UNIVERSITÀ
DI PAVIA

Department of Economics and Management

Master Program in Finance

TEMPORAL FUSION TRANSFORMERS IN CRYPTOCURRENCY MARKETS

Supervisor: Prof. Paolo Giudici

Student: Daniyal RanjbarYajlou

Academic Year 2025-2026

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Paolo Giudici, for his guidance and valuable feedback throughout the development of this thesis.

I would also like to thank Alessandro Piergallini for his support and helpful insights during the writing process of this thesis.

I also thank the Department of Economics and Management for providing a rigorous academic environment that supported this research.

Finally, I am grateful to my family for their continuous support and encouragement during my studies.

Table of Contents

List of Figures and Tables

List of Tables

List of Figures

Abstract

This thesis investigates the effectiveness of advanced deep learning architectures for multi-horizon cryptocurrency return forecasting and evaluates whether improvements in predictive accuracy translate into economically meaningful trading performance. The analysis focuses on three cryptocurrencies—Bitcoin (BTC), Ethereum (ETH), and Dogecoin (DOGE)—using daily data from January 2018 to December 2025.

Three forecasting models are compared within a unified experimental framework: a traditional statistical benchmark (SARIMAX), a recurrent neural network (LSTM), and an attention-based architecture, the Temporal Fusion Transformer (TFT). Forecast accuracy is evaluated across short (H=3), medium (H=14), and long (H=30) horizons using RMSE, MAE, R^2, and directional accuracy. Economic relevance is assessed through a systematic backtesting framework incorporating transaction costs and risk-adjusted performance measures.

Empirical results show that the TFT achieves the lowest RMSE in seven out of nine asset–horizon configurations, demonstrating a consistent statistical advantage, particularly for Ethereum and long-horizon Dogecoin forecasts. However, forecasting superiority does not uniformly translate into trading profitability. The strongest economic outcome is observed for Dogecoin on the long horizon, where the TFT strategy achieves a total return of 70.2% and an annualized Sharpe ratio of 6.92. In contrast, trading results for Bitcoin remain weak across models despite improvements in predictive accuracy.

Overall, the findings indicate that attention-based multi-horizon architectures provide measurable gains in statistical forecasting performance, but economic value depends critically on asset characteristics, volatility structure, and signal design. The study highlights the importance of jointly evaluating statistical and financial performance when assessing advanced forecasting models in cryptocurrency markets.

Riassunto

Questa tesi analizza l'efficacia di architetture avanzate di deep learning per la previsione multi-orizzonte dei rendimenti delle criptovalute e valuta se i miglioramenti nella precisione predittiva si traducano in risultati economicamente significativi nelle strategie di trading. L'analisi si concentra su tre criptovalute Bitcoin (BTC), Ethereum (ETH) e Dogecoin (DOGE) utilizzando dati giornalieri dal gennaio 2018 al dicembre 2025.

Tre modelli di previsione sono confrontati all'interno di un framework sperimentale unificato: un benchmark statistico tradizionale (SARIMAX), una rete neurale ricorrente (LSTM) e un'architettura basata su meccanismi di attenzione, il Temporal Fusion Transformer (TFT). L'accuratezza delle previsioni è valutata su orizzonti breve (H = 3), medio (H = 14) e lungo (H = 30) utilizzando RMSE, MAE, R^2 e accuratezza direzionale. La rilevanza economica è analizzata attraverso un framework sistematico di backtesting che incorpora costi di transazione e misure di performance aggiustate per il rischio.

I risultati empirici mostrano che il TFT ottiene il valore più basso di RMSE in sette delle nove combinazioni asset–orizzonte, evidenziando un vantaggio statistico consistente, in particolare per Ethereum e per le previsioni di Dogecoin nel lungo periodo. Tuttavia, la superiorità nella previsione non si traduce uniformemente in maggiore redditività nelle strategie di trading. Il risultato economico più rilevante si osserva per Dogecoin sull'orizzonte lungo, dove la strategia basata su TFT raggiunge un rendimento totale del 70,2% e uno Sharpe ratio annualizzato pari a 6,92. Al contrario, i risultati di trading per Bitcoin rimangono deboli per tutti i modelli, nonostante i miglioramenti nella precisione predittiva.

Nel complesso, i risultati indicano che architetture multi-orizzonte basate su meccanismi di attenzione offrono miglioramenti misurabili nella performance statistica delle previsioni, ma il valore economico dipende in modo critico dalle caratteristiche dell'asset, dalla struttura della

volatilità e dal design del segnale di trading. Lo studio evidenzia quindi l'importanza di valutare congiuntamente le prestazioni statistiche e finanziarie nell'analisi di modelli avanzati di previsione nei mercati delle criptovalute.

List of Abbreviations

ARIMA	Autoregressive Integrated Moving Average
ATR	Average True Range
BTC	Bitcoin
CAGR	Compound Annual Growth Rate
DA	Directional Accuracy
DOGE	Dogecoin
EMA	Exponential Moving Average
ETH	Ethereum
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GRN	Gated Residual Network
H	Forecast Horizon
LSTM	Long Short-Term Memory
MA	Moving Average
MACD	Moving Average Convergence Divergence
MAE	Mean Absolute Error
OHLCV	Open, High, Low, Close, Volume
R^2	Coefficient of Determination
RNN	Recurrent Neural Network
RMSE	Root Mean Squared Error
RSI	Relative Strength Index
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Variables
TFT	Temporal Fusion Transformer
VSN	Variable Selection Network

Chapter 1– Introduction

1.1 Background and Context

Forecasting financial asset prices remains a central challenge in quantitative finance due to the complex, dynamic, and often non-linear nature of financial markets. Price movements are influenced by multiple interacting factors and frequently exhibit volatility clustering and time-varying dependencies that are difficult to capture using simple linear models. Over time, forecasting methodologies have evolved from traditional econometric approaches toward more flexible machine learning and deep learning frameworks capable of modeling complex temporal structures.

In traditional financial markets, deep learning models have been increasingly applied to improve predictive performance by capturing non-linear patterns in historical price data. In particular, architectures designed for sequential and multi-horizon forecasting have attracted attention, as they allow predictions to be generated across multiple future time steps. Such capabilities are especially relevant in financial contexts where decisions depend on both short-term fluctuations and longer-term trends.

Cryptocurrency markets represent a distinct and comparatively recent asset class. Unlike traditional securities, cryptocurrencies trade continuously and are characterized by pronounced volatility, structural breaks, and sensitivity to technological developments and investor sentiment. These features create additional challenges for forecasting models originally developed in more stable financial environments.

While early studies on cryptocurrency price prediction relied primarily on classical econometric time-series models, more recent research increasingly applies machine learning and deep learning techniques. Nevertheless, systematic cross-asset evaluations of advanced multi-horizon architectures remain comparatively limited. In particular, relatively little empirical

evidence exists on how such models perform across cryptocurrencies with different economic roles and behavioral characteristics.

1.2 Research Problem

Despite methodological advances, forecasting cryptocurrency prices remains a difficult and unresolved problem. The extreme volatility and heterogeneous drivers of digital assets pose significant challenges for both traditional statistical models and more recent machine learning approaches. Models that perform adequately in conventional financial markets may exhibit instability or limited robustness when applied to cryptocurrency data.

Moreover, much of the existing literature emphasizes statistical prediction accuracy without explicitly assessing whether improvements in forecasting performance translate into economically meaningful trading outcomes. From a financial perspective, predictive accuracy alone is insufficient if it does not result in improved risk-adjusted performance when forecasts are used for decision-making.

Although advanced deep learning architectures have demonstrated strong performance in traditional financial markets, their application to cryptocurrency markets remains relatively limited in comparison, particularly for models specifically designed for multi-horizon forecasting. As a result, there is comparatively less systematic evidence regarding their ability to generate reliable and economically interpretable trading signals across cryptocurrencies with structurally different characteristics.

The central problem addressed in this thesis is therefore twofold. First, it is unclear whether an advanced multi-horizon forecasting architecture can achieve robust predictive performance in highly volatile cryptocurrency markets. Second, even if predictive accuracy improves, it

remains uncertain whether such improvements translate into economically meaningful trading performance across heterogeneous digital assets.

1.3 Research Questions

To address this problem, the thesis investigates the following primary research question:

Does an advanced multi-horizon deep learning model provide superior forecasting performance for cryptocurrencies, and does its effectiveness vary across assets with different economic and behavioral characteristics, specifically Bitcoin, Ethereum, and Dogecoin?

This primary question is operationalized through several secondary questions:

1. **Comparative Forecasting Accuracy:** How does forecasting accuracy, measured through metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), differ across Bitcoin, Ethereum, and Dogecoin, and how does the model compare with established benchmark approaches?
2. **Economic Value of Forecasts**: To what extent can the generated forecasts be translated into economically meaningful trading strategies, and how do these strategies perform relative to benchmark models in terms of risk-adjusted measures such as the Sharpe ratio, maximum drawdown, and total return?
3. **Cross-Asset Performance Differences**: Are there systematic differences in model performance across cryptocurrencies with distinct market roles and behavioral drivers?
4. **Stability of Model Performance**: How stable is forecasting and trading performance over the evaluation period?

1.4 Contribution and Scope

This thesis contributes to the literature on cryptocurrency forecasting by providing a structured empirical assessment of an advanced multi-horizon deep learning architecture within a comparative framework that integrates both statistical and financial evaluation criteria. Rather than focusing exclusively on predictive accuracy, the analysis jointly evaluates forecasting performance and trading outcomes, thereby linking methodological advances with economically interpretable results.

A further contribution lies in the cross-asset comparison across Bitcoin, Ethereum, and Dogecoin, which represent distinct categories within the cryptocurrency ecosystem. By explicitly accounting for asset heterogeneity, the study provides insight into how model effectiveness may vary across cryptocurrencies with different economic roles and behavioral dynamics.

The scope of the analysis is limited to these three cryptocurrencies and to daily historical market data obtained from publicly available sources. The evaluation framework focuses on time-series forecasting and rule-based trading strategies. Reported trading performance is computed net of a fixed proportional transaction cost, implemented as a cost of 0.1% per unit of turnover. The analysis does not incorporate bid–ask spreads, slippage, taxation, or regulatory constraints. These simplifying assumptions allow for controlled model comparison but do not attempt to fully replicate real-world trading conditions.

1.5 Structure of the Thesis

The thesis is organized as follows. Chapter 2 reviews the relevant literature on cryptocurrency markets and financial time-series forecasting, with particular attention to deep learning and attention-based architectures. Chapter 3 presents the data, models, and methodological

framework employed in the empirical analysis. Chapter 4 reports the forecasting and trading results and discusses the findings in relation to the research questions. Chapter 5 concludes by summarizing the main results, acknowledging limitations, and outlining directions for future research.

Chapter 2 – Literature Review and Theoretical Background

This chapter reviews the theoretical and empirical literature that forms the foundation of the present study. Cryptocurrency forecasting lies at the intersection of financial market research and advanced time-series modeling, requiring insights from both domains.

The chapter first outlines key structural features of cryptocurrency markets that are relevant for price modeling, including their trading mechanisms and cross-asset heterogeneity. It then surveys major forecasting approaches used in financial time-series analysis, with attention to their underlying assumptions and limitations in complex market environments. Finally, the discussion turns to recent developments in attention-based and multi-horizon architectures, highlighting their conceptual relevance for structured time-series forecasting.

By integrating these strands of literature, the chapter establishes the theoretical context for the empirical framework developed in the subsequent chapters.

2.1 The Cryptocurrency Market: Characteristics and Challenges

Cryptocurrency markets differ substantially from traditional financial markets in terms of structure, trading mechanisms, and price formation processes. Since the introduction of Bitcoin, cryptocurrencies have evolved into a diverse ecosystem of digital assets with varying economic roles, technological foundations, and investor bases. These differences have important implications for price dynamics and, consequently, for the design and performance of forecasting models.

Unlike traditional assets such as equities or bonds, cryptocurrencies operate in a largely decentralized environment and trade continuously without market closures. This structural distinction affects liquidity patterns, volatility behavior, and the transmission of information

across markets. As a result, modeling approaches developed for conventional financial assets may not be directly applicable to cryptocurrencies without appropriate adaptation.

Market Structure and Trading Characteristics

One of the defining features of cryptocurrency markets is their continuous 24-hour trading structure. Transactions occur across global exchanges without centralized coordination, allowing prices to always adjust in response to new information. While this feature enhances market accessibility, it also contributes to irregular trading patterns and abrupt price movements.

