XFinInsight-VolRisk: A Human-in-the-Loop Explainable Artificial Intelligence Framework for Multi-Asset Portfolio Volatility Forecasting

Prof. C.K.Jha,
Professor and Dean,
Department of Computer Science,
Banasthali Vidyapith,
Rajasthan, India
ckjha@banasthali.in

Preetha V K,
Research Scholar,
Department of Computer Science,
Banasthali Vidyapith,
Rajasthan, India
preethadamodaran@gmail.com

Abstract
Accurate forecasting of portfolio-level volatility is central to risk budgeting, margin setting, and regulatory capital calculation, yet the artificial intelligence (AI) models that increasingly support this task remain difficult for risk managers to interrogate and trust. Building on the XFinInsight paradigm for interpretable financial artificial intelligence, this paper introduces XFinInsight-VolRisk, a Human-in-the-Loop Explainable Artificial Intelligence framework for forecasting the 5-day-ahead annualized realized volatility of a multi-asset equity portfolio. Using twelve years (2015-2024) of real daily Open-High-Low-Close-Volume data for a five-asset growth-equity portfolio sourced from Yahoo Finance, the framework constructs a feature set spanning multi-horizon realized volatility, cross-asset return correlation, technical risk indicators, return-distribution moments, and a rolling GARCH(1,1) conditional-volatility forecast. Five forecasting approaches, an interpretable Linear Regression baseline, Random Forest, XGBoost, a compact from-scratch Long Short-Term Memory (LSTM) network, and the classical GARCH(1,1) model, are evaluated under a five-fold walk-forward (expanding-window) cross-validation protocol to avoid look-ahead bias. All learned models substantially outperform the naive persistence and GARCH(1,1) baselines, reducing walk-forward root-mean-square error (RMSE) by 7.6-18.3% relative to persistence; the interpretable Linear Regression baseline, structurally related to the Heterogeneous Autoregressive Realized Volatility (HAR-RV) model, achieves the lowest mean RMSE (0.142) and highest directional accuracy (72.4%) of all learned models. SHapley Additive exPlanations (SHAP) applied to the XGBoost model identify rolling cross-asset correlation and average true range as the dominant drivers of portfolio-risk forecasts, complementing the Linear Regression model's reliance on multi-horizon realized-volatility persistence. A rule-based Human-in-the-Loop validation layer flags 55.2% of test-period forecasts for review on the basis of model disagreement or forecast-magnitude discontinuity, reducing RMSE on the flagged subset by 19.9%. A component-wise ablation study shows that technical risk indicators contribute the largest single improvement to forecast accuracy, while the isolated contribution of cross-asset correlation features is more limited than SHAP attribution alone would suggest. These findings demonstrate that predictive accuracy, model interpretability, and analyst oversight can be jointly embedded in a portfolio-risk forecasting pipeline without recourse to synthetic data or post-hoc justification.

Keywords: Explainable Artificial Intelligence, Portfolio Volatility Forecasting, Human-in-the-Loop Learning, SHAP, GARCH, Walk-Forward Validation, Financial Risk Management.

Doi:10.48047/ijiee.2025.15.6.29

1 Introduction

Portfolio volatility is the primary quantitative input to risk budgeting, value-at-risk (VaR) estimation, margin requirements, and the capital-adequacy calculations that regulators impose on banks, asset managers, and insurers [1], [2]. Because volatility clusters and shifts regime in response to macroeconomic shocks, sector rotations, and changes in the correlation structure across holdings, forecasting it accurately is a persistently difficult problem, and errors propagate directly into mispriced risk limits and capital buffers [3]. Classical econometric models, most notably the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) family introduced by Bollerslev [4] and the Heterogeneous Autoregressive Realized Volatility (HAR-RV) model of Corsi [5], remain the industry standard because their parametric structure is transparent and auditable, even though their single-asset, single-horizon design does not naturally accommodate the cross-asset correlation dynamics that drive risk at the portfolio level.

The growth of machine learning (ML) in quantitative finance has produced volatility and risk models with substantially higher predictive capacity, drawing on ensemble tree methods and deep sequence models that can absorb heterogeneous technical, fundamental, and cross-asset features simultaneously [6], [7]. However, this predictive gain has come at the cost of transparency: ensemble and deep-learning volatility models are largely black boxes, and risk managers, auditors, and regulators increasingly require that any AI system informing capital decisions be explainable, auditable, and subject to expert override [8], [9]. This tension between predictive accuracy and interpretability is especially acute for portfolio-level risk, where a forecast error is not merely a prediction miss but a direct input to a regulatory capital figure.

Explainable Artificial Intelligence (XAI) techniques, particularly SHapley Additive exPlanations (SHAP) [10], [11] and Local Interpretable Model-agnostic Explanations (LIME) [12], have been applied to post-hoc interpretation of financial ML models, revealing which features drive a given prediction without altering the underlying model. While valuable, purely post-hoc explanation does not, by itself, give a human expert a formal mechanism to intervene in the model's output when the explanation reveals an implausible or ungrounded forecast. The XFinInsight research program addresses this gap by embedding interpretability throughout the modeling pipeline and coupling it with a Human-in-the-Loop (HIL) validation layer through which domain experts can review, question, and correct model outputs before they are used downstream [12].

A companion study in this research program, XFinInsight-HI, applied this dual-track philosophy, interpretable-baseline-plus-complex-model, SHAP explainability, and rule-based HIL validation, to single-stock price prediction using Random Forest, XGBoost, and LSTM models trained on technical indicators. That study is deliberately restricted to single-asset price-level forecasting; it does not address the correlation structure, diversification dynamics, or realized-volatility persistence that determine risk at the level of a portfolio rather than a single position. The present study extends the XFinInsight framework to a structurally distinct and practically important problem, multi-asset portfolio volatility forecasting, where the target is not the price of one instrument but the forward realized variability of a weighted combination of instruments, and where cross-asset correlation is itself a first-class predictive feature rather than an absent dimension.

This paper introduces XFinInsight-VolRisk, a Human-in-the-Loop Explainable Artificial Intelligence framework for forecasting 5-day-ahead annualized realized volatility of a multi-asset equity

Doi:10.48047/ijiee.2025.15.6.29

portfolio. The framework acquires real daily OHLCV data for a five-asset growth-equity portfolio (Amazon, Alphabet, Meta Platforms, Tesla, and Salesforce) from Yahoo Finance, engineers a feature set spanning multi-horizon realized volatility, cross-asset correlation, technical risk indicators, return-distribution moments, and a rolling GARCH(1,1) conditional-volatility forecast, and evaluates five forecasting approaches under a walk-forward cross-validation protocol designed to avoid the look-ahead bias that a single chronological train/test split can introduce in a non-stationary volatility series. SHAP explainability and Linear Regression coefficients are compared as dual interpretability tracks, a rule-based HIL layer flags and corrects high-disagreement forecasts, and a component-wise ablation study quantifies the marginal contribution of each feature group to predictive accuracy.

The contributions of this paper are fourfold. First, it extends the XFinInsight interpretable-AI paradigm from single-asset price prediction to multi-asset portfolio volatility forecasting, a distinct problem class with direct relevance to risk management and regulatory compliance. Second, it evaluates five forecasting approaches, GARCH(1,1), Linear Regression, Random Forest, XGBoost, and a from-scratch LSTM, under a rigorous walk-forward protocol on twelve years of real market data, and reports both favorable and unfavorable findings honestly, including cases in which the interpretable baseline outperforms the more complex ensembles. Third, it applies SHAP and standardized linear coefficients as complementary interpretability tracks and shows that they can disagree about feature importance in informative ways. Fourth, it introduces a rule-based Human-in-the-Loop validation layer and a systematic ablation study that together quantify not only whether the framework works, but which of its components are responsible for the observed gains.

2 Related Work

Volatility forecasting has a long econometric lineage. The ARCH model of Engle [13] and its generalization, GARCH [4], remain widely used because their conditional-variance recursion is transparent and their parameters are directly interpretable as persistence and shock-sensitivity coefficients. Corsi's HAR-RV model [5] extended this tradition by regressing realized volatility on its own lagged values at daily, weekly, and monthly horizons, a structure that this paper's multi-horizon realized-volatility feature set (5-, 10-, 21-, and 63-day windows) deliberately mirrors, and that motivates treating the paper's Linear Regression model as a HAR-RV-consistent interpretable baseline rather than an arbitrary linear fit.

Machine learning approaches to volatility forecasting have generally reported accuracy gains over GARCH-family baselines by incorporating a wider information set. Ensemble tree methods such as Random Forest and XGBoost have been applied to realized-volatility and value-at-risk forecasting with reported improvements in out-of-sample error over GARCH benchmarks [6], and LSTM-based sequence models have been used to capture volatility clustering and regime persistence directly from return sequences [7], [14]. A recurring limitation across this literature, noted explicitly by Rudin [15], is that the reported accuracy gains are rarely accompanied by a systematic account of which engineered features or model components are actually responsible for the improvement, leaving practitioners unable to distinguish a robust signal from an artifact of a particular train/test split.

