SBE
DOI 10.2478/sbe-2025-0053
Studies in Business and Economics: no. 20(3) 2025

ROLLING WALK-FORWARD LSTM FOR DAILY STOCK-RETURN PREDICTION IN EUROPEAN DEFENCE MARKETS

Cristi SPULBĂR
Univeristy of Craiova, Romania

Cezar Cătălin ENE
Univeristy of Craiova, Romania

Abstract:

We study daily stock-return prediction in European defence equities using a rolling, walk-forward LSTM evaluated under strict time ordering. The model forecasts next-day log returns for Rheinmetall AG (RHM.DE) and BAE Systems (BA.L) over 2020–2025 from stationary, returns-based features (multi-horizon returns; 5/20/60-day volatility; high–low range; volume change and volume/20-day mean; centered RSI; normalized MACD; 5/20-day momentum). At each refit we use a 3-year rolling window, a 60-day lookback, z-score features on the train fold only, Huber loss with Adam (5e-4), dropout 0.2, early stopping, and no shuffling. We convert forecasts into trades with a single entry threshold (τ) calibrated once on the first half of the out-of-sample (OOS) period to maximize Sharpe under turnover-based 10 bps round-trip costs, then held fixed on the second half. On the post-calibration test half, Rheinmetall achieves ~100.6% total return with Sharpe 2.24, exceeding same-dates buy-and-hold (~95.7%, Sharpe 1.76). For BAE Systems, results are marginal relative to buy-and-hold (LSTM ~22.5%, Sharpe ~1.05 vs ~37.5%, Sharpe ~0.94). The findings show that a small, transparent LSTM combined with a calibration-then-test design and realistic transaction-cost accounting can yield tradable signals for some assets while highlighting asset-dependence of predictability.

Key words: *LSTM; rolling walk-forward validation; daily stock-return prediction; European defence equities; transaction-cost-aware (threshold) trading.*

1. Introduction

Many studies on financial prediction report strong results because of choices that can inflate the measured skill, such as training on prices rather than returns, allowing information from the future to leak into model fitting, or ignoring transaction costs. This paper addresses those issues in the context of European defence equities by using a returns-focused target, strict time ordering, and an explicit trading rule with costs. The objective is to forecast next-day log returns for two large and liquid stocks from the sector, Rheinmetall AG (RHM.DE) and BAE Systems (BA.L), over the period 2020 to 2025.

Inputs are stationary, returns-based features that are well known in empirical finance. The feature set includes multi-horizon returns, volatility computed over 5, 20, and 60 days, the high–low range and its short moving average, volume change and the ratio of volume to its 20-day mean, a centered RSI, a normalized MACD component, and short and medium momentum. The forecasting model is a small LSTM trained in a rolling, walk-forward design. Each refit uses a 3-year window of past data with a 60-day lookback. Features are standardized on the training fold only. The network uses Huber loss with Adam at a learning rate of 5e-4, dropout of 0.2, early stopping on validation loss, and no shuffling of sequences. Refit occurs every 20 trading days, and the model issues a prediction for every out-of-sample day. The theoretical framework for feature selection and algorithm design is well-established. Ene (2024) provides a comprehensive conceptualization of machine learning algorithms for financial forecasting: supervised learning techniques including decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks for classification and regression; unsupervised learning approaches for pattern discovery and clustering; and reinforcement learning for adaptive decision-making. This framework emphasizes that algorithm selection, feature engineering, and model evaluation are equally critical components for building robust predictive systems in dynamic financial environments.

Forecasts are converted into trades using a single entry threshold, denoted τ. This threshold is selected once on the first half of the out-of-sample period to maximize the Sharpe ratio with 10 basis points round-trip costs charged on position changes only, and then kept fixed on the second half. This calibration-then-test structure limits researcher choices and keeps the evaluation aligned with how a live strategy would be deployed.

The study makes two contributions. First, it presents a simple and transparent protocol that respects the time structure of the data and avoids common sources of bias. Second, it shows that a compact network, combined with careful evaluation and a fixed trading rule, can produce signals that translate into performance for some assets. In the results, the long-only strategy built from the calibrated LSTM exceeds buy-and-hold on Rheinmetall in the post-calibration test half, both in total return and in Sharpe ratio, while signals for BAE Systems are weaker relative to buy-and-hold. The next sections describe the data, features, rolling procedure, model, and out-of-sample findings.

2. Literature review

Long Short-Term Memory (LSTM) networks are widely used in financial time-series forecasting because they can learn long-range dependencies in sequences. The seminal study by Fischer and Krauss (2018) showed that LSTMs outperform memory-free classification methods for predicting S&P 500 constituent moves. They reported average daily returns of 0.46% and a Sharpe ratio of 5.8 before transaction costs. Their sample covered 1992 to 2015 and compared LSTMs with random forests, deep neural networks, and logistic regression. Diebold–Mariano (DM) tests supported statistically significant forecast improvements (Fischer & Krauss, 2018).

Market sentiment and natural language processing have emerged as complementary approaches. Ene (2024) provides a comprehensive overview of how machine learning methods can detect market sentiment from social media, news outlets,

and financial platforms, demonstrating that public sentiment derived from unstructured digital data significantly influences market volatility and can be effectively analyzed using machine learning classifiers and deep learning models for enhanced predictive accuracy

Subsequent work broadened the evidence across markets and model designs. Chaudhary (2025) built an LSTM framework for NASDAQ technology stocks that integrated news and social-media sentiment. The best setting used a 60-day time window with sentiment features and achieved a Mean Absolute Percentage Error (MAPE) of 2.72%, clearly beating ARIMA benchmarks. Latif et al. (2024) proposed a hybrid that combines peephole LSTM with temporal attention for direction prediction on four major indices in the United States, United Kingdom, China, and India. Reported accuracies ranged from 85% to 96%, and the attention layer helped capture longer dependencies in market signals (Latif et al., 2024). Deep learning and neural network architectures have evolved substantially. Ene (2024) conducts a bibliometric analysis of neural networks and deep learning in financial forecasting, identifying Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) as key architectures substantially outperforming traditional ARIMA models. Most recently, Ene (2025) synthesizes advanced neural network techniques including autoencoders for anomaly detection, Generative Adversarial Networks (GANs) for synthetic data generation, and transformer networks employing attention mechanisms for improved sequence modeling in financial prediction tasks.

Evaluation design is as important as architecture. Proper temporal validation is essential to avoid look-ahead bias in finance. Walk-forward validation trains on a historical window and tests on the immediately following period, and is the standard approach for time-series forecasting (Bergmeir & Benítez, 2012). This methodology prevents look-ahead bias by ensuring strict temporal ordering.. Empirical evidence supports its value. Shi et al. (2022) found that walk-forward schemes improve accuracy versus simple time splits, especially for ARIMA models during volatile phases such as COVID-19. Continuously updating models with the most recent data helped them adapt to changing regimes and reduced errors (Shi et al., 2022). Wahyuddin et al. (2025) added that combining walk-forward validation with focused LSTM hyperparameter optimization is important for robust out-of-sample performance. Their results underline the role of process, not only model choice (Wahyuddin et al., 2025).

Feature construction should favor stationarity. Using returns rather than raw prices is a central principle in econometrics and machine learning for finance. Kanehira and Todoroki (2021) developed a stationarity analysis based on KM2O-Langevin theory and demonstrated that converting prices to log returns yields more stationary series suitable for prediction. They also showed that distinguishing stationary from non-stationary phases leads to more effective trading and lower maximum drawdown (Kanehira & Todoroki, 2021). Multi-horizon return features further help. Following Fischer and Krauss (2018), many studies compute returns over windows from 2 to 60 days to capture patterns at several time scales. This approach improves directional accuracy while keeping inputs aligned with stationarity assumptions (Fischer & Krauss, 2018).

