Interpretable Machine Learning for Volatility Forecasting: A Regime-Aware Evaluation of ARIMA, LSTM, and StockMixer Across U.S. and Australian Equity Markets

Gurudeep Singh Dhinjan
Student ID: 24555981

Bachelor of Computing Science (Honours)
Major: Enterprise Systems
Honours Thesis: Financial Risk Forecasting using Interpretable AI Models

University of Technology Sydney (UTS): Faculty of Engineering and Information Technology

Supervisor: Dr. Gnana Bharathy

Abstract

The ability to accurately forecast short-term financial volatility is critical for risk management, investment strategy, and regulatory compliance. This study proposes a regime-aware, interpretable forecasting framework that evaluates the predictive performance of traditional statistical models (ARIMA) and advanced machine learning architectures (LSTM, StockMixer) across U.S. and Australian equity markets. The research employs a dual-task design regression and classification using a walk-forward cross-validation protocol to simulate realistic deployment and mitigate data leakage. Feature engineering integrates technical indicators, inter-asset correlations, and systemic risk measures such as the VIX. Model interpretability is enhanced using SHapley Additive exPlanations (SHAP), providing both global and local feature attributions. Results demonstrate that modern deep learning models outperform ARIMA in predictive accuracy and regime sensitivity, while SHAP-based diagnostics support transparency and ethical AI deployment. The findings offer actionable insights for institutions seeking robust, transparent volatility forecasting solutions in globally connected financial markets.

Keywords

Volatility Forecasting | Financial Time Series | LSTM | StockMixer | ARIMA | Walk-Forward Cross-Validation | SHAP | Explainable AI | Regime-Aware Modeling | Australian and U.S. Equity Markets

Chapter 1: Introduction

1.1 Background and Context

The COVID-19 pandemic in early 2020 exposed significant structural vulnerabilities within global financial markets. Within a matter of weeks, equity markets around the world experienced extreme volatility, with the S&P 500 declining by over 30 percent from its peak in March 2020. The Australian ASX 200 followed with similarly sharp losses. These disruptions were not isolated anomalies, but

rather part of a broader pattern of systemic fragility shaped by complex geopolitical interdependencies, high-frequency trading dynamics, and macroeconomic uncertainty. Algorithmic trading, which now constitutes over 70 percent of daily trading volume in U.S. markets, further amplifies these dynamics, often exacerbating market swings through rapid and reactive order execution. In such an environment, the ability to forecast short-term market risk, particularly price volatility, has emerged as a critical capability for investors, fund managers, risk officers, and financial regulators.

Volatility, defined as the degree of variation in an asset's price over time, serves both as a quantitative measure of risk and a forward-looking indicator for investment strategy. In the Australian context, volatility patterns are influenced by several structural factors. These include a high concentration of resource-sector equities, a relatively smaller domestic market driven by compulsory superannuation contributions, and close economic and geographic ties to the Asia-Pacific region. Together, these features contribute to unique volatility behaviours that often diverge from those observed in the United States, particularly during global economic shocks or commodity price fluctuations. As a result, improving the accuracy and interpretability of short-term volatility forecasting is not only of academic interest, but also holds practical significance for enhancing investment decision-making and regulatory oversight in both domestic and international markets.

Traditional approaches to volatility forecasting have relied on statistical time series models such as the Autoregressive Integrated Moving Average (ARIMA) and the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) families. These models offer strong theoretical foundations and are analytically interpretable. However, their applicability is limited by underlying assumptions such as stationarity, linearity, and the normality of residuals. These assumptions are frequently violated in real-world financial data, which are often characterised by regime shifts, structural breaks, and non-linear dependencies. Empirical research has consistently shown that these classical models tend to underperform in dynamic, high-frequency, and multi-asset environments, especially under conditions of market stress or during transitional periods between volatility regimes.

In response to these limitations, there has been increasing adoption of machine learning (ML) and deep learning (DL) techniques in financial time series forecasting. Models such as Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and ensemble methods including XGBoost have demonstrated superior performance over traditional techniques in capturing complex, non-linear temporal dependencies. More recently, novel lightweight architectures such as StockMixer have shown promising results by utilising Multi-Layer Perceptron (MLP) based mechanisms to model interactions across time, indicators, and assets in an efficient and scalable manner. These architectures reduce the computational burden associated with deep sequential models while retaining predictive power across various market conditions.

Despite these advancements, a major challenge remains: the interpretability of complex ML and DL models. In critical financial applications, where decisions must be transparent, auditable, and justifiable to stakeholders and regulators, black-box models are insufficient. In response, the field of Explainable Artificial Intelligence (XAI) has gained momentum, with methods such as SHapley Additive exPlanations (SHAP) now widely used to attribute model predictions to individual input features. While SHAP enhances transparency by providing global and local explanation capabilities, it is not a comprehensive solution. SHAP does not inherently address issues such as fairness, causal interpretability, or systemic bias. Therefore, while valuable, it must be employed within a broader ethical framework that supports responsible AI governance.

Regulatory interest in AI explainability has also grown significantly. Jurisdictions such as the European Union, through the proposed AI Act, have begun classifying many financial AI applications as high-risk systems, requiring detailed documentation, interpretability, and human oversight. Similarly, regulatory bodies such as the Australian Securities and Investments Commission (ASIC) and the U.S. Securities and Exchange Commission (SEC) have increased scrutiny of model governance, particularly in relation to risk modelling, automated decision-making, and the use of non-traditional data. These developments underscore the need for financial models that are not only accurate but also interpretable, fair, and compliant with evolving regulatory standards.