XAI methods have increasingly been applied within finance to address this opacity. SHAP, grounded in cooperative game theory [10], and LIME [11] have been used to explain credit-risk, fraud-detection, and price-prediction models [16], [17], and comparative reviews report that SHAP is the most widely adopted attribution method for financial ML, though the literature notes

Doi:10.48047/ijiee.2025.15.6.29

it is typically applied post-hoc, after model selection, rather than embedded within the modeling pipeline itself [18]. Human-in-the-Loop machine learning, surveyed by Mosqueira-Rey et al. [19], formalizes the broader principle that domain-expert review can be incorporated as an explicit stage of a model's inference pipeline rather than treated as an informal, outside-the-model process; this paper's HIL validation layer operationalizes that principle for portfolio-risk forecasting specifically.

Relatively little prior work combines these three strands, econometric volatility structure, ensemble/deep-learning prediction, and embedded explainability with expert review, for the multi-asset portfolio setting specifically. Studies that explain single-asset financial predictions with SHAP [16], [17] do not generally incorporate cross-asset correlation as a first-class feature, and portfolio-level risk studies that do model correlation structure [1], [2] have generally not paired their forecasts with feature-level explainability or a formal validation-and-correction mechanism. This gap, a unified pipeline for multi-asset portfolio volatility that is simultaneously accurate, explainable at the feature level, and subject to a documented human-review mechanism, is the specific research gap addressed by XFinInsight-VolRisk

Table 1 Summary of the Research Gap

Focus Area in Existing Studies	Limitation Identified	Research Gap	How XFinInsight-VolRisk Addresses It
Econometric volatility models (GARCH, HAR-RV)	Transparent but single-asset; no cross-asset correlation input	No portfolio-level, multi-asset formulation	Portfolio return series with explicit cross-asset correlation features
ML/DL volatility and price forecasting	Higher accuracy but post-hoc or absent explainability	Feature contribution to gains rarely isolated	Component-wise ablation study quantifying each feature group
Post-hoc XAI for financial ML (SHAP, LIME)	Explains predictions after the fact; not embedded in modeling pipeline	No mechanism for expert correction	Rule-based HIL validation layer with documented flag-and-correct rule
Human-in-the-loop ML (general)	Formalized conceptually; rarely instantiated for portfolio risk	Absence of a concrete, reproducible portfolio-risk instantiation	Reproducible, rule-based HIL layer evaluated on real market data

3. Methodology

This study proposed XFinInsight-VolRisk, an Explainable Artificial Intelligence model in the Human-in-the-Loop format that predicts the five-day ahead annualized realized volatility of a multi-assets equity portfolio. The framework includes data gathering and harmonization, portfolio construction, feature engineering, predictive modeling, walk-forward validation, explainability evaluation, Human-in-the-Loop forecast evaluation, and performance evaluation. We utilize five large-cap growth stocks – Amazon, Alphabet, Meta Platforms, Tesla, and Salesforce – and examine linear regression, random forest, XGBoost, LSTM, and GARCH(1,1) forecasts.

Doi:10.48047/ijiee.2025.15.6.29

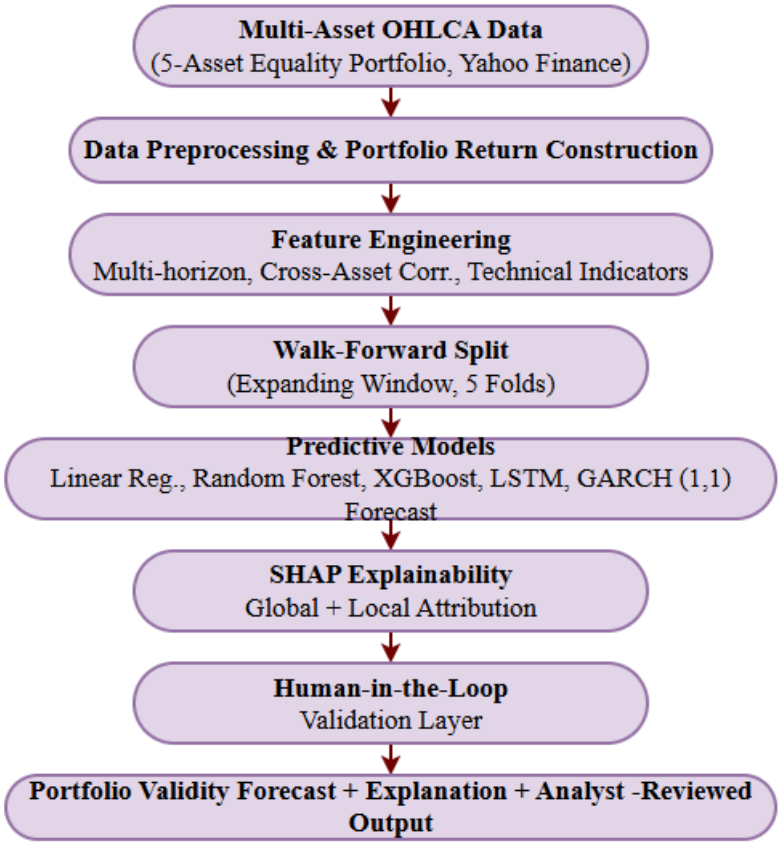

Figure 1 Flowchart of the XFinInsight-VolRisk Framework

2.1 Data Collection and Preprocessing

Open, High, Low, Close, Adjusted Close, and Volume data were gathered for the following firms: Amazon.com Inc. (AMZN), Alphabet Inc. (GOOGL), Meta Platforms Inc. (META), Tesla Inc. (TSLA), and Salesforce Inc. (CRM). The initial dataset was from the common trading period ranging from May 2012 until November 2024. But the final dataset after generating lagged variables, rolling GARCH models, and forward volatility forecasts consisted of 2,393 days from 22 May 2015 until 21 November 2024.

The adjusted closing prices of the assets were aligned using the intersection of their trading dates. Missing values arising from non-common trading days were excluded to ensure that all portfolio constituents had valid observations on each date. The continuously compounded return of asset i on day t was calculated as Eq. (1)

$$r_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right) \quad (1)$$

where $P_{i,t}$and $P_{i,t-1}$represent the adjusted closing prices of asset $i$on days $t$and $t-1$, respectively.

2.2 Portfolio Construction

An equal-weighted portfolio was constructed to isolate the volatility-forecasting problem from portfolio-weight optimisation. The daily portfolio return was computed as

Doi:10.48047/ijiee.2025.15.6.29

$$r_{p,t} = \sum_{i=1}^{N} w_i \, r_{i,t}, \tag{2}$$

subject to

$$\sum_{i=1}^{N} w_i = 1, \; w_i = \frac{1}{N} \tag{3}$$

For the five-asset portfolio, $N = 5$, and therefore $w_i = 0.20$for each constituent. Transaction costs, rebalancing costs, and portfolio-weight optimisation were not incorporated because the primary objective was to evaluate volatility-forecasting performance rather than trading profitability.

A portfolio net asset value series was generated for calculating price-based technical indicators:

$$NAV_t = NAV_{t-1} \exp\left(r_{p,t}\right) \tag{4}$$

where NAV_0denotes the initial portfolio value.

2.3 Target Variable Construction

The forecasting target was the annualized realized volatility of the portfolio over the following five trading days. For a forecast generated on day t, the target variable was calculated as

$$y_t = \sqrt{252 \left[\frac{1}{H-1} \sum_{k=1}^{H} \left(r_{p,t+k} - \bar{r}_{p,t}^{(H)} \right)^2 \right]} \tag{5}$$

where $H = 5$represents the forecast horizon and

$$\bar{r}_{p,t}^{(H)} = \frac{1}{H} \sum_{k=1}^{H} r_{p,t+k} \tag{6}$$

The factor 252 annualizes the daily volatility under the conventional assumption of approximately 252 trading days per year. Only information available up to day $t$was used to construct the explanatory variables, while returns from $t + 1$ to $t + 5$ were used exclusively for constructing the prediction target.

2.4 Feature Engineering

Five groups of explanatory variables were constructed: multi-horizon realized-volatility features, cross-asset correlation features, technical risk indicators, return-distribution moments, and a GARCH-informed volatility feature.

