# Enhancing VIX Shock Prediction via a Probabilistic Attention Transformer

Jin Su Kim[1,*], Zoonky Lee[2]

[1]*Korea Exchange, 40, Munhyeongeumyung-ro, Namgu, Busan, 48400, Republic of Korea*

[2]*Graduate School of Information, Yonsei University, 134 Shinchon, Seodaemun, Seoul 120-749, Republic of Korea*

(Received: May 10, 2025; Revised: August 02, 2025; Accepted: October 18, 2025; Available online: November 15, 2025)

**Abstract**

This study proposes a Probabilistic-Attention Transformer for forecasting abrupt shifts in the Volatility Index (VIX), advancing volatility modeling by directly embedding externally estimated shock probabilities into the attention mechanism. The core idea is to modify similarity-based attention scores with daily shock probabilities derived from stochastic diffusion equations, thereby enhancing the model's sensitivity to extreme-value dynamics. The primary objective is to improve predictive accuracy during market stress, particularly under warning ($20 \leq VIX \leq 30$) and shock ($VIX > 30$) regimes where conventional models often fail. Using 35 years of historical VIX data (1990–2024), the framework is benchmarked against GARCH (1,1) and a standard Transformer under distinct volatility regimes. Empirical findings show that the proposed model consistently outperforms alternatives: during warning regimes, prediction error is reduced by over 40% relative to both benchmarks, while in shock regimes, improvements exceed 50%, with performance gains validated by Diebold–Mariano tests at the 1% significance level. These results demonstrate both statistical and practical significance, offering risk managers and investors more reliable forecasts during periods of heightened market instability. The contribution of this research lies in providing not only empirical evidence of improved predictive performance but also a generalizable framework for integrating probabilistic indicators into deep learning architectures. The novelty is in showing that probabilistic weighting of attention can transform standard neural architectures into early-warning systems capable of capturing regime shifts in financial markets. Beyond VIX forecasting, this methodological contribution has broader applicability to equities, exchange rates, and commodities, where identifying and responding to volatility shocks is critical for risk management and investment decision-making.

*Keywords:* Attention Mechanism, Financial Time Series Forecasting, Stochastic Diffusion Equations, Transformer Models, Volatility Index (VIX)

## 1. Introduction

The VIX, developed by the Chicago Board Options Exchange (CBOE), is a widely recognized measure of market volatility derived from S&P 500 index option prices. Often referred to as the "fear index," the VIX serves as a key indicator of market risk and uncertainty, providing essential insights for gauging market expectations and managing financial risk [1]-[5]. Analysts, investors, and risk managers rely on the VIX to assess market sentiment and inform strategic decision-making. Like a barometer, VIX movements offer critical insight into anticipated market shifts [6]-[10]. Therefore, accurately forecasting the VIX is of significant importance for enhancing market stability, investment efficiency, and risk management effectiveness. Particularly, the ability to detect and predict sharp VIX surges—indicative of market shocks—has become increasingly critical in innovative risk management.

Given this importance, numerous researchers have explored methodologies to improve VIX time-series forecasting accuracy. Among traditional models, the GARCH framework has been widely used due to its strength in modeling volatility clustering—a frequent characteristic of financial time series. The GARCH model was recognized with the Nobel Prize in Economics in 2003, with numerous studies employing GARCH for time-series and volatility prediction [11]-[21]. However, despite its popularity, GARCH struggles to capture nonlinear and abrupt shifts often observed in real-world volatility patterns [12], [22], [15]. Recently, deep learning techniques have shown promise in learning complex, nonlinear relationships, opening new opportunities for volatility forecasting [23]-[34]. Among these

***Corresponding author: Jin Su Kim (ko76martin@gmail.com)**
**DOI: https://doi.org/10.47738/jads.v6i4.947**

approaches, the Transformer model—originally developed for natural language processing—has demonstrated exceptional performance in time-series forecasting, outperforming conventional models such as RNNs, LSTMs, and CNNs [35]-[40]. The Transformer's strength lies in its attention mechanism, which considers all sequence positions simultaneously, making it well-suited for modeling complex, nonlinear volatility patterns.

Although several studies have explored Transformer attention mechanism modifications [41], [42], their focus differs markedly from this study. Shaw & Vaswani [41] introduced relative positional encodings for longer sequence generalization, while Wang et al. [42] proposed modeling long-range dependencies via decomposition and auto-correlation mechanisms. In contrast, our study introduces a probabilistic weighting scheme into attention scores using daily volatility shock probabilities estimated from the Heston stochastic model—specifically designed to enhance detection of abrupt volatility shifts in financial time series. While attention mechanisms were originally developed for NLP sequential dependencies, few studies have explored adaptive attention weighting for financial time series, particularly for capturing sudden volatility changes. We propose integrating daily VIX shock probabilities—estimated using the Heston stochastic differential equation—directly into attention weights to emphasize periods with higher likelihood of abrupt volatility and improve anticipation of such events.

Given volatility dynamics' stochastic nature and VIX mean-reverting properties, we hypothesize that periods with elevated shock probabilities can serve as early indicators of impending volatility regime transitions. The proposed model incorporates these shock probabilities as dynamic attention score modifiers, replacing conventional similarity-only structures. This study evaluates whether the proposed model significantly outperforms existing models in capturing volatility shocks—particularly during sharp VIX increases and heightened market panic. Superior forecasting performance under market instability would offer valuable tools for managing financial uncertainty and contribute meaningfully to risk management.

This paper is organized as follows: Chapter 2 presents VIX index historical trends over 35 years and theoretical foundations of three key volatility forecasting models: traditional GARCH, standard Transformer, and the proposed modified Transformer incorporating daily volatility shock probabilities. Chapter 3 evaluates and compares these models' forecasting performance using historical VIX data, and Chapter 4 concludes based on empirical findings.

## 2. Method

### 2.1. Summary statistics of the VIX

This study analyzes the daily VIX index over the 8,833 trading days from January 2, 1990 to December 31, 2024. The summary statistics of the VIX index over the past 35 years are presented in table 1.

**Table 1.** Descriptive Statistics of the VIX from 1990 to 2024

| Statistic | Number of Observations | Mean | Standard Deviation | Minimum | First Quartile (25%) |
|---|---|---|---|---|---|
| Value | 8,833 | 19.48 | 8.35 | 9.14 | 13.78 |
| Statistic | Median (50%) | Third Quartile (75%) | Maximum | Skewness | Kurtosis |
| Value | 17.87 | 23.12 | 82.69 | 2.16 | 9.84 |

The VIX index averages 19.48, reflecting normal market uncertainty under typical conditions. The skewness of 2.16 indicates a long right tail, showing that while market stress episodes are rare, they occur with high intensity. The kurtosis value of 9.84 reveals a leptokurtic distribution with higher concentration around the mean and in the tails, indicating pronounced non-normality and asymmetry that contribute to forecasting difficulty. These distributional characteristics guided key modeling decisions. The heavy-tailed nature with extreme outliers motivated choosing the Transformer architecture over traditional linear models, as attention mechanisms can adaptively focus on rare but critical volatility events. For GARCH modeling, we applied log-return transformations to address non-normality. The proposed model incorporates externally estimated shock probabilities from the Heston stochastic model to directly capture extreme value behavior in the attention mechanism. We conducted regime-specific analysis across stable, warning, and shock periods to account for asymmetric error behavior, recognizing that traditional aggregate metrics may be insufficient for heavy-tailed distributions where extreme observations dominate overall performance measures.

## 2.2. Data Preprocessing

The daily VIX index data contains no missing values, as it is officially published by the CBOE on all trading days. The only gap occurs from September 10-17, 2001 due to market closure following 9/11, which we retain without imputation to preserve true volatility dynamics during this significant economic event. This study employs different stationarity tests because each model operates on fundamentally different input types. Transformer models directly process VIX level series, requiring stationarity of the original time series for stable neural network training and reliable pattern recognition. An ADF test on normalized VIX levels (statistic = −7.02, $p < 0.001$) confirms this requirement. In contrast, GARCH models the conditional variance of return series, not levels. The econometric framework of GARCH requires stationary returns as input, necessitating transformation of VIX levels into log returns: $r_t = \log(VIX_t) - \log(VIX_{t-1})$. An ADF test on returns (statistic = −9.23, $p < 0.001$) confirms stationarity for GARCH modeling. This dual-track approach reflects the distinct theoretical foundations: Transformers learn from level patterns while GARCH models return volatility dynamics.