Another significant limitation in existing volatility forecasting literature is the widespread use of static or randomly partitioned test sets. These approaches fail to reflect the non-stationary and path-dependent nature of real-world financial data, leading to overfitted models and inflated performance metrics. In contrast, walk-forward cross-validation offers a more rigorous and realistic evaluation method. This approach tests models sequentially over time using rolling training and validation windows, thereby simulating a real-time forecasting environment. When combined with volatility regime segmentation and multi-asset evaluation, walk-forward analysis provides a robust framework for assessing model generalisability, stability, and performance across varying market conditions.

Taken together, these trends highlight the need for volatility forecasting models that are accurate, interpretable, and adaptable to real-world constraints. This study addresses these challenges by proposing a forecasting framework that integrates modern ML and DL methods with SHAP-based interpretability, walk-forward cross-validation, and regime-aware analysis. By applying this framework across both U.S. and Australian financial assets, the study aims to evaluate model robustness, interpretability, and cross-market generalisability, thereby contributing both to the academic literature and to the development of practically deployable AI-driven financial tools.

1.2 Problem Statement

Short-term volatility forecasting remains a persistent challenge in financial modelling due to the complex, noisy, and highly non-stationary nature of financial time series. Asset returns are frequently influenced by external shocks such as geopolitical events, central bank policy shifts, and macroeconomic announcements, all of which can cause abrupt and unpredictable changes in market structure. These factors introduce regime-dependent dynamics that are difficult to capture with traditional forecasting techniques. As a result, volatility prediction must be viewed not merely as a statistical optimisation problem, but as a path-dependent, context-aware task requiring adaptive and interpretable models.

Econometric models such as Autoregressive Integrated Moving Average (ARIMA) and Generalised Autoregressive Conditional Heteroskedasticity (GARCH) offer analytical tractability and a strong theoretical foundation. However, their performance tends to deteriorate in high-frequency, multi-asset, or cross-regime settings due to their reliance on fixed lag structures, linear assumptions, and stationarity. Such models struggle to accommodate the structural breaks, long memory effects, and non-linear relationships that are prevalent in real-world financial systems. Although useful in stable periods, their predictive utility declines rapidly in volatile or transitional market phases.

More recently, deep learning models such as Long Short-Term Memory (LSTM) networks have been applied to financial time series forecasting with encouraging results. These models excel in learning complex temporal dependencies and non-linear patterns. However, many are trained on an individual asset basis and fail to incorporate broader market context, such as cross-asset correlations

and systemic signals. As a result, they may achieve high in-sample accuracy yet perform poorly under distributional shifts or during regime transitions (Gajamannage and Park, 2022). Lightweight deep learning architectures like StockMixer attempt to bridge this gap by explicitly modelling inter-stock and temporal interactions, yet require further empirical validation in live-market simulations.

A major methodological gap in the existing literature is the continued use of static holdout datasets or random train-test splits, which fail to reflect the temporal dependencies and evolving structure of financial markets. These approaches can lead to inflated performance metrics and poor generalisability. In contrast, this study adopts a walk-forward cross-validation (WFCV) framework, which simulates realistic deployment conditions by sequentially training and evaluating models across rolling time windows. This design enables more accurate assessment of model robustness, adaptability, and regime-specific performance.

In addition to methodological challenges, there are interpretability and ethical considerations that must be addressed. The financial sector increasingly operates under regulatory mandates requiring AI systems to be transparent, auditable, and fair. Tools such as SHapley Additive exPlanations (SHAP) offer partial interpretability by attributing model outputs to input features. However, SHAP alone does not provide comprehensive ethical transparency. It cannot uncover proxy bias, ensure demographic fairness, or validate causal reasoning. Therefore, while SHAP is used in this study to enhance model transparency, it is acknowledged as a necessary but insufficient component of responsible AI deployment.

This study also recognises potential class imbalance in volatility regime labels. In real financial markets, stable (sideways) regimes tend to dominate, while high-volatility (crisis or spike) regimes are relatively rare. This imbalance can bias model learning, affecting both prediction accuracy and interpretability across regimes. Although preprocessing techniques and walk-forward splits are used to mitigate this issue, the impact of label imbalance remains a relevant limitation for generalisability.

Finally, feature selection plays a critical role in the design of effective forecasting models. This study incorporates inter-asset correlation measures and market-wide risk indicators such as the VIX (Volatility Index) and pairwise correlations like AAPL-GSPC. These features were selected due to their strong theoretical and empirical associations with systemic market risk and volatility propagation. Prior research has demonstrated that sudden changes in the VIX or breakdowns in inter-asset correlation structures often precede volatility surges, making them essential inputs for any forward-looking risk model.

To address these interrelated challenges, this thesis proposes a forecasting framework that integrates interpretable machine learning and deep learning models, regime-aware volatility labelling, inter-asset correlation features, walk-forward cross-validation, and SHAP-based diagnostics. The framework is applied across selected equities in the U.S. and Australian markets to investigate forecasting performance, cross-market transferability, and practical feasibility under realistic institutional conditions.

1.3 Research Aim and Objectives

Research Aim

This research aims to evaluate the practical effectiveness, robustness, and interpretability of selected AI-based models specifically LSTM and StockMixer against traditional statistical methods such as ARIMA for short-term volatility forecasting. The study focuses on representative assets from the Australian and U.S. markets and employs realistic walk-forward cross-validation and SHAP explainability to assess each model's utility in financial risk management.

Objectives

- Evaluate and compare the forecasting performance of ARIMA, LSTM, and StockMixer models on selected AU and US financial assets using real-world data.
- Assess each model's interpretability using SHAP, with a focus on global feature importance and a targeted example of local explanation.
- Analyze model behavior under different market regimes using a limited number of walk-forward cross-validation folds.
- Investigate model generalizability between selected Australian and U.S. stocks under similar risk labeling and evaluation frameworks.