In addition, cryptocurrency markets are characterized by varying degrees of liquidity across assets and exchanges. While major cryptocurrencies such as Bitcoin and Ethereum tend to exhibit relatively high liquidity, smaller or less established assets often experience thin trading volumes and higher price sensitivity. The absence of centralized market makers and the fragmentation across trading platforms further amplify these effects.

These structural characteristics complicate the identification of stable statistical relationships in cryptocurrency price series. Periods of relative calm may be followed by sudden and pronounced price swings, often triggered by external events such as regulatory announcements, security breaches, or technological developments. Consequently, forecasting models must be able to adapt to rapidly changing market conditions (Baur et al., 2018; Corbet et al., 2018).

Volatility and Non-Stationarity

High volatility is one of the most widely documented features of cryptocurrency markets. Empirical studies consistently show that cryptocurrency returns exhibit greater variability than those of traditional financial assets, along with pronounced volatility clustering. Such behavior

violates key assumptions underlying many classical time-series models, including constant variance and stable distributions.

In addition to high volatility, cryptocurrency price series often display strong non-stationarity. Structural breaks, regime shifts, and changing market participation are common, particularly during periods of rapid market expansion or contraction. These dynamics reduce the effectiveness of models that rely on fixed parameters or long-term historical averages.

The combination of volatility and non-stationarity poses significant challenges for forecasting. Models that fail to account for time-varying relationships may produce unstable predictions or perform well only during specific market phases. This has motivated the exploration of more flexible modeling frameworks capable of capturing evolving temporal dependencies (Chu et al., 2017; Corbet et al., 2019; Katsiampa, 2017).

Heterogeneity Across Cryptocurrency Assets

Although cryptocurrencies are often discussed as a single asset class, substantial heterogeneity exists across individual assets. Major cryptocurrencies differ in terms of market capitalization, underlying use cases, and the relative importance of speculative versus fundamental drivers. These differences can lead to distinct price behaviors and forecasting challenges.

Bitcoin is commonly regarded as a store-of-value asset and is often compared to digital gold. Its price dynamics are influenced by macro-level factors, investor adoption, and perceptions of scarcity. Ethereum, by contrast, functions as a platform for decentralized applications and smart contracts, with its value linked more closely to network activity and technological development. Other cryptocurrencies, such as Dogecoin, are frequently driven by speculative behavior and social sentiment, leading to more abrupt and less predictable price movements.

This heterogeneity suggests that a single forecasting approach may not perform uniformly across all cryptocurrencies. Models that capture longer-term trends may be more effective for

some assets, while others may require greater sensitivity to short-term fluctuations and sentiment-driven dynamics. Recognizing these differences is essential for evaluating model performance and interpreting empirical results (Brandvold et al., 2015; Grobys et al., 2021).

Implications for Price Forecasting

The structural features, volatility patterns, and heterogeneity of cryptocurrency markets collectively contribute to the complexity of price forecasting in this domain. Traditional forecasting models, which often assume stable statistical relationships and well-defined market regimes, may struggle to adapt to the rapidly changing dynamics observed in cryptocurrency data.

As a result, there is growing interest in forecasting approaches that can accommodate non-linearity, time-varying dependencies, and multi-scale temporal patterns. Understanding the unique challenges posed by cryptocurrency markets provides a necessary foundation for assessing the suitability of advanced forecasting models, which are discussed in the subsequent sections of this chapter.

2.2 Financial Time-Series Forecasting Models

The problem of forecasting financial time series has been extensively studied in the literature, leading to the development of a wide range of modeling approaches. These methods differ in their underlying assumptions, flexibility, and capacity to model non-linear dependencies and time-varying relationships in financial data. This section reviews the main classes of forecasting models relevant to financial and cryptocurrency price prediction, highlighting their respective strengths and limitations.

2.2.1 Traditional Statistical Models

Traditional statistical models have long formed the foundation of financial time-series analysis. Among the most widely used approaches are autoregressive and moving average models, including the Autoregressive Integrated Moving Average (ARIMA) framework. These models describe the current value of a time series as a linear function of its past values and past forecast errors, under the assumption that the underlying data-generating process is stationary after appropriate transformations.

ARIMA-type models have been successfully applied to a variety of economic and financial time series, particularly in relatively stable market environments. Their appeal lies in their interpretability, well-established theoretical properties, and relatively low computational complexity. However, their performance often deteriorates in the presence of strong non-linearity, structural breaks, and time-varying volatility.

In the context of cryptocurrency markets, these limitations become more evident. Digital asset prices frequently exhibit structural breaks, parameter instability, and pronounced fluctuations over time, which challenge the linearity and stability assumptions underlying traditional econometric models. While ARIMA and related approaches are commonly used as benchmark specifications, their forecasting performance may deteriorate when relationships in the data evolve over time (Box et al., 2015).

2.2.2 Machine Learning Approaches

To overcome the limitations of linear statistical models, researchers have increasingly turned to machine learning techniques. These approaches aim to model complex and potentially non-linear relationships between inputs and outputs without relying on strict parametric

assumptions. Commonly used machine learning models in financial forecasting include support vector machines, decision trees, and ensemble methods such as random forests.

Machine learning models have demonstrated improved predictive performance in various financial applications, particularly when relationships between variables are non-linear or when interactions among predictors are important. Their flexibility allows them to incorporate a wide range of explanatory variables beyond historical prices, such as technical indicators or sentiment measures.

Nevertheless, many traditional machine learning models treat observations as independent or rely on limited temporal structures. While lagged variables can be included as features, these models are often not specifically designed to capture long-range temporal dependencies inherent in sequential data. This limitation reduces their effectiveness for tasks where the temporal ordering of observations plays a central role (Hastie et al., 2009).

2.2.3 Deep Learning and Sequential Models

Deep learning models represent a further evolution in financial time-series forecasting, particularly for sequential data. Recurrent neural networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, are explicitly designed to model temporal dependencies by maintaining internal memory states that evolve over time.

LSTM networks address key limitations of standard RNNs, particularly the vanishing and exploding gradient problems, thereby enabling the learning of longer-term dependencies. As a result, they have been widely adopted in financial forecasting applications, including stock price prediction and cryptocurrency return modeling. Empirical evidence frequently suggests that LSTMs outperform traditional statistical and machine learning models in capturing

nonlinear and time-dependent patterns in financial data (Fischer & Krauss, 2018; Hochreiter & Schmidhuber, 1997; Nelson et al., 2017).

Despite these advantages, recurrent architectures exhibit several limitations. Training can be computationally intensive, and performance may be sensitive to hyperparameter specification. Moreover, standard LSTM implementations are typically designed for single-horizon prediction and may face challenges when required to model multiple future steps simultaneously. In addition, their internal representations often lack transparency, limiting interpretability in financial applications where understanding the temporal structure of predictions can be important.

These limitations have motivated the development of attention-based architectures that aim to capture long-range dependencies more flexibly while supporting multi-horizon forecasting within a unified framework. Such models are discussed in the following section.

2.3 Attention Mechanisms and Transformer Models

As financial time series data become increasingly complex and high-dimensional, limitations of recurrent neural network architectures have motivated the development of alternative modeling approaches. While recurrent models such as LSTMs are designed to capture temporal dependencies through sequential processing, they may struggle to efficiently model long-range relationships, particularly when relevant information is distributed across distant time steps.

Attention mechanisms were introduced to address these limitations by allowing models to selectively focus on the most relevant parts of an input sequence when generating predictions. Rather than relying solely on a fixed internal memory, attention-based models assign weights to different time steps, enabling a more flexible representation of temporal dependencies. This

mechanism has proven particularly effective in tasks where long-term context plays a critical role.

The introduction of the Transformer architecture marked a significant shift in sequence modeling. Transformers replace recurrent structures entirely with self-attention mechanisms, allowing all elements of a sequence to be processed in parallel. By computing relationships between all pairs of time steps, self-attention enables the model to capture both short-term patterns and long-range dependencies more effectively than traditional recurrent networks.

In financial forecasting applications, attention-based models offer several potential advantages. First, they provide a more flexible way to model non-linear and non-local dependencies in time-series data, which are common in financial markets. Second, attention mechanisms can enhance interpretability by highlighting which past observations contribute most strongly to a given prediction. These features are particularly relevant in environments characterized by high volatility and regime changes.

Despite their advantages, generic Transformer models are not without limitations when applied directly to financial time series. They are typically designed for generic sequence modeling tasks and may require substantial adaptation to handle structured temporal data, multiple covariates, and forecasting across multiple horizons. These considerations have motivated the development of specialized attention-based architectures tailored specifically to time-series forecasting problems.

One such architecture is the Temporal Fusion Transformer, which integrates attention mechanisms with additional components designed to handle multi-horizon forecasting and heterogeneous inputs. The theoretical foundations and key design principles of this model are discussed in the following section (Lim & Zohren, 2021; Vaswani et al., 2017).

2.4 Temporal Fusion Transformer

The Temporal Fusion Transformer (TFT), introduced by Lim et al. (2021), is a structured attention-based architecture designed for interpretable multi-horizon time-series forecasting. Unlike generic Transformer models, the TFT is specifically constructed to handle heterogeneous inputs and produce forecasts across multiple future time steps within a single unified framework.

Multi-Horizon Forecasting Formulation

The Temporal Fusion Transformer models multi-horizon forecasting using a structured input representation that explicitly distinguishes among historical targets, time-varying covariates, and static covariates. Following Lim et al. (2021).

The forecasting structure can be expressed as:

$$\hat{y}_i(q,t,\tau) = f_q(\tau, \hat{y}_{i,t-k:t}, z_{i,t-k:t}, x_{i,t:t+\tau}, s_i) \qquad (2.1)$$

where:

- $\hat{y}_i(q,t,\tau)$ is the predicted value for time series iat time t, for forecast horizon τ and quantile q
- f_q denotes the model used to compute the quantile forecast
- τ is the forecast horizon
- $\hat{y}_{i,t-k:t}$ represents historical target values from time $t-k$to t
- $z_{i,t-k:t}$ denotes time-varying unknown inputs (also referred to as observed historical covariates)
- $x_{i,t:t+\tau}$ denotes time-varying known inputs available up to the forecast horizon (e.g., calendar variables)

- s_i represents static covariates associated with series i

This formulation highlights three key characteristics of the TFT architecture:

1. **Explicit separation of input categories**: The model distinguishes among historical observed inputs, known future inputs, and static covariates.
2. **Static conditioning**: Static features $s_i$condition the network throughout the architecture, influencing variable selection networks, gating layers, and attention mechanisms.
3. **Direct multi-horizon prediction**: The model produces forecasts for horizon τdirectly rather than recursively, thereby reducing error propagation.

   Although the present study focuses on point forecasts rather than quantile estimation, this formulation provides a general representation of the TFT structure and its handling of heterogeneous inputs.

Encoder–Decoder Structure

The Temporal Fusion Transformer integrates recurrent processing, variable selection, and gating mechanisms within an encoder–decoder framework specifically designed for structured time-series forecasting.

The **LSTM encoder** processes historical inputs—including past target values and time-varying observed covariates—to capture sequential dependencies and local temporal dynamics. Prior to recurrent processing, variable selection networks identify and weight the most relevant input features, allowing the model to adaptively emphasize informative signals across time.

The **LSTM decoder** operates on time-varying known future inputs and the encoded historical context. This stage generates future-conditioned temporal representations that are subsequently refined through attention mechanisms. By incorporating recurrent processing on both sides of

the architecture, the TFT differs from standard sequence-to-sequence models and ensures coherent multi-horizon forecasting within a unified structure.

To enhance stability and learning efficiency, the architecture employs **gated residual networks and Add & Norm layers**, which regulate information flow and mitigate overfitting. These components allow the model to dynamically filter signals while preserving gradient stability during training.

Together, variable selection, dual LSTM blocks, gating layers, and attention mechanisms form a structured hybrid architecture capable of modeling complex temporal relationships while maintaining interpretability.

Variable Selection Networks

A key innovation of the TFT is the use of Variable Selection Networks (VSNs), which dynamically weight input variables according to their relevance.

For a given input vector x_t.Variable selection can be conceptually represented as:

$$\tilde{x}_t = \sum_{i=1}^{m} \alpha_{i,t}\, x_{i,t} \qquad (2.2)$$

where:

- $x_{i,t}$ is the i-th input feature
- $\alpha_{i,t}$ is a learned importance weight
- m is the number of input variables.

These weights are generated through gating mechanisms, allowing the model to adaptively emphasize relevant features across time.

The TFT distinguishes among:

- Static covariates

- Time-varying known inputs
- Time-varying observed inputs

  Each category is processed separately before integration.

