A close up of a logo

Description automatically generated

Faculty of Business and Economics

Chair of Business Information Systems, esp. Business Engineering OR   
Information Systems in Trade and Industry

**Thesis/Paper Title**

**Research Paper (Belegarbeit)**

**Seminar Paper**

*Course Name*

|  |  |  |
| --- | --- | --- |
|  | Submitted by: | Last name, first name(s) |
|  | Email: | [firstname.lastname@mailbox.tu-dresden.de](mailto:firstname.lastname@mailbox.tu-dresden.de) |
|  | Student ID: | 1234567 |
|  | Study program: | Program name |
|  | Chair: | Prof. Dr. Martin Wiener |
|  | Supervisor: | Supervisor name |
|  | Working period: | DD.MM.YYYY – DD.MM.YYYY |

Abstract

The abstract introduces the study topic and research objectives/questions (including a brief motivation), outlines the methodological approach used, as well as summarizes the study’s key findings and contributions. The length of the abstract is limited to one page (max).

Table of Contents

[List of Figures III](#_Toc124513004)

[List of Tables IV](#_Toc124513005)

[List of Formulas V](#_Toc124513006)

[List of Abbreviations VI](#_Toc124513007)

[1 Introduction (Heading 1) 1](#_Toc124513008)

[2 Template Use 1](#_Toc124513009)

[2.1 Headings (Heading 2) 1](#_Toc124513010)

[2.2 Standard Text and Accentuations 1](#_Toc124513011)

[2.3 Bulleted Lists and Numberings 1](#_Toc124513012)

[2.4 Figures, Tables and Formulas 2](#_Toc124513013)

[2.4.1 Figures (Heading 3) 2](#_Toc124513014)

[2.4.2 Tables 2](#_Toc124513015)

[2.4.3 Formulas 3](#_Toc124513016)

[2.5 Footnotes 3](#_Toc124513017)

[2.6 Headers 3](#_Toc124513018)

[2.7 Abbreviations 3](#_Toc124513019)

[2.8 References 3](#_Toc124513020)

[2.9 Index Tables 3](#_Toc124513021)

[2.10 Declaration of Academic Integrity and Originality 4](#_Toc124513022)

[References 5](#_Toc124513023)

[Appendix A: Appendix Title VII](#_Toc124513024)

[Appendix B: Appendix Title VIII](#_Toc124513025)

List of Figures

[Figure 1: Meaningful figure caption (Source, 1979, p. 25) 2](#_Toc35525153)

[Figure 2: Pointer on how to set the correct page number for the appendix 4](#_Toc35525154)

List of Tables

[Table 1: Meaningful table caption 2](#_Toc35525155)

[Table 2: Use of Arabic and Roman page numbers 4](#_Toc35525156)

List of Formulas

[Formula 1: Binomial theorem 3](#_Toc35525157)

# Introduction

Time series forecasting functions as a vital instrument throughout different fields especially in finance and economics because it enables precise predictions which guide stock price prediction and demand forecasting and economic analysis (Makridakis et al., 2018). The reliability of forecasts stands as the highest priority in these high-risk fields because small prediction inaccuracies can create major financial repercussions (Fischer & Krauss, 2018). The production of "hallucinations" which are unrealistic predictions remains a challenge for advanced forecasting models when dealing with volatile market conditions or unexpected trends (Qin et al., 2017). The occurrence of hallucinations damages decision-making processes because forecasting models need to deliver both accurate and stable predictions.

The development of time series forecasting models resulted in multiple approaches each offering unique advantages and constraints. The statistical forecasting community endorses ARIMA models because they work well with stationary data that has linear patterns but these models lack ability to handle non-linear or complex data sets (Box & Jenkins, 1976; Siami-Namini et al., 2018). The Long Short-Term Memory (LSTM) network represents a recurrent neural network which effectively detects complex temporal dependencies and non-linear relationships thus making it suitable for financial forecasting tasks where patterns tend to be irregular (Hewamalage et al., 2021; Fischer & Krauss, 2018). The predictive power of LSTMs leads to abnormal output during market volatility which results in hallucinations when dealing with unpredictable market conditions. Prophet as a Facebook-developed tool provides a combination of trend and seasonality modeling which results in business-friendly forecasting solutions (Taylor & Letham, 2018). Prophet shows strong performance in handling seasonal patterns, but it tends to simplify complex data structures especially during periods of high data variability. Research needs to address the gap in understanding how different models perform regarding accuracy and stability when exposed to varied conditions.