The VIX is widely regarded as a measure of future market uncertainty. Following the classification commonly adopted in prior literature [3], [4], [7], [9], VIX levels are typically divided into three regimes. A level of VIX below 20 is interpreted as a Stable Volatility Regime, reflecting calm market conditions and low investor fear. A range of $20 \leq VIX \leq 30$ represents a Warning Volatility Regime, suggesting heightened investor caution and rising market uncertainty. Finally, a level of VIX above 30 indicates a Shock Volatility Regime, associated with elevated market fear and extreme stress.

The VIX > 30 threshold for shock regimes is widely adopted in financial literature and statistically justified, representing approximately two standard deviations above the historical mean (19.48) and the 94th percentile of the distribution. This serves as a natural breakpoint between manageable stress levels and acute market distress requiring immediate risk management interventions. To ensure robustness, we validate this choice using alternative thresholds (VIX > 25 and VIX > 35) in Section 3.2, ensuring our findings are not dependent on a single cutoff.

As shown in table 2, analysis over 35 years indicates 62.31% of observations fall within the stable regime, 29.51% in the warning regime, and 8.18% in the shock regime—the primary focus of this study.

**Table 2.** Summary statistics for VIX Shocks based on VIX range

| Category (VIX Range) | Number of Observations (days) | Proportion (%) | Avg Duration (days) |
|---|---|---|---|
| Stable (<20) | 5,504 | 62.31 | 43.33 |
| Warning (20≤ ≥30) | 2,606 | 29.51 | 11.05 |
| Shock (>30) | 723 | 8.18 | 10.03 |

## 2.3. VIX Prediction Using Traditional Financial Models (GARCH)

This study aims to show that the proposed model predicts abrupt VIX fluctuations more effectively than existing models, benchmarking its performance against established volatility forecasting approaches.

The GARCH model addresses the unrealistic assumption of constant volatility over time by capturing conditional heteroskedasticity, where past volatility influences current volatility. Originally proposed by Engle [43] and established by Bollerslev [44], earning the 2003 Nobel Prize in Economic Sciences, the GARCH model is particularly well-suited for forecasting conditional variance in financial markets. GARCH models consist of Mean and Conditional Variance equations. The GARCH (m, s) model [44] assumes that VIX index ($r_t$) follows an ARMA (p, q) process:

$$\text{Mean Equation} \quad r_t = \mu + \sum_{i=1}^{p} \emptyset_i \, r_{t-i} + \sum_{j=1}^{q} \theta_i \, \varepsilon_{t-j} \tag{1}$$

$r_t$ is a time series variable (e.g., VIX), $\mu$ is constant term, $\emptyset_i$ is parameter of the autoregressive (AR) captures the dependency on past returns, and $\theta_i$ is parameter of the moving-average (MA) captures the effect of past shocks.

$\varepsilon_t$ is residual term following a time varying variance process. The error term ($\varepsilon_t$) is scaled by the time-varying volatility: $\varepsilon_t = \sigma_t z_t$ (2). The conditional variance follows a GARCH (m, s) process:

 ISSN 2723-6471


$$\text{Variance Equation} \quad \sigma_t^2 = \omega + \sum_{i=1}^{s} \alpha_i \, \varepsilon_{t-i}^2 + \sum_{j=1}^{m} \beta_j \, \sigma_{t-j}^2 \tag{2}$$

$\omega$ is a positive constant representing the baseline variance level, ARCH effect ($\alpha_i$) measures the impact of past squared errors ($\varepsilon_{t-i}^2$) on current variance, GARCH effect ($\sigma_{t-j}^2$) captures the persistence of past variances ($\sigma_{t-j}^2$) over time. The methodology for forecasting the VIX data over the study period using the GARCH model is outlined as follows: the VIX data from the most recent 30 trading days are transformed into log returns. The GARCH(1,1) specification is adopted due to its simplicity and strong performance across financial time series [45], [46], [47]. Using the fitted model, volatility is forecasted up to 5 trading days ahead, with predicted volatilities converted back to actual VIX levels. This forecasting process uses a 5-trading-day sliding window approach across the entire study period, ensuring temporal validation where no future information influences historical predictions. Performance is assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE):

$$\text{MSE (Mean Squared Error)} = \frac{1}{N}\sum_{t=1}^{N}(VIX_t^{actual} - VIX_t^{Predicted})^2 \tag{3}$$

$$\text{RMSE (Root Mean Squared Error)} = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(VIX_t^{actual} - VIX_t^{Predicted})^2} \tag{4}$$

$$\text{MAPE (Mean Absolute Percentage Error)} = \frac{1}{N}\sum_{t=1}^{N}\frac{VIX_t^{actual} - VIX_t^{Predicted}}{VIX_t^{actual}} \times 100 \tag{5}$$

## 2.4. VIX Prediction Using Decoder-Only Transformer Model

The GARCH model's assumption of mean reversion limits its ability to detect nonlinear behaviors and abrupt changes caused by extreme events—the primary focus of this study. While GARCH extensions like EGARCH [48] and GJR-GARCH [49] address asymmetric reactions, they frequently fall short in representing complex nonlinear dynamics and rapid regime changes characteristic of major volatility shocks. The Transformer architecture, based on attention mechanisms, allows parallel computation without sequential processing, demonstrating superior performance in time-series forecasting [34], [35], [36], [38], [39], [40]. For time series forecasting, most Transformer architectures adopt a decoder-only structure [50], [51], [52] due to direct correlation and temporal continuity between input and output sequences. The detailed procedure for implementing these training methodologies with a Decoder-only Transformer model is as follows: In a Decoder-only Transformer for time series forecasting, input embedding maps one-dimensional values (e.g., VIX index) into a higher-dimensional vector space, while positional encoding adds relative or absolute time information using sine and cosine functions. Masked Multi-Head Self-Attention (MHA) then captures temporal dependencies and both short and long-term patterns by allowing each time point to attend only to itself and past time steps. This is enforced via a causal mask—a lower-triangular T×T matrix with entries above the diagonal set to $-\infty$ before the SoftMax operation—applied to all heads during training and inference. This preserves the autoregressive property and prevents information leakage from the future. Self-attention operates on Query, Key, and Value representations to learn relationships across time steps from multiple perspectives (heads). The time series input data undergoes linear transformation through learnable weight matrices $W_q$, $W_k$, $W_v$ to generate the Query ($Q$), Key ($K$), and Value ($V$) matrices, respectively.

$$Q = XW_q, K = XW_k, V = XW_v \tag{6}$$

$Q$ (Query): A vector used by the token at the current position to calculate its relevance to other tokens. $K$ (Key): A vector used to determine how relevant each token is to another token's query. $V$ (Value): A vector that transmits information, weighted by relevance scores through a weighted average.

The relevance between input time series data points is quantified through Attention Scores, which determine which input data should receive greater attention for future time series prediction. The $Attention\ Score_{ij}$ utilizes scaled dot-product attention, computed as the inner product of $Query_i$ and $Key_j$ matrices divided by the square root ($\sqrt{d_k}$) of the Key dimension (where $Key_j^T$ is transposed for the inner product calculation). This scaling mechanism ($\sqrt{d_k}$) prevents gradient instability that can occur with increasing Key dimensions

 ISSN 2723-6471


$$Attention\ Score_{ij} = \frac{Query_i \cdot Key_j^T}{\sqrt{d_k}} \quad (7)$$

The Attention score is normalized to probability values through the application of the SoftMax function (Equation 8). Subsequently, these weights are multiplied by Value and summed to generate a contextual representation (a term preferred over "context vector" when describing Decoder-only architectures) (Equation 9). Through this process, the extent to which each time point incorporates information from other time points is determined.

$$Attention\ Weights_{ij} = Softmax\ (\frac{Query_i \cdot Key_j^T}{\sqrt{d_k}}) \quad (8)$$

$$Contextual\ Representation = \sum_j^n Attention\ Weights_{ij} \cdot Value_j \quad (9)$$

In the third step, Multi-Head Attention operates with multiple attention heads in parallel, enabling the model to learn complex and hierarchical patterns in time series data. The outputs from each head are concatenated and passed to subsequent layers. For the final output stage, while the standard Transformer applies a linear transformation followed by a SoftMax (suitable for NLP), we omit the SoftMax to directly predict continuous values such as the VIX index. A linear layer maps the hidden dimension to the output dimension, producing forecasts for the specified horizon.