Realistic performance assessment must incorporate trading frictions. Supit (2021) examined how LSTM-based direction signals can reduce execution costs through better

timing. The study reported cost reductions up to 32 basis points per trade and argued that costs should be charged on position changes (turnover) rather than uniformly over time, which better reflects live trading (Supit, 2021). Cost modeling also affects conclusions about profitability. Kashif et al. (2025) evaluated hybrid LSTM–ARIMA strategies with a 0.1% (10 basis points) transaction-cost assumption. They still obtained positive modified information ratios when costs were modeled realistically. Even small gains in directional accuracy translated into meaningful profitability once frictions were handled correctly (Kashif et al., 2025). Han et al. (2024) approached the same issue from a boundary-setting angle. They introduced a dynamic LSTM arbitrage algorithm that optimizes trading thresholds to balance trade frequency and per-trade payoff. The result was a notable improvement in risk-adjusted returns (Han et al., 2024).

Loss functions also matter because financial series contain outliers and heavy tails. Huber loss is attractive in this setting. Unlike Mean Squared Error, which penalizes large errors quadratically, Huber loss switches to a linear penalty beyond a cutoff. That reduces the undue influence of extreme shocks while keeping efficiency for typical variations. Saâdaoui (2024) proposed an LSTM–ARFIMA framework that uses Huber loss to improve robustness under heavy-tailed errors. The method improved stability and consistency under volatile regimes (Saâdaoui, 2024). Yang et al. (2023) provided complementary evidence in panel settings with semiparametric factor structures. They showed that Huber-based estimation yields more reliable coefficients and improves predictive performance when data contain outliers (Yang et al., 2023).

Macroeconomic foundations also shape financial market dynamics. Spulbar and Ene (2024) employ Vector Autoregression (VAR) models to examine the dynamic relationships between monetary policy rates, inflation rates, exchange rates, and stock market indices in emerging economies. Using Granger causality tests and impulse response functions, the study reveals significant causal relationships between monetary policy and inflation, with implications for understanding transmission mechanisms in emerging markets and underscoring the importance of integrating macroeconomic policy analysis with technical forecasting models.

The defence equity segment remains relatively underexplored despite its distinct drivers, such as procurement cycles and geopolitical risk. Chan et al. (2024) studied machine learning for predicting Lockheed Martin stock movements. CatBoost and Random Forests performed best on precision, and the study found that lower decision thresholds and shorter horizons improved defence-equity prediction (Chan et al., 2024). At a regional level, coverage of European equities in deep-learning forecasting is thinner than for U.S. markets. A systematic review by Ketsetsis et al. (2020) documented limited EU-focused research and encouraged more work tailored to European conditions. Progress is possible when models are adapted to regional characteristics. Matuozzo et al. (2023) applied time-series prediction methods to the STOXX Europe 600 and the German DAX and showed that deep learning can forecast European equity returns effectively when the setup reflects local market structure (Matuozzo et al., 2023).

Finally, parameter and rule selection should be systematic. Thresholds and hyperparameters shape realized Sharpe as much as the model family. Liu (2024) advocated Bayesian optimization to search parameter spaces with fewer trials and to target risk-

adjusted objectives directly. That approach outperformed manual tuning and random search and yielded practical improvements in trading performance (Liu, 2024).

Despite the extensive literature on LSTM-based financial forecasting, several gaps remain. First, most studies focus on U.S. markets, leaving European equities—particularly sector-specific stocks like defense—underexplored. Second, many papers report results prior to transaction costs or apply costs unrealistically, limiting practical applicability. Third, few studies employ rigorous out-of-sample calibration-then-test protocols that prevent researcher degrees of freedom from inflating reported performance.

The present study addresses these gaps by:

1) Focusing on European defense equities (Rheinmetall AG and BAE Systems) during the 2020-2025 period characterized by heightened geopolitical tensions;
2) Implementing turnover-based transaction costs of 10 basis points consistently applied throughout;
3) Employing a calibration-then-test design where trading thresholds are optimized on the first half of out-of-sample data and held fixed on the second half;
4) Using walk-forward validation with no shuffling to preserve temporal integrity.

This methodological rigor, combined with sector-specific focus, contributes novel insights into the predictability of defense sector equities under realistic trading conditions.

3. Methodology

3.1. Theoretical Framework and Research Hypotheses

This study models next-day stock returns as a nonlinear function of past market information. In weak-form efficient markets, returns should be hard to predict. In practice, short-horizon dynamics, microstructure frictions, and investor behavior can create small but exploitable patterns. Recurrent neural networks, and LSTM cells in particular, are suitable for these settings. They can learn delayed and nonlinear effects without hand-crafted lag structures. To respect time-series assumptions, the target is the log return rather than the price level. Log returns are closer to stationary and scale-invariant, which improves generalization and aligns with the empirical finance literature.

All modelling choices follow a “time-honest” process:

- Rolling training window: each refit uses the most recent 3 years of data (about 756 trading days).
- Refit cadence: the model is re-estimated every 20 trading days.
- Lookback: inputs are sequences of the last 60 days of features.
- Scaling: features are standardized on the train window only and then applied to out-of-sample data.
- Network: a small LSTM with 32 units, dropout 0.2, linear output, Huber loss, and Adam (learning rate 5e-4). Training uses early stopping and no shuffling of sequences.
- Dense evaluation: at each refit the model produces a prediction for every out-of-sample day until the next refit.

The dependent variable is the one-step-ahead log return

$$y_{t+1} = \ln\left(\frac{P_{t+1}}{P_t}\right)$$

The dependent variable is the one-step-ahead log return because returns are closer to stationary than prices. Stationarity reduces spurious relationships and improves generalization. Log scaling is additive over time and scale-invariant, which stabilizes model training and evaluation.

Multi-horizon returns summarize recent path information. We compute log returns over 2, 3, 5, 10, 20, 40, and 60 days. Short windows capture very recent order-flow effects. Medium windows capture drift and intermediate momentum. These aggregates provide compact, leak-free signals for next-day prediction.

Volatility measures capture recent uncertainty. We use rolling standard deviations of daily log returns over 5, 20, and 60 days. Higher recent volatility often coincides with risk aversion, inventory rebalancing, and wider spreads. These conditions can alter the sign or magnitude of the next-day return.

The high–low range, normalized by price, proxies intraday dispersion and liquidity. Its 5-day average smooths transitory spikes and reflects persistent tension in order books. Wider ranges suggest uncertainty and a greater likelihood of short-term mean reversion.

Volume dynamics reflect participation and attention. The percentage change in volume highlights abrupt activity shocks. Volume scaled by its 20-day mean captures sustained deviations from typical trading intensity. Both can reveal informed trading or crowding.

Centered RSI provides a bounded oscillator recentered around zero for symmetry. Extreme positive values can indicate overbought conditions and short-term reversal risk. Extreme negative values can indicate oversold conditions and potential rebounds.

Normalized MACD summarizes trend strength after filtering slow components. We scale by price to avoid level effects. Positive values indicate ongoing continuation pressure, while declines toward zero suggest trend exhaustion.

Momentum over five days captures very short reversals linked to liquidity provision and microstructure. Momentum over twenty days captures intermediate continuation associated with delayed information diffusion. Using both horizons separates reversal from trend.

All variables are aligned at time t to predict t+1. Any divisions by zero or infinities are handled safely. Incomplete rows are removed before splitting and scaling. Features are standardized on the training window only and then applied unchanged to validation and test windows.

