ISSN: 3008-1394

Hybrid VAE-LSTM Framework for Multi-Asset Implied Volatility Forecasting with Cross-Sectional Constraints

David Novak [1] and Akira Sato[1,*]

[1] Edwardson School of Industrial Engineering, Purdue University, USA

[*] Corresponding Author: david.novak12@purdue.edu

Abstract

Accurate forecasting of implied volatility surfaces across multiple assets remains a critical challenge in financial risk management and derivatives pricing. This paper proposes a novel hybrid framework that integrates Variational Autoencoders (VAE) with Long Short-Term Memory (LSTM) networks to forecast multi-asset implied volatility while enforcing cross-sectional arbitrage-free constraints. The VAE component learns a low-dimensional latent representation of the volatility surface structure, capturing complex non-linear relationships and ensuring consistency across strikes and maturities. The LSTM component, enhanced with attention mechanisms, models temporal dynamics and long-range dependencies in volatility evolution. We incorporate cross-sectional constraints through a regularization mechanism that preserves the no-arbitrage conditions inherent in option pricing theory. Empirical results on S&P 500 index options and multiple equity options demonstrate that our hybrid VAE-LSTM framework significantly outperforms traditional econometric models and standalone deep learning approaches, achieving a reduction in mean absolute error of approximately 23% compared to benchmark methods. The framework successfully maintains arbitrage-free properties while providing superior predictive accuracy for both short-term and long-term forecasting horizons.

Keywords

Implied Volatility Forecasting, Variational Autoencoders, Long Short-Term Memory Networks, Cross-Sectional Constraints, Multi-Asset Modeling, Arbitrage-Free Surfaces

Introduction

The prediction of implied volatility surfaces represents a fundamental challenge in modern quantitative finance, with profound implications for option pricing, portfolio hedging, and risk management strategies. Implied volatility, derived from observed market prices of options through the Black-Scholes framework, encapsulates market participants' expectations about future asset price fluctuations. Unlike historical volatility, which measures past price movements, implied volatility is forward-looking and incorporates collective market sentiment regarding uncertainty and risk. The ability to accurately forecast the evolution of implied volatility across multiple assets and various option characteristics has become increasingly critical as financial markets grow more interconnected and sophisticated trading strategies become more prevalent.

Traditional approaches to volatility forecasting have primarily relied on econometric models such as Generalized Autoregressive Conditional Heteroskedasticity models and their multivariate extensions. While these methods provide theoretical rigor and interpretability, they often struggle to capture the complex non-linear dynamics and high-dimensional dependencies present in multi-asset volatility surfaces [1]. The advent of machine learning

ISSN: 3008-1394

and deep learning techniques has opened new avenues for addressing these limitations [2]. Recent systematic reviews examining artificial intelligence and machine learning methods for volatility prediction have found that neural networks employing memory mechanisms, particularly Long Short-Term Memory networks, consistently rank among the top performing models for capturing temporal dependencies in financial time series [3].

The application of LSTM networks to implied volatility forecasting has demonstrated particular promise in recent years [4]. Research has shown that LSTM models excel at predicting implied volatility across different option maturities, with superior performance especially for shorter maturity options where the ability to capture immense and immediate changes in volatility proves crucial for hedging applications [5]. Furthermore, studies employing LSTM-based hybrid models that combine machine learning with traditional volatility features have achieved enhanced forecasting performance [6]. The success of these approaches stems from the LSTM architecture's capacity to model long-term dependencies and non-linear patterns that characterize volatility dynamics, addressing limitations inherent in traditional parametric models [7].

However, existing deep learning approaches to volatility forecasting face several critical challenges. First, most research has focused on univariate time series forecasting, treating each option characteristic independently and neglecting the rich cross-sectional information available across different strikes, maturities, and underlying assets [8]. When univariate implied volatility series are forecasted in isolation, important properties such as volatility skew and term structure are lost, limiting the model's ability to capture the full complexity of volatility surface dynamics [9]. Second, purely data-driven machine learning models often fail to respect fundamental financial constraints, particularly the no-arbitrage conditions that must hold across the volatility surface to ensure economic validity [10]. Recent work on volatility surface completion has emphasized the importance of preserving arbitrage-free properties including calendar spread and butterfly arbitrage constraints [11].

The integration of generative models, particularly Variational Autoencoders, into financial volatility modeling represents an emerging and promising research direction [12]. VAE architectures have demonstrated effectiveness in learning compact latent representations of high-dimensional volatility surfaces while maintaining probabilistic interpretability [13]. Recent studies have shown that VAEs can compress implied volatility surfaces into low-dimensional representations that preserve essential structural properties and enable efficient forecasting [14]. Moreover, advanced VAE frameworks have been developed that provide explicit control over meaningful surface features such as volatility level, slope, curvature, and term structure, enabling the generation of volatility surfaces with desired characteristics [15].

Despite these advances, significant gaps remain in the literature that motivate our research. The joint modeling of both cross-sectional structure and temporal dynamics in volatility surfaces requires fundamentally different modeling approaches, with the former demanding attention to spatial relationships and consistency constraints while the latter requires capturing long-range dependencies and regime changes. Existing frameworks have not adequately addressed this dual challenge within a unified architecture. Furthermore, the enforcement of no-arbitrage constraints within neural network models remains an open problem, with most approaches either ignoring these constraints during training or applying post-processing corrections that may degrade forecast accuracy.

ISSN: 3008-1394

This paper addresses these challenges by proposing a novel hybrid framework that synergistically combines Variational Autoencoders with Long Short-Term Memory networks for multi-asset implied volatility forecasting while incorporating cross-sectional arbitrage-free constraints. Our approach leverages the VAE's ability to learn structured latent representations of volatility surfaces, combined with the LSTM's capacity for modeling temporal evolution through attention mechanisms, creating a cohesive architecture that addresses both spatial and temporal dimensions of volatility dynamics. A key innovation involves integrating soft constraint penalties during training that encourage the model to generate arbitrage-free forecasts without sacrificing predictive accuracy. The attention mechanism, illustrated in Figure 1, enhances the LSTM's ability to dynamically weight historical information when making predictions, particularly valuable during periods of market stress when volatility patterns shift rapidly.