The following section outlines the core questions that drive the investigation.

- To examine the volatility forecasting performance of ML/DL models (LSTM, StockMixer) and traditional models (ARIMA) across various financial market conditions.
- To assess forecasting accuracy, interpretability, and ethical transparency using SHAP explanations and performance benchmarks.
- To evaluate model behavior and generalizability using walk-forward cross-validation across distinct market regimes.
- To investigate model transferability across Australian and U.S. financial instruments and regimes.

1.4 Research Questions

This study is guided by the following primary and secondary research questions, which aim to explore the forecasting performance, interpretability, and practical viability of advanced AI-based models compared to conventional statistical methods under realistic financial market conditions.

Primary Research Question

To what extent can hybrid AI-based forecasting models, specifically Long Short-Term Memory (LSTM) networks and StockMixer, enhance short-term volatility prediction accuracy, interpretability, and ethical transparency in financial markets when evaluated under realistic, time-aware simulation frameworks, relative to traditional statistical approaches such as ARIMA?

Secondary Research Questions

1. How does the volatility forecasting performance of LSTM and StockMixer compare with that of ARIMA when applied to selected financial instruments in the Australian and U.S. equity markets?
2. In what ways do forecasting accuracy, model robustness, and regime sensitivity vary across different market conditions, including bull, bear, and sideways volatility regimes?
3. To what extent can SHapley Additive exPlanations (SHAP) contribute to improving the transparency, interpretability, and ethical accountability of AI-driven volatility forecasting models?
4. What insights can be drawn regarding the potential for model generalisability across markets, specifically when exploring whether forecasting models trained on Australian assets exhibit transferable performance characteristics when applied to analogous U.S. assets, and vice versa?

These research questions collectively underpin the methodological framework adopted in this study. They serve to evaluate not only the predictive capabilities of modern forecasting models but also their real-world relevance, interpretability, and cross-market applicability under a structured and institutionally credible simulation environment.

1.5 Significance of the Study

This research makes a dual contribution to both academic scholarship and professional

financial practice by proposing a realistic, interpretable, and context-aware framework for short-term volatility forecasting. Rather than seeking exhaustive coverage of models and asset classes, the study adopts a focused, hypothesis-driven approach using a strategically selected subset of forecasting techniques and financial instruments. This design enables rigorous testing of key concepts such as model explainability, regime sensitivity, and potential cross-market applicability, while ensuring the feasibility and replicability of the research.

By examining assets from both the U.S. and Australian equity markets, the study addresses a gap in the existing literature surrounding cross-national volatility forecasting. While it does not claim universal transferability, it offers preliminary insights into whether forecasting models exhibit stable performance across distinct economic environments and market regimes. This exploratory cross-market lens contributes to an improved understanding of model generalisability, especially in the context of heterogeneous financial systems.

From a practitioner perspective, the proposed framework delivers:

- A risk-sensitive volatility forecasting tool tailored for institutional use, capable of generating confidence-informed predictions under evolving market conditions.
- Explainable outputs using SHAP diagnostics to enhance model transparency, supporting compliance with emerging financial AI governance standards such as the EU AI Act, ASIC guidance, and SEC expectations.
- A flexible feature design incorporating inter-asset and market-wide indicators, supporting use across diverse asset types and sectors without requiring fundamental model redesign.

From an academic perspective, the study contributes:

- A novel integration of SHAP-based interpretability, regime-aware evaluation, and walk-forward cross-validation within a single forecasting pipeline.
- Comparative evidence evaluating traditional statistical models, deep learning architectures, and hybrid approaches under realistic deployment scenarios.
- A reproducible, modular research framework that can be extended to other markets, asset classes, or risk horizons.
- A response to the growing reproducibility and explainability challenges in financial machine learning, through transparent methodology, justifiable feature engineering, and ethical interpretability practices.

Overall, the proposed methodology represents a step toward the development of responsible, transparent, and operationally viable AI-driven tools for volatility forecasting. It bridges the gap between predictive performance and real-world deployment constraints, contributing to both the academic discourse on financial forecasting and the practical advancement of interpretable AI in high-stakes domains.

1.6 Thesis Structure Overview

This thesis is organised into five chapters, each building upon the last to develop, evaluate, and critically reflect on a regime-aware, interpretable framework for short-term volatility forecasting:

- **Chapter 2: Methodology** Describes the overall research design, including data acquisition, preprocessing, feature engineering, volatility labelling, and model selection. It also outlines the walk-forward cross-validation protocol, SHAP-based interpretability strategy, and ethical considerations embedded within the RiskPipeline framework.
- **Chapter 3: Results and Analysis** Presents empirical findings from both regression and classification experiments. This includes comparative performance metrics across models, visual analyses

using SHAP to interpret model behaviour, and a breakdown of forecasting accuracy across different market regimes and asset classes.

- **Chapter 4: Discussion** Interprets the significance of the results in light of the research questions and broader literature. It reflects on methodological choices, evaluates model strengths and weaknesses, discusses ethical implications, and considers practical applications in institutional finance.
- **Chapter 5: Conclusion** Summarises the main contributions of the thesis, revisits the initial objectives, and outlines the theoretical, methodological, and practical implications of the study. It also identifies limitations and proposes future research directions to enhance generalisability, interpretability, and long-term robustness.

Each chapter is designed to progressively develop the case for interpretable, responsible, and empirically grounded volatility forecasting in real-world financial contexts.