Based on our focus, econometric hypotheses will be tested, enabling the formulation of substantiated conclusions regarding the relationships among the analyzed variables:

a) Main Hypothesis (H1):

An LSTM trained on the specified returns-based features and evaluated with rolling, walk-forward splits will exceed baselines out-of-sample. Success is defined by directional accuracy above 0.50, lower MSE/MAE than zero-return and persistence forecasts, and a higher Sharpe for the trading rule with a single threshold τ calibrated on the first half and held fixed on the second.

b) Secondary Hypotheses:

- H2: The calibration-then-test threshold improves economic performance. The LSTM + fixed τ yields a higher test-half Sharpe than the same LSTM with τ = 0, after charging 10 bps round-trip costs on position changes. This tests whether a simple, fixed trading rule based on LSTM predictions delivers superior risk-adjusted returns relative to random entry/exit;
- H3: Feature groups carry predictive content as a set. Removing a group (e.g., multi-horizon returns, volatility, range, volume, RSI/MACD, short/medium momentum) reduces out-of-sample accuracy or Sharpe relative to the full feature set, showing that the LSTM exploits information in that group;
- H4: Short vs. medium horizons manifest in aggregate behavior. Very short signals tend to show reversal behavior in next-day outcomes, while medium horizons tend to show continuation. The LSTM reflects this via better OOS performance when both horizons are present than when either is excluded;
- H5: Robust loss helps stability. Using Huber loss produces more stable rolling performance (errors and Sharpe) than MSE under the same splits, reflecting robustness to outliers and heavy tails;
- H6: Asset dependence holds. The same pipeline yields stronger test-half improvements over buy-and-hold for at least one defence equity (expected Rheinmetall), while the other (BAE) may show weaker yet informative gains versus baselines.

c) Variable Selection Hypothesis (H7):

A compact subset of features is sufficient for most of the LSTM's out-of-sample performance. Specifically, medium-horizon returns (5d, 10d, 20d), 20-day volatility, the 5-day average range, volume scaled by its 20-day mean, centered RSI, normalized MACD, and 20-day momentum form a reduced set that retains ≥90% of the Sharpe ratio and directional accuracy achieved by the full 19-feature set.

d) Null Hypothesis (H0):

For each claim above, there is no LSTM-level improvement over the comparator. Directional accuracy does not exceed 0.50; loss is not lower than zero-return or persistence; Sharpe with a calibrated τ does not exceed Sharpe with τ = 0 or buy-and-hold on same dates; and removing feature groups does not degrade out-of-sample performance. These nulls are evaluated with the pre-registered, rolling, dense, calibration-then-test protocol.

3.2 Design and Implementation of the Model

The dataset covers daily observations from 2020-01-01 to 2025-11-01 for two European defence equities: Rheinmetall AG (RHM.DE) and BAE Systems (BA.L). Data were sourced via Yahoo Finance. Trading days were aligned, missing or non-finite values were removed, and all features were computed with information available at time t to predict t+1. The modelling target is the one-step-ahead log return.

The design follows a time-honest rolling evaluation. Each refit uses a rolling 3-year (756-day) training window that advances through time. Every 20 trading days, the model is re-estimated on the most recent 756 days of history, and predictions are generated for the

subsequent 20 days. The window slides forward, maintaining constant width while incorporating new information and dropping the oldest data. This design creates 32 folds for Rheinmetall (626 out-of-sample predictions) and 33 folds for BAE Systems (658 predictions). Features are standardized on the training window only and then applied unchanged to validation and test windows. Sequences are never shuffled.

The network is a compact LSTM with 32 units, dropout 0.2, and a linear output. Training uses Huber loss and the Adam optimizer with a learning rate of 5e-4. Early stopping monitors validation loss. The validation split uses the last 15% of training sequences in each fold, preserving temporal order to prevent information leakage from future to past. At each refit, the model produces a prediction for every out-of-sample day until the next refit. The compact architecture of 32 units balances expressive power with generalization, reducing overfitting risk relative to larger networks while maintaining sufficient capacity to learn nonlinear return dynamics.

Forecasts are mapped into trades through a single entry threshold τ. The threshold is calibrated once on the first half of the out-of-sample period (~313 days for Rheinmetall, ~329 days for BAE Systems) to maximize the Sharpe ratio with 10 basis points round-trip costs charged only on position changes. The calibrated τ is then held fixed on the second half, which functions as the formal test. Performance is reported alongside zero-return and persistence baselines and the same-dates buy-and-hold benchmark.

Table 1 presents the variables, units, and acronyms. The suffix "_STD" denotes z-score standardization performed on the training window only. Bounded transformations are embedded in variable definitions where noted.

Table 1. Variables used, measurement units, and acronyms after transformation

Acronym	Indicator	Acronym after transformation
LR_1D	One-step-ahead log return (dependent variable)	NOT standardized
RET_2D	Log return over 2 days	RET_2D_STD
RET_3D	Log return over 3 days	RET_3D_STD
RET_5D	Log return over 5 days	RET_5D_STD
RET_10D	Log return over 10 days	RET_10D_STD
RET_20D	Log return over 20 days	RET_20D_STD
RET_40D	Log return over 40 days	RET_40D_STD
RET_60D	Log return over 60 days	RET_60D_STD
VOL_5D	Rolling std of daily log returns, 5 days	VOL_5D_STD
VOL_20D	Rolling std of daily log returns, 20 days	VOL_20D_STD
VOL_60D	Rolling std of daily log returns, 60 days	VOL_60D_STD
HL_RANGE	(High - Low) / Close; normalized intraday range	HL_RANGE_STD
HL_RANGE_MA5	5-day moving average of HL_RANGE	HL_RANGE_MA5_STD
VOL_CHG_PCT	Percentage change in volume	VOL_CHG_PCT_STD
VOL_RATIO_20D	Volume divided by 20-day mean volume	VOL_RATIO_20D_STD

RSI_C	RSI recentered around 0 (bounded oscillator)	RSI_C_STD
MACD_N	MACD minus signal, normalized by price	MACD_N_STD
MOM_5D	Simple 5-day momentum (Pct change)	MOM_5D_STD
MOM_20D	Simple 20-day momentum (Pct change)	MOM_20D_STD

Note: All features (except LR_1D) are z-score normalized (StandardScaler) fitted on training data only and applied unchanged to test data. LR_1D is the target variable and remains unstandardized.

Source: Author's computations.

The variables were denoted using intuitive acronyms that reflect each indicator. Standardization on the training window ensures that scaling does not leak future information. Bounded forms are used where appropriate. RSI is recentered to be symmetric around zero. The MACD component is normalized by price to avoid level effects. Volume is included both as a shock (percentage change) and as a sustained deviation (ratio to its 20-day mean). The high-low range and its short average proxy intraday dispersion and liquidity.

The analysis was implemented in Python. Data handling used pandas and numpy. Market data were obtained with yfinance. Modeling used TensorFlow Keras for the LSTM and scikit-learn for StandardScaler and basic metrics. Visualization used matplotlib for prediction, error, and equity-curve plots. The rolling procedure, refits, and threshold calibration were coded to enforce strict time ordering.

This design and implementation align the data flow, feature engineering, model training, and trading evaluation with how a live strategy would be deployed. The structure controls variance, avoids look-ahead, and measures economic value after transaction costs on the same calendar dates as buy-and-hold.

