Intraday Gold Price Forecasting Using Hybrid Machine Learning And Transformer Models Across Multiple Timeframes

Mohammad Ali Jaber

Supervised by Dr. Vitaliy Milke

A thesis submitted in fulfillment of the requirements
for the degree MSc Computer Science

School of Computing and Information Science
Anglia Ruskin University
18/9/2025

Declaration

I, **Mohammad Ali Jaber**, declare that the work in this dissertation titled *"Intraday Gold Price Forecasting Using Hybrid Machine Learning and Transformer Models Across Multiple Timeframes"* is carried out by me. This work has not been submitted to Anglia Ruskin University or any other educational institution for the award of a degree or educational qualification. I also declare that the information published in this dissertation has been obtained and presented in accordance with academic rules and ethical conduct. Any information obtained from other sources has been properly referenced.

Table of Contents

List of Figures

List of Tables

Abstract

This study presents a comprehensive comparative analysis of three advanced machine learning architectures for multi-horizon forecasting of gold prices (XAU/USD) using high-frequency intra-day gold price data from Dukascopy, at 5-minute, 15-minute, and 1-hour intervals. In particular, two hybrid models, long-short-term memory networks integrated with Extreme Gradient Boosting (XGBoost) and CATBoost, and the state-of-the-art Informer time-series transformer architecture were implemented and rigorously evaluated. The empirical findings show that hybrid LSTM-boosting models consistently outperform the Informer model, demonstrating superior predictive accuracy and generalisability across multiple prediction horizons.

Chapter 1

Introduction

1.1 Overview

Gold occupies a distinctive position in global finance as both a monetary asset and a safe haven during macroeconomic stress, attracting persistent intra-day liquidity in major venues (e.g., XAU/USD) and facilitating rapid price discovery. During periods of uncertainty, such as the COVID-19 pandemic, the hedging and diversification properties of gold strengthened, underscoring its appeal to both institutional and retail traders (Wen, Tong & Ren 2022). At high frequencies, price dynamics are driven by order-flow imbalances, information releases, and cross-asset spillovers from the U.S. dollar and energy markets, producing bursts of volatility, microstructure frictions, and regime shifts that complicate short-horizon forecasting (Liang et al. 2023).

In parallel, forecasting methodology has evolved from statistical models toward machine learning (ML) and deep learning (DL). Recurrent neural networks such as LSTM capture temporal dependencies and have been widely applied to commodity prices, including gold, while gradient-boosted trees (e.g., XGBoost, CatBoost) remain state-of-the-art in tabular features extracted from time series (Jabeur et al. 2024, Grinsztajn et al. 2022). More recently, Transformer architectures tailored for time series, such as Informer and PatchTST, have targeted long-range dependencies with efficient attention and patching mechanisms (Zhou et al. 2021, Nie et al. 2022). However, empirical results remain mixed, with several surveys noting that Transformers do not uniformly dominate classical or hybrid approaches across horizons and datasets (Wen, Zhou, Zhang, Chen, Ma, Yan & Sun 2022).

Despite abundant work on daily or longer horizons, intra-day gold forecasting is less explored relative to equities or crypto, and faces additional challenges: microstructure noise (bid–ask bounce), diurnal seasonality, and sensitivity to evaluation protocol (e.g., back-tests, leakage) (Zhang 2020). Rigorous, walk-forward, block-wise validation is therefore essential for assessing true generalisation under non-stationarity; yet, many studies still rely on random splits or single-period tests (Wahyuddin et al. 2025). This dissertation addresses this gap by explicitly focusing on intra-day horizons forecasting for XAU/USD and contrasting modern Transformer with hybrid LSTM boosted models under a strict walk-forward design.

1.2 Problem Background

Intra-day forecasting of XAU/USD is operationally valuable for market making, execution, and risk control, but remains difficult due to three intertwined issues. First, according to Hansen & Lunde (2006), high-frequency data are contaminated by market microstructure noise (e.g., discrete pricing, order-book frictions), which biases volatility and adversely affects supervised learning unless carefully handled. Second, intra-day regimes change rapidly around macro news and cross-asset shocks (e.g., USD, oil), creating non-stationarity and concept drift (Jurdi 2020). Third, evaluation practices are inconsistent; many studies use random splits, static windows, or single train–test partitions, which can inflate apparent accuracy by leaking temporal information and under-representing hard market states (Wang & Ruf 2022).

Studies that address these challenges are promising, but their findings are incomplete. LSTM-based models demonstrate effectiveness in commodity forecasting, but rarely tackle intra-day gold or microstructure noise (Ferreira & Medeiros 2021). Gradient-boosted trees (XGBoost / CATBoost) excel on tabular features and remain competitive or superior on many real-world datasets, especially with moderate sample sizes (Grinsztajn et al. 2022). Transformer-style forecasters (e.g., Informer; PatchTST) address long-horizon context with efficient attention and patching, yet recent surveys caution that Transformers are not universally superior and can underperform simpler or hybrid baselines without careful design (Kim et al. 2025).

From these gaps emerge three drivers for this study: (i) a relative paucity of systematic intra-day XAU/USD forecasting studies across multiple horizons; (ii) a lack of direct, like-for-like comparisons between hybrid LSTM-plus-boosting pipelines and modern Transformers; and (iii) insufficient use of leakage-resistant, walk-forward block splitting. The working hypothesis is that hybrid pipelines using LSTM to encode sequential structure and boosting (XGBoost/CatBoost) to model non-linearities in learned features will match or outperform a strong Transformer (Informer) on intra-day XAU/USD when evaluated with rigorous walk-forward validation and realistic targets (high bid; low ask).

1.3 Research Aim

This research aims to advance the application of machine learning methods for intra-day gold price forecasting by systematically benchmarking and evaluating different model architectures. In particular, the study focuses on comparing hybrid deep learning-tree-based models, namely LSTM-XGBoost and LSTM-CatBoost, against a Transformer-based forecaster (Informer). These models were selected because they represent distinct methodological paradigms: recurrent networks coupled with gradient-boosted decision trees on one hand, and attention-driven sequence models on the other. By examining their relative strengths and weaknesses, the research seeks to identify which approach is most effective for capturing intra-day price dynamics in the gold market.

The study is motivated by the need to address the challenges of short-term forecasting under noisy, high-frequency conditions where microstructure effects can undermine prediction accuracy. Through empirical experimentation on intra-day gold price data, the research aims to determine whether hybrid architectures or Transformer-style forecasters provide superior performance across multiple prediction horizons. The overarching goal is to contribute to the methodological literature on financial time series forecasting by clarifying the comparative efficacy of state-of-the-art machine learning approaches in the context of gold price prediction.

1.4 Research Question

The growing complexity of financial markets, combined with the volatility of gold (XAU/USD), makes intra-day forecasting a particularly challenging task. Traditional time series models often fail to capture microstructural noise, while recent advances in deep learning and attention-based architectures offer promising yet inconclusive results. Against this backdrop, this study seeks to identify which modelling approach and data resolution are best suited for robust intra-day predictions. This dissertation is driven by the following core question:

> For intra-day XAU/USD forecasting, do hybrid gradient-boosted LSTM pipelines (LSTM+XGBoost, LSTM+CatBoost) outperform a modern Transformer-based model (Informer) across 5, 15, 30, and 60-minute horizons, and which candlestick resolution (5, 15, or 60-minute) provides the most effective training input under a walk-forward validation framework?

1.5 Research Objectives

- To design a leakage-resistant, walk-forward evaluation protocol with block splitting for intra-day XAU/USD across 5, 15, and 60-minute horizons.
- To implement three models: LSTM+XGBoost, LSTM+CatBoost, and Informer on high-bid and low-ask targets using consistent preprocessing and feature interfaces.
- To evaluate and compare the predictive performance of the three models across multiple intra-day horizons (5, 15, 30, and 60 minutes) using RMSE, MAE, and R^2.
- To investigate the impact of data granularity (5, 15, and 60-minute candlesticks) on model training effectiveness and forecasting reliability.

1.6 Research Scope

1.6.1 In Scope

The study focuses on intra-day forecasting of XAU/USD using Dukascopy historical data (ask/bid) from 1 Jan 2024 to 29 Mar 2025, aggregated to 5, 15, and 60-minute candlesticks. Prediction tasks target the next 5, 15, 30, and 60 minutes, depending on source frequency. Models include LSTM+XGBoost, LSTM+CatBoost, and Informer. The evaluation employs walk-forward, block-wise splits with RMSE, MAE, and R^2 reported separately for high bid and low ask prices.

1.6.2 Out Of Scope

While this study aims to provide a rigorous comparison of machine learning models for intra-day gold forecasting, several areas remain outside its scope. The analysis does not extend to other asset classes such as equities, cryptocurrencies, or commodities beyond gold, which may exhibit different dynamics and forecasting challenges (Encean & Zinca 2022). Similarly, macroeconomic indicators, news sentiment, and geopolitical shocks (factors known to influence gold price volatility) are excluded, as the focus is restricted to pure time-series modelling (Baur & Lucey 2010). Furthermore, this work does not incorporate practical trading considerations such as transaction costs, slippage, or execution strategies, nor does it engage in portfolio optimisation or asset allocation frameworks (Loras 2024). Finally, the study does not provide financial advice or policy recommendations; its contributions are limited to methodological evaluation and precision of the forecast.

1.7 Abridged Methodology

This study builds a unified pipeline for intra-day XAU/USD forecasting by first aggregating tick-level bid and ask data into 5, 15, and 60-minute candlesticks. After merging bid and ask series, the data are cleaned, timestamp-aligned, and enriched with engineered features such as volume deltas, price spreads, and binary indicators for new trading days and weeks. Values are then normalised, prices divided by 1,000 and volumes by their training-set maxima, and memory usage is halved through downcasting, producing two dataset versions tailored respectively for LSTM-based hybrids and the Informer.

To ensure realistic performance estimates, the framework employs walk-forward validation rather than random splits. Two custom splitting functions generate sequential training and testing blocks: one transforms 3D tensors into (X_train, X_test, y_train, y_test) for the LSTM-XGBoost and LSTM-CATBoost hybrids, while the other slices the data into blocks and saves them as CSV files to feed the Informer's native loader. This preserves temporal dependencies, prevents data leakage, and mimics live forecasting, where models are continuously updated as new data arrive.

Three models are compared under identical schemes: (1) an LSTM feature extractor combined with XGBoost regressors; (2) the same LSTM front end coupled with CatBoost regressors; and (3) Informer, a Transformer-based encoder–decoder optimised for long-sequence time series. All are evaluated across four horizons (5, 15, 30, and 60 minutes), depending on the tick data used, using mean absolute error, root mean squared error, and the coefficient of determination to measure both accuracy and generalisability.

1.8 Contribution

This study contributes to the intra-day forecasting literature on XAU/USD by delivering the first systematic, walk-forward evaluation of hybrid gradient-boosted LSTM pipelines against a state-of-the-art Transformer forecaster. The research demonstrates that hybrid models (particularly LSTM+CATBoost) consistently outperform the Informer across multiple horizons, with the 15-minute resolution emerging as the most effective training input. Beyond the empirical results, the project advances methodological practice by implementing a dual-target framework (high bid and low ask) under leakage-resistant walk-forward splits, ensuring greater robustness and reproducibility than many prior studies. Collectively, these findings close the gap identified in Section 1.2 by showing that simpler hybrid pipelines can surpass complex Transformer architectures in intra-day commodity forecasting, thus providing both academic insight and practical relevance for trading and risk management applications.

Chapter 2

Literature Review

2.1 Overview

Intra-day forecasting of gold prices (XAU/USD) remains a complex and under-explored challenge in financial econometrics. Gold is both a commodity and a monetary asset, exhibiting strong safe-haven behaviour and sensitivity to macroeconomic shocks, particularly movements in the US dollar, interest rates, and geopolitical risk (Baur & Lucey 2010). While traditional econometric models such as ARIMA and GARCH families have long been applied to precious metals, their assumptions of stationarity and linear dependence often fail to capture the high-frequency volatility and regime shifts that characterise intra-day markets (Yaziz et al. 2013).

Over the past decade, advances in machine learning and deep learning have reshaped financial forecasting research. Recurrent neural networks, especially long-short-term-memory (LSTM) architectures, have been widely adopted for capturing sequential dependencies in commodity and FX prices (Nelson et al. 2017, Fischer & Krauss 2018). In parallel, tree-based ensemble learners such as XGBoost and CATBoost have proven highly effective for tabular feature sets extracted from financial time series, often outperforming stand-alone deep learning in structured prediction tasks (Chen & Guestrin 2016, Prokhorenkova et al. 2018). More recently, Transformer-based architectures have emerged as promising alternatives, designed to model long-range temporal dependencies with efficient attention mechanisms. Models such as the Informer (Zhou et al. 2021) and PatchTST (Nie et al. 2022) demonstrate strong results in benchmark time series tasks, although their superiority in financial intra-day contexts remains inconclusive.

A further complication arises from diverse approaches. Empirical studies differ widely in data resolution, choice of evaluation protocol, and target horizons. Several reviews highlight that inconsistent use of random splits, static windows, or short back-tests might lead to a loss of generality, especially in non-stationary intra-day markets (Lim & Zohren 2021). Consequently, time-series cross-validation has been recommended for temporal model evaluation under structural breaks and concept drift (Bergmeir & Benítez 2012).

Against this backdrop, this study situates itself at the intersection of hybrid LSTM approaches (LSTM+XGBoost, LSTM+CATBoost) and Transformer-based forecasting (Informer). It aims to evaluate these architectures under consistent intra-day settings and across multiple horizons, applying a rigorous walk-forward validation framework, thereby addressing gaps in the literature around methodological domain-specific applications to XAU/USD.

