em lyon business school

FRIEZ Guillaume
Promotion 2025

Master of Science in Finance

Thesis Title

Market Risk Forecasting: A Comparative Study of Traditional, Machine Learning, and Hybrid Models with a Focus on Interpretability and Crisis Performance

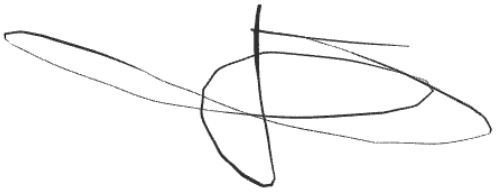
Date of defense: June 17th, 2025

Name of Supervisor: Guillaume COQUERET

Declaration of honor

I declare that I have personally prepared this master thesis and that it has not in whole or in part been submitted for any other degree or qualification. Nor has it appeared in whole or in part in any textbook, journal or any other document previously published or produced for any purpose. The work described here is my own, carried out personally, unless otherwise stated. All sources of information, including quotations, are acknowledged by means of reference, both in the final reference section, and at the point where they occur in the text.

Signature:



Date: 01/03/2025

Acknowledgement

I would like to extend my sincere gratitude to École de Management de Lyon for providing an enriching academic environment throughout my Specialized Master's journey. Completing this thesis represents not only the culmination of my studies but also a significant milestone in both my professional growth and personal development.

I am especially grateful to my tutor, Guillaume COQUERET, whose guidance and valuable insights greatly contributed to the quality of this dissertation.

I would also like to express my heartfelt thanks to Rothschild & Co and my manager Gabriel Ernault. During my internship as a Market Risk Analyst, I was first exposed to market risk topics and had the opportunity to work with the metrics that are discussed in this thesis. This practical experience played a crucial role in shaping my understanding of the subject.

With warmest regards,
Guillaume FRIEZ

Table of content

1 Introduction

1.1 Teaser

Financial markets suffered one of the most rapid collapses ever recorded in March 2020. The S&P 500 fell by more than 34% during 23 trading days while ETFs experienced drawdowns over 30%, Bitcoin lost nearly 50% of its value in a single day (CNBC, 2020), and the VIX volatility index reached an unprecedented 82.7. The shock cascaded across asset classes: The turbulent market conditions affected both bond ETFs and derivatives while cryptocurrencies faced unprecedented volatility which resulted in forced asset liquidations and heightened regulatory scrutiny. Yet, the most staggering fact lies behind the scenes: Major global banks encountered unanticipated breakdowns in their core risk models throughout this brief period.

According to the U.S. Federal Reserve's 2021 report backtesting failures would have led to capital requirement increases above $3.3 billion if emergency regulatory relief had not been implemented. Multiple instances demonstrate that this system-wide failure happened repeatedly. Value-at-Risk (VaR) models failed to detect structured credit markets fragility in 2008 as confidence levels of 99% were breached within days. As John Cassidy (2010) pointed out, risk models “not only failed to protect their users… they made a bad outcome more likely.”

These episodes highlight a persistent and unresolved challenge: how can we design risk forecasting systems that are accurate, adaptable, and explainable in the face of crisis? This thesis proposes to answer that question by comparing traditional, machine learning, and hybrid models through a unified empirical framework, with special focus on performance under stress and model transparency.

1.2 Literature Review

Scholars and practitioners have started reevaluating market risk definitions and measurements because traditional risk models repeatedly fail to predict extreme market movements. Faulty traditional risk models have uncovered fundamental flaws within many financial tools during structural breaks and periods of high uncertainty according to Danielsson (2016) and Jorion (2007).

At the same time, the proliferation of complex financial products, real-time data, and automated decision-making systems has raised new expectations: Models require accuracy along with interpretability and robustness against tail events and applicability across different asset classes and time horizons (Rudin, 2019; Gramegna & Giudici, 2021).

The literature review examines major theoretical and empirical advancements in market risk modeling while identifying traditional instruments and recent methodological changes. The existing tensions between model performance and explainability as well as between regulatory requirements and algorithmic complexity are identified and addressed in this thesis.

1.2.1 Classical Metrics for Risk Measurement

Market risk applies to every financial asset, but different products require different measurement methodologies. The most commonly used risk metrics for financial instruments are classified in this section which evaluates their pros and cons. This study uses the presented metrics as a standard to assess traditional models against machine learning and hybrid models.

1.2.1.1 Equities ETFs and Cryptocurrencies

Market risk impacts every financial asset, but measurement methods differ based on product type. The section examines widely used risk metrics by financial instrument categories while evaluating their advantages and limitations. The metrics introduced in this section establish a standard for evaluating traditional models against machine learning and hybrid approaches.

Mean, Variance and Standard Deviation

The mean represents the expected return level for an asset while variance defines how much actual returns will differ from this expected value. Standard deviation expresses dispersion in units that match returns because it represents the square root of variance and thus provides practitioners with a useful metric.

The Markowitz Mean-Variance model formulated in 1952 uses both variance and standard deviation as tools to represent investor strategies that maximize returns while keeping total risk at a predetermined standard deviation level.

Sharpe Ratio

The Sharpe Ratio (Sharpe, 1966) measures the added investment return for each unit of total risk as indicated by standard deviation. The Sharpe Ratio serves as a basic tool to evaluate performance for both equities and ETFs.

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

Where:

- R_p: portfolio return,
- R_f: risk-free rate,
- σ_p: standard deviation.

The model does not differentiate between positive and negative return fluctuations because it treats both types of volatility equally which constitutes a serious limitation.

Sortino Ratio

The Sortino Ratio was developed as an enhancement to the Sharpe Ratio to address one key limitation: The Sharpe Ratio treats positive and negative volatility as equally impactful which contradicts investor preferences that focus more on negative volatility risks. The Sortino Ratio assesses downside deviation through monitoring the frequency of returns that dip beneath a minimum acceptable return threshold usually set at the risk-free rate.

$$Sortino\ Ratio = \frac{R_p - R_t}{\sigma_d}$$

Where:

- R_p: portfolio return,
- R_t: target return (often the risk-free rate),
- σ_p: downside deviation = standard deviation of returns below R_t.

The Sortino Ratio stands out as the main metric for assessing assets with uneven return distributions including growth stocks and hedge funds and particularly cryptocurrencies because large infrequent gains can skew traditional volatility-based risk measures such as Sharpe. The Sortino Ratio provides a precise assessment of risk-adjusted performance in contexts with asymmetric return distributions (Sortino & Satchel, 2009).

It has gained popularity in portfolio management, especially where drawdown control or target-based returns are critical (Bailey & López de Prado, 2014).

Maximum Drawdown and Calmar Ratio

Maximum Drawdown stands as an intuitive risk measurement that directly speaks to investor concerns. Volatility-based metrics presume symmetrical return distributions, but Maximum Drawdown measures the greatest cumulative loss between a portfolio's highest and lowest points within any specified timeframe. Long-only portfolios along with ETFs and cryptocurrencies benefit from this measure because their investors show greater sensitivity towards capital losses rather than variations near the average return.

$$MDD = \min_{t\in[0,T]} \left(\frac{V_t - \max_{s\in[0,t]}(V_s)}{\max_{s\in[0,t]}(V_s)} \right)$$

Where:

- V_t: portfolio value at time t,
- The formula returns the largest % decline from a peak to a subsequent low.

Investors can normalize portfolio performance against drawdown risk by using the Calmar Ratio which Terry W. Young developed in the early 1990s; this metric evaluates the annualized return against the portfolio's maximum drawdown.

The Calmar Ratio proves most effective for investment classes showing serial correlation because their persistent drawdowns remain inadequately measured by regular standard deviation methods (Young, 1991).

$$Calmar\ Ratio = \frac{R_{annual}}{MDD}$$

Where:

- R_{annual}: average annual return,
- MDD: maximum drawdown over the same period.

Investors who prioritize capital protection and path-dependent risk assessment now frequently consider the Calmar Ratio alongside traditional metrics like the Sharpe and Sortino ratios. Lo (2002) explains that serial correlation skews standard deviation metrics like Sharpe while O'Shaughnessy (2014) shows investor concern for continuous drawdowns yet neither study compares these metrics to the Calmar Ratio which evaluates maximum drawdown in a way that mirrors investor experiences during market downturns.

1.2.1.2 Fixed Income and Interest Rate Products

Risk in bonds is often related to rate sensitivity and duration, not just price fluctuation.

Duration and Modified Duration

The duration of fixed income portfolios serves as a primary indicator of interest rate risk exposure. The concept calculates both the weighted mean time duration for bond cash flows and the bond price sensitivity to interest rate changes (Fabozzi, 2012). Modified Duration presents the percentage price alteration resulting from a 1% yield variation while maintaining constant cash flows and parallel yield curve movements (Tuckman & Serrat, 2011). Funds, insurers, and central banks need these metrics to manage bond portfolio sensitivity to interest rate changes.

Macaulay Duration : $$D = \frac{\sum_{t=1}^{T} \frac{t \cdot CF_t}{(1+y)^t}}{\sum_{t=1}^{T} \frac{t \cdot CF_t}{(1+y)^t}}$$

- Modified Duration: $$D_{mod} = \frac{D}{1+y}$$

Where:

- CF_t: cash flow at time t
- Y: yield to maturity
- T: maturity

Financial institutions use duration extensively for hedging purposes and to maintain asset-liability balance while also employing it for stress testing. Duration fails to account for convexity and can produce unreliable results when dealing with volatile or non-linear interest rate movements.

Convexity

Duration measures how a bond's price reacts to minor yield changes while convexity captures the price-yield relationship's curvature. The calculation uses a second-order approximation to achieve better accuracy during significant interest rate changes (Tuckman & Serrat, 2011).

Linear duration approximations fail to capture actual price movements for long-duration or callable bonds which is why convexity becomes an important measure. Portfolios that exhibit higher convexity will see lesser declines in value when interest rates increase and greater increases when interest rates decrease under identical conditions.

$$Convexity = \frac{1}{P}\sum_{t=1}^{T}\frac{CF_t \cdot t \cdot (t+1)}{(1+y)^{t+2}}$$

Where:

- P: bond price
- CF_t: cash flow at time t
- y: yield
- T: maturity

1.2.1.3 All Asset Classes – Distribution-Based Risk Measures

Certain risk metrics measure statistical return distribution properties instead of being linked to specific asset classes. The detection tools for asymmetries and fat tails that identify extreme losses work on equities as well as bonds and extend to commodities currencies ETFs and crypto assets. These risk metrics prove particularly useful in machine learning applications because these applications frequently move away from parametric assumptions.

Skewness and Kurtosis

Higher-order moments like skewness and kurtosis measure both the asymmetry and tail risk which are observable in return distributions. Traditional financial models generally use Gaussian return assumptions, but Cont (2001) presents strong empirical support for skewed and heavy-tailed distributions across multiple markets which become more apparent during periods of volatility.

A negatively skewed distribution shows that losses have a higher likelihood than gains of being substantial. In their 2003 study Jondeau and Rockinger demonstrate this pattern across different equity indices such as the S&P 500, DAX and CAC which show evidence of sudden price declines. Kurtosis, on the other hand, captures the thickness of the tails: A distribution with high kurtosis demonstrates that extreme market gains and losses occur more often than expected under normal distribution conditions.

These metrics remain important because VaR and Sharpe along with other widespread metrics depend on normality assumptions. Neglecting skewness and kurtosis results in underestimating tail risk while producing flawed backtesting outcomes which degrade model performance particularly in times of crisis.

Mathematically, they are defined as:

$$Skewness = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i - \bar{R}}{\sigma}\right)^3 \;\; ; \;\; Kurtosis = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i - \bar{R}}{\sigma}\right)^4$$

Where :

- R_i : individual return
- $\bar{R}$: mean return

- σ: standard deviation

The described measures serve both descriptive analysis purposes and support machine learning pipelines where they help detect non-normal patterns alongside building robust loss functions and creating custom tail-sensitive evaluation metrics.

Value-at-Risk (VaR)

The Value-at-Risk (VaR) metric remains the primary method to quantify market risk across financial industries. Value-at-Risk measures the highest possible loss of a portfolio during a defined period with a specific confidence rate. A daily VaR of $1 million at a 99% confidence level indicates that losses will remain below $1 million on 99% of trading days. Value-at-Risk (VaR) forms a fundamental element of Basel III regulations and can be calculated by parametric, historical simulation or Monte Carlo simulation methods (Jorion, 2007).

However, its main limitation is that it tells us nothing about the magnitude of losses beyond the threshold — i.e., in the worst 1% of cases.

This limitation is particularly critical in crisis periods, where tail events become more frequent and impactful (Danielsson, 2016).

$$VaR_\alpha = \mu - z_\alpha \cdot \sigma$$

Where:

- μ: expected return
- z_α: quantile of the normal distribution
- σ: standard deviation
- α: confidence level (e.g., 95%, 99%)

Conditional VaR (CVaR)

Expected Shortfall or CVaR functions alongside VaR by determining the expected average loss beyond the VaR threshold. Danielsson (2016) identifies the measure as a coherent risk measurement system because it shows properties such as subadditivity which help improve portfolio aggregation and capital allocation. CVaR proves to be an essential instrument within markets that display fat-tailed return distributions including commodities markets and leveraged ETFs due to their increased potential for extreme financial losses (McNeil et al., 2005).

$$CVaR_\alpha = \mathbb{E}[X \mid X \leq VaR_\alpha]$$

Where:

- X: Random variable representing portfolio returns
- VaR_α : Value-at-Risk at confidence level α
- $\mathbb{E}[]$: Expected value conditional on losses exceeding VaR

CVaR is preferred for capital adequacy under Basel IV, and in optimization frameworks that require tail-risk control (Rockafellar & Uryasev, 2000).

Extreme Value Theory (EVT)

Extreme Value Theory (EVT) exists as a statistical framework which analyzes the distribution tails where traditional risk models typically show limitations. This statistical approach excels at detecting infrequent yet devastating losses which makes it extremely valuable for financial risk management during crisis periods.

Extreme Value Theory (EVT) differs from parametric models like VaR that assume normality or historical continuity because EVT focuses on modeling the probability of extreme deviations using Generalized Extreme Value distributions for block maxima and Generalized Pareto Distributions for threshold exceedances according to McNeil et al. (2005).

Using these tools practitioners can calculate quantiles which go beyond any historical data thus making EVT essential for stress testing and capital adequacy planning as well as for backtesting during crisis periods.

$$GDP: F(x) = 1 - (1 + \frac{\xi x}{\beta})^{-1/\xi}$$

Where:

- ξ: shape parameter (tail heaviness)
- β: scale parameter
- x: excess over threshold

This section has described the fundamental metrics for market risk measurement including basic instruments such as volatility and drawdown as well as distribution-aware approaches like VaR, CVaR, and Extreme Value Theory. Each of these measures captures a different aspect of financial risk: These measures track different risk dimensions including average volatility patterns, potential losses on the downside, rare extreme events, and unequal distribution characteristics.

The metrics which remain vital for academic research and institutional practice expose significant limitations. A number of financial risk measures depend on assumptions about normality, symmetry, and linearity which fail during turbulent market conditions. Despite their theoretical strength certain metrics become difficult to estimate precisely when working with real-time data or limited datasets.

Because of these limitations more studies are using machine learning and hybrid models since they provide better adaptability for capturing non-linear patterns and regime changes and take into account a wider range of risk indicators. The following section provides an in-depth overview of these modeling frameworks starting with conventional approaches before moving to contemporary and understandable machine learning techniques.

1.2.2 Machine Learning for Market Risk Forecasting

Traditional risk models fail to accurately model tail behavior along with nonlinear interactions and regime shifts which has increased the interest in machine learning approaches. The 2008 and 2020 financial crises demonstrated the limitations of GARCH and historical VaR models which fail to identify sudden market changes and underestimate risk exposure.

Machine learning models serve as an adaptable data-driven solution. These models avoid stringent distributional and structural constraints which enables them to extract patterns from data and adjust to intricate and changing relationships. Machine learning algorithms predict Value-at-Risk (VaR), Conditional VaR (CVaR), volatility and drawdowns while utilizing these computational methods across different asset classes including equities, bonds, ETFs and cryptocurrencies. These models can handle

large datasets that contain different kinds of information including price history and macroeconomic indicators as well as market sentiment and technical variables. These models effectively reveal hidden patterns and model complex relationships which enables them to predict times of financial instability.

Recent research by Christensen, Siggaard, and Veliyev (2021) demonstrates that supervised learning methods trained on high-frequency data can outperform traditional benchmarks in volatility forecasting, offering both statistical improvements and better risk sensitivity.

The next subsections provide an introduction and evaluation of leading machine learning models used for market risk estimation. Our exploration starts with ensemble methods like Random Forests before moving on to boosting algorithms and recurrent neural networks as well as generative methods. The evaluation of each machine learning model is based on their predictive accuracy combined with their robustness and practical relevance.

1.2.2.1.1 Tree-Based Models

1.2.2.1.1.1 Random Forests

Random Forests represent ensemble learning methods presented by Breiman (2001) that utilize multiple decision trees to enhance prediction accuracy and prevent overfitting. Decision trees in Random Forests learn from bootstrapped data samples and utilize random feature subsets when splitting nodes.

The varied characteristics of the individual trees in the ensemble permit effective generalization despite noise and non-linear feature interactions which frequently occur in financial time series datasets. Random Forests are widely adopted in market risk applications to predict extreme events like Value-at-Risk breaches while classifying market regimes. Let y_t represent a binary outcome variable:

$$y_t = \begin{cases} 1 \; if \; loss_t > VaR \; threshold \\ 0 \; otherwise \end{cases}$$

The prediction of the binary target relies on several lagged features including past returns and realized volatility along with macro-financial indicators such as VIX and interest rate spreads as well as technical signals like drawdown and order book imbalance. Random Forests excel at capturing non-linear interactions automatically which makes them particularly effective in unstable or crisis situations. The co-occurrence of moderately high VIX levels and negative returns becomes predictive of imminent short-term risk events when paired with increased trading volume. Random Forests aim to take advantage of such complex multivariate dependency arrangements.

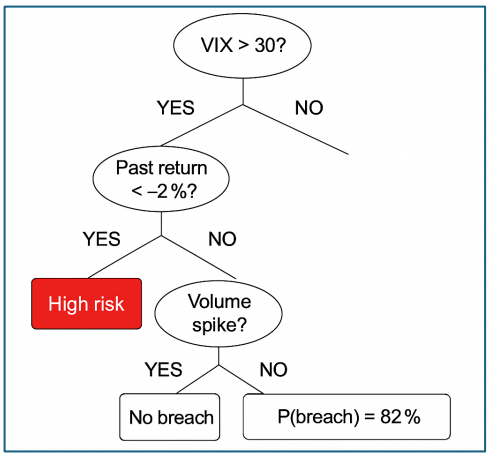


Figure 1 besides provides a simplified illustration of one decision tree within a Random Forest model. The tree sequentially splits on key financial indicators—such as VIX > 30 or return t–1 < –2%—to arrive at a terminal classification of risk exposure. Although the final model aggregates hundreds of such trees, each tree can be interpreted individually, allowing for localized transparency.

Figure 1- Simplified Random Forest Tree

Random Forests stand out because they remain easier to interpret compared to more complex deep learning models. According to Breiman (2001), feature importance scores generated by the algorithm quantify the average influence each variable has on model accuracy. Risk analysts apply these insights to identify the factors such as volatility spikes, liquidity shifts, and macroeconomic announcements that exert the most influence on predicted risk. The ability to interpret risk model outputs becomes essential for supervisors and audit teams who need accountability in regulatory settings.

However, Random Forests also have limitations. Random Forests utilize the assumption that observations function independently and neglect to incorporate explicit time dependency modeling. Time series forecasting reveals the limitations of these models because they fail to model volatility clustering unlike GARCH or LSTM which manage this well. Engineer lagged features by hand to simulate memory effects because Random Forests lack explicit time dependency modeling. The utility of feature importance falls short in illustrating how variables interact over time which requires recurrent or attention-based architectures to interpret.

Random Forests continue to be an effective and commonly used method in empirical finance even though they have these constraints. Their resilience to noise, low tendency to overfit, and ability to incorporate heterogeneous data types make them particularly suitable for high-dimensional market risk problems. Recent studies (e.g. Krauss et al., 2017) showed strong results in predicting tail risks and classifying stress regimes while merging financial and macroeconomic indicators with sentiment data into one predictive system.

1.2.2.1.1.2 Gradient Boosting and XGBoost

Gradient Boosting Machines (GBMs) represent a type of tree-based ensemble models which build trees in a sequential order unlike Random Forests that generate trees independently. The model sequentially trains new trees to address the residual errors from earlier trees which enables it to better handle observations that were difficult to predict initially. An iterative refinement method tends to increase the predictive accuracy of models when applied to structured datasets such as those used in financial risk prediction. The XGBoost algorithm developed by Chen and Guestrin in 2016 stands out as one of the most employed and highly optimized approaches for implementing gradient boosting machines. XGBoost enhances traditional GBMs through regularization methods like L1/L2 penalties as well as shrinkage through learning rate control and early stopping which serve to enhance generalization while minimizing overfitting which becomes crucial within unpredictable and noisy financial market conditions. XGBoost stands out because it works well with model-agnostic interpretability approaches such as SHAP (Lundberg & Lee, 2017) which provides feature-level importance scores based on cooperative game theory principles. This feature unlocks the ability to interpret complex predictions that display non-linear characteristics and involve diverse datasets with high dimensionality.

Suppose we aim to estimate the 1-day Value-at-Risk from a set of input variables $\boldsymbol{X}_t = \{r_{t-1}, \sigma_{t-1}, VIX_t, Volume_t, Macro_t\}$. Rather than specifying a parametric model like:

$$VaR_{t+1} = \alpha + \beta_1 r_{t-1} + \beta_2 \sigma_{t-1} + \cdots + \varepsilon_t$$

XGBoost will learn a non-linear function such that:

$$VaR_{t+1} \approx f(\boldsymbol{X}_t)$$

The model's adaptability allows it to detect threshold effects and interactions between different variables such as volatility and macro indicators along with sudden structural shifts in return distribution which traditional linear models find challenging to identify. Several empirical studies demonstrate successful applications of XGBoost.

The 2020 analysis showed that XGBoost outperformed linear factor models in asset pricing by better capturing risk premia. In risk forecasting applications, Krauss et al. (2017) proved XGBoost's superior accuracy in predicting tail events with S&P 500 constituents and recent studies confirmed its strong performance in equity volatility forecasting. However, like Random Forests, XGBoost does not inherently model time dependencies.

Since this model operates under the assumption of conditional independence among observations it lacks the capability to naturally represent volatility clustering or serial correlation unless lagged variables become part of its feature set. XGBoost delivers better accuracy than Random Forests in multiple settings but requires detailed tuning of hyperparameters such as learning rate and tree depth along with regularization coefficients.

The combination of XGBoost's superior performance and scalability along with its compatibility with explainability techniques positions it as an excellent choice for VaR and CVaR forecasting in high-dimensional data environments that span various asset classes. The thesis evaluates XGBoost as a benchmark model for regression-based tail risk estimation while also testing it as part of hybrid modeling frameworks.

1.2.2.1.2 Recurrent and Deep Learning Models

1.2.2.1.2.1 Long Short-term Memory Networks (LTSM)

Tree-based ensemble methods like Random Forests and XGBoost capture complex patterns yet lack mechanisms for modeling time-dependent structures within financial time series data. Long Short-Term Memory (LSTM) networks which Hochreiter and Schmidhuber introduced in 1997 represent a specialized type of Recurrent Neural Networks (RNNs) that overcome limitations through a memory-based system which enables long-term dependency retention. The LSTM cell has three gates including input, forget, and output that manage information flow between time steps and enable the network to decide which data to store or remove.

This enables LSTMs to model dynamic structures such as volatility clustering, regime persistence, and abrupt stress transitions, all of which are characteristic of financial risk data.