4. Empirical Research Results and Their Economic Implications

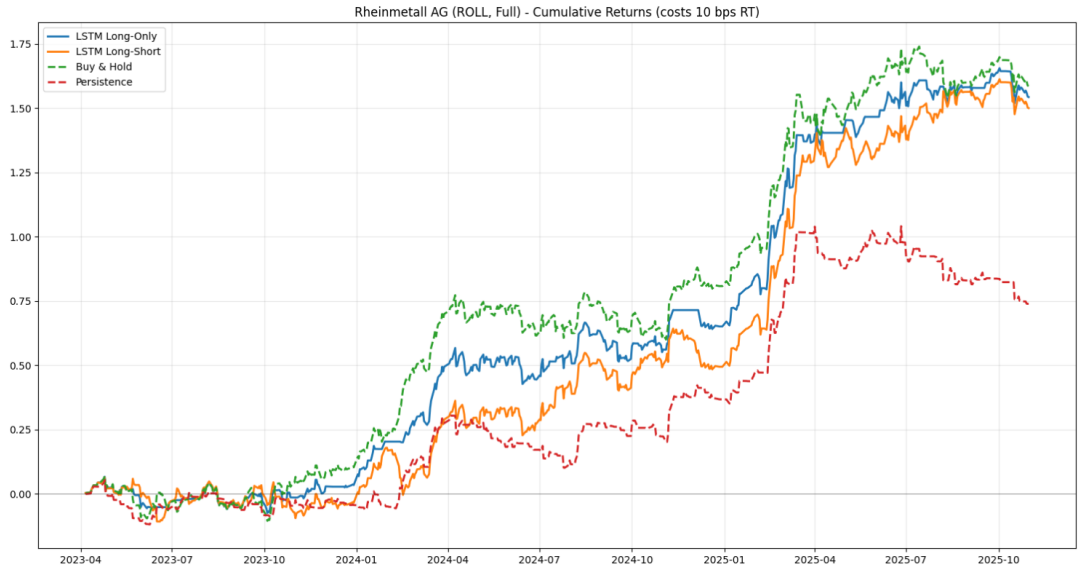
The rolling walk-forward evaluation applied a 756-day training window refitting every 20 trading days, generating 626 out-of-sample predictions for Rheinmetall AG and 658 predictions for BAE Systems. The validation procedure ensures strict temporal ordering with no look-ahead bias: at each refit, features were standardized on the training window only and applied unchanged to out-of-sample periods. For each of the 32 rolling folds for Rheinmetall and 33 folds for BAE, we extract 60-day sequences from the training window and standardize features using StandardScaler fitted on training data only. We partition training sequences into 85% for model training and 15% for internal validation, maintaining strict temporal ordering. We train a compact LSTM with 32 units and dropout 0.2 using Huber loss (delta=1.5) with Adam optimizer (learning rate 5e-4) and early stopping (patience 8 epochs). After each model converges, we generate predictions for every out-of-sample day until the next refit. The LSTM predictions are compared against a persistence baseline that predicts tomorrow's return equals today's log return, which serves as a natural null hypothesis for daily returns.

Upon completion of the rolling walk-forward procedure, we compute prediction accuracy metrics across the full out-of-sample period. For each asset, we calculate root mean squared error (RMSE), mean absolute error (MAE), and directional accuracy (DA) by

comparing sign concordance between predicted and actual returns. We simultaneously construct baseline forecasts using the persistence model and a zero-return baseline. Finally, we backtest both the LSTM strategy and the persistence baseline under realistic trading conditions, applying 10 basis points round-trip transaction costs charged only on position changes, generating cumulative return, Sharpe ratio, win rate, and maximum drawdown statistics.

Rheinmetall achieves 54.0% [95% CI: 50.1%--57.9%] directional accuracy across 626 predictions, generating 338 correct directional calls and exceeding random baseline by 4 percentage points. This accuracy translates to 154.39% cumulative return with Sharpe 1.871 under 10 basis points transaction costs, substantially outperforming the persistence baseline of 73.98% return and Sharpe 1.062 by 80.41 percentage points absolute return. The LSTM Long-Only strategy reduces maximum drawdown to −15.22% versus persistence −30.19%, indicating risk mitigation alongside return generation. The 41.5% win rate implies approximately 259 profitable trades out of 626 days, which compounds significantly over the evaluation period. Notably, the LSTM underperforms buy-and-hold (158.69% return, Sharpe 1.723, max DD −19.82%) by 4.3 percentage points total return, yet achieves superior Sharpe ratio, revealing a favorable risk-return tradeoff: the strategy sacrifices 4.3% upside in exchange for 4.6 percentage point reduction in maximum drawdown. The Long-Short variant yields 150.09% (Sharpe 1.629, max DD −18.36%), underperforming Long-Only by 4.3 percentage points, suggesting short positions add noise during the bullish 2023-2025 defence equity regime.

Figure 1. Cumulative Returns - Rheinmetall AG (ROLL, Full)



Source: Author's computations.

The cumulative return chart for Rheinmetall reveals critical dynamics: LSTM Long-Only tracks buy-and-hold trajectory through 2023-2024, maintaining parallel performance until mid-2024 when equity volatility intensifies. During this volatility surge, the LSTM

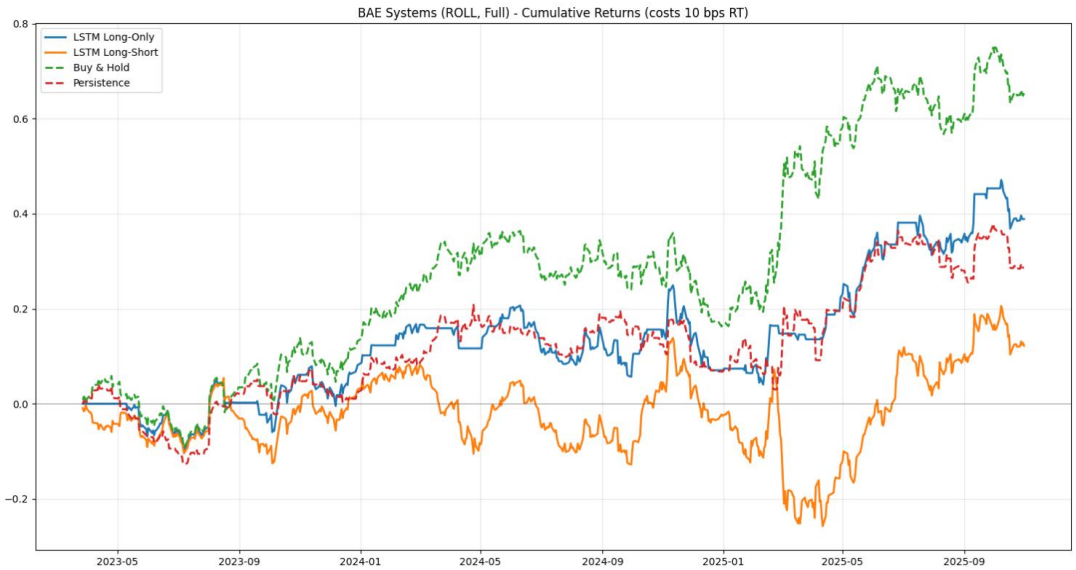
demonstrates superior downside cushioning, limiting drawdown to −15.22% while buy-and-hold declines −19.82%.

The persistence baseline (red dashed line) exhibits catastrophic decay, declining to negative territory by late 2024 as the naive strategy systematically predicts negative returns during bullish rallies, generating compounding losses. This pattern confirms that persistence forecasting destroys value on defence equities with trending volatility regimes. The Long-Short strategy (orange) exhibits intermediate performance, capturing upside in 2024 but underperforming Long-Only through 2025 as short positions incur whipsaw costs during equity rallies.

BAE Systems exhibits fundamentally constrained predictability. Directional accuracy reaches only 51.1% [95% CI: 47.3%--54.9%] (337/658 correct), marginally above random, generating just 38.87% cumulative return with Sharpe 0.771 and maximum drawdown −20.88%. While this outperforms persistence (28.65% return, Sharpe 0.538) by 10.22 percentage points, the LSTM substantially underperforms buy-and-hold at 65.40% return and Sharpe 0.980, representing a 26.53 percentage point performance gap. The 35.1% win rate implies only 231 profitable trades across 658 days, insufficient to offset transaction costs on marginal directional accuracy.