Doi:10.48047/ijiee.2025.15.6.29

2.4.1 Multi-Horizon Realized Volatility

Trailing portfolio realized volatility was calculated over 5-, 10-, 21-, and 63-day windows. For a window of h days, annualized realized volatility was defined as

$$RV_{p,t}^{(h)} = \sqrt{252\left[\frac{1}{h-1}\sum_{j=0}^{h-1}\left(r_{p,t-j} - \bar{r}_{p,t}^{(h)}\right)^2\right]} \tag{7}$$

Where

$$\bar{r}_{p,t}^{(h)} = \frac{1}{h}\sum_{j=0}^{h-1} r_{p,t-j} \tag{8}$$

These variables capture short-, medium-, and long-horizon volatility persistence and follow the general logic of the Heterogeneous Autoregressive Realized Volatility model.

2.4.2 Cross-Asset Correlation Features

The pairwise correlation between assets $i$and $j$over a 63-day rolling window was computed as

$$\rho_{ij,t}^{(63)} = \frac{\sum_{k=0}^{62}(r_{i,t-k}-\bar{r}_{i,t})(r_{j,t-k}-\bar{r}_{j,t})}{\sqrt{\sum_{k=0}^{62}(r_{i,t-k}-\bar{r}_{i,t})^2}\sqrt{\sum_{k=0}^{62}(r_{j,t-k}-\bar{r}_{j,t})^2}} \tag{9}$$

The average pairwise correlation was then calculated as

$$\bar{\rho}_t^{(63)} = \frac{2}{N(N-1)}\sum_{i=1}^{N-1} \quad \sum_{j=i+1}^{N}\rho_{ij,t}^{(63)} \tag{10}$$

For five assets, this measure averages the correlations of the ten unique asset pairs. A higher value indicates stronger co-movement and a reduction in the portfolio's diversification benefit.

The average asset-level realized volatility was calculated as

$$\bar{RV}_t^{(21)} = \frac{1}{N}\sum_{i=1}^{N} RV_{i,t}^{(21)} \tag{11}$$

while the cross-sectional dispersion of asset volatility was computed as

$$D_{RV,t}^{(21)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(RV_{i,t}^{(21)} - \bar{RV}_t^{(21)}\right)^2} \tag{12}$$

2.4.3 Technical Risk Indicators

The technical feature group consisted of the 14-day Relative Strength Index, 20-day Bollinger Band width, normalized Average True Range, and a 63-day trading-volume z-score.

Doi:10.48047/ijiee.2025.15.6.29

The Relative Strength Index was calculated as

$$RSI_t = 100 - \frac{100}{1+RS_t},\ RS_t = \frac{\text{Average Gain}_{14,t}}{\text{Average Loss}_{14,t}} \tag{13}$$

The Bollinger Band width was computed from the portfolio net asset value as

$$BBW_t^{(20)} = \frac{UB_t - LB_t}{MA_t^{(20)}} \tag{14}$$

Where, $UB_t = MA_t^{(20)} + 2SD_t^{(20)}$, and $LB_t = MA_t^{(20)} - 2SD_t^{(20)}$.

For asset i, the true range was defined as

$$TR_{i,t} = \max\left[H_{i,t}, -, L_{i,t}|H_{i,t} - C_{i,t-1}||L_{i,t} - C_{i,t-1}|\right] \tag{15}$$

The 14-day normalized Average True Range was

$$ATR\%_{i,t}^{(14)} = \frac{\frac{1}{14}\sum_{k=0}^{13} TR_{i,t-k}}{C_{i,t}} \tag{16}$$

The portfolio-level ATR feature was obtained by averaging $ATR\%_{i,t}^{(14)}$across the five assets.

The volume z-score for asset $i$was computed as

$$Z_{i,t}^{Vol} = \frac{Vol_{i,t} - \mu_{Vol,i,t}^{(63)}}{\sigma_{Vol,i,t}^{(63)}} \tag{17}$$

and the portfolio-level volume anomaly indicator was obtained by averaging the five asset-level z-scores.

2.4.4 Return-Distribution Features

Rolling skewness and excess kurtosis were calculated over the preceding 21 trading days. Skewness was defined as

$$Skew_t^{(21)} = \frac{1}{21}\sum_{k=0}^{20}\left(\frac{r_{p,t-k} - \bar{r}_{p,t}^{(21)}}{s_{p,t}^{(21)}}\right)^3 \tag{18}$$

while excess kurtosis was calculated as

$$Kurt_t^{(21)} = \frac{1}{21}\sum_{k=0}^{20}\left(\frac{r_{p,t-k} - \bar{r}_{p,t}^{(21)}}{s_{p,t}^{(21)}}\right)^4 - 3 \tag{19}$$

Doi:10.48047/ijiee.2025.15.6.29

These variables capture asymmetry and heavy-tailed behaviour that cannot be represented by variance alone.

2.4.5 GARCH-Informed Feature

A GARCH(1,1) model was estimated using a trailing window of 756 trading days and re-estimated every 63 trading days. The portfolio return was represented as

$$r_{p,t} = \mu + \varepsilon_t,\ \varepsilon_t = \sigma_t z_t \tag{20}$$

where $z_t$is a standardized innovation. The conditional variance followed

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{21}$$

subject to

$$\omega > 0,\ \alpha \geq 0,\ \beta \geq 0,\ \alpha + \beta < 1 \tag{22}$$

The one-step-ahead GARCH volatility forecast was computed as

$$\hat{\sigma}_{t+1} = \sqrt{\hat{\omega} + \hat{\alpha}\varepsilon_t^2 + \hat{\beta}\sigma_t^2} \tag{23}$$

This forecast was evaluated as a standalone econometric prediction and was also supplied as an engineered input to the machine-learning models.

Five feature groups were engineered from the aligned OHLCV panel, summarized in Table 2. Rows with insufficient trailing or leading history for any feature or target were dropped, leaving 2,393 daily observations spanning 22 May 2015 to 21 November 2024 for model training and evaluation.

Table 2 Engineered Feature Groups

Feature Group	Features	Window(s)	Rationale
Multi-horizon realized volatility	port_rvol_5d, 10d, 21d, 63d	5/10/21/63 days	HAR-RV-consistent volatility persistence at multiple horizons
Cross-asset correlation	avg_pairwise_corr_63d, mean_asset_rvol_21d, dispersion_asset_rvol_21d	63/21 days	Captures diversification collapse and systemic co-movement
Technical risk indicators	port_rsi_14, port_boll_width_20, mean_atr_pct_14, mean_volume_z_63d	14/20/63 days	Momentum, band-width, true-range, and volume-anomaly signals
Return-distribution moments	port_ret_skew_21d, port_ret_kurt_21d	21 days	Tail-risk and asymmetry not captured by variance alone
Econometric baseline feature	garch_vol_forecast	756-day rolling fit, refit every 63 days	One-step-ahead GARCH (1,1) conditional-volatility forecast

Doi:10.48047/ijiee.2025.15.6.29

Algorithm 1: Multi-Asset Portfolio Feature Engineering

Input: aligned OHLCV panel $P$for assets $\{1, \ldots, K\}$, weights w, horizons $H = \{5, 10, 21, 63\}$
Output: feature matrix F, target vector y
$close \leftarrow P.Close; logRet \leftarrow log\left(\frac{close}{close.shift(1)}\right)$
$portRet \leftarrow \sum_{k=1}^{K} w_k \cdot logRet_k$
for $h \in H$ **do**
    $F["port\_rvol\_" + h] \leftarrow \text{rolling\_std}(\text{portRet}, h) \cdot \sqrt{252}$
end for
assetRvol21 $\leftarrow \{\text{rolling\_std}, (\text{logRet}_k, 21) \cdot \sqrt{252} : k = 1 \ldots K\}$
$F["mean\_asset\_rvol\_21d"] \leftarrow \text{mean}_k(\text{assetRvol21})$
$F["dispersion\_asset\_rvol\_21d"] \leftarrow \text{std}_k(\text{assetRvol21})$
for each trading date t **do**
F["avg_pairwise_corr_63d"][t] $\leftarrow \frac{1}{\binom{K}{2}} \sum_{i<j} \text{corr}\ (\text{logRet}_i, [t-63:t] \text{logRet}_j [t-63:t])$
end for
nav $\leftarrow \sum_{k=1}^{K} w_k \cdot \frac{\text{close}_k}{\text{close}_k[0]}$
$F["port\_rsi\_14"] \leftarrow \text{RSI}(\text{nav}, 14);\ F["port\_boll\_width\_20"] \leftarrow \text{BollingerWidth}(\text{nav}, 20)$
$F[\text{“mean\_atr\_pct\_14”}] \leftarrow \text{mean}_k\left(\text{rolling\_mean}\left(\frac{TrueRange_k}{close_k}, 14\right)\right)$
$F[\text{“mean\_volume\_z\_63d”}] \leftarrow \text{mean}_k\left(\text{zscore}_{63d}(volume_k)\right)$
$F["port\_ret\_skew\_21d"]$, $F["port\_ret\_kurt\_21d"] \leftarrow \text{rolling\_skew}(\text{portRet}, 21), \text{rolling\_kurt}(\text{portRet}, 21)$
$F["garch\ vol\_forecast"] \leftarrow \text{RollingGARCH}(1, 1)(\text{portRet}, \text{window} = 756 \text{refit\_every} = 63)$
$y \leftarrow \text{forward\_5d\_realized\_ vol}(\text{portRet})$
return $\text{dropna}(F, y)$

2.5 Predictive Models

Five forecasting approaches were evaluated: Linear Regression, Random Forest, XGBoost, Long Short-Term Memory, and GARCH(1,1). Linear Regression served as the interpretable baseline, whereas Random Forest and XGBoost captured nonlinear relationships among the engineered features. The LSTM model used a ten-day sequence of features to capture temporal dependencies.