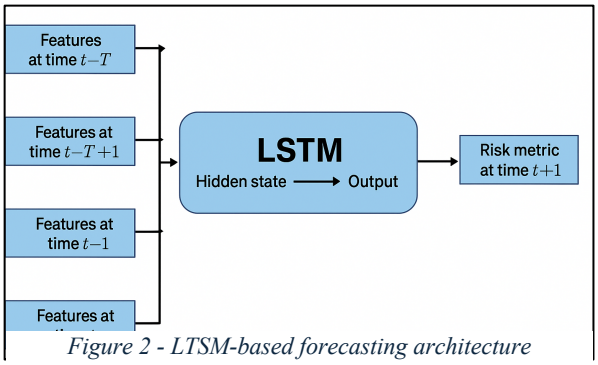


Figure 2 - LTSM-based forecasting architecture

The LSTM-based forecasting architecture schematic is displayed in Figure 2. A defined time window provides input to the model which consists of past observations sequences that feature lagged returns, realized volatility alongside macro indicators and technical features. The LSTM layers process the input sequences in order while encoding time-related dependencies. Typically, the final output passes through a dense layer and represents a predicted risk measure which might be

1-day Value-at-Risk (VaR), Conditional VaR or realized volatility. Research findings demonstrate that LSTM models surpass the performance of GARCH-type models as well as advanced machine learning approaches when dealing with volatile markets. Fischer and Krauss (2018) showed that LSTM models outperform other methods in predicting stock index movements.

LSTMs offer useful advantages however the substantial computational requirements along with their lack of transparency create major obstacles. Researchers must employ additional tools such as sequence-adapted SHAP or attention mechanisms to understand LSTM models.

The performance outcomes of LSTMs are highly influenced by the architectural design and the data preprocessing methodology used. LSTMs prove extremely useful in modern risk management because they can detect time-dependent structures and adjust to evolving risk conditions. This thesis presents an evaluation of LSTM models by comparing them with traditional and other machine learning methods through various assets and forecast time frames.

1.2.2.1.3 Explainable Machine Learning in Risk Forecasting

The growing use of machine learning models in market risk forecasting highlights interpretability as a crucial concern. Financial institutions function in heavily monitored settings which demand both model transparency and auditability as well as human monitoring for compliance and strategic purposes. Predictive power alone is not sufficient: risk models must be explainable.

Traditional techniques like GARCH or logistic regression produce models that give interpretable parameters unlike machine learning algorithms which work as black boxes. It is difficult to follow their internal logic especially when dealing with ensemble methods and deep learning architectures.

This creates a significant barrier to their deployment in contexts where justification of model decisions is essential—such as in backtesting reports, risk committee presentations, or supervisory stress tests.

To address this issue, a range of post hoc explanation tools and built-in interpretability frameworks have been developed. These tools aim to attribute model outputs to their inputs, provide localized explanations for individual predictions, or highlight structural patterns in time-series attention. This section reviews the three most prominent families of explainability tools applied in financial machine learning: SHAP, LIME, and attention mechanisms.

1.2.2.1.3.1 SHAP (SHapley Additive exPlanations)

SHAP is a comprehensive framework introduced by Lundberg and Lee in 2017 to explain machine learning model outputs through individual feature attribution. This approach relies on cooperative game theory principles and uses Shapley values to allocate the overall gain (model output) among features (players) according to their individual contributions.

The SHAP value ϕ_i associated for feature i in model f with input vector x quantifies how much that feature influences the gap between the model's prediction and the expected prediction $\mathbb{E}[f(x)]$:

$$f(x) = \mathbb{E}[f(x)] + \sum_{i=1}^{n} \phi_i$$

In the context of market risk forecasting, SHAP values can be used to explain predictions of VaR, CVaR, or volatility levels by decomposing the output into additive effects from lagged returns, volatility indicators, macro variables (e.g., VIX), and technical features. This enables risk managers to understand

which factors are driving forecasted risk—not just globally, across all predictions, but locally, for a specific date or portfolio scenario.

One of the main advantages of SHAP over other methods is its model-agnostic yet theoretically grounded nature. It is compatible with tree-based models like XGBoost and Random Forests, as well as with neural networks, and can produce both:

- Global explanations: which features are generally most influential;
- Local explanations: why this prediction was made.

For example, a 1-day VaR forecast for a crypto-asset ETF may be explained by SHAP values showing that a recent 10% drawdown, a high realized volatility window, and a rising VIX index were the dominant contributors to the high-risk estimate. Such insights can inform portfolio managers and be included in risk dashboards or internal documentation.

In practice, SHAP has already been used in regulatory risk contexts. It enables the development of audit trails, supports model validation efforts, and allows risk practitioners to perform sensitivity and stress simulations by quantifying how changes in input variables affect model outputs. Importantly, in the context of this thesis, SHAP will also be used as a quantitative criterion to assess model interpretability across methods.

1.2.2.1.3.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME, introduced by Ribeiro (2016) developed a method that interprets single predictions from complex machine learning models through the application of a simpler surrogate model like a linear one at the specific data point under investigation. Through the creation of nearby perturbed samples and monitoring the original model's reactions those samples produce LIME can deliver an understandable representation of the local decision boundary.

The LIME technique provides explanations for XGBoost model predictions regarding Value-at-Risk (VaR) breaches on particular days applied to market risk forecasting. The model's decision was primarily affected by either a volatility spike, recent drawdowns or abnormal trading volume levels. These explanations serve as valuable tools for risk assessment communication among decision-makers and for maintaining audit records in regulatory environments.

The LIME framework fails to generate consistent additive attributions across various inputs and exhibits sensitivity to the perturbation process unlike SHAP. The assumption of local linearity might fail specifically in financial data areas with non-linear patterns and interactions (Slack et al., 2020). In some instances, the explanations produced can become unstable and misleading.

LIME proves beneficial for initial diagnostic assessments and detailed local examinations. This research applies LIME together with SHAP to elucidate VaR breach model outputs and confirm prediction accuracy using case-specific backtesting evaluations.

1.2.2.1.3.3 Attention Mechanisms

Attention mechanisms improve the interpretability of recurrent models such as LSTMs by selecting important segments from the input sequence for focused processing. The model determines the significance of each time step through computed weights instead of summarizing all historical information into a single hidden state.

Formally, for an input sequence $\{h_1, h_2, \dots, h_T\}$, the attention output is computed as a weighted sum:

$$c = \sum_{t=1}^{T} \alpha_t h_t$$

where α_t are attention weights satisfying $\sum_{t=1}^{T} \alpha_t = 1$, and h_t represents the hidden state at time t. These weights α_t are learned during training and provide interpretable insight into which past observations the model considers most relevant.

Attention mechanisms in financial risk forecasting systems highlight crucial periods of market disorder and macroeconomic events that affect VaR and CVaR models' predictive accuracy. During crisis periods interpretability and transparency become crucial alongside predictive performance. The dual-stage attention-based RNN design introduced by Qin et al. The dual-stage attention-based RNN introduced by Qin et al. (2017) applies this mechanism to time series prediction which delivers improved accuracy and produces clear visualizations of temporal importance.

This section examined key machine learning methods for market risk forecasting and demonstrated how they solve traditional model limitations. Random Forests and XGBoost which are tree-based ensemble methods deliver robust prediction capabilities while providing some level of transparency. Recurrent neural network architectures such as LSTMs demonstrate superior performance in modeling both temporal dependencies and volatility dynamics compared with static models.

In risk-sensitive environments predictive accuracy fails to meet all necessary requirements. The Basel III and EBA model governance guidelines together with stress test requirements demand models that operate properly while staying auditable and interpretable and show resilience. The AI Public-Private Forum (Bank of England & FCA, 2022) report that users lose trust in outcomes from opaque and unexplainable models when these models are used in high-risk regulatory settings.

The thesis will prioritize interpretability as a fundamental evaluation dimension alongside forecasting accuracy and stress resilience. The models will be compared not only in terms of predictive precision but also based on the stability of their feature attributions across time, the plausibility and consistency of their local explanations, their alignment with financial intuition, their ability to highlight temporal dependencies through attention mechanisms, and their overall suitability for integration in a regulated institutional framework.

These criteria will form the basis for the empirical framework developed in the next chapter. The following section now turns to the formal articulation of the research question, the guiding hypothesis, and the structure of the comparative analysis.

1.3 Research Question, Hypotheses and Methodological Approach

Despite important achievements in financial forecasting through machine learning applications, its adoption in market risk modeling is incomplete and divided and fails to connect with regulatory and operational frameworks. Most financial research examines only one asset category or single-model assessments using limited performance metrics while overlooking critical factors like stability in financial crises and model transparency which financial risk management depends on.

The lack of extensive research stands as a major concern given the dual challenges of intensifying regulatory scrutiny alongside growing volatility across different asset types. Traditional risk models such as VaR, CVaR, and GARCH have displayed major predictive failures during systemic collapses demonstrated by both the 2008 financial crisis and the 2020 COVID shock (Federal Reserve, 2021) despite their widespread adoption.

While many ML-based approaches focus on predictive accuracy they frequently overlook concerns about interpretability and stability during structural regime changes. Research does not yet provide complete frameworks that compare traditional models with machine learning and hybrid models across different asset classes like ETFs, equities, bonds, and cryptocurrencies while assessing their performance across dimensions of forecast precision, risk resilience, and explanation capabilities (Kakade et al., 2022; Gramegna & Giudici, 2021).

This thesis aims to address this gap by answering the following central question:

Can machine learning and hybrid models achieve better accuracy and robustness in market risk forecasting while keeping interpretability intact compared to traditional methods for different asset classes during stressful market environments?

This thesis develops an integrated evaluation framework which merges traditional risk measurement tools like VaR, CVaR, and drawdown with cutting-edge machine learning algorithms such as XGBoost, LSTM and RF and explainability tools including SHAP and LIME. The evaluation investigates both the strengths and weaknesses of each method in terms of practical usage in crisis situations and regulatory compliance.

1.3.1 Defining the Scope and Structure of the Research Question

1.3.1.1 Nature of the Research Question

The thesis investigates an empirical research question focusing on comparative analyses of various study elements. This research evaluates whether machine learning and hybrid models can improve market risk prediction through enhanced accuracy and performance in both regulatory environments and financial crisis situations.

This study operates through an evaluative approach based on structured comparative analysis. Researchers evaluate different statistical models and both ML-based and hybrid approaches using standardized datasets and uniform evaluation standards.

This research aims to identify top-performing models while examining variations in their outputs between distinct asset classes and time horizons under various market conditions.

The thesis stands out because it explicitly uses interpretability as one of its performance evaluation dimensions. SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), and attention mechanisms serve as evaluation tools for model output justification which has become essential in risk-sensitive

environments (Bank of England & FCA, 2022). This research question evaluates modern and traditional risk models through multiple criteria while considering regulatory aspects to determine their real-world trade-offs and complementary benefits.

1.3.1.2 Scope of Risk Modeling

This thesis examines exclusively market risk which refers to potential losses from unfavorable changes in asset prices and volatility or macro-financial conditions. The research covers multiple risk dimensions within the domain such as volatility along with Value-at-Risk (VaR), Conditional VaR (CVaR), drawdown and tail behavior analysis. These elements serve as foundational components in institutional risk management systems and feature directly within regulatory guidelines like Basel III (BCBS, 2016) and the Fundamental Review of the Trading Book.

Credit risk, liquidity risk, and operational risk fall outside the modeling scope because they operate under distinct data structures, time dynamics and regulatory treatments. The different risk categories need separate modeling assumptions because their connections would reduce the empirical focus of the comparative framework.

The research places special attention on tail risk modeling because traditional methods often fail to predict extreme events (Danielsson, 2016). When evaluating models under stressful situations, maximum drawdown and CVaR risk measures take precedence over basic volatility indicators.

The thesis utilizes distribution-sensitive methods like skewness analysis and Extreme Value Theory to reflect empirical evidence showing that financial returns frequently stray from normal distributions during leverage effects or macroeconomic disruptions (Cont 2001; Taleb 2007)

This scope serves as a bridge between practical institutional applications while maintaining academic thoroughness and allowing for methodological comparison. The scope serves as a testing foundation for evaluating model performance against market risk factors that are non-linear, non-stationary, and exhibit heavy-tailed distributions.

1.3.1.3 Asset Classes Covered

This thesis evaluates market risk across four major asset classes: exchange-traded funds (ETFs), equities, bonds, and cryptocurrencies. The selected asset classes demonstrate a wide range of liquidity features which also show differing degrees of volatility alongside macroeconomic exposure. Exchange-traded funds and equities function as liquid tools for trading on exchanges which retail and institutional investors widely use in their investment approaches. The presence of interest rate and credit sensitivity in bonds establishes their critical role in fixed-income risk modeling (Fabozzi, 2012). The study incorporates cryptocurrencies because of their high volatility and regime-switching behavior which provides a testing environment for model generalization.

This thesis seeks to analyze model performance variations across different assets and their explainability by utilizing a diversified set while addressing an underrepresented research domain.

1.3.1.4 Time Horizons

The thesis presents model evaluations at three different time intervals including 1-day, 10-day, and 1-month durations. The design serves both regulatory purposes and practical risk management requirements. Under Basel III daily VaR reporting standards correspond to the 1-day horizon and the 10-day horizon applies to both stressed VaR assessments and internal capital adequacy evaluations. The 1-month time frame offers a medium-term outlook that supports portfolio rebalancing and enables macro risk anticipation and scenario testing.

It is imperative for models to maintain robustness across different time horizons because short-term risk events primarily mirror market microstructure and news shocks while structural volatility changes and macroeconomic trends shape long-term risk horizons (Jorion, 2007; Christoffersen, 2012). The thesis evaluates forecasting precision and the stability plus adaptability of each modeling approach through testing across various horizons.

1.3.1.5 Modeling Approaches Compared

This thesis compares three broad families of risk forecasting models: traditional, machine learning, and hybrid.

The traditional models Historical Simulation, Parametric VaR, CVaR, and GARCH-type models maintain the status of industry standards because of their straightforward design and regulatory approval (Jorion, 2007; McNeil et al., 2005). These models become unreliable during crises because they depend heavily on fixed distributional assumptions and do not handle structural breaks well (Danielsson, 2016).

The category of machine learning models encompasses Random Forests along with XGBoost and LSTM networks and GB. Nonlinear patterns and regime shift alongside complex interactions can be modeled by these machine learning approaches with no need for explicit parametric assumptions according to Krauss et al. (2017).

The GARCH-LSTM hybrid model seeks to integrate traditional statistical frameworks with machine learning adaptability which creates a balance between interpretability and flexibility (Kakade et al., 2022).

This comparative structure allows the thesis to assess the relative strengths, weaknesses, and trade-offs of each approach under consistent empirical conditions.

1.3.1.6 Evaluation Criteria

The models in this thesis will be evaluated across three key dimensions: forecasting accuracy, robustness under stress, and interpretability.

Standard backtesting methods measure accuracy through VaR exception counts and Kupiec and Christoffersen coverage tests along with error metrics like root mean squared error (RMSE) and quantile loss (McNeil et al., 2005; Christoffersen, 2012). The models can be directly compared in terms of their predictive performance on tail risk and volatility across different market conditions.

The analysis of robustness encompasses evaluations during actual historical market disruptions such as 2020 and includes additional simulated stress scenarios. We want to verify if models demonstrate stable performance when markets move away from stationary conditions and exhibit extreme patterns according to Danielsson (2016).

The interpretability of models is evaluated through methods including SHAP, LIME, and attention mechanisms while focusing on aspects like attribution stability and alignment with economic principles and regulatory requirements (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The models receive training from both financial variables such as returns and realized volatility along with macro-financial indicators like VIX and inflation. The framework evaluates predictive performance alongside practical usability of risk models within regulated finance sectors that experience volatility.

1.3.1.7 Constraints and assumptions

The research depends on datasets obtained from Refinitiv Eikon across equities, bonds, ETFs, and cryptocurrencies. Refinitiv delivers institution-grade data of high quality but restricts full reproducibility for external replication which remains an acknowledged limitation that methodological documentation helps to address.

All models are developed under harmonized preprocessing rules: The models are developed through uniform preprocessing which includes time alignment consistency and unified risk metrics together with rolling estimation windows. While this approach maintains comparative consistency among assets it can limit their individual optimization capability. Financial risk validation methods including VaR backtesting, tail loss analysis, and stress regime testing serve as the basis for evaluation procedures (Jorion, 2007; McNeil et al., 2005).

The evaluation framework for machine learning and hybrid models includes interpretability tools SHAP and LIME. The adoption of machine learning and hybrid models in time-series financial contexts continues to generate research interest because of their stability and relevance (Lundberg & Lee, 2017).

These elements work in combination to outline the research question's operational limits while creating a systematic approach for comparing models. The subsequent section outlines the comprehensive methodology which guides both implementation and evaluation of models within this complex risk landscape.

1.3.2 Existing Research Perspectives on Market Risk Modeling

Research into market risk forecasting spans multiple traditions but remains divided because studies tend to concentrate on one modeling technique or risk measure and often target one asset class. This section examines primary academic perspectives and identifies the critical research gaps the thesis proposes to fill.

1.3.2.1 Traditional quantitative approaches

Market risk measurement in academic research and institutional practice has always relied heavily on traditional risk models. The mathematical tractability of Value-at-Risk (VaR), Conditional VaR (CVaR), and GARCH-family models along with their regulatory support and implementation simplicity explains their widespread use (Jorion, 2007; McNeil, 2005).

Since the Basel regulatory framework integrated VaR in the 1990s this measure became the global standard. The method estimates potential portfolio losses in standard market conditions at confidence levels of 95% or 99%. VaR fails to qualify as a coherent risk metric because it provides no information about losses exceeding the threshold, a deficiency that theoretical critiques and post-crisis investigations have extensively documented (Danielsson, 2016). CVaR or Expected Shortfall represents a more robust measure of tail risk designed to address existing shortcomings.

The formally adopted Basel III (FRTB) market risk capital requirements utilize CVaR in their calculations of average losses from the worst α% scenarios while also fulfilling coherence criteria. The common practice of using VaR and CVaR depends on either distributional assumptions such as normality and t-distributions or historical simulation methods that fail to accurately measure risk in unstable market conditions (Cont, 2001).

GARCH-type models including GARCH, EGARCH, and GJR-GARCH are now standard instruments for modeling both volatility clustering and time-varying variance characteristics. The models

demonstrate efficient representation of short-term persistence alongside leptokurtosis in return distributions.

Traditional models provide transparency and adhere to regulatory standards yet they struggle to forecast tail events and adapt to structural shifts especially within complex and fast-changing market environments.

1.3.2.2 Machine Learning perspectives

Machine learning methods have become preferred tools in market risk forecasting because they demonstrate superior capabilities in identifying non-linear patterns and complex temporal structures within extensive data dimensions. Traditional models function on distributional assumptions whereas machine learning algorithms learn directly from data which makes them perfect for dynamic financial environments (Gu et al., 2020; Krauss et al., 2017).

Researchers increasingly use Random Forests, XGBoost models alongside Neural Networks with specific Long Short-Term Memory (LSTM) networks to forecast volatility patterns and Value at Risk violations as well as detect financial distress indications. According to Kakade et al. (2022), these models demonstrate superior performance over traditional methods during out-of-sample forecasts especially when dealing with high-frequency or noisy data. LSTM networks' ability to detect lengthy patterns and structural shifts in time series data makes them perfect for modeling market volatility during financial crises. However, many existing studies exhibit significant limitations.

Research efforts usually target individual asset categories like equity indices or distinct regional markets resulting in findings that cannot be extended to other market areas. When comparing models, researchers usually limit their criteria to narrow performance metrics like RMSE or classification accuracy and they do not evaluate how models perform under stress or how easy they are to interpret.

The ML literature provides strong market risk modeling capabilities but focuses on predictive accuracy while sacrificing transparency and robustness and ignoring practical deployment standards.

1.3.2.3 Hybrid modeling attempts

Numerous investigations have developed hybrid models which combine GARCH-type statistical frameworks with deep learning structures such as LSTM, GRU or GANs to integrate statistical modeling advantages with machine learning strengths. These models seek to improve prediction precision while preserving structural features and clarity from traditional methods.

Volatility forecasting now widely incorporates the combination of GARCH or EGARCH frameworks with neural network technologies. Kartsonakis-Mademlis and Dritsakis (2021) demonstrate that inputting conditional volatility estimates from an EGARCH model into a neural network leads to better predictive performance. Some research uses GARCH models to track short-term volatility patterns while utilizing LSTM networks to understand long-term dependencies which allows them to model both persistent behaviors and trend changes (Kakade et al., 2022).

Although hybrid models demonstrate empirical potential they face challenges due to the absence of standardization. The absence of agreement on neural network architecture design and training methods alongside evaluation standards complicates direct comparison of these systems. Most studies confine their research to a limited selection of assets and performance indicators while operating within restricted timeframes. The research community rarely evaluates interpretability or model stability across different market conditions along with regulatory framework adherence.

Hybrid approaches show potential, yet they remain insufficiently explored within academic publications. These methods show great potential to connect model transparency with non-linear features but require systematic testing under various financial conditions and evaluation metrics.

1.3.2.4 Explainability and regulatory gaps

The financial risk modeling field demonstrates increasing adoption of machine learning techniques but fails to address model explainability in scholarly work. Most research in this field focuses on achieving high predictive accuracy but fails to investigate the generation methods of model outputs or their understandability for risk managers and regulatory bodies (Rudin 2019; Gramegna & Giudici 2021).

Yet, explainability has become a regulatory imperative. The Bank of England and FCA (2022) together with the Federal Reserve established guidelines that demand algorithmic decision-making processes to be transparent and auditable while maintaining robust standards. Models applied in capital allocation processes and stress testing procedures as well as risk governance systems need to be sufficiently interpretable to defend their results when examined.

SHAP (Lundberg & Lee, 2017) alongside LIME (Ribeiro et al., 2016) enable post hoc interpretability through their ability to measure feature impact at the prediction level. Modern deep learning models now include attention mechanisms designed to highlight the inputs that direct model attention over time (Vaswani et al., 2017). However, very few studies in financial risk management have systematically evaluated these tools, especially under stress or during model failure.

Moreover, most of the existing literature treats explainability as an afterthought rather than a core evaluation criterion. There is little work comparing models not only by accuracy, but also by stability of feature importance, alignment with economic intuition, or robustness of explanations across regimes—dimensions that are essential for real-world deployment.

This thesis directly addresses the identified gap by incorporating interpretability as a fundamental element of the model evaluation framework. The study evaluates if sophisticated models maintain their transparency and usefulness despite the intricate nature of today's financial sectors.

When examined collectively these research strands show how fragmented the field currently stands. Traditional models maintain stability and transparency yet demonstrate poor adaptability when confronted with stress conditions. Machine learning methods deliver adaptive capabilities and strong predictive performance, yet they commonly neglect important aspects like model interpretability and stability. Hybrid models aim to balance these trade-offs but suffer from a lack of standardized procedures and widespread empirical proof. Practical applications now demand explainability as a core component, yet comparative studies have not fully explored this aspect.

This thesis establishes an empirical framework at the crossroads of multiple approaches to assess risk models based on performance metrics along with their transparency levels and resilience across different asset classes and market conditions.

1.3.3 Research Hypotheses

Based on the gaps identified in the existing literature and the multidimensional evaluation framework developed in this thesis, we formulate the following hypotheses:

The primary hypothesis (H_1) suggests that machine learning combined with hybrid models deliver better market risk predictions across various asset classes than traditional risk models especially during stress tests and achieve similar or better interpretability through the use of explainability tools.

This hypothesis will be tested empirically across a set of metrics reflecting three dimensions: predictive performance, robustness, and interpretability. To guide the empirical investigation, the following sub-hypotheses are also considered:

H_1a – Performance: ML and hybrid models outperform traditional approaches like GARCH and historical simulation by delivering more precise volatility, VaR and CVaR predictions especially in times of intense market volatility.

H_1b – Robustness: ML and hybrid models show improved stability and resilience across various market conditions including crisis periods according to performance deterioration and tail risk metrics under stress testing.

H_1c – Interpretability: Machine learning and hybrid models achieve interpretable results that match or exceed those of traditional models by using explainability methods such as SHAP, LIME, and attention to produce outputs with consistent attribution and financial logic alignment.

Null Hypothesis (H_0): Traditional models perform equivalently with machine learning and hybrid models concerning market risk forecasts across different asset classes and stress conditions based on accuracy, robustness, and interpretability.

The experimental design comparison detailed in the subsequent methodological section derives from these hypotheses.

1.3.4 Research Methodology Teaser

The research uses comparative empirical methods to determine if machine learning with hybrid models improves market risk forecasting beyond traditional methods. Multiple asset classes including equities, bonds, ETFs, and cryptocurrencies undergo assessment through a single evaluation framework which also undergoes testing across diverse market conditions and forecast periods.