Chapter 2: Methodology

2.1 Research Design

This study adopts a quantitative, comparative research design to evaluate the predictive performance and interpretability of various forecasting models for short-term market volatility. The objective is to simulate realistic financial forecasting conditions and assess how well traditional statistical methods and machine learning-based models generalize across distinct market regimes and geographies.

The research follows a dual-task setup consisting of: (1) a regression task to forecast short-term risk as a continuous variable, defined as the 5-day rolling standard deviation of asset log returns (Volatility5D), and (2) a classification task where volatility is categorized into discrete levels (Low, Medium, High) via quantile binning. This dual-task structure provides a comprehensive evaluation of model utility in both continuous and categorical risk contexts.

The study is grounded in historical financial data sourced from U.S. and Australian equity markets, spanning multiple asset types (e.g., index ETFs and individual stocks). Models are trained and validated using walk-forward cross-validation (WFCV), a sequential data-splitting method that emulates real-world deployment by preventing data leakage and preserving the temporal order of financial observations. This design emphasizes practical applicability by aligning evaluation strategies with how forecasting tools would be used in institutional finance.

The overall methodology emphasizes model interpretability, generalizability, and computational feasibility. All experiments are conducted in Python using a modular pipeline (RiskPipeline) implemented within a Windows Subsystem for Linux (WSL) environment on a Dell XPS 15 Signature Edition laptop (Intel i7-8750H CPU, 16GB RAM, NVIDIA GTX 1050 Ti Max-Q). This setup reflects a realistic computational environment for academic research and ensures the reproducibility of all results.

2.2 Data Description

The dataset used in this study comprises historical daily price data for six financial instruments representing both the U.S. and Australian equity markets. The selected assets include three U.S. securities: the S&P 500 Index (GSPC), Apple Inc. (AAPL), and Microsoft Corporation (MSFT), and three Australian assets: the iShares Core S&P/ASX 200 ETF (IOZ.AX), Commonwealth Bank of Australia (CBA.AX), and BHP Group Limited (BHP.AX). This combination ensures exposure across diverse sectors, geographies, and market structures, enabling a balanced assessment of model performance under varying conditions.

The dataset spans the period from January 2017 to March 2024, offering a comprehensive temporal window that includes major global

financial events, such as the COVID-19 market crash, subsequent recovery phases, and inflation-induced volatility episodes. Data was primarily sourced from Yahoo Finance and Quandl, both of which are widely used for high-frequency financial analysis and offer consistent, well-documented historical data.

For the purposes of forecasting, daily adjusted close prices are transformed into log returns, which are then used to compute the primary target variable: the 5-day rolling standard deviation of returns (Volatility5D). This measure is used for the regression task, capturing short-term realised volatility dynamics. For the classification task, volatility values are discretised into three categories Low, Medium, and High using a quantile-based binning approach designed to reflect relative risk conditions.

This study acknowledges that the quantile binning process can still result in class imbalance, especially in rolling time windows where certain market regimes (e.g., low-volatility or sideways periods) dominate. Such imbalance can bias model learning and performance evaluation, particularly if the majority class disproportionately influences the training process. While techniques such as stratified sampling and balanced walk-forward splits are used to mitigate this issue, the potential for volatility regime skew remains a limitation and is carefully monitored throughout the modelling pipeline.

In addition to price-derived features, external market-wide indicators are included to capture broader macro-financial signals. These include the CBOE Volatility Index (VIX) and its daily rate of change (VIX_change), which serve as proxies for global investor sentiment and systemic risk. These macroeconomic features are timestamp-aligned and merged with the primary dataset to preserve temporal fidelity.

The final dataset is cleaned, normalised, and temporally aligned across all assets to ensure structural consistency. Missing values are imputed using forward-filling, and anomalous values are retained where relevant to preserve the integrity of volatility spikes. Data engineering and preprocessing are encapsulated within a modular, version-controlled RiskPipeline class to ensure full reproducibility, facilitate experimentation, and maintain methodological transparency across all evaluation phases.

2.3 Feature Engineering

Feature engineering forms the foundation of effective time series modelling by enabling learning algorithms to capture relevant patterns, structural dynamics, and contextual signals. In this study, the feature engineering process was carefully designed to balance asset-specific characteristics with broader market indicators, supporting both predictive accuracy and interpretability. All feature transformations and calculations were implemented within the custom-built, modular RiskPipeline class, ensuring full reproducibility, scalability, and consistency across experimental runs.

1. Technical and Statistical Features

The core set of asset-specific features was selected based on established practices in financial econometrics and quantitative trading. These include:

- **Lagged log returns (Lag1 to Lag3)**: Capture short-term autocorrelations and mean-reversion effects.
- **Rate of Change over 5 days (ROC5)**: Measures momentum and directional shifts.
- **Moving Averages (MA10 and MA50)**: Identify prevailing market trends over short and medium time frames.
- **Rolling Standard Deviation (RollingStd5)**: A proxy for recent realised volatility.
- **Moving Average Ratio (MA_ratio)**: Computed as MA10 divided by MA50, this ratio signals potential trend shifts or convergence-divergence patterns.

These features are widely used in financial time series prediction due to their ability to capture price momentum, volatility clustering, and short-horizon predictive signals. They also support model generalisability across assets by relying on standardised return-based transformations.

2. Macroeconomic and Systemic Features

To incorporate market-wide risk sentiment and regime awareness, two external features are introduced:

- **CBOE Volatility Index (VIX)**: Represents the market's expectation of near-term volatility in U.S. equities, often referred to as the "fear index".
- **VIX_change**: The daily percentage change in the VIX, z-score standardised to ensure comparability across time and reduce scale sensitivity.