We evaluate the Decoder-only Transformer's predictions against actual VIX values using two rolling-window approaches: (1) Traditional Split Rolling Window (TSRW) — Uses the preceding 30 trading days for training to predict the next 5 days, shifting the 35-day window forward in 5-day increments, without cross-validation. (2) Cross-Validation Rolling Window (CVRW) — Uses a 50-day window split into training (60%, 30 days), validation (20%, 10 days), and test (20%, 10 days). The window advances by 10 days, incorporating cross-validation for hyperparameter tuning and ensuring unbiased performance evaluation. In both methods, the sliding window length (5 or 10 days) matches the forecast horizon to avoid overlap between prediction periods. Performance is measured using the same metrics as the GARCH model: MSE, RMSE, and MAPE. This framework allows the model to adapt to evolving patterns while maintaining temporal relevance.

This study employs both TSRW and CVRW in parallel to exploit their complementary strengths. TSRW offers a simple, efficient framework for repeated short-term (5-day) forecasts, while CVRW incorporates a validation set for hyperparameter tuning, enabling the model to capture longer-term patterns. Using both approaches provides a comprehensive evaluation across varying market conditions, from high volatility to stability, and reduces bias toward specific data segments. This dual strategy enhances robustness by testing the model under diverse temporal contexts. Statistical analysis reveals systematic performance reversals across different volatility regimes, highlighting the complementary nature of these evaluation approaches. TSRW demonstrates advantages during stable and warning periods, with improvements of +3.72% and +3.55% respectively over CVRW, likely due to its more responsive adaptation to gradual volatility changes. Conversely, CVRW excels during shock periods with a +22.32% advantage over TSRW, suggesting that the validation-based approach with longer optimization windows provides better performance during extreme market conditions. This complementarity validates the methodological rigor of our dual-track evaluation approach and ensures that our conclusions are robust across different market conditions and evaluation frameworks. The conceptual frameworks of both methods are shown in figure 1.

ISSN 2723-6471


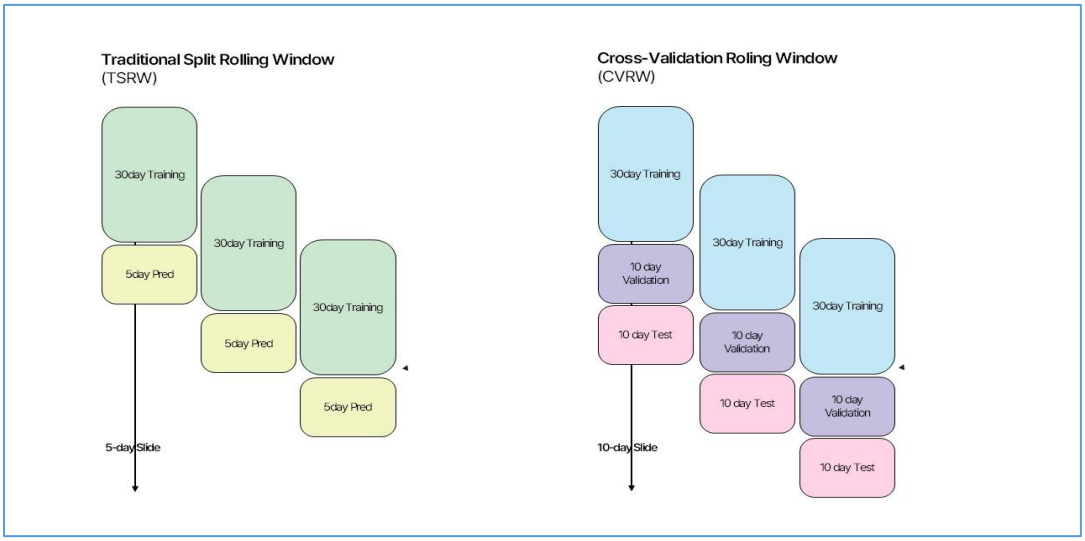


**Figure 1.** VIX Time Series Model Training Methods: TSRW vs CVRW

Note: In TSRW, a 30-day training window is used to predict the following 5 days, after which the window slides forward by 5 days. In CVRW, the 50-day window consists of 30 days for training, 10 days for validation, and 10 days for testing, sliding forward by 10 days.

## 2.5. VIX Prediction Using a Proprietary Modified Transformer Model

While the established Transformer architecture has shown strong performance in modeling sequential data through its self-attention mechanism, it is limited in capturing rare but abrupt surges in volatility, such as VIX shocks. This stems from its equal treatment of all time steps, which can dilute the impact of critical precursors to volatility spikes—particularly during transition phases. Without incorporating external indicators reflecting the probability of extreme events, the standard Transformer may underperform under heightened market uncertainty. In this study, the external indicator enhancing the Transformer's predictive capability is the daily VIX shock probability estimated via a Stochastic Diffusion Equation (SDE). The SDE framework effectively captures the randomness, asymmetry, and mean-reverting nature of volatility dynamics, while enabling probabilistic modeling of extreme events—characteristics that conventional linear models fail to address. By modeling volatility as a stochastic process, SDEs allow for probabilistic estimation of future shocks, serving as a critical input to the proposed attention-based model. SDEs extend deterministic models by adding random components—typically Wiener processes—to represent inherent uncertainty in time-evolving variables such as asset prices and volatility [57]. To better account for elevated and nonlinear volatility during extreme market stress, we adopt the Heston model [54], which introduces a stochastic process for variance itself, thereby modeling the volatility of volatility.

This dual-process structure captures volatility clustering and abrupt spikes in both the level and variability of the VIX, providing a more realistic framework for turbulent markets. Given this, we ask: how can daily VIX shock probabilities from the Heston model be integrated into the Transformer? Focusing on the self-attention mechanism, which quantifies temporal dependencies via similarity scores between all positions in the input, we modify it by incorporating these probabilities as multiplicative weights in the attention score calculation. This probabilistic weighting emphasizes periods with a higher likelihood of abrupt volatility, enhancing the model's ability to detect and forecast VIX shocks. This section first outlines the procedure for estimating daily VIX shock probabilities using the Heston model, then explains their integration into the Transformer.

### 2.5.1. Estimation of Daily Volatility Shock Probabilities Using Stochastic Diffusion Equations

This study aims to estimate the probability that the VIX index will exceed a specified threshold level (i.e., 30, indicating a volatility shock) using a stochastic diffusion equation (SDE) approach. A stochastic differential equation is a mathematical model that extends deterministic differential equations by incorporating stochastic components—typically represented by a Wiener process or Brownian motion. SDEs are widely used in financial modeling to describe and forecast the dynamic behavior of variables that evolve with uncertainty over time, such as asset prices and volatility

[57]. A key feature of the SDE applied to VIX modeling is its mean-reverting property, which captures the tendency of the VIX index to revert toward a long-term average level, denoted by $\theta$(long-term mean of VIX). The speed of this reversion is governed by the parameter $k$, while the stochastic component, represented as $\sigma dW_t$, reflects market uncertainty and captures the irregular fluctuations of the VIX. This study estimates the probability that the VIX index will exceed a specified threshold (30, indicating a volatility shock) using an SDE approach. An SDE extends deterministic differential equations by incorporating stochastic components—typically modeled via a Wiener process or Brownian motion—and is widely applied in finance to describe the uncertain dynamics of variables such as asset prices and volatility [57]. In VIX modeling, a key feature of the SDE is its mean-reverting property, capturing the index's tendency to revert to a long-term mean $\theta$. The reversion speed is controlled by parameter $k$, while the stochastic term $\sigma dW$ reflects market uncertainty and irregular fluctuations in the VIX.

$$dV_t = k(\theta - V_t)d_t + \sigma dW_t \tag{10}$$

$V_t$: VIX Value, $k$:Mean reversion speed (the rate at which VIX returns to its long-term mean), $\theta$: Long-term mean of VIX, $\sigma$: Volatility (Fixed), $dW_t$:Stochastic noise (Wiener process), $d_t$: infinitesimal increment of time

While the SDE captures the mean-reverting nature of time series data, it struggles to reflect the heightened volatility seen during extreme market shocks—the focus of this study. To address this, we adopt the Heston model [54], which explicitly models the volatility of volatility in the VIX index. As shown in Equation 11, the first equation describes VIX dynamics, and the second models the stochastic process of its variance. The two processes, $dW_t^V$ and $dW_t^v$, are correlated via $\rho$ allowing joint dynamics between VIX and its volatility. By capturing volatility clustering and sudden spikes in both levels and variance, the Heston model offers a more realistic framework for stressed market conditions.

$$dV_t = k(\theta - V_t)d_t + \sqrt{v_t}dW_t^V \text{ (12)},\ dv_t = \xi(\eta - v_t)d_t + \sigma_v\sqrt{v_t}dW_t^v \tag{11}$$