Gating Mechanisms

The architecture incorporates gated residual networks (GRNs), which regulate information flow and stabilize training. A simplified representation is:

$$GRN(x) = LayerNorm(x + Gate(\phi(x))) \qquad (2.3)$$

where:

- $\phi(x)$ represents a nonlinear transformation,
- $\mathrm{Gate}(\cdot)$ controls the degree of information passed forward.

  These gating layers improve robustness and interpretability.

Self-Attention Mechanism

To capture long-range dependencies, the TFT applies multi-head self-attention in the decoder. The standard attention formulation is:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2.4)$$

where:

- Q= queries
- K= keys
- V= values
- d_k= dimension scaling factor

This mechanism allows the model to weigh historical time steps when generating forecasts for each horizon.

Interpretability and Practical Relevance

Through attention weights and variable selection scores, the TFT provides insight into which variables and time steps contribute most strongly to predictions. This property is particularly relevant in financial applications, where model transparency is important.

Overall, the TFT integrates recurrent encoding, variable selection, gating layers, and attention blocks within an encoder–decoder framework to support structured multi-horizon forecasting. These architectural characteristics make it well suited for financial time-series applications involving heterogeneous inputs and evolving temporal relationships (Lim et al., 2021).

2.5 Synthesis and Identification of the Research Gap

The literature reviewed in this chapter highlights both the progress made in financial time-series forecasting and the challenges that remain unresolved, particularly in the context of cryptocurrency markets. Existing research clearly documents that cryptocurrencies exhibit structural features, volatility patterns, and behavioral drivers that distinguish them from traditional financial assets. These characteristics complicate forecasting tasks and limit the effectiveness of models developed under assumptions of stability and linearity.

The evolution of forecasting methodologies reflects ongoing attempts to address these challenges. Traditional statistical models offer interpretability and theoretical grounding but struggle to adapt to non-stationary and highly volatile environments. Machine learning approaches provide greater flexibility but often lack explicit mechanisms for modeling long-range temporal dependencies. Deep learning architectures, particularly recurrent neural

networks such as LSTMs, represent a significant advancement by explicitly accounting for sequential structure, yet they still face limitations in handling multi-horizon forecasting and complex interactions among heterogeneous inputs.

Recent developments in attention-based models and Transformer architectures offer a promising alternative by enabling more flexible representations of temporal dependencies and improving interpretability through attention mechanisms. The Temporal Fusion Transformer, in particular, has been proposed as a specialized solution for structured time-series forecasting, combining multi-horizon prediction capabilities with mechanisms for handling diverse input types and providing insight into model behavior.

Despite these advances, the application of such models to cryptocurrency markets remains limited in the existing literature. Many studies focus on a single cryptocurrency or evaluate models solely on statistical accuracy, without considering their economic implications when used in trading strategies. Furthermore, there is a lack of comparative analyses that examine how advanced forecasting models perform across cryptocurrencies with different economic roles and behavioral characteristics.

This thesis addresses these gaps by applying a multi-horizon attention-based forecasting model to multiple cryptocurrency assets and evaluating its performance within a unified and reproducible framework. By jointly assessing forecasting accuracy and trading performance across Bitcoin, Ethereum, and Dogecoin, the study seeks to contribute empirical evidence on the suitability and practical relevance of advanced deep learning models in heterogeneous cryptocurrency markets. The next chapter builds on this theoretical foundation by describing the methodological framework and models employed in the empirical analysis.

Chapter 3 – Methodology and Model Framework

This chapter describes the methodological framework adopted to address the research questions outlined in Chapter 1. The objective is to provide a transparent and reproducible account of the data, models, and evaluation procedures used in the empirical analysis, without anticipating or discussing the results.

The chapter begins by describing the data sources and preprocessing steps applied to construct the final dataset used across all experiments. It then presents the forecasting models considered in the study, including a traditional statistical benchmark, a recurrent neural network architecture, and an advanced attention-based deep learning model. These models are selected to represent different levels of model complexity and to allow for a structured comparison across methodologies.

Subsequently, the chapter outlines the experimental design, including the training and testing procedure, the multi-horizon forecasting setup, and the evaluation metrics used to assess forecasting accuracy. Finally, the chapter describes the trading strategy framework employed to translate model forecasts into buy, hold, or sell signals and to evaluate their economic performance through back-testing.

By clearly separating methodological choices from empirical findings, this chapter establishes the foundation for the results and discussion presented in the following chapter.

3.1 Data Description and Preprocessing

The empirical analysis is based on historical daily market data for three cryptocurrencies: Bitcoin (BTC), Ethereum (ETH), and Dogecoin (DOGE). Data was retrieved from Yahoo Finance (tickers BTC-USD, ETH-USD, DOGE-USD) and consolidated into a single dataset

that is used consistently across all forecasting and trading experiments. The sample covers the period from 1 January 2018 to 28 December 2025, yielding 2,919 daily observations per asset.

Table 1. Cryptocurrency Dataset Overview

Asset	Ticker	Frequency	Sample Period	Observations
Bitcoin	BTC-USD	Daily	2018-01-01 – 2025-12-28	2,919
Ethereum	ETH-USD	Daily	2018-01-01 – 2025-12-28	2,919
Dogecoin	DOGE-USD	Daily	2018-01-01 – 2025-12-28	2,919

Source: Author's compilation based on Yahoo Finance data.

The forecasting target is the daily log return. Let P_t denote the closing price on day t.

The target variable is defined as:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \qquad (3.1)$$

This corresponds to the dataset column log_ret_1.

where P_t is the closing price at time t. Models therefore predict future log returns directly. Price levels are not forecasted directly; instead, predicted prices are reconstructed from predicted returns when needed for visualization or trading evaluation.

Although the models forecast returns directly, implied prices are reconstructed only for visualization and for translating forecasts into trading outcomes. Given a predicted return $\hat{r}_t$, the implied one-step-ahead price is:

$$\hat{P}_t = P_{t-1} \cdot \exp(\hat{r}_t) \qquad (3.2)$$

The dataset includes standard OHLCV variables together with a set of engineered technical indicators commonly used in empirical financial forecasting. In the Temporal Fusion Transformer (TFT) implementation, inputs are structured according to the architecture's distinction between **static covariates**, **time-varying known inputs**, and **time-varying observed inputs**. Static covariates are limited to the asset identifier (symbol).

A per-asset time index is constructed as the number of calendar days elapsed since the first observation of each asset:

$$\text{time\_idx}_t = \text{date}_t - \min(\text{date}) \qquad (3.3)$$

measured in days and computed separately for each cryptocurrency.

Missing values in continuous predictors are handled using forward filling, followed by zero-imputation for any remaining gaps, consistent with the empirical implementation.

Table 2. TFT inputs: static and time-varying known variables

Variable	Description
symbol	Asset identifier (BTC, ETH, DOGE); used as a static categorical covariate
time_idx	Per-asset time index (days since first observation)
dow	Day of week (0 = Monday, …, 6 = Sunday)
month	Calendar month (1–12)
sin_doy	Seasonal encoding: $\sin(2\pi \cdot \text{doy}/365.25)$
cos_doy	Seasonal encoding: $\cos(2\pi \cdot \text{doy}/365.25)$

Source: Author's calculations.

Table 3. TFT inputs: time-varying observed variables

Variable	Description
log_close	Log closing price: ln (close).
log_ret_1	1-day log return (forecast target)
log_ret_7	7-day log return
ma_7 / ma_21 / ma_50	7/21/50-day moving averages of close
vol_7 / vol_21 / vol_60	Rolling standard deviation of log returns (7/21/60-day windows)
zscore_21	Z-score of close using 21-day rolling statistics
rsi_14	14-day Relative Strength Index
macd	Moving Average Convergence Divergence indicator
macd_signal	Signal line of MACD
hl_range	High–low range scaled by close: $(high - low)/close$
oc_range	Open–close range scaled by close: $(open - close)/close$
atr_14	14-day Average True Range
trend_strength_30	30-day trend-strength proxy
fear_greed	Market sentiment index
mcap_usd, tvol_usd, TxCnt, AdrActCnt, FeeTotNtv	External market/on-chain variables and related transforms.

Source: Author's calculations.

To document the distributional properties of the target variable, Table 3.2 reports descriptive statistics for daily log returns.

Table 4. Descriptive statistics of daily log returns (log_ret_1)

Asset	Mean	Std. Dev.	Min	Max	Median
BTC	0.000638	0.034261	-0.464730	0.171821	0.000741
ETH	0.000459	0.044992	-0.550732	0.230695	0.000588
DOGE	0.000902	0.066913	-0.515118	1.516328	-0.000952

Source: Author's calculations.

Dogecoin exhibits the highest volatility and extreme positive returns, while Bitcoin shows comparatively lower dispersion, consistent with its larger market capitalization and maturity.

All models use an identical chronological split to ensure consistent out-of-sample evaluation and to prevent look-ahead bias. The split (stored in the dataset column split) is defined as follows:

- **Training:** 2018-01-01 to 2025-01-02
- **Validation:** 2025-01-03 to 2025-07-01
- **Test:** 2025-07-02 to 2025-12-28

Each partition contains the same number of observations across assets (2,559 training; 180 validations; 180 test observations). This uniform structure guarantees comparability across models and cryptocurrencies.

3.2 Experimental Design and Forecasting Setup

The experimental design is structured to ensure a consistent comparison across forecasting models and cryptocurrency assets. All models are evaluated using identical data partitions, forecasting horizons, and evaluation metrics. This framework ensures that performance differences can be attributed to model characteristics rather than variations in data handling or experimental setup.

The forecasting target is the daily log return $r_t$as defined in Eq. (3.1). Models predict future log returns directly. Price levels are reconstructed via Eq. (3.2) only when needed for visualization or trading evaluation.

The forecasting problem is formulated as a **multi-horizon supervised learning task**. For each asset, the model generates forecasts for multiple future time steps simultaneously. Three prediction horizons are considered:

- **Short horizon:** $\mathrm{H} = 3$days
- **Medium horizon:** $\mathrm{H} = 14$days
- **Long horizon:** $\mathrm{H} = 30$days

For each horizon, the input window (encoder length) is adjusted to reflect the temporal context used for prediction:

- Short horizon: encoder length = 90 days
- Medium horizon: encoder length = 180 days
- Long horizon: encoder length = 365 days

This setup allows the models to incorporate different amounts of historical information depending on the forecasting horizon.

Model training follows a chronological split of the dataset into training, validation, and test subsets. Hyperparameter tuning is conducted using the validation set. Final performance is evaluated exclusively on the hold-out test set, which remains unseen during model training and selection.

To ensure realistic evaluation, forecasts are generated using only information available up to the prediction date. The temporal ordering of observations is strictly preserved, and no future information is used during training or inference. This procedure prevents look-ahead bias and information leakage.

By standardizing the forecasting target, horizon structure, encoder lengths, and data partitions across models and assets, the experimental design provides a controlled framework for assessing predictive accuracy and economic performance.

3.3 SARIMAX Model

As a statistical benchmark, this thesis employs the Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) model, which extends the ARIMA framework by allowing the inclusion of external regressors (Box et al., 2015).

The SARIMAX model is widely used in empirical time-series analysis due to its transparent structure and relatively strong baseline performance in linear forecasting settings.

In this study, the SARIMAX model is applied to the daily log return series r_t, which serves as the common forecasting target across all models. The model is estimated separately for each cryptocurrency (BTC, ETH, and DOGE). The general specification is given by:

$$r_t = c + \sum_{i=1}^{p} \phi_i \, r_{t-i} + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j} + \boldsymbol{\beta}^{\top} \mathbf{x}_t + \varepsilon_t \qquad (3.5)$$

where $\mathrm{x}_t$denotes the vector of exogenous regressors and $\varepsilon_t$is a white-noise disturbance term.

No seasonal component is included in the benchmark specification. Accordingly, the seasonal order is fixed to

$$(P, D, Q, s) = (0,0,0,0) \qquad (3.6)$$

The non-seasonal differencing order is also fixed at $d = 0$, since the target variable is already defined as a return series rather than a price level. This avoids redundant differencing and ensures consistency with the underlying stationarity properties of log returns.

The exogenous regressor set is restricted to calendar-based variables, consistent across assets:

$$x_t = \{\text{dow}_t, \text{month}_t, \sin(\text{doy}_t), \cos(\text{doy}_t)\} \qquad (3.7)$$

These regressors are standardized using parameters estimated on the training split and then applied consistently to validation and test samples.