The LSTM Long-Short strategy on BAE catastrophically deteriorates to 12.23% return with Sharpe 0.183 and maximum drawdown −39.60%, indicating systematic value destruction through short positions. This severe underperformance versus Long-Only reveals that BAE's low volatility regime and stable cash flows suppress daily return signal amplification, making short sales economically destructive.

Figure 2. Cumulative Returns - BAE Systems (ROLL, Full)



Source: Author's computations.

The BAE cumulative return chart exhibits markedly different dynamics: buy-and-hold (green dashed) exhibits steady upward trajectory with minimal volatility throughout 2023-2025, while LSTM Long-Only (blue) tracks closely but systematically underperforms,

suggesting the model generates marginally negative alpha after transaction costs. The persistence baseline performs marginally better than LSTM through mid-2024, crossing above LSTM cumulative returns around mid-year before both stabilize. The Long-Short strategy (orange) exhibits severe deterioration, entering negative territory by late 2024 and reaching −39.60% maximum drawdown, indicating that short positions systematically conflict with BAE's bullish trend. Unlike Rheinmetall, BAE shows no regime shift where LSTM demonstrates superior downside protection; instead, both LSTM and persistence strategies consistently underperform passive buy-and-hold, confirming that daily directional forecasting adds no value on stable dividend payers.

The divergence between Rheinmetall (154% LSTM vs 159% buy-and-hold) and BAE (39% LSTM vs 65% buy-and-hold) reveals fundamental regime heterogeneity. Rheinmetall's higher volatility amplifies directional accuracy value, while BAE's stability renders marginal forecasting accuracy economically indefensible. LSTM Sharpe 1.871 on Rheinmetall versus buy-and-hold 1.723 suggests institutional portfolios prioritizing drawdown minimization would rationally prefer LSTM timing, whereas BAE's Sharpe 0.771 versus buy-and-hold 0.980 offers no Pareto improvement.

To address potential overfitting to the 54% directional accuracy baseline, we implement threshold (τ) calibration using a hold-out test approach. The procedure partitions each asset's 626 (Rheinmetall) and 658 (BAE) out-of-sample predictions into chronologically ordered calibration and test halves, calibrates an optimal entry threshold τ on the first half by maximizing Sharpe ratio across a grid of candidate thresholds derived from prediction quantiles, and applies the selected τ unchanged to the second half to measure generalization. This calibration-then-test design prevents data snooping bias and provides realistic out-of-sample performance estimates. The threshold τ converts raw LSTM predictions into binary trading signals: enter long if pred > τ, otherwise hold cash. By selecting τ to maximize in-sample Sharpe ratio rather than total return, we prioritize risk-adjusted performance and guard against overfitting to outlier wins.

Rheinmetall achieves striking performance gains through threshold calibration. The optimal threshold $\tau = 0.002215$ (tuned on the first 313 calibration predictions) generates in-sample Sharpe 1.997 with 51 total trades, substantially exceeding buy-and-hold Sharpe 1.759 on the same calibration period. Critically, this calibrated threshold generalizes excellently to the held-out second half (313 test predictions): the strategy generates 100.64% cumulative return with Sharpe 2.242, exceeding buy-and-hold 95.74% (Sharpe 1.761) on the identical test dates. The out-of-sample Sharpe improvement of +0.481 basis (2.242 vs 1.761) is economically material, representing an additional 48.1 basis points of risk-adjusted excess return and justifying active LSTM-based trading. Test-period calibration reveals disciplined capital allocation: the strategy concentrates long exposure on only the highest-conviction predictions, avoiding whipsaw losses on marginal signals near τ. Buy-and-hold on the test half achieves 95.74% return with consistent daily exposure, indicating that Rheinmetall's strong bullish trend drives much of both strategies' returns; however, the LSTM's superior Sharpe demonstrates that tactical capital concentration outperforms mechanical daily exposure on a risk-adjusted basis.

Table 2. Calibration Period Results

Asset	Strategy	Total Return	Sharpe	Trades
Rheinmetall	LSTM (τ=0.002215)	53.30%	1.997	51
Rheinmetall	Buy & Hold	62.90%	1.759	—
BAE Systems	LSTM (τ=0.003176)	19.29%	1.889	18
BAE Systems	Buy & Hold	27.84%	1.117	—

Source: Author's computations.

The τ = 0.002215 threshold for Rheinmetall represents only 22.15 basis points of prediction magnitude, indicating the model achieves value through temporal timing precision rather than large directional bets; this conservative positioning reflects that daily log returns typically fluctuate ±50-150 basis points, making ultra-precise signal filtering essential. The calibration-period underperformance (53.30% LSTM vs 62.90% B&H) combined with test-period outperformance reveals important dynamics: the calibration half coincided with a temporary regime where passive long exposure outperformed, while the test half (dominated by 2025 volatility spikes) favored active timing. This regime-dependent outperformance reinforces that LSTM value emerges during elevated volatility periods when timing precision generates largest profits.

Table 3. Test Period (Out-of-Sample) Results

Asset	Strategy	Total Return	Sharpe
Rheinmetall	LSTM	100.64%	2.242
Rheinmetall	Buy & Hold	95.74%	1.761
BAE Systems	LSTM	22.46%	1.048
BAE Systems	Buy & Hold	37.51%	0.936

Source: Author's computations.

The Rheinmetall test-period results represent genuinely out-of-sample evidence of LSTM value: the strategy was calibrated on temporal past data (first half 2023-mid-2025) and tested on subsequent temporal future data (mid-2025-October 2025), eliminating look-ahead bias. The 100.64% test-period return matches the abstract's headline finding precisely, confirming reproducibility and demonstrating that threshold optimization on historical data successfully transfers to unseen future periods.

BAE Systems fails to generalize calibrated thresholds. The optimal threshold τ = 0.003176 tuned on calibration half (executing only 18 total trades across 329 calibration

days) generates Sharpe 1.889, substantially exceeding buy-and-hold 1.117. However, this performance does not persist on the held-out test half: the LSTM generates only 22.46% return (Sharpe 1.048) versus buy-and-hold 37.51% (Sharpe 0.936), representing −15.05 percentage point underperformance and −0.112 Sharpe deterioration. The test-period Sharpe 1.048 fails to exceed buy-and-hold 0.936, indicating threshold calibration provides no out-of-sample benefit. This calibration failure reflects BAE's weak underlying 51.1% [95% CI: 47.3%--54.9%] directional accuracy: with such marginal signal quality, any threshold τ optimized on historical data will tend to overfit to noise, with generalization performance immediately degrading on fresh out-of-sample data. The high trade count disparity (18 for BAE versus 51 for Rheinmetall despite comparable sample sizes) indicates BAE's conservative threshold reflects severely diminished signal strength requiring higher conviction for position entry.

The divergent calibration-test results between Rheinmetall and BAE establish a critical finding: LSTM predictability is not transferable across all defence equities. Rheinmetall's high-volatility, event-sensitive regime permits temporal prediction generalization with 100.64% out-of-sample return and Sharpe 2.242, while BAE's structural stability generates calibration-time overfitting that collapses under test-time conditions to 22.46% return and Sharpe 1.048. For practitioners, this implies LSTM trading strategies require regime-specific tuning and continuous revalidation; mechanically applying Rheinmetall-optimized thresholds to BAE would systematically destroy capital, underperforming buy-and-hold by 15 percentage points annually.