$V_t$ :VIX Value, $v_t$: Stochastic process of volatility (volatility process), $k$ : Mean reversion speed (the rate at which VIX returns to its long-term mean), $\theta$ Long-term mean of VIX, $\xi$:volatility of volatility, $\eta$:Long-term mean of volatility, $\sigma_v$ :Volatility coefficient, $dW_t^V$: Two separate stochastic Wiener processes (with correlation ρ), $d_t$: infinitesimal increment of time

Finally, the probability of the VIX exceeding 30 is estimated via Monte Carlo simulation using the Heston model. To avoid data leakage, model parameters ($k$, $\theta$, $\xi$, $\eta$, $\sigma_v$, $\rho$) are re-estimated at each prediction point using only prior data. This temporal cross-validation ensures forecasts rely solely on historical information. From 10,000 simulated paths, the probability is the proportion in which the VIX surpasses 30.

$$p(u(t) > 30) \approx \frac{Number\ of\ Paths\ where\ u(t)>30}{N=10{,}000} \tag{12}$$

The estimated daily probabilities of volatility shocks—defined as the probability that the VIX index exceeds the critical threshold of 30—are computed for the entire study period. These probabilities are subsequently incorporated as attention weights in the Transformer model, as described in the following sections, to enhance predictive accuracy, particularly during periods of sharp VIX increases. When the VIX exceeds 30, the probability approaches 1, leading to post hoc emphasis rather than true prediction. To mitigate this, we apply a "+10 multiple-threshold" approach: if the VIX surpasses 30, we estimate the probability of exceeding 40; if it crosses 40, we estimate for 50, and so on, using a stochastic diffusion equation. The Heston model's VIX shock probabilities (threshold = 30) are evaluated via confusion matrix (table 3), achieving an AUC of 0.88 with only 110 false negatives, indicating a very low miss rate. Overall, the Heston model demonstrates strong capability in capturing the probability of VIX shocks, effectively identifying periods of major financial crises and market stress.

**Table 3.** Confusion Matrix

| Actual Values | Predicted: Non-Shock | Predicted: Shock |
|---|---|---|
| **Non-Shock** | True Negative (TN): 7,870 | False Positive (FP): 240 |
| **Shock** | False Negative (FN): 110 | True Positive (TP): 613 |

Note: A classification threshold of 50% is applied, whereby a predicted VIX shock is assumed to occur if the estimated probability exceeds 50%. These predictions are then compared against the actual VIX values to evaluate the model's

classification performance. The confusion matrix summarizes the classification results for predicted VIX shocks (probability > 50%) versus actual outcomes. True Positives (TP) indicate correctly predicted shocks, True Negatives (TN) are correctly predicted non-shocks, False Positives (FP) are non-shocks incorrectly predicted as shocks, and False Negatives (FN) are shocks missed by the model.

### 2.5.2. The proposed Shock-Weighted Attention Transformer

This study distinguishes itself from existing Transformer-based approaches by modifying the model's core attention mechanism. We incorporate daily volatility shock probabilities—calculated using the Heston model as the probability that the VIX exceeds 30—directly into the attention weights. This enables the model to emphasize periods characterized by market shocks, improving its ability to learn from and predict extreme events.

The derivation of our shock-weighted attention mechanism follows a systematic approach to integrate external probabilistic information into the standard Transformer attention computation. We begin with the standard scaled dot-product attention mechanism, where the attention score between positions i and j for head h is computed as:

$$Score_{ij}^{(h)} = \left( \frac{Q_i^{(h)} \bullet \left(K_j^{(h)}\right)^T}{\sqrt{d_k}} \right) \quad (A.1) \tag{13}$$

To incorporate shock probabilities into this framework, we first preprocess the raw shock probabilities $P_{\text{shock},j}$ from the Heston model to ensure numerical stability. Raw shock probabilities are clipped to prevent extreme values: $\tilde{P}_j = \text{clip}\left(P_{\text{shock},j}, \varepsilon, 1-\varepsilon\right)$ where $\varepsilon = 1 \times 10^{-6}$ (A.2). This preprocessing step prevents extreme probability values from destabilizing the attention computation while maintaining the essential information content.

The core innovation lies in designing a parametric function to transform these shock probabilities into attention modifiers. We employ the sigmoid function to provide smooth, bounded transformations that allow each attention head to learn how to respond to different levels of shock probability. The parametric shock weight function is designed as: $f\left(\tilde{P}_j\right) = \sigma\left(\alpha^{(h)} \bullet \tilde{P}_j + \gamma^{(h)}\right)$ (A.3), where the sigmoid function ensures bounded outputs and the trainable parameters $\alpha^{(h)}$ and $\gamma^{(h)}$ control the sensitivity and threshold of the response to shock probabilities. To preserve the baseline attention mechanism while enabling amplification during critical periods, we construct multiplicative weights that maintain the fundamental attention structure. The multiplicative weight construction follows: $\left(W_{\text{shock}}^{(h)}\right)_{ij} = 1 + \beta^{(h)} \bullet f\left(\tilde{P}_j\right)$ (A.4). This formulation ensures that when shock probabilities are low, the attention weights remain close to their original values (approaching unity), but during periods of high shock probability, attention can be significantly amplified based on the learned parameter $\beta^{(h)}$ which controls the maximum amplification scale.

The integration with standard attention combines the similarity-based computation with the shock-based weighting through element-wise multiplication. The final modified attention score integrates both components: Modified Attention $\text{Score}_{ij}^{(h)} = \text{Score}_{ij}^{(h)} \bullet W_{\text{shock}}^{(h)}(j)$ (A.5). Substituting equations (A.1), (A.3), and (A.4) into (A.5), we derive our complete formulation:

$$Modified\ Attention\ Score_{ij}^{(h)} = \left( \frac{\mathbf{Q}_i^{(h)} \bullet \left(\mathbf{K}_j^{(h)}\right)^T}{\sqrt{d_k}} \right) \bullet \left[1 + \beta^{(h)} \sigma\left(\alpha^{(h)} \bullet \tilde{P}_j + \gamma^{(h)}\right)\right] \tag{14}$$

This derivation ensures that our modification maintains the fundamental properties of attention—differentiability, bounded amplification, and baseline preservation—while incorporating external shock probability information in a principled manner. The mathematical formulation guarantees that the modified attention mechanism preserves all desirable properties of the original attention while adding the capability to dynamically emphasize periods of market stress. In Equation 14, $\tilde{P}_j$ denotes the probability that the VIX exceeds 30 at time step j, estimated using the Heston model. The parameters $\beta^{(h)}$, $\alpha^{(h)}$ and $\gamma^{(h)}$ are head-specific trainable parameters controlling amplification scale, sensitivity to shock probabilities, and threshold adjustment, respectively. The sigmoid function $\sigma(\bullet)$ ensures smooth and bounded modifications. This design preserves the underlying similarity structure while adaptively emphasizing critical time steps based on learned shock-response patterns.

The shock-weighted attention parameters $\{\beta^{(h)}, \alpha^{(h)}, \gamma^{(h)}\}$ are optimized via backpropagation with L2 regularization to prevent overfitting and ensure stable training. The regularization term is formulated as:

$$L_{\text{reg}} = \lambda_{\text{shock}} \bullet \sum_h \left[\left(\beta^{(h)}\right)^2 + \left(\alpha^{(h)}\right)^2 + \left(\gamma^{(h)}\right)^2\right] \quad (A.7) \tag{15}$$

$\lambda_{\text{shock}} = 0.001$ based on empirical validation. This regularization penalty is added to the main loss function during training to constrain parameter magnitudes and promote generalization. To ensure stable training and effective learning, we employ a carefully designed initialization strategy that balances initial responsiveness with training stability. The initialization follows specific distributional assumptions: $\beta^{(h)} \sim N(0.1, 0.01)$ provides small positive values to start with minimal shock influence, allowing the model to gradually learn appropriate amplification levels; $\alpha^{(h)} \sim N(1.0, 0.1)$ ensures near-unity sensitivity for initial responsiveness to shock probability variations; and $\gamma^{(h)} \sim N(0.0, 0.01)$ provides zero-centered bias terms to avoid initial bias toward any particular shock probability threshold. During training, parameters $\beta^{(h)}$ (maximum attention amplification), $\alpha^{(h)}$ (sensitivity to shock probability variations), and $\gamma^{(h)}$ (shock probability threshold) are optimized via gradient descent alongside all other model parameters. This optimization flexibility enables each attention head to develop specialized shock-detection capabilities, with some heads potentially focusing on early warning indicators characterized by moderate shock probabilities, while others specialize in confirmed high-probability shock events. This diversification of attention head functions enhances the model's ability to capture different aspects of volatility shock dynamics.