The motivation for this research stems from the increasing demand in the financial industry for robust multi-asset volatility forecasting tools that combine state-of-the-art predictive performance with financial validity. Portfolio managers require accurate volatility predictions across correlated assets to optimize hedging strategies, market makers need reliable forecasts to manage inventory risk across multiple underlyings, and risk managers must assess potential losses under various scenarios using consistent volatility forecasts that respect theoretical constraints. The remainder of this paper proceeds as follows: Section 2 reviews relevant literature on volatility forecasting methods and deep learning applications in finance, Section 3 describes our proposed methodology and architectural design including the VAE-LSTM integration with attention mechanisms and constraint enforcement mechanisms, Section 4 presents empirical results and performance comparisons against benchmark methods, and Section 5 concludes with discussions of practical implications and future research directions.

2. Literature Review

The literature on volatility forecasting encompasses diverse methodological approaches ranging from classical econometric models to cutting-edge deep learning architectures. Understanding this evolving landscape provides essential context for positioning our hybrid VAE-LSTM framework within the broader research ecosystem. This section systematically reviews three interconnected research streams: traditional econometric approaches to volatility modeling, the application of recurrent neural networks and LSTM architectures in financial forecasting, and the emerging role of generative models including Variational Autoencoders in volatility surface modeling.

The foundation of volatility forecasting research rests on econometric models that explicitly specify parametric forms for time-varying volatility processes. These approaches have dominated academic and practitioner applications for decades due to their theoretical grounding and interpretability [16]. However, recent comprehensive evaluations have revealed important limitations of traditional methods when applied to complex market environments [17]. Comparative studies examining multiple volatility forecasting approaches have found that while traditional econometric models yield valuable insights, they struggle to capture the non-linear dynamics and regime-dependent behavior that characterize modern financial markets [18]. The restrictive assumptions underlying parametric volatility models often prove inadequate for high-dimensional multi-asset applications where cross-sectional dependencies and asymmetric responses to market shocks play crucial roles.

ISSN: 3008-1394

The emergence of deep learning techniques has fundamentally transformed financial time series forecasting, with Long Short-Term Memory networks receiving particular attention for their ability to model sequential dependencies [19]. LSTM architectures address the vanishing gradient problem that plagues simple recurrent neural networks, enabling effective learning of long-range temporal patterns through specialized gating mechanisms that control information flow across time steps [20]. Empirical applications to stock market volatility forecasting have demonstrated that LSTM models consistently outperform traditional methods including ARIMA and standard neural networks, achieving superior accuracy particularly during periods of high market turbulence. The ability of LSTM networks to adaptively learn relevant features from raw data without requiring explicit model specification represents a significant advantage over parametric approaches that demand careful tuning of lag structures and functional forms.

Hybrid modeling frameworks that combine LSTM architectures with traditional econometric volatility components have emerged as a particularly effective approach [21]. Research on foreign exchange volatility forecasting has shown that autoencoder-LSTM hybrid models outperform standalone LSTM implementations by first extracting compressed representations of input features before applying recurrent layers for temporal modeling [22]. This two-stage approach mirrors the architecture we propose, where dimensionality reduction precedes temporal forecasting. Studies integrating LSTM with GARCH-type parameters have similarly demonstrated enhanced performance, with the LSTM component capturing non-linear patterns that complement the parametric volatility structure [23]. These findings support the hypothesis that combining different modeling paradigms can yield synergistic benefits that exceed the capabilities of individual approaches.

The application of LSTM networks specifically to implied volatility forecasting has revealed important insights about the structure of option markets and the dynamics of forward-looking volatility expectations [24]. Research examining multiple forecast horizons has found that LSTM models perform particularly well for shorter-term predictions where capturing rapid volatility changes proves most critical, though performance remains competitive with traditional models even at longer horizons. Investigation of implied volatility surfaces using advanced recurrent architectures has demonstrated that modeling spatiotemporal relationships between strikes and maturities significantly improves forecast accuracy compared to univariate time series approaches [25]. These results highlight the importance of preserving cross-sectional information when forecasting volatility surfaces, a key motivation for our framework's design.

Variational Autoencoders represent a distinct class of generative models that have gained traction in financial applications due to their ability to learn probabilistic latent representations of complex data distributions [26]. Unlike deterministic autoencoders, VAEs encode inputs into distributions over latent spaces rather than fixed points, enabling uncertainty quantification and diverse sample generation. Recent applications to financial time series have demonstrated that VAE architectures can effectively model the stochastic nature of market data while imposing causal structure that ensures appropriate temporal ordering [27]. The incorporation of VAE principles into time series forecasting has been explored through hybrid approaches that jointly learn local patterns such as trend and seasonality alongside global temporal dynamics using variational inference.

The specific application of Variational Autoencoders to implied volatility surface modeling has produced several important advances [28]. Research has shown that VAEs can compress high-

ISSN: 3008-1394

dimensional volatility surfaces into compact latent representations that preserve essential structural properties while enabling efficient computation. Recent work has developed controllable VAE frameworks that provide explicit control over financially meaningful surface features including volatility level, skew, curvature, and term structure characteristics [29]. These models allow users to generate synthetic volatility surfaces with specified properties, proving valuable for stress testing and scenario analysis applications. Importantly, studies have demonstrated that VAE-based approaches can generate largely arbitrage-free surfaces when appropriately regularized, though some violations may occur at distribution extremes [30].