The forecasting models demonstrate strong versatility when applied to various application domains which extend beyond financial use. The combination of LSTM and Prophet proves effective for pharmaceutical sales forecasting because they track non-linear demand patterns which stem from seasonal changes and health-related external influences (Mohammed et al., 2024). The environmental sciences utilize ARIMA, LSTM and Prophet models to predict PM2.5 concentrations where LSTM demonstrates superior performance in detecting complex temporal patterns (Karnati et al., 2025). The models have been implemented in epidemiology for COVID-19 spread forecasting where Prophet's ability to model holiday effects proved beneficial (Chen et al., 2024). The energy sector utilizes ARIMA and LSTM models to forecast oil production while Prophet model’s server resource consumption in cloud computing environments to optimize auto-scaling systems (Chia et al., 2024). The multiple applications across different domains strengthen the need to assess their performance in unstable and unpredictable situations particularly during critical decision-making processes.

The main objective is to perform a comparative analysis of three leading models in time series forecasting: ARIMA, LSTM, and Prophet, with a focus on the evaluation of both accuracy and hallucination tendencies. Volatility is defined as the degree of variation in a time series over time and is a critical factor in the analysis Volatility, defined as the degree of variation in a time series over time, is a critical factor in the analysis (Sbai et al., 2024). Financial time series often exhibit varying levels of volatility, with stable periods showing predictable trends and volatile periods characterized by abrupt and irregular price movements. These contrasting conditions significantly impact the performance of forecasting models, challenging their ability to adapt to non-linear and unpredictable patterns.

The research addresses the following question:

*How do ARIMA, LSTM, and Prophet models compare in terms of forecasting accuracy and hallucination tendencies when applied to time series data under different levels of volatility?*

The research question will be answered through prediction interval checks to detect overconfidence and Z-score analysis to detect extreme deviations from historical norms. The forecasting accuracy of each model will be assessed through standard accuracy metrics which include RMSE, MAE and MAPE (Hyndman & Koehler, 2006). The analysis reveals the accuracy-stability trade-offs across the models which provides important information about their reliability in real-world applications. The analysis evaluates model performance throughout periods of lower and higher volatility to determine their ability to manage volatility fluctuations.

The research contributes to time series forecasting by analyzing accuracy and hallucination tendencies across different conditions. The study of both stable and volatile periods reveals how each model performs under difficult conditions which makes the results highly important for financial and economic applications with high stakes (Hewamalage et al., 2021). The research provides practitioners with guidance to choose models that produce accurate forecasts while maintaining stable realistic predictions for better forecasting practices.

# Theoretical Background

## Time Series Forecasting

Time series forecasting involves analyzing and using past data points, observed sequentially over time, to predict future values. This process is critical across various industries, including finance, economics, energy, and retail, where it serves as the backbone for decision-making in domains such as stock price prediction, demand forecasting, resource planning, and economic trend analysis (Makridakis et al., 2018). Time series data typically exhibit recurring patterns, such as trends, seasonality, and cyclicality, which forecasting models aim to capture and project. For instance, in finance, trends might reflect long-term growth in stock prices, while seasonality could represent predictable fluctuations tied to specific times of the year, such as the “January effect,” where stock returns tend to increase during the first month of the year (Kajol et al., 2020).

The distinctive patterns in time series data distinguish it from other data types. The process of forecasting comes with multiple difficulties. The established patterns in time series data become difficult to predict because external shocks and market volatility and non-linear dependencies create disruptions. Advanced models struggle to produce hallucinations which this study defines as predictions that exceed realistic levels or deviate substantially from past trends because of overfitting or noise or poor generalization (Qin et al., 2017; Li et al., 2023). An LSTM model trained on Tesla stock data could generate a 50 % daily price increase prediction during market volatility although such an extreme price fluctuation is unlikely to occur in real-world market situations. Decision-makers need forecasting models that deliver accurate results while providing interpretable and realistic predictions because hallucinations from these models can lead them toward incorrect choices in critical situations.

## Role of AI and Recent Developments

The field of time series forecasting underwent a fundamental transformation through Artificial Intelligence (AI) developments during the last ten years. ARIMA models together with other statistical methods have traditionally provided the basis for predicting stationary data that follows linear patterns. The modern datasets present rising complexity and non-linear patterns and volatility which reveal weaknesses in traditional forecasting approaches. Deep learning advances within AI have introduced effective alternatives which successfully model complex temporal relationships and unpredictable patterns in financial economic and operational data (Goodfellow, Bengio, & Courville, 2016).

AI has revolutionized forecasting through the creation and widespread implementation of Recurrent Neural Networks (RNNs) and their advanced versions including Long Short-Term Memory (LSTM) networks. LSTM models solve the vanishing gradient problem which restricted previous RNNs by enabling the detection of long-term dependencies and non-linear relationships in time series data (Hochreiter & Schmidhuber, 1997). The new capabilities have created possibilities for precise forecasting in highly volatile environments that traditional models cannot handle.