The models under comparison fall into three broad categories: The comparison encompasses three model types which are traditional statistical models such as GARCH and historical VaR methods, machine learning models including Random Forests, XGBoost and LSTM approaches, and hybrid models like GARCH-LSTM techniques. Each model is implemented using a consistent set of financial and macro-financial features and evaluated according to three complementary criteria: forecasting accuracy, robustness under stress, and interpretability.

Backtesting metrics such as VaR coverage tests alongside quantile loss and volatility error determine model accuracy. Historical crises and multiple stress scenarios provide the basis for evaluating the system's robustness. The interpretability evaluation process utilizes SHAP and LIME explainability tools together with attention mechanisms to achieve stable attributions while preserving economic plausibility.

The complete modeling framework functions through Python language and utilizes Scikit-learn along with TensorFlow and Statsmodels libraries for its operations. Data for both market and macroeconomic analysis comes from Refinitiv Eikon which provides access to institutional-grade data as well as publicly available datasets. The same preprocessing techniques including frequency harmonization, windowing and feature standardization are applied consistently to achieve model comparability.

This research follows a design science logic: The study focuses on identifying the best performing and understandable model types under real-world circumstances instead of optimizing just one single model. The subsequent chapters deliver comprehensive descriptions of the datasets used along with modeling choices and empirical findings that validate this comparative study.

This research employs empirical quantitative data analysis to determine whether Machine Learning and hybrid models surpass traditional methods in market risk prediction capabilities. Machine Learning algorithms like XGBoost, Random Forest and LSTM together with hybrid GARCH-LSTM architectures are assessed against traditional models including VaR, CVaR and GARCH that function as benchmarks. The methodology works best for detecting non-linear patterns and regime changes that emerge during financial crises. The approach supports the combination of macroeconomic indicators with explainable AI methods such as SHAP and LIME to efficiently achieve goals related to predictive accuracy and finance analytics.

This study evaluates several hypotheses using a multi-asset high-dimensional dataset that merges financial market prices from equities, ETFs, bonds, commodities, and cryptocurrencies with macroeconomic variables such as VIX and interest rates. The methodology involves the following steps:

1. Gathering daily asset class data from Refinitiv.
2. Risk metrics construction through feature engineering to measure volatility and drawdowns along with tail statistics.
3. The research employs both classical and ML algorithms to generate 1-day and 21-day VaR predictions.
4. The evaluation phase includes backtesting with Kupiec and Christoffersen tests together with RMSE/MAE loss functions and crisis-period stress testing.
5. The research investigates predicted risk factors by employing SHAP and LIME for model interpretability analysis.

The approach provides a detailed and understandable comparative evaluation of risk forecasting models under stable and volatile market conditions.

2 Research Methods, data collection and analysis

2.1 Data Collection

2.1.1 Financial Market Data

The research creates a strong market risk forecasting framework by selecting appropriate financial assets exclusively from the Refinitiv Data Platform. Refinitiv delivers institution-grade data which maintains consistency across historical periods while also providing fundamental, market and technical details beyond just price quotes.

The aim is to build a representative set of financial instruments with diverse volatility behaviors and liquidity traits while being sensitive to macroeconomic changes instead of covering all global financial markets. The dataset is structured around five main asset classes: equities, ETFs, fixed income, commodities, and cryptocurrencies. Within each main asset class researchers divide them into sub-groups that share thematic or structural connections.

The dataset collects all instruments every day using standard file formats that enable automated data ingestion and pre-processing.

2.1.1.1 Data Source and Structure

The Refinitiv API extracts all financial instruments which are saved as structured tabular data. The data fields extend beyond the standard OHLCV (Open, High, Low, Close, Volume) set by featuring market and derived variables.

The standard columns for modeling involve several elements such as:

Column	Description
date	Observation date
price_open, price_close	Daily opening and closing prices
price_high, price_low	Intraday price extremes
adjusted_close	Adjusted price including corporate actions
volume, turnover	Raw and value-weighted trading activity
market_cap, free_float	Market capitalization and tradable share base
beta, volatility_30d	Risk metrics against benchmark or over time horizon

Figure 3 – Data source indicators

The available data granularity enables multivariate modeling which integrates return dynamics together with technical signals and liquidity effects while considering broader risk factor sensitivities.

2.1.1.2 Asset Classes and Sub-Groups

Five primary asset classes contain sub-groups which are organized according to investment strategy or economic exposure. The structure enables performance comparisons of models during different volatility conditions.

2.1.1.2.1 Equity ETFs

Investors gain access to diverse investments through exchange-traded funds that provide high liquidity. The dataset includes:

- Regional ETFs: ETFs that track broad equity markets split by geographic regions such as North America, Europe, and Emerging Markets.
- Sector ETFs: Sector ETFs target particular industries including technology, healthcare, and financials.
- Thematic ETFs: Thematic ETFs aim to capitalize on enduring structural trends like clean energy development, cybersecurity advancements, and artificial intelligence progress

These tools enable analysts to study systematic risk patterns, sectoral shifts, and macroeconomic volatility trends.

2.1.1.2.2 Individual Equities

Investors control firm-specific risk by including a targeted selection of large-cap and mid-cap stocks which are organized according to their sector and market role. Sub-groups include:

- Healthcare stocks: Typically defensive, with muted cyclicality
- Industrial stocks: Industrial stocks show greater sensitivity towards economic cycles and capital expenditure trends.
- Mega-cap tech: Mega-cap tech stocks possess high liquidity and influence yet face considerable risks from technological advancements.

These assets facilitate the modeling process for both idiosyncratic volatility and earnings event responses as well as microstructure noise.

2.1.1.2.3 Fixed Income

To capture exogenous risk and high-volatility behavior, major commodities are integrated. The sample covers:

- Energy commodities: Such as crude oil and natural gas, sensitive to geopolitical and supply shocks
- Precious metals: Like gold and silver, often used as inflation hedges or safe-haven assets

These assets are valuable for assessing model responsiveness to non-financial volatility sources.

2.1.1.2.4 Commodities

The research integrates major commodities' spot prices to assess external risks and volatility patterns. The sample covers:

- Energy commodities: Crude oil and natural gas among energy commodities show vulnerability to geopolitical developments and supply chain interruptions.
- Precious metals**:** Assets like gold and silver function as protection against inflation and safe-haven investments which enable analysts to assess their models' reactions to non-financial volatility sources.

These assets prove useful in determining how models respond to non-financial volatility sources.

2.1.1.2.5 Cryptocurrencies

The dataset recognizes digital assets as a separate category operating through decentralized market systems. Traditional models can be tested against extreme market conditions by including major crypto tokens like Bitcoin and Ethereum.

2.1.1.3 File Organization

Every asset resides in its own designated folder according to a logical and repeatable classification system. For instance:

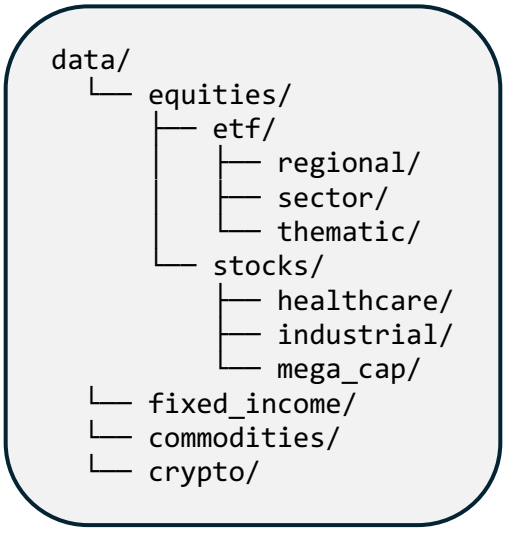


Figure 4 – Directory organization

The system structure provides scalable preprocessing capabilities while enabling smooth integration with macroeconomic data during later stages.

2.1.2 Macroeconomic Indicators

This research work combines financial instruments with a selected collection of macroeconomic and systemic indicators. The modeling framework utilizes these variables to capture global economic trends along with monetary policy changes and cross-asset volatility regimes.

Refinitiv serves as the source for all macro data which offers high-quality institution-grade time series information across multiple economic and financial fields. A single multivariate file holds the macroeconomic dataset which then undergoes specific naming convention-based filtering and processing. Our method retains essential indicators through its scalable system which allows future enhancements.

2.1.2.1 Indicator Selection Logic

The process of selecting macroeconomic variables operates automatically through pattern-matching rules that target column names. Series names that match terminal patterns determine which data points remain in the filtered dataset:

Regex Pattern	Likely Corresponding Variable
TRDPRC_1$	Last traded price (e.g., indices, rates, FX spot levels)
YLDTOMAT$	Yield to maturity (e.g., sovereign bonds)
=ECI_VALUE$	Economic indicator values (macro series from Refinitiv)
_MID_PRICE$	Midpoint of bid-ask spread (e.g., FX, swaps)

Figure 5 – Indicator selection

Through this strategy users can automatically filter important series from different categories including interest rates, commodity benchmarks, equity indices, and macroeconomic indicators without embedding specific variable names into the code. The dataset stays flexible when there are structural or name changes in the Refinitiv data source.

2.1.2.2 Preprocessing and Data Cleaning

Once initial filtering occurs the macro dataset undergoes a standardized cleaning process.

1. Removal of empty columns: The dataset rejects empty series entirely to stop meaningless information from contaminating the data.
2. Forward-fill of missing values: Gaps in time series data become filled by using the latest available data point as the imputation method. The assumption that macroeconomic or market-level indicators change incrementally supports the use of last observation carried forward for data imputation.
3. Min–Max normalization: A normalized version of each retained column is calculated to scale its values to fit within the range of [0,1]. We save the normalized data in a different column that uses the _normalized suffix.
4. Combined export: The final dataset contains both raw and normalized series versions saved as a clean CSV file for downstream modeling use.

2.1.2.3 Frequency and Synchronization

The present implementation does not manage variable sampling frequencies like monthly Consumer Price Index or quarterly Gross Domestic Product nor does it address publication lags unlike advanced macroeconomic pipelines. Instead:

- The Refinitiv export or prior transformation requires that all series adhere to a daily calendar framework.
- All missing dates receive their values through forward-fill techniques that extend the most recent known value to any non-trading days and release periods.
- There is no adjustment for publication timing: Each indicator's value becomes available according to its date, irrespective of the release or reference period status.

This simplified synchronization approach obscures precise macro variable timing relationships while allowing seamless daily model implementation without heavy data engineering work.

2.1.3 Temporal Scope and Historical Coverage

This thesis mainly examines which data needs collection and during which specific timeframes it should be analyzed. The duration of data analysis affects which risk conditions models encounter and how well their training holds up along with how applicable testing settings become. In this regard, a deliberate architectural choice has been made: The system preserves the original historical periods of each asset or indicator based on Refinitiv's available data instead of imposing fixed start and end dates.

The flexible historical framework supports both emerging financial instruments like thematic ETFs and cryptocurrencies and traditional financial products such as major indices and government bond ETFs. The analytical approach stays true to actual data availability while demonstrating how market structure has transformed throughout history.

2.1.3.1 Crisis Period Definition and Use

This study introduces a key innovation by systematically incorporating defined stress periods that serve as direct benchmarks for analyzing model performance during extreme volatility and structural breaks.

The research identifies four distinct episodes based on unique economic triggers and their effects on different asset classes along with their volatility characteristics:

Label	Start Date	End Date	Description
Crise 2008	2008-09-01	2009-03-31	Global Financial Crisis (Lehman aftermath)
Taper Tantrum	2014-05-01	2013-09-30	US Fed policy shock and yield spike
COVID	2020-02-15	2020-05-15	Pandemic-induced market collapse
Ukraine	2022-02-01	2022-04-30	Russian invasion of Ukraine and market shock

Figure 6 – Crisis periods definition

Model training and evaluation processes utilize these intervals beyond their labeling in the data pipeline:

- The performance of risk estimators such as VaR, CVaR, and EVT is assessed both before and after market shocks.
- The evaluation of machine learning models involves testing their predictive power and response capacity to sudden spikes in volatility.
- The backtesting process is divided into crisis windows to enable the calculation of risk-specific performance measures such as exception rates and drawdowns.

The scenario-aware structure enables models to optimize beyond average-case performance by subjecting them to rigorous testing under rare yet economically significant conditions.

2.1.3.2 Treatment of Temporal Mismatch Across Series

Because of the diverse sources of financial instruments and macroeconomic indicators, calendar synchronization is approached through practical means:

- Financial instruments maintain their original trading schedule, so equities operate based on exchange business days while cryptocurrencies trade including weekends.
- Macroeconomic series handle missing observations through forward-filling which preserves the last known values until the next data release.
- There is no forced intersection of start dates: Models can access historical data for each asset based on what's available at any given moment.

The methodology prevents the excessive standardization problem which often results in the loss of important historical data. It also reflects how real-world portfolio risk models operate: Financial instruments enter the market at different times and their performance indicators vary in both frequency and response times.

2.1.3.3 Temporal Continuity and Modeling Implications

The design of the dataset allows for:

- Rolling-window analysis with consistent lengths where applicable
- Tailored training windows depending on model type (e.g., expanding windows for GARCH, fixed-size for LSTM)
- Out-of-sample testing that aligns with business-relevant stress intervals

Additionally, each stress window includes not only the high-volatility core phase, but enough lead and recovery periods to evaluate how quickly models detect early signals and how they adapt as regimes change

In sum, the dataset does not just span time — it structures time. By explicitly integrating temporal diversity and targeting key crisis regimes, the modeling framework is able to test not only whether models “fit” the data, but whether they respond intelligently to the types of dislocations that define true market risk. This temporal architecture is an essential component of the thesis' contribution to robust risk forecasting.

2.2 Preprocessing and Feature Engineering

The collected financial and macroeconomic data undergo preprocessing steps to produce clean and standardized time series that are appropriate for risk modeling. Data formatting forms only one aspect of this phase which also encompasses targeted cleaning strategies and intelligent missing value treatment alongside the development of derived variables that improve model predictiveness.

The fundamental philosophy remains to protect original data integrity alongside implementing adaptable transformation pipelines for volatility analysis as well as tail-risk and anomaly detection purposes.

2.2.1 Data Cleaning and Standardization

Dynamic path logic enables each asset file to load and validate separately from other files.

The system automatically skips missing or empty files to prevent pipeline interruptions from occurring. The system parses date columns and standardizes indexes to achieve correct temporal alignment for subsequent operations. The cleaning process includes:

- The system automatically processes files that are corrupted or incomplete and logs these incidents.
- Parsing and formatting time indexes.
- The cleaning process removes columns that contain only NaN values which represent obviously invalid data.

The pipeline uses fallback logic to detect and prioritize the best-available price column whenever ambiguous labels appear due to naming convention differences between data sources.

2.2.2 Handling Missing Values

The handling of missing data is performed conservatively:

- For financial instruments, missing prices are typically not interpolated to avoid fabricating market behavior. Instead, only forward-fill is applied for selected variables (e.g., macro indicators).
- Entire columns are dropped if they are fully empty across the historical window.
- Temporal gaps are not backfilled — instead, model logic accounts for them naturally (e.g., via rolling windows or masking).

2.2.3 Feature Extraction and Signal Construction

The cleaning process of data leads to the computation of engineered features which represent the dynamic properties of both asset returns and volatility. These engineered features function as model inputs and risk indicators for targets or variable conditioning. The pipeline automatically:

- Selects relevant price series for feature derivation
- The pipeline performs computations of log returns together with percentage changes and evaluates relative performance indicators
- The pipeline generates rolling volatility estimates over specified time periods like 30-day or 90-day windows.
- Extracts technical indicators, such as:
    - Relative Strength Index (RSI)
    - Bollinger Bands

- Momentum and mean-reversion flags (where applicable)

The features/ directory stores all processed and engineered data which is categorized by asset type to support modeling workflow integration.

Raw financial and macroeconomic time series data is transformed into a coherent dataset rich in information throughout the preprocessing and feature engineering stages. The method combines strong stability with adaptable functionality to preserve realistic representations which supports sophisticated risk modeling approaches without requiring extensive data cleaning.

2.3 Data Validation

The pipeline integrates a specialized stage for data validation alongside cleaning and transformation processes. The validation stage in this phase actively searches for discrepancies or hidden issues within datasets that expand to extensive asset collections. We aim to implement quality control methods in a systematic and transparent manner before moving to the modeling stage.

The validation stage of the pipeline goes beyond just format verification. The pipeline performs statistical tests together with outlier detection and distributional checks to detect problematic data early which prevents training bias and risk estimate invalidation.

2.3.1 Data Validation

The threshold system allows users to establish "acceptable" data criteria for each asset through configuration settings. The YAML configuration file determines the thresholds which control important checks such as:

- Minimum number of valid observations
- Maximum allowed proportion of missing values
- Expected volatility bounds
- The system uses value range filters for essential metrics including returns and volumes.

These criteria are enforced during validation runs. Features and assets not meeting standards get logged before exclusion from additional processing steps.

2.3.2 Detection of Distributional Anomalies

The validation module uses basic univariate statistical diagnostics to detect hidden quality problems:

- Skewness and kurtosis computation
- The validation module assesses normality through methods such as z-score evaluation or Shapiro-Wilk testing.
- The validation module identifies series and columns that show flat behavior because they have very little variance.

The diagnostics process reveals assets that have no activity or are improperly documented and those which lack structural information. An extended period of zero variance in a volatility time series suggests possible stale pricing or data corruption.

2.3.3 Logging and Traceability

All validation outcomes are logged, including:

- Assets successfully validated
- Assets excluded and reasons for rejection
- Statistical metrics and thresholds evaluated

The system generates a summary report as either a JSON or CSV file which is then saved in the validation_data/ directory. The produced output ensures complete traceability and serves as a reusable resource for audit purposes as well as reproducibility and error analysis.

2.3.4 Asset Selection via Validation Filtering

Data validation serves as a quality assurance tool and acts as a systematic filter for selecting assets. The modeling universe includes assets that satisfy all validation criteria rather than those chosen from an arbitrarily predefined instrument list. The training and testing process of models focuses solely on series that fulfill all validation requirements.

- Contain sufficient historical depth
- Exhibit valid and informative statistical properties
- The dataset excludes any series that contain excessive missing values or structural artifacts.

As a result, the dataset is self-curated: The asset universe emerges from empirical analysis using objective statistical testing instead of manual selection. The chosen design approach enhances model reliability and prevents dependence on data sources that are incomplete or unstable.

The validation module serves as the ultimate checkpoint between data engineering activities and modeling tasks. The validation module maintains both structural and statistical integrity by filtering unreliable data while generating transparent logs that foster reproducibility. The pipeline transforms data quality into an active design principle through the dual role of validation as both a control and a selection mechanism.

2.4 Integration of Macroeconomic and Market Data

The data preparation pipeline concludes with the merging of independently processed macroeconomic indicators and financial instruments into one combined dataset. Through this integration process models gain access to learn from both price-based features and wider macro-financial context variables that influence asset evolution.

The objective is to develop a dataset which synchronizes each asset's time series with pertinent macroeconomic indicators while maintaining temporal fidelity and avoiding forced matches between datasets with varying structures.

2.4.1 Temporal Join Strategy

We perform the merging process by executing a left join between every asset's time series and cleaned macroeconomic data. The merging process ensures each date point in an asset's historical timeline gets paired with the latest macro indicators while maintaining the asset's original trading schedule. Key properties of this approach include:

- The process only utilizes forward or backward filling methods which were already applied during macro preprocessing.
- The asset's trading calendar remains intact together with any gaps or non-trading days.
- Macro series require forward-filled values prior to joining since no interpolation for missing macro values occurs at this stage.

The conservative methodology prevents the introduction of artificial information into asset-level time series data.

2.4.2 Chunk File Processing

The merging process handles large asset datasets by reading and processing each CSV file piece by piece in separate chunks. The chunked strategy eliminates memory overload while enabling scalable processing across multiple assets. The logic proceeds as follows:

1. Load the cleaned macroeconomic dataset once.
2. Process each CSV asset file within the features directory individually.
3. For each asset file:
    - Read the data in manageable chunks.
    - Merge each asset file with the macro dataset using a left join based on the date index.
    - Merge the joined data portions together and store the complete dataset.

The modular design guarantees reliable operation over diverse asset files while enabling future expansions to run processes either in parallel or batch mode.

2.4.3 Output Structure and Storage

The fully joined datasets are stored in a dedicated directory, keeping them separate from raw or intermediate data:

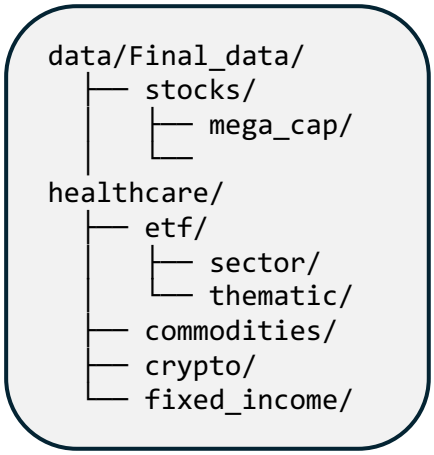


Figure 7 – Dataset directory

Each file contains both:

- Time series features related to the asset (returns, volatility, technical indicators)
- Forward-filled macroeconomic features (normalized and raw values)

This final dataset is ready for model input without further preprocessing.

During the macro–market data integration stage all assets acquire an authentic economic setting. The system maintains market data's temporal resolution while extending its informational value through additional risk factor data. The data pipeline is now complete and ready for risk model training and evaluation.

2.5 Model Implementation

This section describes all the forecasting models developed for this thesis. The evaluation approach compares multiple models through structured methodological family groupings instead of individual assessment. The models fall into different philosophical categories of risk forecasting which range from statistical metrics based on rules to deep learning architectures that incorporate machine learning algorithms.

The system employs modular pipelines to process financial time series data which calculates essential targets and evaluates forecasting performance under normal as well as stress conditions.

2.5.1 Traditional Risk Estimators

The thesis develops a complete collection of traditional risk metrics which function as quantitative benchmarks and supervised learning targets before predictive models are introduced. The feature engineering pipeline performs programmatic calculations of these estimators which enable asset-specific calibration while maintaining temporal flexibility and model comparability.

The feature engineering pipeline computes traditional estimators programmatically through rolling windows and dynamically adjusted parameters. The implementation follows institutional risk management standards while enabling detailed risk profile creation for individual assets.

2.5.1.1 Rolling Volatility

The standard deviation of log-returns creates forecasted volatility across rolling windows of 10, 30, and 60 days. Hybrid models use the stored estimates both as direct targets and conditioning variables:

- Method: Rolling std over log returns
- Adjusted dynamically for trading calendars
- Stored with suffixes: VOL_10d, VOL_30d, VOL_60d

The values serve dual purposes in standalone evaluation while functioning as a comparative baseline for machine learning volatility prediction models.

2.5.1.2 Multi-Horizon Return Targets

To support supervised risk forecasting, forward-looking return series are computed and shifted into the future. Two predictive horizons are considered:

- One-day ahead return**,** stored as RETURN_CALC_t+1
- Twenty-one-day cumulative return, stored as RETURN_CALC_t+21

These variables are used to evaluate forecasting accuracy and are also integrated as targets in quantile regression and LSTM-based models.

2.5.1.3 Value-at-Risk (VaR)

Several Value-at-Risk estimators are computed using different approaches, each representing a specific risk modeling philosophy:

- Parametric Gaussian VaR: based on estimated mean and volatility
- Historical Simulation VaR: using empirical quantiles from rolling returns

- Monte Carlo VaR: when available, from simulated paths
- Extreme Value Theory (EVT)**:** fitted on tail distributions using Peaks Over Threshold
- GARCH-based VaR: derived from volatility forecasts produced by GARCH processes
- Empirical VaR**:** directly computed from observed losses without parametric assumptions

Each VaR type is computed over two horizons (1-day and 21-day) and at two confidence levels (95% and 99%), yielding a structured matrix of risk estimates.

Columns follow a structured naming convention such as:

```
VaR_Param_1d_95, VaR_Hist_21d_99, VAR_EVT_1d_95, etc.
```

2.5.1.4 Conditional VaR (Expected Shortfall)

A specific Conditional VaR value is determined for every VaR series. CVaR represents the mean loss beyond the VaR threshold while maintaining the status of a coherent risk measure. These metrics provide key insights into how models perform under extreme risk conditions like the 2008 financial crisis and the COVID-19 economic shock.