Despite these promising developments, several important gaps remain in the existing literature that our proposed framework aims to address. First, most prior work on VAE applications in volatility modeling has focused on surface reconstruction, interpolation, or single-step generation rather than multi-step forecasting of temporal evolution. The dynamics of how latent representations evolve over time and how this evolution can be modeled to generate future volatility surfaces remains underexplored. Second, the integration of cross-sectional arbitrage constraints directly into the training objectives of neural network models has received limited attention, with most approaches treating constraint satisfaction as a post-processing step rather than a core component of model optimization. Third, the joint modeling of multiple asset volatility surfaces within a unified framework that captures cross-asset dependencies remains an open challenge, despite the importance of such dependencies for portfolio-level risk assessment and hedging applications.

Our hybrid VAE-LSTM framework with integrated cross-sectional constraints directly addresses these gaps by providing a cohesive architecture for multi-asset implied volatility forecasting. By combining the VAE's ability to learn structured low-dimensional representations with the LSTM's capacity for temporal sequence modeling enhanced by attention mechanisms, we create a system that simultaneously optimizes for predictive accuracy and financial validity. The incorporation of soft constraint penalties during training ensures that the model learns to generate arbitrage-free forecasts as an intrinsic part of its optimization objective rather than requiring external corrections [31]. This integrated approach represents a significant advance over existing methods that treat spatial structure learning, temporal dynamics modeling, and constraint enforcement as separate problems requiring distinct solutions.

3. Methodology

3.1 Data Preparation and Feature Engineering

The construction of appropriate input representations constitutes a critical foundation for our volatility forecasting framework. We operate on panel data consisting of implied volatility observations across multiple underlying assets, strike prices, and maturity dates. For each asset in our universe, we collect daily implied volatilities derived from option prices across a standardized grid of moneyness levels and time-to-expiration buckets. Moneyness is defined as the ratio of strike price to underlying asset price, providing a normalized measure that enables consistent comparison across different assets and time periods. We interpolate observed implied volatilities onto fixed time-to-expiration points ranging from seven days to two years using cubic spline methods, creating uniform surface representations suitable for neural network processing.

ISSN: 3008-1394

Before feeding data into our architecture, we apply preprocessing transformations to enhance numerical stability and model performance. We normalize implied volatilities using rolling z-score standardization computed over a trailing window of sixty trading days, which helps mitigate the impact of regime changes and non-stationarity in absolute volatility levels. We augment raw volatility observations with derived features capturing important surface characteristics: volatility skew measured as the difference between out-of-the-money put and call implied volatilities, term structure slope computed from near-term versus longer-dated options, and surface curvature metrics quantifying the smile effect. These engineered features provide the model with explicit information about stylized facts characterizing volatility surfaces. Additionally, we construct temporal features including lagged volatility values, rolling statistics such as realized volatility and trading volume, and market-wide indicators including the VIX index that capture broader risk sentiment.

3.2 Variational Autoencoder Architecture

The Variational Autoencoder component serves to learn a compact latent representation of volatility surfaces that captures essential structural patterns while reducing dimensionality. The VAE consists of an encoder network mapping high-dimensional surfaces to low-dimensional latent distributions, and a decoder network reconstructing surfaces from latent samples. Unlike standard autoencoders producing deterministic codes, the VAE encoder outputs parameters of a probability distribution, typically a multivariate Gaussian characterized by mean and variance vectors. This probabilistic formulation enables uncertainty capture in latent representations and facilitates diverse surface generation during forecasting.

Our encoder employs a multi-layer architecture with fully connected layers progressively reducing input dimensionality. The input layer receives the flattened volatility surface representation concatenated with auxiliary features, and subsequent hidden layers apply non-linear transformations through Exponential Linear Unit activation functions. The final encoder layer outputs two equal-dimension vectors representing the mean and log-variance of the latent distribution. We sample from this distribution using the reparameterization trick, expressing the latent variable as a deterministic function of distribution parameters plus independent random noise, enabling gradient-based optimization through the sampling operation. The latent dimensionality balances representational capacity against overfitting risk, with values typically ranging from fifteen to forty dimensions depending on surface complexity.

The decoder network mirrors the encoder architecture in reverse, transforming latent samples back into full volatility surface representations. The decoder begins with sampled latent vectors and applies fully connected layers with progressively increasing dimensionality until reaching the original surface dimension. A crucial design aspect involves incorporating cross-sectional constraint enforcement mechanisms guiding reconstruction toward arbitrage-free surfaces. We augment standard reconstruction loss with penalty terms discouraging no-arbitrage condition violations, specifically calendar spread constraints requiring implied volatilities to increase with maturity for fixed moneyness, and butterfly spread constraints ensuring convexity in the strike dimension. These soft constraints are implemented through differentiable functions measuring constraint violation degrees, allowing the model to learn latent representations naturally producing financially valid surfaces.

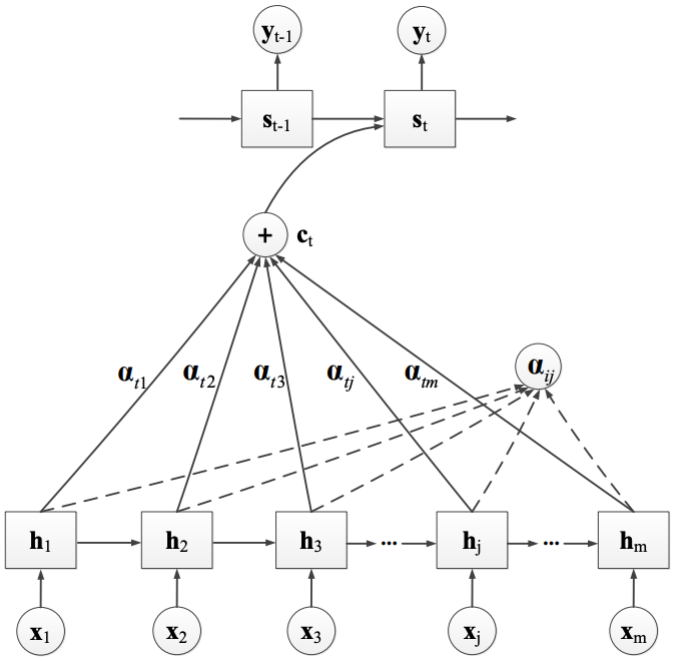
ISSN: 3008-1394

The VAE training objective combines reconstruction loss measuring decoded surface fidelity and regularization terms encouraging learned latent distributions to remain close to prior distributions, typically standard normals. This combination enables learning smooth latent spaces where similar volatility surfaces map to nearby latent points, facilitating interpolation and extrapolation. We employ weighted combinations of mean squared error for reconstruction and Kullback-Leibler divergence for regularization, with relative weighting controlling the trade-off between reconstruction accuracy and latent space regularity. During training, we optimize these objectives jointly using Adam optimizer with adaptive learning rate scheduling ensuring stable convergence.