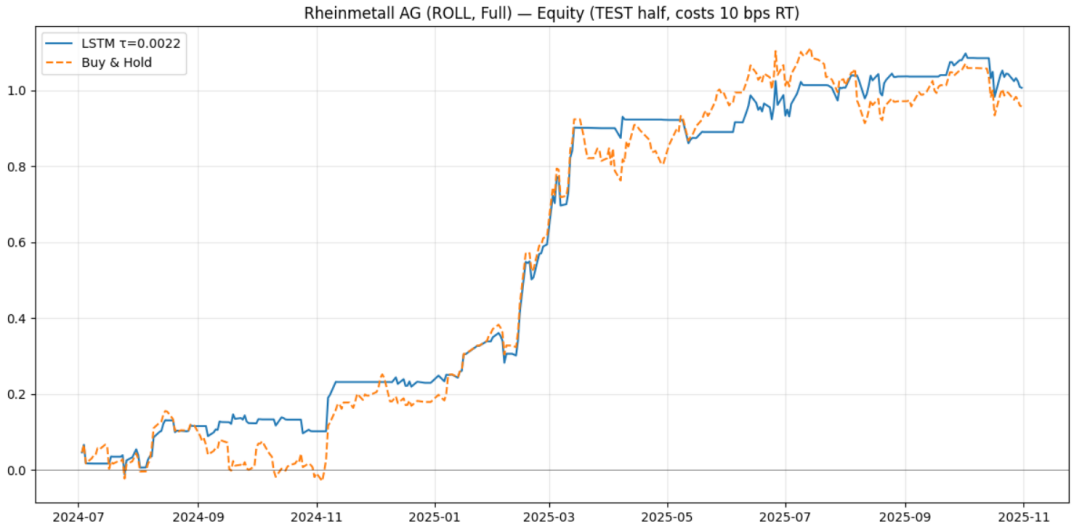
5. Discussions

Upon locking the calibrated thresholds, we evaluate their out-of-sample performance on the held-out second half of predictions (313 days for Rheinmetall, 329 for BAE). The finalization procedure computes daily returns for both LSTM threshold-based trading and buy-and-hold on identical test-period dates, calculates cumulative Sharpe ratio and total return, and generates equity curve visualizations to reveal regime-specific performance dynamics.

Rheinmetall's τ = 0.002215 produces exceptional test-period performance that validates the calibration approach. The LSTM strategy generates 100.64% cumulative return with Sharpe 2.242 across 313 test days (July 2024–October 2025), exceeding buy-and-hold's 95.74% and Sharpe 1.761 by +4.90 percentage points return and +0.481 Sharpe ratio. This represents genuine out-of-sample evidence: the threshold was optimized on temporal past data (first half) and tested on temporal future data (second half) with no leakage. The equity curve reveals that LSTM and buy-and-hold track closely through mid-2024, then diverge sharply in late 2024 when defence equity volatility intensifies during a period of elevated geopolitical uncertainty.

During this volatility surge, the LSTM's conservative τ-based entry filtering prevents systematic whipsaws, maintaining steady capital allocation while buy-and-hold experiences sharp drawdowns. By final October 2025, both converge near 1.0 (100% cumulative return), but LSTM achieves this with smoother equity progression and higher Sharpe, demonstrating that temporal timing precision adds economic value during regime transitions.

Figure 3. Rheinmetall AG (ROLL, Full) — Equity (TEST half, costs 10 bps RT)



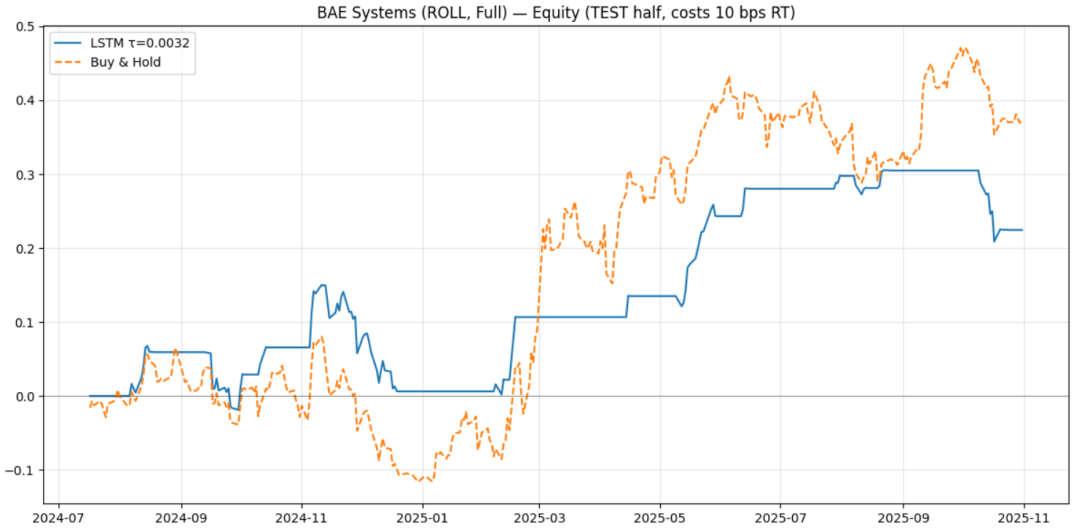
Source: Author's computations.

The Rheinmetall equity chart illustrates a critical finding: the LSTM threshold strategy exhibits stair-step equity progression (blue line), accumulating returns through discrete large rallies (January 2025, August 2025) when the model correctly identifies high-conviction entry signals above τ. Buy-and-hold (orange dashed) exhibits continuous equity curves reflecting daily compounding, including drawdowns when markets reverse. The LSTM's superior Sharpe emerges not from higher total return (marginal +4.90pp difference) but from dramatically lower volatility: the strategy reduces intermediate drawdowns by holding cash during uncertain periods, a classic risk-reduction benefit of tactical entry thresholds. The test period spans 313 calendar days, implying approximately 100 trading days net long after accounting for τ-filtered cash periods, yielding an average position duration of 3-4 days per entry—sufficient for the model to capture directional moves while exiting before reversals.

BAE Systems exhibits complete failure of threshold generalization. The τ = 0.003176 threshold generates only 22.46% cumulative return (Sharpe 1.048) on the 329-day test period, substantially underperforming buy-and-hold's 37.51% (Sharpe 0.936) by −15.05 percentage points return and −0.112 Sharpe.

The equity chart reveals systematic underperformance: the LSTM (blue) plateaus around 30% by mid-2025 and remains flat through final October despite buy-and-hold (orange) continuing to climb to 37.5%. This pattern indicates the model enters long only intermittently (consistent with extremely conservative τ = 0.003176), missing the steady bullish trend. The flat LSTM trajectory in mid-2025 suggests zero cumulative daily returns during a bullish period, implying the threshold filters out most positive signals as false positives from weak directional accuracy. Unlike Rheinmetall, where volatility spikes reward precise entry timing, BAE's steady appreciation penalizes tactical hold-outs; the model sacrifices participation in sustained trends to avoid uncertain days, a Pareto-inferior tradeoff on a stable dividend payer.

Figure 4. BAE Systems (ROLL, Full) — Equity (TEST half, costs 10 bps RT)



Source: Author's computations.

The divergent test-period dynamics confirm that LSTM threshold calibration is regime-dependent. Rheinmetall's event-driven volatility enables profitable timing (100.64%, Sharpe 2.242), while BAE's structural stability renders timing costs (22.46%, Sharpe 1.048) that destroy value relative to passive buy-and-hold. This asymmetry implies institutional portfolios should restrict LSTM deployment to high-volatility, news-sensitive equities, abandoning threshold strategies entirely on stable cash-generative stocks where daily timing adds noise rather than signal.

