution integration. Financial signals arrive through prices, order flow, macro releases, news, and cross-asset spillovers, each with different timing conventions and noise characteristics. Future research should pay more attention to causally valid synchronization, confidence-aware fusion, and hierarchical routing across temporal scales. The success of sentiment-aware models [21], decomposition hybrids [19,24], and state-space or multiscale architectures [12,27,30,32] suggests that the next gains will come from better coordination among representations rather than from a single universal backbone.

The third challenge concerns foundation models and transfer learning. Large-model approaches [4,8,17,22,26] are promising, but finance raises unresolved questions about pretraining corpora, domain mismatch, continual adaptation, and probabilistic calibration under rare events. A stock index model that sees many generic time series may still fail on policy shocks or crisis dynamics that are economically unique. Accordingly, foundation-model research in finance should emphasize domain adaptation, uncertainty estimation, and retrieval or prompting mechanisms grounded in economically meaningful context rather than assuming that scale alone solves the forecasting problem.

Finally, interpretability and responsible deployment remain central. As models become deeper and more hybridized, it becomes harder to explain whether predictions arise from stable economic relationships, transient co-movements, or artifacts of preprocessing. This matters not only for trust, but also for model maintenance: without interpretable failure modes, it is difficult

to know when a forecasting system should be retrained, down-weighted, or suspended. The most valuable future systems will therefore combine strong predictive machinery with diagnostics that help analysts understand regime sensitivity, feature relevance, and uncertainty boundaries.

A related research direction is human-in-the-loop forecasting. In many practical settings, stock index models are not used autonomously but as inputs to analyst judgment, asset allocation meetings, or risk committees. This creates design opportunities that are underexplored in the literature, such as models that surface the dominant temporal regime, compare current conditions with retrieved historical analogues, or explain whether a forecast is driven by price action, exogenous news, or cross-market contagion. As forecasting architectures become more powerful, making them easier to interrogate may generate more practical value than pursuing another marginal accuracy gain on a fixed benchmark.

7. Conclusion

The literature reviewed here shows that stock index forecasting has progressed from linear univariate and multivariate benchmarks to a broad ecosystem of deep neural, hybrid, state-space, and foundation-model approaches. Each stage addressed a real limitation of the previous one: ARIMA clarified baseline dependence structure, GARCH modeled changing volatility, classical machine learning captured nonlinear feature interactions, recurrent and convolutional networks automated sequence representation, hybrid systems handled multi-scale nonstationarity, and foundation-model approaches opened the possibility of broader transfer and probabilistic generalization. At the same time, the field has learned that architectural novelty is not a substitute for disciplined evaluation.

For stock index forecasting, the central research problem is no longer simply how to fit a more expressive sequence model. It is how to align model inductive bias with financial data structure, preserve robustness under distribution shift, integrate exogenous information without leakage, and evaluate usefulness in economically meaningful terms. Viewed in that light, the most promising future direction is a hybrid one: transparent statistical reasoning for diagnostics, domain-aware deep architectures for representation, and carefully adapted foundation-model components for transfer, uncertainty, and multimodal context. Such an agenda is more demanding than chasing benchmark gains, but it is also more likely to produce forecasting systems that remain useful outside the narrow conditions of a single experiment.

References

[1] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv*:1803.01271.

[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. *https://doi.org/10.1023/A:1010933404324*

[4] Cao, D., Jia, F., Arik, S. O., Pfister, T., Zheng, Y., Ye, W., & Liu, Y. (2024). TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. *In Proceedings of the International Conference on Learning Representations* (ICLR).

[5] Chen, S.-A., Li, C.-L., Arik, S. O., Yoder, N. C., & Pfister, T. (2023). TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*.

[6] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724–1734). https://doi.org/10.3115/v1/D14-1179

[7] Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., & Yu, R. (2023). Long-term forecasting with TiDE: Time-series dense encoder. *arXiv preprint arXiv*:2304.08424.