2.2 Financial Forecasting Approaches

Recurrent Neural Networks (RNNs), particularly LSTM and Gated Recurrent Unit (GRU) architectures, have become foundational tools in time series modelling due to their ability to capture sequential dependencies. Unlike traditional statistical methods such as ARIMA or GARCH, which assume linearity and stationarity, LSTMs can model complex non-linear temporal relationships, making them particularly suited for volatile financial data (Hochreiter & Schmidhuber 1997, Fischer & Krauss 2018). In the context of commodities and foreign exchange, LSTM-based models have demonstrated considerable effectiveness. For instance, Fischer & Krauss (2018) applied LSTM networks to S&P 500 component data from 1992 to 2015 and demonstrated that LSTMs significantly outperformed memory-free models like Random Forests and Deep Neural Nets in directional prediction accuracy and Sharpe ratio, while Livieris et al. (2020) highlighted their capacity for capturing long-range dependencies in exchange rate forecasting. More recently, applications to commodity markets, including crude oil and gold, confirm their ability to handle seasonality and volatility (Sezer et al. 2020).

However, several limitations constrain their broader applicability. LSTMs are prone to overfitting when trained on high-frequency intra-day data, especially in the presence of microstructure noise. They also exhibit high computational cost and training instability compared to tree-based or attention-based alternatives (Lim et al. 2021). These shortcomings have motivated hybrid approaches, where LSTMs are combined with ensemble learners such as gradient boosting to leverage both temporal representation learning and structured feature exploitation. Hybrid boosted architectures, such as LSTM+XGBoost and LSTM+CATBoost, have emerged as effective alternatives. XGBoost, introduced by Chen & Guestrin (2016), is widely recognised for its scalability and strong generalisation on tabular financial features. By pairing XGBoost with LSTM, researchers have reported improved predictive accuracy and robustness in financial time series (Oukhouya et al. 2024). Similarly, CATBoost, developed by Prokhorenkova et al. (2018), handles categorical and imbalanced feature distributions effectively, making it a strong complement to LSTM outputs. Studies show that such hybrid pipelines often outperform stand-alone deep learning models by combining sequential pattern extraction with non-linear feature enhancement, thus reducing the overfitting and instability challenges of pure LSTM architectures (Livieris et al. 2020).

More recently, Transformer-based architectures have been proposed as alternatives to recurrent models in time series forecasting. Unlike RNNs, Transformers rely on self-attention mechanisms to capture long-range dependencies without recurrent connections, offering better predictions and efficiency. Informer, introduced by Zhou et al. (2021), uses probSparse attention and a generative decoder to handle long-sequence forecasting with reduced computational complexity, while PatchTST (Nie et al. 2022) introduces a patching mechanism to enhance temporal context learning. Although these models demonstrate strong benchmark performance across multiple domains, their superiority in financial intra-day forecasting remains inconclusive. Several surveys highlight that while Transformers excel in long-horizon forecasting, their performance does not consistently dominate hybrid or recurrent approaches in volatile, high-frequency financial markets (Lim & Zohren 2021).

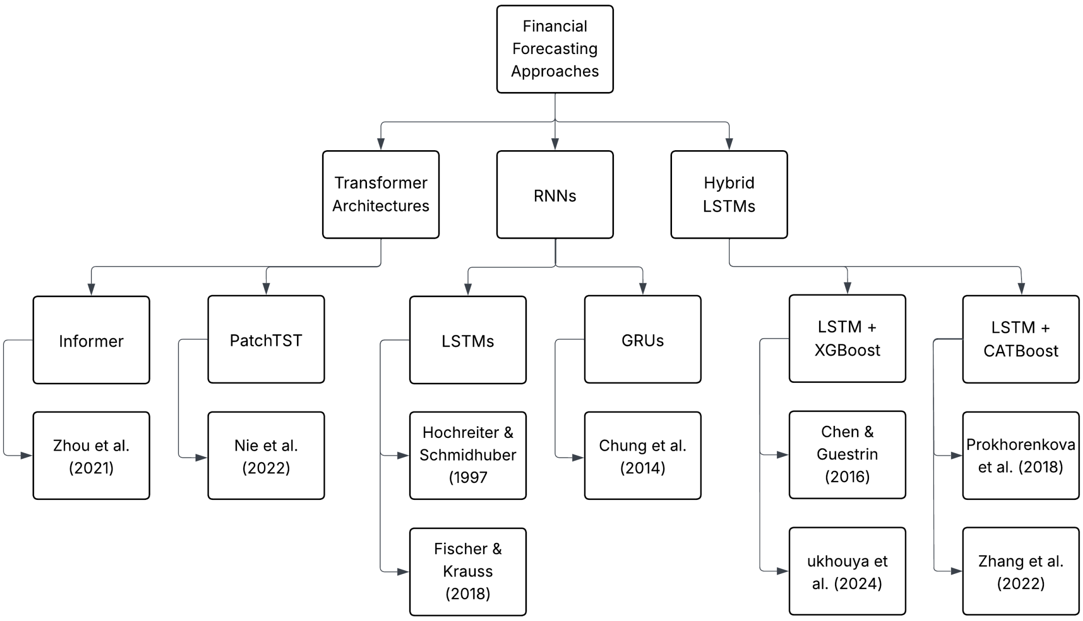


Figure 2.1: Taxonomy of financial forecasting approaches

Figure 2.1 above illustrates the main categories of machine learning approaches applied in financial time series forecasting. Recurrent neural networks (RNNs), such as LSTMs (Hochreiter & Schmidhuber 1997, Fischer & Krauss 2018) and

GRUs (Chung et al. 2014), form the foundational class of sequential models. Hybrid boosted models extend LSTM capabilities by integrating gradient boosting methods, including XGBoost (Chen & Guestrin 2016, Oukhouya et al. 2024) and CATBoost (Prokhorenkova et al. 2018, Zhang et al. 2022), to enhance performance. Also, transformer-based architectures such as Informer (Zhou et al. 2021) and PatchTST (Nie et al. 2022) have been included for efficiently capturing long-range dependencies in financial data.

2.3 Recurrent Neural Networks

RNNs are foundational architectures in sequential modelling and have been widely applied in financial forecasting tasks. Their ability to capture temporal dependencies makes them more suitable for time series predictions than traditional econometric models like ARIMA and GARCH, which assume stationarity and linearity. However, vanilla RNNs suffer from vanishing and exploding gradient problems, which limit their effectiveness in modelling long sequences. To overcome these challenges, advanced recurrent architectures such as LSTM and GRUs were introduced, both of which have become standard tools in financial time series modelling.

2.3.1 LSTMs

LSTM networks, introduced by Hochreiter & Schmidhuber (1997), address the vanishing gradient problem by incorporating memory cells and gating mechanisms that regulate the flow of information. LSTMs have been successfully applied in financial domains, particularly for capturing long-range dependencies in volatile markets. For example, Fischer & Krauss (2018) demonstrated the superiority of LSTMs over traditional ML models in predicting directional movements of the S&P 500 index. These results highlight the effectiveness of LSTMs for financial forecasting.

2.3.2 GRUs

Gated Recurrent Units simplify the LSTM architecture by combining the input and forget gates into a single update gate, thereby reducing computational cost while retaining strong performance (Chung et al. 2014). GRUs have been shown to perform comparably to LSTMs in many time series forecasting tasks, sometimes outperforming them in short-sequence settings due to their simpler structure. In financial forecasting, GRUs have been used in high-frequency trading scenarios where computational efficiency is critical.

2.4 Hybrid LSTMs

While the original LSTM has been highly effective in capturing long-term dependencies in sequential data, various modifications have been introduced to enhance its performance, reduce computational complexity, and address domain-specific challenges.

2.4.1 LSTM + XGBoost

XGBoost, introduced by Chen & Guestrin (2016), is a gradient boosting algorithm known for scalability and superior performance in tabular data. When paired with LSTMs, the hidden state representations produced by the LSTM layers can be used as features for XGBoost, enabling both temporal dependency capture and boosted feature refinement. In the study by Oukhouya et al. (2024), an LSTM–XGBoost hybrid was applied to forecasting daily prices of major international stock market indices using a grid-search-optimised pipeline. The hybrid model significantly outperformed stand-alone LSTM and XGBoost models, demonstrating superior predictive accuracy and robustness. This evidence reinforces the value of combining sequential deep learning with boosting ensembles for more reliable forecasts in financial domains.

2.4.2 LSTM + CATBoost

CATBoost, developed by Prokhorenkova et al. (2018), is also a gradient boosting framework specifically optimised for categorical and imbalanced datasets. Its combination with LSTM models has been explored to enhance predictive performance in financial markets. A study by Zhang et al. (2022) is found in short-term electricity spot price forecasting, where a hybrid model based on bidirectional LSTM (BiLSTM) and CATBoost demonstrated superior predictive performance compared to stand-alone deep learning or boosting models. By leveraging BiLSTM's ability to capture temporal dependencies and CATBoost's capacity for robust feature learning, the hybrid model achieved improved accuracy and generalisation. These findings suggest that similar architectures can be effectively adapted for financial time series.

2.5 Transformer Architectures

The widespread use of Transformers in natural language processing over the past few years has led to their extensive use in other areas like computer vision and time series forecasting. A transformer is a general sequence modelling tool that can approximate any function, according to its theoretical foundation. Furthermore, Transformer's multi-head mechanism enhances modelling capabilities and generalisation performance by enabling the model to process data in multiple subspaces in parallel Wu (2024).

2.5.1 Informer

Informer, proposed by Zhou et al. (2021), introduces the ProbSparse attention mechanism and a generative decoder to address the inefficiency of standard Transformers in long-sequence forecasting. By selectively attending to the most informative queries, Informer reduces computational complexity from quadratic to near-linear, making it feasible for intra-day financial data with thousands of time-steps. Benchmark results show Informer's effectiveness in electricity and weather datasets, while early applications to financial forecasting suggest its potential in handling structural breaks and high volatility. However, comparative studies highlight that its performance advantage over LSTMs and hybrid models in financial markets is not yet conclusive. Duan & Ke (2024) indicated that a fusion strategy combining the advantages of LSTM and Informer models is expected to enhance prediction accuracy and computational efficiency.

2.5.2 PatchTST

PatchTST, introduced by Nie et al. (2022), applies vision-inspired patching techniques to time series forecasting. PatchTST reduces noise sensitivity and improves temporal context learning by grouping subsequences into patches rather than processing the entire sequence token by token. On long-horizon forecasting benchmarks, this method has demonstrated strong performance. Although there is currently little empirical validation on intra-day commodity and foreign exchange markets, PatchTST exhibits promise in capturing seasonal patterns and local dependencies in financial contexts.

2.6 Literature Review Summary

Three general modelling approaches for intra-day financial forecasting are highlighted in the reviewed literature. First, despite frequent overfitting and instability problems, recurrent neural networks remain fundamental for capturing temporal dependencies. Second, by fusing structured feature exploitation with sequential representation learning, hybrid boosted LSTM architectures that employ XGBoost or CATBoost show promise, producing increased robustness and accuracy. Third, while transformer-based models like PatchTST and Informer are effective at managing long-range dependencies, their benefits in high-frequency gold markets are still unclear. In conclusion, these studies highlight the value of thorough evaluation procedures and inspire this work's concentration on comparing hybrid boosted models and Transformers for intra-day XAU/USD forecasting under a walk-forward validation.

Table 2.1: Literature review summary

#	Study	Method Description	Pros	Cons
1	**Fischer & Krauss (2018)**	LSTM applied to S&P 500 directional prediction	Captures sequential dependencies; high Sharpe ratio	Overfitting risk on high-frequency data
2	**Chung et al. (2014)**	GRU introduced as a simplified RNN variant	Fewer parameters faster training	May underperform on very long sequences
3	**Oukhouya et al. (2024)**	Hybrid LSTM+XGBoost for stock forecasting	Boosts accuracy over stand-alone LSTM or XGBoost	Requires complex tuning
4	**Zhang et al. (2022)**	Hybrid BiLSTM+CatBoost for electricity price forecasting	Combines sequential learning with boosting robustness	Applied outside finance; transferability uncertain
5	**Zhou et al. (2021)**	Informer with ProbSparse attention for long-sequence forecasting	Scales efficiently; reduces attention cost	Mixed results in volatile intra-day data
6	**Nie et al. (2022)**	PatchTST with temporal patching for time series	Competitive benchmark performance	Limited evidence in financial forecasting

Chapter 3

Methodology

3.1 Overview

High volatility, non-linearity, and the presence of both short-term and long-term dependencies in asset price movements are some of the main characteristics of financial markets. When applied to intra-day or high-frequency data, traditional econometric and statistical forecasting models frequently fail to adequately capture these dynamics. This leads to a research problem where precise and trustworthy forecasting is still very difficult, particularly in situations where institutions and investors depend on predictive insights to make decisions.

To address this problem, the present study evaluates three modern forecasting approaches: (i) a hybrid model that combines LSTM networks with Extreme Gradient Boosting (XGBoost)

(ii) a hybrid model that combines LSTM with CATBoost

(iii) the Informer, a time-series transformer-based architecture

The methodological framework employs walk-forward validation, ensuring that the models are tested in a realistic, time-consistent manner that mirrors actual financial prediction scenarios. In addition, experiments are applied on three different candlesticks to examine the robustness and generalisability of the models across varying market conditions. To further test adaptability, this study explores multiple prediction horizons (next 5, 15, 30, and 60 minutes' prices), thereby assessing how each model performs when forecasting across different temporal scales. The targeted prices in this research are high bid and low ask prices, for which a trader can maximise selling value or minimise buying cost.

This design enables a thorough comparison between transformer-based architectures and hybrid deep learning–boosting models. To determine which method offers the most consistent predictive performance in unstable financial environments, the study combines walk-forward evaluation with cross-dataset and multi-horizon testing. In the end, the proposed framework contributes both to the academic understanding of hybrid and transformer forecasting models and to the practical implications for high-frequency trading systems.