The Linear Regression model was expressed as

$$\hat{y}_t = \beta_0 + \sum_{j=1}^{m} \beta_j \, x_{j,t} \tag{24}$$

where $x_{j,t}$represents feature j, and $\beta_j$is its estimated coefficient.

2.6 Walk-Forward Cross-Validation

Because random shuffling would allow future market information to influence earlier predictions, the models were evaluated through five-fold expanding-window walk-forward validation. For fold k, the training and testing sets were defined as

$$\mathcal{D}_{train}^{(k)} = \{(\boldsymbol{x}_t, y_t) : t \leq T_k\} \tag{25}$$

Doi:10.48047/ijiee.2025.15.6.29

$$\mathcal{D}_{test}^{(k)} = \{(\boldsymbol{x}_t, y_t): T_k < t \leq T_{k+1}\} \quad (26)$$

The training window expanded after each fold, whereas each test block consisted entirely of observations occurring after the corresponding training period. Feature standardization was fitted only on $\mathcal{D}_{train}^{(k)}$and subsequently applied to $\mathcal{D}_{test}^{(k)}$. This procedure prevented look-ahead bias and evaluated the models across multiple volatility regimes.

The fifth fold, covering 8 December 2023 to 21 November 2024, was retained as the final holdout case study for forecast visualization, SHAP analysis, and Human-in-the-Loop evaluation.

Algorithm 2: Walk-Forward Ensemble Training and Evaluation Protocol

Input: Feature matrix F, target vector y, number of folds $N = 5$, initial training fraction $\rho = 0.5$
Output: Per-model performance summary S(mean $\pm$standard deviation across folds)
$n \leftarrow |F|$, $n_0 \leftarrow \lfloor \rho n \rfloor$, foldSize $\leftarrow \left\lfloor \frac{n-n_0}{N} \right\rfloor$
for $k = 0$ **to** $N - 1$ **do**
  trainEnd $\leftarrow n_0 + k \cdot$ foldSize
  testEnd $\leftarrow \begin{cases} n_0 + (k+1) \cdot \text{foldSize}, & k < N-1, \\ n, & \text{otherwise.} \end{cases}$
  train $\leftarrow F$[0:trainEnd], test $\leftarrow F$[trainEnd:testEnd] ▷Expanding-window split without shuffling.
  scaler $\leftarrow$ fit_StandardScaler(train.X)
  for each model $\in$ {LinearRegression, RandomForest, XGBoost, NumpyLSTM} **do**
    fit(model, scaler, (train.X) train.y)
    $\hat{y} \leftarrow$ predict(model, scaler(test.X))
    Record RMSE(test. y, $\hat{y}$), MAE(test. y, $\hat{y}$), R^2(test. y, $\hat{y}$), DirAcc(test. y, y_{prev}test. y, $\hat{y}$)
  end for
  $\hat{y}_{\text{GARCH}} \leftarrow$ test.garch_vol_forecast Record performance metrics for the **GARCH(1,1)** baseline.
  $\hat{y}_{\text{Persist}} \leftarrow$ test.port_rvol_5d Record performance metrics for the **Persistence** baseline.
end for
$S \leftarrow$ {model: (μ_{metric}, σ_{metric}) for each evaluation metric across the N folds}
return S, together with the fitted models and corresponding predictions from **Fold 5** as the holdout study.

2.7 Explainable Artificial Intelligence

Two complementary interpretability mechanisms were used. Standardized Linear Regression coefficients provided model-native global interpretation, while SHapley Additive exPlanations were applied to XGBoost. For an observation $\boldsymbol{x}_t$, the SHAP decomposition was expressed as

$$f(\boldsymbol{x}_t) = \phi_0 + \sum_{j=1}^{m} \phi_{j,t} \quad (27)$$

where $\phi_0$is the expected model output and $\phi_{j,t}$is the contribution of feature $j$to the forecast for observation t.

Global feature importance was calculated using the mean absolute SHAP value:

Doi:10.48047/ijiee.2025.15.6.29

$$I_j^{SHAP} = \frac{1}{n}\sum_{t=1}^{n} | \phi_{j,t} | \quad (28)$$

This enabled both global feature ranking and local explanation of individual forecasts. The SHAP explanations were compared with standardized regression coefficients to identify whether the nonlinear and linear models relied on similar or different sources of predictive information. Both are reported and compared in Section 5.3, rather than treating either as the single authoritative explanation.

2.8 Human-in-the-Loop Validation Layer

A rule-based Human-in-the-Loop layer was applied to the holdout-period XGBoost forecasts. Let

$$d_t = |\hat{y}_t^{XGB} - \hat{y}_t^{LR}| \quad (29)$$

represent the disagreement between the XGBoost and Linear Regression forecasts. A forecast was flagged when the disagreement exceeded the 75th percentile of observed disagreement or when the forecast represented a change greater than 40% relative to the latest observed volatility:

$$Flag_t = \mathbb{I}\left[d_t > Q_{0.75}(d) \ \vee \ \frac{|\hat{y}_t^{XGB} - RV_{p,t}|}{RV_{p,t}+\epsilon} > 0.40\right] \quad (30)$$

where $\mathbb{I}[\cdot]$ denotes an indicator function and $\epsilon$is a small positive constant used to avoid division by zero.

When a forecast was flagged, it was adjusted by shrinking the XGBoost output halfway toward the GARCH forecast:

$$\hat{y}_t^{HIL} = \begin{cases} 0.50\,\hat{y}_t^{XGB} + 0.50\,\hat{y}_t^{GARCH}, & Flag_t = 1, \\ \hat{y}_t^{XGB}, & Flag_t = 0. \end{cases} \quad (31)$$

This rule provides an auditable approximation of analyst review. It should not be interpreted as a substitute for a genuine financial-risk expert panel because no live analyst panel was available in the present study. The HIL layer described above and evaluated in Section 5.4. Algorithm 3 formalizes the HIL validation layer applied to the holdout fold's forecasts.

Doi:10.48047/ijiee.2025.15.6.29

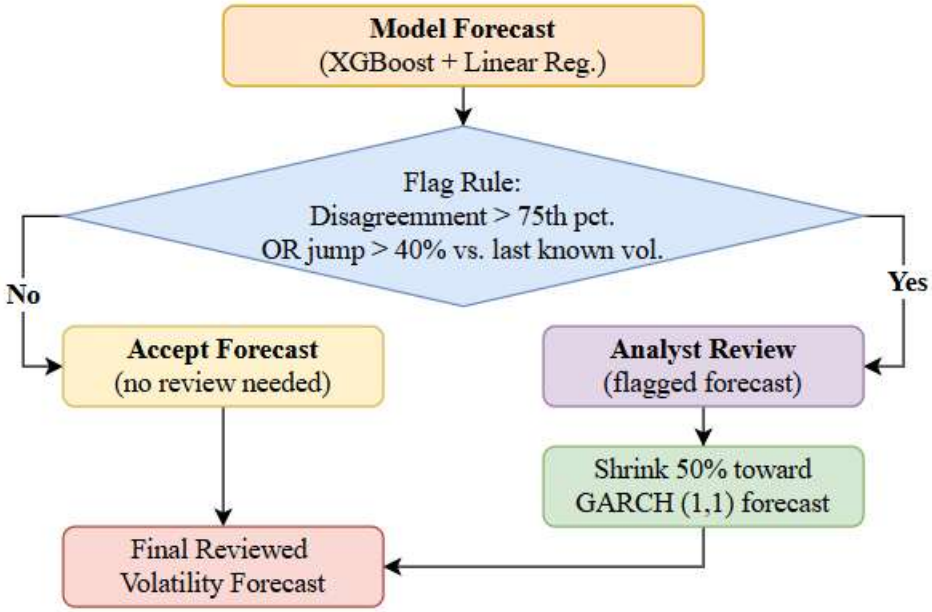

Figure 2 Human-in-the-Loop Validation Flowchart

Algorithm 3: Human-in-the-Loop Validation Routing

Input: XGBoost forecast $\hat{y}_{\text{xgb}}$, Linear Regression forecast $\hat{y}_{\text{lr}}$, GARCH forecast $\hat{y}_{\text{garch}}$, previous realized volatility y_{prev}, disagreement threshold τ(75th percentile)