For each cryptocurrency, the ARMA order $(p, d, q)$is selected via a small validation-based search over the following candidate set:

$$(p, d, q) \in \{(1,0,1), (2,0,1), (1,0,2), (2,0,2)\} \qquad (3.8)$$

Selection is performed using the **validation RMSE** computed from one-step-ahead forecasts generated over the full validation period. The model with the lowest validation RMSE is retained and then re-estimated on the combined training and validation sample before producing test forecasts.

The selected orders obtained in the empirical implementation are:

- **BTC:** $(2, 0, 1)$
- **ETH:** $(2, 0, 2)$
- **DOGE:** $(2, 0, 2)$

(These choices arise from the validation RMSE criterion described above.)

To maintain comparability with the multi-horizon framework used throughout the thesis, SARIMAX produces forecasts over horizons $H \in \{3,14,30\}$. For each test date t, the fitted SARIMAX model generates an H-step-ahead forecast path:

$$(\hat{r}_{t+1}, \hat{r}_{t+2}, \dots, \hat{r}_{t+H})$$

Forecasts are generated in a walk-forward manner. After each step, the model state is updated by appending the newly observed return without refitting parameters, ensuring that only information available up to time t is used when predicting future observations. This approach preserves temporal ordering and avoids look-ahead bias while allowing sequential out-of-sample forecasting.

Overall, SARIMAX serves as a parsimonious linear benchmark against which the performance gains of more flexible machine learning and deep learning models can be assessed.

3.4 Long Short-Term Memory (LSTM) Model

As a deep learning benchmark, this thesis employs a Long Short-Term Memory (LSTM) network, a recurrent neural network architecture designed to model sequential data and retain information over time through gated memory cells (Hochreiter and Schmidhuber, 1997). The inclusion of the LSTM serves to provide a nonlinear sequence-based benchmark that is more flexible than traditional statistical models, yet structurally simpler than attention-based multi-horizon architectures such as the Temporal Fusion Transformer (TFT).

Consistent with the overall experimental framework described in **Section 3.2**, the LSTM is trained to forecast daily log-returns rather than price levels. The forecasting target is the daily log return, defined in **Eq. (3.1)**.

Multi-horizon forecasting is implemented by defining a vector-valued target:

$$y_t = (r_{t+1}, r_{t+2}, \dots, r_{t+H}) \qquad (3.10)$$

where r_{t+h} represents the log-return at forecast horizon h, and $H \in \{3,14,30\}$corresponds to the short-, medium-, and long-horizon buckets.

These targets are precomputed and stored in the shared dataset as:

target_h1, ... ,target_hH.

This ensures full comparability with SARIMAX and TFT within the unified multi-horizon evaluation framework.

The LSTM model is estimated separately for each cryptocurrency (BTC, ETH, DOGE). Inputs consist of engineered predictors derived from daily OHLCV data, constructed in a strictly causal manner (using only information available up to time t).

The feature vector follows the naming convention of the shared Excel dataset and includes:

All continuous predictors are converted to numeric format. Missing values are handled using forward filling followed by zero-imputation where necessary, consistent with the preprocessing pipeline described in Section 3.2.

Table 5. Engineered technical indicators used as time-varying unknown inputs

Feature group	Variable	Description
Moving averages	ma_7/ ma_21/ ma_50	7/21/50-day moving average of closing price
Volatility measures	vol_7/ vol_21/ vol_60	7/21/60-day rolling standard deviation of daily log returns
Range & trend proxies	hl_range	Daily high–low range scaled by close: $(high_t - low_t)/close_t$
	trend_strength_30	30-day trend-strength indicator (rolling trend proxy derived from historical prices)

Source: Author's calculations.

At each decision time t, the model input is a historical feature matrix:

$$X_t \in \mathbb{R}^{T \times F} \quad (3.11)$$

where Fdenotes the number of input features and Trepresents the lookback window (encoder length).

To maintain methodological consistency across models, the LSTM encoder length matches the TFT configuration:

The TFT is evaluated using the same multi-horizon configuration defined earlier in the chapter (short, medium, and long horizons) and the same horizon bucket definitions adopted for the SARIMAX benchmark. For each bucket, the model produces a vector of Hfuture log-return forecasts, conditional on a fixed-length historical encoder window, ensuring that performance differences across models reflect architectural choices rather than differences in horizon design.

All preprocessing is performed using only the training split to prevent information leakage. The chronological partition (train/validation/test) is identical across models and is read directly from the shared dataset column split.

Input features are scaled using a **RobustScaler**, which is less sensitive to outliers and appropriate for financial time series. Targets are scaled using a **StandardScaler**, fitted on the flattened training target matrix and applied consistently to validation and test sets.

The LSTM is implemented as a stacked bidirectional recurrent network followed by a feed-forward prediction head.

Architecture configuration:

- LSTM:
    - 2 layers
    - Hidden size: 128
    - Bidirectional
    - Dropout: 0.2 (between layers)
- Normalization:
    - LayerNorm applied to sequence hidden states
- Prediction head:
    - Linear → ReLU → Dropout → Linear
    - Output dimension: H

The final hidden representation (last time step) is used to generate the multi-horizon forecast vector:

$$\hat{y}_t \in \mathbb{R}^H$$

Model training is performed using the AdamW optimizer with a learning rate of $3 \times 10^{-4}$and weight decay of 10^{-5}, minimizing the Huber loss. A batch size of 128 is used, with a maximum of 12 training epochs and early stopping applied based on validation RMSE (patience of four epochs). Gradient clipping with a maximum norm of 0.5 is employed to enhance numerical stability.

Calibration and Out-of-Sample Forecasting

After training, predictions are generated for validation and test sets. Forecasts are inversely transformed to the log-return scale.

To reduce systematic bias across horizons, a per-horizon affine calibration is estimated on the validation set:

$$\tilde{y}_{t,h} = a_h \hat{y}_{t,h} + b_h \qquad (3.12)$$

where $a_h$and $b_h$are estimated separately for each horizon h. The calibrated forecasts are then applied to the test set.

Final forecasting metrics (RMSE, MAE, R^2, and directional accuracy) are computed on calibrated test predictions.

Role Within the Model Comparison Framework

The LSTM provides a nonlinear sequence-based benchmark within the shared multi-horizon experimental design. Because the target definition, encoder lengths, scaling procedure, data partitions, and evaluation metrics are standardized across SARIMAX, LSTM, and TFT, any observed performance differences can be attributed to model architecture rather than inconsistencies in data handling.

In this framework, the LSTM represents an intermediate level of complexity between the linear SARIMAX benchmark and the attention-based Temporal Fusion Transformer, allowing for a structured comparison across modeling paradigms.

3.5 Temporal Fusion Transformer (TFT) Model

The primary forecasting model employed in this study is the Temporal Fusion Transformer (TFT), an attention-based deep learning architecture specifically designed for multi-horizon time-series forecasting (Lim et al., 2021). The TFT extends standard sequence models by combining recurrent layers, gating mechanisms, variable selection networks, and interpretable self-attention blocks within a unified encoder–decoder framework.

Consistent with Section 3.2, the TFT is trained to forecast daily log returns and outputs multi-horizon predictions as defined in Eq. (3.4), with $H \in \{3,14,30\}$.

corresponds to the short-, medium-, and long-horizon forecasting buckets defined earlier.

Input Structure

The TFT distinguishes among three categories of inputs: static covariates, time-varying known inputs, and time-varying observed inputs.

The only static categorical variable is **symbol** (BTC, ETH, DOGE), which enables the model to learn asset-specific representations through embedding layers while allowing parameter sharing across cryptocurrencies.

Time-varying known real variables, available to both the encoder and decoder, include the per-asset time index (**time_idx**), calendar indicators (**dow**, **month**), and cyclical encodings of day-of-year (**sin_doy**, **cos_doy**). These variables capture weekly and annual seasonality and are deterministically known for all future horizons.

Time-varying unknown real variables consist of historically observed market quantities provided to the encoder. In addition to the target return series, the model incorporates engineered technical indicators derived from OHLCV data to capture trend, volatility, and momentum dynamics. Where available, external sentiment and on-chain indicators are also included. Continuous predictors are converted to numeric format, forward-filled, and zero-imputed when necessary to ensure a complete input matrix.

Table 6. TFT time-varying unknown (observed) real variables

Variable	Brief description
log_close	Log closing price
log_ret_1	1-day log return (target)
log_ret_7	7-day log return
ma_7 / ma_21 / ma_50	7/21/50-day moving average
vol_7 / vol_21 / vol_60	7/21/60-day return volatility
zscore_21	21-day standardized price deviation
rsi_14	14-day Relative Strength Index
macd	MACD indicator
macd_signal	MACD signal line
hl_range	(High − Low) / Close
oc_range	(Open − Close) / Close
atr_14	14-day Average True Range
trend_strength_30	30-day trend strength proxy
fear_greed	Sentiment index
mcap_usd, tvol_usd, TxCnt, AdrActCnt, FeeTotNtv	External market/on-chain variables

Source: Author's calculations.

Encoder Configuration and Data Structure

The TFT uses the same multi-horizon configuration defined earlier in Section 3.2 (p.23), ensuring comparability with the SARIMAX benchmark. For each horizon bucket, the encoder length (historical lookback window) is fixed according to the previously established configuration.

The dataset is structured as a grouped panel of time series, where each cryptocurrency represents a separate entity within a unified modeling framework. The target variable is the daily log return (**log_ret_1**), and all series are normalized at the asset level to account for scale differences across cryptocurrencies. This group-wise normalization allows the model to learn comparable patterns while preserving cross-sectional heterogeneity.

A strict chronological split is applied using the predefined training, validation, and test partitions to prevent information leakage and ensure genuinely out-of-sample evaluation.

Model Architecture

The Temporal Fusion Transformer integrates multiple components within a unified forecasting framework. The architecture consists of:

- Variable Selection Networks (VSNs) for static and time-varying inputs,
- LSTM-based local sequence encoders and decoders,
- Gated residual connections to regulate information flow,
- A multi-head self-attention block applied within the decoder,
- A fully connected output layer producing H-dimensional forecasts.

The self-attention mechanism follows the standard scaled dot-product formulation introduced in Section 2.4 (see p.18). Rather than repeating the equation here, we refer to the earlier presentation for the formal mathematical expression. In the present implementation, attention

is applied after recurrent processing to capture long-range temporal dependencies relevant for multi-horizon prediction.

Hyperparameter Configuration and Training

A limited grid search is conducted for each horizon bucket using the validation set to perform hyperparameter optimization. The parameter grid includes:

Table 7. TFT Hyperparameter Grid

Hidden Size	Attention Heads	Dropout	Learning Rate
16	1	0.1	1e-3
32	2	0.1	3e-4
64	4	0.1	3e-4

Source: Author's calculations.

For the long-horizon bucket, hidden size is capped at 32 to control model complexity.

After hyperparameter selection, the model is trained using the RMSE loss function for a maximum of 25 epochs, with early stopping applied based on validation performance (patience = 5). All experiments are conducted with a fixed random seed (42) to ensure reproducibility.

Calibration and Evaluation

For the short-horizon bucket, an affine calibration is estimated on the validation set (Eq. (3.12)) and applied to test forecasts. Model performance is evaluated using RMSE, MAE, R^2, and directional accuracy, both in aggregate and separate for each forecast horizon.

Trading Signal Construction

Multi-horizon forecasts are aggregated into a single trading score using exponentially decaying weights:

$$w_h = \frac{0.5^{h/\tau}}{\sum_{j=1}^{H} 0.5^{j/\tau}} \qquad (3.13)$$

where τ is the half-life parameter specific to each horizon bucket.

The resulting score is transformed into long-only or long-short trading positions depending on the asset, incorporating:

- Rolling z-score normalization
- Rolling quantile thresholds
- Transaction cost of 0.1% per trade

Performance metrics include total return, Sharpe ratio (annualized with factor 365), maximum drawdown, CAGR, and exposure.

Explainability

The TFT's interpretability mechanisms are used to extract:

- Variable importance (static, encoder, decoder blocks)
- Attention weight heatmaps

These outputs provide insight into the relative contribution of different features and historical time steps in generating forecasts.

3.6 Evaluation Metrics

This section defines the statistical criteria used to evaluate the forecasting performance of the SARIMAX, LSTM, and Temporal Fusion Transformer (TFT) models. As described in Section 3.3, the forecasting target is the daily log return r_{t+h}, defined in Eq. (3.1), evaluated at forecast horizon $h = 1, \dots, H$. All metrics are computed on the out-of-sample test set to ensure an unbiased assessment of predictive performance.

Forecast accuracy is evaluated using standard point-forecast error measures. Let $\mathrm{r_{t+h}}$ denote the realized log-return at horizon h, and $\hat{\mathrm{r}}_{\mathrm{t+h}}$ the corresponding forecast.