3.2 Research Framework

Figure 3.1 below illustrates the research framework adopted in this study, which defines the step-by-step pipeline from raw data acquisition to model deployment. The process begins with the collection of XAU/USD historical bid and ask prices at multiple resolutions (5-minute, 15-minute, and 60-minute candlesticks). The raw data is subjected to preprocessing, including merging, cleaning, feature engineering, and normalisation, to ensure quality and consistency. Shifted targets are then created to enable multi-horizon forecasting (5, 15, 30, and 60 minutes ahead), after which a walk-forward cross-validation approach is applied to preserve the temporal structure of financial time series and avoid data leakage. Three models—LSTM+XGBoost, LSTM+CatBoost, and Informer—are prepared and trained within this framework. Hyperparameter tuning and validation are performed to optimise model performance before evaluation. Finally, the results are assessed using standard error metrics, and the best-performing models are considered for potential deployment. This systematic framework ensures reproducibility, scalability, and fair comparison across all forecasting approaches.

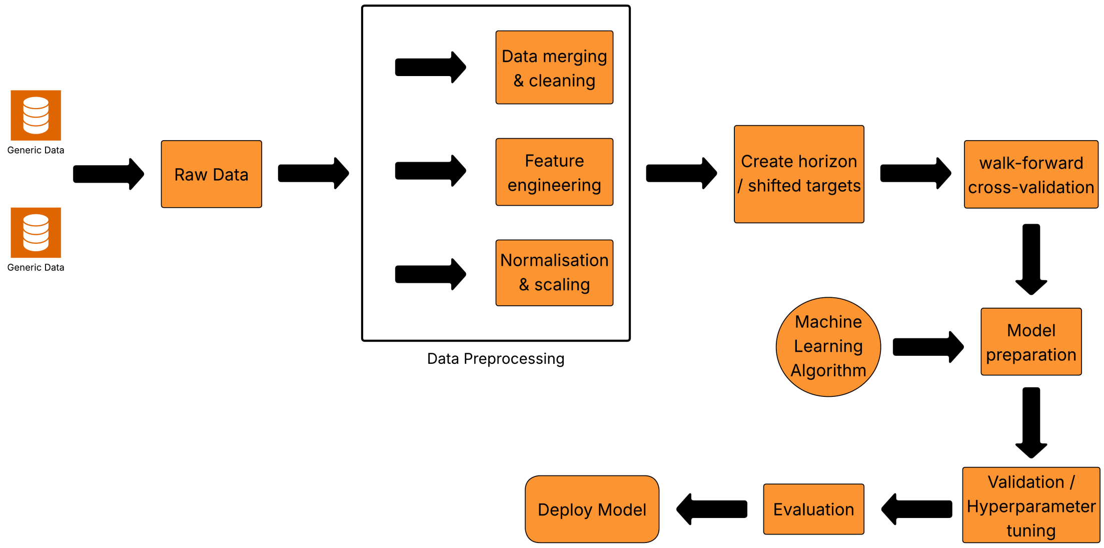


Figure 3.1: Overall research framework

3.3 Dataset

The raw data used in this research were obtained from Dukascopy's historical market data service, which provides high-frequency tick-level bid and ask price records for foreign exchange and commodities. For this study, the XAU/USD pair was selected over the period 1st January 2024 to 29th March 2025. For each time frame, two distinct datasets were downloaded (bid and ask) and later merged to ensure alignment between both sides of the market. Each dataset initially contains candlestick values directly computed from the tick data, with the following attributes for every time window:

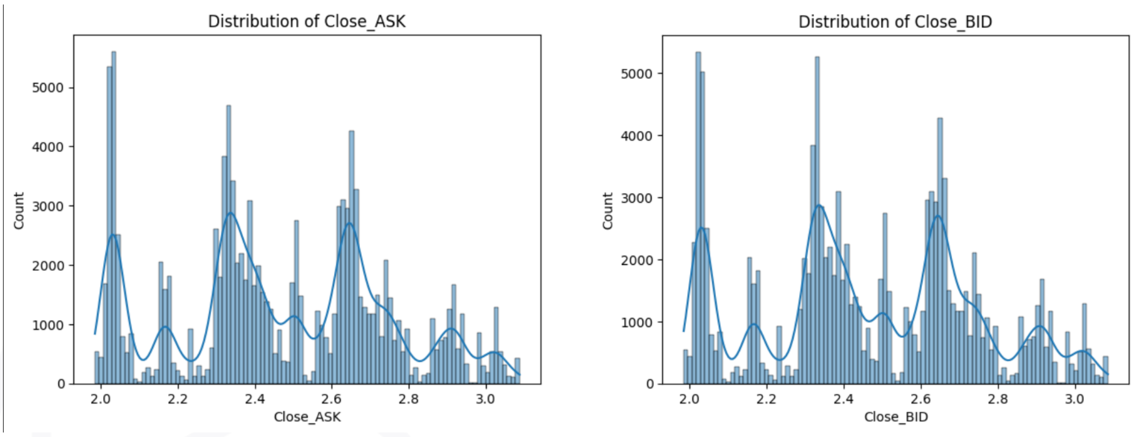
- Open – the first recorded price within the interval
- High – the maximum observed price within the interval
- Low – the minimum observed price within the interval
- Close – the final recorded price within the interval
- Volume - the number of price changes(ticks) during the candle's period

This procedure was repeated for three separate granularities of candlestick aggregation:

- 5-minute candles: around 130,000 rows
- 15-minute candles: around 43,000 rows
- 60-minute candles: around 11,000 rows

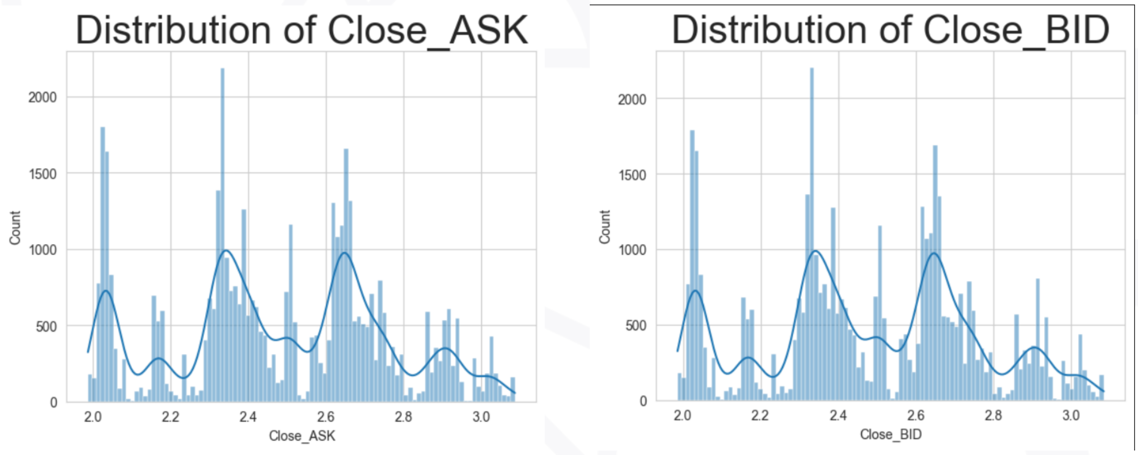
The histograms in Figures 3.2, 3.3, and 3.4 below illustrate the distribution of closing ask and bid prices for the XAU/USD pair across the three different time frames. A consistent and key observation across all six distributions is their nearly identical shape between ask and bid prices, indicating an extremely narrow and stable ask-bid spread.

From a financial standpoint, this narrow ask-bid spread implies that the XAU/USD asset was trading with strong participation from both buyers and sellers throughout the period, allowing transactions to occur with minimal difference between the purchase and sale price.



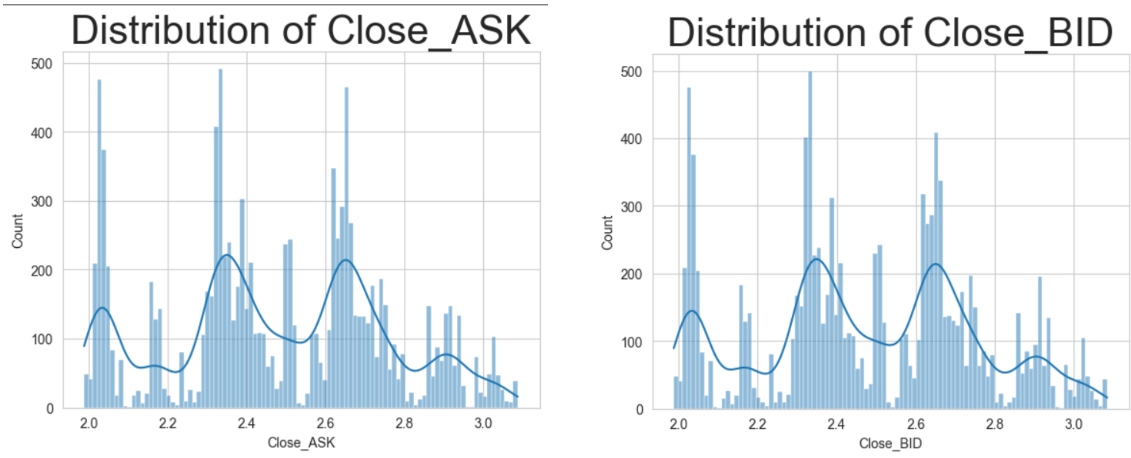
(a) Distribution of Close_ASK (5-min) **(b)** Distribution of Close_BID (5-min)

Figure 3.2: Distribution of close prices (Ask and Bid) for the 5-minute dataset.



(a) Distribution of Close_ASK (15-min) **(b)** Distribution of Close_BID (15-min)

Figure 3.3: Distribution of close prices (Ask and Bid) for the 15-minute dataset.



(a) Distribution of Close_ASK (1-hr) **(b)** Distribution of Close_BID (1-hr)

Figure 3.4: Distribution of close prices (Ask and Bid) for the 1-hour dataset.

3.4 Data Preprocessing

The preprocessing stage was crucial to ensure that the raw market data could be used effectively by the forecasting models.

3.4.1 Merging Datasets

The first step involved merging the bid and ask datasets on the common timestamp column; this step was necessary to ensure synchronisation in the data.

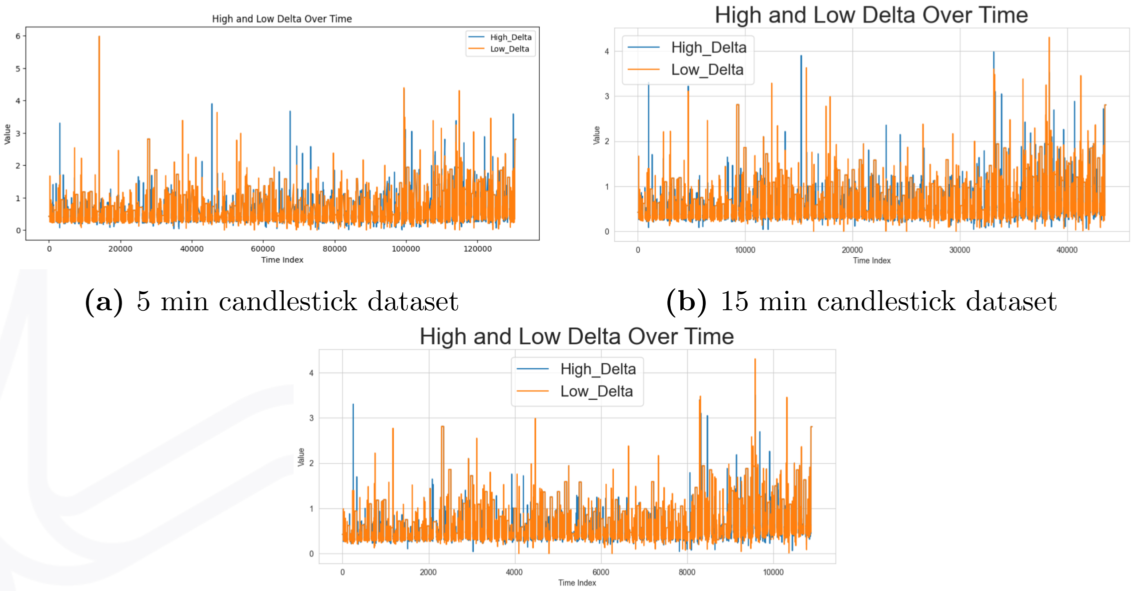
3.4.2 Missing Values

After merging, the dataset was checked for null values and timestamp consistency, and rows with missing values were dropped. This ensured that the time series remained continuous without gaps that could disturb training.

3.4.3 Feature Engineering

To enrich the dataset beyond the standard OHLC (Open, High, Low, Close) and volume values, several engineered features were created. Specifically, the Volume Delta (Volume_ASK - Volume_BID) and its absolute form were added to capture the imbalance in buying and selling pressure. Similarly, price deltas were computed across all four OHLC values to reflect the spread between bid and ask prices, a critical indicator of short-term liquidity. Additionally, two binary features were derived: New_Day (indicating when a new trading day begins) and New_Week (marking the transition between weekly cycles). These temporal features provide context for periodic patterns that may affect price behaviour.

Two target variables were then created (Y_High_BID and Y_Low_ASK). These serve as the prediction objectives, reflecting the maximum price at which a trader could sell (high bid price) and the minimum price at which a trader could buy (low ask).