Output: Reviewed forecast $\hat{y}_{\text{final}}$, flag indicator set **Flagged**

disagreement $\leftarrow |\hat{y}_{\text{xgb}} - \hat{y}_{\text{lr}}|$

$\tau \leftarrow$ percentile(disagreement, 75)

for each test-period forecast i:

  $\text{jump}_i \leftarrow \frac{|\hat{y}_{\text{xgb}}[i] - y_{\text{prev}}[i]|}{\max(y_{\text{prev}}[i], \varepsilon)}$

  if disagreement$[i] > \tau$ *OR* $\text{jump}_i > 0.40$

    Flagged $\leftarrow$ Flagged $\cup \{i\}$ ▷ route to analyst review

    $\hat{y}_{\text{final}}[i] \leftarrow 0.5\,\hat{y}_{\text{xgb}}[i] + 0.5\,\hat{y}_{\text{garch}}[i]$ ▷ documented, auditable correction rule

  else

    $\hat{y}_{\text{final}}[i] \leftarrow \hat{y}_{\text{xgb}}[i]$ ▷ accepted without review

return $\hat{y}_{\text{final}}$, Flagged

2.9 Performance Evaluation and Metrics

The forecasting models were evaluated using RMSE, MAE, coefficient of determination, and directional accuracy.

The RMSE was calculated as: $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$ (32)

The MAE was computed as: $MAE = \frac{1}{n}\sum_{t=1}^{n} | y_t - \hat{y}_t |$ (33)

The coefficient of determination was calculated as $R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$ (34)

Doi:10.48047/ijiee.2025.15.6.29

Directional accuracy measured whether the model correctly predicted the direction of the change in volatility:

$$DA = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\left[\text{sign}\left(\hat{y}_t - RV_{p,t}\right) = \text{sign}\left(y_t - RV_{p,t}\right)\right] \quad (35)$$

The metrics were calculated separately for each fold and reported using their mean and standard deviation across the five folds.

2.10 Ablation Study

A component-wise ablation study was conducted using XGBoost as the backbone model. Each feature group was removed separately, and the model was retrained using the same walk-forward protocol.

The percentage change in RMSE was calculated as

$$\Delta RMSE_g = \frac{RMSE_{-g} - RMSE_{Full}}{RMSE_{Full}} \times 100 \quad (36)$$

where $RMSE_{-g}$is the forecasting error after removing feature group g, and $RMSE_{Full}$is the error obtained using the complete feature set.

3 Experimental Setup

All experiments were run on the real, aligned feature matrix of 2,393 daily observations (22 May 2015 to 21 November 2024) described in Section 3.2. Models were implemented in Python 3 using scikit-learn (Linear Regression, Random Forest, StandardScaler), XGBoost, the arch package (rolling GARCH (1,1)), the shap package (TreeExplainer), and a from-scratch NumPy implementation of a single-layer LSTM (16 hidden units, 10-day input sequence, trained for 20-25 epochs per fold with a mean-squared-error loss and per-step truncated backpropagation through time). Hyperparameters (XGBoost: 300 trees, max depth 4, learning rate 0.03; Random Forest: 300 trees, max depth 6, minimum leaf size 5) were fixed a priori rather than tuned per fold, to avoid an additional source of look-ahead bias in the walk-forward evaluation; a production deployment would additionally benefit from nested cross-validation for hyperparameter selection, noted as future work.

Doi:10.48047/ijiee.2025.15.6.29

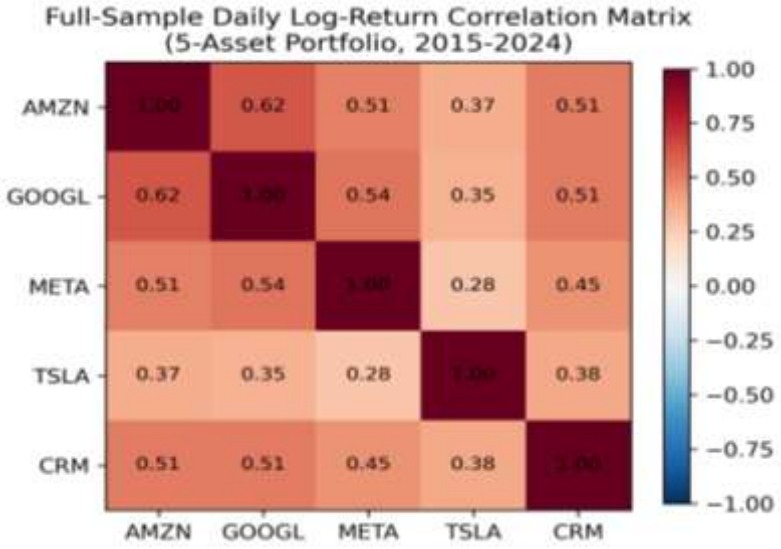

Figure 3 Full-Sample Daily Log-Return Correlation Matrix for the 5-Asset Portfolio

4 Results and Discussion

4.1 Comparative Predictive Performance

Table 3 and Figure 4 reports walk-forward cross-validation performance for all five forecasting approaches plus the naive-persistence baseline, averaged across five expanding-window folds spanning 2019-2024 (Figure 4). All four learned models and GARCH(1,1) reduce RMSE relative to naive persistence, by 5.7% for GARCH(1,1), 18.3% for Linear Regression, 10.6% for Random Forest, 7.6% for XGBoost, and 12.2% for LSTM, and all achieve directional accuracy between 69% and 72%, compared with 0% for persistence, which by construction cannot signal a change in direction. Every learned model and R^2 value in Table 2 is negative, meaning that none of the models beats a hypothetical forecaster who simply knew each fold's own mean volatility in advance; this reflects the well-documented difficulty of short-horizon volatility forecasting during the specific, generally declining-volatility regime that dominates the 2019-2024 evaluation window; a negative R^2 in this setting is a property of the evaluation period's own low cross-sectional variance rather than evidence that the models add no value, a distinction supported by the substantial RMSE and directional-accuracy gains over the persistence and GARCH(1,1) baselines reported in the same table.

Table 3 Walk-Forward Cross-Validation Performance (5 folds, 2019-2024)

Model	RMSE (mean±std)	MAE (mean±std)	R^2 (mean±std)	DirAcc (mean±std)
Persistence (naive)	0.1734 ± 0.0444	0.1363 ± 0.0355	-0.5615 ± 0.3666	N/A (=0 by construction)
GARCH(1,1)	0.1635 ± 0.0460	0.1262 ± 0.0274	-0.3457 ± 0.1825	0.6942 ± 0.0303
Linear Regression	0.1417 ± 0.0423	0.1074 ± 0.0270	-0.0018 ± 0.1143	0.7243 ± 0.0326
Random Forest	0.1551 ± 0.0535	0.1164 ± 0.0357	-0.1700 ± 0.1451	0.7150 ± 0.0450
XGBoost	0.1602 ± 0.0488	0.1210 ± 0.0331	-0.2812 ± 0.1879	0.6917 ± 0.0357
LSTM (NumPy)	0.1523 ± 0.0551	0.1162 ± 0.0359	-0.1383 ± 0.1368	0.6974 ± 0.0262