The requirement for forecasting solutions that users can understand led to the creation of Prophet models. Prophet which Taylor and Letham (2018) introduced brings together statistical modeling with AI principles to detect trends and seasonality automatically while providing easy-to-use parameter adjustment capabilities. The hybrid approach made forecasting accessible to practitioners who lacked advanced technical skills so they could implement robust forecasting models in business and operational settings.

Model innovations have been accompanied by parallel developments in evaluation practices. The current research approach uses three error metrics RMSE, MAE and MAPE to evaluate forecasting accuracy (Hyndman & Koehler, 2006). The evaluation now includes assessing model robustness against hallucinations which refers to unrealistic predictions especially during volatile market conditions.

The current advancements demonstrate how AI systems combine with conventional statistical models to develop hybrid systems which maintain both accuracy and interpretability and stability. The current forecasting trend demonstrates how AI serves to develop models which provide transparency alongside adaptability and resilience in various real-world conditions.

## Conceptual Foundations of Forecasting Models

### ARIMA (AutoRegressive Integrated Moving Average)

The ARIMA model which Box and Jenkins established in the 1970s continues to be one of the primary statistical methods people use for time series forecasting (Box & Jenkins, 1976). The model consists of three fundamental elements: autoregression (AR) which establishes relationships between current data points and previous values; differencing (I) which stabilizes non-stationary data by extracting trends; and moving average (MA) which uses past forecast errors to generate better future predictions. These components within ARIMA allow the model to produce forecasts from identified linear patterns in the data (Box & Jenkins, 1976).

ARIMA achieves widespread use because of its easy interpretation and straightforward design which makes it suitable for financial and environmental applications as well as educational analytics (Chanchí-Golondrino et al., 2025). Financial institutions commonly employ ARIMA models to analyze stock returns and currency exchange rates when the data shows stable stationarity (Siami-Namini et al., 2018). Real-world applications face challenges because the model assumptions create limitations. ARIMA models fail to detect nonlinear relationships between data points while being highly reactive to sudden changes in trends or volatility thus needing extensive preprocessing.

### LSTM (Long Short-Term Memory)

#### Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational structures inspired by the biological neural networks of the human brain. An ANN consists of interconnected layers of nodes (neurons) that process input data by applying activation functions and weighted connections. These networks iteratively adjust their weights to minimize prediction errors and improve performance through a process called backpropagation (LeCun, Bengio, & Hinton, 2015). ANNs are particularly effective at learning complex, non-linear relationships, making them suitable for a wide range of applications such as image recognition, classification tasks, and pattern detection. However, traditional feedforward ANNs are not designed to model temporal dependencies, as they assume that inputs are independent and identically distributed, limiting their effectiveness on sequential data such as time series.

#### Recurrent Neural Networks

Traditional ANNs gain extended capabilities through Recurrent Neural Networks (RNNs) which implement loops in their network architecture to maintain information across time steps. The structural design of RNNs allows them to store hidden state information from previous inputs which makes them ideal for modeling sequential data including time series forecasting and speech recognition and natural language processing (Goodfellow, Bengio, & Courville, 2016). The standard ANN lacks the ability of RNNs to leverage historical context for making current predictions. Basic RNNs struggle with long-term dependencies because training introduces the vanishing gradient problem. The inability of RNNs to detect long-term relationships in sequences led to the creation of advanced network architectures including Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997).

#### LSTM

Long Short-Term Memory (LSTM) networks have revolutionized the analysis of sequential and non-linear data through their advanced time series forecasting abilities. The vanishing gradient problem which affects traditional RNNs can be solved by LSTM models which operate as recurrent neural networks. According to Hochreiter & Schmidhuber (1997) the memory cells together with the input, forget and output gates allow the model to store vital information across long time sequences by discarding unnecessary data. The architecture of LSTM provides better detection of extensive relationships in complex datasets including financial market data because it maintains tracking of historical patterns which influence future trends across extended periods. LSTM has become widely used for tasks such as stock price forecasting and demand and sales forecasting because of its ability to model complex data relationships (Fischer & Krauss, 2018; Nelson et al., 2017; Siami-Namini et al., 2018).

LSTMs come with certain performance constraints. The training process of LSTMs demands substantial computational power and resources while simultaneously being susceptible to overfitting when the training data includes noise or volatility according to Brownlee (2020). The model produces unrealistic predictions that differ from actual trends because of overfitting which results in hallucinations. LSTM faces difficulties with generalization during volatile market periods because sudden market shocks and irregularities disrupt learned patterns (Chimmula & Zhang, 2020).