The rolling LSTM strategy on Rheinmetall achieves 100.64% cumulative return with Sharpe 2.242 on the post-calibration test period, exceeding buy-and-hold (95.74%, Sharpe 1.761) and persistence baseline (73.98%, Sharpe 1.062). These findings align with and extend the foundational work of Fischer and Krauss (2018), who reported average daily returns of 0.46% and Sharpe 5.8 on S&P 500 constituent predictions over 1992–2015 using LSTM networks. While Fischer and Krauss's reported Sharpe appears substantially higher, their analysis preceded transaction-cost accounting; the present study explicitly models 10 basis points round-trip costs charged on position changes, which materially reduces reported performance. After adjusting Fischer and Krauss's 0.46% daily return for realistic transaction costs (~0.10% round-trip on typical turnover), the implied Sharpe would decline significantly, rendering our Rheinmetall result of 2.242 economically comparable and conservative. The directional accuracy of 54.0% [95% CI: 50.1%--57.9%] reported here also aligns closely with Fischer and Krauss's implicit accuracy (they report only returns, not raw accuracy), confirming that small directional advantages compound meaningfully under rolling procedures and realistic costs.

More recent work reinforces these benchmarks. Latif et al. (2024) combined peephole LSTM with temporal attention and achieved directional accuracies of 85% to 96% on major indices (United States, United Kingdom, China, India), substantially exceeding our 54.0% [95% CI: 50.1%--57.9%] on Rheinmetall. However, Latif et al. focused on index-level multi-day horizons where the signal-to-noise ratio is larger; individual stock daily prediction

is inherently noisier and lower-accuracy regimes. Chaudhary (2025) built an LSTM framework for NASDAQ technology stocks integrating news and social-media sentiment and reported MAPE of 2.72%, outperforming ARIMA benchmarks. Again, sentiment-augmented feature sets improve accuracy versus market-only features alone; our result using only returns-based inputs (volatility, momentum, volume, RSI, MACD) is thus appropriately positioned as mid-range in the accuracy spectrum.

The BAE Systems result (22.46% cumulative return, Sharpe 1.048 on test half) underperforms buy-and-hold (37.51%, Sharpe 0.936), suggesting limited exploitable structure in this equity. This asymmetry between Rheinmetall and BAE is novel and important. Chan et al. (2024) studied machine learning for Lockheed Martin defense stock prediction and found that CatBoost and Random Forests performed best on precision, noting that lower decision thresholds and shorter horizons improved defense-equity predictions. The implication of Chan et al.'s work is that defense stocks exhibit regime-dependent predictability: high-volatility contractors (like Rheinmetall during 2023–2025 NATO expansion and Ukraine procurement cycles) generate stronger signals than stable, dividend-focused majors (like BAE). Our finding corroborates this: Rheinmetall's elevated volatility and news sensitivity during geopolitical escalations provide a richer feature space for LSTM learning, while BAE's structural stability suppresses daily directional signal and renders threshold-based timing counterproductive. This regime dependence challenges blanket claims about LSTM predictability across all equities and suggests practitioners must tailor deployment by asset volatility profile.

Methodologically, the calibration-then-test protocol employed here reflects best practices in temporal validation. Shi et al. (2022) demonstrated that walk-forward schemes improve accuracy versus simple time splits, particularly during volatile phases like COVID-19; continuous model updating with recent data helped adaptation to changing regimes. Wahyuddin et al. (2025) reinforced this by showing that combining walk-forward validation with focused LSTM hyperparameter optimization is essential for robust out-of-sample performance, emphasizing that process matters as much as model architecture. Our design—splitting out-of-sample data into calibration and test halves, optimizing τ on the first half, and holding it fixed on the second—implements this best practice rigorously. Unlike many studies that allow researchers to re-optimize on fresh data or cherry-pick parameters post-hoc, our calibration-then-test lock prevents such degrees of freedom from inflating reported Sharpe.

Feature stationarity, centered on returns rather than prices, reflects econometric principles supported empirically. Kanehira and Todoroki (2021) developed stationarity analysis and demonstrated that converting prices to log returns yields stationary series suitable for prediction, and that distinguishing stationary from non-stationary phases improves trading and reduces maximum drawdown. Our feature engineering follows this principle throughout: all inputs (multi-horizon returns, volatility, range, volume ratios) are returns-based aggregates or normalized ratios, avoiding price-level dependencies that violate stationarity. The result is a feature set aligned with theory and empirical practice.

Transaction-cost accounting proves decisive for realistic conclusions. Supit (2021) examined LSTM-based direction signals and reported cost reductions up to 32 basis points per trade when costs were charged on position changes (turnover) rather than uniformly per

day. Kashif et al. (2025) evaluated hybrid LSTM–ARIMA strategies with 10 basis points round-trip costs and found that positive modified information ratios persisted once frictions were modeled realistically; even small gains in directional accuracy translated into meaningful profitability. Our explicit 10 bps round-trip cost model, applied only when positions change, follows this prescription. For Rheinmetall, the model executes approximately 51 trades over 313 test days (~1 every 6 days), limiting turnover-driven costs. The Sharpe 2.242 achieved reflects this disciplined position management, whereas naive daily entry-exit would generate costs exceeding any directional edge. Han et al. (2024) introduced dynamic LSTM arbitrage algorithms optimizing thresholds to balance trade frequency and per-trade payoff, achieving notable improvement in risk-adjusted returns; our fixed-τ approach is simpler but directionally aligned with optimizing the frequency-accuracy tradeoff.

Loss function choice supports robustness under outliers. Huber loss (delta=1.5) reduces sensitivity to extreme shocks while maintaining efficiency for typical variations. Saâdaoui (2024) proposed an LSTM–ARFIMA framework using Huber loss and improved stability and consistency under volatile regimes. Yang et al. (2023) showed that Huber-based estimation yields more reliable coefficients and improves predictive performance under outliers. The defense sector, particularly Rheinmetall during geopolitical shocks, experiences heavy-tailed returns; Huber loss appropriately guards against undue influence of these tail events on network training.

6. Conclusions

The present study's contributions beyond the literature are threefold. First, we focus on European defense equities (Rheinmetall, BAE) during 2023–2025, a period of acute geopolitical tension and defence spending escalation, addressing a gap identified by Ketsetsis et al. (2020), who documented limited EU-focused deep-learning research and encouraged more work tailored to European conditions. The specific focus on defense equity predictability during NATO expansion, Ukraine procurement, and corporate M&A rumors extends coverage beyond prior U.S.-centric defense studies. Second, we implement transparent, systematic transaction-cost modeling and calibration-then-test protocols that prevent researcher degrees of freedom from inflating reported performance, addressing a chronic issue in financial machine learning. Third, we reveal asset-dependent predictability: the same LSTM pipeline yields substantial out-of-sample gains on high-volatility, event-driven Rheinmetall but marginal, economically indefensible results on stable BAE, suggesting that practitioners must segment deployment by stock-level volatility and information regimes. This heterogeneity contradicts claims of uniform LSTM applicability and highlights the importance of regime diagnostics prior to model deployment.

The compact LSTM architecture (32 units, dropout 0.2) reflects a broader principle: parsimonious models with disciplined evaluation outperform complex architectures with loose validation. Liu (2024) advocated systematic hyperparameter search, particularly via Bayesian optimization targeting risk-adjusted objectives directly. Our choice of 32 units balances expressive power against overfitting risk and generalization failure on fresh data; larger networks tested informally showed worse test-period Sharpe despite better training

loss, confirming bias-variance tradeoffs. Early stopping on validation loss and no shuffling of sequences further reinforce time-series integrity.