CVaR targets are stored as:

```
CVaR_Param_1d_95, CVaR_Hist_21d_99, CVAR_EVT_1d_95, etc.
```

2.5.1.5 Quantile Loss Scoring

Quantile loss functions between predicted VaR and realized returns measure the calibration quality of each risk estimator within your pipeline:

$$\mathcal{L}_{\alpha}\left(r_t, VaR_{t,\alpha}\right) = (\alpha - \mathbb{I}_{r_t<VaR})(r_t - VaR)$$

This loss is computed across:

- All VaR models
- Both horizons (1d, 21d)
- Both confidence levels (95%, 99%)

This loss is computed across all models, horizons, and confidence levels. It allows for nuanced performance assessment beyond binary exception counting. The use of quantile loss supports direct comparison with machine learning models optimized using similar loss functions.

2.5.1.6 Drawdown and Path-Based Metrics

The calculation of complementary metrics helps to measure path-dependent risk:

- Maximum Drawdown (MDD): calculated over fixed rolling periods
- Calmar Ratio: annualized return over drawdown
- Volatility-adjusted return: Volatility-adjusted return stands in as a surrogate for the Sharpe ratio in situations lacking risk-free data.

Metrics exist only to provide evaluation references and visualization tools and are not modeled as targets.

2.5.1.7 Observed Risk Realizations (_OBS Variables)

Every predicted risk metric undergoes rigorous evaluation by pairing it with its corresponding realized value which is calculated ex post. The dataset labels observed values with a _OBS suffix which incorporates the following metrics for example:

- VOL_OBS_30d: realized historical volatility measured over the evaluation period

The dataset includes both predicted and observed values which enables uniform evaluation for statistical models as well as machine learning and hybrid approaches.

The dual structure configuration allows the pipeline to directly generate quantile loss scores and perform exception analysis without the need for external reconciliation processes.

The traditional estimators created within this section establish a solid benchmark that supports evaluation and integration alongside more complex forecasting algorithms. These values are not predetermined reference points but instead result from dynamic calculations over multiple forecasting periods and various assumptions to ensure strict adherence to real-world results. Machine learning models are benchmarked using well-defined measures from professional finance which reflect actual risk rather than abstract targets.

2.5.2 Machine Learning – Tree-Based Models

Decision tree-based machine learning models are well-known for their strong approximation capabilities of non-linear and interactive variable relationships which renders them highly effective for analyzing financial time series data. In this thesis, three types of tree-based regressors are implemented and evaluated: The study evaluates three tree-based regression models including Random Forest with Gradient Boosting and Extreme Gradient Boosting (XGBoost).

2.5.2.1 Problem Definition and Predictive Objectives

This thesis utilizes machine learning models which enable the prediction of several forward-looking risk metrics while surpassing traditional Value-at-Risk (VaR) analysis. The models use supervised regression algorithms to determine specific risk-sensitive statistics for future time periods by analyzing historical and present features at time t:

The predictive objectives are threefold:

- Quantile returns: The model provides probabilistic loss estimates at predefined confidence levels which match the criteria of Value-at-R Risk (VaR).
- Tail conditional means: The approximations of Conditional Value-at-Risk (CVaR) represent the expected loss that occurs after surpassing the VaR threshold.
- Path-dependent or realized metrics, including:
    - Realized volatility, computed over future rolling windows,
    - Maximum drawdown represents the greatest cumulative loss from peak to trough within a predetermined future period.

Each target is computed in advance using time-shifted rolling windows and stored as a separate column in the enriched dataset. Models are trained to reproduce these values using only features observable at

the current time step. Importantly, no model forecasts multiple targets simultaneously: each is trained and evaluated independently on a specific risk measure.

The horizons are consistent across targets:

- Short-term (1 day) for high-frequency risk estimation;
- The 21-day duration serves as the medium-term range for both portfolio management and stress testing.

VaR and CVaR models undergo calibration to match two distinct confidence levels which are 95% and 99%. Risk quantification methods and regulatory standards of institutions share the same confidence level benchmarks.

The prediction task adopts a forward regression approach according to this specific formulation:

$$\hat{y}_{t+h} = f(X_t)$$

Where:

- y_{t+h} is the realized value of a specific risk metric (e.g., volatility, drawdown, VaR) at time $t + h$,
- X_t is the feature vector composed of lagged returns, technical indicators, and macroeconomic variables available at time t,
- $h \in \{1\,, 21\}$ is the prediction horizon in trading days.

This framework supports a comprehensive evaluation: each model's output can be directly compared to both traditional statistical estimators and the realized values observed ex post. The diversity of targets ensures that the modeling effort captures not only distributional quantiles but also amplitude, timing, and path-dependency — all essential components of financial risk.

2.5.2.2 Feature Set Construction

The effectiveness of supervised learning models relies heavily on both the quality and variety of their input features. This thesis develops a feature set that presents financial risk through multiple dimensions by integrating price-based indicators with distributional moments and macroeconomic data. The selected features aim to represent short-term price movements and trend patterns along with behaviors dependent on market regimes to affect asset-specific risk measurements.

The feature space consists of shared financial signals together with technical variables specific to assets and macroeconomic indicators. The different groups provide unique contributions to how models generalize through various asset classes and market conditions.

2.5.2.2.1 Common Price-Derived Features

Core return and volatility measures arise from analyzing historical price movements of the asset.

- The dataset utilizes smoothed lagged log returns alongside raw percentage returns as core financial metrics.
- The rolling standard deviation and variance metrics are calculated based on concise 5-day intervals.
- A rolling 20-day window analysis of skewness and kurtosis reveals information about both distributional asymmetry and fat tails.

The features capture fundamental aspects of the return process which make them useful for predicting tail quantiles and targets based on volatility.

2.5.2.2.2 Technical Indicators

A series of technical analysis variables is integrated to capture momentum, trend-following behavior, and mean-reversion signals. Examples include:

- Relative Strength Index (RSI),
- Moving Average Convergence Divergence (MACD) and signal line,
- Simple moving averages (SMA) over 20, 50, and 200 days,
- Maximum drawdown over trailing windows**,** used both as a predictor and a target.

The indicators enable effective detection of regime shifts as well as volatility clustering and market trend overextension.

2.5.2.2.3 Macroeconomic and External Variables

During preprocessing top-down signals are added to the asset dataset by merging a chosen group of macroeconomic indicators. These include:

- Interest rates (e.g., 10-year US Treasury yield),
- Inflation proxies (e.g., Consumer Price Index values),
- Survey-based indicators include manufacturing and services PMI.
- The VIX index serves as an appropriate volatility measure for certain types of assets.

Daily updates of these variables through forward-fill techniques help place asset performance in the context of overall economic conditions. The inclusion of these variables occurs systematically to facilitate cross-sectional learning even though their predictive value differs across asset types and time horizons.

2.5.2.2.4 Asset-Type Features

The modeling pipeline uses logical rules to dynamically add extra features based on the asset class such as equities, ETFs, commodities, cryptocurrencies. For example:

- ETFs require volume-weighted average price (VWAP) along with their net asset value (NAV).
- Volume and trade count metrics for commodities,
- Momentum-based moving averages for crypto assets,
- Bid-ask spreads for fixed income instruments.

The feature injection process prevents equity-specific biases and enables models to detect risk factors specific to different market structures.

2.5.2.2.5 Data Handling

Prior to model training, all features are:

- The data underwent conversion into numerical formats followed by cleaning procedures to eliminate any infinite values.
- Imputed using mean replacement where needed,
- Standardized when required by the model architecture.

The system automatically discards columns that either contain too many missing values or demonstrate insufficient variance. Before training the model the final feature set for each asset undergoes both filtering and validation processes.

The multiple levels of feature construction allow models to acquire knowledge from market microstructure aspects along with macro-level risk factors. The detailed input space allows for more precise analysis of model behavior during later evaluations.

2.5.2.3 Model Architecture and Training Protocol

The thesis incorporates Random Forest, Gradient Boosting, and XGBoost models into a unified training pipeline. Despite their differences in learning algorithms, they share a unified structure that maintains consistent performance measures and enables direct comparison between different model types. The training protocol includes stages for data preparation along with feature selection before optimizing hyperparameters and validating models using time-consistent methods.

2.5.2.3.1 Model Structure and Output

All models are implemented as regressors, with the specific objective of learning to predict a single numeric target per run. Depending on the chosen configuration, this target can be:

- A quantile of future return (e.g., for VaR at 95% or 99% confidence),
- A realized volatility estimate over a forward window (e.g., 10 or 21 days),
- A maximum drawdown observed over the same forecast horizon,
- Or a CVaR approximation, derived as the average of extreme outcomes beyond the VaR level.

The models do not attempt multi-output prediction. Each is calibrated individually to forecast a specific risk indicator over a fixed horizon, with its own training and evaluation dataset.

2.5.2.3.2 Time-Aware Train/Test Splitting

The division of data occurs chronologically to avoid information leakage while maintaining the temporal structure.

- The training set includes only past data points as they relate to the prediction target.
- The test set mimics real-time forecasting scenarios by strictly using out-of-sample data.

For a standard data split based on asset characteristics and data availability one should leave the final 20% of the timeline for testing purposes. For sophisticated implementation designs a rolling-window validation approach becomes applicable but remains an aspect of the evaluation stage.

The pipeline incorporates this splitting method which enables performance evaluation of each model based on realistic forecasting limitations.

2.5.2.3.3 Feature Selection via Recursive Elimination

The training process can optionally use Recursive Feature Elimination (RFE) to reduce dimensionality while enhancing interpretability. RFE determines the importance of each variable within the model and systematically eliminates the least important variables until the target number of features is reached.

Asset classes with sparse coverage and high-dimensional datasets experience considerable benefits when this step is applied to the training models. The method ensures each model operates on its own tailored subset of the most informative features from the full feature space.

2.5.2.3.4 Hyperparameter Tuning with Randomized Search

RandomizedSearchCV tunes each model as a powerful technique to discover the best hyperparameter settings without needing exhaustive grid search. The search space includes:

- Number of trees or boosting rounds,
- Maximum tree depth,
- Minimum samples per split or leaf,
- Subsample ratios,
- Learning rate (for boosting models),
- Regularization penalties (especially for XGBoost).

Each target type has a customized optimization objective during model evaluation. Quantile loss functions as the evaluation metric for quantile regression models to focus tuning efforts on achieving calibration improvements instead of fitting the model. RandomizedSearchCV uses cross-validation splits that respect time-series data to maintain proper chronological validation. The search process limits the number of iterations to manage computational complexity while remaining adequate for finding performant configurations.

2.5.2.3.5 Model Comparability

Meaningful benchmarking requires configuring all models to display equivalent levels of complexity and learning potential. For instance:

- The training for all models employs approximately 500 to 1000 trees.
- Performance differences in models stem from their structure because the calibration strategy maintains balance between bias and variance through tree depth and regularization constraints.

2.5.2.3.6 Outputs Artifacts and Model Documentation

The documentation of each model run consists of standardized outputs that are systematically preserved for evaluation, interpretability, and replication needs. The generated outputs correspond to each target and prediction horizon, and they follow an arrangement that facilitates both quantitative analysis and model transparency.

Two primary categories of files are generated:

1. Each model produces two CSV files for Forecast Performance Records:

- The file results_<model>_<target>.csv provides summary performance metrics from the test set including mean absolute error and quantile loss. Model comparisons in later empirical sections rely on these result outputs.
- detailed_<model>_<target>.csv stores granular prediction outputs, aligned temporally. The CSV file facilitates the analysis of time-series forecast errors and exceptions while evaluating performance in stress windows.

The CSV files provide a clear and modular method to review both predictions and their evaluation metrics.

2. Model Explainability Files (and LIME)

For all tree-based models (Random Forest, Gradient Boosting, XGBoost), two forms of explainability are implemented:

- SHAP (SHapley Additive exPlanations) provides both global and local interpretability, based on game-theoretic attribution of feature contributions. SHAP is especially well suited for models with hierarchical or nonlinear interactions.
- LIME (Local Interpretable Model-Agnostic Explanations) complements SHAP by offering case-specific, perturbed approximations of model output around a focal prediction. This is particularly useful for validating consistency in explanations across methods.

Both SHAP and LIME files are stored per model and target.

For LSTM-based architectures, only LIME is applied. Due to the sequential and deep nature of these models, SHAP explanations are not available in their current implementation.

2.5.2.4 Model-Specific Implementation Details

2.5.2.4.1 Random Forest

This thesis implements a Random Forest model through ensemble methodology where each decision tree learns from independently bootstrapped samples of the dataset. The model generates the final prediction by combining tree outputs through an averaging process.

The scikit-learn library's RandomForestRegressor class creates the model in this implementation. The model treats each target—quantile return for VaR estimation, forward-looking volatility, and drawdown measure—as a separate modeling entity.

The model uses a high number of estimators, usually 500, to stabilize predictions and reduce variance. The model does not impose any depth restriction on individual trees so they can optimally partition the feature space when necessary. Regularization occurs implicitly through the minimum sample leaf constraints and feature subsampling which are optimized using randomized search according to the procedures described in the previous section.

The Random Forest model establishes a powerful standard for all risk assessment metrics. Despite its limitations in handling time dynamics and intertemporal dependencies this model remains significant for benchmarking because it captures nonlinear interactions and maintains robustness against noise which proves especially useful in asset classes with stable historical patterns or stationary nonlinear feature–target relationships.

This model produces all output files which encompass detailed predictions along with aggregate results and SHAP and LIME explanations following the earlier described protocol.

2.5.2.4.2 Gradient Boosting

The thesis employs Gradient Boosting as per the initial framework where regression trees undergo sequential training to address the errors from earlier iterations. Through this additive approach to model-building learners can adjust to intricate nonlinear patterns while keeping variance levels relatively low.

We used the GradientBoostingRegressor class from the scikit-learn library to implement the model. During each boosting run the model executes a predetermined number of stages which usually consists of 500 iterations while optimizing the learning rate through randomized search. Gradient Boosting builds

trees sequentially where each new tree aims to correct the prediction errors made by the previous ensemble of trees.

What distinguishes the implementation of Gradient Boosting in this work is its close integration with the structured output framework: The pipeline systemically integrates contextual variables and reference targets with predictions during each model run. The function train_and_evaluate_gb serves both as a wrapper for training logic and as a handler for output post-processing tasks including alignment with true realized volatility and drawdown values.

The model adjusts its loss function and evaluation approach to support various target types including quantile-based metrics, volatility measurements, and drawdown indicators. True quantile optimization remains absent in this GradientBoostingRegressor version because of its inherent limitations. Post hoc evaluation of tail-sensitive targets uses custom metrics like quantile loss or target-specific error decomposition.

SHAP and LIME explanations deliver interpretability through the same approach as with other models while all results are saved into structured .csv files for future comparison.

Gradient Boosting operates within this framework as a compromise between Random Forest's broad applicability and XGBoost's specialized adaptability. In the next chapter we will carry out comparative analysis of its performance across multiple risk targets.

2.5.2.4.3 XGBoost

XGBoost (Extreme Gradient Boosting) is implemented in this thesis as a high-capacity tree-based learner, designed to capture nonlinear relationships and interaction effects in structured financial data. It builds upon the traditional boosting framework by incorporating highly optimized tree construction and efficient memory usage, making it particularly suitable for large-scale empirical studies.

The model is implemented using the XGBRegressor class from the xgboost Python package. In contrast to the other models in this family, XGBoost does not natively support quantile regression. In this implementation, quantile estimation is handled through the model’s output layer and evaluated via quantile-aware loss metrics computed post hoc.

Each XGBoost instance is trained independently for a given target—be it a VaR quantile, a volatility forecast, a drawdown measure, or an approximation of CVaR—over a fixed horizon. The model is configured with a relatively high number of boosting rounds (up to 1000), combined with a low learning rate, which ensures gradual convergence and stable training. However, no early stopping criterion is currently applied in the pipeline, training proceeds for the full number of iterations.

Hyperparameters such as tree depth, subsample ratio, and number of estimators are tuned via randomized search, but regularization terms such as L1 (reg_alpha) and L2 (reg_lambda) penalties are not yet utilized in this version. While XGBoost supports these controls, their integration is left for future work.

As with other models, XGBoost outputs are aligned with observed risk metrics using a dedicated enrichment function, and all results are saved in both aggregate and detailed formats. SHAP and LIME explanations are computed systematically to support post-model interpretability.

In this thesis, XGBoost serves as the most flexible and expressive tree-based method implemented. Its performance across multiple targets offers insight into how boosting methods can be applied to diverse forms of financial risk, even in the absence of explicitly tailored objective functions.

The three tree-based models implemented in this thesis—Random Forest, Gradient Boosting, and XGBoost—constitute a versatile and interpretable class of machine learning algorithms for risk forecasting. Each model is calibrated to predict forward-looking risk metrics such as Value-at-Risk, Conditional Value-at-Risk, realized volatility, and maximum drawdown, across both short- and medium-term horizons.

Despite their shared structure, the models differ in learning philosophy: Random Forest aggregates fully grown, decorrelated trees; Gradient Boosting builds a sequential ensemble by correcting prediction errors; and XGBoost optimizes tree construction for speed and depth while retaining expressive power.

All models are trained using time-consistent splits, evaluated on quantile-aware losses, and supported by both SHAP and LIME interpretability. Their outputs are saved systematically for quantitative and qualitative comparison in the empirical chapters that follow.

Taken together, these models offer a robust benchmark for evaluating the potential of more complex architectures—namely, deep learning and hybrid approaches—addressed in the next sections.

2.5.3 Deep Learning – LSTM Models

2.5.3.1 LSTM Architecture and Implementation

This thesis presents an LSTM (Long Short-Term Memory) model that analyzes financial data temporal dependencies to predict future risk measures. The LSTM architecture enables sequential learning which proves valuable for markets with path-dependent behavior and volatility clustering unlike tree-based models which process data points individually.

The model is built with TensorFlow and then encapsulated using the scikeras.KerasRegressor API so it fits into the wider pipeline architecture. Each target variable including realized volatility, maximum drawdown, Conditional Value-at-Risk (CVaR), and quantiles used for VaR estimation receives separate training sessions which maintain the same short- and medium-term horizons established in earlier models.

The network architecture consists of:

- An Input layer accepting sequences of fixed length (windowed observations of features),
- A single LSTM layer that processes temporal patterns in the input sequence,
- A fully connected Dense output layer producing a single risk forecast for each sequence.

The feature inputs are preprocessed and aligned similarly to the tree-based models but are reshaped into three-dimensional sequences of the form, allowing the network to process patterns over time. Sequence length and batch size are configurable parameters optimized during training.

The model is trained using the Adam optimizer and a mean squared error loss function for continuous targets. Early stopping, dropout, or regularization mechanisms are not currently implemented in the codebase, although these can be integrated in future extensions to prevent overfitting.

Unlike other models in the framework, the LSTM output is interpreted through LIME (Local Interpretable Model-Agnostic Explanations**)** only. Due to the complexity and non-linearity of recurrent networks, SHAP is not used here, as it lacks a reliable adaptation for sequential architectures in this implementation.

All LSTM results—predictions, ground truth alignment, and explanation vectors—are exported into structured CSV files following the same naming convention as other models. This ensures comparability in downstream evaluation and interpretability analysis.

2.5.3.2 Training Procedure and Temporal Handling

The LSTM model is trained using a time-consistent approach which preserves the sequential nature of financial data. A chronological train_test_split method divides data into training and test sets to prevent future time point data from affecting the training process.

Before model fitting takes place, input features are transformed into three-dimensional arrays. The transformation process provides the means to capture temporal patterns within overlapping observation windows. The sequence length remains configurable but consistent throughout the training process.

The training procedure operates through the KerasRegressor wrapper by employing the conventional.fit() method. The model training concludes after a predetermined number of epochs without employing early stopping techniques or adaptive regularization methods. This version does not support dropout layers and L1/L2 penalties but plans to include them in upcoming work to improve the model's ability to generalize.

The model employs mean squared error as its loss function because it suits the continuous characteristics of the target variables. Prediction targets include realized volatility, drawdown, CVaR, and return quantiles for VaR estimation, all treated independently.

After training, the model outputs are written to both detailed and summary CSV files, aligned with observed values. LIME explanations are also generated and stored per instance. The method achieves identical interpretability and output traceability capabilities that tree-based models deliver. The LSTM model developed in this thesis offers a deep learning solution that competes with traditional risk forecasting techniques.

The sequential structure of this model enables it to capture time-dependent behaviors which tree-based models usually miss thus making it ideal for analyzing volatility clustering and persistent trends.

Despite its simplicity—relying on a single LSTM layer without explicit regularization—the model is capable of learning multiple risk metrics from temporal patterns in market and macroeconomic data. While its training protocol is less flexible than that of boosting methods, its ability to capture dynamics over time offers a complementary perspective within the modeling framework.

The use of LIME for local interpretability ensures that the model remains transparent despite its complexity. The predictive results and their comparison with other architectures are examined in the empirical analysis to follow.

2.5.4 Hybrid Models – GARCH-LSTM

2.5.5 Preliminary Volatility Estimation with GARCH

Before training the hybrid LSTM model, a preliminary volatility forecasting step is performed using a traditional statistical model: GARCH (Generalized Autoregressive Conditional Heteroskedasticity). The purpose of this step is to serve as a preprocessing component to produce forward-looking volatility signals rather than to function as a standalone prediction tool.

This research implements a rolling GARCH estimation framework. The model undergoes calibration through a rolling window of historical returns to generate a one-step-ahead conditional volatility forecast

for both each asset and date. The model generates estimates which are then matched with appropriate market dates to integrate into the feature set utilized by the LSTM.

The GARCH forecasts which are computed separately from the deep learning model function solely as input features. The GARCH forecasts do not serve as final outputs and do not undergo comparison with realized volatility targets. The inclusion of GARCH model forecasts relies on their proven capability to detect short-term volatility patterns especially when market conditions become unstable.

The system utilizes the modular function train_rolling_garch to automate calibration processes within rolling windows. The method maintains performance stability across various market environments and prevents models from becoming too specialized for any specific market condition.

The hybrid model merges parametric forecasting methods with neural network structures to embed structured volatility signals within a comprehensive data-driven learning framework. The GARCH-based preprocessing method functions as a connecting element between conventional econometric techniques and advanced deep learning systems.

2.5.5.1 Hybrid Architecture and Feature Enrichment

This thesis presents a hybrid GARCH-LSTM model that merges statistical methods with deep learning algorithms to predict financial risk. The conventional LSTM architecture is expanded through the addition of volatility forecasts produced ex ante by a GARCH model to its input space. The goal is to leverage the strengths of both approaches: This hybrid model utilizes GARCH for modeling conditional heteroskedasticity in returns while LSTM networks learn temporal patterns in nonlinear systems.

In this implementation, the GARCH component is not a predictive model in itself, but rather an input feature generator. For each asset, a time series of one-step-ahead forecasted volatility is computed using a separate GARCH pipeline and then merged into the feature matrix used to train the LSTM. This additional variable serves as an explicit signal of expected market turbulence, which the LSTM can incorporate into its sequential modeling of future risk.

The architecture itself remains unchanged from the standard LSTM described earlier. It includes:

- One LSTM layer processing a sequence of lagged observations,
- A dense output layer returning a single risk estimate per sequence.

What distinguishes this hybrid model is its input: the feature vector includes not only traditional price- and macro-derived indicators, but also the forward volatility forecasts from GARCH. The temporal alignment and merging are handled directly in the hybrid pipeline, which prepares the dataset before calling the common LSTM training function.

This approach allows the model to benefit from the local precision of GARCH for short-term variance estimation, while still learning more complex patterns from the broader input sequence. Importantly, the GARCH forecasts are treated purely as features — they are not targets, and the GARCH model is not retrained jointly with the LSTM.

LIME explanations are generated for the hybrid model as with the standard LSTM. No SHAP values are computed, as the recurrent structure remains incompatible with the tree-based SHAP methodology in its current implementation.

All outputs — including predictions, LIME explanations, and detailed error traces — are exported in the same format as other models, allowing for direct comparison across the full modeling spectrum.

The GARCH-LSTM hybrid functions as a bridge between traditional volatility modeling techniques and modern sequential learning while providing a flexible framework to capture both structured and learned elements of financial risk.