(a) 5 min candlestick dataset

(b) 15 min candlestick dataset

(c) 1 hr candlestick dataset

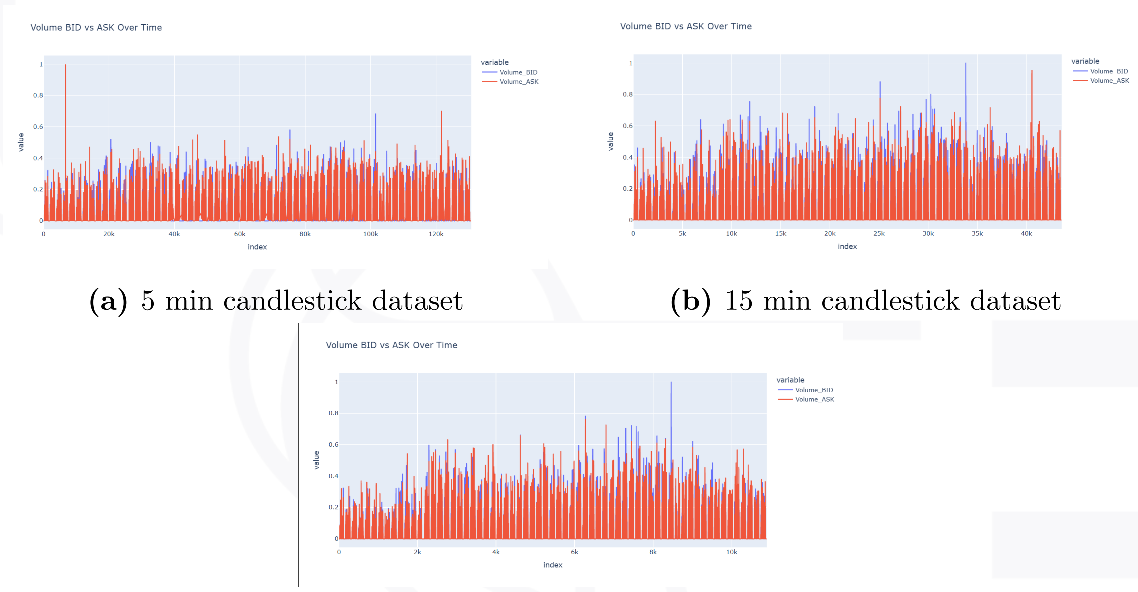
Figure 3.5: High and Low Delta Over Time

Figure 3.5 illustrates the evolution of the High_Delta and Low_Delta values across the three datasets (5-minute (3.5a), 15-minute (3.5b), and 60-minute (3.5c)). Across all time frames, the deltas generally fluctuate within a narrow band close to one, with occasional sharp spikes that might reflect periods of heightened market volatility. In the 15-minute and 60-minute datasets, the deltas are more frequent and noisy, capturing the rapid micro-movements of the market, while in the 5-minute dataset, these patterns smooth out, though large fluctuations are still visible. These deltas are significant from a financial standpoint because they measure the magnitude of price fluctuations over each time period, acting as an indicator of short-term volatility that can affect trading decisions and the performance of forecasting models.

3.4.4 Normalisation

The preprocessing also included data normalisation. The dataset was divided into training (80%) and testing (20%) sets, with the training set used to extract the scaling parameters. Price values were rescaled by dividing by 1000, shifting the range to approximately 1.98 to 2.79, while volume values were normalised relative to their maximum value calculated from the training set. Delta features didn't need any scaling since they had a maximum value of 5.981 and a minimum value of 0.001.

The plots in figure 3.6 show that the normalised trading volumes of bids and asks exhibit high frequency variations, reflecting the short-term dynamics of order flow in the market. Both bid and ask volumes remain mostly below 0.6 (after normalisation), with occasional spikes suggesting periods of heightened liquidity demand, often coinciding with increased market activity such as news releases or overlapping trading seasons. These confirm that smaller candlesticks effectively capture intra-day volatility and liquidity shocks, making it suitable for modelling rapid market movements.



(a) 5 min candlestick dataset

(b) 15 min candlestick dataset

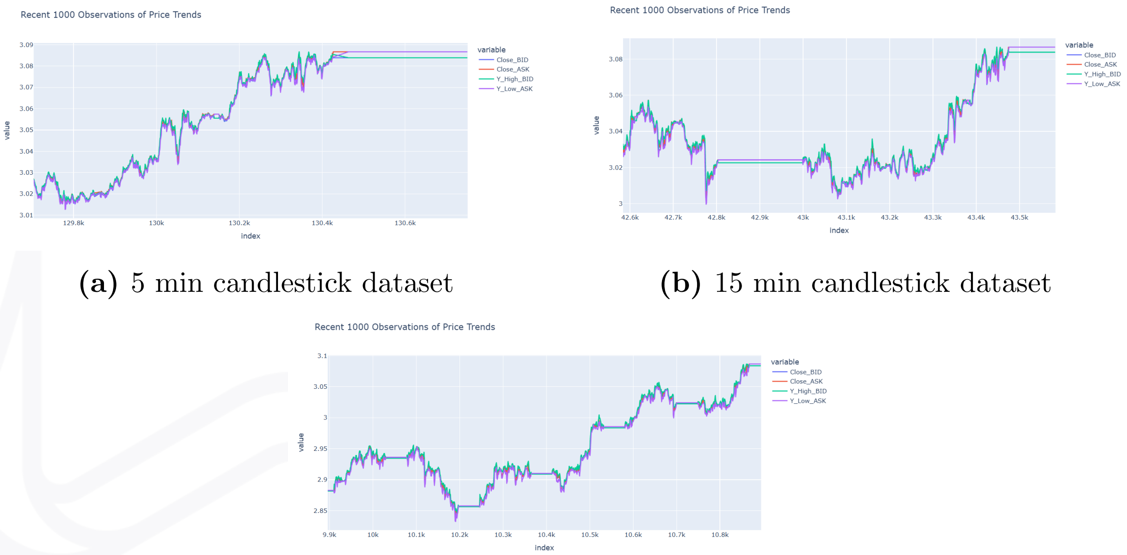
(c) 1 hr candlestick dataset

Figure 3.6: Volume Bid vs Ask Over Time

3.4.5 Memory Optimisation

Then, to improve memory efficiency, numerical columns were downcast from 64-bit to 32-bit floats and from standard integers to smaller integer types. This optimisation reduced memory usage by half, which is particularly important given the large counts in the datasets.

Finally, since different model architectures required different inputs, two dataset versions were saved: one with target columns for the LSTM-based models and one without target columns for the Informer model, which generates predictions without explicit label columns in training.



(a) 5 min candlestick dataset

(b) 15 min candlestick dataset

(c) 1 hr candlestick dataset

Figure 3.7: Recent 1000 Observations of Gold Price Trends

The four-line charts in Figure 3.7 present the recent 1000 observations of XAU/USD price trends. Each graph reveals consistent upward momentum in both Close_BID and Close_ASK values, with Y_High_BID and Y_Low_ASK providing a broader view of intra-period volatility. Notably, segments of horizontal lines, especially in the shorter time frames, suggest intervals of off-hours trading, where market activity slows and prices remain static due to reduced liquidity. Financially, traders might interpret this setup as a signal for continued upward momentum, especially if supported by volume and breakout confirmations.

3.5 Cross Validation

For this study, walk-forward validation was employed as the evaluation strategy, as it is particularly well-suited for time series forecasting tasks in financial markets, as time series data exhibits strong temporal dependencies that must be preserved.
To implement walk-forward validation, two dedicated functions were designed to meet the structural requirements of the models employed in this study. For the hybrid models (LSTM-XGBoost and LSTM-CATBoost), the function walk_forward_split was implemented. This function takes as input the already preprocessed three-dimensional data tensors representing the features (X) and targets (y), along with the training and testing block sizes, where targets are already shifted based on the forecasting horizon (next 5, 15, 30, or 60 minutes). The function then outputs four three-dimensional arrays: X_train, X_test, y_train, and y_test, which correspond to sequential blocks of training and testing sets. This ensures that at each iteration,

the model is trained on past data and evaluated on unseen future data while maintaining the chronological order of the series.

For the Informer model, which has stricter input requirements, a separate function named walk_forward_split_for_informer was developed. Instead of working directly with input tensors, this function takes the full data frame that's already been preprocessed for informer training, as discussed in section 3.4, the training and testing block sizes, and the directory in which to store blocks. At each iteration, the function extracts the relevant portion of the data frame, covering both the training and testing horizons, and saves it as an independent CSV file. This design was necessary to integrate with the Informer's native data-loading pipeline, which reads directly from file-based datasets. The function then returns the total number of blocks created.

By tailoring the walk-forward validation process to both model categories, this study ensured consistency across experiments while addressing model-specific requirements. This design not only preserved temporal dependencies and avoided data leakage but also closely simulated real-world forecasting conditions where financial models are applied incrementally as new data becomes available.

Table 3.1: Walk-forward block configuration for each data frame

Dataset	**Total Length**	**Train Size**	**Test Size**	**Number of blocks**
5-minute	118,096	16,500	2,500	7
15-minute	43,584	10,000	2,000	4
60-minute	10,896	5,000	800	2

3.6 Model Development

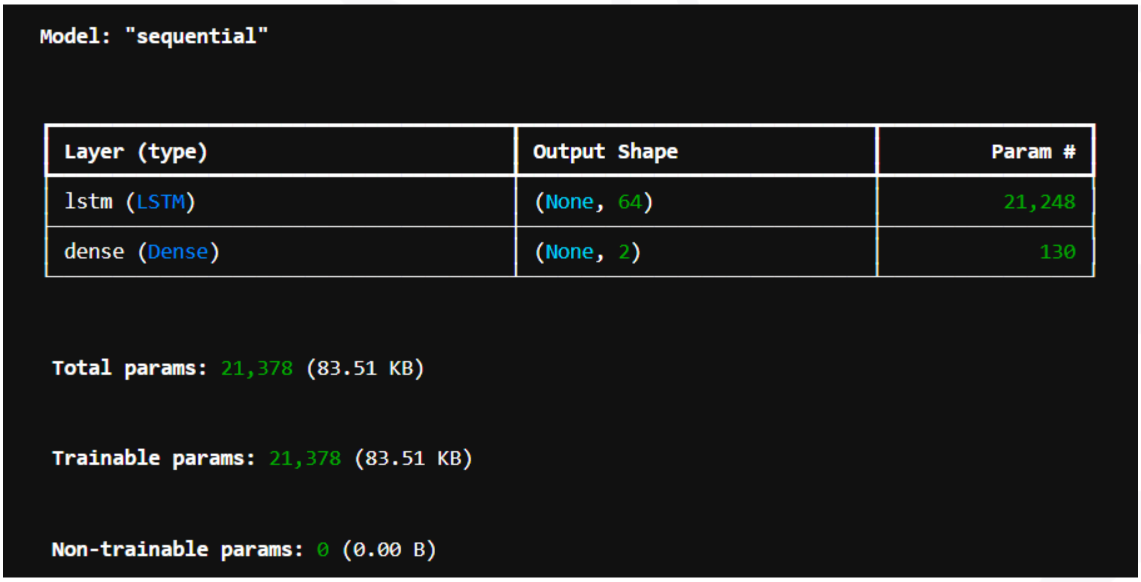
This study employed three distinct predictive models: LSTM-XGBoost, LSTM-CATBoost, and Informer. Each was trained and evaluated on the same datasets under the same walk-forward validation framework to ensure consistency and comparability of results. The following subsections describe the design, implementation, and rationale of each model in detail.

3.6.1 Hybrid LSTM-XGBoost

The first model combined an LSTM network with XGBoost regressors. The LSTM-XGBoost hybrid model was designed to leverage the sequential learning capabilities of LSTM networks and the robust predictive power of the Extreme Gradient Boosting (XGBoost) algorithm. The LSTM network consisted of a single LSTM layer with 64 hidden units and the Rectified Linear Unit (ReLU) activation function, followed by a fully connected dense layer with two outputs corresponding to the high bid and low ask prices. The model was trained using the Adam optimiser, a popular first-order gradient-based optimisation algorithm, and mean squared error (MSE) loss, with early stopping applied to prevent overfitting. Then, under a loop over the block sets, the LSTM is trained and used as a feature extractor: it transforms the 3D inputs into 2D feature representations. These features were then passed to two independent XGBoost regressors, one for predicting the high bid price and one for predicting the low ask price. For each walk-forward block, evaluation metrics (RMSE, MAE, R^2) are calculated and appended into a results array.

3.6.2 Hybrid LSTM-CATBoost

The second model followed the same pipeline as the LSTM–XGBoost hybrid but replaced XGBoost with CATBoost regressors.



```
Model: "sequential"

┌──────────────────────────────┬──────────────────────┬─────────────┐
│ Layer (type)                 │ Output Shape         │     Param # │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ lstm (LSTM)                  │ (None, 64)           │      21,248 │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ dense (Dense)                │ (None, 2)            │         130 │
└──────────────────────────────┴──────────────────────┴─────────────┘

 Total params: 21,378 (83.51 KB)

 Trainable params: 21,378 (83.51 KB)

 Non-trainable params: 0 (0.00 B)
```

Figure 3.8: LSTM Network Summary

Figure 3.8 presents the architectural summary of the LSTM network used, comprising the two primary layers. The first layer is an LSTM unit with 64 neurons, producing an output shape of (None, 64) and containing 21,248 parameters. This layer is responsible for capturing temporal dependencies and sequential patterns within the input data. The subsequent dense (fully connected) layer has 2 neurons, resulting in an output shape of (None, 2) and contributing 130 parameters.

This lightweight yet efficient architecture is well-suited for sequence modelling tasks, particularly for feature extraction.

3.6.3 Informer

The third model employed the Informer architecture, a state-of-the-art Transformer-based model optimised for long-sequence time-series forecasting. The implementation is based directly on the official GitHub repository (Zhou et al. 2020), with modifications to adapt it to the datasets and multiple forecast horizons used in this study. Informer follows an encoder–decoder architecture. The encoder extracts global temporal dependencies and contextual information from past sequences, while the decoder generates forecasts for the required horizons.

Similarly to the hybrid models described above, Informer is trained by looping over the walk-forward blocks, where each block represented a distinct training–testing window. After training on each block, predictions are generated for the corresponding test set, and the evaluation metrics are appended to a results array. At the end of the process, the results from all blocks are averaged, providing a stable and reliable estimate of the model's forecasting performance across the entire dataset.