### Prophet

Prophet, developed by Facebook, offers a practical and user-friendly approach to time series forecasting, especially for datasets with strong seasonal and trend components. It uses a decomposable model framework, splitting the time series into three main components: trend, seasonality, and holiday effects (Taylor & Letham, 2018). The trend component models long-term changes in the data, such as growth or decline, while the seasonal component accounts for recurring patterns, such as daily, weekly, or yearly fluctuations. The holiday effect captures irregular spikes or drops tied to specific events, such as holidays or promotional campaigns. Prophet is particularly appealing to non-experts due to its straightforward parameter tuning and interpretability. For instance, business analysts can easily specify known seasonalities or trend shifts, and the model will incorporate these adjustments seamlessly. However, while Prophet excels at handling predictable patterns, it may oversimplify datasets with irregular or complex behaviors. This limitation becomes apparent in financial forecasting or during volatile periods, where sudden, non-seasonal fluctuations challenge Prophet’s assumptions, leading to potential inaccuracies or unrealistic predictions (Taylor & Letham, 2018; Makridakis et al., 2018).

## Evaluation Metrics

The assessment of forecasting model effectiveness demands strict performance metrics which measure prediction accuracy against actual results. Time series forecasting demands thorough evaluation because it determines both model predictive strength and real-world market reliability. Multiple statistical tools exist to evaluate forecast accuracy because they provide unique insights about model performance. The measurement of forecast accuracy depends on error-based metrics to calculate prediction-value differences, but other methods analyze forecast error distribution and extremity. High-volatility financial environments require these metrics because small prediction errors can create major consequences (Makridakis et al., 2018; Hyndman & Koehler, 2006). The research uses traditional accuracy metrics alongside anomaly detection techniques to evaluate both average performance and extreme deviations known as hallucinations.

### Accuracy Metrics

Evaluating the accuracy of forecasting models is critical to determine their performance and reliability. Accuracy metrics quantify the difference between the model’s predictions and the actual observed values, providing insights into how well the model captures underlying data patterns. In this study, three widely used metrics—Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE)—are employed to ensure a comprehensive evaluation of model performance.

* Root Mean Squared Error (RMSE):

RMSE is a widely used metric that calculates the square root of the average squared differences between predicted and actual values. It is mathematically defined as:

where represents the predicted value, is the actual observed value, and is the total number of observations. RMSE is particularly sensitive to large errors, as squaring the residuals magnifies their impact. This makes it suitable for applications where minimizing significant deviations is critical, such as financial forecasting or demand planning. However, RMSE can be influenced by scale, meaning it may not be directly comparable across datasets with different ranges (Hyndman & Koehler, 2006).

* Mean Absolute Error (MAE):

MAE calculates the average of the absolute differences between predicted and actual values, without squaring the errors:  
MAE provides the average magnitude of prediction errors, offering an intuitive and scale-independent measure of accuracy. Unlike RMSE, it treats all errors equally, regardless of their size.

* Mean Absolute Percentage Error (MAPE):

MAPE expresses errors as a percentage of actual values, allowing for comparisons across datasets with different scales. It is defined as:

MAPE provides an intuitive, relative measure of accuracy, making it particularly useful in business applications where interpreting errors as a percentage is more meaningful. However, MAPE has limitations: it becomes unstable when actual values are near zero and can disproportionately penalize under-predictions compared to over-predictions (Kim & Kim, 2016).

By combining these three metrics, this study ensures a balanced evaluation of model performance, capturing both the magnitude and relative impact of forecast errors.

### Hallucination Detection Techniques

The evaluation of accuracy together with hallucination detection represents essential steps for determining the reliability of forecasting models. The evaluation of hallucinations aims to detect extreme or unrealistic predictions which helps determine forecasting model stability and reliability. The research uses two established methods to detect hallucinations in forecast outputs which include prediction interval checks and Z-score analysis. The statistical tools serve as common methods for anomaly detection and forecasting validation according to Dong et al. (2024)

* Prediction Interval Check:

These methods provide a range within which the model expects future values to fall, based on the level of uncertainty in its predictions. For example, a 95% prediction interval indicates that the model is 95% confident the actual value will fall within the specified range. If actual values frequently fall outside these intervals, it signals overconfidence or poor uncertainty estimation by the model. This method is particularly useful for models like Prophet and ARIMA, which naturally generate prediction intervals as part of their outputs (Taylor & Letham, 2018). By analyzing these intervals, researchers can detect hallucinations that occur when models fail to account for variability during volatile periods.

* Z-Score Analysis:

Z-score analysis evaluates how far a prediction deviates from the mean of historical data, measured in terms of standard deviations. The Z-score for a predicted value X is calculated as:

where is the historical mean and is the standard deviation. Predictions with absolute Z-scores greater than 3 are typically flagged as extreme anomalies. This method is effective for identifying hallucinations, especially in volatile periods, as it highlights outputs that significantly diverge from typical historical behavior. Recent research has directly linked Z-score-based anomaly detection to hallucination mitigation in time series models (Dong et al., 2024).