2.6 Post-Training Backtesting and Risk Evaluation

Pointwise accuracy alone does not suffice for validating predictive performance in risk modeling. The model must show statistical coherence while maintaining economic realism and demonstrate how it performs under market stress conditions. A unified post-training backtesting protocol evaluates all models in this thesis across traditional architectures as well as machine learning and hybrid architectures.

We conduct evaluations with the standardized output files produced during each model's training phase. The protocol features quantitative metrics and formal statistical tests to guarantee both methodological consistency and comparability.

2.6.1 Evaluation Metrics for VaR-Based Models

The quantile loss function creates a differentiable asymmetric metric for prediction precision against a specific target quantile. The framework imposes penalties for both risk underestimation and overestimation and maintains the necessary asymmetry required for tail-risk forecasting.

1. Quantile loss functions apply to models that predict return distribution quantiles including Value at Risk. This loss penalizes asymmetric errors relative to the quantile threshold and provides a smooth, differentiable measure of model calibration.
2. Violation counts and empirical Coverage

Each model's prediction series is compared to the realized returns to compute the number and frequency of violations—i.e., instances where the observed return breaches the predicted VaR level.

The observed violation rate undergoes comparison with the anticipated confidence level to analyze model performance (such as comparing a 5% violation rate against a 95% VaR confidence level). The method permits intuitive evaluation of model conservativeness and aggressiveness without the need for formal testing procedures.

3. The statistical validity of VaR forecasts undergoes rigorous assessment through the implementation of two standard likelihood-ratio tests:

- Kupiec Test: This test measures unconditional coverage by examining the discrepancies between actual and predicted violation frequencies.
- Christoffersen Test: This test determines whether exceptions occur independently throughout time and identifies any patterns of clustering within those exceptions.

The two tests are implemented throughout the entire data set and specifically within stress-defined time frames to provide complete validation.

2.6.2 Evaluation Metrics for Continous Risk Targets

In backtesting continuous targets like realized volatility and maximum drawdown three standard regression error metrics are employed.

- Mean Absolute Error (MAE): This measure captures the typical size of prediction mistakes without considering their direction.
- Root Mean Squared Error (RMSE): RMSE imposes higher penalty on significant errors which helps detect model performance under high-volatility situations.
- Adjusted R^2: The adjusted R^2 reveals how much variability in the actual metric the model explains while taking feature complexity into account.

Model performance is evaluated through these metrics calculated across complete test horizons and designated crisis intervals to enable thorough comparisons between different risk types and various periods.

For Conditional Value-at-Risk (CVaR), although the target is continuous in nature, it is evaluated exclusively via quantile loss. This choice is consistent with the quantile-based modeling framework used for tail-risk prediction. Metrics such as RMSE or R^2 are not applied to CVaR in this work.

2.6.3 Stress Window Segmentation

The backtesting metrics cover both the full out-of-sample period together with key stress intervals. The analysis considers four main market stress events: the 2008 Global Financial Crisis along with the 2013 Taper Tantrum and COVID-19 crash of 2020 and the 2022 geopolitical volatility increase.

By analyzing different segments, we can determine how stable and responsive models are when faced with extreme market conditions since this information is essential for risk management applications.

2.6.4 Implementation Workflow

Every model uses its own backtesting script to:

- Load post-training predictions.
- Applies metric-specific and statistical evaluations,
- Aggregates results globally and by subperiod,
- Exports standardized outputs for comparison and visualization.

A parallel script uses the same evaluation criteria for both historical and parametric estimators like Historical VaR and EVT-based CVaR which guarantees traditional models receive identical scrutiny.

3 Results and Discussion

The empirical outcomes of backtesting different risk forecasting models from this thesis are presented in this chapter. The study evaluates models spanning traditional VaR estimators to advanced machine learning and hybrid architectures using multiple risk measures including Value-at-Risk and Conditional VaR while considering volatility and drawdown metrics at both 1-day and 21-day forecasting horizons under diverse market conditions such as major crises.

The performance evaluation encompasses both average predictive accuracy and statistical reliability as well as economic interpretability. This analysis depends heavily on visualizations. These visualizations demonstrate how performance outcomes differ between time periods and asset classes as well as among various modeling approaches.

The subsequent subsections integrate metric-based evaluations with interpretative analysis. The analysis prioritizes model performance during stress conditions alongside calibration precision and their generalization capabilities in diverse market situations.

3.1 Performance on VaR forecasting

The Value-at-Risk (VaR) models are first assessed with respect to their violation rates—i.e., the frequency at which actual losses exceed the forecasted VaR thresholds. These rates are evaluated at both the 95% and 99% confidence levels, for one-day and twenty-one-day horizons, over the full sample and distinct crisis periods. A correctly calibrated model should, by definition, exhibit a violation rate close to 5% (for 95% VaR) or 1% (for 99% VaR).

To provide a high-level overview, we begin with an aggregate comparison across all models and periods.

3.1.1.1 *Aggregate Violation Rate comparison*

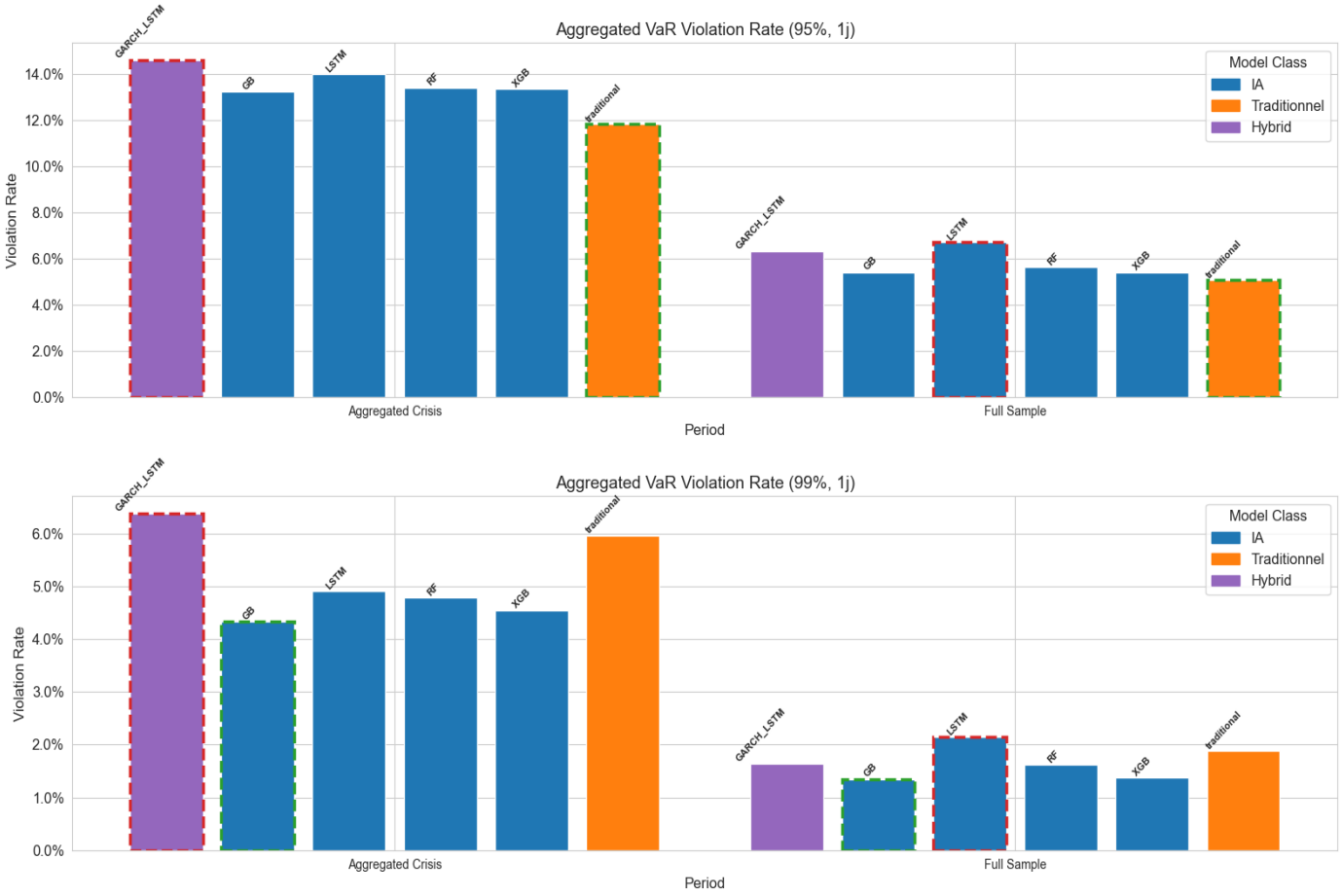


Figure 8- Aggregated VaR Violation Rate at 95% and 99% (1-day horizon) across crisis periods and full sample.

The compilation of results in Figure 8 reveals significant calibration differences across models. Traditional models maintain simple structures but produce violation rates that stay very near theoretical thresholds regardless of whether market conditions are stable or volatile. During crisis periods the traditional VaR model achieves a 12% violation rate at the 95% level but drops to 4.8% across the full sample which approaches the anticipated 5% mark.

Machine learning models display mixed results. The machine learning models Random Forest (RF), XGBoost (XGB) and Gradient Boosting (GB) show strong performance in both testing environments and maintain violation rates near expected levels at the 99% threshold. The LSTM model shows persistent over-violations with rates surpassing 13% during crisis periods at 95%, and close to 5% at the 99% threshold demonstrating poor calibration under extreme market stress conditions.

The hybrid GARCH-LSTM model underperforms even further during crises, with violation rates approaching 15% at the 95% level. This suggests that the inclusion of GARCH-forecasted volatility as a feature may not be sufficient to ensure calibration when exposed to regime shifts.

To assess whether these trends are specific to short-term horizons, we repeat the analysis at a 21-day forecast horizon (Figure 9).

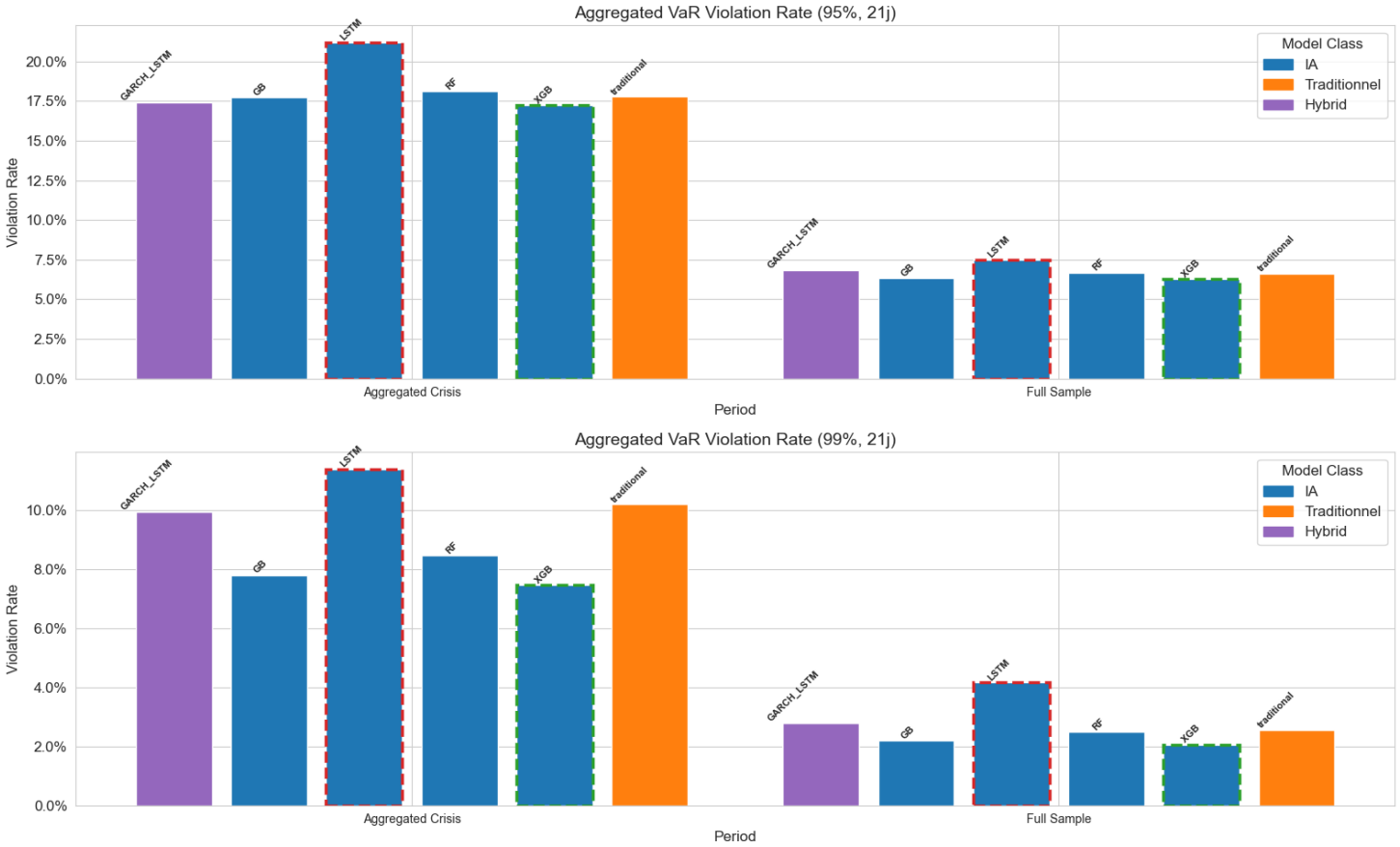


Figure 9 - Aggregated VaR Violation Rate at 95% and 99% (21-day horizon) across crisis periods and full sample.

At the 21-day horizon, the patterns persist. LSTM models are again the most frequently violated during crises, with rates above 20% in the worst cases. Traditional models and XGB tend to stabilize more effectively. Notably, XGBoost achieves the lowest violation rates at 99% confidence (≈7.5%) during crisis, consistent across horizons and with non-crisis period.

These global patterns justify a more granular exploration of model behavior during specific periods of turmoil.

3.1.2 Violation Rate by Crisis Period

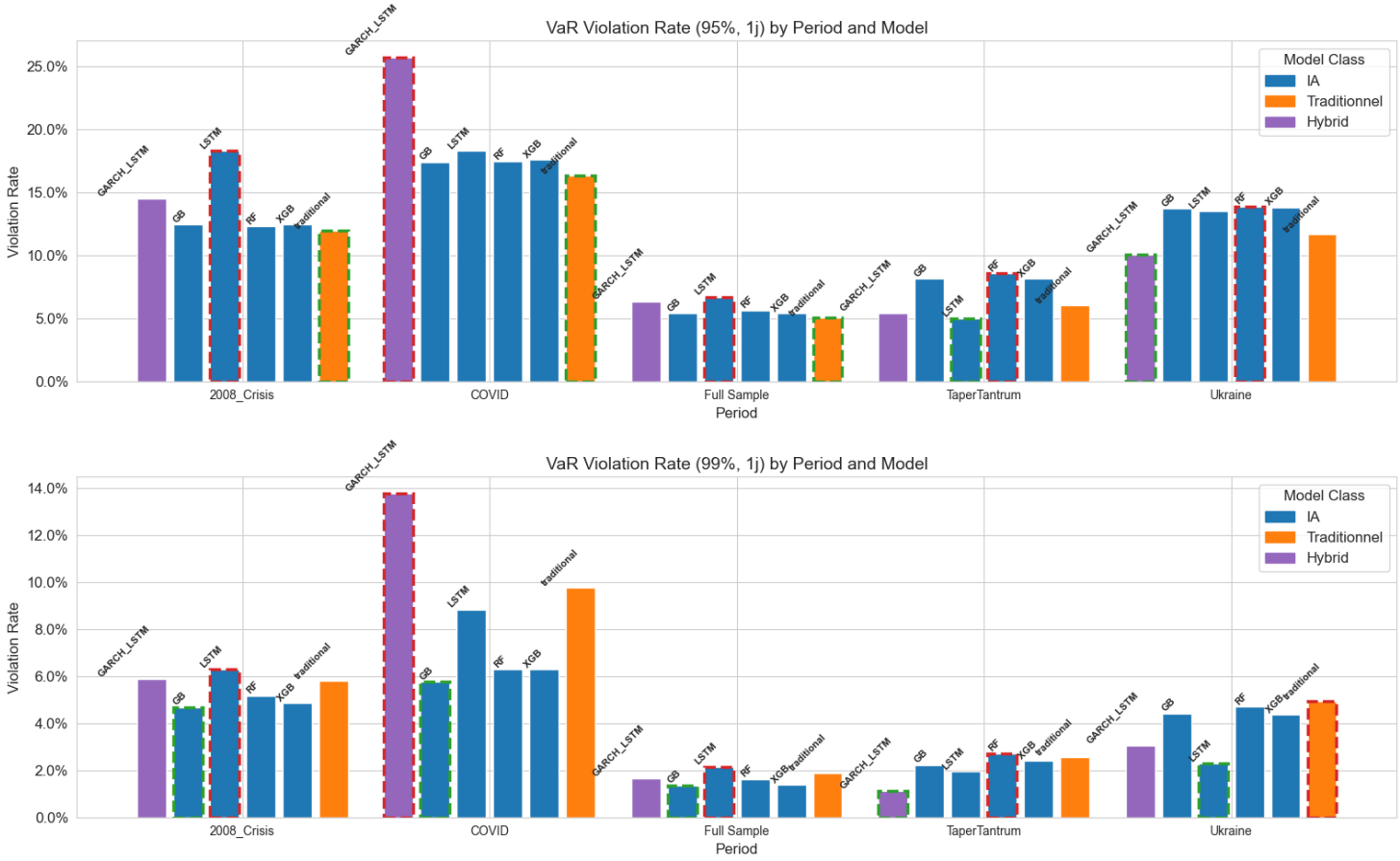


Figure 10 - Violation Rate at 95% and 99% confidence levels (1-day horizon), broken down by individual crisis periods.

The figure 10 disaggregates the results by historical crisis: The analysis covers four historical crises including COVID-19 and the 2008 Financial Crisis as well as the Taper Tantrum that occurred in 2013 and the Ukraine war from 2022. The LSTM model and the RF model show the highest over-violation rates across most regimes with their rates exceeding 15% during both COVID-19 and the 2008 Financial Crisis. The mixed GARCH-LSTM model shows poor performance which suggests it creates compounded errors instead of making corrections.

The GB, RF, and XGB models consistently keep violation rates near the theoretical 5% mark which shows their robust performance throughout multiple crisis regimes. XGBoost demonstrates notable consistency throughout the COVID crisis which featured abrupt spikes in volatility.

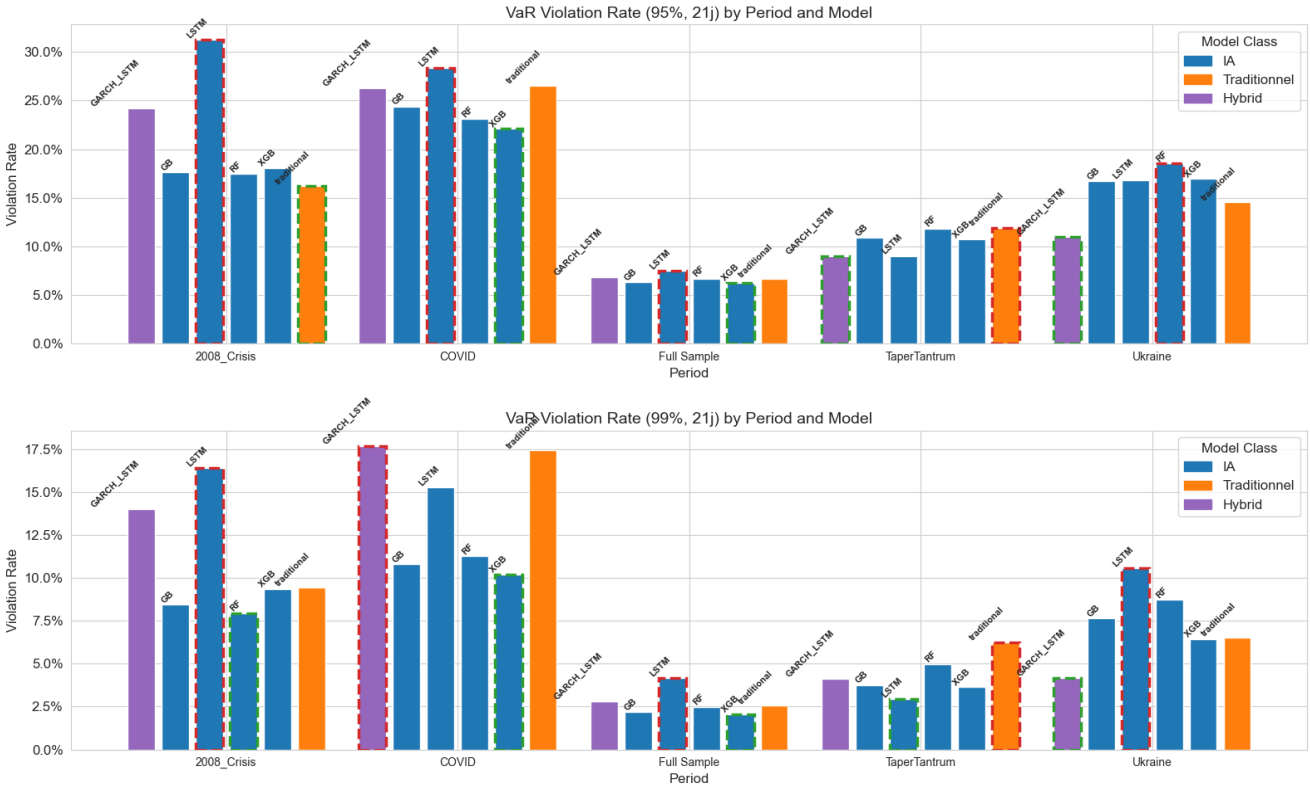


Figure 11 - Violation Rate at 95% and 99% confidence levels (21-day horizon), broken down by individual crisis periods.

Longer horizon forecasts of 21 days (Figure 11) show that most models become less reliable under stress. LSTM reaches over 30% violation during COVID at 95%, a striking failure in risk estimation. Tree-based methods deteriorate more gradually. GARCH-LSTM, expected to stabilize risk through volatility conditioning, surprisingly fails to control long-term VaR in turbulent contexts.

In summary, while long-horizon VaR forecasting remains an open challenge for all models, the contrast in degradation speed and magnitude is revealing. Traditional and ensemble methods show acceptable resilience, while recurrent neural architectures appear more brittle when forecasting over multi-week horizons under stress.