3.7 Evaluation Metrics

To assess the performance of the proposed models (LSTM-XGBoost, LSTM-CATBoost, and Informer), a set of standard regression metrics will be employed. These metrics capture both the accuracy of predictions and the models' ability to generalise across different forecasting horizons (5, 15, 30, and 60 minutes).

3.7.1 Mean Absolute Error (MAE)

The Mean Absolute Error measures the average magnitude of the errors in the predictions. It is calculated by taking the absolute differences between predicted and actual values and averaging them.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the actual value, and $\hat{y}_i$ is the predicted value. MAE is useful because its error units match the target value's units, changes in MAE are linear and intuitive, and it gives equal weight to all errors without being skewed by large outliers.

3.7.2 Root Mean Squared Error (RMSE)

The Root Mean Squared Error measures the average magnitude of errors between predicted and actual values, providing a score in the same units as the data, which makes it an effective metric for evaluating the accuracy of a model, especially a regression model. It's calculated by finding the difference between predictions and truths, squaring those differences, averaging them, and then taking the square root.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

RMSE is particularly useful in financial forecasting tasks where large deviations can be more critical than smaller ones.

3.7.3 Coefficient of Determination

The Coefficient of Determination, or R^2 score, measures how well a statistical model explains the variability in a dependent variable. It is a value between 0 and 1, where a higher score indicates a better fit of the model to the data. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

where $\bar{y}_i$ represents the mean of the observed values.

Chapter 4

Results

4.1 Overview

This chapter presents the experimental results obtained from the three forecasting models: the hybrid LSTM-XGBoost, the hybrid LSTM-CatBoost, and the Informer. Results are reported across three datasets sampled at 5-minute, 15-minute, and 1-hour intervals, with forecasts generated for horizons of 5, 15, 30, and 60 minutes ahead, depending on the dataset.

Performance is measured using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2), separately for the High_BID and Low_ASK targets. All reported values correspond to averages across the walk-forward validation blocks.

Overall, the results suggest that hybrid approaches are more robust for shorter datasets and horizons, while the Informer showed potential for short-term forecasts on the high-frequency 5-minute data.

The following sections provide a detailed breakdown of results for each model and dataset, supported by tables and visualisations to highlight performance patterns across horizons.

4.2 5-Minute Dataset Results

The 5-minute dataset provided the highest-frequency view of the market in this study, with more than 100 thousand rows covering bid and ask movements. This fine-grained data enabled the models to capture rapid short-term dynamics, but it also introduced noise due to market microstructure effects. Each model was evaluated on the prediction of two separate targets: the next High Bid and the next Low Ask.

For the next-5-minute horizon, the LSTM-CAT hybrid clearly outperformed the other models, achieving an RMSE of 0.0097 for both High Bid and Low Ask, paired with an MAE of 0.0061. This translated into strong explanatory power, with R^2 values of 0.7992 (High Bid) and 0.7934 (Low Ask). In comparison, LSTM-XGB recorded slightly weaker metrics (RMSE ≈ 0.012/0.012, MAE ≈ 0.006/0.007, $R^2 \approx$ 0.619/0.630). Interestingly, although Informer produced much larger errors (RMSE ≈ 0.279/0.272, MAE ≈ 0.17/0.162), its R^2 values (0.7637/0.7976) remained competitive, suggesting that the model was able to capture general directional movements, but with poorer precision in the magnitude of predictions.

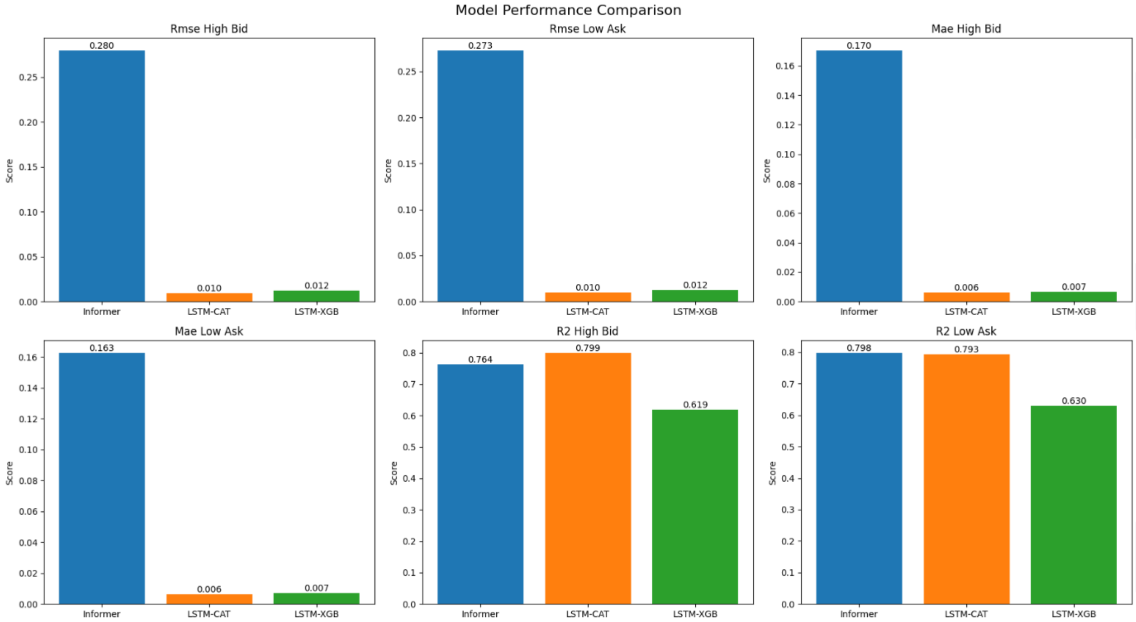


Figure 4.1: Next 5-minute Metrics Achieved Using 5-minute Dataset Histograms

As the forecasting horizon extended, the hybrids remained consistently superior. At the next-15-minute horizon, LSTM-CAT achieved RMSE values of 0.0095 (High Bid) and 0.0097 (Low Ask), with corresponding R^2 scores of 0.8142 and 0.8088, slightly surpassing LSTM-XGB (RMSE ≈ 0.009/0.009, $R^2 \approx$ 0.802/0.796). Informer's error values increased (RMSE ≈ 0.319/0.322), with weaker R^2 scores (≈ 0.57/0.56), showing difficulties in handling noisy intra-day volatility at this horizon.

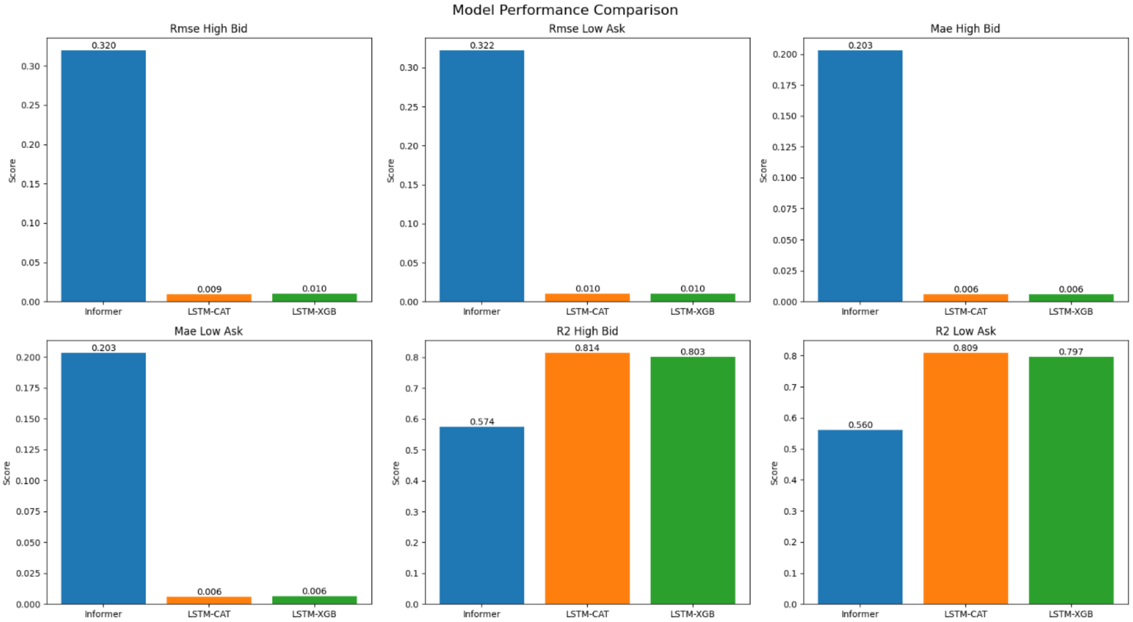
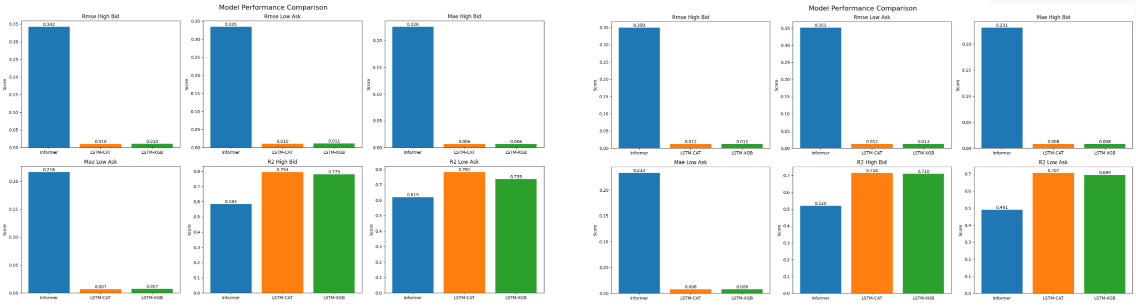


Figure 4.2: Next 15-minute Metrics Achieved Using 5-minute Dataset Histograms

At longer horizons (30 and 60 minutes), both hybrids maintained good stability: LSTM-CAT sustained R^2 values above 0.7, and LSTM-XGB remained between 0.69–0.77. While Informer degraded further, with RMSE values above 0.3 and R^2 scores dropping near 0.49–0.61. Overall, the results confirm that while the Informer architecture struggles with high-frequency data, the hybrid LSTM models, particularly LSTM-CAT, are well-suited to short-term financial forecasting with both High Bid and Low Ask targets.



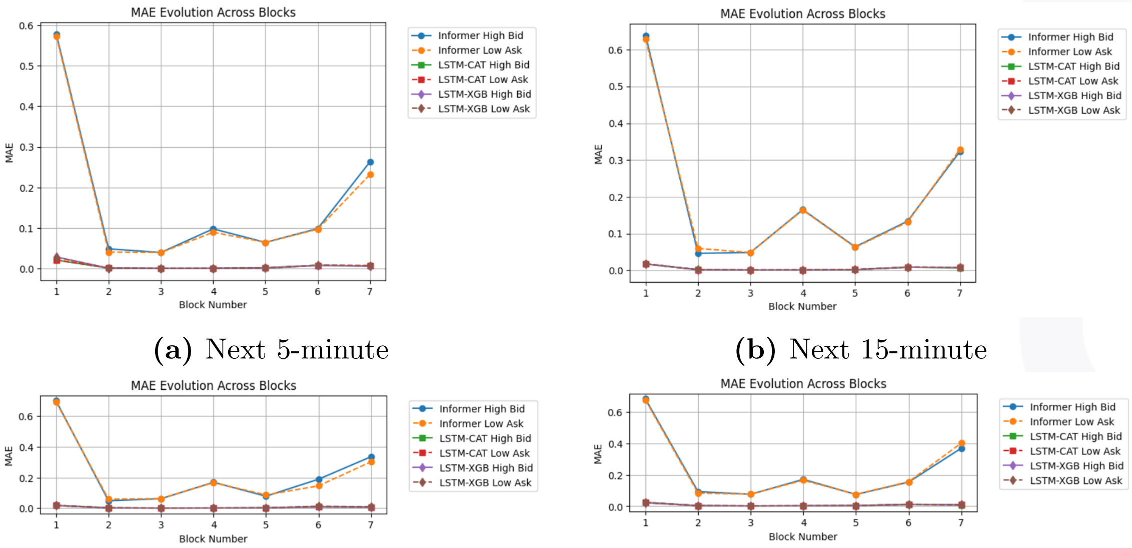
(a) Next 30-minute Metrics

(b) Next 60-minute Metrics

Figure 4.3: Next 30-minute and 60-minute Metrics Achieved Using 5-minute Dataset Histograms

Table 4.1: Summary of metrics achieved using 5-minute dataset

Model	Next 5-min	Next 15-min	Next 30-min	Next 60-min
LSTM-XGB	rmse high_bid: 0.0124 rmse low_ask: 0.0123 mae high_bid: 0.0068 mae low_ask: 0.0070 r2 high_bid: 0.6191 r2 low_ask: 0.6303	rmse high_bid: 0.0097 rmse low_ask: 0.0099 mae high_bid: 0.0058 mae low_ask: 0.0061 r2 high_bid: 0.8028 r2 low_ask: 0.7968	rmse high_bid: 0.0104 rmse low_ask: 0.0113 mae high_bid: 0.0064 mae low_ask: 0.0071 r2 high_bid: 0.7793 r2 low_ask: 0.7354	rmse high_bid: 0.0122 rmse low_ask: 0.0125 mae high_bid: 0.0079 mae low_ask: 0.0080 r2 high_bid: 0.7100 r2 low_ask: 0.6938
LSTM-CAT	rmse high_bid: 0.0097 rmse low_ask: 0.0097 mae high_bid: 0.0061 mae low_ask: 0.0061 r2 high_bid: 0.7992 r2 low_ask: 0.7934	rmse high_bid: 0.0095 rmse low_ask: 0.0097 mae high_bid: 0.0057 mae low_ask: 0.0059 r2 high_bid: 0.8124 r2 low_ask: 0.8088	rmse high_bid: 0.0101 rmse low_ask: 0.0105 mae high_bid: 0.0062 mae low_ask: 0.0066 r2 high_bid: 0.7973 r2 low_ask: 0.7809	rmse high_bid: 0.0120 rmse low_ask: 0.0122 mae high_bid: 0.0079 mae low_ask: 0.0080 r2 high_bid: 0.7157 r2 low_ask: 0.7067
Informer	rmse high_bid: 0.2795 rmse low_ask: 0.2726 mae high_bid: 0.1702 mae low_ask: 0.1626 r2 high_bid: 0.7637 r2 low_ask: 0.7976	rmse high_bid: 0.3196 rmse low_ask: 0.3220 mae high_bid: 0.2028 mae low_ask: 0.2033 r2 high_bid: 0.5739 r2 low_ask: 0.5598	rmse high_bid: 0.3421 rmse low_ask: 0.3346 mae high_bid: 0.2255 mae low_ask: 0.2162 r2 high_bid: 0.5838 r2 low_ask: 0.6190	rmse high_bid: 0.3495 rmse low_ask: 0.3510 mae high_bid: 0.2312 mae low_ask: 0.2334 r2 high_bid: 0.5202 r2 low_ask: 0.4907