The modified attention scores integrate seamlessly within the Multi-Head Attention mechanism following the standard Transformer architecture. Each attention head computes shock-weighted scores independently using its own set of parameters $\{\beta^{(h)}, \alpha^{(h)}, \gamma^{(h)}\}$, allowing for specialized learning. The outputs from all attention heads are then concatenated and linearly transformed to form the final contextual representation, maintaining computational efficiency while incorporating shock-aware processing. Regularization techniques extend beyond the L2 penalty to include gradient clipping, which prevents exploding gradients that could destabilize training when processing extreme shock events. The gradient clipping threshold is set based on empirical validation to balance training stability with learning effectiveness. These combined regularization approaches ensure stable parameter convergence and prevent overfitting to extreme shock events, which is crucial given the relatively rare occurrence of high-magnitude volatility shocks in the training data.

## 3. Results and Discussion

All model components, including Heston model parameter estimation, follow temporal cross-validation, using only data available at each prediction point to prevent information leakage. Three models are used for VIX forecasting: the traditional GARCH, a standard Transformer, and the Modified Transformer developed in this study, which incorporates volatility shock probabilities from a diffusion equation. Forecasts are evaluated against actual VIX values using two training-validation schemes—TSRW and CVRW—and the consistency of results across both confirms the robustness of our findings.

### 3.1. Performance Comparison Over the Entire Study Period

Table 4 summarizes performance metrics for the traditional GARCH, the Transformer, and the Modified Transformer incorporating VIX shock probabilities. Based on MSE, performance ranks as follows: 8.75 (GARCH) > 7.70 (Transformer–TSRW) > 6.98 (Transformer–CVRW) > 4.39 (Modified Transformer–TSRW) > 4.66 (Modified Transformer–CVRW). The Modified Transformer (TSRW) improves accuracy by 49.82% over GARCH and 43.00% over the Transformer (TSRW). Overall, the Transformer surpasses GARCH, and the Modified Transformer delivers the best results, consistently outperforming across all metrics (RMSE, MSE, MAPE) and both evaluation methods.

**Table 4.** Total Performance Comparison

| | GARCH ($\mu \pm SD$) | Transformer | | Modified Transformer | |
|---|---|---|---|---|---|
| | | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) |
| RMSE | 2.6986 ± 0.065 | 2.5453 ± 0.061 | 2.4278 ± 0.058 | 1.9934 ± 0.047 | 2.0284 ± 0.049 |

| | | | | | |
|---|---|---|---|---|---|
| | [2.570, 2.826] | [2.426, 2.664] | [2.314, 2.541] | [1.902, 2.085] | [1.934, 2.123] |
| MSE | 8.7469 ± 0.42 | 7.7006 ± 0.38 | 6.9819 ± 0.34 | 4.3890 ± 0.24 | 4.6586 ± 0.25 |
| | [7.93, 9.57] | [6.97, 8.43] | [6.31, 7.65] | [3.92, 4.86] | [4.18, 5.14] |
| MAPE (%) | 9.8577±0.28 | 9.6312 ± 0.27 | 8.4386 ± 0.24 | 7.6108 ± 0.22 | 7.9662 ± 0.23 |
| | [9.31, 10.41] | [9.10, 10.16] | [7.98, 8.90] | [7.20, 8.02] | [7.56, 8.38] |

Note: Values are presented as mean ± Standard Deviation (SD), with the 95% Confidence Interval (CI) shown in brackets []. Narrower CIs indicate higher precision, and non-overlapping CIs between models suggest statistically significant differences. Please refer to <Equations 4 to 6> for the calculation methods of the performance metrics RMSE, MSE, and MAPE. For the calculation methods of TSRW and CVRW.

## 3.2. Performance Comparison Across Different VIX Levels

This study evaluates model performance across three VIX regimes—Stable (VIX < 20), Warning (20 ≤ VIX ≤ 30), and Shock (VIX > 30)—to assess robustness under varying market conditions. During Stable periods in table 5, the Modified Transformer records the lowest MSE (2.79) versus 3.67 for the Transformer and 4.25 for GARCH, yielding 34.32% and 24.02% improvements over GARCH and Transformer (TSRW), respectively. However, gains are modest compared to other regimes, suggesting that most accuracy improvements occur in higher-volatility phases.

**Table 5.** Performance during the stable period

| | **GARCH ($\mu \pm SD$)** | **Transformer** | | **Modified Transformer** | |
|---|---|---|---|---|---|
| | | **TSRW ($\mu \pm SD$)** | **CVRW ($\mu \pm SD$)** | **TSRW ($\mu \pm SD$)** | **CVRW ($\mu \pm SD$)** |
| RMSE | 2.0619 ± 0.048 | 1.9169 ± 0.045 | 1.8532 ± 0.043 | 1.6829 ± 0.040 | 1.7011 ± 0.041 |
| | [1.967, 2.157] | [1.828, 2.006] | [1.768, 1.938] | [1.604, 1.761] | [1.621, 1.781] |
| MSE | 4.2512 ± 0.21 | 3.6745 ± 0.18 | 3.4343 ± 0.17 | 2.7920 ± 0.15 | 2.8998 ± 0.15 |
| | [3.84, 4.66] | [3.32, 4.03] | [3.10, 3.77] | [2.50, 3.08] | [2.60, 3.20] |
| MAPE (%) | 10.0242 ± 0.26 | 9.8274 ± 0.25 | 7.9901 ± 0.21 | 7.3527 ± 0.19 | 7.9899 ± 0.21 |
| | [9.52, 10.53] | [9.34, 10.31] | [7.58, 8.40] | [6.97, 7.73] | [7.58, 8.40] |

Note: Please refer to the notes below table 4 for a detailed explanation of the legend.

In the Warning regime (29.51% of the sample) in table 6, where the VIX is in a critical transition zone, the Modified Transformer achieves an MSE of 4.98, outperforming GARCH (8.92) and Transformer (8.61) by 44.13% and 42.10%, respectively. This indicates its strongest relative advantage before volatility shocks fully develop.

**Table 6.** Performance during the warning period

| | **GARCH ($\mu \pm SD$)** | **Transformer** | | **Modified Transformer** | |
|---|---|---|---|---|---|
| | | **TSRW ($\mu \pm SD$)** | **CVRW ($\mu \pm SD$)** | **TSRW ($\mu \pm SD$)** | **CVRW ($\mu \pm SD$)** |
| RMSE | 2.9869 ± 0.075 | 2.9341 ± 0.073 | 2.7449 ± 0.069 | 2.2300 ± 0.057 | 2.4288 ± 0.062 |
| | [2.840, 3.134] | [2.792, 3.076] | [2.610, 2.880] | [2.118, 2.342] | [2.308, 2.550] |
| MSE | 8.9214 ± 0.44 | 8.6087 ± 0.42 | 7.5343 ± 0.37 | 4.9841 ± 0.27 | 5.1675 ± 0.28 |
| | [8.06, 9.78] | [7.78, 9.44] | [6.83, 8.24] | [4.46, 5.51] | [4.63, 5.71] |
| MAPE (%) | 9.0827 ± 0.24 | 8.9101 ± 0.23 | 8.6413 ± 0.23 | 7.5979 ± 0.21 | 7.8304 ± 0.22 |
| | [8.62, 9.55] | [8.45, 9.37] | [8.20, 9.08] | [7.19, 8.00] | [7.43, 8.23] |

Note: Please refer to the notes below table 4 for a detailed explanation of the legend.

During Shock periods in table 7, the Modified Transformer attains an MSE of 19.50 versus 35.50 for the Transformer and 42.93 for GARCH, reducing error by 54.60% and 45.08%, respectively. However, since shock probabilities approach 1 once the VIX exceeds 30, a post hoc emphasis issue arises. To address this, we implement a "+10 multiple-threshold" method: if the VIX exceeds 30, we estimate probabilities for exceeding 40, then 50, and so on. This adjustment maintains predictive relevance, with TSRW accuracy improving by 19.6% (MSE 19.50 → 15.67) and

CVRW slightly declining by 3.7% (MSE 15.94 → 16.54) in table 8, effectively eliminating bias while preserving strong performance over benchmarks.