## 2.4 Applications of ARIMA, LSTM, and Prophet Models Beyond Financial Forecasting

Time series forecasting models including ARIMA, LSTM and Prophet have been widely used across various fields beyond financial markets, demonstrating their versatility and importance in solving real-world problems. Each model has its own strengths and weaknesses, which make them suitable for different types of data and forecasting tasks, and can be applied in healthcare, logistics, energy, retail, and environmental monitoring.

ARIMA models demonstrate their effectiveness by analyzing stationary and linear data in various domains. They serve healthcare by predicting infectious disease spread through their analysis of historical case count patterns (Chien, Yu, & Schootman, 2010). The application in environmental studies enables researchers to predict air pollution levels which provides essential information for policy development (Ghanem & Zhang, 2014). ARIMA models have proven effective in urban water demand forecasting which enables cities to develop plans for resource distribution and infrastructure development (Adamowski, 2008). They apply their forecasting capabilities to agricultural settings where they predict crop yields using historical planting data and weather patterns.

The advanced version of traditional recurrent neural networks known as LSTM networks demonstrates exceptional performance in processing complex non-linear sequential data with long-term dependencies. The application in healthcare includes predicting patient admissions and emergency department demand and disease progression through analysis of electronic health records (Shickel et al., 2018). LSTM models serve transportation systems by predicting real-time traffic flow which helps cities manage congestion and develop smart urban planning strategies (Ma et al., 2015). The energy sector uses LSTM extensively for predicting irregular patterns in wind and solar power outputs because these sources exhibit volatile behavior (Marino, Amarasinghe, & Manic, 2016). The logistics industry uses a forecasting system that combines ARIMA with LSTM models to predict cargo volume by utilizing ARIMA for linear trend detection and LSTM for non-linear residual modeling (Liu, Zhang, & Wang, 2022).

The Prophet model developed by Facebook has become popular because of its interpretability and ease of use especially for business practitioners. The model has applications that go beyond finance to other fields that need basic automated forecasting pipelines. Prophet has been used in retail to forecast demand spikes and sales volume and to accommodate holiday effects and seasonality (Taylor & Letham, 2018). Prophet has been used in the tourism sector to forecast tourist arrival numbers and to capture seasonal patterns and external event-driven influences (Antipov & Pokryshevskaya, 2020). Moreover, during the COVID-19 pandemic, Prophet was used to forecast case numbers, which shows its usefulness in epidemiological modeling where flexible handling of trend shifts and missing data was critical (Ribeiro et al., 2020).

The wide adoption of ARIMA, LSTM, and Prophet in healthcare, logistics, energy, retail, and environmental forecasting shows their versatility and robustness. By showing their effectiveness outside of finance, these models are shown to be useful tools in solving complex forecasting problems in dynamic and uncertain environments. Their application diversity further increases their relevance in comparative forecasting research, including the present study.

# Methodology

The aim of this study is to compare the performance of three time series forecasting models—ARIMA, LSTM, and Prophet—on financial datasets characterized by varying levels of volatility. The study evaluates both forecasting accuracy and the propensity of the models to generate hallucinations. By focusing on datasets with stable and volatile price behaviors, the research highlights each model’s suitability for specific market conditions.

Volatility is a measure of price variation over time and is a critical factor in financial forecasting. Stable price movements, typical of broad market indices like the S&P 500, allow models to leverage predictable patterns. Conversely, volatile stocks, such as Tesla, challenge forecasting models with abrupt fluctuations and irregular patterns. Understanding how models perform under these contrasting conditions is crucial for selecting the right forecasting approach for specific financial applications. This study captures these dynamics by comparing models across datasets with varying volatility levels.

## Dataset Selection

The choice of datasets is a critical component of this study, as it directly impacts the evaluation of forecasting models. To assess the performance of ARIMA, LSTM, and Prophet models, two distinct financial datasets representing contrasting volatility profiles are used: the S&P 500 Index (stable) and Tesla, Inc. (volatile). This selection is designed to simulate real-world financial forecasting scenarios, where data stability and volatility vary significantly across different assets and market conditions. By comparing model performance on these datasets, this study provides insights into how well these models generalize under varying market dynamics.

### S&P 500 Index

The S&P 500 Index, maintained by S&P Global, is one of the most prominent financial indices globally. It tracks the performance of 500 publicly traded large-cap companies listed in the United States. As a benchmark for the overall health of the U.S. economy and equity markets, the index is widely regarded as a leading indicator for investment performance and economic trends. The index covers approximately 80% of the total market capitalization of the U.S. stock market, making it a representative measure of the broader market (Standard & Poor’s, 2023).

While generally considered stable, the S&P 500 is not immune to periods of volatility. Events like the 2008 Global Financial Crisis, the 2020 COVID-19 pandemic, and inflationary pressures in recent years have led to significant fluctuations in the index’s value. However, its long-term trend has historically been upward, driven by the consistent growth of the U.S. economy and innovation within its constituent companies (Bodie et al., 2021).