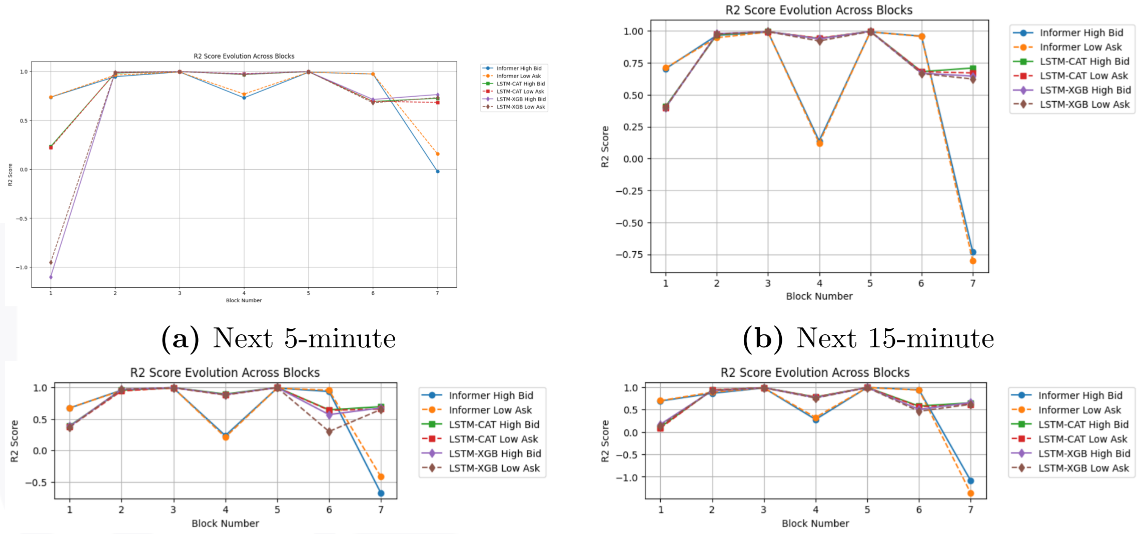
(a) Next 5-minute

(b) Next 15-minute

(c) Next 30-minute

(d) Next 60-minute

Figure 4.4: MAE Evolution Across All Blocks Using 5-minute Dataset



(a) Next 5-minute

(b) Next 15-minute

(c) Next 30-minute

(d) Next 60-minute

Figure 4.5: R^2 Evolution Across All Blocks Using 5-minute Dataset

4.3 15-Minute Dataset Results

The 15-minute dataset contained around 43,000 rows, providing a balance between detail and aggregation. This time frame reduced market noise while still retaining enough granularity to reflect intra-day fluctuations. As with the 5-minute dataset, both targets (High Bid and Low Ask) were forecasted simultaneously across horizons of 15, 30, and 60 minutes.

For the next-15-minute horizon, the LSTM-CAT hybrid delivered the strongest overall performance, with RMSE values of only 0.0051 for both High Bid and Low Ask, accompanied by extremely low MAE values of 0.0032 and 0.0031. The R^2 scores reached 0.9703 and 0.9714, respectively, indicating that nearly all variance in the series was explained by the model. This significantly surpassed LSTM-XGB, which, although it achieved strong results (RMSE $\approx$ 0.008/0.009, $R^2 \approx$ 0.883/0.872), still lagged in precision. Informer achieved respectable explanatory power at this horizon ($R^2 \approx$ 0.89/0.88), but with much higher errors (RMSE $\approx$ 0.167/0.174), showing it could track trends but lacked the predictive sharpness of the hybrids.

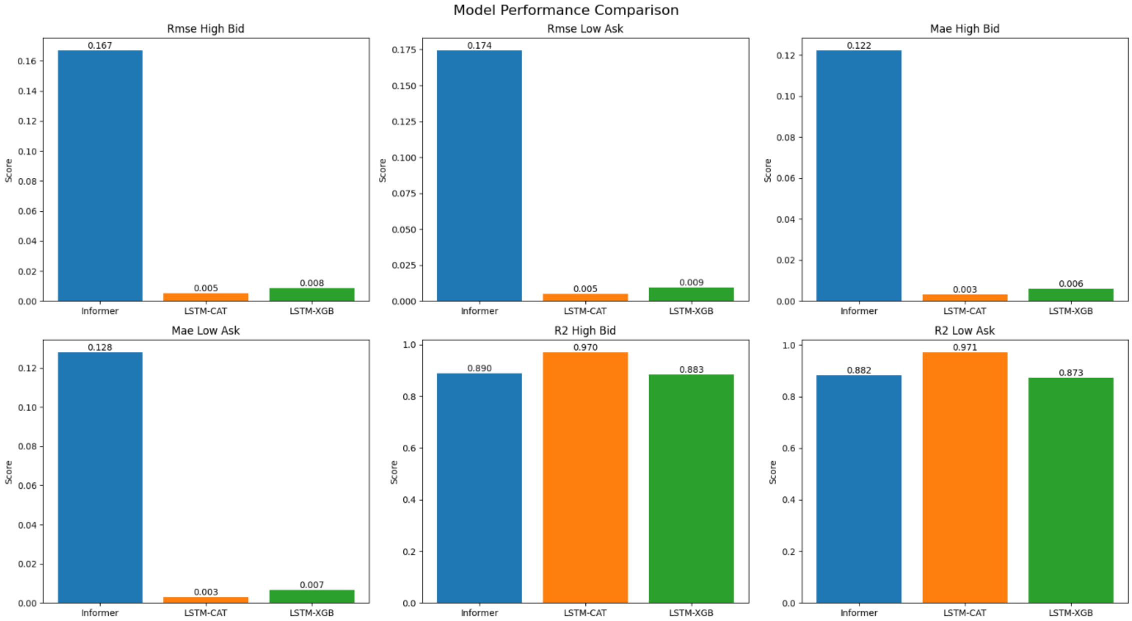


Figure 4.6: Next 15-minute Metrics Achieved Using 15-minute Dataset Histograms

At the next 30-minute horizon, results were more mixed. LSTM-XGB outperformed LSTM-CAT, recording RMSE values of 0.0079/0.0078 and $R^2 \approx 0.932/0.933$, whereas LSTM-CAT dropped slightly in performance ($R^2 \approx 0.86/0.85$). Informer's results weakened substantially, with errors above 0.24 and R^2 scores around 0.61/0.65.

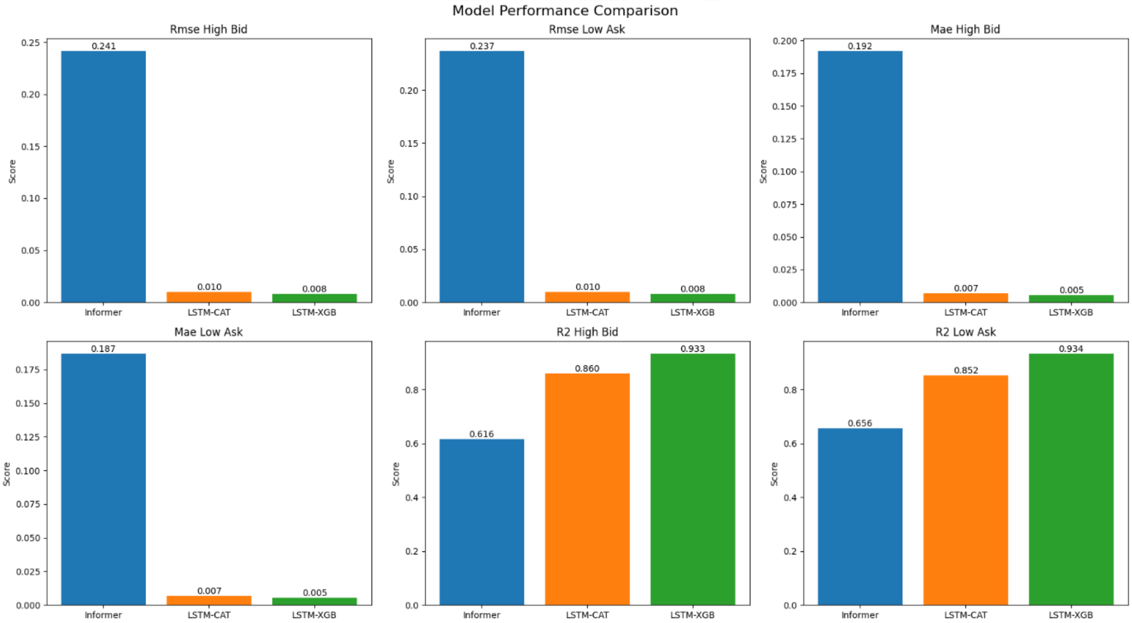


Figure 4.7: Next 30-minute Metrics Achieved Using 15-minute Dataset Histograms

Finally, at the 60-minute horizon, both hybrids converged to very strong performance: LSTM-XGB achieved RMSE values of 0.007/0.007 with $R^2 \approx 0.951/0.945$, while LSTM-CAT remained competitive (RMSE $\approx 0.0071/0.0073$, $R^2 \approx 0.948/0.945$).

Informer again trailed (RMSE ≈ 0.213/0.211, $R^2 \approx 0.72/0.73$).

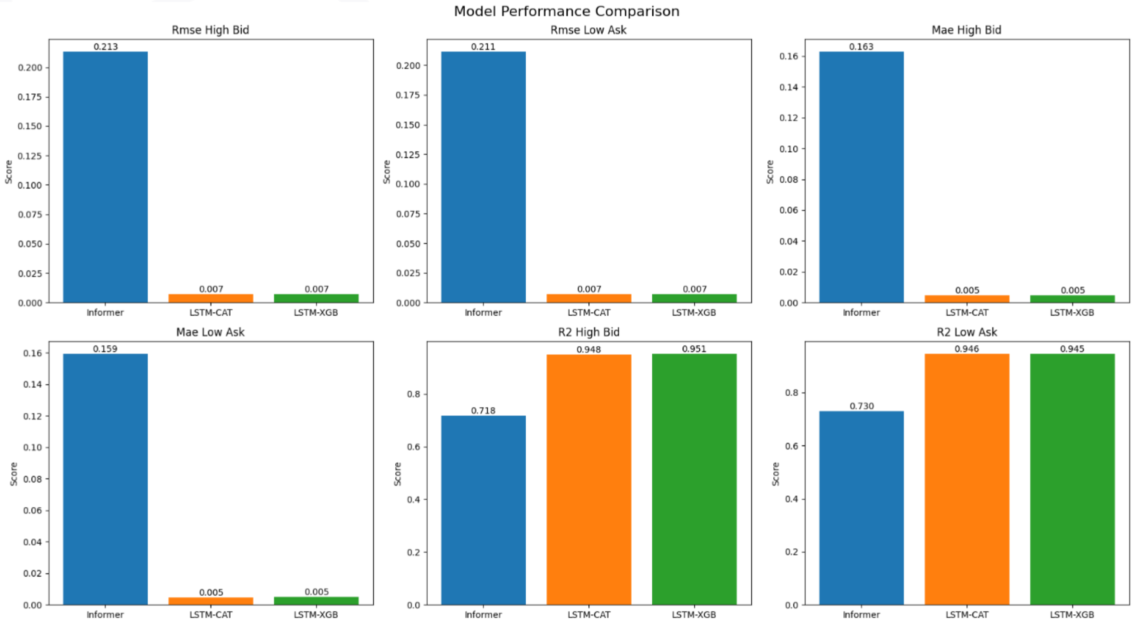
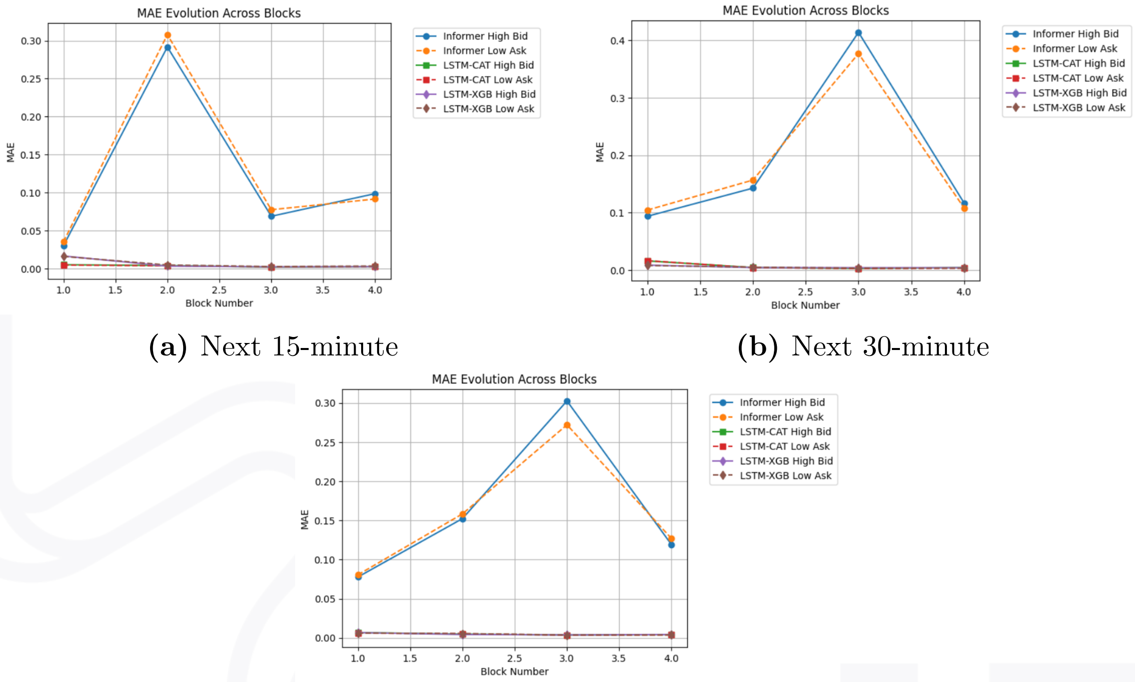


Figure 4.8: Next 60-minute Metrics Achieved Using 15-minute Dataset Histograms

These results highlight that for the 15-minute dataset, both hybrids dominate across all horizons, with LSTM-CAT excelling in very short-term forecasts (15 minutes ahead) and LSTM-XGB excelling at longer horizons (30–60 minutes ahead). Informer, though competitive in explanatory power at the shortest horizon, was significantly less precise overall.