The **Root Mean Squared Error (RMSE)** is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}(\mathrm{r_i} - \hat{\mathrm{r}}_\mathrm{i})^2} \qquad (3.14)$$

where Nis the number of out-of-sample observations (aggregated across horizons when reported as an overall metric). RMSE penalizes larger errors more heavily due to the squared term and is therefore sensitive to extreme forecast deviations.

The **Mean Absolute Error (MAE)** is given by:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} |\,\mathrm{r_i} - \hat{\mathrm{r}}_\mathrm{i}\,| \qquad (3.15)$$

MAE provides a scale-consistent measure of average forecast deviation and is less sensitive to outliers compared to RMSE.

In addition, the **coefficient of determination (**$\mathrm{R^2}$**)** is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(r_i - \hat{r}_i)^2}{\sum_{i=1}^{N}(r_i - \bar{r})^2} \qquad (3.16)$$

where r̄denotes the mean of realized returns in the evaluation sample. Although $R^2$is often modest in financial return forecasting, it provides a standardized measure of explanatory power relative to a naive mean forecast benchmark.

Because trading decisions depend primarily on the direction of predicted returns rather than their exact magnitude, directional accuracy is also reported. The **Directional Accuracy (DA)** metric is defined as:

$$\mathrm{DA} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{1}\left(\mathrm{sign}(\mathrm{r_i}) = \mathrm{sign}(\hat{\mathrm{r}}_\mathrm{i})\right) \qquad (3.17)$$

where $\mathbf{1}(\cdot)$ is an indicator function equal to one when the predicted and realized return share the same sign and zero otherwise. This measure captures the model's ability to correctly anticipate upward versus downward price movements.

Given the multi-horizon forecasting setup, performance is evaluated at two levels. First, metrics are computed **separately for each prediction horizon** $\mathrm{h} = 1, \ldots, \mathrm{H}$ to assess how forecast accuracy evolves with increasing forecast distance. Second, an **overall metric** is calculated by aggregating prediction errors across all horizons, providing a single summary measure for model comparison within each horizon bucket (short, medium, long).

All models are evaluated using identical test samples and data partitions to ensure strict comparability. The reported forecasting metrics form the statistical basis for the subsequent trading strategy evaluation presented in the following section.

3.7 Trading Strategy and Backtesting Framework

While the previous section evaluates statistical forecasting accuracy, the practical relevance of the models is assessed through a systematic trading and backtesting framework. The objective is to examine whether improvements in predictive performance translate into economically meaningful returns after accounting for transaction costs and realistic trading constraints.

Signal Construction

The forecasting models generate multi-horizon predictions of future log returns $(\hat{r}_{t+1}, \dots, \hat{r}_{t+H})$. To convert these forecasts into a single trading signal at time t, an aggregated score is constructed using exponentially decaying weights across horizons. The horizon weights w_h are defined according to the exponentially decaying scheme introduced in **Eq. (3.13)**, where τrepresents the half-life parameter specific to each horizon bucket (short, medium, long).

The aggregated forecast score is then computed as:

$$S_t = \sum_{h=1}^{H} w_h \hat{r}_{t+h} \qquad (3.18)$$

This weighting scheme places greater emphasis on nearer-term predictions while still incorporating longer-horizon information.

To stabilize the signal and account for time-varying volatility, the score is standardized using a rolling z-score transformation:

$$Z_t = \frac{S_t - \mu_t}{\sigma_t} \qquad (3.19)$$

Position Sizing and Threshold Rules

Trading positions are determined using adaptive threshold rules designed to control the frequency of trades. A rolling quantile threshold is estimated over a historical window such that approximately a target proportion of observations (e.g., 8%) generate active signals.

For Bitcoin and Ethereum, a **long-only strategy** is implemented. A positive position is taken when the standardized score exceeds the threshold, while the position is zero otherwise.

For Dogecoin, a **long–short strategy** is applied. Positions are taken in the direction of the signal when the absolute standardized score exceeds the threshold, allowing both positive (long) and negative (short) exposures.

Positions are scaled proportionally to the signal strength and bounded within the interval $[-1,1]$in the long–short case or $[0, 1]$in the long-only case.

Return Calculation and Transaction Costs

Strategy returns are computed using the next-day realized simple return derived from log-returns:

$$R_{t+1} = e^{r_{t+1}} - 1 \qquad (3.20)$$

Let p_t denote the position held at time t. The gross strategy return is:

$$R_{t+1}^{gross} = p_t \cdot R_{t+1} \qquad (3.21)$$

Transaction costs are incorporated as proportional costs applied to turnover.

If $\Delta p_t = | p_t - p_{t-1} |$ denotes portfolio turnover and $c$the proportional cost rate (set to 0.1%), net returns are computed as:

$$R_{t+1}^{net} = p_t \cdot R_{t+1} - c \cdot \Delta p_t \qquad (3.22)$$

All reported trading performance metrics are calculated **net of transaction costs**.

To avoid look-ahead bias, positions are determined using only information available up to time t, and returns are realized at $t + 1$.

Performance Metrics

Trading performance is evaluated using standard portfolio statistics computed on the out-of-sample test period.

The **Total Return** is calculated as:

$$\text{Total Return} = \prod_{t=1}^{T}(1 + R_t^{net}) - 1 \qquad (3.23)$$

The **Sharpe Ratio (annualized)** is defined as:

$$\text{Sharpe} = \sqrt{365} \cdot \frac{\bar{R}}{\sigma_R} \qquad (3.24)$$

where $\bar{R}$and $\sigma_R$denote the mean and standard deviation of daily net returns, respectively.

The **Maximum Drawdown** measures the largest peak-to-trough decline in cumulative equity over the evaluation period.

The **Compound Annual Growth Rate (CAGR)** is computed as:

$$\mathrm{CAGR} = \left(\prod_{t=1}^{T} (1 + R_t^{net}) \right)^{\frac{365}{T}} - 1 \qquad (3.25)$$

In addition, **Exposure** is reported as the proportion of days during which the strategy holds a non-zero position.

Warm-Up Period and Robustness

Because rolling statistics are used for signal standardization and threshold estimation, an initial warm-up period is excluded from performance evaluation. This ensures that all reported results are based on fully initialized signals.

The same trading framework and parameter settings are applied consistently across SARIMAX, LSTM, and TFT forecasts. Consequently, differences in trading performance can be attributed to differences in predictive quality rather than to variations in the trading rules.

This framework provides a structured link between statistical forecasting accuracy and practical financial performance, enabling a comprehensive assessment of the economic value of advanced time-series forecasting models in cryptocurrency markets.

Chapter 4 – Empirical Results and Discussion

This chapter presents the empirical findings of the study and evaluates the performance of the forecasting models introduced in Chapter 3. The objective is to assess both the predictive accuracy and the economic relevance of the proposed approaches when applied to cryptocurrency markets.

The analysis focuses on three cryptocurrencies—Bitcoin, Ethereum, and Dogecoin—and compares the performance of three modeling frameworks: a traditional statistical benchmark (SARIMAX), a recurrent neural network model (LSTM), and an attention-based deep learning architecture (Temporal Fusion Transformer). All models are evaluated under a unified multi-horizon forecasting setup, ensuring consistency across assets and methodologies.

The results are organized into two main dimensions. First, forecasting performance is examined using statistical accuracy measures, including RMSE, MAE, R^2, and directional accuracy. These metrics provide insight into the models' ability to predict future log-returns across short-, medium-, and long-term horizons. Second, the economic value of the forecasts is assessed by translating model outputs into trading strategies and evaluating their net performance using risk-adjusted measures such as total return, Sharpe ratio, maximum drawdown, and exposure.

In addition to overall model comparison, particular attention is given to cross-asset differences. Since Bitcoin, Ethereum, and Dogecoin exhibit distinct structural and behavioral characteristics, the analysis investigates whether forecasting and trading performance varies systematically across assets.

Finally, for the Temporal Fusion Transformer, interpretability results are presented to examine variable importance and attention patterns, providing additional insight into how the model processes temporal information.

The findings reported in this chapter provide empirical basis for evaluating the research questions posed in Chapter 1 and for assessing the practical relevance of advanced deep learning methods in cryptocurrency forecasting.

4.1 Forecasting Performance

This section presents the out-of-sample forecasting performance of the SARIMAX, LSTM, and Temporal Fusion Transformer (TFT) models. Model accuracy is evaluated on the test set using Root Mean Squared Error (RMSE) as the primary metric, complemented by Mean Absolute Error (MAE), coefficient of determination (R^2), and directional accuracy.

The comparison is conducted separately for each cryptocurrency (Bitcoin, Ethereum, and Dogecoin) and for each prediction horizon bucket (short: H = 3, medium: H = 14, long: H = 30). The detailed results are reported in Tables 8–10 and visually illustrated in Figures 1–3.

Across the nine asset–horizon configurations reported in Tables 8–10, the Temporal Fusion Transformer achieves the lowest RMSE in seven out of nine cases, indicating a broad statistical advantage over SARIMAX and LSTM.

BITCOIN SECTION

For Bitcoin, the forecasting results are reported in Table 8.

Table 8. Forecasting Performance: Bitcoin

Horizon	Model	RMSE	MAE	Directional Accuracy
Short	SARIMAX	0.0192	0.0145	0.4607
	LSTM	0.0193	0.0146	0.4400
	TFT	**0.0175**	0.0128	0.4009
Medium	SARIMAX	0.0196	0.0149	0.4731
	LSTM	**0.0194**	0.0146	0.5732

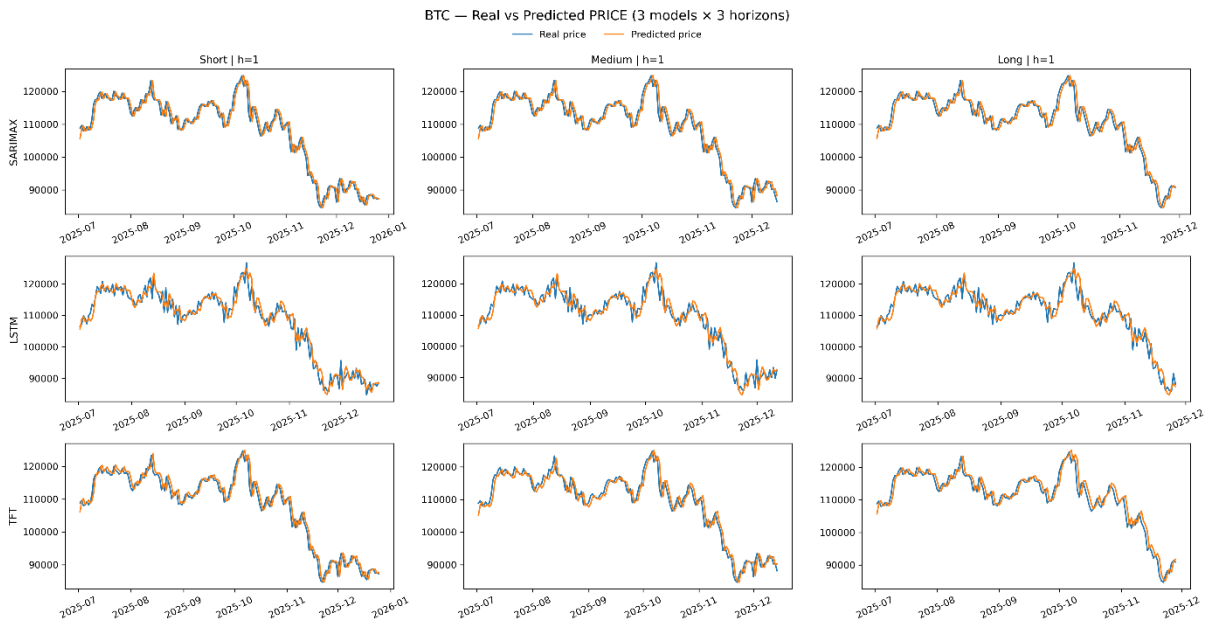
Horizon	Model	RMSE	MAE	Directional Accuracy
	TFT	0.0210	0.0179	0.6992
	SARIMAX	0.0190	0.0145	0.4702
Long	LSTM	0.0190	0.0145	0.4932
	TFT	**0.0148**	0.0116	0.2162

Source: Author's calculations.

As shown in Table 8, the TFT model achieves the lowest RMSE in the short horizon (0.0175) and the long horizon (0.0148), outperforming both SARIMAX and LSTM. In the medium horizon, however, LSTM records the lowest RMSE (0.0194), slightly improving upon SARIMAX (0.0196) and outperforming TFT (0.0210).