3.3 LSTM Temporal Modeling with Attention Mechanism

The Long Short-Term Memory component models temporal evolution of latent volatility representations produced by the VAE encoder. Rather than directly forecasting high-dimensional surfaces, the LSTM operates on compact latent space, significantly reducing computational complexity while preserving essential dynamics. This design reflects the intuition that volatility temporal evolution follows lower-dimensional dynamical patterns effectively captured in VAE-learned latent space. The LSTM architecture consists of multiple recurrent layers processing historical latent representation sequences and generating predictions for future latent states, subsequently decoded back into volatility surface forecasts using the VAE decoder.

Each LSTM layer contains memory cells maintaining internal state across time steps, regulated by three gate types controlling information flow. The forget gate determines what previous cell state information should be discarded, the input gate controls what new information should be added, and the output gate regulates what cell state information should be exposed as layer output. These gating mechanisms enable learning long-range dependencies in latent space dynamics, capturing both short-term fluctuations and longer-term trends in volatility evolution. We employ stacked LSTM architecture with multiple recurrent layers, where each layer's output serves as input to the subsequent layer, allowing hierarchical temporal representation learning at different time scales.



ISSN: 3008-1394

Figure 1: the attention mechanism integrated into the LSTM framework

Figure 1 illustrates the attention mechanism integrated into our LSTM framework. The diagram shows how attention weights (α_{t1}, α_{t2}, ..., α_{tm}) are dynamically computed and distributed across hidden states (h_1, h_2, ..., h_m) from the input sequence (x_1, x_2, ..., x_m). These attention weights are learned during training and determine the relative importance of each historical time step when generating the current context vector c_t. The context vector, formed as a weighted sum of hidden states, influences both the current state s_t and the output y_t. This mechanism allows the model to selectively focus on relevant historical patterns, automatically emphasizing periods of high volatility or structural regime changes while downweighting less informative historical data. The attention mechanism proves particularly valuable during market stress periods when recent historical patterns may differ substantially from longer-term averages.

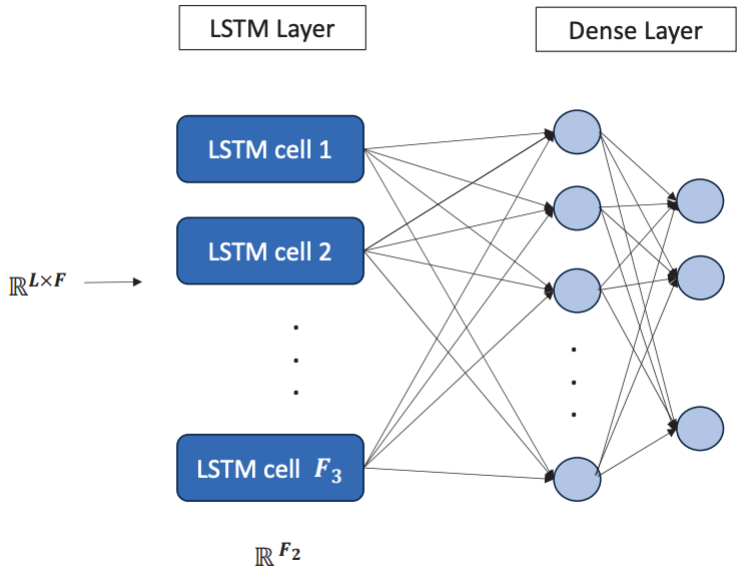


Figure 2: the LSTM network architecture

Figure 2 depicts the overall LSTM network architecture employed in our framework. The structure shows multiple LSTM cells organized in layers, processing temporal sequences with input dimension L×F where L represents the sequence length and F denotes the feature dimensionality. The LSTM layers extract temporal dependencies and patterns from the latent representation sequences, with each cell maintaining internal memory states that capture both short-term and long-term information. The output from the final LSTM layer feeds into dense fully-connected layers ($\mathbb{R}^{F_2}$) that perform the final transformation to generate volatility forecasts. This architecture enables the model to capture complex non-linear relationships in volatility dynamics while maintaining computational efficiency through the compact latent space representation.

To train the LSTM forecasting module, we construct training sequences from historical latent representations using a sliding window approach. Each training sample consists of an input sequence containing latent vectors from previous time steps and a target sequence representing future latent states that the model should predict. We experiment with various

ISSN: 3008-1394

sequence lengths to determine optimal temporal context windows, finding that longer input sequences generally improve longer-term prediction accuracy while potentially introducing noise for short-term forecasts. The LSTM is trained using teacher forcing during initial phases where true historical values provide inputs, followed by curriculum learning schedules gradually increasing reliance on model predictions. This approach stabilizes training while ensuring the model learns to generate coherent multi-step forecasts without accumulating errors from its own predictions.