For this study, the S&P 500 Index was chosen due to its relatively stable price movements, making it ideal for testing forecasting models under low-volatility conditions. Its seasonality and clear trends, influenced by economic cycles and corporate performance, provide an excellent dataset for assessing the baseline performance of ARIMA, LSTM, and Prophet models. Furthermore, its stability reduces the likelihood of erratic predictions, allowing for focused evaluation of model accuracy.

### Tesla, Inc. Stock

Tesla, Inc. (TSLA) is a global leader in the electric vehicle (EV) and renewable energy sectors, recognized for its innovative approach to technology and sustainability. Founded in 2003 and led by CEO Elon Musk, Tesla has transformed the automotive industry, becoming a dominant player in the global EV market. In addition to its vehicles, the company’s focus on clean energy solutions, including solar panels and energy storage systems, has further solidified its position as a pioneer in sustainable technology. Tesla’s market capitalization has grown substantially in recent years, often surpassing that of traditional automakers, reflecting its perceived potential for long-term growth and innovation (Fischer & Krauss, 2018).

Tesla’s stock is widely known for its high volatility, making it one of the most dynamic equities in the global financial market. While its long-term trajectory has shown remarkable growth, short-term price movements are often unpredictable. External factors such as regulatory developments, CEO statements, production milestones, and macroeconomic trends significantly impact its valuation. Events like product announcements, quarterly earnings releases, and global shifts in demand for EVs often trigger sharp fluctuations in its stock price (Zhang et al., 2021).

This volatility is amplified by Tesla’s visibility in the media and its popularity among retail and institutional investors. For instance, public statements by Elon Musk on social media have led to abrupt price changes, further emphasizing the sensitivity of Tesla’s stock to market sentiment. Despite these challenges, the stock remains a favorite among traders and long-term investors, highlighting its appeal as a high-risk, high-reward asset (Chen et al., 2022).

## Data Collection

The accuracy and reliability of forecasting models heavily depend on the quality and relevance of the input data. To ensure robust and meaningful comparisons of the ARIMA, LSTM, and Prophet models, this study carefully selected two distinct financial datasets that capture contrasting market behaviors: the S&P 500 Index and Tesla, Inc. (TSLA) stock prices. These datasets were sourced from Yahoo Finance, a widely recognized and reputable platform for financial market data. Yahoo Finance provides comprehensive, historical financial data, including daily price movements, volume, and other market indicators, making it highly suitable for academic research and financial analysis.

Yahoo Finance was chosen as the primary data source for several reasons. First, it offers free and convenient access to a wide range of financial data, including equities, indices, and commodities. Second, its data is regularly updated and widely used in both academic research and industry analyses, ensuring reliability and relevance. Additionally, Yahoo Finance allows the retrieval of historical data over extensive periods, which is critical for performing time series analysis across different market conditions. Its consistent data format also simplifies the data extraction and preprocessing steps, making it a practical choice for this research.

## Timeframe Selection

The selection of an appropriate timeframe is a crucial element in evaluating the performance of time series forecasting models, particularly when dealing with financial data characterized by varying levels of volatility. A well-chosen timeframe must encompass both stable and volatile market conditions to allow for a comprehensive assessment of each model’s forecasting accuracy and its susceptibility to generating hallucinations, or extreme and unrealistic predictions.

This study utilizes historical financial data spanning from January 1, 2015, to December 31, 2024. This eight-year period was intentionally selected to provide diverse market conditions for rigorous model testing. During this timeframe, global markets experienced a wide range of economic events that contributed to both stability and volatility. For instance, the 2015–2016 market slowdown, primarily driven by declining oil prices and a decelerating Chinese economy, introduced moderate market volatility (Makridakis et al., 2018). This was followed by a relatively stable period of global economic growth between 2017 and 2019. However, this stability was disrupted by the COVID-19 pandemic in 2020, which led to an unprecedented market crash followed by a rapid recovery fueled by extensive fiscal stimulus measures and market optimism (Qin et al., 2017). Recently, the year 2022 was marked by significant volatility due to rising inflation and aggressive interest rate hikes implemented by central banks to combat economic instability (Fischer & Krauss, 2018).

The selection of this extended timeframe ensures that the forecasting models are exposed to a wide array of market dynamics, enabling them to learn and adapt to both predictable trends and unexpected disruptions. Moreover, this period provides a sufficiently large dataset, which is critical for training deep learning models such as the Long Short-Term Memory (LSTM) network. LSTM models require extensive data to effectively learn complex temporal dependencies and avoid overfitting (Fischer & Krauss, 2018). Additionally, this timeframe captures long-term market trends and recurring seasonal patterns, which is particularly advantageous for models like Prophet that are specifically designed to incorporate trend and seasonality components into their forecasts (Taylor & Letham, 2018).