Directional accuracy does not fully align with RMSE. For example, in the medium horizon, TFT achieves the highest directional accuracy (0.6992) despite having the highest RMSE in that bucket.

Figure 1. BTC forecast plots



Source: Author's calculations.

The visual comparison in Figure 1 confirms that differences across models are relatively modest in medium horizons, while the TFT exhibits closer alignment in short and long horizons.

ETHEREUM SECTION

For Ethereum, forecasting performance is summarized in Table 9.

Table 9. Forecasting Performance: Ethereum

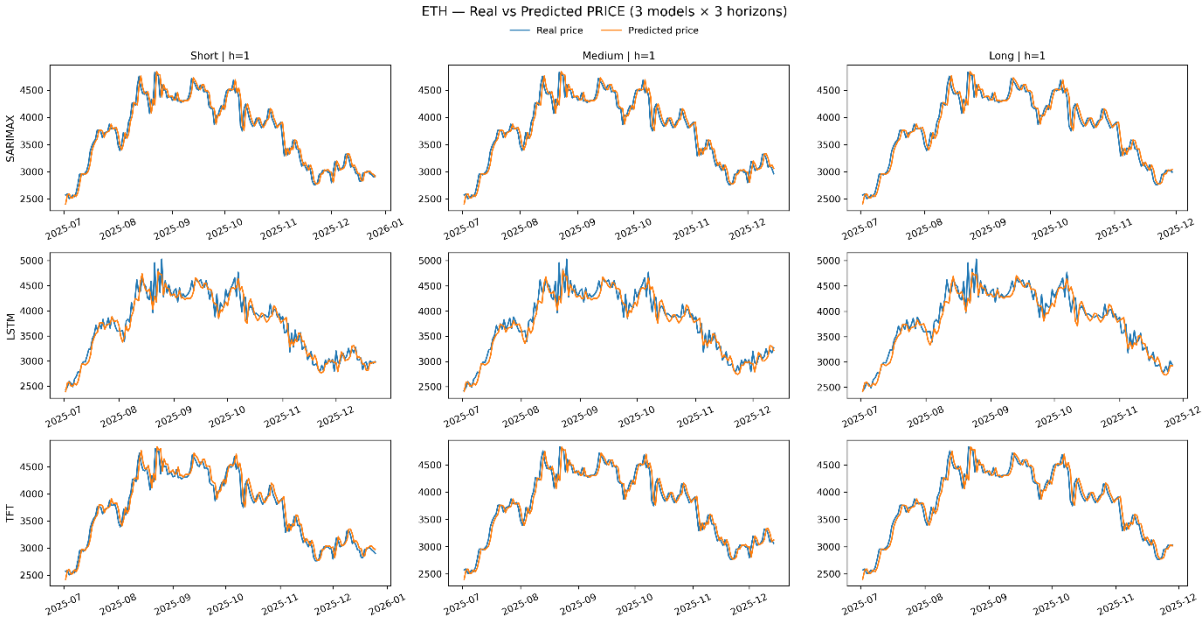
Horizon	Model	RMSE	MAE	Directional Accuracy
Short	SARIMAX	0.0362	0.0264	0.5169
	LSTM	0.0371	0.0276	0.5029
	TFT	**0.0351**	0.0274	0.4189
Medium	SARIMAX	0.0369	0.0273	0.5269
	LSTM	0.0380	0.0289	0.4573
	TFT	**0.0324**	0.0250	0.2773
Long	SARIMAX	0.0368	0.0269	0.5232
	LSTM	0.0386	0.0286	0.4932
	TFT	**0.0252**	0.0146	0.7387

Source: Author's calculations.

As shown in Table 9, the TFT model achieves the lowest RMSE in all three horizons: 0.0351 (short), 0.0324 (medium), and 0.0252 (long). The improvement is particularly pronounced in the long horizon, where TFT reduces RMSE substantially relative to SARIMAX (0.0368) and LSTM (0.0386).

Directional accuracy presents a different pattern. For instance, in the medium horizon, SARIMAX records higher directional accuracy (0.5269) compared to TFT (0.2773), illustrating that lower squared error does not necessarily imply improved directional prediction.

Figure 2. ETH forecast plots



Source: Author's calculations.

Figure 2 visually reinforces the numerical advantage of TFT, particularly in medium and long horizons.

DOGECOIN SECTION

Forecasting results for Dogecoin are reported in Table 10.

Table 10. Forecasting Performance: Dogecoin

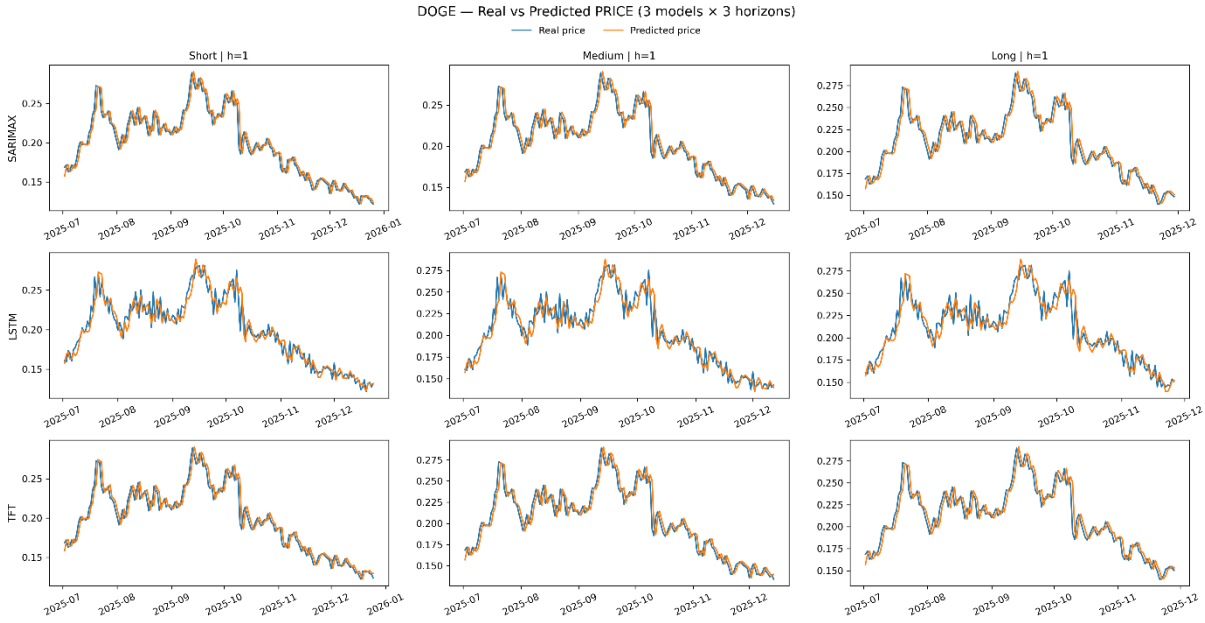
Horizon	Model	RMSE	MAE	Directional Accuracy
Short	SARIMAX	0.0470	0.0362	0.5000
	LSTM	**0.0470**	0.0359	0.5257
	TFT	0.0471	0.0383	0.3784
Medium	SARIMAX	0.0478	0.0368	0.5090
	LSTM	0.0479	0.0367	0.5122
	TFT	**0.0460**	0.0387	0.2813
Long	SARIMAX	0.0486	0.0371	0.5298
	LSTM	0.0491	0.0374	0.5000
	TFT	**0.0360**	0.0277	0.2372

Source: Author's calculations.

As shown in Table 10, results are more heterogeneous. In the short horizon, RMSE differences are minimal, with LSTM (0.0470) and SARIMAX (0.0470) performing marginally better than TFT (0.0471). In the medium and long horizons, however, TFT achieves the lowest RMSE (0.0460 and 0.0360, respectively), with a substantial improvement in the long horizon relative to SARIMAX (0.0486) and LSTM (0.0491).

Directional accuracy does not systematically favor TFT. In the long horizon, SARIMAX attains higher directional accuracy (0.5298) compared to TFT (0.2372), again illustrating the distinction between magnitude accuracy and directional performance.

Figure 3. DOGE forecast plots



Source: Author's calculations.

The visual evidence in Figure 3 confirms that the TFT model better captures large movements in the long horizon.

Overall, Tables 8–10 and Figures 1–3 demonstrate that the Temporal Fusion Transformer provides a consistent advantage in multi-horizon return prediction, particularly for Ethereum and long-horizon Dogecoin forecasts. Nevertheless, directional accuracy does not uniformly favor TFT, highlighting the multidimensional nature of forecast evaluation.

4.2 Trading Performance Comparison

While Section 4.1 evaluates statistical forecasting accuracy (Tables 8–10), the economic relevance of the models is assessed through the trading backtesting framework summarized in Table 11.

Table 11. Trading Performance Summary (All Assets & Horizons)

Asset	Horizon Bucket	Model	Total Return	Sharpe (ann., 365)	Max Drawdown	CAGR	Exposure	Horizon
BTC	Short	SARIMAX	-0.08	-4.359	-0.079	-0.31	0.049	H=3
BTC	Short	LSTM	-0.045	-1.801	-0.053	-0.189	0.088	H=3
BTC	Short	TFT	-0.037	-1.304	-0.058	-0.102	0.181	H=3
BTC	Medium	SARIMAX	0	0	0	0	0	H=14
BTC	Medium	LSTM	0	0	0	0	0	H=14
BTC	Medium	TFT	0	0	0	0	0	H=14
BTC	Long	SARIMAX	0	0	0	0	0	H=30
BTC	Long	LSTM	-0.058	-3.746	-0.076	-0.339	0.132	H=30
BTC	Long	TFT	-0.07	-5.547	-0.07	-0.105	0.08	H=30
DOGE	Short	SARIMAX	-0.121	-2.006	-0.197	-0.437	0.11	H=3
DOGE	Short	LSTM	0.088	1.409	-0.119	0.468	0.138	H=3
DOGE	Short	TFT	-0.234	-2.325	-0.24	-0.535	0.11	H=3
DOGE	Medium	SARIMAX	-0.139	-2.09	-0.182	-0.538	0.127	H=14
DOGE	Medium	LSTM	0.028	2.234	-0.001	0.156	0.014	H=14
DOGE	Medium	TFT	0	0	0	0	0	H=14
DOGE	Long	SARIMAX	0.036	0.924	-0.089	0.266	0.164	H=30
DOGE	Long	LSTM	0.203	2.135	-0.123	2.58	0.226	H=30
DOGE	Long	TFT	0.702	6.918	-0.001	1.26	0.134	H=30
ETH	Short	SARIMAX	-0.026	-1.669	-0.027	-0.11	0.049	H=3
ETH	Short	LSTM	-0.043	-0.605	-0.112	-0.181	0.212	H=3
ETH	Short	TFT	-0.172	-5.237	-0.172	-0.418	0.11	H=3
ETH	Medium	SARIMAX	0	0	0	0	0	H=14
ETH	Medium	LSTM	-0.056	-2.625	-0.073	-0.263	0.058	H=14
ETH	Medium	TFT	0	0	0	0	0	H=14
ETH	Long	SARIMAX	0	0	0	0	0	H=30
ETH	Long	LSTM	0	0	0	0	0	H=30
ETH	Long	TFT	0.071	4.33	-0.001	0.11	0.05	H=30

Source: Author's calculations.

The table reports total return, annualized Sharpe ratio (365-day scaling), maximum drawdown, CAGR, and exposure across all assets and horizon buckets.

The trading results reveal a markedly different ranking compared to the forecasting metrics. Although the TFT model achieves the lowest RMSE in seven out of nine asset–horizon configurations (Section 4.1), this statistical advantage does not translate uniformly into superior economic performance.

Bitcoin

For **Bitcoin**, trading performance is generally weak across all models and horizons.

In the **short horizon (H=3)**:

- TFT achieves the highest (least negative) total return at **–3.67%**, with a Sharpe ratio of **–1.30** and exposure of **18.11%**.
- LSTM produces a total return of **–4.49%** (Sharpe = –1.80).
- SARIMAX performs worst, with a total return of **–7.99%** (Sharpe = –4.36).

In the **medium horizon (H=14)**, all three models produce **zero total return and zero Sharpe ratio**, reflecting zero exposure. This indicates that the signal-generation mechanism did not activate positions during the evaluation period. Consequently, no economically meaningful differentiation can be established for this bucket.

In the **long horizon (H=30)**:

- LSTM records a total return of **–5.83%** (Sharpe = –3.75),
- TFT records **–6.95%** (Sharpe = –5.55),
- SARIMAX remains inactive (0% return, 0% exposure).