Table 4.2: Summary of metrics achieved using 15-minute dataset

Model	**Next 15-min**	**Next 30-min**	**Next 60-min**
LSTM-XGB	rmse high_bid: 0.0084	rmse high_bid: 0.0079	rmse high_bid: 0.0070
	rmse low_ask: 0.0094	rmse low_ask: 0.0078	rmse low_ask: 0.0074
	mae high_bid: 0.0060	mae high_bid: 0.0055	mae high_bid: 0.0048
	mae low_ask: 0.0065	mae low_ask: 0.0051	mae low_ask: 0.0048
	r2 high_bid: 0.8835	r2 high_bid: 0.9328	r2 high_bid: 0.9510
	r2 low_ask: 0.8726	r2 low_ask: 0.9337	r2 low_ask: 0.9450
LSTM-CAT	rmse high_bid: 0.0051	rmse high_bid: 0.0099	rmse high_bid: 0.0071
	rmse low_ask: 0.0051	rmse low_ask: 0.0100	rmse low_ask: 0.0073
	mae high_bid: 0.0032	mae high_bid: 0.0070	mae high_bid: 0.0048
	mae low_ask: 0.0031	mae low_ask: 0.0067	mae low_ask: 0.0046
	r2 high_bid: 0.9703	r2 high_bid: 0.8598	r2 high_bid: 0.9484
	r2 low_ask: 0.9714	r2 low_ask: 0.8523	r2 low_ask: 0.9458
Informer	rmse high_bid: 0.1670	rmse high_bid: 0.2414	rmse high_bid: 0.2132
	rmse low_ask: 0.1743	rmse low_ask: 0.2367	rmse low_ask: 0.2114
	mae high_bid: 0.1222	mae high_bid: 0.1919	mae high_bid: 0.1628
	mae low_ask: 0.1280	mae low_ask: 0.1867	mae low_ask: 0.1594
	r2 high_bid: 0.8902	r2 high_bid: 0.6162	r2 high_bid: 0.7181
	r2 low_ask: 0.8819	r2 low_ask: 0.6557	r2 low_ask: 0.7296

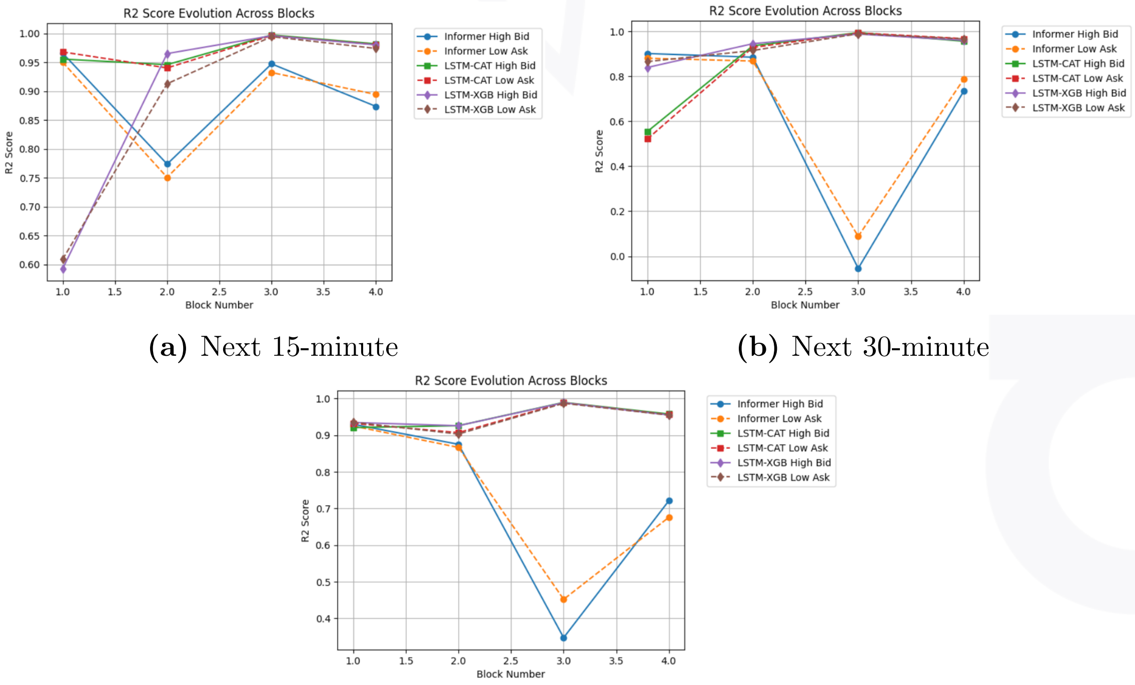


(a) Next 15-minute

(b) Next 30-minute

(c) Next 60-minute

Figure 4.9: MAE Evolution Across All Blocks Using 15-minute Dataset



(a) Next 15-minute

(b) Next 30-minute

(c) Next 60-minute

Figure 4.10: R^2 Evolution Across All Blocks Using 15-minute Dataset

4.4 1-Hour Dataset Results

The 60-minute dataset was the smallest of the three, containing around 10,896 rows. This lower data density provided a broader view of market movements, but at the cost of losing intra-hour granularity. Forecasting was restricted to the next 60-minute horizon, as shorter horizons could not be derived from this coarser dataset. Both hybrid models delivered comparable performance. LSTM-XGB achieved RMSE values of 0.0352 (High Bid) and 0.0380 (Low Ask), with MAE $\approx 0.024/0.026$ and $R^2 \approx 0.53/0.46$. LSTM-CAT produced nearly identical outcomes (RMSE $\approx 0.035/0.038$, $R^2 \approx 0.52/0.46$). These values reflect the increased difficulty of prediction at this resolution, where volatility is more pronounced and fewer data points are available to capture complex intra-hour dynamics.

Informer struggled substantially on this dataset, recording RMSE values of 0.3671/0.3398, MAE $\approx 0.29/0.27$, and lower $R^2 \approx 0.37/0.46$. This indicates that the architecture, while suited for long-sequence forecasting, does not translate effectively when trained on coarser financial datasets with fewer samples.

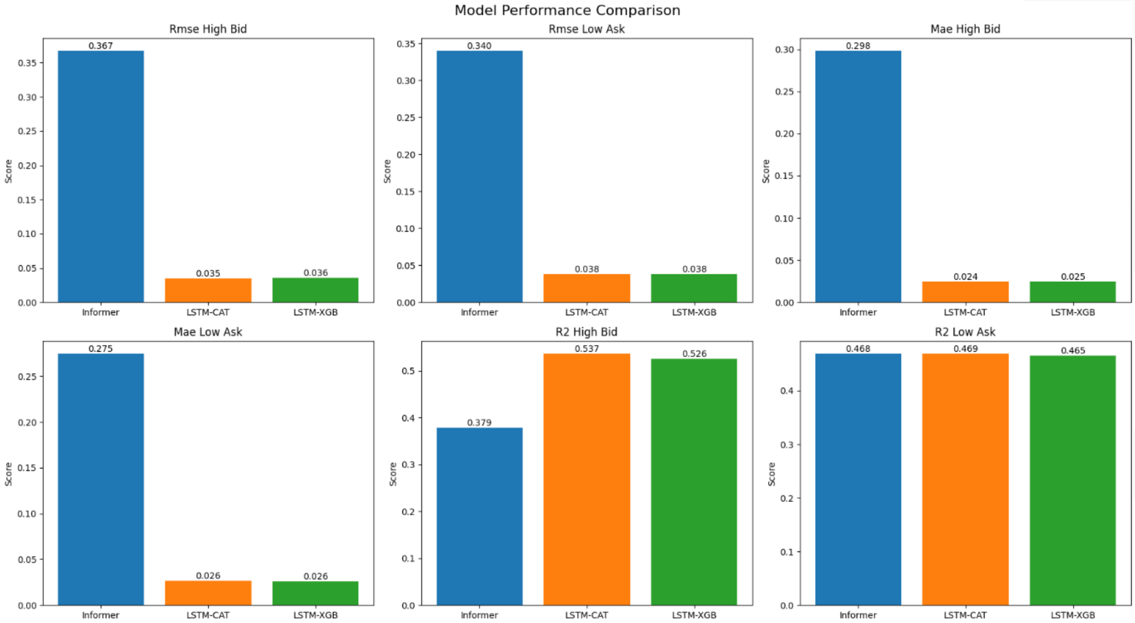


Figure 4.11: Next 60-minute Metrics Achieved Using 1-hour Dataset Histograms

Table 4.3: Summary of metrics achieved using 1-hour dataset

Model	**Next 60-min**
LSTM-XGB	rmse high_bid: 0.0352 rmse low_ask: 0.0380 mae high_bid: 0.0242 mae low_ask: 0.0263 r2 high_bid: 0.5365 r2 low_ask: 0.4688
LSTM-CAT	rmse high_bid: 0.0356 rmse low_ask: 0.0381 mae high_bid: 0.0246 mae low_ask: 0.0261 r2 high_bid: 0.5262 r2 low_ask: 0.4655
Informer	rmse high_bid: 0.3671 rmse low_ask: 0.3398 mae high_bid: 0.2982 mae low_ask: 0.2746 r2 high_bid: 0.3791 r2 low_ask: 0.4683