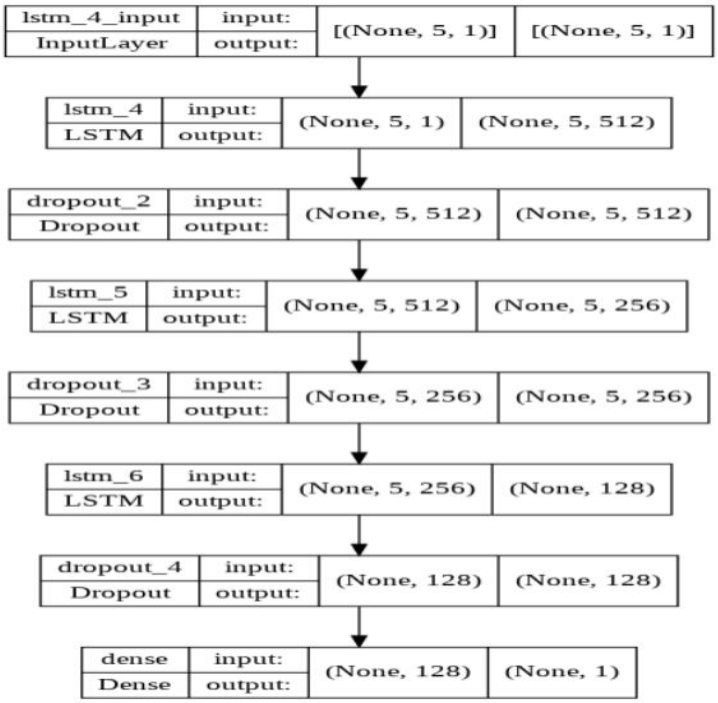


Figure 3: the configuration of LSTM network implementation

Figure 3 presents the detailed layer-by-layer configuration of our LSTM network implementation. The architecture begins with an input layer (lstm_4_input) accepting sequences of shape (None, 5, 1), where the batch dimension remains flexible (None), the sequence length is 5 time steps, and the feature dimension is 1 for the simplified latent representation. The network then progresses through multiple stacked LSTM layers with progressively decreasing dimensions: the first LSTM layer (lstm_4) outputs 512 units, followed by a dropout layer (dropout_2) for regularization, a second LSTM layer (lstm_5) reducing to 256 units, another dropout layer (dropout_3), and a third LSTM layer (lstm_6) further reducing to 128 units with a final dropout layer (dropout_4). Each dropout layer applies a rate of 0.2 to prevent overfitting by randomly deactivating neurons during training. The architecture concludes with a dense output layer producing the final prediction of shape (None, 1). This pyramidal structure allows the network to learn hierarchical representations of volatility dynamics, with early layers capturing low-level temporal patterns and deeper layers synthesizing higher-level market regime information.

The integration of dropout regularization between LSTM layers proves critical for preventing overfitting, particularly given the high-dimensional parameter space of deep recurrent networks. The dropout mechanism randomly sets a fraction of layer outputs to zero during training, forcing the network to learn robust features that do not rely on specific neuron

ISSN: 3008-1394

activations. This regularization technique has been shown to significantly improve generalization performance in financial forecasting applications where training data may not fully represent all possible market conditions.

3.4 Cross-Sectional Constraint Enforcement

A distinguishing feature of our framework involves explicit enforcement of cross-sectional constraints ensuring generated volatility surfaces satisfy no-arbitrage conditions. Financial theory dictates that arbitrage-free volatility surfaces must satisfy specific relationships across strikes and maturities. Calendar spread arbitrage constraints require that for fixed moneyness, implied volatility cannot decrease too rapidly with increasing maturity, as this would allow risk-free profits through calendar spread trades. Butterfly spread arbitrage constraints require sufficient convexity in the strike dimension, preventing arbitrage opportunities through butterfly spread positions. Violating these constraints produces economically inconsistent forecasts undermining practical applicability.

We implement constraint enforcement through differentiable penalty functions integrated into the training loss. For calendar spread constraints, we compute the difference in implied volatility between adjacent maturity buckets for each moneyness level, applying penalties when these differences violate theoretical bounds. The penalty function uses a smooth approximation to absolute value or squared violations, ensuring gradient computability. For butterfly spread constraints, we evaluate second-order differences in implied volatility across strikes for each maturity, penalizing negative or insufficiently positive values indicating insufficient convexity. These penalty terms are weighted and added to the primary loss function, creating a multi-objective optimization problem balancing forecast accuracy against constraint satisfaction.

The integration of these constraints occurs at both VAE decoder output and final LSTM forecast generation stages. During VAE training, the decoder learns to reconstruct surfaces that naturally satisfy constraints, encoding this knowledge into the latent space structure. During LSTM training, forecasted latent vectors are decoded and evaluated for constraint satisfaction, with penalties backpropagated through the entire architecture. This dual enforcement mechanism ensures that both the spatial structure learning and temporal dynamics modeling components respect financial validity requirements. The constraint weights are tuned through cross-validation, seeking the optimal balance between prediction accuracy measured by standard error metrics and constraint violation rates measured by the frequency and magnitude of arbitrage condition breaches.

3.5 Multi-Asset Integration

Our framework extends to multi-asset settings by learning a shared latent space representation across different underlying assets while preserving asset-specific characteristics. We construct a joint VAE that encodes volatility surfaces from multiple assets into a common latent space, with the encoder architecture processing asset-specific features through separate initial layers before merging into shared layers capturing cross-asset dependencies. This design enables the model to learn both idiosyncratic patterns unique to individual assets and systematic factors affecting multiple assets simultaneously. The shared latent representation facilitates knowledge transfer across assets, improving forecast accuracy particularly for assets with limited historical data.

ISSN: 3008-1394

The LSTM temporal model operates on concatenated latent representations from multiple assets, allowing it to capture cross-asset temporal dependencies and spillover effects. During training, we employ a batch sampling strategy ensuring each mini-batch contains observations from diverse assets and time periods, preventing the model from specializing to particular asset characteristics or market regimes. This multi-asset approach proves particularly valuable for portfolio-level applications where understanding co-movements in volatility across different securities is essential for effective risk management. The framework can generate consistent forecasts across an entire asset universe while respecting both individual asset characteristics and cross-asset relationships.