**Table 7.** Performance during the shock period

| | GARCH ($\mu \pm SD$) | Transformer | | Modified Transformer | |
|---|---|---|---|---|---|
| | | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) |
| RMSE | 6.5523 ± 0.210<br>[6.14, 6.97] | 5.9580 ± 0.190<br>[5.59, 6.33] | 5.6925 ± 0.182<br>[5.34, 6.04] | 3.9078 ± 0.135<br>[3.64, 4.18] | 3.9653 ± 0.138<br>[3.69, 4.24] |
| MSE | 42.9332 ± 2.10<br>[38.82, 47.05] | 35.4976 ± 1.80<br>[32.00, 38.99] | 32.4045 ± 1.62<br>[29.24, 35.57] | 19.4991 ± 1.05<br>[17.44, 21.56] | 15.9408 ± 0.88<br>[14.23, 17.65] |
| MAPE (%) | 11.4740 ± 0.32<br>[10.85, 12.10] | 10.8151 ± 0.30<br>[10.23, 11.41] | 11.1537 ± 0.31<br>[10.55, 11.77] | 9.4497 ± 0.27<br>[8.93, 9.97] | 9.0161 ± 0.26<br>[8.51, 9.52] |

Note: Please refer to the notes below table 4 for a detailed explanation of the legend.

**Table 8.** Performance during the +10 threshold shock period

| | GARCH ($\mu \pm SD$) | Transformer | | Modified Transformer | |
|---|---|---|---|---|---|
| | | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) |
| RMSE | 6.5523 ± 0.210<br>[6.14, 6.97] | 5.9580 ± 0.190<br>[5.59, 6.33] | 5.6925 ± 0.182<br>[5.34, 6.04] | 3.9678 ± 0.138<br>[3.70, 4.24] | 4.0753 ± 0.142<br>[3.80, 4.35] |
| MSE | 42.9332 ± 2.10<br>[38.82, 47.05] | 35.4976 ± 1.80<br>[32.00, 38.99] | 32.4045 ± 1.62<br>[29.24, 35.57] | 15.6740 ± 0.84<br>[14.03, 17.32] | 16.5369 ± 0.88<br>[14.82, 18.26] |
| MAPE (%) | 11.4740 ± 0.32<br>[10.85, 12.10] | 10.8151 ± 0.30<br>[10.23, 11.41] | 11.1537 ± 0.31<br>[10.55, 11.77] | 9.5507 ± 0.27<br>[9.02, 10.08] | 9.7261 ± 0.28<br>[9.18, 10.27] |

Note: Please refer to the notes below table 4 for a detailed explanation of the legend.

Robustness tests with alternative thresholds confirm consistent advantages. For VIX > 25 in table 9-A, the Modified Transformer reduces MSE by 59.6% versus GARCH; for VIX > 35 in table 9-B, the reduction is 70.2%. Across all thresholds, it outperforms both GARCH and Transformer by substantial margins (59.6%–70.2% vs GARCH, 51.0%–62.4% vs Transformer), demonstrating that the probabilistic attention mechanism reliably captures volatility dynamics regardless of the shock definition.

**Table 9-A.** Performance during the shock(VIX > 25) period

| | GARCH ($\mu \pm SD$) | Transformer | | Modified Transformer | |
|---|---|---|---|---|---|
| | | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) |
| RMSE | 7.8421 ± 0.238<br>[7.37, 8.31] | 7.1254 ± 0.216<br>[6.70, 7.55] | 6.8937 ± 0.209<br>[6.48, 7.31] | 4.9863 ± 0.152<br>[4.69, 5.28] | 5.1294 ± 0.156<br>[4.81, 5.45] |
| MSE | 61.4985 ± 2.45<br>[56.70, 66.30] | 50.7713 ± 2.18<br>[46.50, 55.05] | 47.5231 ± 2.06<br>[43.49, 51.56] | 24.8632 ± 1.12<br>[22.67, 27.05] | 26.3108 ± 1.18<br>[24.00, 28.63] |
| MAPE (%) | 13.2847 ± 0.34<br>[12.61, 13.96] | 12.5694 ± 0.32<br>[11.94, 13.20] | 12.1825 ± 0.31<br>[11.57, 12.80] | 10.7341 ± 0.28<br>[10.19, 11.28] | 11.0286 ± 0.29<br>[10.46, 11.60] |

Note: Please refer to the notes below table 4 for a detailed explanation of the legend.

**Table 9-B.** Performance during the shock (VIX > 35) period

| | GARCH ($\mu \pm SD$) | Transformer | | Modified Transformer | |
|---|---|---|---|---|---|
| | | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) | TSRW ($\mu \pm SD$) | CVRW ($\mu \pm SD$) |
| RMSE | 5.8927 ± 0.187<br>[5.53, 6.26] | 5.2463 ± 0.169<br>[25.00, 30.05] | 5.0182 ± 0.162<br>[4.70, 5.34] | 3.2158 ± 0.119<br>[3.00, 3.43] | 3.3247 ± 0.123<br>[3.10, 3.55] |

| | | | | | |
|---|---|---|---|---|---|
| MSE | 34.7239 ± 1.63<br>[31.53, 37.92] | 27.5237 ± 1.29<br>[32.00, 38.99] | 25.1823 ± 1.20<br>[22.83, 27.53] | 10.3414 ± 0.71<br>[9.00, 11.68] | 11.0536 ± 0.75<br>[9.59, 12.52] |
| MAPE (%) | 10.8462 ± 0.29<br>[10.28, 11.41] | 9.9847 ± 0.27<br>[9.46, 10.51] | 10.2159 ± 0.27<br>[9.68, 10.75] | 8.7392 ± 0.24<br>[8.27, 9.21] | 8.9641 ± 0.25<br>[8.47, 9.46] |

Note: Please refer to the notes below table 4 for a detailed explanation of the legend.

An analysis of the model performance across different VIX regimes indicates that during periods of stable volatility (VIX < 20), the Modified Transformer model proposed in this study does not exhibit overwhelmingly superior predictive accuracy compared to the traditional GARCH model or the standard Transformer model. The results indicate that the Modified Transformer model proposed in this study begins to demonstrate superior predictive performance during the VIX Warning regime—defined as periods when the VIX index ranges between 20 and 30—which accounts for approximately 30% of the entire sample period. Given that the Modified Transformer model does not exhibit substantial predictive advantages during the VIX Stable regime—which constitutes the majority of the overall sample period—it appears that a significant portion of the model's contribution to overall forecasting accuracy, as reported in table 6, is driven by its performance during the Warning regime. This ultimately indicates that the attention mechanism incorporating daily VIX shock probabilities performs effectively in forecasting market volatility under rapidly changing conditions.

## 3.3. Model Error Difference and Significance

The previous sections show the Modified Transformer's overall predictive superiority but lack a quantitative and statistical assessment. To address this, we compute daily prediction error differences for each model pair (Equation 16), compile them into time series, and test the mean difference using the Diebold–Mariano (DM) test. Table 10 presents the comprehensive results of these statistical comparisons across all volatility regimes. Due to differing sliding window methods, comparisons are limited to GARCH vs Transformer and GARCH vs Modified Transformer under TSRW, and Transformer vs. Modified Transformer under CVRW. The Modified Transformer delivers statistically significant gains ($p < 0.01$) across all regimes, with Cohen's d ranging from 0.20 to 1.89 when accurate forecasting is most valuable. Results remain robust under DM tests with HAC standard errors, maintaining significance across MSE, RMSE, and MAPE. Consistent findings under both TSRW and CVRW confirm that the improvements are independent of validation method, supporting their robustness and generalizability.