The inclusion of VIX and VIX_change is based on robust empirical evidence linking spikes in the VIX to heightened investor uncertainty and impending volatility surges across both domestic and international markets. These features serve as early warning signals for systemic risk and contribute significantly to model interpretability, particularly when used in conjunction with SHAP explanations.

3. Inter-Asset Relationship Features

Financial markets are inherently interconnected, and cross-asset relationships often contain valuable predictive information. To capture co-movement and potential contagion effects, the study introduces rolling correlation features between key asset pairs:

- AAPL & GSPC correlation
- IOZ & CBA correlation
- BHP & IOZ correlation

These rolling correlations are computed over a 30-day sliding window and updated dynamically throughout the training and evaluation pipeline. Their inclusion is motivated by both academic and practitioner literature, which highlights that breakdowns or spikes in asset correlations often precede volatility clusters or market-wide stress. For instance, a sharp decline in the correlation between a major stock (like Apple) and a broad index (like the S&P 500) may signal decoupling or sector-specific risk.

Including these features enables models to capture latent market structure, improve regime awareness, and simulate how shocks may propagate across sectors or geographies. They also provide interpretable diagnostics when combined with feature attribution methods, supporting institutional use cases such as portfolio stress testing and diversification analysis.

4. Preprocessing and Diagnostic Controls

All features are aligned using consistent timestamp joins across assets, with forward-filling applied to handle non-overlapping trading days and missing entries. NaN values are imputed using forward-fill methods to preserve temporal continuity. Z-score normalisation is applied to all continuous features to ensure that models are not biased by differences in scale. Outliers are flagged using interquartile range and z-score thresholds but are retained during training to preserve volatility realism, particularly around extreme market events.

5. Feature Summary Table

To support transparency and replication, Table 2.1 summarises all engineered features, categorising them by type and forecasting role.

Table 2.1: Summary of Engineered Features

Feature Name	Type	Purpose / Forecasting Role
Lag1, Lag2, Lag3	Technical	Capture recent return autocorrelations and mean reversion tendencies
ROC5	Technical	Measures 5-day momentum and trend direction
MA10, MA50	Technical	Detect short- and medium-term price trends
RollingStd 5	Technical	Proxy for recent realised volatility
MA_ratio	Technical	Identifies trend strength or reversal (MA10 / MA50)
VIX	Macroeconomic	Systemic market risk sentiment (U.S. market proxy)
VIX_change	Macroeconomic	Captures changes in investor sentiment and near-term volatility expectations
AAPL–GSPC corr	Inter-Asset	Detects decoupling between stock and index (U.S. correlation risk)
IOZ–CBA corr	Inter-Asset	Captures financial sector co-movement in the Australian market
BHP–IOZ corr	Inter-Asset	Reflects mining sector alignment with the broader Australian equity market

This multi-level feature engineering strategy ensures that models have access to both micro-level (asset-specific) and macro-level (market-wide) signals. By capturing momentum, volatility, systemic risk, and inter-asset dependencies, the framework supports more accurate, robust, and interpretable volatility forecasts across market regimes and geographies. It also aligns closely with institutional priorities such as explainability, model governance, and cross-market risk generalisation.

2.4 Modeling Approach

This study adopts a multi-task modelling framework to evaluate predictive performance and interpretability across both regression and classification formulations of the short-term volatility forecasting problem. The modelling

strategy is designed to systematically compare traditional statistical approaches, deep learning architectures, and ensemble methods under realistic, temporally consistent evaluation protocols.

2.4.1 Task Overview

Two distinct but complementary forecasting tasks are defined:

- **Regression Task**: Predicting the 5-day realised volatility (Volatility5D) as a continuous variable, with the goal of estimating magnitude and directional risk levels.
- **Classification Task**: Predicting the categorical volatility regime (Low, Medium, High), derived via quantile binning, to assess the model's capacity to differentiate between volatility states.

By modelling volatility from both continuous and discrete perspectives, the study aims to assess how different algorithmic classes perform under varied objectives and evaluation metrics.

2.4.2 Selected Models for Regression

The following models were selected for regression based on their relevance to financial time series literature, suitability for short-term forecasting, and diversity of modelling paradigms:

- **ARIMA (Autoregressive Integrated Moving Average)**
  A foundational econometric model that provides strong interpretability and statistical grounding. ARIMA is used as a baseline to benchmark the effectiveness of more complex models and to demonstrate the limitations of linear assumptions in non-stationary regimes.
- **LSTM (Long Short-Term Memory Network)**
  A deep learning architecture capable of capturing long-term dependencies and non-linear temporal interactions. LSTM models are well-suited to financial time series due to their ability to learn latent patterns across multiple time steps without requiring manual lag specification.
- **StockMixer**
  A recent MLP-based model specifically designed for financial time series, StockMixer uses parallel pathways to mix temporal, indicator, and cross-stock signals efficiently. Its inclusion supports evaluation of performance and interpretability in a low-latency, resource-efficient context (Fan & Shen, 2024).

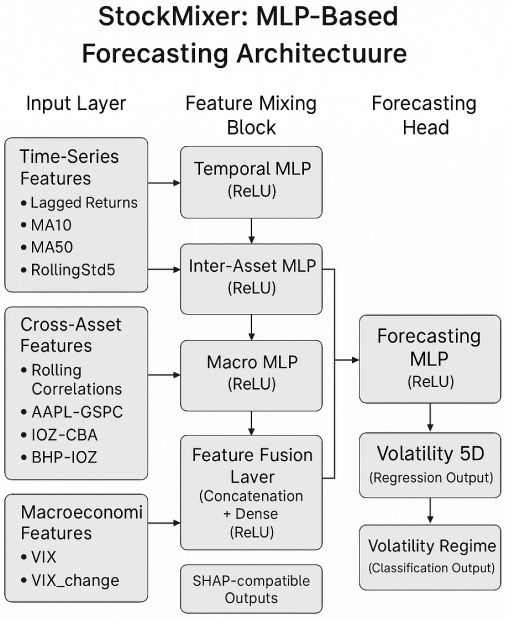

Figure 2.2 presents the proposed StockMixer architecture used in this study, highlighting its three-stage input processing, feature fusion, and dual-task forecasting outputs.