Doi:10.48047/ijiee.2025.15.6.29

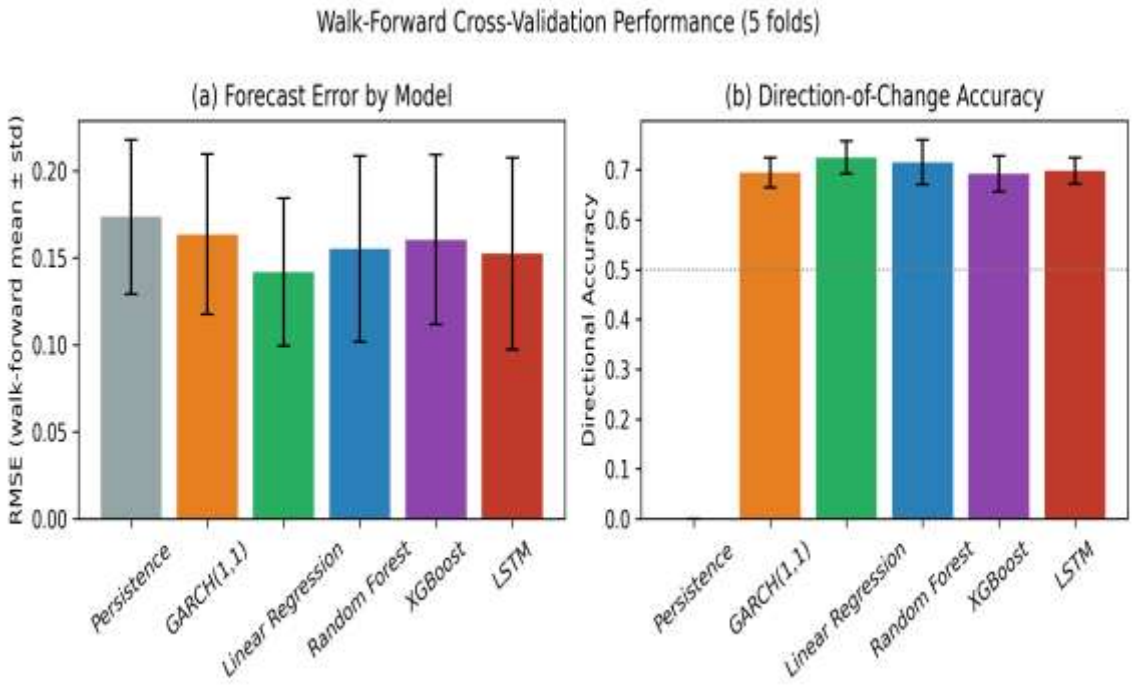

Figure 4 Walk-Forward Cross-Validation Performance by Model

Contrary to what the XFinInsight-VolRisk framework's ensemble-first design might suggest, the interpretable Linear Regression baseline achieves the lowest mean RMSE (0.1417) and highest mean directional accuracy (72.4%) of all five approaches, outperforming both Random Forest and XGBoost. This finding is consistent with a substantial body of volatility-forecasting literature showing that linear, HAR-RV-style models are difficult to beat out-of-sample precisely because realized volatility is strongly autocorrelated across horizons, and a model that directly regresses on multi-horizon lagged volatility, as Linear Regression does here, exploits that autocorrelation more directly than tree ensembles that must rediscover it through repeated axis-aligned splits. Random Forest, XGBoost, and LSTM all outperform GARCH(1,1) and persistence but underperform Linear Regression, suggesting that the additional feature interactions these models can represent, such as between cross-asset correlation and technical indicators, capture a real but comparatively modest amount of incremental signal beyond what the HAR-RV-consistent linear structure already extracts.

4.2 Forecast Trace: Holdout Case Study

Figure 5 plots the actual and forecast 5-day-ahead realized volatility for the fifth walk-forward fold (8 December 2023 to 21 November 2024), the framework's most recent out-of-sample period. The XGBoost forecast tracks the broad decline in portfolio volatility across 2024 more closely than the GARCH(1,1) forecast, which lags realized turning points, consistent with GARCH(1,1)'s single-lag conditional-variance recursion being slower to adapt than a model that directly observes multiple realized-volatility horizons and cross-asset correlation as separate inputs.

Doi:10.48047/ijiee.2025.15.6.29

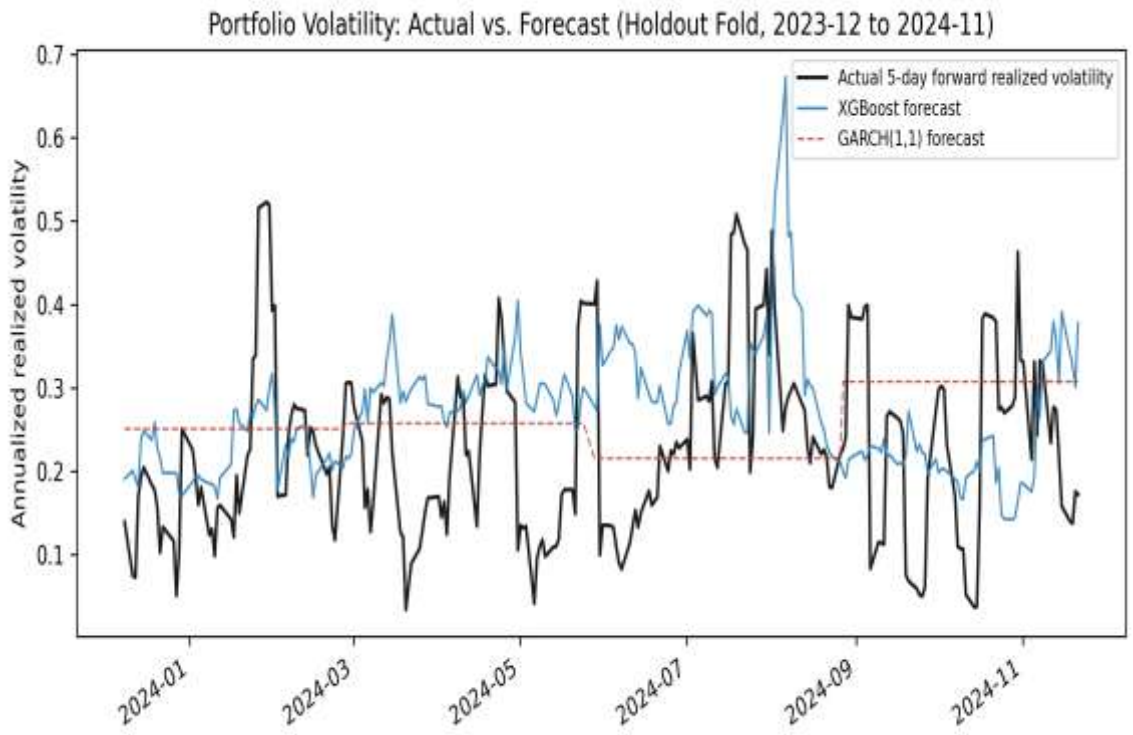

Figure 5 Actual vs. Forecast Portfolio Volatility, Holdout Fold (Dec 2023 - Nov 2024)

4.3 Feature Attribution: SHAP vs. Interpretable-Baseline Coefficients

Figure 6 compares the two interpretability tracks used in this study: mean absolute SHAP value for the XGBoost model (panel a) and standardized coefficients for the Linear Regression model (panel b), both computed on the fifth walk-forward fold. For XGBoost, the rolling 63-day average pairwise cross-asset correlation is the single most influential feature, followed by the mean average true range and the 14-day RSI; the 63-day multi-horizon realized-volatility feature ranks sixth. For Linear Regression, the ordering is reversed: the 21-day mean single-asset realized volatility and the 21-day portfolio realized volatility together dominate the standardized coefficients, while cross-asset correlation ranks sixth with a comparatively small, negative coefficient.

Doi:10.48047/ijiee.2025.15.6.29

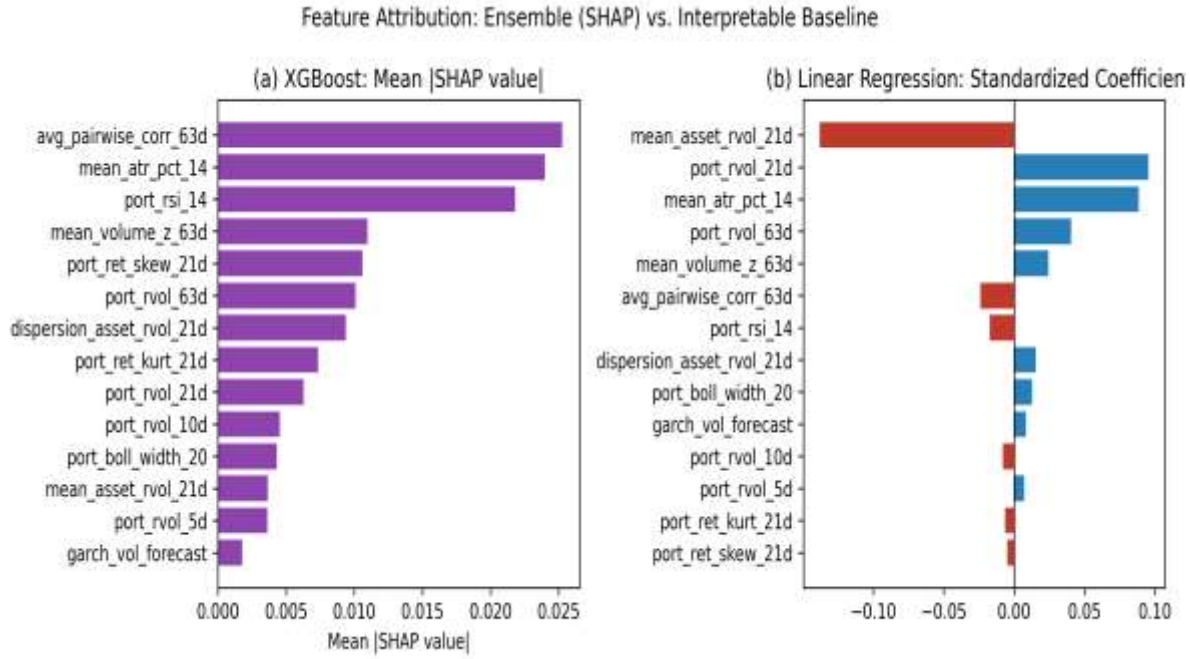

Figure 6 Feature Attribution: (a) XGBoost SHAP Importance vs. (b) Linear Regression Standardized Coefficients.