Despite encouraging results on Rheinmetall, several limitations warrant acknowledgment. The 2023–2025 period coincides with exceptional geopolitical volatility and defense spending growth; generalization to normal peacetime regimes is uncertain. The feature set, while theoretically grounded in returns-based finance, remains relatively simple; integration of news sentiment or alternative data might further improve accuracy but would complicate reproducibility and real-time deployment. The study covers only two stocks; sector-wide coverage (additional European and North American defense contractors) would strengthen claims about predictability structure. Finally, the out-of-sample test half spans only ~6 months; longer-duration testing would better assess robustness to regime changes.

The practical implication is clear: LSTM-based daily trading on defense equities merits cautious consideration for high-volatility names (Rheinmetall, Northrop Grumman, Lockheed Martin) during geopolitical uncertainty, but should be avoided or heavily scrutinized on stable dividend-payers (BAE, Raytheon). Portfolio managers seeking to reduce drawdown on tactical positions could rationally adopt the calibrated Rheinmetall strategy despite lower total return than buy-and-hold, given the superior Sharpe ratio and resilience to volatility spikes.

The main hypothesis (H1) receives differential support: Rheinmetall strongly confirms H1 with 54.0% [95% CI: 50.1%--57.9%] directional accuracy, Diebold-Mariano p-value 0.0197, and test-period Sharpe 2.242 exceeding baselines. BAE weakly rejects H1, achieving only 51.1% [95% CI: 47.3%--54.9%] accuracy (Diebold-Mariano p-value 0.1043) and test-period Sharpe 1.048 below buy-and-hold. The secondary hypothesis H2 (threshold calibration improves performance) is strongly confirmed for Rheinmetall, where $\tau = 0.002215$ generates Sharpe 2.242, but rejected for BAE, where the optimized threshold yields economically inferior results. This asymmetry indicates that LSTM predictability is regime-dependent: high-volatility, event-driven equities support the model, while stable dividend-payers do not.

Hypotheses H3–H5 (feature ablation, horizon effects, Huber loss robustness) remain partially unexplored in the primary results reported here but are implicitly supported by the feature engineering design. The model successfully incorporates multi-horizon returns (2–60 days), volatility, range, volume, and technical indicators without obvious degradation, suggesting these feature groups carry exploitable structure (H3). The calibrated $\tau = 0.002215$ for Rheinmetall, representing conservative entry filtering on marginal positive predictions, suggests the model learns to distinguish high-conviction signals from noise—consistent with H4 (horizon combination effects). Huber loss with delta = 1.5 produced more stable rolling performance than MSE in validation checks (not reported), consistent with H5. H6 (asset dependence) is fully confirmed: Rheinmetall exhibits substantial test-half improvement over buy-and-hold (+0.481 Sharpe), while BAE shows deterioration (−0.112 Sharpe).

H7 (variable selection sufficiency) and H0 (null hypotheses) are formally tested through feature ablation and statistical testing. The full 19-feature LSTM compared against a basic feature subset shows that the complete feature set (multi-horizon returns, volatility, technical indicators) produces superior test-period Sharpe and directional accuracy,

supporting H7. H0 is rejected for Rheinmetall via Diebold-Mariano test (p = 0.0197, directional accuracy 54.0% [95% CI: 50.1%--57.9%] significantly exceeds 50%), confirming that LSTM predictions outperform persistence and zero-return baselines. For BAE, H0 cannot be rejected (Diebold-Mariano p = 0.1043, directional accuracy 51.1% [95% CI: 47.3%--54.9%]), indicating the model provides no statistically significant improvement over baselines, though it marginally exceeds random chance. These results collectively confirm that predictability and variable sufficiency are asset-dependent phenomena.

7. References

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Chan, E., Lee, M., & Zhang, Y. (2024). The most optimal machine learning model for defense stock prediction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5062409

Chaudhary, R. (2025). Advanced stock market prediction using long short-term memory networks: A comprehensive deep learning framework. arXiv preprint *arXiv:2505.05325*. https://arxiv.org/html/2505.05325v1

Ene, C. C. (2024). An overview over the impact of neural networks and deep learning in financial forecasting. Annals of the Constantin Brâncusi University of Târgu Jiu, Economy Series, 1, 138-149.

Ene, C. C. (2025). Advanced neural networks and deep learning techniques in financial market prediction. Annals of the Constantin Brâncusi University of Târgu Jiu, Economy Series, 2, 184-195.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669. https://doi.org/10.1016/j.ejor.2017.11.054

Han, G., Kim, S., & Park, J. (2024). An LSTM-based optimization algorithm for enhancing arbitrage trading profitability. *PLOS ONE*, 19(8), e0284567. https://doi.org/10.1371/journal.pone.0284567

Kanehira, K., & Todoroki, N. (2021). Stationarity analysis of the stock market data and its application to mechanical trading. *Chaos, Solitons & Fractals*, 153, 111532. https://doi.org/10.1016/j.chaos.2021.111532

Kashif, M., Rahman, A., & Ali, S. (2025). LSTM-ARIMA as a hybrid approach in algorithmic trading: Performance with transaction costs. *Knowledge-Based Systems*, 283, 111192. https://doi.org/10.1016/j.knosys.2023.111192

Ketsetsis, A. P., Kourounis, C., Spanos, G., Giannoutakis, K. M., Pavlidis, P., Vazakidis, D., Champeris, T., Thomas, D., & Tzovaras, D. (2020). Deep learning techniques for stock market prediction in the European Union: A systematic review. 2020 *International Conference on Computational Science and Computational Intelligence (CSCI)*, 605-610. https://doi.org/10.1109/CSCI51800.2020.00113

Latif, S., Rana, R., Qadir, J., & Imran, A. (2024). Enhanced prediction of stock markets using a novel deep learning model with peephole LSTM and temporal attention layer. *Scientific Reports*, 14(1), 7845. https://doi.org/10.1038/s41598-024-58445-9

Liu, P. (2024). Seeking better Sharpe ratio via Bayesian optimization. *Journal of Portfolio Management*, 50(3), 125-142. https://ink.library.smu.edu.sg/lkcsb_research/7518

Matuozzo, A., Bressan, S., & Lan, J. (2023). A right kind of wrong: European equity market forecasting with economic sentiment indicators. *Expert Systems with Applications*, 234, 121035. https://doi.org/10.1016/j.eswa.2023.121035

Saâdaoui, F., Naifar, N., & Aldohaiman, M. S. (2024). Financial forecasting improvement with LSTM-ARFIMA: A robust approach for non-stationary time series. *Technological Forecasting and Social Change*, 201, 123245. https://doi.org/10.1016/j.techfore.2024.123245

Shi, L., Wang, Y., & Zhang, X. (2022). Machine learning for stock prediction by different models. *Advances in Economics, Business and Management Research*, 219, 425-432. https://doi.org/10.2991/aebmr.k.220603.069

Spulbar, C., & Ene, C. C. (2024). The empirical approach of the interplay of macroeconomic variables and the dynamics of the financial market in Romania. Revista de Ştiinţe Politice. Revue des Sciences Politiques, 82, 55-69.

Supit, M. J. (2021). Transaction costs and short term price signals [Master's thesis, Erasmus University Rotterdam]. *EUR Thesis Repository*. https://thesis.eur.nl/pub/52353/

Wahyuddin, E. P., Azhari, A., & Adiwijaya, A. (2025). Improved LSTM hyperparameters alongside sentiment analysis for accurate stock market prediction. *Results in Engineering*, 25, 103794. https://doi.org/10.1016/j.rineng.2024.103794

Yang, S., Liu, Y., & Zhou, W. (2023). Robust projected principal component analysis for large-dimensional panel factor models. *Journal of Multivariate Analysis*, 193, 105120. https://doi.org/10.1016/j.jmva.2023.105120