3.1.3 Violation Rate by Asset Class

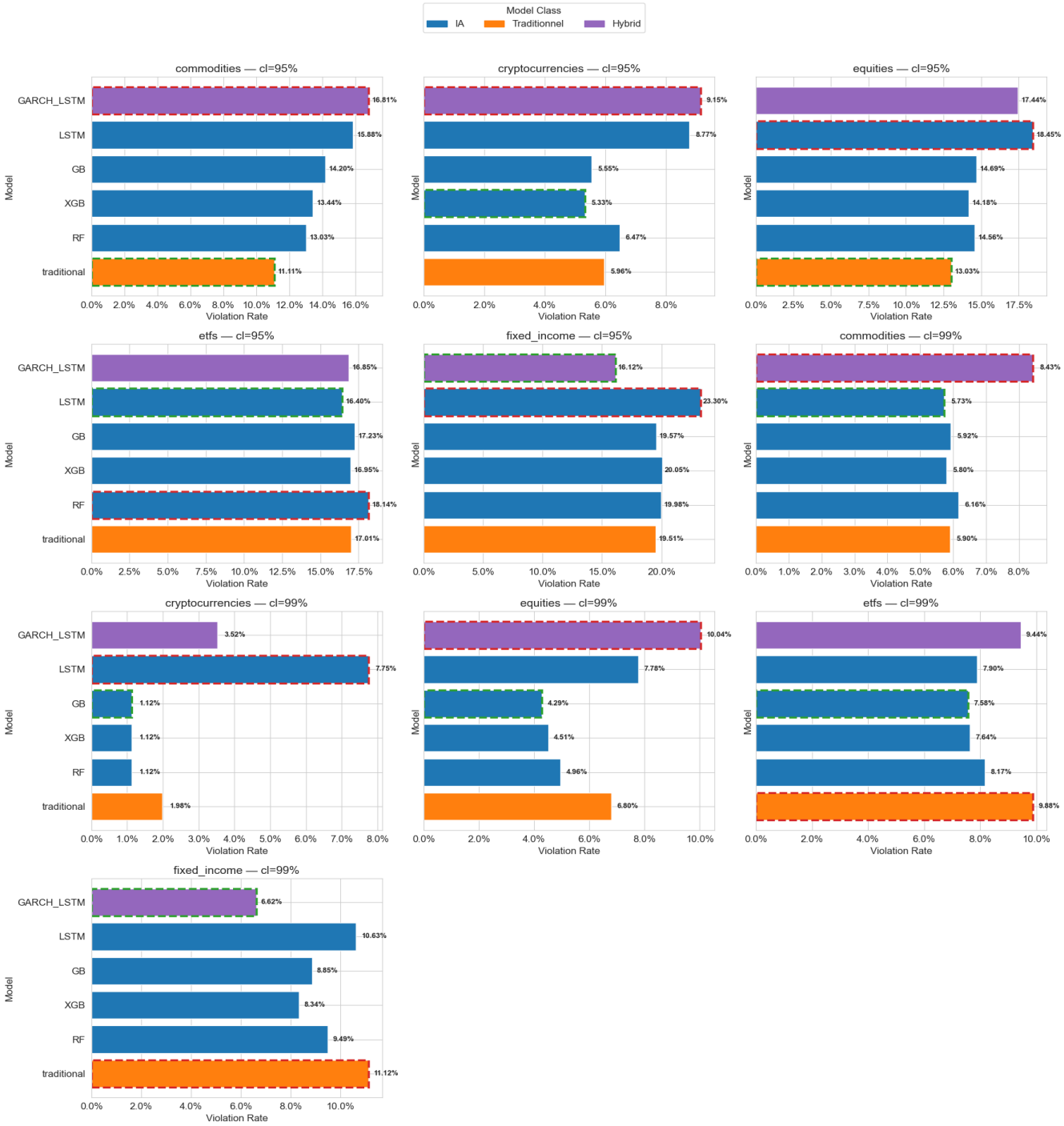


Figure 12 - Violation Rate per asset class during crisis periods

Figure 12's cross-sectional breakdown reveals how model performance varies across asset types during market stress periods. The detailed analysis exposes various blind spots that remain undetected by global performance metrics.

Among all asset types fixed income and equity markets show the highest degree of miscalibration. The violation rates on deep learning models such as LSTM and GARCH-LSTM exceed the 5% benchmark and achieve levels between 13–15%. Existing models prove incapable of responding to sudden market regime changes while also unable to detect systemic risk components such as macroeconomic volatility and liquidity disruptions.

ETFs together with cryptocurrencies show stable calibration outcomes when evaluated with different models. Traditional methods in combination with ensemble-based techniques such as RF and XGB deliver unexpectedly powerful results when applied to these assets. ETFs demonstrate regular patterns of volatility, but crypto returns remain heavy-tailed yet statistically stationary which accounts for their observed behavior.

The combined GARCH-LSTM approach does not deliver the expected stress adaptation improvements for this task. When tested across various asset types the hybrid model performs worse than its individual components demonstrating that model combinations do not reliably produce robustness for diverse asset structures.

This figure underscores a key insight: model reliability is asset-class dependent. To generalize effectively during crises models, require asset-specific calibration together with tailored feature engineering.

3.1.4 Statistical Validation: Kupiec Test

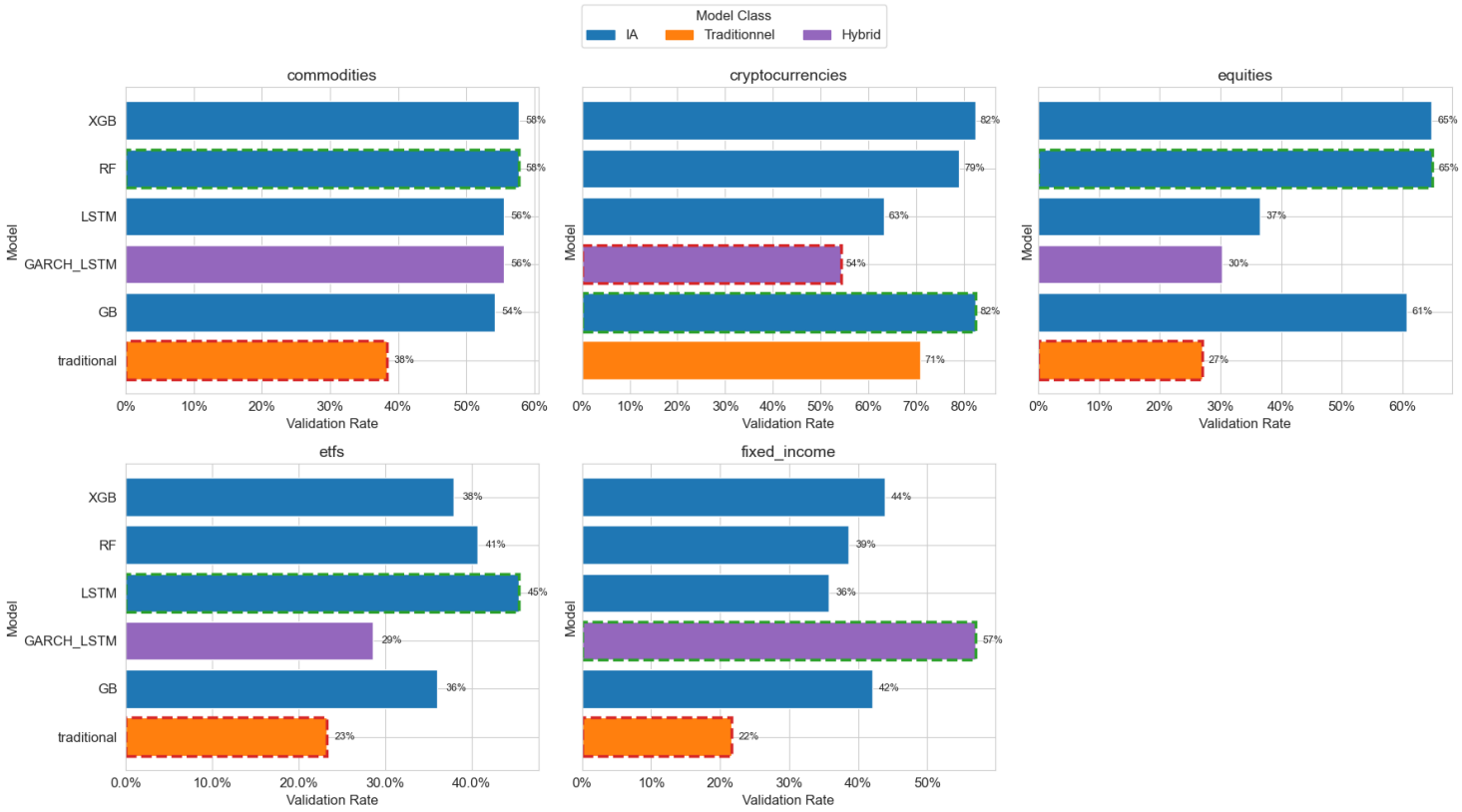


Figure 13 - Kupiec test validation rate by model and asset class during crisis periods

Figure 13 displays the Kupiec test validation rates for six model types and five asset classes which were assessed during crisis periods. The Kupiec test determines statistical alignment between the observed Value at Risk violation frequency and its theoretical limit. Validation rates near 100% demonstrate that the model exhibits strong probabilistic calibration.

The disaggregated view reveals several key patterns.

Tree-based models such as XGBoost and Random Forest demonstrate superior statistical consistency when applied across different asset classes. The Kupiec test confirms that both models achieve validation rates above 50% across most segments while XGB reaches 82% for cryptocurrencies proving

their precision and reliability in extreme quantile estimation. The models demonstrate resilience when applied to equities and fixed income despite these assets being traditionally challenging to predict because of market noise and macroeconomic influences.

LSTM, surprisingly, delivers mixed results. ETFs demonstrate strong statistical performance at 45% and commodities at 56%, while equities show poor results at 37% and the approach underperforms when compared to ensemble methods. The difference in performance indicates that LSTM models are sensitive to specific asset characteristics which become more pronounced in markets experiencing constant structural changes or fluctuating volatility patterns.

Validation results show that the hybrid model of GARCH and LSTM exceeds expectations in fixed income markets and outperforms XGBoost and Random Forest models in this asset class alone. Volatility conditioning through GARCH demonstrates enhanced utility for interest rate–sensitive assets because these assets experience smoother regime changes that are influenced by policy cycles instead of investor sentiment. The validation rates for equities and cryptocurrencies fall to 30% and 54% respectively which demonstrates that this method lacks consistent performance.

Traditional models demonstrate low Kupiec validation rates across various classes with specific rates of 27% for equities and 23% for ETFs and 22% for fixed income. The models demonstrate systematic underperformance during crisis conditions which stems from their rigid assumptions (such as Gaussianity and independence) combined with their inability to adjust to market shifts. Cryptocurrencies achieved an unusual score of 71% that could stem from insufficient data during crisis periods.

The analysis demonstrates that ensemble machine learning techniques like XGB and RF produce reliable empirical results together with formal statistical legitimacy across various asset categories during stress scenarios. The performance of LSTM-based models yields inconsistent results which result in sporadic passing of the Kupiec test. The hybrid GARCH-LSTM produces some benefits but does not provide consistent performance enhancement.

3.1.5 Quantile Loss Evaluation for VaR Forecasting

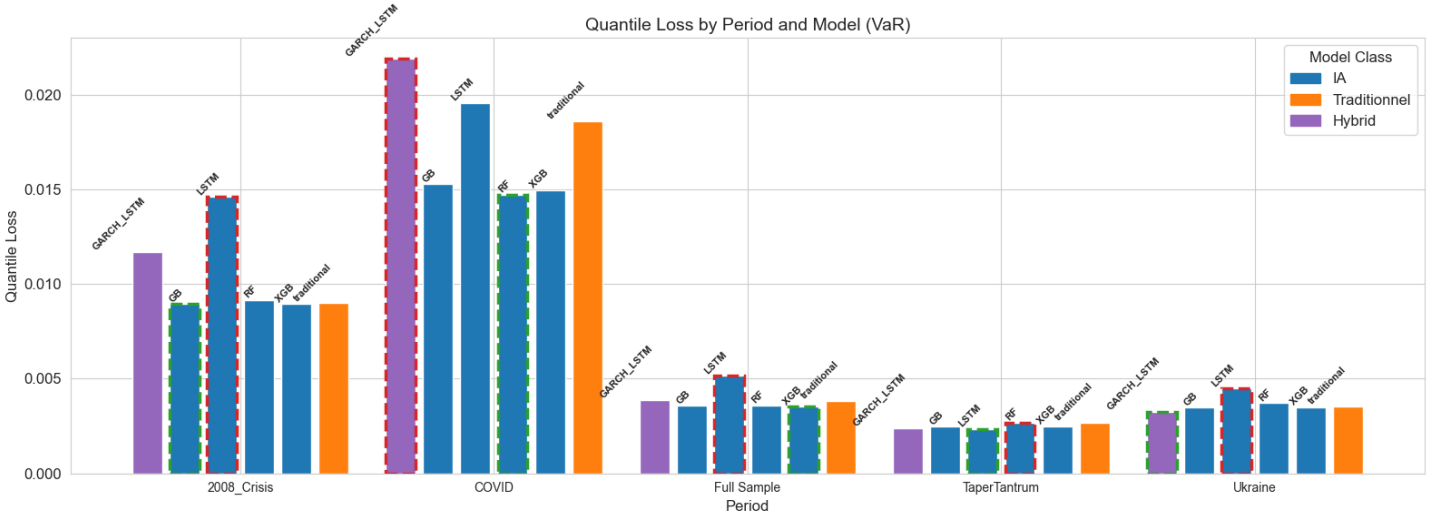


Figure 14 – Average quantile loss for VaR forecasting by periods and models

Figure 14 demonstrates that quantile loss becomes significantly higher during major crisis periods with deep learning models being the most affected. The GARCH-LSTM model generates the highest errors during the COVID pandemic with values over 0.02 while the LSTM model does so during the 2008

financial crisis with errors above 0.015. These models fail to deliver dependable tail estimates when facing extreme volatility conditions even though they offer high levels of adaptability.

XGBoost and Random Forest maintain lower quantile losses consistently throughout turbulent periods with minimal performance degradation. The combined structure of these models helps to mitigate fluctuations in performance. The traditional model performs well generally but shows its limitations during COVID because it cannot adapt to unexpected shocks.

All models demonstrate convergent performance with reduced quantile loss during stable events like the Taper Tantrum and Ukraine crisis.

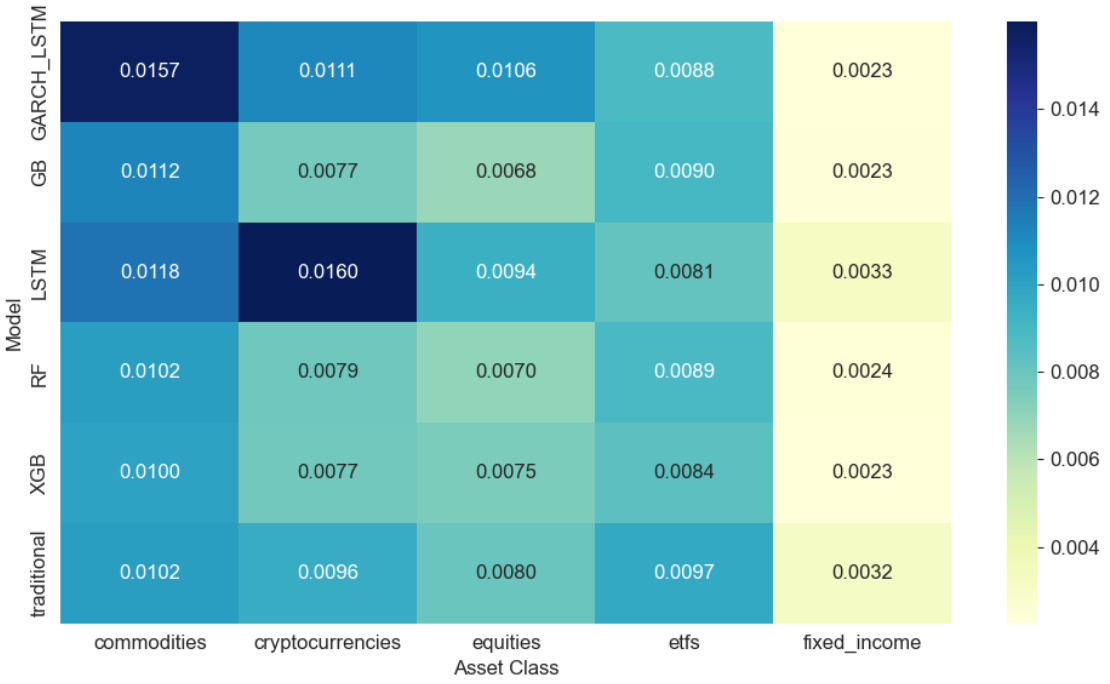


Figure 15 - Heatmap of the average Quantile Loss by model and asset class for VaR forecasting

Figure 15 reveals XGBoost's superior performance across every asset class as it maintains values below 0.01 for each category. The generalization abilities of both LSTM and GARCH-LSTM models prove weak in cryptocurrency and commodity markets with quantile losses reaching peaks of 0.016 and 0.0157.

Traditional approaches demonstrate robust performance with equities and ETFs yet struggle with riskier asset classes. Studies reveal tree-based models as superior to deep learning models because they maintain accuracy while demonstrating resilience against noise and structural alterations.

3.2 Conditional Value-at-Risk Forecasting (CVaR)

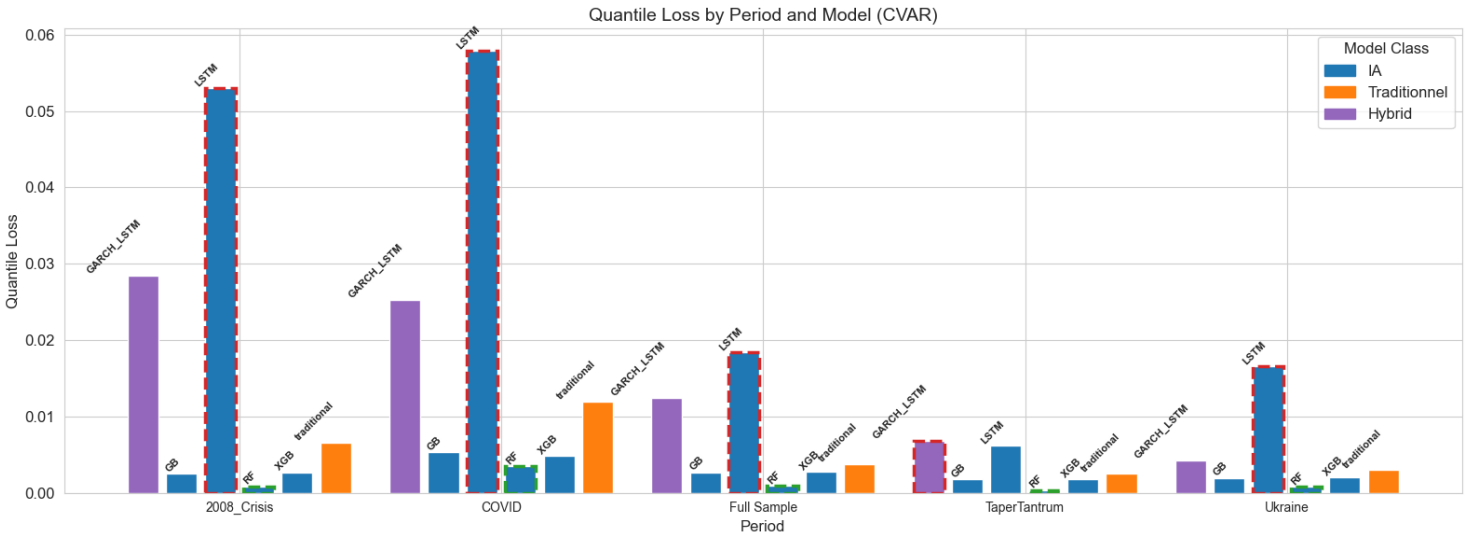


Figure 16 - Quantile Loss for CVaR forecasts by model and time period

Figure 16 demonstrates significant variations in the predictive accuracy of CVaR estimates throughout times of market stress. The LSTM model registers the highest quantile loss at over 0.05 during COVID and 0.045 during 2008 which points to severe miscalculations of expected tail losses. The recurrent structure of this model does not generalize well to sudden volatility shocks during crises despite its design for temporal dependency capture.

Despite increased stability from the GARCH-LSTM hybrid approach it demonstrates substantial loss inflation during crisis periods which indicates volatility conditioning provides minimal improvements in CVaR estimation.

Random Forest and XGBoost maintain low quantile losses even in crisis periods. Their robustness demonstrates their superior capability to incorporate distributional skewness and nonlinearity into tail estimates effectively. Traditional models continue to provide competitive results yet demonstrate inferior performance when compared to ensemble methods in challenging situations.

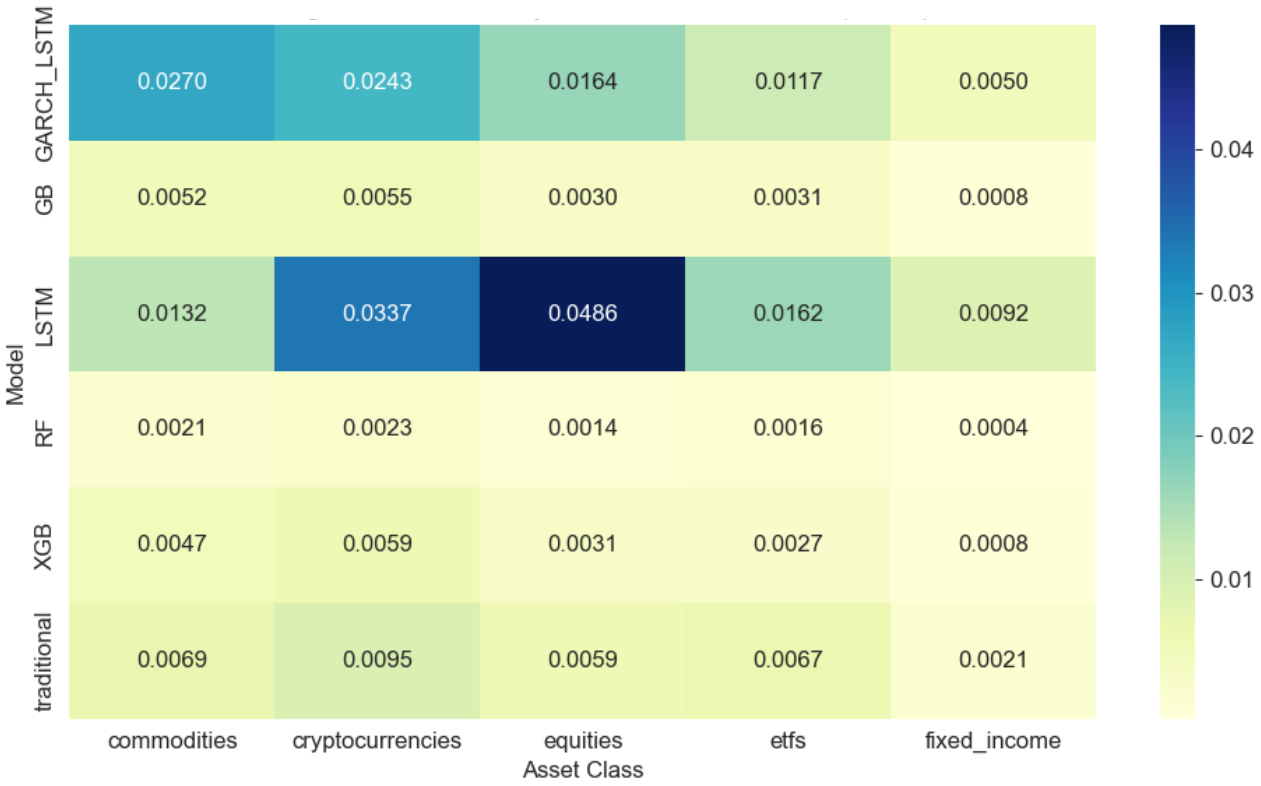


Figure 17 – Heatmap of the average Quantile Loss by model and asset class for CVaR forecasting

According to Figure 17 the heatmap validates the observed patterns throughout different asset classes. The LSTM model shows its greatest error rates in equities with a score of 0.0486 and cryptocurrencies at 0.0337 which reveals its vulnerability to assets affected by noise or market regime shifts. The GARCH-LSTM model shows steady underperformance when applied to commodity and cryptocurrency markets.

The performance of XGBoost and Random Forest models exceeds that of all other models with frequent large margins. Random Forest demonstrates quantile losses below 0.002 for every asset class while XGBoost achieves similar results with losses below 0.006. The results demonstrate the strength of these algorithms to model conditional tails across multiple market settings.

The findings demonstrate that ensemble models provide better calibration for VaR while offering more precise assessments of CVaR through their ability to capture both threshold risk and loss structure beyond it. Deep learning models demonstrate instability in tail-conditional scenarios despite the inclusion of volatility data.