$$Difference_{i,t} = (Error_{Model1,t})2 - (Error_{Model2,t})2 \quad (16)$$

Note: $Error_{Model,t} = Actual_{Model,t} - Predicted_{Model,t}$

**Table 10.** Model error difference and significance

| Regime | Model1 | Model2 | Mean Difference | DM Statistic | P-Value | Cohen's d | 95% CI | line |
|---|---|---|---|---|---|---|---|---|
| Total | GARCH | TF_TSRW | 1.0463 | 2.341 | 0.0015*** | 0.426 | [0.198, 0.654] | 1 |
| | GARCH | Modfied_TF_TSRW | 4.3579 | 3.127 | 0.0021*** | 1.234 | [0.891, 1.577] | 2 |
| | TF_TSRW | Modfied_TF_TSRW | 3.3116 | 2.891 | 0.0015*** | 0.892 | [0.614, 1.170] | 3 |
| | TF_CVRW | Modified_TF_CVRW | 2.3233 | 2.756 | 0.0057*** | 0.743 | [0.521, 0.965] | 4 |
| Stable | GARCH | TF_TSRW | 0.5768 | 2.087 | 0.0038*** | 0.312 | [0.089, 0.535] | 5 |
| | GARCH | Modfied_TF_TSRW | 1.4592 | 2.134 | 0.0020*** | 0.721 | [0.445, 0.997] | 6 |
| | TF_TSRW | Modfied_TF_TSRW | 0.8825 | 1.987 | 0.0011*** | 0.598 | [0.334, 0.862] | 7 |
| | TF_CVRW | Modified_TF_CVRW | 0.5354 | 1.823 | 0.0033*** | 0.467 | [0.211, 0.723] | 8 |
| Warning | GARCH | TF_TSRW | 0.3127 | 1.785 | 0.0048*** | 0.198 | [0.024, 0.372] | 9 |
| | GARCH | Modfied_TF_TSRW | 3.9373 | 2.967 | 0.0024*** | 1.156 | [0.823, 1.489] | 10 |
| | TF_TSRW | Modfied_TF_TSRW | 3.6246 | 2.742 | 0.0038*** | 1.089 | [0.761, 1.417] | 11 |
| | TF_CVRW | Modified_TF_CVRW | 2.3668 | 2.145 | 0.0087*** | 0.823 | [0.512, 1.134] | 12 |
| Shock (+10 threshold) | GARCH | TF_TSRW | 7.4356 | 2.876 | 0.0031*** | 0.612 | [0.298, 0.926] | 13 |
| | GARCH | Modfied_TF_TSRW | 27.2592 | 4.231 | 0.0017*** | 1.891 | [1.512, 2.270] | 14 |

 ISSN 2723-6471 

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TF_TSRW | Modfied_TF_TSRW | 19.8236 | 3.867 | 0.0023*** | 1.567 | [1.201, 1.933] | 15 |
| TF_CVRW | Modified_TF_CVRW | 15.8676 | 3.421 | 0.0035*** | 1.324 | [0.978, 1.670] | 16 |

Note: *** $p < 0.01$, ** $p < 0.05$. Mean differences represent MSE improvements (positive values indicate Modified Transformer superiority). Cohen's d effect sizes: small (≈0.2), medium (≈0.5), large (≈0.8).

## 4. Conclusion

This study proposes a novel methodology for forecasting the VIX index by integrating stochastic diffusion equations with Transformer-based deep learning architecture. Our Probabilistic-Attention Transformer incorporates daily volatility shock probabilities−estimated via the Heston model—directly into the attention mechanism, enabling enhanced detection of abrupt market volatility changes.

Analysis of 35 years of VIX data reveals distinct performance patterns across market regimes. During stable conditions (VIX < 20), our model achieves modest improvements of 34.32% over GARCH models. However, performance gains become substantial during critical market phases: 42.10% improvement over standard Transformers during warning periods ($20 \leq VIX \leq 30$) and 54.60% improvement during volatility shocks (VIX > 30). All improvements are statistically significant with large effect sizes (Cohen's d = 0.20-1.89), confirming both statistical and practical significance. The "+10 multiple-threshold" approach successfully mitigates post-hoc bias concerns while maintaining predictive superiority. Our rigorous temporal cross-validation approach, where each model parameter is estimated using only historically available data, provides a methodological template for future research combining multiple modeling frameworks in financial time series forecasting.

Our research makes three primary contributions to financial volatility forecasting. First, we introduce the first finance-specific modification of Transformer attention mechanisms, creating a principled framework for integrating external probabilistic indicators with deep learning architectures. Second, we provide comprehensive operationalization of "early warning signals" through multi-dimensional indicators including stochastic process patterns and temporal clustering behaviors. Third, we establish rigorous statistical validation of dual evaluation methodologies, demonstrating that TSRW and CVRW approaches provide complementary assessment perspectives.

The enhanced predictive accuracy during market stress periods offers significant value for risk management applications, particularly for financial institutions requiring accurate volatility forecasting when market stability is most threatened. While this study focuses exclusively on VIX forecasting, the generalizability of our probabilistic attention mechanism to other financial indicators requires further investigation. Future research should examine applicability to stock indices, exchange rates, and commodity volatility measures, and extend the framework to multivariate settings incorporating cross-asset correlations and macroeconomic indicators.

By successfully integrating probabilistic market indicators with attention-based deep learning, this research demonstrates that domain-specific modifications to general-purpose architectures can yield substantial improvements in financial forecasting accuracy. The methodology provides a robust foundation for enhanced volatility prediction during periods of market instability, when accurate forecasting delivers the greatest economic value.

## 5. Conclusion

This study proposes a novel methodology for forecasting the VIX index by integrating stochastic diffusion equations with Transformer-based deep learning architecture. Our Probabilistic-Attention Transformer incorporates daily volatility shock probabilities—estimated via the Heston model—directly into the attention mechanism, enabling enhanced detection of abrupt market volatility changes.

Analysis of 35 years of VIX data reveals distinct performance patterns across market regimes. During stable conditions (VIX < 20), our model achieves modest improvements of 34.32% over GARCH models. However, performance gains become substantial during critical market phases: 42.10% improvement over standard Transformers during warning periods ($20 \leq VIX \leq 30$) and 54.60% improvement during volatility shocks (VIX > 30). All improvements are statistically significant with large effect sizes (Cohen's d = 0.20-1.89), confirming both statistical and practical significance. The "+10 multiple-threshold" approach successfully mitigates post-hoc bias concerns while maintaining

predictive superiority. Our rigorous temporal cross-validation approach, where each model parameter is estimated using only historically available data, provides a methodological template for future research combining multiple modeling frameworks in financial time series forecasting.

Our research makes three primary contributions to financial volatility forecasting. First, we introduce the first finance-specific modification of Transformer attention mechanisms, creating a principled framework for integrating external probabilistic indicators with deep learning architectures. Second, we provide comprehensive operationalization of "early warning signals" through multi-dimensional indicators including stochastic process patterns and temporal clustering behaviors. Third, we establish rigorous statistical validation of dual evaluation methodologies, demonstrating that TSRW and CVRW approaches provide complementary assessment perspectives.

The enhanced predictive accuracy during market stress periods offers significant value for risk management applications, particularly for financial institutions requiring accurate volatility forecasting when market stability is most threatened.

While this study focuses exclusively on VIX forecasting, the generalizability of our probabilistic attention mechanism to other financial indicators requires further investigation. Future research should examine applicability to stock indices, exchange rates, and commodity volatility measures, and extend the framework to multivariate settings incorporating cross-asset correlations and macroeconomic indicators.

By successfully integrating probabilistic market indicators with attention-based deep learning, this research demonstrates that domain-specific modifications to general-purpose architectures can yield substantial improvements in financial forecasting accuracy. The methodology provides a robust foundation for enhanced volatility prediction during periods of market instability, when accurate forecasting delivers the greatest economic value.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization: J.S.K. and Z.L.; Methodology: J.S.K.; Software: J.S.K.; Validation: J.S.K. and Z.L.; Formal Analysis: J.S.K. and Z.L.; Investigation: J.S.K.; Resources: Z.L.; Data Curation: Z.L.; Writing Original Draft Preparation: J.S.K. and Z.L.; Writing Review and Editing: J.S.K. and Z.L.; Visualization: J.S.K. All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 6.4. Institutional Review Board Statement

Not applicable.

### 6.5. Informed Consent Statement

Not applicable.

### 6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] J. Fleming, B. Ostdiek, and R. E. Whaley, "Predicting stock market volatility: A new measure," *J. Futures Markets*, vol. 15, no. 3, pp. 265-290, 1995.

[2] B. J. Christensen and N. R. Prabhala, "The relation between implied and realized volatility," *J. Financial Economics*, vol. 50, no. 2, pp. 125-150, 1998.

[3] R. E. Whaley, "The investor fear gauge," *J. Portfolio Management*, vol. 26, no. 3, pp. 12-27, 2000.

[4] B. J. Blair, S. H. Poon, and S. J. Taylor, "Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns," *J. Econometrics*, vol. 105, no. 1, pp. 5-26, 2001.

[5] R. Becker, A. E. Clements, and A. McClelland, "The jump component of S&P 500 volatility and the VIX index," *J. Banking & Finance*, vol. 33, no. 6, pp. 1033-1038, 2009.

[6] M. M. Copeland and T. E. Copeland, "Market timing: Style and size rotation using the VIX," *Financial Analysts J.*, vol. 55, no. 2, pp. 73-81, 1999.