## Volatility Segmentation

To thoroughly evaluate the forecasting models under different market dynamics, the dataset is segmented into stable and volatile periods. This segmentation allows for a more focused analysis of how each model responds to various market conditions. Stable periods are defined by consistent market growth and low price fluctuations, exemplified by the economic expansion from 2017 to 2019 (Makridakis et al., 2018). In contrast, volatile periods are identified using the rolling standard deviation of daily returns, a common method for detecting fluctuations in market volatility (Qin et al., 2017). Notably, the market crash between February and March 2020 due to the COVID-19 pandemic and the market correction in late 2018 serve as prominent examples of high-volatility periods.

This segmentation strategy ensures that the models are not only evaluated on their overall forecasting accuracy but also on their ability to remain stable and avoid generating hallucinations during periods of extreme market movement.

## Selection of Data Frequency

In time series forecasting, selecting the appropriate data frequency is a critical factor in determining the effectiveness of predictive models. The frequency of data collection influences how well forecasting models capture market dynamics, adapt to rapid changes, and generate reliable predictions. Since this study relies on financial data from Yahoo Finance, it is essential to explore the available data frequency options and provide a clear justification for selecting daily data for analysis.

Yahoo Finance offers several data frequency options that cater to different analytical needs and forecasting goals.

Intraday data (1-minute to 60-minute intervals) captures price movements within a single trading day at frequent intervals. This high-frequency data is suitable for high-frequency trading (HFT) and short-term market analysis where identifying micro-level price changes is crucial. However, the use of intraday data comes with significant computational complexity and the risk of overfitting due to noise in the data (Makridakis et al., 2018).

Daily frequency data provides the open, high, low, close (OHLC) prices and trading volume for each trading day. It offers a well-balanced view of market trends and volatility while maintaining a manageable data size, making it a widely used standard for financial forecasting (Fischer & Krauss, 2018).

Weekly data consolidates daily price movements into a single data point per week. This frequency is advantageous for detecting long-term trends but may smooth out important short-term fluctuations, limiting its effectiveness in capturing rapid market changes.

Monthly data reflects long-term market behavior and is often used for macroeconomic analysis and strategic financial planning. However, its coarse granularity makes it unsuitable for capturing timely market shifts necessary for short-term and medium-term forecasting.

The decision to use daily data frequency for analyzing the S&P 500 Index and Tesla, Inc. (TSLA) stock is grounded in its ability to provide an optimal balance between capturing market dynamics and ensuring computational efficiency.

Daily data captures the fast-paced and often unpredictable nature of financial markets. Financial assets are highly sensitive to daily developments such as earnings reports, economic policies, and geopolitical events. These factors can cause sudden price movements that models must be able to detect and respond to effectively. By using daily data, the models are better equipped to reflect these short-term market behaviors and provide timely predictions that are relevant for decision-making (Fischer & Krauss, 2018).

Additionally, daily data strikes a practical balance between detail and computational feasibility. While intraday data offers a finer resolution of market activities, it introduces massive data volumes that can overwhelm forecasting models—especially deep learning models like Long Short-Term Memory (LSTM) networks, which are already computationally intensive. Daily data, on the other hand, contains enough information to capture meaningful market patterns without burdening the models with unnecessary noise or excessive processing requirements (Makridakis et al., 2018).

Another advantage of daily data is its flexibility in supporting different forecasting horizons. Financial models must often accommodate both short-term and medium-term forecasts depending on the user’s objectives. Daily data allows the models to provide near-term forecasts while still offering insight into broader market trends. This flexibility makes daily data particularly valuable for a wide range of financial forecasting applications (Taylor & Letham, 2018).

Moreover, the use of daily data aligns with established practices in both academic research and the financial industry. It is the standard frequency for financial time series forecasting, making the results of this study comparable with existing literature and directly applicable to real-world financial analysis and decision-making (Makridakis et al., 2018).

By selecting daily data, this study ensures that the forecasting models can effectively capture meaningful market movements and trends. This choice supports the study’s objective of evaluating model performance under both stable and volatile conditions while maintaining computational efficiency and providing actionable insights.

## Training and Testing Split

To objectively assess the forecasting performance of the ARIMA, LSTM, and Prophet models, the dataset is divided into two distinct parts: a training set and a testing set. This division is crucial for evaluating how well each model can learn from historical data and accurately predict future market trends under both stable and volatile conditions.

The training set includes data from January 1, 2015, to December 31, 2022. This extended period provides the models with comprehensive exposure to a wide range of market conditions. These conditions include the 2015–2016 global market slowdown, the period of steady economic growth from 2017 to 2019, the COVID-19 market crash and recovery in 2020, and the inflation-driven volatility of 2022. By training on such diverse market scenarios, the models are equipped to recognize different financial data patterns, including long-term trends, seasonality, and sudden market disruptions. This variety in the training data allows the models to develop a more robust understanding of market behavior and enhances their adaptability to real-world dynamics.