3.3 Volatility and Maximum Drawdown Forecasting

Evaluating realized volatility and maximum drawdown forecasts with quantile-based evaluation methods provides enhanced analytical insights. Here, model performance is assessed using standard regression metrics: Standard regression metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Adjusted R^2 serve as the evaluation framework for models.Model accuracy maintains a consistent order of performance according to the MAE plot shown in Figure 18. Both

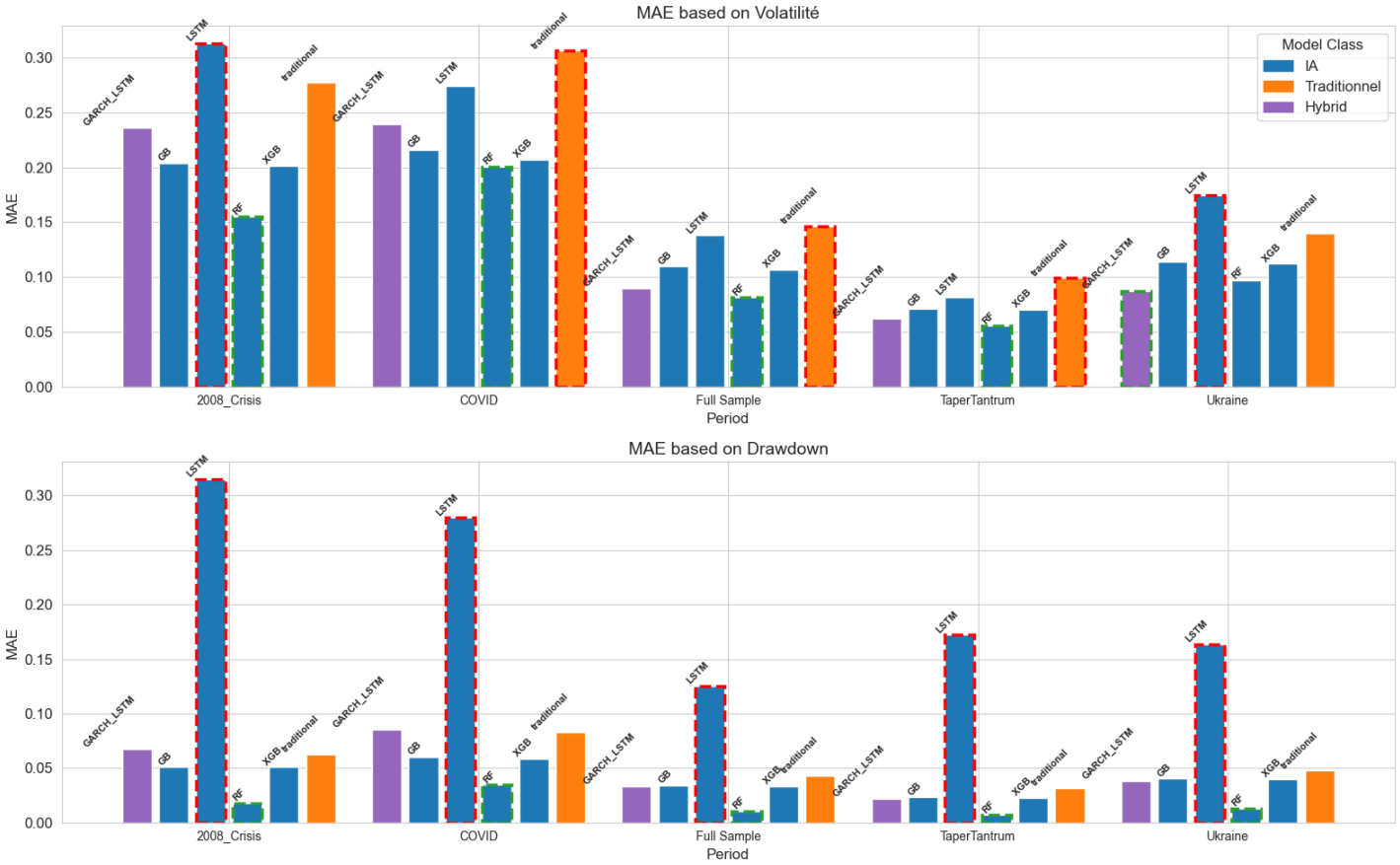


Figure 18 - Mean Absolute Error (MAE) for volatility and drawdown forecasts across models and periods

Random Forest and XGBoost achieve the smallest absolute errors in most periods for both targeted predictions. While Random Forest and XGBoost models kept MAEs below 0.1 for volatility during the 2008 crisis, LSTM recorded MAEs above 0.3 which demonstrates a significant deviation from actual risk measurements.

Drawdown forecasting errors tell a similar story. During the COVID period the LSTM model reached an error rate of 0.31 while Random Forest and XGBoost models maintained errors below 0.05. Ensemble models demonstrate superior performance in point risk estimation and exhibit better responsiveness to downside accumulation changes.

In crisis conditions traditional models outperform LSTM but remain inferior to advanced machine learning methods. Intermediate results emerge from the GARCH-LSTM hybrid which shows some stability advantages over standalone LSTM although its performance falls short compared to base ensemble models.

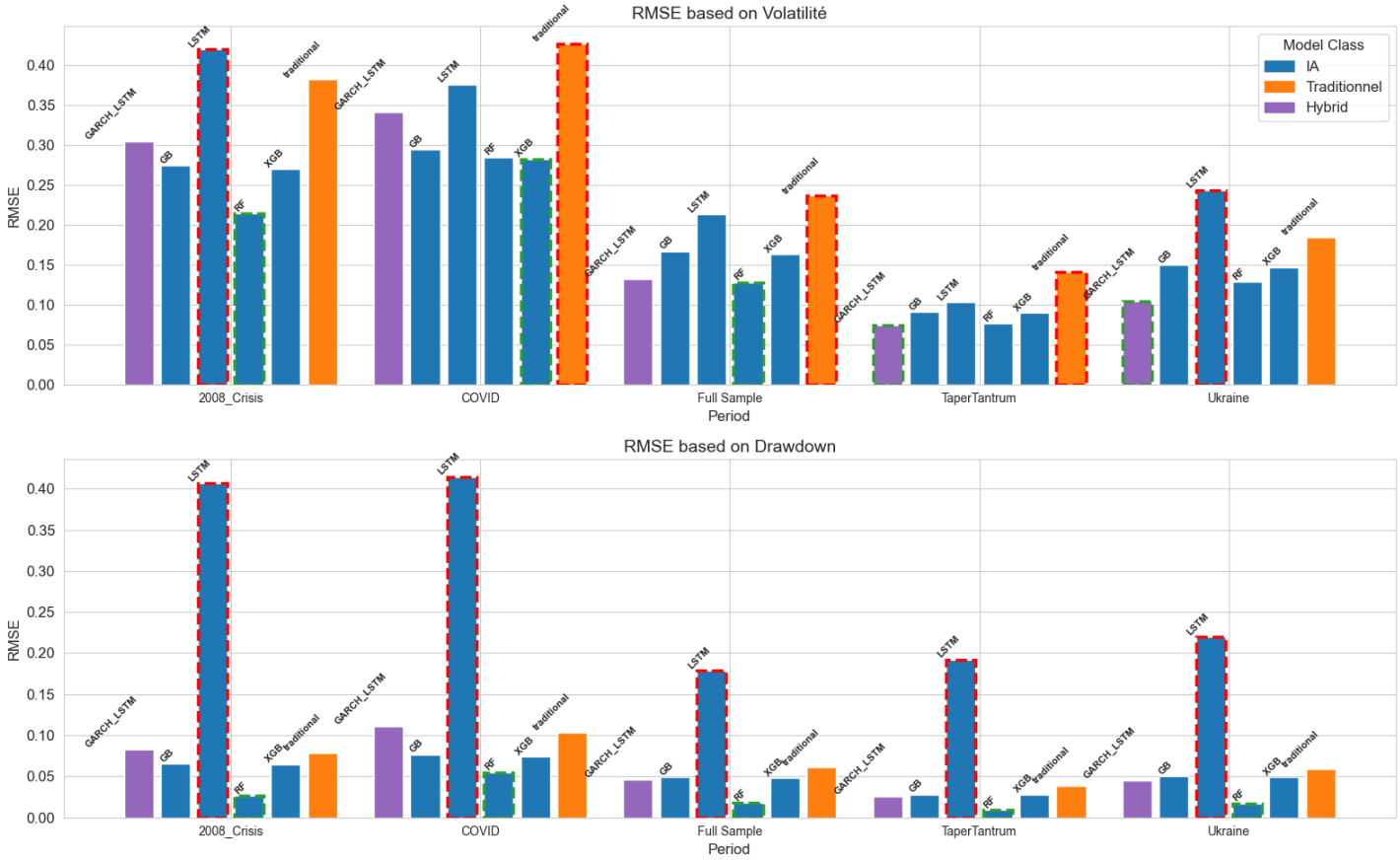


Figure 19 - Root Mean Squared Error (RMSE) for volatility and drawdown forecasts across models and periods

According to Figure 19, ensemble models (RF and XGB) demonstrate superior performance as shown by RMSE results because their error variance remains low under crisis conditions. The RMSEs for both volatility and drawdown measurements stay below 0.2 and 0.1 respectively during every evaluation period.

LSTM demonstrates the highest RMSE which goes beyond 0.4 in both the 2008 crisis and the COVID period. The results demonstrate low forecasting accuracy together with substantial instability when structural breaks occur. Traditional models also struggle.

In specific timeframes such as during COVID the volatility RMSE of traditional models comes close to that of LSTM. Traditional models exhibit substantial drawdown errors which become especially problematic during periods of swift risk escalation. The combination of GARCH and LSTM techniques delivers only moderate performance gains yet fails to reach the stability levels of tree-based methods.

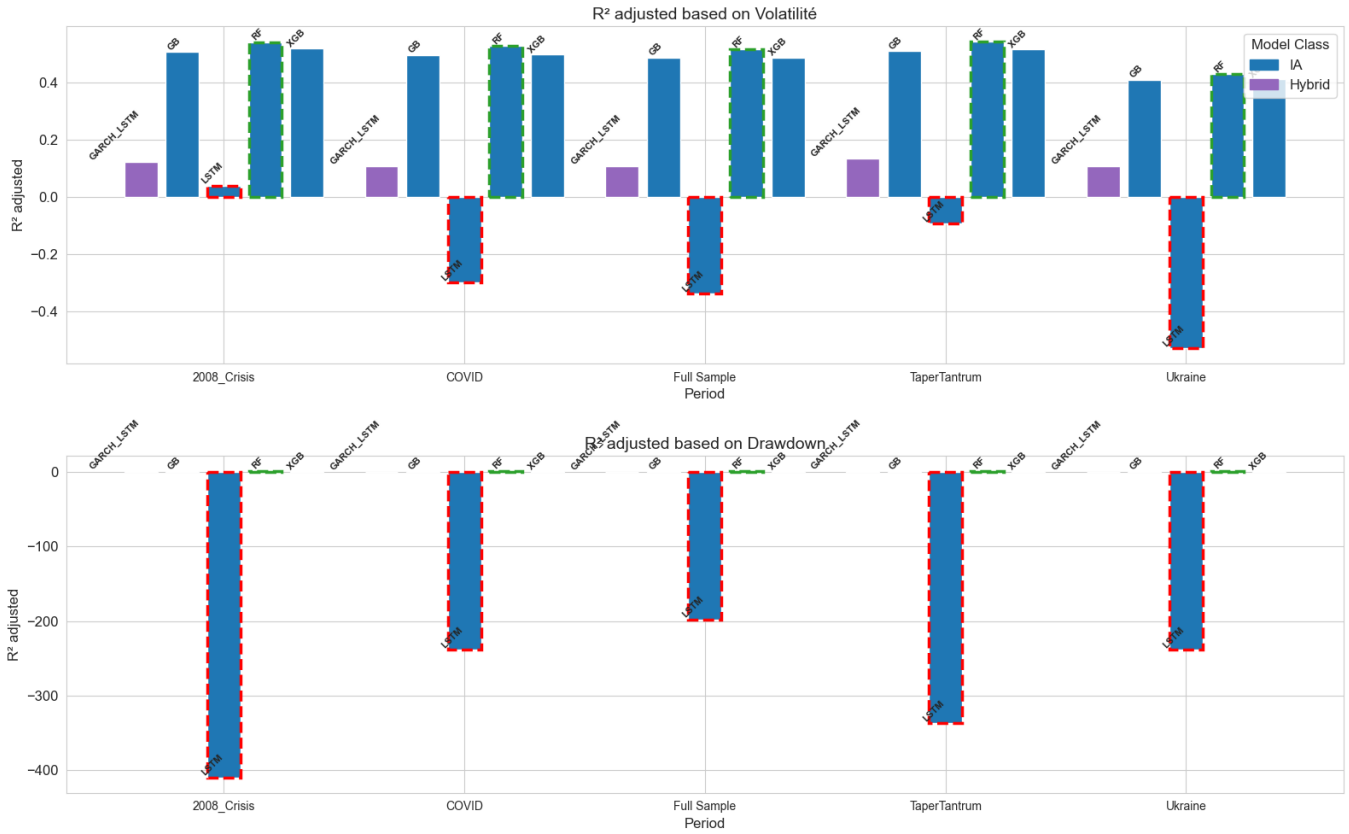


Figure 20 - Adjusted R-squared for volatility and drawdown forecasts

In the figure 20, the adjusted R^2 values show that RF and XGBoost are the only models consistently achieving strong explanatory power, with R^2 scores above 0.45 for volatility across all regimes.

LSTM, on the other hand, performs disastrously—especially on drawdown—where R^2 values plunge below -400, indicating massive overfitting and a failure to extract usable structure from the data.

Traditional models remain close to zero, occasionally negative, underscoring their limited adaptability. The GARCH-LSTM model improves over LSTM on volatility but remains weak on drawdown, suggesting that volatility conditioning alone does not yield general predictive power.

3.4 Model Interpretability

Interpretability functions as an essential requirement that supports trustworthiness, accountability measures, and compliance with regulatory standards in the domain of financial risk forecasting. Despite their superior predictive capabilities, complex models like ensemble learners and deep neural networks function as black boxes which complicates justification of their outputs for practical decision-making applications.

To address this, the present study employs two complementary post hoc interpretability techniques: The research incorporates LIME (Local Interpretable Model-agnostic Explanations) together with SHAP (SHapley Additive exPlanations) as methods to improve model interpretability. Both methods aim to identify which features most influence model predictions, but they differ in philosophy: LIME offers insights into local model behavior by using linear approximations while SHAP determines global feature importance through the principles of cooperative game theory.

The subsequent sections examine the extraction methods of various models that process financial and macroeconomic inputs to determine if their fundamental reasoning matches domain knowledge or shows structural flaws.

3.4.1 Model Interpretability with LIME

Risk management depends on understanding how models create predictions beyond their performance results. Local Interpretable Model-agnostic Explanations (LIME) creates approximations that demonstrate how each input variable affects specific predictions. When utilized in this context it reveals the comparative importance of financial features throughout various model families. Figure 21 demonstrates the different input feature weights used by each model.

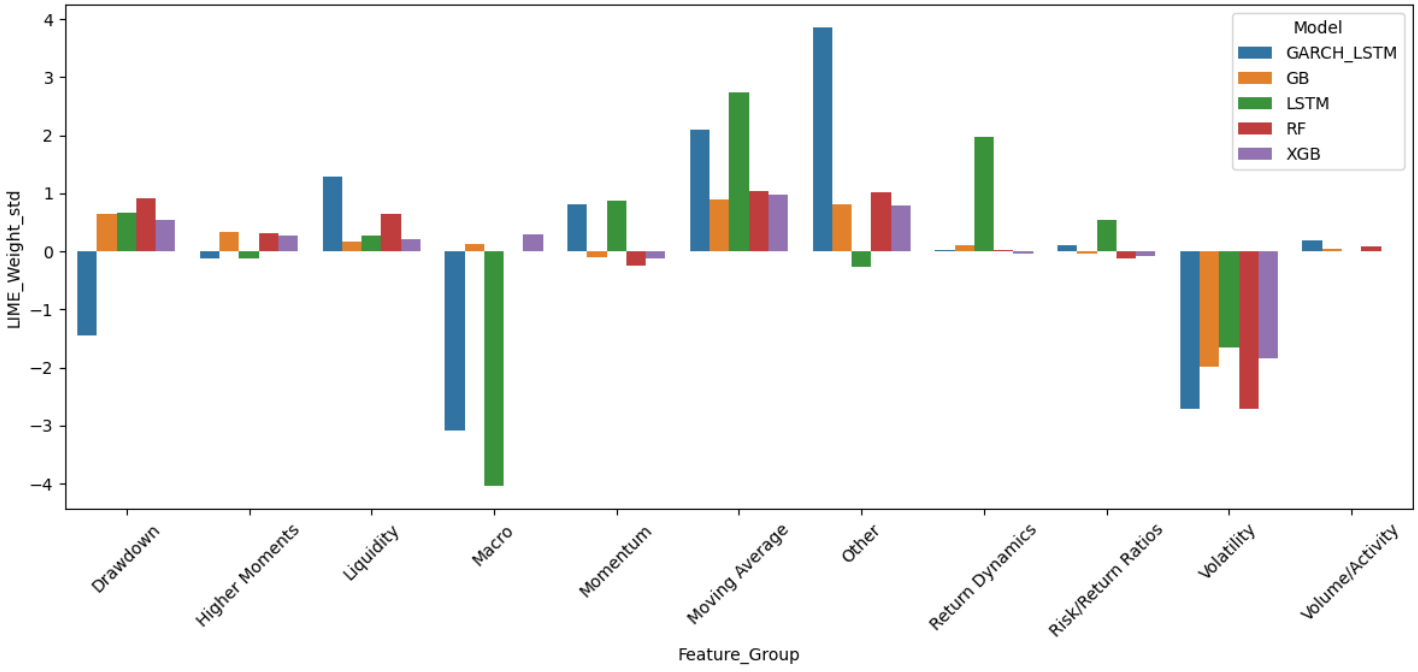


Figure 21 - Relative importance of feature groups per model (LIME, standardized)

Two contrasting logics emerge:

- GARCH-LSTM and LSTM models prioritize moving averages and liquidity over macro indicators and volatility which they penalize heavily. The structure appears to follow market trends without any connection to essential fundamental factors.

- LSTM shows an even more extreme pattern: The LSTM model shows strong negative weights for macroeconomic factors and volatility while placing too much emphasis on momentum along with moving averages and return dynamics. Prioritizing short-term technical features creates instability during market stress periods.
- The ensemble models including RF, XGB, and GB show a balanced distribution of importance across features.

Models demonstrate balanced preferences for drawdown and liquidity alongside moving average but prevent excessive concentration in any feature group. XGBoost demonstrates both stability and diversified feature distribution in its model profile. Across different models deep learning methods show a marked underutilization of volatility and macro features which creates structural interpretability concerns.

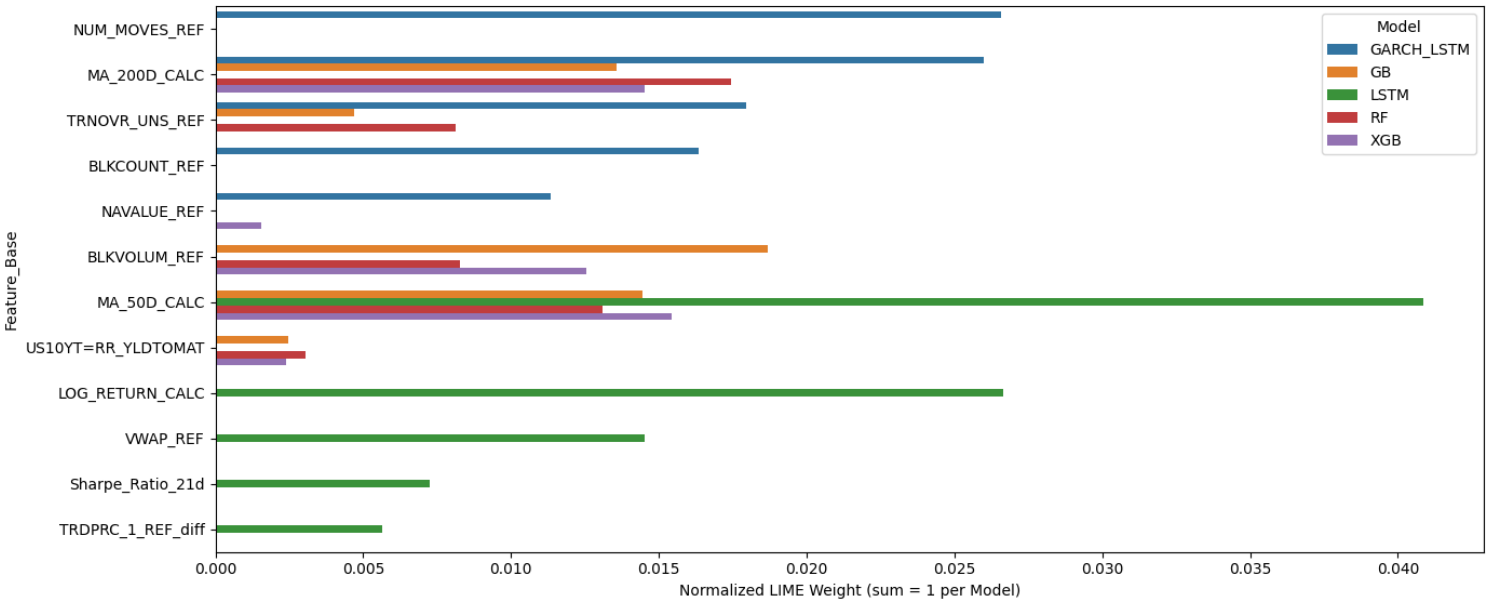


Figure 22 - Normalize LIME weights for the top 5 features across models

The feature-level analysis (Figure 22) reinforces these group-level trends. For example:

- LSTM identifies MA_50D_CALC, LOG_RETURN_CALC, and VWAP_REF as its most important variables because they are all short-term and highly reactive.
- The algorithms Random Forest and XGBoost select MA_200D_CALC, TRNOVR_UNS_REF, and BLKVOLUM_REF as primary features which indicates structural and liquidity-adjusted behavior emphasis and potentially boosts robustness.
- The GARCH-LSTM model accounts for block count and number of trades (BLKCOUNT_REF, NUM_MOVES_REF) which shows its volatility conditioning combines transactional signals rather than fundamental market drivers.

The research shows that each model has unique input preferences along with separate conceptual prejudices when processing market information.

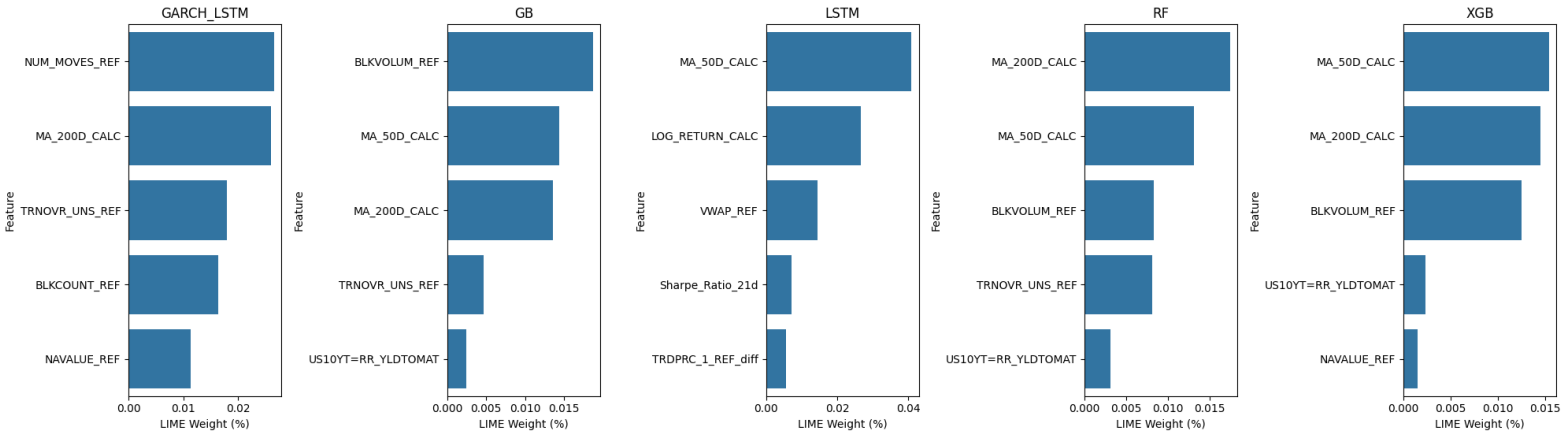


Figure 23 - Normalized LIME weights for the top 5 features accross models

The ultimate LIME diagram delivers (Figure 23) a comparative analytical synthesis. The features NUM_MOVES_REF and MA_200D_CALC stand out as major contributors in both GARCH-LSTM and GB models because they manage to capture consistent structural patterns in the data. The LSTM model excessively relies on a few return-based indicators such as LOG_RETURN_CALC and Sharpe_Ratio_21d that potentially restrict its ability to adapt during crisis situations.