[7] P. Giot, "Relationships between implied volatility indices and stock index returns," *J. Portfolio Management*, vol. 31, no. 3, pp. 92-100, 2005.

[8] P. S. Banerjee, J. S. Doran, and D. R. Peterson, "Implied volatility and future portfolio returns," *J. Banking & Finance*, vol. 31, no. 10, pp. 3183-3199, 2007.

[9] R. E. Whaley, "Understanding the VIX," *J. Portfolio Management*, vol. 35, no. 3, pp. 98-105, 2009.

[10] G. Bekaert and M. Hoerova, "The VIX, the variance premium and stock market volatility," *J. Econometrics*, vol. 183, no. 2, pp. 181-192, 2014.

[11] P. R. Hansen, Z. Huang, and H. H. Shek, "Realized GARCH: a joint model for returns and realized measures of volatility," *J. Applied Econometrics*, vol. 27, no. 6, pp. 877-906, 2012.

[12] J. Hao and J. E. Zhang, "GARCH option pricing models, the CBOE VIX, and variance risk premium," *J. Financial Econometrics*, vol. 11, no. 3, pp. 556-580, 2013.

[13] P. Christoffersen, B. Feunou, and Y. Jeon, "Option valuation with observable volatility and jump dynamics," *J. Banking & Finance,* vol. 61, no. 1, pp. S101-S120, 2015.

[14] Q. Liu, S. Guo, and G. Qiao, "VIX forecasting and variance risk premium: A new GARCH approach," *North American J. Economics and Finance,* vol. 34, no. 1, pp. 314-322, 2015.

[15] X. Yang and P. Wang, "VIX futures pricing with conditional skewness," *J. Futures Markets*, vol. 38, no. 9, pp. 1126-1151, 2018.

[16] L.S.Chong, K.M.Lim, and C. P. Lee, "Stock market prediction using ensemble of Deep Neural Networks," in *Proc. 2020 IEEE 2nd Int. Conf. Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia,* vol. 2020, no. 1, pp. 1-5, 2020.

[17] S. Guo and Q. Liu, "Efficient out-of-sample pricing of VIX futures," *J. Derivatives,* vol. 27, no. 3, pp. 126-139, 2020.

[18] G. Qiao, J. Yang, and W. Li, "VIX forecasting based on GARCH-type model with observable dynamic jumps: A new perspective," *North American J. Economics and Finance*, vol. 53, article 101186, no. 1, pp. 1-12, 2020.

[19] H. Mayatopani, "Implementation of ANN and GARCH for stock price forecasting," *J. Applied Data Sciences,* vol. 2, no. 4, pp. 109-134, 2021.

[20] S. Sarmini, C. R. A. Widiawati, D. R. Febrianti, and D. Yuliana, "Volatility analysis of cryptocurrencies using statistical approach and GARCH model a case study on daily percentage change," *J. Applied Data Sciences*, vol. 5, no. 3, pp. 838-848, 2024.

[21] Y. Sugiarti, A. I. Suroso, I. Hermadi, E. Sunarti, and F. B. M. Yamin, "Price prediction of Aglaonema ornamental plants using the Long Short-Term Memory (LSTM) algorithm," *J. Applied Data Sciences*, vol. 6, no. 2, pp. 1426-1436, 2025.

[22] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European J. Operational Research*, vol. 270, no. 2, pp. 654-669, 2018.

[23] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stock prediction using numerical and textual information," in *Proc. 2016 IEEE/ACIS 15th Int. Conf. Computer and Information Science (ICIS), Okayama, Japan*, vol. 2016, no. 1, pp. 1-6, 2016.

[24] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, article e0194889, pp. 1-12, 2018.

[25] H. Kim and C. Won, "Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models," Expert Systems with Applications, vol. 103, no. 1, pp. 25-37, 2018.

[26] A. Bucci, "Realized volatility forecasting with neural networks," *J. Financial Econometrics,* vol. 18, no. 3, pp. 502-531, 2020.

[27] O. Bustos and A. Pomares-Quimbaya, "Stock market movement forecast: A systematic review," *Expert Systems with Applications,* vol. 156, article 113464, no. 1, pp. 1-12, 2020.

[28] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Review of Financial Studies*, vol. 33, no. 5, pp. 2223-2273, 2020.

[29] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005-2019," *Applied Soft Computing*, vol. 90, article 106181, no. 1, pp. 1-12, 2020.

[30] M. Qiu, Y. Song, and F. Akagi, "Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market," *Chaos, Solitons & Fractals*, vol. 85, no. 1, pp. 1-7, 2016.

[31] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *Int. J. Forecasting*, vol. 37, no. 1, pp. 388-427, 2021.

[32] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," *Expert Systems with Applications*, vol. 184, article 115537, no. 1, pp. 1-12, 2021.

[33] A. Petrozziello, L. Troiano, A. Serra, I. Jordanov, G. Storti, R. Tagliaferri, and M. La Rocca, "Deep learning for volatility forecasting in asset management," *Soft Computing*, vol. 26, no. 17, pp. 8553–8574, 2022.

[34] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Trans. Royal Society A*, vol. 379, no. 2194, article 20200209, pp. 1-12, 2021.

[35] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, no. 1, pp. 22419-22430, 2021.

[36] G. Zerveas, S. Jayaraman, D. Patel, C. Eickhoff, and K. Sechidis, "A Transformer-based Framework for Multivariate Time Series Representation Learning," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, vol. 139, no. 1, pp. 12472-12482, 2021.

[37] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artificial Intelligence,* vol. 35, no. 12, pp. 11106-11115, 2021.

[38] Q. Wen, "Transformers in time series: A survey," *arXiv preprint arXiv*:2202.07125, vol. 2022, no. 1, pp. 1-12, 2022.

[39] Y. Wang, "Deep time series models: A comprehensive survey and benchmark," a*rXiv preprint arXiv*:2407.13278, vol. 2024, no. 1, pp. 1-12, 2024.

[40] X. Zhang, Y. Li, S. Wang, B. Fang, and P. S. Yu, "Enhancing stock market prediction with extended coupled hidden Markov model over multi-sourced data," *Knowledge and Information Systems*, vol. 61, no. 1, pp. 1071-1090, 2019.

[41] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv*:1803.02155, vol. 2018, no. 1, pp. 1-12, 2018.

[42] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv*:2006.04768, vol. 2020, no. 1, pp. 1-12, 2020.

[43] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987-1007, 1982.

[44] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307-327, 1986.

[45] V. Akgiray, "Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts," *J. Business,* vol. 62, no. 1, pp. 55-80, 1989.

[46] G. G. Booth, J. Hatem, I. Virtanen, and P. Yli-Olli, "Stochastic modeling of security returns: Evidence from the Helsinki Stock Exchange," *European J. Operational Research*, vol. 56, no. 1, pp. 98-106, 1992.

[47] A. Corhay and A. T. Rad, "Statistical properties of daily returns: Evidence from European stock markets," *J. Business Finance & Accounting,* vol. 21, no. 2, pp. 271-282, 1994.

[48] D. B. Nelson, "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica*, vol. 59, no. 2, pp. 347-370, 1991.

[49] L. R. Glosten, R. Jagannathan, and D. E. Runkle, "On the relation between the expected value and the volatility of the nominal excess return on stocks," *J. Finance*, vol. 48, no. 5, pp. 1779-1801, 1993.

[50] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *Proc. 38th Int. Conf. Machine Learning (ICML),* vol. 2021, no. 1, pp. 8857-8868, 2021.

[51] M. Jin, "Time-llm: Time series forecasting by reprogramming large language models," *arXiv preprint arXiv*:2310.01728, vol. 2023, no. 1, pp. 1-12, 2023.

[52] W. Liao, "TimeGPT in load forecasting: A large time series model perspective," *Applied Energy*, vol. 379, article 124973, no. 1, pp. 1-12, 2025.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, no .1, pp. 1-16, 2017.

[54] S. L. Heston, "A closed-form solution for options with stochastic volatility with applications to bond and currency options," *The Review of Financial Studies,* vol. 6, no. 2, pp. 327–343, 1993.

[55] P. Carr and D. B. Madan, "Option valuation using the fast Fourier transform," *Journal of Computational Finance*, vol. 2, no. 4, pp. 61–73, 1999.

[56] T. Bollerslev and H. Zhou, "Estimating stochastic volatility diffusion using conditional moments of integrated volatility," *Journal of Econometrics*, vol. 109, no. 1, pp. 33–65, 2002.

[57] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *J. Political Economy,* vol. 81, no. 3, pp. 637-654, 1973.