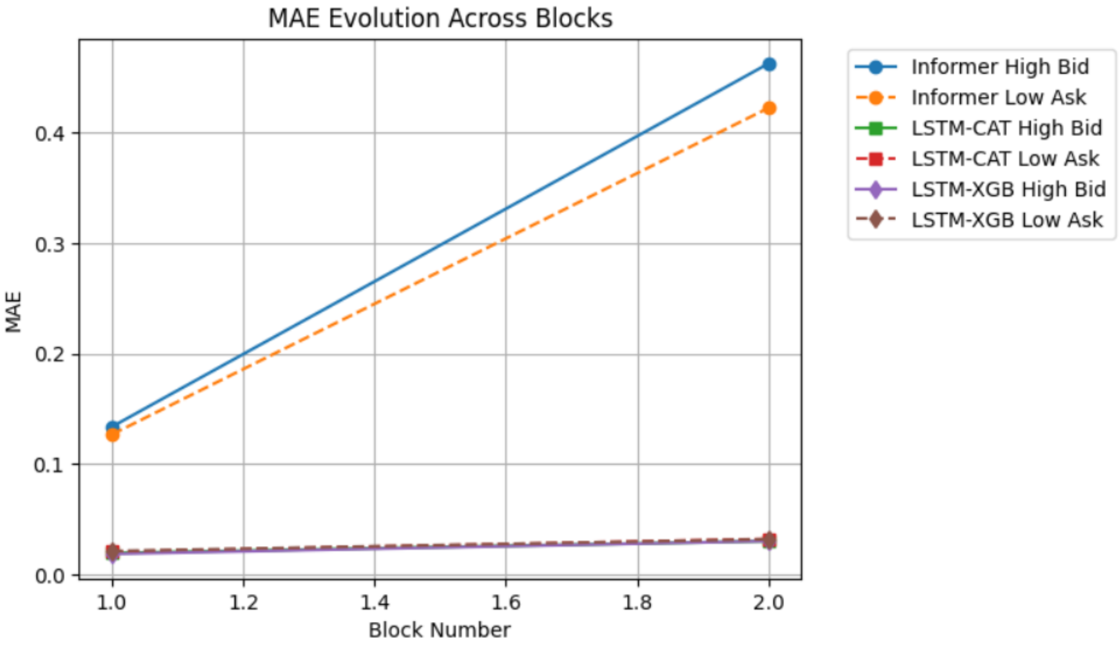


Figure 4.12: MAE Evolution Across All Blocks Using 1-hour Dataset

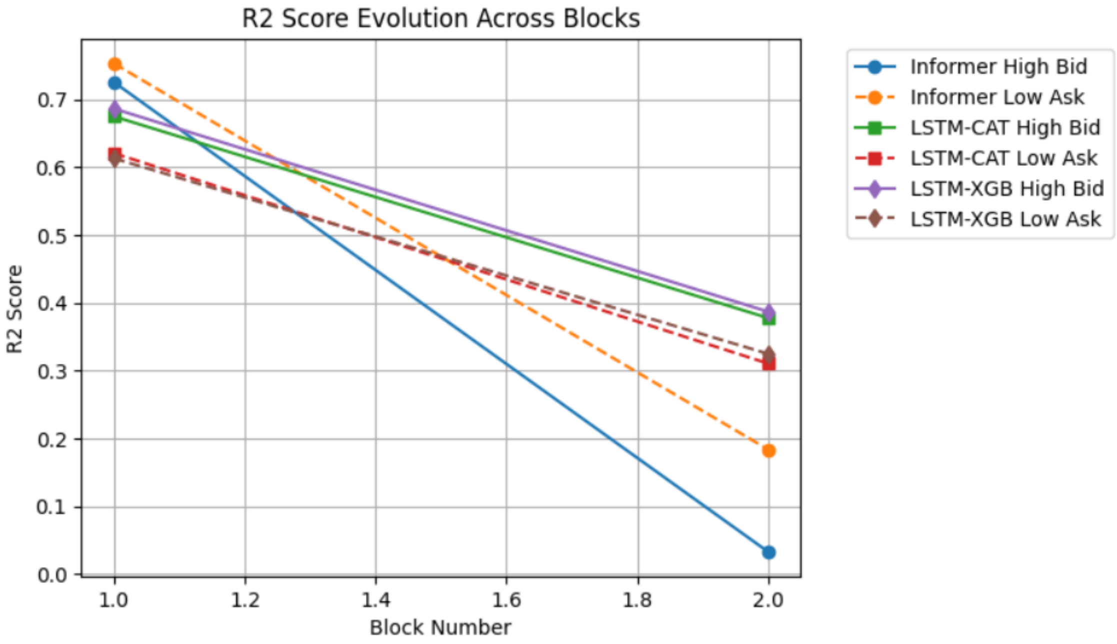


Figure 4.13: R^2 Evolution Across All Blocks Using 1-hour Dataset

4.5 Models Comparison

Across all datasets and horizons, the hybrid LSTM models consistently outperformed the Informer. On the 5-minute dataset, both LSTM-XGB and LSTM-CAT achieved low errors and strong R^2 scores, whereas Informer lagged with high RMSE and MAE values, despite reaching competitive R^2 scores.

Between the hybrids, LSTM-CAT excelled at shorter horizons, while LSTM-XGB was more stable at longer horizons. On the hourly dataset, all models performed poorly, with hybrids delivering only moderate results and Informer performing the worst.

Importantly, results for the two targets (High Bid and Low Ask) were consistent across hybrids, showing robust joint modelling of bid–ask dynamics. Informer showed more variability, further confirming the hybrids' superiority. Overall, LSTM-CAT provided the strongest accuracy at shorter horizons, while LSTM-XGB offered greater resilience as forecasting horizons increased.

Chapter 5

Discussion

5.1 Overview

This research set out to determine whether hybrid gradient-boosted LSTM pipelines (LSTM+XGBoost, LSTM+CATBoost) can outperform a Transformer-based architecture (Informer) for intra-day forecasting of XAU/USD, and to establish which candlestick resolution is most effective under a walk-forward validation framework. By systematically comparing models across three horizons of analysis (architectural choice, dataset granularity, and evaluation methodology), the study contributes new empirical insights to an area of financial forecasting that remains under-explored in the literature, particularly for commodities like gold at intra-day scales.

5.2 Models Performance

The results consistently showed that hybrid LSTM pipelines outperformed Informer across most horizons and datasets. Both LSTM+XGBoost and LSTM+CATBoost achieved lower error scores (RMSE, MAE) and higher R^2 values, confirming the strength of combining recurrent sequence modelling with boosting-based feature exploitation. This supports prior evidence from Fischer & Krauss (2018) and Oukhouya et al. (2024), who found boosted LSTMs superior in equity and stock market forecasting. Informer, although effective in benchmark long-horizon time series (Zhou et al. 2021, Nie et al. 2022), struggled to adapt to the volatility and microstructure noise inherent in intra-day gold trading. Thus, this study narrows the research gap by showing that in high-frequency commodity contexts, boosting-enhanced LSTMs remain more reliable than Transformers.

5.3 Dataset Granularity

The second contribution relates to dataset resolution. Results revealed that 15-minute candlesticks provided the most effective balance between information richness and noise reduction. On this dataset, both hybrid LSTMs achieved R^2 values exceeding 0.95 in some horizons, outperforming results on the 5-minute and 1-hour datasets. The 5-minute data suffered from high microstructure noise, which lowered generalisation despite hybrid models' robustness. Conversely, the 1-hour data smoothed much of the noise but failed to capture the fast-moving intra-day dynamics, which is crucial for short-term trading decisions. These findings contribute to the literature by demonstrating that intermediate resolutions, such as 15-minute candlesticks, offer optimal predictive performance for intra-day gold markets.

5.4 Walk-Forward Validation

Finally, the application of a walk-forward validation framework proved critical in ensuring robustness and preventing overestimation of model performance. Unlike random splits, walk-forward evaluation respects the temporal ordering of data and reflects how a model would perform in live trading conditions. The results showed that even high-performing models exhibited variability across blocks, highlighting the importance of testing over multiple periods rather than relying on a single train–test partition. In doing so, this study addresses a key methodological gap noted in Lim & Zohren (2021), offering a reliable framework that future financial forecasting studies can adopt to improve reliability. However, it should be noted that for the 1-hour dataset, which contained only about 10,000-11,000 rows, dividing the data into blocks may have reduced the training size available to each block; this relative shortage of training samples could have negatively affected model learning, and could explain the weaker performance observed compared to the larger datasets.

Chapter 6

Conclusion and Future Work

Having completed this project, I believe it successfully meets the initial aims and objectives established at the outset. The primary goal was to evaluate whether a hybrid gradient-boosted LSTM (LSTM+XGBoost or LSTM+CATBoost) could outperform a Transformer-based forecaster (Informer) for intra-day gold (XAU/USD) price prediction, across different horizons and candlestick resolutions, under a walk-forward validation. The research has shown that the hybrid LSTMs consistently provided more reliable performance, particularly on the 15-minute dataset, thereby addressing the identified research gap. In contrast, the Informer demonstrated mixed results, highlighting both its potential and its limitations in small, high-frequency financial datasets.

From the technical results and discussions, several key conclusions can be drawn. First, LSTM+CATBoost emerged as the strongest overall performer, delivering the lowest RMSE and highest R^2 score across most horizons. Second, the 15-minute candlestick data proved to be the most effective resolution, balancing sample size and noise. Third, the use of walk-forward validation confirmed the robustness of these findings by ensuring no temporal leakage, an essential step often overlooked in financial prediction studies. Collectively, these outcomes confirm that the artefact not only meets the original technical requirements but also provides a reproducible and rigorous framework for future intra-day forecasting research.

On a personal level, this project has been an important learning process. Technically, I developed advanced skills in time series modelling, hybrid architectures, and Transformer implementations, alongside training and evaluating ML models under a time series cross-validation. I also strengthened my ability to design and execute a reproducible research pipeline. On the other hand, this report was written entirely in LaTeX (via Overleaf), which allowed me to gain valuable experience in professional academic writing and typesetting. In addition, the project improved my soft skills, particularly in critical analysis and structured research communication.

Reflecting on the work, I recognise that more extensive hyperparameter optimisation and greater computational resources could have improved model performance, especially for the Informer and the 1-hour dataset, where the relatively small training size may have slowed down learning. If I were to repeat the project, I would incorporate external drivers, such as news sentiment and macroeconomic indicators, earlier in the process, as well as experiment with larger datasets, as the complexity of transformers needs to be saturated by large datasets to achieve acceptable results. Looking forward, there are several promising avenues for future work. The framework could be extended by testing alternative Transformer architectures (such as PatchTST or FEDformer), scaling experiments across other assets like oil, equities, or cryptocurrencies. Moreover, an ensemble model that combines the best-performing models across datasets could be developed to further enhance accuracy and robustness. Such advancements could strengthen the contribution of this project to both academia and industry.

Bibliography

Baur, D. G. & Lucey, B. M. (2010), 'Is gold a hedge or a safe haven? an analysis of stocks, bonds and gold', *Financial review* **45**(2), 217–229.

Bergmeir, C. & Benítez, J. M. (2012), 'On the use of cross-validation for time series predictor evaluation', *Information Sciences* **191**, 192–213.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014), 'Empirical evaluation of gated recurrent neural networks on sequence modeling', *arXiv preprint arXiv:1412.3555* .

Duan, C. & Ke, W. (2024), 'Advanced stock price prediction using lstm and informer models', *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023* **5**(1), 141–166.

Encean, A.-A. & Zinca, D. (2022), Cryptocurrency price prediction using lstm and gru networks, *in* '2022 International Symposium on Electronics and Telecommunications (ISETC)', IEEE, pp. 1–4.

Ferreira, I. H. & Medeiros, M. C. (2021), 'Modeling and forecasting intraday market returns: a machine learning approach', *arXiv preprint arXiv:2112.15108* .

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.

Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022), 'Why do tree-based models still outperform deep learning on typical tabular data?', *Advances in neural information processing systems* **35**, 507–520.

Hansen, P. R. & Lunde, A. (2006), 'Realized variance and market microstructure noise', *Journal of Business & Economic Statistics* **24**(2), 127–161.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Jabeur, S. B., Mefteh-Wali, S. & Viviani, J.-L. (2024), 'Forecasting gold price with the xgboost algorithm and shap interaction values', *Annals of Operations Research* **334**(1), 679–699.

Jurdi, D. J. (2020), 'Intraday jumps, liquidity, and us macroeconomic news: evidence from exchange traded funds', *Journal of Risk and Financial Management* **13**(6), 118.

Kim, J., Kim, H., Kim, H., Lee, D. & Yoon, S. (2025), 'A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges', *Artificial Intelligence Review* **58**(7), 1–95.

Liang, D., Xu, Y., Hu, Y. & Du, Q. (2023), 'Intraday return forecasts and high-frequency trading of stock index futures: A hybrid wavelet-deep learning approach', *Emerging Markets Finance and Trade* **59**(7), 2118–2128.

Lim, B., Arık, S. Ö., Loeff, N. & Pfister, T. (2021), 'Temporal fusion transformers for interpretable multi-horizon time series forecasting', *International journal of forecasting* **37**(4), 1748–1764.

Lim, B. & Zohren, S. (2021), 'Time-series forecasting with deep learning: a survey', *Philosophical Transactions of the Royal Society A* **379**(2194), 20200209.

Livieris, I. E., Pintelas, E., Stavroyiannis, S. & Pintelas, P. (2020), 'Ensemble deep learning models for forecasting cryptocurrency time-series', *Algorithms* **13**(5), 121.

Loras, R. (2024), 'The impact of transactions costs and slippage on algorithmic trading performance'.

Nelson, D. M., Pereira, A. C. & De Oliveira, R. A. (2017), Stock market's price movement prediction with lstm neural networks, *in* '2017 International joint conference on neural networks (IJCNN)', Ieee, pp. 1419–1426.

Nie, Y., Nguyen, N. H., Sinthong, P. & Kalagnanam, J. (2022), 'A time series is worth 64 words: Long-term forecasting with transformers', *arXiv preprint arXiv:2211.14730* .

Oukhouya, H., Kadiri, H., El Himdi, K. & Guerbaz, R. (2024), 'Forecasting international stock market trends: Xgboost, lstm, lstm-xgboost, and backtesting xgboost models', *Statistics, Optimization & Information Computing* **12**(1), 200–209.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. (2018), ‘Catboost: unbiased boosting with categorical features’, *Advances in neural information processing systems* **31**.

Sezer, O. B., Gudelek, M. U. & Ozbayoglu, A. M. (2020), ‘Financial time series forecasting with deep learning: A systematic literature review: 2005–2019’, *Applied soft computing* **90**, 106181.

Wahyuddin, E. P., Caraka, R. E., Kurniawan, R., Caesarendra, W., Gio, P. U. & Pardamean, B. (2025), ‘Improved lstm hyperparameters alongside sentiment walk-forward validation for time series prediction’, *Journal of Open Innovation: Technology, Market, and Complexity* **11**(1), 100458.

Wang, W. & Ruf, J. (2022), ‘A note on spurious model selection’, *Quantitative Finance* **22**(10), 1797–1800.

Wen, F., Tong, X. & Ren, X. (2022), ‘Gold or bitcoin, which is the safe haven during the covid-19 pandemic?’, *International Review of Financial Analysis* **81**, 102121.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J. & Sun, L. (2022), ‘Transformers in time series: A survey’, *arXiv preprint arXiv:2202.07125* .

Wu, R. (2024), Multivariate financial time series forecasting model based on transformer architecture, *in* ‘2024 3rd International Conference on Smart City Challenges & Outcomes for Urban Transformation (SCOUT)’, IEEE, pp. 155–161.

Yaziz, S., Azizan, N., Zakaria, R. & Ahmad, M. (2013), The performance of hybrid arima-garch modeling in forecasting gold price, *in* ‘20th international congress on modelling and simulation, adelaide’, pp. 1–6.

Zhang, C. (2020), ‘Volatility, noise, and market microstructure: Econometric analysis using high-frequency data’, *Economics) Jia Li, Adviser* .

Zhang, F., Fleyeh, H. & Bales, C. (2022), ‘A hybrid model based on bidirectional long short-term memory neural network and catboost for short-term electricity spot price forecasting’, *Journal of the Operational Research Society* **73**(2), 301–325.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. & Zhang, W. (2020), ‘Informer2020: The GitHub repository for “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting”’, https://github.com/zhouhaoyi/Informer2020. Accessed: 2025-06-19.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. & Zhang, W. (2021), Informer: Beyond efficient transformer for long sequence time-series forecasting, *in* 'Proceedings of the AAAI conference on artificial intelligence', Vol. 35, pp. 11106–11115.