2.4.3 Selected Models for Classification

The classification models reflect a blend of decision-tree ensembles, neural networks, and sequence-based learning:

- **XGBoost (Extreme Gradient Boosting)**
  A high-performance ensemble model known for its effectiveness on structured data. Its robustness to overfitting, built-in regularisation, and ability to model

complex non-linear relationships make it an ideal benchmark for volatility regime classification.

- **Multilayer Perceptron (MLP)** A standard feedforward neural network employed as a deep learning benchmark for tabular classification. Its inclusion provides insight into the marginal benefit of sequence modelling versus static architectures.
- **LSTM Classifier** An adaptation of the LSTM model for categorical output. It allows evaluation of whether sequence-aware models outperform static classifiers in identifying regime transitions, particularly under skewed class distributions.

2.4.4 Baseline Models

To contextualise model performance, two minimal baselines are implemented:

- **Naive Moving Average**: For regression, a simple model that predicts volatility using the previous rolling value. It serves as a realistic lower-bound benchmark reflective of naïve market heuristics.
- **Random Classifier**: For classification, a model that assigns labels at random according to class prior distribution. This establishes a minimum expected accuracy and F1-score.

2.4.5 Implementation and Evaluation Protocol

All models are implemented in Python using industry-standard libraries, including statsmodels, scikit-learn, xgboost, and tensorflow. Hyperparameter tuning is performed using grid search and stratified sampling within a walk-forward cross-validation framework to preserve temporal integrity. For classification tasks, label stratification is applied to address class imbalance in regime labels.

The entire modelling workflow is encapsulated within the modular RiskPipeline framework. This object-oriented structure facilitates reproducibility, experiment tracking, model comparison, and pipeline extensibility across different asset sets, timeframes, and research objectives.

This modelling approach enables a comprehensive and interpretable evaluation of volatility forecasting under both regression and classification formulations. It supports the study's broader aim of benchmarking predictive performance, interpretability (via SHAP), and generalisability across asset classes and volatility regimes.

2.5 Evaluation Strategy

To ensure rigorous and meaningful evaluation of forecasting performance, this study adopts a multi-metric strategy tailored to the dual-task setup. For the regression task, model accuracy is assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). For the classification task, evaluation is based on Accuracy, F1 Score, Precision, and Recall to capture both class-level and overall predictive reliability.

A core component of the evaluation design is the use of walk-forward cross-validation (WFCV), which simulates a real-world deployment setting by training models on rolling windows of past data and testing them on future, unseen periods. This approach prevents data leakage and aligns with best practices in financial forecasting where time-dependent patterns and temporal causality are crucial. The WFCV setup typically includes five sequential splits for each asset, balancing evaluation robustness with computational feasibility.

Additionally, regime-aware analysis is conducted by segmenting test periods into bull, bear, and sideways market regimes based on return slope thresholds. This allows for performance comparison across structurally different market phases, providing insights into

model adaptability and robustness under varying conditions.

Interpretability is assessed using SHapley Additive exPlanations (SHAP), a model-agnostic tool that quantifies the contribution of each feature to individual predictions. Both global and local SHAP analyses are performed. Global SHAP summary plots help identify the most influential features across the entire dataset, while local explanations provide case-specific insights into prediction rationale. SHAP is applied primarily to XGBoost and StockMixer models due to their complexity and relevance in practical applications.

This evaluation strategy ensures that model performance is assessed from both statistical and operational standpoints, reflecting the dual priorities of accuracy and transparency in high-stakes financial environments.

2.6 Ethical Considerations

Ethical considerations are fundamental to the responsible application of artificial intelligence (AI) and machine learning (ML) in financial forecasting. Given the potential influence of predictive models on investment decision-making, portfolio risk management, and regulatory reporting, this study integrates multiple safeguards to ensure transparency, fairness, reproducibility, and alignment with responsible AI development frameworks.

First, the issue of model interpretability is directly addressed through the use of SHapley Additive exPlanations (SHAP). SHAP values enable granular attribution of model outputs to input features, allowing both developers and stakeholders to audit how predictions are formed. This is particularly important in financial domains, where opaque or black-box models risk violating regulatory expectations or contributing to misinformed decisions. However, this study recognises that SHAP, while valuable, does not constitute full ethical transparency. It does not uncover deeper structural issues such as causal misattribution, proxy bias, or the downstream effects of model deployment in dynamic market environments. As such, SHAP is treated as one component within a broader interpretability and auditability framework, rather than a complete ethical solution.

Second, the study examines the ethical implications of feature selection. Features such as the CBOE Volatility Index (VIX) and inter-asset correlation coefficients, while commonly used in practice, are not neutral. For example, the VIX predominantly reflects U.S. investor sentiment and may underrepresent risk dynamics in non-U.S. or emerging markets. Similarly, correlation-based features can inadvertently encode systemic biases, particularly if they reflect short-term co-movements driven by global capital flows or sector-specific anomalies. These risks are acknowledged in the discussion and limitations sections, and future work is encouraged to explore more inclusive and demographically balanced feature sets.