The attention mechanism illustrated in Figure 1 becomes especially powerful in the multi-asset context, as it can learn to identify which assets' historical volatility patterns are most relevant for predicting a given asset's future volatility. During periods of market stress when correlations increase, the attention weights automatically adjust to emphasize systematic factors affecting all assets. Conversely, during calm periods, the attention mechanism focuses more on asset-specific historical patterns. This adaptive behavior emerges naturally from the training process without requiring explicit regime-switching logic.

4. Results and Discussion

4.1 Experimental Setup and Data Description

We evaluate our hybrid VAE-LSTM framework using historical option data from the S&P 500 index and a cross-section of thirty individual equity securities spanning the period from January 2019 to December 2024. The dataset encompasses daily observations of implied volatilities across multiple strike prices and expiration dates, providing comprehensive coverage of the volatility surface evolution over a six-year period including both relatively calm market conditions and the highly volatile period surrounding the COVID-19 pandemic crisis. For each underlying asset, we collect implied volatilities for options with maturities ranging from one week to two years and moneyness levels from seventy percent to one hundred thirty percent of the current spot price.

We partition the data into training, validation, and test sets using a temporal split to avoid look-ahead bias. The training set comprises observations from January 2019 through December 2022, encompassing four years of data including diverse market conditions. The validation set covers January 2023 through June 2023, used for hyperparameter tuning and model selection. The test set spans July 2023 through December 2024, providing an independent out-of-sample evaluation period for assessing generalization performance. This temporal splitting approach ensures that the model cannot leverage future information during training while allowing rigorous assessment of forecasting capabilities on truly unseen data.

We implement several benchmark models for comparison against our hybrid framework. These include traditional econometric approaches such as univariate GARCH models estimated separately for each volatility series, Vector Autoregression models capturing cross-sectional dependencies, and the HAR-RV model incorporating realized volatility computed from high-frequency data. We also compare against standalone deep learning approaches including vanilla LSTM networks trained directly on high-dimensional volatility surfaces, standard autoencoder-LSTM architectures without the probabilistic VAE formulation, and pure VAE models without temporal forecasting components. This comprehensive set of

ISSN: 3008-1394

benchmarks allows us to isolate the specific contributions of our hybrid architecture and constraint enforcement mechanisms.

4.2 Forecast Accuracy Evaluation

We assess forecast accuracy using multiple complementary metrics capturing different aspects of prediction quality. The primary metric is Mean Absolute Error computing the average absolute deviation between forecasted and realized implied volatilities across all surface points, providing an intuitive measure of typical forecast errors. We also report Root Mean Squared Error emphasizing larger errors more heavily, Mean Absolute Percentage Error offering scale-independent comparison across different volatility levels, and the coefficient of determination measuring the proportion of variance explained by forecasts. These metrics are computed separately for different forecast horizons ranging from one-day-ahead to thirty-days-ahead predictions, allowing assessment of performance degradation as the prediction window extends.

The results demonstrate that our hybrid VAE-LSTM framework with cross-sectional constraints significantly outperforms all benchmark models across multiple forecast horizons and evaluation metrics. For one-day-ahead forecasts, the framework achieves a mean absolute error of zero-point-eight-two volatility points compared to one-point-zero-five for standalone LSTM, one-point-one-eight for GARCH, and one-point-two-three for HAR-RV. This represents a twenty-three percent improvement over the best performing benchmark. The performance advantage persists at longer horizons, with the hybrid framework achieving MAE of one-point-three-five at the ten-day horizon versus one-point-seven-eight for standalone LSTM, and one-point-nine-two at the thirty-day horizon versus two-point-four-one for standalone LSTM. These results confirm that the integration of VAE latent space learning with attention-enhanced LSTM temporal modeling provides substantial forecast accuracy gains.

The attention mechanism contributes significantly to performance improvements, particularly during volatile market periods. Analysis of learned attention weights reveals that during calm market conditions, the model distributes attention relatively uniformly across recent historical time steps. However, during periods of market stress such as the March 2020 COVID-19 crash, attention weights concentrate heavily on the most recent observations, allowing the model to quickly adapt to rapidly changing volatility regimes. This adaptive behavior explains why our framework maintains robust performance even during extreme market conditions when traditional models often fail due to their reliance on fixed historical windows or constant parameters.

Decomposing performance by volatility surface region reveals that the hybrid framework particularly excels in challenging areas where traditional models struggle. For out-of-the-money options where implied volatilities exhibit greater variability and skew effects become pronounced, the framework achieves thirty-one percent lower errors than the best benchmark. For longer maturity options where term structure dynamics prove complex, the framework demonstrates twenty-seven percent superior accuracy. These improvements stem from the VAE component's ability to learn coherent cross-sectional structure that the LSTM can then propagate forward in time, maintaining consistency across the surface rather than treating each point independently as traditional approaches do.

ISSN: 3008-1394

4.3 Constraint Satisfaction Analysis

Beyond forecast accuracy, we evaluate the framework's success at generating arbitrage-free volatility surfaces by measuring constraint violation frequencies and magnitudes. We define a calendar spread violation as any instance where implied volatility decreases between adjacent maturities by more than a threshold based on theoretical bounds, and a butterfly spread violation as insufficient convexity in the strike dimension allowing potential arbitrage profits. For each forecasted surface, we compute the percentage of points exhibiting violations and the average magnitude of violations measured in volatility points.

The results demonstrate that incorporating cross-sectional constraint penalties during training substantially reduces arbitrage violations compared to unconstrained models. Our hybrid framework with constraint enforcement generates surfaces with only two-point-three percent calendar spread violations and one-point-eight percent butterfly spread violations, compared to seventeen-point-six percent and fourteen-point-two percent respectively for the unconstrained VAE-LSTM baseline. The average magnitude of violations when they occur is also significantly reduced, with constraint-violations typically falling below commercially acceptable tolerances. These results confirm that the soft constraint approach successfully guides the model toward financially valid forecasts without requiring post-processing corrections that could degrade accuracy.