Overall, Bitcoin trading results are negative across most active configurations. Importantly, despite TFT's forecasting advantage in short and long horizons (Table 9), it does not produce superior economic outcomes for Bitcoin. This illustrates that improved RMSE does not guarantee positive trading profitability.

Ethereum

For **Ethereum**, trading activity is also limited in several configurations.

In the **short horizon (H=3)**:

- SARIMAX achieves the least negative total return of **–2.58%** (Sharpe = –1.67),
- LSTM produces **–4.28%** (Sharpe = –0.61),
- TFT performs worst with **–17.17%** (Sharpe = –5.24).

In the **medium horizon (H=14)**:

- LSTM yields **–5.61%** (Sharpe = –2.62),
- SARIMAX and TFT both generate zero exposure and zero return.

In the **long horizon (H=30)**:

- TFT is the only active strategy, achieving a positive total return of **+7.05%**, with a Sharpe ratio of **4.33**, maximum drawdown of only **–0.10%**, and exposure of **5.04%**.
- Both LSTM and SARIMAX remain inactive (0% exposure).

Thus, while TFT dominates statistically for Ethereum (Table 10), its economic superiority emerges clearly only in the **long horizon**, where it is the only strategy generating positive returns.

Dogecoin

For **Dogecoin**, the trading results are substantially stronger and more differentiated across models.

In the **short horizon (H=3)**:

- LSTM achieves a positive total return of **+8.78%** (Sharpe = 1.41),
- SARIMAX records **–12.11%** (Sharpe = –2.01),
- TFT records **–23.41%** (Sharpe = –2.32).

In the **medium horizon (H=14)**:

- LSTM generates **+2.78%** (Sharpe = 2.23),
- SARIMAX produces **–13.95%** (Sharpe = –2.09),

- TFT remains inactive (0% exposure).

The most pronounced economic outcome appears in the **long horizon (H=30)**:

- TFT achieves a total return of **+70.20%**, with an exceptionally high Sharpe ratio of **6.92**, maximum drawdown of only **–0.10%**, and exposure of **13.45%**.
- LSTM also performs strongly, generating **+20.34%** (Sharpe = 2.14),
- SARIMAX produces **+3.61%** (Sharpe = 0.92).

This represents the strongest economic result observed across all asset–horizon combinations. The long–short framework applied to Dogecoin allows the TFT architecture to exploit both upward and downward return dynamics, and its attention-based multi-horizon structure appears particularly effective in capturing medium- to longer-term trends in this more volatile asset.

Cross-Model Comparison

Across all assets and horizons:

- The highest Sharpe ratio in the entire study is achieved by **TFT for Dogecoin (H=30)** with **Sharpe = 6.92**.
- The second-highest Sharpe ratio is **4.33** for **TFT in Ethereum (H=30)**.
- LSTM delivers competitive performance for Dogecoin (Sharpe = 2.23 in medium horizon; 2.14 in long horizon).
- SARIMAX never achieves the top Sharpe ratio in any asset–horizon combination.

However, several configurations exhibit zero exposure (particularly SARIMAX and TFT in certain Ethereum and Bitcoin settings). These cases demonstrate that trading profitability depends critically on signal activation frequency, not merely forecasting precision.

Interpretation

Taken together, Table 11 demonstrates that predictive accuracy and trading profitability are related but distinct objectives. Although the TFT model achieves the lowest RMSE in seven out of nine configurations (Section 4.1), its economic superiority is concentrated primarily in:

- **Dogecoin (H=30)** strongest overall performance,
- **Ethereum (H=30)** moderate positive performance.

In contrast, for Bitcoin, none of the models produce economically attractive results.

These findings confirm that while attention-based multi-horizon architectures provide a consistent statistical advantage, their economic value depends strongly on asset characteristics, exposure dynamics, and the interaction between signal construction and volatility structure.

4.3 Discussion of Results

The empirical findings reported in Sections 4.1 and 4.2 (Tables 8–11) provide several insights regarding the relationship between model architecture, forecasting accuracy, and economic performance in cryptocurrency markets.

First, the results confirm that model complexity and architectural design materially influence multi-horizon forecasting accuracy. As reported in Tables 8–10, the Temporal Fusion Transformer (TFT) achieves the lowest RMSE in **seven out of nine asset–horizon configurations**. In particular, TFT dominates for Ethereum across all three horizons (Table 9) and for Dogecoin in the medium and long horizons (Table 10). For example, in Ethereum (H = 30), TFT reduces RMSE to **0.0252**, compared to 0.0368 (SARIMAX) and 0.0386 (LSTM). Similarly, in Dogecoin (H = 30), TFT achieves an RMSE of **0.0360**, substantially lower than 0.0486 and 0.0491 for SARIMAX and LSTM, respectively. These magnitudes indicate that the

attention-based multi-horizon structure provides a measurable statistical advantage in modeling nonlinear cryptocurrency return dynamics.

Second, the comparison between LSTM and SARIMAX highlights the incremental benefit of nonlinear modeling. Although LSTM outperforms SARIMAX in selected configurations — for instance, Bitcoin medium horizon (RMSE = 0.0194 vs. 0.0196; Table 8) it does not consistently dominate across assets. Moreover, LSTM does not outperform TFT in any asset–horizon combination when considering the full set of nine configurations. This suggests that recurrent memory alone is insufficient when modeling multi-horizon dependencies and complex feature interactions, which are more effectively captured by the TFT's variable selection and attention mechanisms.

Third, the divergence between statistical forecasting accuracy and economic trading performance is clearly observable when comparing Tables 8–10 with Table 11. Although TFT achieves the lowest RMSE in seven configurations, its economic superiority is concentrated in specific settings rather than universal. The strongest economic outcome in the entire study occurs for **Dogecoin in the long horizon (H = 30)**, where TFT achieves a total return of **70.20%**, a Sharpe ratio of **6.92**, and a maximum drawdown of only –0.10% (Table 11). In contrast, for Bitcoin, all active strategies generate negative total returns across horizons, despite TFT achieving the lowest RMSE in short and long horizons. For example, in Bitcoin (H = 3), TFT records RMSE = 0.0175 (Table 8) but a total return of –3.67% and Sharpe = –1.30 (Table 11). This confirms that improvements in RMSE do not automatically translate into economically superior outcomes.

Fourth, asset-specific behavior plays a decisive role. Bitcoin exhibits relatively stable performance across models, with modest RMSE differences (Table 8) and generally weak trading results (Table 11). Ethereum shows strong statistical gains under TFT — particularly in the long horizon (RMSE = 0.0252; Table 9) yet the most economically meaningful outcome

appears only in the long-horizon trading configuration (Sharpe = 4.33; Table 11), while other horizons remain inactive or negative. Dogecoin displays the clearest economic differentiation, with TFT achieving both the lowest long-horizon RMSE (0.0360; Table 10) and the highest Sharpe ratio (6.92; Table 11). This suggests that more volatile and speculative assets may benefit more substantially from flexible attention-based architectures.

Fifth, the role of the multi-horizon structure is central to understanding these results. The TFT model generates simultaneous forecasts across H future steps, allowing smoother signal aggregation within the trading framework described in Section 3.7.

The superior performance of TFT in Dogecoin (H = 30) suggests that integrating cross-horizon dependencies may enhance stability and directional capture over longer time frames. By contrast, SARIMAX and LSTM, although adapted to the same evaluation framework, do not inherently integrate cross-horizon information in the same structured manner.

Taken together, the empirical evidence supports three conclusions:

1. Attention-based deep learning models provide a consistent statistical forecasting advantage in cryptocurrency markets, as evidenced by lower RMSE in seven of nine configurations (Tables 8–10).
2. Economic gains depend not only on predictive accuracy but also on signal activation, exposure dynamics, and asset-specific volatility characteristics (Table 11).
3. Model performance varies meaningfully across cryptocurrencies, indicating that digital assets should not be treated as a homogeneous asset class in forecasting research.

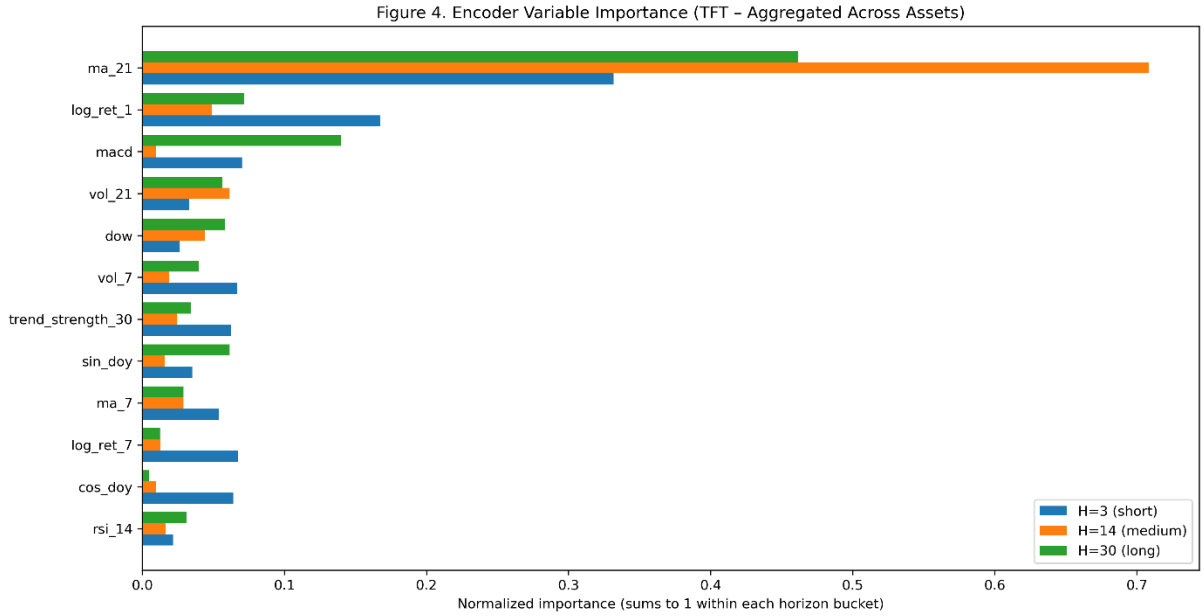
These findings contribute to the broader literature by demonstrating that while advanced deep learning architectures can meaningfully improve return prediction accuracy, their economic value must be evaluated within a realistic trading framework. The joint consideration of statistical metrics (Tables 8–10) and financial performance measures (Table 11) provides a comprehensive assessment of model effectiveness in practice.

4.4 Model Interpretation and Explainability

Beyond statistical accuracy and trading performance, the Temporal Fusion Transformer (TFT) provides interpretable components that allow direct inspection of how the model forms its forecasts. In particular, the variable selection networks and the multi-head attention mechanism offer quantitative measures of feature relevance and temporal weighting. The interpretation presented here integrates the variable-importance rankings and the attention heatmaps extracted from the trained TFT models (see Figures 4–6), providing additional insight into the structural drivers of the forecasting results discussed in Sections 4.1–4.3, where Tables 8–10 summarize forecasting performance and Table 11 summarizes trading performance.

The variable-importance outputs reveal a consistent pattern across assets and horizons. Return-based features and volatility-related indicators receive the highest importance weights in the encoder variable selection network. Short-term lagged returns (e.g., log-return features) are particularly dominant in the short-horizon bucket (H = 3), whereas rolling volatility measures (e.g., 7-day and 21-day standard deviations) become increasingly influential in medium and long horizons. Technical trend indicators, including moving averages and trend-strength measures, exhibit greater importance as the forecasting horizon expands, indicating that longer-term structural information becomes more relevant when predicting further into the future.