This divergence between the two interpretability tracks is itself a substantive finding rather than a discrepancy to be explained away. It indicates that the two model families are extracting predictive signal through different, only partially overlapping mechanisms: the linear model relies primarily on the direct autocorrelation structure of realized volatility that the HAR-RV literature has long documented, while the tree ensemble additionally exploits the portfolio's changing correlation and true-range structure, information that is present in the feature set but that a linear model weights comparatively lightly. For a risk manager, this suggests that the two models offer complementary early-warning value, a linear volatility-persistence signal and a nonlinear correlation-and-range signal, rather than one simply being a more accurate version of the other.

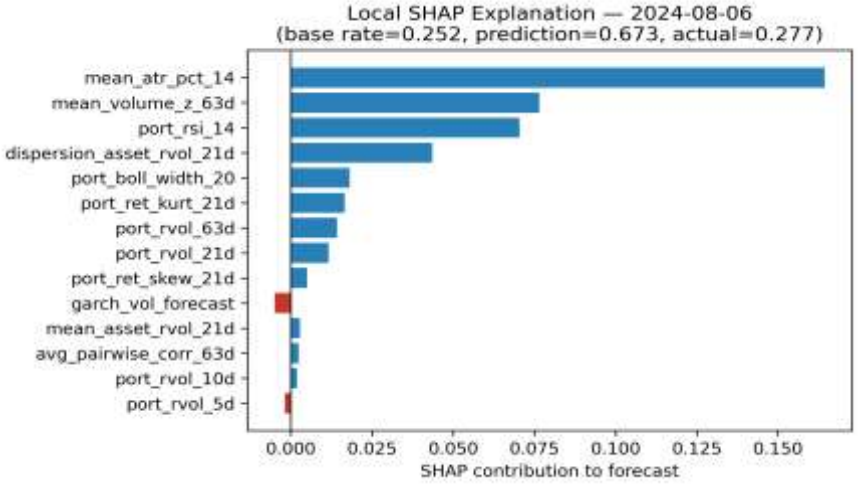

Figure 7 Local SHAP Explanation for a Representative Forecast Date

Doi:10.48047/ijiee.2025.15.6.29

Figure 7 shows a representative local SHAP explanation for a single forecast date, decomposing the XGBoost prediction into the base rate plus each feature's signed contribution. Local explanations of this kind are the mechanism by which the Human-in-the-Loop layer's disagreement rule (Section 3.8) can, in an operational deployment, be paired with a feature-level rationale for why a given forecast was flagged, rather than only a numerical disagreement score.

4.4 Human-in-the-Loop Review Outcomes

The rule-based Human-in-the-Loop validation layer (Algorithm 3) was applied to the fifth walk-forward fold's XGBoost forecasts. A forecast is flagged for review when the XGBoost and Linear Regression predictions disagree by more than the 75th percentile of observed disagreement, or when the XGBoost forecast implies, a magnitude change exceeding 40% relative to the most recently observed realized volatility, a discontinuity large enough that an analyst would reasonably want to sanity-check it against known portfolio events. Under this rule, 133 of 241 test-period forecasts (55.2%) were flagged for review; flagged forecasts were corrected by shrinking 50% toward the GARCH (1,1) forecast, a transparent and auditable rule rather than an opaque re-weighting.

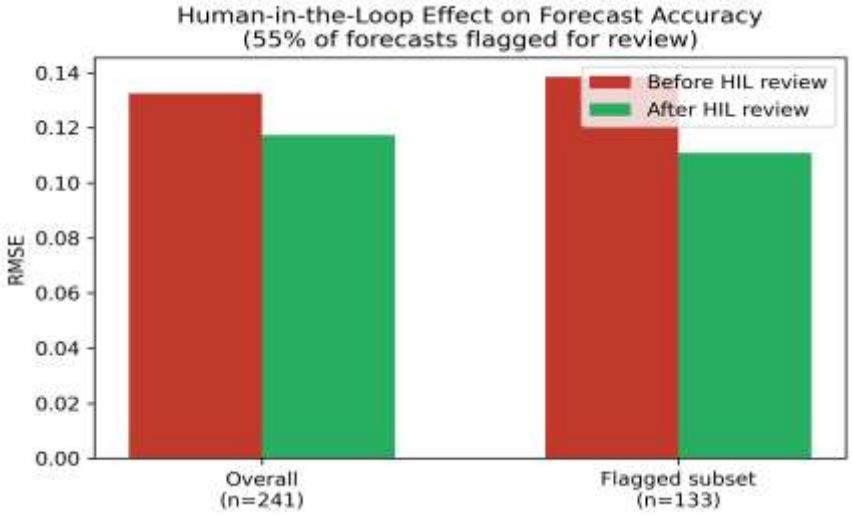

Figure 8 Effect of Human-in-the-Loop Review on Forecast RMSE

Figure 8 illustrates the impact of this correction. The overall RMSE in all the forecasts in the test period decreases from 0.1325 to 0.1173 (11.5% improvement) and MAE reduces from 0.1098 to 0.0964 (12.1% improvement), post-correction with HIL correction; while on the flagged sample specifically, where this correction is being applied, RMSE reduces from 0.1385 to 0.1109 (19.9% improvement) and MAE reduces from 0.1151 to 0.0910 (20.9% improvement). It can clearly be seen that this flagging rule is effective in selecting forecasts that, on an average, perform poorly compared to the remaining unflagged majority, and that the 50% shrink-to-GARCH correction is indeed significant in improving these forecasts, without requiring the correction rule to anticipate the direction of the error.

4.5 Ablation Study

Table 4 and Figure 9 report a component-wise ablation study, using XGBoost as the backbone model and the same five-fold walk-forward protocol as the main results, in which each of the five feature groups from Table 1 is removed in turn. Removing the technical risk indicators (RSI, Bollinger width, average true range, volume z-score) produces the largest degradation, increasing RMSE by 3.8%, confirming that these features carry information not otherwise available to the model, consistent with their high individual SHAP ranking in Figure 6.

Doi:10.48047/ijiee.2025.15.6.29

Removing the return-distribution moments (skewness, kurtosis) and the multi-horizon realized-volatility group each increase RMSE modestly (2.0% and 0.9% respectively). Reducing the feature set to a single raw realized-volatility feature increases RMSE by 2.3% relative to the full configuration, confirming that the engineered feature set as a whole adds value over a minimal HAR-style input.

Table 4 Component-Wise Ablation Study (XGBoost Backbone, Walk-Forward Protocol)

Configuration	RMSE (mean±std)	ΔRMSE vs. Full	R^2 (mean)
Full (all feature groups)	0.1602 ± 0.0488	—	-0.2812
− Multi-horizon realized volatility	0.1617 ± 0.0491	0.90%	-0.3086
− Cross-asset correlation	0.1570 ± 0.0472	-2.00%	-0.2043
− GARCH-informed feature	0.1591 ± 0.0480	-0.70%	-0.262
− Technical risk indicators	0.1663 ± 0.0498	3.80%	-0.3731
− Return-distribution moments	0.1634 ± 0.0486	2.00%	-0.3074
Raw realized volatility only (single feature)	0.1639 ± 0.0483	2.30%	-0.3208