The testing set spans from January 1, 2023, to December 31, 2024, serving as an out-of-sample dataset to evaluate the models’ ability to generalize and make accurate predictions on unseen data. This period includes evolving economic conditions such as the ongoing economic recovery, monetary policy adjustments, and potential geopolitical risks. Evaluating model performance on this recent data allows for a realistic assessment of each model’s capacity to handle current and emerging market trends. This clear distinction between the training and testing datasets aligns with best practices in time series forecasting, ensuring that models are not overfitted to historical data, can adapt effectively to future market changes, and minimize the risk of generating hallucinations in unpredictable conditions.

By structuring the data into these two distinct periods, the study rigorously measures each model’s forecasting accuracy, stability, and reliability across different market environments. This comprehensive evaluation provides valuable insights into how effectively ARIMA, LSTM, and Prophet models can support decision-making in practical financial forecasting applications.

## Evaluation and Analysis of Model Performance

The evaluation of the forecasting models in this study is designed to provide a comprehensive assessment of both forecasting accuracy and the models’ susceptibility to generating hallucinations, particularly under varying levels of market volatility. The analysis focuses on how the ARIMA, LSTM, and Prophet models perform when exposed to both stable and volatile market conditions, providing insights into their reliability and practical application in financial forecasting.

To evaluate the accuracy of each model, the study employs three well-established error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics are selected for their complementary ability to quantify forecasting errors and offer a robust assessment of model performance. RMSE is used to measure the average magnitude of forecasting errors with a particular sensitivity to larger errors, making it suitable for financial data where significant deviations can lead to major financial consequences. MAE provides a straightforward measure of the average absolute difference between predicted and actual values, offering an intuitive understanding of model accuracy across various conditions. MAPE, by expressing prediction errors as a percentage of actual values, allows for a normalized comparison of model performance across different datasets and market scales.

However, measuring accuracy alone does not fully capture the models’ reliability in financial forecasting. This study places significant emphasis on identifying and analyzing hallucinations, defined as extreme or unrealistic predictions that deviate substantially from actual market behavior. Hallucinations are particularly concerning in high-stakes financial applications, as they can lead to misguided decision-making and financial losses.

To detect and evaluate hallucinations, two analytical methods are employed: Prediction Interval Analysis and Z-Score Analysis.

This study assesses how frequently actual values fall outside the models’ forecasted confidence intervals using Prediction Interval Analysis. Models like ARIMA and Prophet naturally generate prediction intervals that reflect the expected range of outcomes. When actual market values frequently breach these intervals, it suggests that the model may be overconfident or failing to account for market uncertainty, increasing the risk of generating hallucinations. By analyzing these deviations, the study can determine how well each model captures the inherent volatility of financial markets and manages uncertainty in its forecasts.

Z-Score Analysis standardizes predictions to detect extreme deviations from historical trends, identifying outliers that diverge from typical market movements. Predictions with exceptionally high or low Z-scores are flagged as hallucinations, indicating that the model has produced forecasts that are not grounded in realistic market behavior. This analysis is particularly critical for evaluating models like LSTM, which, due to their complexity and sensitivity to data, may generate exaggerated predictions in response to market volatility.

During stable periods, models are expected to deliver consistent and accurate forecasts that align with observable trends and seasonal patterns. Conversely, during volatile periods, the models’ capacity to adjust to rapid and unpredictable market shifts will be critically analyzed, with a focus on identifying hallucinations that could compromise their practical utility.

By integrating traditional accuracy metrics with targeted hallucination detection methods, this evaluation framework ensures a comprehensive assessment of the ARIMA, LSTM, and Prophet models. This dual analysis approach allows for an in-depth comparison of how each model balances forecasting accuracy with the risk of generating extreme or unrealistic predictions. The findings from this evaluation will provide valuable insights into the trade-offs between model complexity, accuracy, and reliability, guiding financial analysts and decision-makers in selecting forecasting models that are both accurate and resilient under real-world market conditions.

# Appendix A: Appendix Title

Text

# Appendix B: Appendix Title

Text

**Declaration of Academic Integrity and Originality**

I hereby confirm that I have written this work, entitled

„*Thesis/Paper Title*“

by myself, without any external support, and without the use of any sources other than those cited in the text. All intellectual property taken directly or indirectly from external sources is distinctly marked as such.

The work has not been submitted to another examination office in this or in similar form and has not been published.

I agree that the work will be checked for plagiarism with the help of a software service.

Dresden, DD.MM.YYYY

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

First and last name Signature