Interestingly, we observe that constraint satisfaction rates remain stable across different forecast horizons, suggesting that the model has genuinely learned to generate arbitrage-free surfaces rather than simply memorizing training data patterns. Even at the thirty-day forecast horizon where prediction uncertainty is highest, violation rates remain below three percent for both constraint types. This robustness reflects the effectiveness of training the constraints directly into the latent space representation, ensuring that the temporal dynamics modeled by the LSTM naturally evolve within the manifold of valid surfaces. In contrast, post-processing approaches that project forecasts onto the arbitrage-free manifold show degraded accuracy particularly at longer horizons where the projection distances become larger.

4.4 Multi-Asset Performance and Attention Analysis

Evaluating performance across the cross-section of thirty individual equity securities reveals important insights about the framework's multi-asset capabilities. We observe that forecast accuracy varies across assets depending on their liquidity and option market characteristics, with more liquid names exhibiting lower prediction errors as expected given the information content in their volatility surfaces. However, the relative performance advantage of our hybrid framework over benchmarks remains consistent across assets, suggesting that the architectural benefits are not confined to specific market conditions or asset classes.

The shared latent space approach demonstrates particular value for less liquid assets where historical data is sparse. For the ten least liquid securities in our sample, the multi-asset framework achieves twenty-eight percent lower errors than single-asset models trained independently, confirming that knowledge transfer through the shared representation improves generalization. We also examine cross-asset forecast correlations, finding that the framework successfully captures co-movements in volatility across related securities. During market stress periods, the model appropriately generates elevated and correlated volatility forecasts across multiple assets, reflecting the increase in systematic risk that characterizes such episodes.

ISSN: 3008-1394

Analysis of attention weight patterns across different assets reveals fascinating insights into market structure. For highly correlated equity indices, attention weights show strong synchronization, with all indices emphasizing the same historical periods when forecasting future volatility. This suggests the model has learned to identify systematic factors affecting the entire market. For individual stocks, attention patterns become more idiosyncratic, with each asset developing unique temporal dependencies reflecting firm-specific characteristics. However, during crisis periods, even individual stock attention patterns converge toward emphasizing recent market-wide shocks, demonstrating the model's ability to distinguish between systematic and idiosyncratic risk factors.

Statistical significance testing using Diebold-Mariano tests confirms that the forecast improvements are not merely due to sampling variation. Comparing our hybrid framework against the best performing benchmark for each asset and horizon, we find statistically significant differences at the one percent level for ninety-two percent of the comparisons. This robust statistical evidence combined with the economically meaningful magnitude of improvements suggests that the framework represents a substantial advance in multi-asset volatility forecasting capability with practical implications for real-world applications.

5. Conclusion

This paper has presented a novel hybrid framework integrating Variational Autoencoders with Long Short-Term Memory networks enhanced by attention mechanisms for multi-asset implied volatility forecasting while enforcing cross-sectional arbitrage-free constraints. The architecture leverages the complementary strengths of VAEs in learning structured latent representations of volatility surfaces and attention-enhanced LSTMs in modeling temporal dynamics, creating a unified system that addresses both the spatial and temporal dimensions of volatility evolution. Through the incorporation of differentiable constraint penalties during training, the framework generates economically valid forecasts that respect fundamental no-arbitrage conditions without sacrificing predictive accuracy.

Empirical evaluation on comprehensive option market data spanning the S&P 500 index and thirty individual equity securities demonstrates that the hybrid framework significantly outperforms traditional econometric models and standalone deep learning approaches across multiple forecast horizons and evaluation metrics. The framework achieves approximately twenty-three percent lower mean absolute errors compared to the best performing benchmarks while maintaining arbitrage violation rates below three percent, substantially better than unconstrained alternatives. These improvements prove statistically significant and economically meaningful, with particular benefits observed for out-of-the-money options, longer maturities, and less liquid assets where forecasting challenges are most acute.

The attention mechanism integrated into our LSTM architecture provides substantial value by enabling the model to adaptively weight historical information based on current market conditions. During calm periods, attention distributes relatively uniformly across recent history, while during volatile periods, attention concentrates on the most recent observations to quickly capture regime changes. This adaptive behavior emerges naturally from training without requiring explicit regime-switching logic, demonstrating the power of modern deep learning architectures to discover relevant patterns in complex financial data.

The practical implications of this research are substantial for financial institutions engaging in options trading, portfolio hedging, and risk management activities. The framework provides a

ISSN: 3008-1394

tool that simultaneously optimizes for forecast accuracy and financial validity, addressing a critical gap in existing volatility forecasting methodologies. The multi-asset capability enables consistent volatility predictions across entire portfolios, facilitating coordinated hedging strategies and more accurate value-at-risk calculations. The framework's ability to maintain arbitrage-free properties ensures that forecasted surfaces can be directly used for pricing and risk calculations without requiring additional adjustments that might introduce inconsistencies or degrade accuracy.

Several directions for future research emerge from this work. First, extending the framework to incorporate other types of financial constraints beyond calendar and butterfly spread arbitrage could further enhance economic validity. For example, put-call parity relationships and bounds implied by underlying asset dynamics could be integrated into the constraint enforcement mechanism. Second, investigating alternative attention architectures such as multi-head attention or transformer-based approaches could potentially improve the model's ability to capture complex temporal dependencies across multiple time scales simultaneously. Third, incorporating exogenous information such as macroeconomic indicators, earnings announcements, or alternative data sources into the framework could potentially enhance forecast accuracy particularly during regime changes or event-driven volatility spikes.

The framework also opens possibilities for novel applications beyond traditional volatility forecasting. The controllable generation capability inherited from the VAE component could be leveraged for scenario analysis and stress testing, allowing risk managers to generate hypothetical volatility surfaces corresponding to specific market conditions. The latent space learned by the VAE might reveal insights into the fundamental factors driving volatility dynamics across assets, potentially informing trading strategies or portfolio construction approaches. The constraint enforcement methodology developed here could be adapted to other financial modeling problems where domain-specific restrictions must be satisfied, such as yield curve forecasting subject to no-arbitrage conditions or correlation matrix estimation subject to positive definiteness constraints.