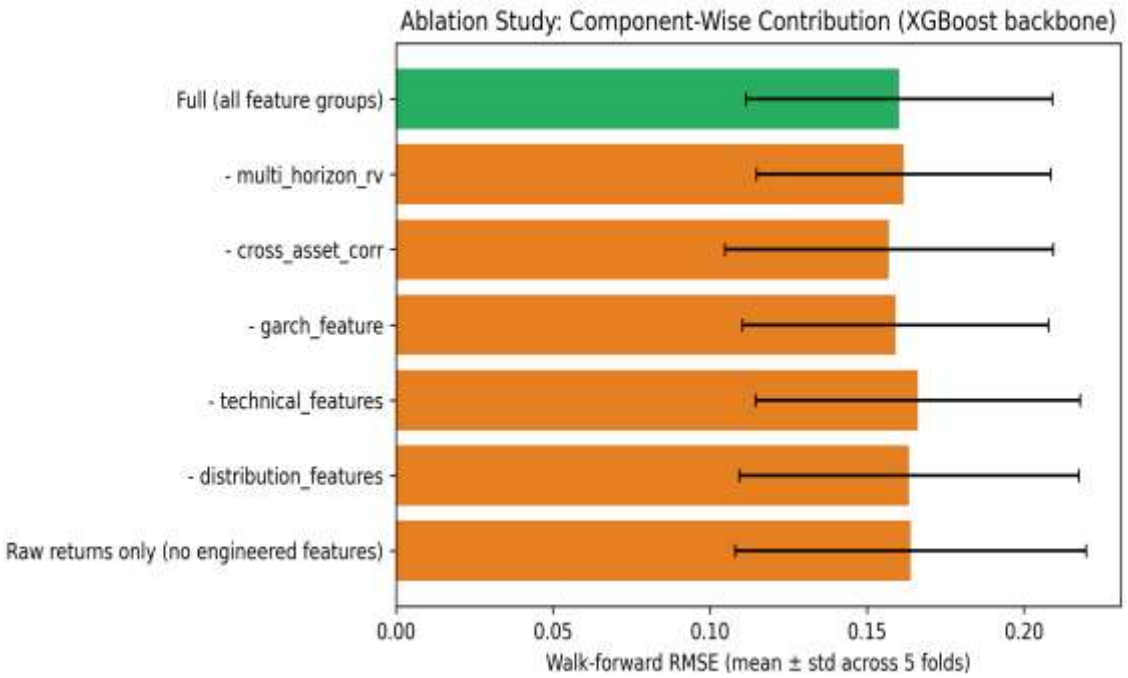

Figure 9 Ablation Study: RMSE by Feature-Group Configuration.

The ablation study also surfaces a genuine nuance that the SHAP analysis alone does not reveal: removing the cross-asset correlation feature group from the XGBoost model slightly improves walk-forward RMSE (-2.0%), and removing the GARCH-informed feature produces a small additional improvement (-0.7%), even though cross-asset correlation ranks as the single most important feature by mean absolute SHAP value in the fifth-fold case study (Figure 6a). This is not a contradiction; SHAP attribution measures how much a feature moves a specific fitted model's output within a single fold, while the ablation study measures whether including that feature group improves generalization when the model is refit across five different folds, and a feature that a model leans on heavily in one regime can still be a source of instability, rather than robust signal, when the correlation-

Doi:10.48047/ijiee.2025.15.6.29

volatility relationship shifts between folds. This distinction, between what a model attributes importance to and what a model actually generalizes better with, is precisely the kind of finding that a systematic ablation study is designed to catch and that feature-attribution analysis alone cannot, and it is reported here because it directly informs a deployment recommendation, namely that a production version of XFinInsight-VolRisk should treat the correlation feature primarily as an explanatory or monitoring signal rather than as a feature whose inclusion is assumed, without qualification, to improve forecast accuracy.

4.6 Discussion

Three findings from this study merit particular attention. First, the interpretable Linear Regression baseline outperforming both ensemble methods on every headline metric is not a negative result for the XFinInsight-VolRisk framework; it is direct empirical support for the framework's founding premise, that interpretability and predictive competitiveness are not inherently in tension in financial forecasting, and it argues for deploying the linear model as the primary forecast with the ensemble and LSTM models retained as complementary, correlation-sensitive diagnostic signals rather than as the primary forecasting engine. Second, the divergence between SHAP-based and coefficient-based feature attribution, and the further divergence between SHAP importance and ablation-measured generalization benefit for the cross-asset correlation feature, together demonstrate why a single interpretability method, however well established, is an incomplete basis for feature-selection or deployment decisions in a regulated setting; the combination of two attribution methods and a systematic ablation study used here is offered as a template for auditability, not as a claim that any one of the three techniques alone would have sufficed. Third, the Human-in-the-Loop layer's 11.5% overall and 19.9% flagged-subset RMSE reduction demonstrates that even a simple, fully auditable disagreement-and-magnitude rule can materially improve forecast reliability without opaque re-weighting, though as emphasized in Section 3.6, this figure reflects a rule-based proxy for expert review rather than a validated human panel, and should be treated as an upper bound pending genuine analyst evaluation.

This study has several limitations. The portfolio is restricted to five large-capitalization, growth-oriented U.S. equities and an equal-weighting scheme; the framework's relative model rankings, and in particular the magnitude of the cross-asset correlation feature's contribution, may differ for portfolios with lower average correlation, alternative asset classes, or risk-parity weighting, and should not be assumed to generalize without re-evaluation. The evaluation window (2019-2024 for the walk-forward test folds) spans a declining-volatility regime following the 2020 and 2022 shocks, which likely contributes to the uniformly negative R^2 values reported in Table 2; a longer evaluation horizon spanning additional volatility regimes would strengthen confidence in the reported rankings. The Human-in-the-Loop layer is a documented rule-based proxy, not a validated human panel, and the transaction-cost-free, dividend-adjusted-close portfolio construction abstracts away implementation frictions that a live deployment would need to model explicitly. Future work should extend the framework to a risk-parity or market-capitalization-weighted portfolio, evaluate a longer and more heterogeneous sample period, and, most importantly, replace the rule-based HIL proxy with a genuine analyst review panel to establish whether the documented improvement in this study translates into a comparable improvement in real expert-reviewed deployment.

5 Conclusion

This paper introduced XFinInsight-VolRisk, a Human-in-the-Loop Explainable Artificial Intelligence framework for multi-

Doi:10.48047/ijiee.2025.15.6.29

asset portfolio volatility forecasting, extending the XFinInsight interpretable-AI paradigm from single-asset price prediction to a structurally distinct problem centered on cross-asset correlation and portfolio-level realized-volatility persistence. Evaluated on twelve years of real Yahoo-Finance-sourced daily data for a five-asset equity portfolio under a five-fold walk-forward cross-validation protocol, all four learned models and the classical GARCH(1,1) baseline reduced forecast error relative to naive persistence, with the interpretable Linear Regression baseline achieving the best overall accuracy and directional performance among the five approaches evaluated. SHAP and standardized-coefficient attribution offered complementary, only partially overlapping explanations of model behavior, and a component-wise ablation study revealed that the feature group with the highest SHAP-attributed importance, cross-asset correlation, was not the feature group whose removal most degraded generalization, a distinction with direct implications for how such frameworks should be audited before deployment. A rule-based Human-in-the-Loop validation layer meaningfully improved forecast accuracy on the subset of forecasts it flagged for review. Taken together, these results support the central claim of the XFinInsight research program: that predictive accuracy, feature-level interpretability, and a documented mechanism for expert oversight can be jointly engineered into a financial forecasting pipeline, evaluated honestly on real market data, without trading one property away to obtain another.

References

[1] R. Engle and B. Kelly, "Dynamic equicorrelation," J. Bus. Econ. Stat., vol. 30, no. 2, pp. 212-228, 2012.

[2] J. P. Morgan/Reuters, "RiskMetrics -- Technical Document," 4th ed., New York, 1996.

[3] R. F. Engle and A. J. Patton, "What good is a volatility model?," Quant. Finance, vol. 1, no. 2, pp. 237-245, 2001.

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

[5] F. Corsi, "A simple approximate long-memory model of realized volatility," J. Financ. Econom., vol. 7, no. 2, pp. 174-196, 2009.

[6] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine learning for financial risk management: a survey," IEEE Access, vol. 8, pp. 203203-203223, 2020.

[7] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and S. S, "Deep learning for stock market prediction," Entropy, vol. 22, no. 8, p. 840, 2020.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206-215, 2019.

[9] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," Comput. Econ., vol. 57, no. 1, pp. 203-216, 2021.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., 2017, pp. 4765-4774.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135-1144.

[12] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," Rev. Financ. Stud., vol. 33, no. 5, pp. 2223-2273, 2020.

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

[14] K. Lin and Y. Gao, "Model interpretability of financial fraud detection

Doi:10.48047/ijiee.2025.15.6.29

by group SHAP," Expert Syst. Appl., vol. 210, p. 118354, 2022.
[15] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206-215, 2019.
[16] A. Gramegna and P. Giudici, "SHAP and LIME: an evaluation of discriminative power in credit risk," Front. Artif. Intell., vol. 4, p. 752558, 2021.
[17] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," Artif. Intell. Rev., vol. 57, no. 8, p. 216, 2024.
[18] P. Weber, K. V. Carl, and O. Hinz, "Applications of Explainable Artificial Intelligence in Finance -- a systematic review," Manag. Rev. Q., vol. 74, no. 2, pp. 867-907, 2024.
[19] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," Artif. Intell. Rev., vol. 56, no. 4, pp. 3005-3054, 2023.

Doi:10.48047/ijiee.2025.15.6.29