Third, the dataset used in this study consists exclusively of publicly available market data sourced from reputable APIs such as Yahoo Finance and Quandl. No private, personal, or confidential information has been used, and all preprocessing steps are implemented within a version-controlled and reproducible Python pipeline (RiskPipeline). This ensures full traceability, transparency, and reusability of results in accordance with open science and responsible data handling practices.

Finally, this research is conducted strictly within a simulation environment and is intended for academic exploration only. The study does not offer financial advice or prescriptive insights for live investment scenarios. All model forecasts are retrospective and evaluated in out-of-sample, backtested conditions. No claims are made regarding the operational deployment of the models without further validation, stress testing, or regulatory oversight.

Taken together, these safeguards and acknowledgements reflect a commitment to

developing ethical, transparent, and responsible AI models in financial data science. While the models proposed offer technical innovation, their design and interpretation are grounded in an awareness of the broader societal, economic, and regulatory context in which they may eventually be applied.

2.7 Limitations and Assumptions

While this study presents a rigorously structured and interpretable framework for short-term volatility forecasting, several limitations and underlying assumptions must be acknowledged to ensure transparent contextualisation of the findings. These limitations span computational, methodological, and conceptual dimensions, and reflect both practical constraints and design trade-offs inherent in applied financial machine learning research.

1. Computational Constraints

All experiments were conducted on a Dell XPS 15 Signature Edition laptop equipped with an Intel i7-8750H CPU, 16GB RAM, and an NVIDIA GTX 1050 Ti Max-Q GPU. While sufficient for prototyping and experimentation, this environment imposes limitations on the scalability of deep learning architectures, the granularity of walk-forward cross-validation folds, and the breadth of hyperparameter optimisation. As a result, model depth, ensemble scope, and grid search resolution were strategically constrained to maintain experimental feasibility. Future work leveraging high-performance computing environments could explore larger model variants, longer lookback periods, and deeper cross-validation windows for improved performance and robustness.

2. Data Granularity and External Coverage

This study uses daily adjusted closing prices from Yahoo Finance and Quandl, which are widely regarded as reliable and publicly accessible sources. However, this data excludes intraday price dynamics, volatility spikes, and high-frequency patterns that may be informative in ultra-short-term forecasting contexts such as algorithmic or intraday trading. Furthermore, due to inconsistencies in data availability, macroeconomic indicators such as the Australian Volatility Index (AXVI) and industry-specific sentiment scores were excluded. This limits the external interpretability and granularity of macro-financial regime signals, especially for Australian assets.

3. Feature Engineering Assumptions

Feature transformations, including rolling averages, lagged returns, and inter-asset correlations, rely on fixed window sizes (e.g., 5-day, 30-day). While these choices are grounded in empirical conventions and literature, they may not be optimal across all market conditions or assets. Market volatility is inherently heteroskedastic, and the ideal lookback horizon may vary by regime, sector, or macroeconomic context. Similarly, extreme market events are retained in the dataset to preserve volatility realism, but they may disproportionately influence learning dynamics in models sensitive to scale or outliers.

4. Regime Labeling and Target Construction

Market regime labels (bull, bear, sideways) were assigned using slope-based heuristics applied to historical return windows. While this approach is computationally tractable and widely used in prior studies, it simplifies the multidimensional nature of market dynamics, which are influenced by sentiment, liquidity, monetary policy, and behavioural factors. As such, the regime classification framework should be interpreted as an approximate segmentation rather than a definitive market taxonomy. In future work, regime inference could be enhanced using probabilistic or unsupervised learning techniques.

5. Generalisability and Transferability

The study evaluates model performance on a selected basket of six large-cap equities three

from the U.S. and three from the Australian market. Although chosen to represent a balance of geographies and sectors, this limited sample may not generalise to smaller-cap stocks, fixed income instruments, cryptocurrencies, or other geographies. Additionally, while the study explores cross-market performance, it does not claim broad transferability. Instead, any evidence of transferability between Australian and U.S. assets should be interpreted as exploratory and context-dependent, rather than conclusive.

6. Concept Drift and Stability Over Time

Financial markets are non-stationary environments subject to structural change, policy shifts, and evolving investor behaviour. As such, models trained on historical data may experience concept drift, where the relationships between features and target variables change over time. Although walk-forward cross-validation is employed to simulate live deployment, it does not fully guarantee robustness against long-term drift or exogenous shocks. Periodic retraining, online learning, or adaptive recalibration mechanisms would be necessary in production settings.

Despite these limitations, this study has made every effort to ensure methodological transparency, computational reproducibility, and practical relevance. Design decisions are explicitly documented, ethical risks are acknowledged, and results are presented within the appropriate contextual bounds. These limitations do not detract from the core contributions of the study but rather serve to delineate the scope and inform avenues for future improvement.

References

Albeladi, K., Zafar, B., & Mueen, A. (2023). Time series forecasting using LSTM and ARIMA. The Science and Information (SAI) Organisation. https://thesai.org/Downloads/Volume14No1/Paper_33-Time_Series_Forecasting_using_LSTM_and_ARIMA.pdf

Antulov-Fantulin, N., Cauderan, A., & Kolm, P. N. (2024). A dynamic regime-switching model using gated recurrent straight-through units. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4810879

Arsenault, P.-D., Wang, S., & Patenaude, J.-M. (2024). A survey of Explainable Artificial Intelligence (XAI) in financial time series forecasting. arXiv. https://arxiv.org/html/2407.15909v1

Bieganowski, B., & Slepaczuk, R. (2024). Supervised autoencoder MLP for financial time series forecasting. arXiv. https://arxiv.org/abs/2404.01866

Christensen, K., Siggaard, M., & Veliyev, B. (2022). A machine learning approach to volatility forecasting. ResearchGate. https://www.researchgate.net/publication/363007775_A_Machine_Learning_Approach_to_Volatility_Forecasting

Fan, J., & Shen, Y. (2024). StockMixer: A simple yet strong MLP-based architecture for stock price forecasting. In Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24). https://ojs.aaai.org/index.php/AAAI/article/download/28681/29322

Gajamannage, K., & Park, Y. (2022). Real-time forecasting of time series in financial markets using sequentially trained many-to-one LSTMs. arXiv. https://arxiv.org/abs/2205.04678

Huang, X., You, P., Gao, X., & Cheng, D. (2023). Stock price prediction based on ARIMA-GARCH and LSTM. Atlantis Press. https://www.atlantis-press.com/

Kortian, T., & O'Regan, J. (1996). Australian financial market volatility: An exploration of cross-country and cross-market linkages.