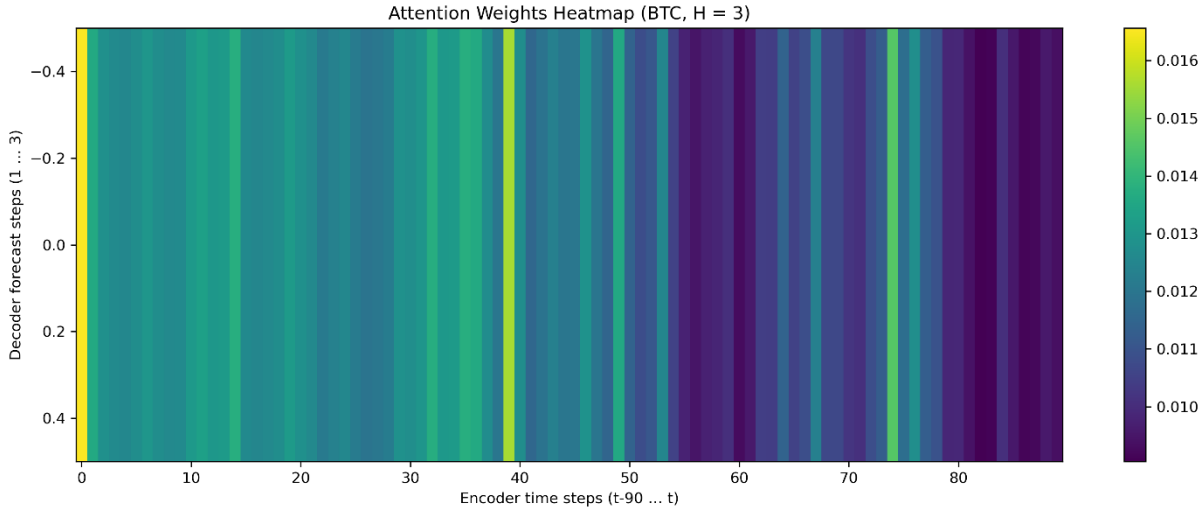
Figure 4. Encoder Variable Importance (TFT – Aggregated Across Assets)



Source: Author's calculations (TFT encoder variable selection network).

Calendar-based variables (day-of-week and cyclical transformations such as sine and cosine of day-of-year) receive comparatively lower but non-negligible importance weights. Their contribution suggests the presence of mild seasonal structure, but the dominance of return and volatility features confirms that the model primarily relies on endogenous price dynamics rather than deterministic calendar effects.

Figure 5. Attention Weights Heatmap (Bitcoin, H = 3)

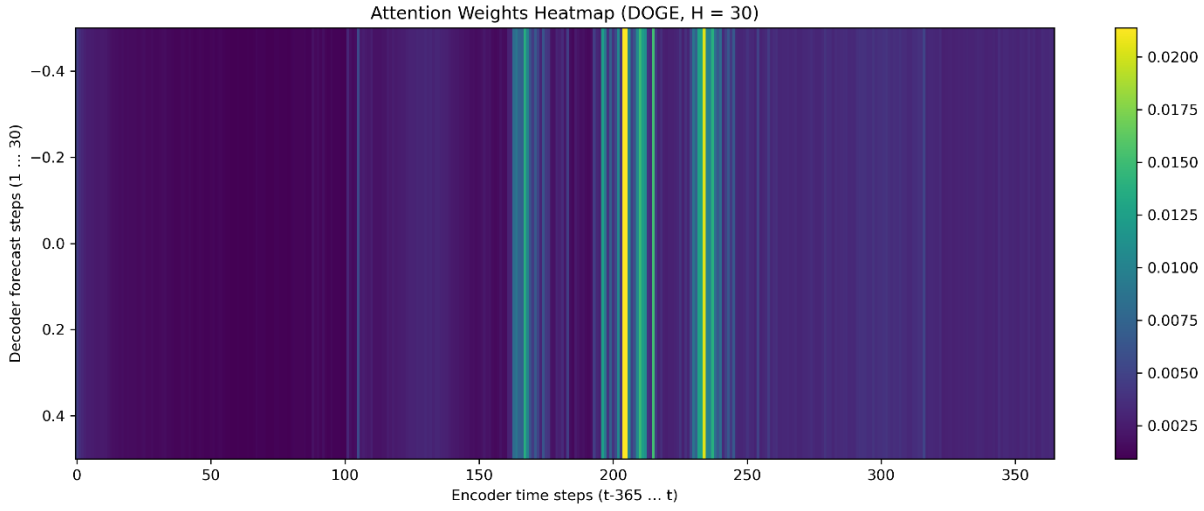


Source: Author's calculations (TFT attention mechanism output).

Importantly, these variable-importance rankings are consistent with the quantitative forecasting results reported in Tables 8–10. For example, the strong long-horizon performance of TFT for Dogecoin (RMSE = 0.0360; Table 10) coincides with higher importance assigned to volatility and trend-strength variables, which are particularly relevant in highly volatile assets. Similarly, Ethereum's substantial RMSE improvement in the long horizon (0.0252; Table 9) aligns with a broader distribution of importance across both momentum and volatility features. These patterns suggest that the TFT's statistical advantage is not arbitrary but reflects systematic prioritization of economically meaningful predictors.

The attention heatmaps provide complementary insight into how the model allocates weight across historical time steps. Across all assets, attention weights exhibit a clear recency bias: the most recent encoder observations receive the highest weights, especially in the short-horizon bucket. This pattern is consistent with financial market behavior, where recent information typically has greater predictive relevance.

Figure 6. Attention Weights Heatmap (Dogecoin, H = 30)



Source: Author's calculations (TFT attention mechanism output).

As the forecasting horizon increases, the attention distribution becomes more dispersed across earlier time steps. In medium and long horizons, the model assigns non-negligible weight to a broader range of past observations, effectively extending its temporal memory. This horizon-dependent dispersion is particularly visible in Ethereum and Dogecoin heatmaps (see Figures 4–6), where long-horizon forecasts incorporate more distant historical information.

Asset-specific differences are also observable. For Bitcoin, attention weights decline smoothly as one moves backward in time, indicating relatively stable temporal dynamics. Ethereum exhibits broader dispersion in medium horizons, suggesting more complex medium-term dependencies. Dogecoin occasionally displays concentrated attention on specific historical segments, consistent with episodic volatility and speculative behavior. These temporal weighting patterns align with the asset-specific performance differences documented in Section 4.3, where Dogecoin demonstrates the strongest economic differentiation under the TFT model.

Taken together, the variable-importance rankings and attention patterns support three key interpretations. First, the TFT model emphasizes economically intuitive predictors recent returns, volatility, and trend measures — rather than relying on spurious inputs. Second, the model dynamically adjusts its effective memory length depending on the forecasting horizon, integrating longer historical context when predicting further ahead. Third, asset-specific temporal structures are reflected in distinct attention distributions, reinforcing the conclusion that cryptocurrencies should not be treated as a homogeneous asset class.

It is important to note that these interpretability measures are relative rather than causal. Variable-importance scores reflect the contribution of inputs within the internal architecture of the model, and attention weights represent normalized relevance within the self-attention mechanism. They should therefore be interpreted as indicators of model behavior rather than evidence of structural economic causality. Nonetheless, compared to traditional recurrent

architectures such as LSTM, the TFT provides substantially greater transparency, strengthening confidence in the structural validity of the forecasting framework.

Overall, the explainability analysis reinforces the empirical findings reported earlier. The statistical superiority of TFT in seven out of nine forecasting configurations (Tables 8–10) and its strong long-horizon economic performance in Dogecoin (Table 11) are consistent with a model architecture that selectively emphasizes volatility, trend persistence, and recent dynamics while adaptively adjusting its temporal focus.

Chapter 5 – Conclusion

This thesis investigated whether an advanced multi-horizon deep learning architecture the Temporal Fusion Transformer (TFT) provides superior forecasting performance in cryptocurrency markets and whether improvements in predictive accuracy translate into economically meaningful trading outcomes. The analysis was conducted using daily data for Bitcoin, Ethereum, and Dogecoin over the period 2018–2025, within a unified and reproducible experimental framework comparing SARIMAX, LSTM, and TFT models across short (H = 3), medium (H = 14), and long (H = 30) horizons.

The empirical findings provide clear evidence that model architecture plays a decisive role in multi-horizon forecasting performance. The TFT achieves the lowest RMSE in seven out of nine asset–horizon configurations, demonstrating a consistent statistical advantage over both the linear benchmark (SARIMAX) and the recurrent neural network benchmark (LSTM). The strongest forecasting improvements are observed for Ethereum across all horizons and for Dogecoin in the medium and long horizons. These results indicate that attention-based architectures are well suited to capturing nonlinear and heterogeneous temporal dynamics characteristic of cryptocurrency returns.

However, the analysis also reveals that statistical superiority does not uniformly translate into economic profitability. Trading results show substantial variation across assets and horizons. The most pronounced economic outcome emerges for Dogecoin in the long horizon (H = 30), where the TFT strategy achieves a total return of 70.2% and an annualized Sharpe ratio of 6.92, representing the strongest performance across all configurations. For Ethereum, economically meaningful gains appear primarily in the long horizon. In contrast, trading results for Bitcoin remain weak across models, despite measurable improvements in forecasting accuracy. These

findings confirm that predictive accuracy and trading profitability are related but distinct objectives, and that economic value depends critically on signal construction, exposure dynamics, and asset-specific volatility characteristics.

The cross-asset comparison further demonstrates that cryptocurrencies should not be treated as a homogeneous asset class. Model effectiveness varies meaningfully across Bitcoin, Ethereum, and Dogecoin, reflecting differences in market maturity, volatility structure, and behavioral drivers. Assets characterized by higher volatility and speculative dynamics appear to benefit more substantially from flexible attention-based forecasting frameworks.

Beyond performance evaluation, the interpretability analysis strengthens confidence in the structural validity of the TFT model. Variable importance rankings emphasize economically intuitive predictors such as recent returns, volatility measures, and trend indicators, while attention patterns reveal horizon-dependent temporal weighting. These findings suggest that the model's statistical advantage reflects structured learning of relevant financial features rather than arbitrary complexity.

Despite these contributions, several limitations should be acknowledged. The analysis is restricted to three cryptocurrencies and daily frequency data, limiting generalization to other digital assets or intraday settings. The trading framework incorporates proportional transaction costs but does not account for slippage, liquidity constraints, or market impact. Moreover, hyperparameter optimization is conducted within a predefined grid, and alternative architectural specifications may yield different outcomes. Finally, the evaluation period represents a specific historical regime, and model stability under future structural shifts remains an open question.

Future research may extend this framework in several directions. First, expanding the asset universe to include additional cryptocurrencies or token categories would allow broader cross-sectional analysis. Second, incorporating alternative data sources such as high-frequency order book information or sentiment indicators may enhance predictive performance. Third,

exploring adaptive or regime-switching attention mechanisms could improve robustness under structural breaks. Finally, integrating portfolio-level optimization across assets may provide additional insight into the economic relevance of multi-horizon deep learning forecasts.

In conclusion, this thesis demonstrates that attention-based multi-horizon architectures provide a consistent statistical forecasting advantage in cryptocurrency markets. However, the translation of predictive gains into trading profitability depends on asset-specific dynamics and strategy design. A comprehensive evaluation that jointly considers statistical accuracy and economic performance is therefore essential when assessing advanced forecasting models in financial applications.

Bibliography

1. Baur, D. G., Dimpfl, T., & Kuck, K. (2018). Bitcoin, gold and the US dollar: A replication and extension. *Finance Research Letters, 25*, 103–110. https://doi.org/10.1016/j.frl.2017.10.012
2. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons.
3. Brandvold, M., Molnár, P., Vagstad, K., & Valstad, O. C. A. (2015). Price discovery on Bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money, 36*, 18–35. https://doi.org/10.1016/j.intfin.2015.02.010
4. Chu, J., Chan, S., Nadarajah, S., & Osterrieder, J. (2017). GARCH modelling of cryptocurrencies. *Journal of Risk and Financial Management, 10*(4), 17. https://doi.org/10.3390/jrfm10040017
5. Corbet, S., Lucey, B., Urquhart, A., & Yarovaya, L. (2019). RETRACTED: Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis, 62*, 182–199. https://doi.org/10.1016/j.irfa.2018.09.003
6. Corbet, S., Meegan, A., Larkin, C., Lucey, B., & Yarovaya, L. (2018). Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Economics Letters, 165*, 28–34. https://doi.org/10.1016/j.econlet.2018.01.004
7. 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
8. Grobys, K., Junttila, J., Kolari, J. W., & Sapkota, N. (2021). On the stability of stablecoins. *Journal of Empirical Finance, 64*, 207–223. https://doi.org/10.1016/j.jempfin.2021.09.002

9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

10. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

11. Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters, 158*, 3–6. https://doi.org/10.1016/j.econlet.2017.06.023

12. Lim, B., Arik, 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. https://doi.org/10.1016/j.ijforecast.2021.03.012

13. Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379*(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

14. McNally, S., Roche, J., & Caton, S. (2018). Predicting the price of Bitcoin using machine learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (pp. 339–343). IEEE. https://doi.org/10.1109/PDP2018.2018.00060

15. Nelson, D. M. Q., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 1419–1426). IEEE. https://doi.org/10.1109/IJCNN.2017.7966019

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need* (arXiv:1706.03762). arXiv. https://arxiv.org/abs/1706.03762