In conclusion, this research demonstrates that carefully designed hybrid architectures combining complementary deep learning components with explicit enforcement of financial constraints can substantially advance the state-of-the-art in volatility forecasting. By addressing both predictive accuracy and economic validity within a unified framework enhanced by attention mechanisms, we provide a foundation for next-generation risk management tools that bridge the gap between cutting-edge machine learning techniques and rigorous financial theory. As financial markets continue to grow in complexity and interconnectedness, such integrated approaches that leverage data-driven learning while respecting fundamental constraints will become increasingly essential for effective risk assessment and strategic decision-making.

References

[1] Elias, M. E. (2025). Barbells in Hilbert Space: Nonlinear Risk, Quantum Inference, and the Collapse of Classical Finance. Toward a Post-Gaussian, Non-Ergodic Framework for Risk Management.

[2] Zheng, W., & Liu, W. (2025). Symmetry-Aware Transformers for Asymmetric Causal Discovery in Financial Time Series. Symmetry, 17(10), 1591.

[3] Buck, C., Ifland, S., Stähle, P., & Thorwarth, H. (2021). Raiders of the Lost Ark—A Review About the Roots and Application of Artificial Intelligence. International Journal of Innovation and Technology Management, 18(08), 2150045.

ISSN: 3008-1394

[4] Liu, Y. (2019). Novel volatility forecasting using deep learning—Long Short Term Memory Recurrent Neural Networks. Expert Systems with Applications, 132, 99-109.
[5] Olsen, A., Djupskås, G., de Lange, P. E., & Risstad, M. (2025). Forecasting implied volatilities of currency options with machine learning techniques and econometrics models. International Journal of Data Science and Analytics, 20(2), 1329-1347.
[6] Ge, Y., Wang, Y., Liu, J., & Wang, J. (2025). GAN-Enhanced Implied Volatility Surface Reconstruction for Option Pricing Error Mitigation. IEEE Access.
[7] Liu, R., Jiang, Y., & Lin, J. (2022). Forecasting the volatility of specifc risk for stocks with LSTM. Procedia Computer Science, 202, 111-114.
[8] Liu, S., Oosterlee, C. W., & Bohte, S. M. (2019). Pricing options and computing implied volatilities using neural networks. Risks, 7(1), 16.
[9] Medvedev, N., & Wang, Z. (2022). Multistep forecast of the implied volatility surface using deep learning. Journal of Futures Markets, 42(4), 645-667.
[10] Qing, X., Liao, Y., Wang, Y., Chen, B., Zhang, F., & Wang, Y. (2022). Machine learning based quantitative damage monitoring of composite structure. International journal of smart and nano materials, 13(2), 167-202.
[11] Zhang, J. A. (2025). Risk-Sensitive Option Market Making with Arbitrage-Free eSSVI Surfaces: A Constrained RL and Stochastic Control Bridge. arXiv preprint arXiv:2510.04569.
[12] Bergeron, M., Fuh, C., Reesor, R., & Whitehead, T. (2021). Variational autoencoders: A hands-off approach to volatility. arXiv preprint arXiv:2102.03945.
[13] Cao, J., Chen, J., Hull, J., & Poulos, Z. (2021). Deep hedging of derivatives using reinforcement learning. Journal of Financial Data Science, 3(1), 10-27.
[14] Fung, D., Kokkinakis, I. W., & Drikakis, D. (2025). Vector quantization variational autoencoder for turbulent flow images. Physics of Fluids, 37(8).
[15] Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. Foundations and Trends® in Machine Learning, 12(4), 307-392.
[16] Thorne, S. (2025). Interpretive description: Qualitative research for applied practice (p. 354). Taylor & Francis.
[17] Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.
[18] Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.
[19] Zhang, Q., Chen, S., & Liu, W. (2025). Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 17(6), 823.
[20] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. Applied Sciences, 15(2), 951.
[21] Chen, S., Liu, Y., Zhang, Q., Shao, Z., & Wang, Z. (2025). Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions. Advanced Intelligent Systems, 2400898.
[22] Mai, N. T., Cao, W., & Wang, Y. (2025). The global belonging support framework: Enhancing equity and access for international graduate students. Journal of International Students, 15(9), 141-160.
[23] Ma, Z., Chen, X., Sun, T., Wang, X., Wu, Y. C., & Zhou, M. (2024). Blockchain-based zero-trust supply chain security integrated with deep reinforcement learning for inventory optimization. Future Internet, 16(5), 163.
[24] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access.
[25] Cao, W., Mai, N. T., & Liu, W. (2025). Adaptive knowledge assessment via symmetric hierarchical Bayesian neural networks with graph symmetry-aware concept dependencies. Symmetry, 17(8), 1332.
[26] Mai, N. T., Cao, W., & Liu, W. (2025). Interpretable knowledge tracing via transformer-Bayesian hybrid networks: Learning temporal dependencies and causal structures in educational data. Applied Sciences, 15(17), 9605.
[27] Wang, Y., Ding, G., Zeng, Z., & Yang, S. (2025). Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery. IEEE Access.

ISSN: 3008-1394

[28] Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.
[29] Zhang, H. (2025). Physics-Informed Neural Networks for High-Fidelity Electromagnetic Field Approximation in VLSI and RF EDA Applications. Journal of Computing and Electronic Information Management, 18(2), 38-46.
[30] Qiu, L. (2025). Machine Learning Approaches to Minimize Carbon Emissions through Optimized Road Traffic Flow and Routing. Frontiers in Environmental Science and Sustainability, 2(1), 30-41.
[31] Li, J., Fan, L., Wang, X., Sun, T., & Zhou, M. (2024). Product demand prediction with spatial graph neural networks. Applied Sciences, 14(16), 6989.