Reserve Bank of Australia. https://www.rba.gov.au/publications/rdp/1996/9609/bond-share-and-foreign-exchange-markets-descriptive-statistics-and-correlations.html

Liu, J. (2023). A hybrid model integrating LSTM with multiple GARCH-type models for volatility and VaR forecast. EAI Endorsed Transactions. https://eudl.eu/pdf/10.4108/eai.6-1-2023.2330313

Moreno-Pino, F., & Zohren, S. (2024). DeepVol: Volatility forecasting from high-frequency data with dilated causal convolutions. arXiv. https://arxiv.org/abs/2210.04797

Rahimikia, E., & Poon, S.-H. (2020). Machine learning for realised volatility forecasting. ResearchGate. https://www.researchgate.net/publication/344634525_Machine_Learning_for_Realised_Volatility_Forecasting

Reisenhofer, R., Bayer, X., & Hautsch, N. (2022). HARNet: A Convolutional Neural Network for realized volatility forecasting. arXiv. https://arxiv.org/abs/2205.07719

Salih, A., et al. (2024). A perspective on explainable artificial intelligence methods: SHAP and LIME. arXiv. https://arxiv.org/abs/2305.02012

Shi, Z., Hu, Y., Mo, G., & Wu, J. (2023). Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction. arXiv. https://arxiv.org/abs/2204.02623

Siami-Namini, S., & Siami-Namin, A. (2018). Forecasting economics and financial time series: ARIMA vs. LSTM. arXiv. https://arxiv.org/abs/1803.06386

Thakur, N., & Sharma, A. (2024). Ethical considerations in AI-driven financial decision making. Journal of Management and Public Policy. https://jmpp.in/wp-content/uploads/2024/06/Neha-Thakur-and-Aryan-Sharma.pdf

Zhang, C., et al. (2023). Volatility forecasting with machine learning and intraday commonality. arXiv. https://arxiv.org/abs/2202.08962

Zhao, P., et al. (2024). From GARCH to neural network for volatility forecast. arXiv. https://arxiv.org/abs/2402.06642

Appendix

Glossary of Terms

Term	Definition
AI (Artificial Intelligence)	The field of computer science focused on creating systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, and pattern recognition.
ARIMA (Autoregressive Integrated Moving Average)	A classical time series forecasting model that combines autoregression, differencing (integration), and moving average components to model linear temporal relationships.
Asset	A financial instrument, such as a stock, ETF, or index, that represents ownership or a claim to future returns.

Backtesting	A method of evaluating a model by testing it on historical data to simulate its performance in a real-world scenario.
Classification Task	A machine learning task where the goal is to predict discrete labels or categories (e.g., high, medium, or low volatility regimes).
Concept Drift	A phenomenon where the statistical relationships between input features and target variables change over time, often reducing model accuracy in dynamic environments.
Cross-Asset Correlation	A measure of the degree to which two different financial assets move in relation to each other over time.
Deep Learning (DL)	A subset of machine learning that uses neural networks with multiple layers to model complex patterns in high-dimensional data.
ETF (Exchange-Traded Fund)	A type of investment fund traded on stock exchanges, holding a diversified portfolio of assets.
Feature Engineering	The process of transforming raw data into meaningful input variables (features) that improve model performance and interpretability.
LSTM (Long Short-Term Memory)	A type of recurrent neural network designed to capture long-range dependencies in sequential data, widely used in time series forecasting.
MA (Moving Average)	A smoothing technique that averages past prices or returns over a fixed time window to highlight trends.
MA_ratio	The ratio of short-term to long-term moving averages, used to indicate trend strength or reversal signals.
ML (Machine Learning)	A branch of AI focused on developing algorithms that improve their performance on a task through data-driven learning.
Model Interpretability	The degree to which a human can understand the cause of a decision made by a model.

Quantile Binning	A technique to transform continuous variables into categorical bins based on percentile thresholds.
Regression Task	A machine learning task focused on predicting continuous numerical values (e.g., predicted volatility levels).
SHAP (SHapley Additive exPlanations)	A model-agnostic interpretability technique that explains model outputs by attributing importance scores to each feature based on cooperative game theory.
StockMixer	A recent MLP-based deep learning architecture designed to mix temporal, cross-indicator, and inter-stock information efficiently in financial forecasting tasks.
Time Series	A sequence of data points indexed in chronological order, often used for forecasting in economics and finance.
Volatility	A statistical measure of the dispersion of returns for a given asset, often used as a proxy for financial risk.
VIX (Volatility Index)	A popular measure of the market's expectation of near-term volatility, calculated from S&P 500 index options.
Walk-Forward Cross-Validation (WFCV)	A model evaluation technique where training and validation sets are created sequentially to simulate real-time deployment and preserve time order.
Z-score Standardisation	A method of rescaling features by subtracting the mean and dividing by the standard deviation, resulting in a standard normal distribution (mean 0, std 1).