[8] Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 10148–10167)*. PMLR.

[9] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

[10] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

[11] Ge, Q. (2025). Enhancing stock market forecasting: A hybrid model for accurate prediction of S&P 500 and CSI 300 future prices. *Expert Systems with Applications*, 260, Article 125380. https://doi.org/10.1016/j.eswa.2024.125380

[12] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv*:2312.00752.

[13] He, Z., Zhang, H., & Por, L. Y. (2024). A comparative study on deep learning models for stock price prediction. *In Image Processing, Electronics and Computers: Advances in Transdisciplinary Engineering*. IOS Press. https://doi.org/10.3233/ATDE240502

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. *https://doi.org/10.1162/NECO.1997.9.8.1735*

[15] Hoseinzadeh, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, 129, 273–285. https://doi.org/10.1016/j.eswa.2019.03.029

[16] Ilbert, R., Odonnat, A., Feofanov, V., Virmaux, A., Paolo, G., Palpanas, T., & Redko, I. (2024). SAMformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. *In Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 20924–20954). PMLR.

[17] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., & Wen, Q. (2024). Time-LLM: Time series forecasting by reprogramming large language models. *In Proceedings of the International Conference on Learning*

Representations (ICLR).

[18] Kim, T., Kim, J., Tae, Y., Park, C., Choi, J.-H., & Choo, J. (2022). Reversible instance normalization for accurate time-series forecasting against distribution shift. *In Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.

[19] Li, S., Tang, G., Chen, X., et al. (2024). Stock index forecasting using a novel integrated model based on CEEMDAN and TCN-GRU-CBAM. *IEEE Access*, 12, 122524–122543. https://doi.org/10.1109/ACCESS.2024.3452426

[20] Liu, W. J., Ge, Y. B., & Gu, Y. C. (2024). News-driven stock market index prediction based on trellis network and sentiment attention mechanism. *Expert Systems with Applications*, 250, Article 123966. https://doi.org/10.1016/j.eswa.2024.123966

[21] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted transformers are effective for time series forecasting. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[22] Liu, Y., Zhang, H., Li, C., Huang, X., Wang, J., & Long, M. (2024). Timer: Generative pre-trained transformers are large time series models. *In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 32369–32399)*. PMLR.

[23] Mu, S., Liu, B., Gu, J., et al. (2024). Research on stock index prediction based on the spatiotemporal attention BiLSTM model. *Mathematics*, 12(18), Article 2812. https://doi.org/10.3390/math12182812

[24] Mutinda, J. K., & Geletu, A. (2025). Stock market index prediction using CEEMDAN-LSTM-BPNN-decomposition ensemble model. *Journal of Applied Mathematics*, 2025, Article 7706431. https://doi.org/10.1155/jama/7706431

[25] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[26] Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Darvishi Bayazi, M. J., Adamopoulos, G., Riachi, R., Hassen, N., Bilos, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2023). Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv*:2310.08278.

[27] Shi, Z. (2024). MambaStock: Selective state space model for stock prediction. *arXiv preprint arXiv*:2402.18959.

[28] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica,* 48(1), 1–48. https://doi.org/10.2307/1912017

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems 30* (pp. 5998–6008).

[30] Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J. Y., & Zhou, J. (2024). TimeMixer: Decomposable multiscale mixing for time series forecasting. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[31] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *In Proceedings of the European Conference on Computer Vision* (pp. 3–19). https://doi.org/10.1007/978-3-030-01234-2_1

[32] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[33] Xu, R. (2025). Modeling and comparing S&P 500, FTSE and SSEC stock price with ARIMA model. Frontiers in Economics and Management, 6(6), 60–69. *https://doi.org/10.6981/FEM.202506_6(6).0008*

[34] Zhang, J., Ye, L., & Lai, Y. (2023). Stock price prediction using CNN-BiLSTM-attention

model. *Mathematics*, 11(9), Article 1985. https://doi.org/10.3390/math11091985