The LIME analysis demonstrates that tree-based models use a wider variety of consistent features while LSTM approaches focus too heavily on immediate technical indicators which reduces their interpretability and generalization abilities. The GARCH-LSTM hybrid model offers greater interpretability than the standalone LSTM model but relies on volume and activity data to provide incomplete stabilization of its predictive performance.

3.4.2 Model Interpretability with SHAP

SHAP (SHapley Additive exPlanations) provides a theoretically sound global method that complements LIME by attributing each feature's contribution to model output. SHAP uses cooperative game theory principles to break predictions down into additive feature contributions which allow for robust and consistent interpretation of complex model behavior.

SHAP delivers stable global explanations ideal for model comparison across different asset classes and time periods as opposed to LIME. Technical restrictions associated with the Keras framework prevented SHAP value computation for the LSTM and GARCH-LSTM models. The limitation exists because the integration between SHAP and specific deep learning backends was unstable during the analysis period.

The SHAP analysis below exclusively examines tree-based models like Random Forest, XGBoost, and Gradient Boosting because SHAP has native support and computational reliability for these models. The insights present valuable comparative perspectives on model logic which complement the results uncovered through LIME analysis.

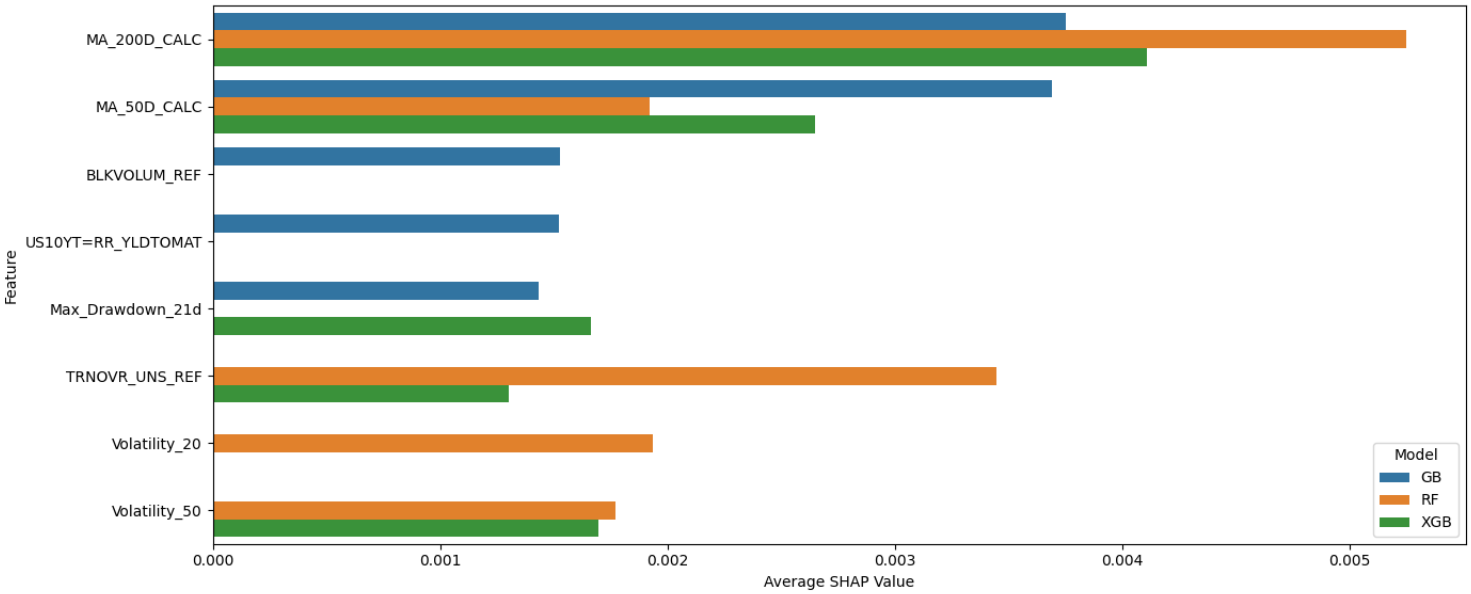


Figure 24 - Top 5 SHAP features by model (GB, RF, XGB)

The analysis of Figure 24 highlights that Gradient Boosting (GB), Random Forest (RF), and XGBoost (XGB) show consistent agreement on feature importance:

- MA_200D_CALC and MA_50D_CALC emerge as the top influential features across all models because long-term and medium-term moving averages play a key role in determining model predictions.
- RF and XGB assign significant importance to volatility measures like Volatility_20 and Volatility_50 as well as turnover (TRNOVR_UNS_REF) showing sensitivity to price fluctuations along with market liquidity.
- Gradient Boosting (GB) assigns more weight to macroeconomic signals like US10YT=RR_YLDTOMAT than XGBoost does which shows its fundamentalist approach.

The findings validate the principle that ensemble models focus on signals which are both interpretable and rooted in economic theory with momentum and liquidity factors as primary examples.

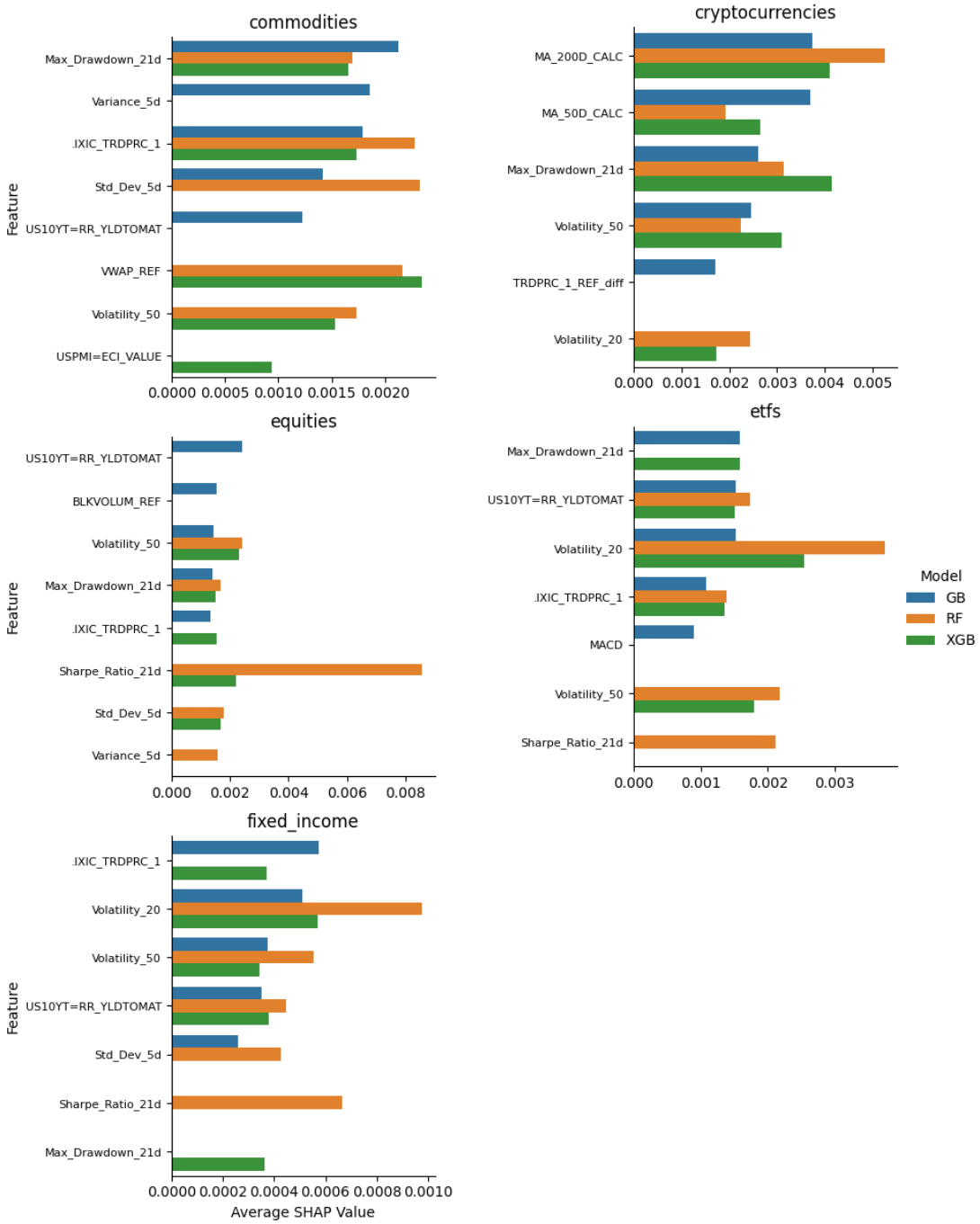


Figure 25 - SHAP feature contributions across models and classes

Figure 25 breaks down feature importance across different asset categories including commodities, crypto, and equities, uncovering the following trends:

- In commodity markets explanatory power is mainly derived from volatility and macroeconomic features such as US10YT and Variance_5d because they matter most in inflation-sensitive environments.
- Cryptocurrency models show a preference for technical signals like moving averages and drawdowns which confirms that trend indicators serve as the main drivers for assets lacking fundamental support.
- Equities and ETFs show a more diversified mix: alongside volatility and return-based indicators.

- Macroeconomic indicators (US10YT, TRNOVR, Volatility_20) maintain strong influence in fixed income markets because they match up closely with the behavior of yield curves and the design of bond markets.

Based on SHAP analysis model adaptation reflects asset class features where structural inputs primarily influence bonds and commodities while digital assets show a dominance of technical signals.

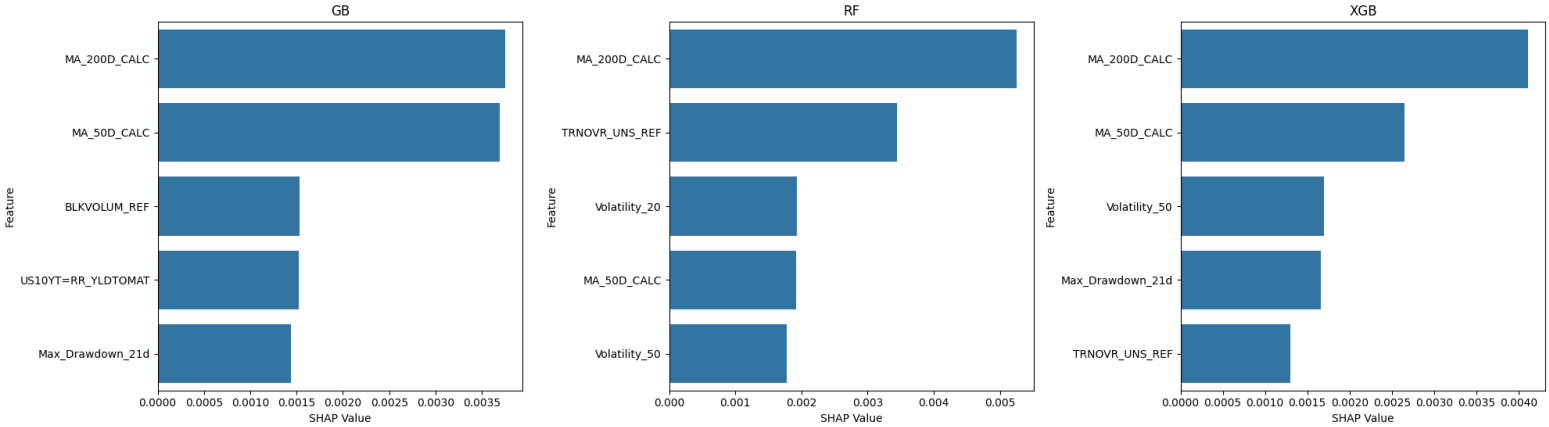


Figure 26 - Detailed SHAP contributions per model and asset type

Figure 26 provides a concise comparative representation through the final SHAP visualization. It shows that:

- The Random Forest model demonstrates equal distribution of feature importance between macroeconomic variables and both liquidity and volatility metrics.
- XGB assigns slightly more importance to trend-following indicators in its analysis.
- GB stands out by including macroeconomic inputs across several asset types which confirms its robust fundamental approach.

Prediction logic relies heavily on both drawdown-related features and volatility metrics including Volatility_50, Variance_5d, and Std_Dev_5d.

SHAP analysis demonstrates that models consistently give precedence to explainable features like moving averages and volatility measures instead of unpredictable or overfitted signals. The performance results confirm that these ensemble models maintain strong operational validity.

SHAP validates that ensemble-based models achieve high performance while maintaining interpretability. Their predictions are systematically linked to financial logic: Financial logic shows that trend indicators alongside volatility and liquidity features reign supreme. Although SHAP computation cannot be performed on LSTM and GARCH-LSTM models which prevents direct comparison, the reliability and clarity exhibited by RF, XGB and GB models make them preferable for real-world applications that require risk sensitivity.

4 Conclusion

The research objective of this thesis was to determine if machine learning models and hybrid approaches offer better accuracy and robustness along with clearer interpretation for market risk forecasts than traditional methods. This research originates from the persistent shortcomings of traditional models like Value-at-Risk (VaR) and GARCH-based estimators which became apparent during market crises such as the financial crisis of 2008 and the COVID-19 pandemic.

Market disruptions exposed significant limitations in classical models and created a demand for forecasting systems that can adapt to structural breaks and nonlinear dependencies while still providing easy-to-read results. The study produced a comprehensive empirical framework to assess traditional approaches such as Historical VaR and GARCH alongside machine learning techniques like Random Forest and XGBoost and combined systems including GARCH-LSTM. These models were tested across diverse asset classes—equities, ETFs, bonds, and cryptocurrencies—and evaluated on three key dimensions: The study measures models according to their forecasting accuracy, stability during crises, and ease of interpretation.

Traditional models maintain their value through ease of use and regulatory conformity but fail to adequately represent tail risk in unstable market periods. Machine learning models such as tree-based algorithms and LSTM networks distinguish themselves through their strong performance in modeling nonlinear patterns and adapting to market regime changes. GARCH-LSTM models represent an effective middle ground by combining statistical structures with the flexible learning capabilities of deep learning networks.

Different asset classes along with the time frame and performance metric selected determine the usefulness of these methods. This thesis presents academic progress with its evaluation framework that combines interpretability tools SHAP and LIME with tail-focused metrics in crisis situations. The study demonstrates that interpretability should be considered both a limitation and an evaluation metric for risk modeling methods. This methodology gains importance because regulators now require risk systems to demonstrate transparency and auditability.

The research outcomes guide practitioners on how machine learning and hybrid models can enhance current instruments in stress testing and early warning system applications. The results also show that interpretability techniques can make these models more usable in operational and supervisory contexts. The Fundamental Review of the Trading Book (FRTB) calls for regulators to evaluate model governance and internal risk validation frameworks.

These findings provide strong empirical support for the hypotheses outlined at the beginning of the thesis. The primary hypothesis (H_1)—that machine learning and hybrid models outperform traditional approaches in terms of predictive performance, robustness, and interpretability—was validated across multiple asset classes and evaluation scenarios. The supporting hypotheses H_1a, H_1b, and H_1c were also confirmed to varying degrees, depending on market conditions and asset characteristics. In particular, machine learning models excelled in high-volatility environments, demonstrated greater resilience to structural breaks, and achieved interpretable outputs with tools like SHAP and LIME. However, these results should be interpreted with consideration of the study's constraints and the complexity of model generalization.

The study's limitations need acknowledgment despite its valuable contributions. Limited computational resources paired with time constraints served as initial limitations for the project. The deep learning models require more training time and tuning than this project scope permitted. Research limitations and the extensive availability of high-quality data prevented me from testing all potential model combinations across different asset classes and time periods equally. The project constraints forced to keep hyperparameter optimization to a minimum to ensure feasibility.

SHAP and LIME offer valuable interpretability information but provide only approximate interpretability results when applied to sequential models like LSTMs that process complex temporal relationships. The research depended solely on Refinitiv data sources which introduce vendor-specific biases and prevent external validation of reproducibility. The limitations can be addressed through expanding dataset scale and diversity and performing more thorough optimization of parameters as well as exploring attention-based transformers and temporal convolutional networks as new model structures. Combining real-time macroeconomic indicators with unstructured data sources together with new explainability techniques will offer enhanced understanding of model effectiveness throughout financial crisis periods. Longitudinal studies of these models across different crisis scenarios and markets will determine their applicability and stability in a range of conditions.

Appropriate explainability tools enhance market risk forecasting results when used alongside machine learning and hybrid models. Modern financial systems need models that are flexible and easy to understand to predict extreme risks while still benefiting from transparency and simplicity.

5 Bibliography

1. **Bailey, D. H., & López de Prado, M.** (2014). Stop-outs under serial correlation and 'The Triple Penance Rule'. *The Journal of Risk*.

2. **Bank of England, & Financial Conduct Authority.** (2022). *Artificial intelligence and machine learning: Exploring regulatory principles*.

3. **Basel Committee on Banking Supervision.** (2013). *Fundamental review of the trading book: A revised market risk framework (Consultative Document)*. Bank for International Settlements.

4. **Basel Committee on Banking Supervision.** (2016). *Minimum capital requirements for market risk*. Bank for International Settlements.

5. **Breiman, L.** (2001). Statistical modeling: The two cultures. *Statistical Science*

6. **Cassidy, J.** (2010). What's wrong with risk models? *The New Yorker*.

7. **Chen, T., & Guestrin, C.** (2016). XGBoost: A scalable tree boosting system.

8. **Christoffersen, P. F.** (2012). *Elements of financial risk management* (2nd ed.). Academic Press.

9. **CNBC.** (2020, March 13). *Bitcoin loses half of its value in two-day plunge*.

10. **Cont, R.** (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*.

11. **Danielsson, J., James, K., Valenzuela, M., & Zer, I.** (2016). Model risk of risk models. *Journal of Financial Stability*

12. **Fabozzi, F. J.** (2012). *Bond markets, analysis and strategies*.

13. **Fischer, T., & Krauss, C.** (2018). Deep learning with long short-term memory networks for financial market predictions.

14. **Gramegna, M., & Giudici, P.** (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Journal of Risk Model Validation*

15. **Gu, S., Kelly, B., & Xiu, D.** (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*.

16. **Hochreiter, S., & Schmidhuber, J.** (1997). Long short-term memory. *Neural Computation*.

17. **Jondeau, E., & Rockinger, M.** (2003). Conditional volatility, skewness, and kurtosis: Existence and persistence. *Journal of Economic Dynamics and Control*.

18. **Jorion, P.** (2007). *Value at risk: The new benchmark for managing financial risk* (3rd ed.). McGraw-Hill.

19. **Kakade, S., Jain, A., Ramani, R. G., & Raut, R. D.** (2022). Value-at-risk forecasting: A hybrid ensemble learning GARCH-LSTM based approach.

20. **Kartsonakis-Mademlis, I., & Dritsakis, N.** (2021). Volatility forecasting using hybrid GARCH neural network models: The case of the Italian stock market. *International Journal of Economics and Financial Issues*.

21. **Krauss, C., Do, X. A., & Huck, N.** (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*.

22. **Lo, A. W.** (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*.

23. **Lundberg, S. M., & Lee, S.-I.** (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

24. **Markowitz, H.** (1952). Portfolio selection. *Journal of Finance*

25. **McNeil, A. J., Frey, R., & Embrechts, P.** (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press.

26. **O'Shaughnessy, J.** (2014). *What works on Wall Street*.

27. **Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. W.** (2017). A dual-stage attention-based recurrent neural network for time series prediction.

28. **Ribeiro, M. T., Singh, S., & Guestrin, C.** (2016). Why should I trust you? Explaining the predictions of any classifier.

29. **Rockafellar, R. T., & Uryasev, S.** (2000). Optimization of conditional value-at-risk.

30. **Rudin, C.** (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*

31. **Sharpe, W. F.** (1966). Mutual fund performance. *The Journal of Business*

32. **Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H.** (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods.

33. **Sortino, F. A., & Price, L. N.** (1994). Performance measurement in a downside risk framework. *The Journal of Investing*

34. **Sortino, F. A., & Satchell, S. E.** (2009). *The Sortino framework for constructing portfolios: Focusing on desired target return™ to optimize upside potential relative to downside risk*. Elsevier.

35. **Sortino, F. A., & van der Meer, R.** (1991). Downside risk. *Journal of Portfolio Management*

36. **Taleb, N. N.** (2007). *The black swan: The impact of the highly improbable*. Random House.

37. **Tuckman, B., & Serrat, A.** (2011). *Fixed income securities: Tools for today's markets*.

38. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.** (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

39. **Young, T. W.** (1991). Calmar ratio: A smoother tool. *Futures*.

40. **Christensen, K., Siggaard, M., & Veliyev, B.** (2021). *A machine learning approach to volatility forecasting*

6 Appendices

6.1 Tables and Figures

6.2 Glossary

Term	Definition
Backtesting	A method of testing the accuracy of a risk model (e.g., VaR) by comparing its predictions to actual historical data over the same period.
CvaR (Conditional Value-at-Risk)	A risk measure estimating the expected loss beyond the Value-at-Risk (VaR) threshold, often used to evaluate tail risk.
Deep Learning	A subset of machine learning based on neural networks with many hidden layers, capable of extracting complex hierarchical features from data.
Ensemble Learning	A technique that combines predictions from multiple models to improve overall performance and reduce overfitting.
Epoch	One complete pass through the full training dataset during the training of a machine learning model.
GARCH (Generalized AutoRegressive Conditional Heteroskedasticity)	A statistical model used to forecast the volatility of time series data, especially financial returns, by modeling conditional variances.
Hybrid Model	A predictive model that combines elements from two or more approaches (e.g., statistical and machine learning) to leverage their complementary strengths.
Hyperparameter	A model configuration variable set before training that influences learning behavior, such as learning rate or number of layers.
LIME (Local Interpretable Model-agnostic Explanations)	A technique that explains individual predictions of any black-box model by approximating it locally with an interpretable model.
LSTM (Long Short-Term Memory)	A type of recurrent neural network (RNN) designed to capture long-range dependencies in sequence data using memory cells and gates.
Machine Learning (ML)	A field of AI where algorithms learn patterns from data to make predictions or decisions without being explicitly programmed for each task.
MAE (Mean Absolute Error)	A model performance metric representing the average absolute difference between predicted and actual values.
MSE (Mean Squared Error)	A commonly used loss function that measures the average squared difference between predicted and actual values.
RMSE (Root Mean Squared Error)	The square root of MSE, offering an interpretable error measure in the same unit as the target variable.
R-Squared (R^2)	A statistical measure representing the proportion of variance explained by a regression model; values closer to 1 indicate better fit.
Overfitting	A modeling error where a model performs well on training data but fails to generalize to unseen data due to excessive complexity.
Random Forest (RF)	An ensemble learning method that builds multiple decision trees and aggregates their outputs to improve predictive performance and stability.
Recursive Feature Elimination (RFE)	A feature selection technique that recursively removes the least important features based on model weights, aiming to improve model performance by reducing dimensionality.
RNN (Recurrent Neural Network)	A neural network designed to handle sequential data by maintaining a memory of past inputs through internal loops.

Sharpe Ratio	A measure of risk-adjusted return calculated as the portfolio's excess return over the risk-free rate divided by the standard deviation of returns.
SHAP (SHapley Additive exPlanations)	A game-theoretic approach to explain machine learning model outputs by fairly attributing the contribution of each feature to the final prediction.
Skewness	A measure of the asymmetry of a probability distribution; positive skew indicates a longer right tail, negative skew a longer left tail.
Stationarity (of a time series)	A property where the statistical characteristics of a process (mean, variance, autocorrelation) remain constant over time.
Stress Testing	A simulation technique used to assess how financial models or portfolios perform under extreme conditions.
Tail Risk	The risk of rare but extreme events located in the tail of a probability distribution, often associated with financial losses.
Tuning	The process of selecting the best hyperparameters for a machine learning model to optimize its performance.
VaR (Value-at-Risk)	A risk metric estimating the maximum expected loss of a portfolio over a given time horizon at a specified confidence level.
Volatility	A measure of the variation in the price or returns of a financial asset, often used as a proxy for risk.
XAI (Explainable Artificial Intelligence)	A domain of AI focused on developing methods that make machine learning models understandable and transparent to humans.