Faculty of Media Engineering and Technology
German University in Cairo

Big-Data Analytics Models in Decision Making Process

Bachelor Thesis

by

Ahmed Sameh Sakr

Supervised by

Prof. Dr. Gamal Abdel Shafy

Date

29 May 2025

Declaration

This is to certify that:

(i) the thesis comprises only my original work towards the Bachelor Degree

(ii) due acknowledgment has been made in the text to all other material used

Ahmed Sameh Sakr

29 May 2025

Acknowledgment

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Gamal Abdel Shafy, for his invaluable support, insightful guidance, and continuous encouragement throughout my research paper. His expertise and feedback were fundamental to shaping the direction of this research and helping me overcome the various challenges encountered along the way. I am deeply appreciative of the time and effort he dedicated to mentoring me, and for always providing thoughtful advice and critical perspectives that enriched both the technical and theoretical aspects of this work. This thesis would not have been possible without his mentorship. Finally, I would also like to thank everyone who contributed directly or indirectly to the successful completion of this project, including my professors, colleagues, and family members, for their encouragement and support.

Contents

List of Figures

List of Tables

Abstract

With the current high-speed financial markets, accurate forecasting and sound decision-making are critical to minimizing risk and maximizing return. This research proposes a hybrid method that employs Long Short-Term Memory (LSTM) neural networks combined with Integer Linear Programming (ILP) in an effort to support automated trading decisions. The LSTM model forecasts short-term stock price movements in a x-day horizon, and the ILP module allocates capital effectively across selected assets under a budget constraint. The motivation stems from the need for a data-driven decision support system that not only forecasts market motion but also provides actionable and interpretable portfolio recommendations.

Despite advances in deep learning-based time series forecasting, most prior research stops short at prediction and doesn't tie that into a capital allocation strategy. Also, the bulk of past approaches presume absolute price prediction that's prone to long-term market drift and macroeconomic noise. Comparatively few systems pay attention to actual-world restrictions such as budget restrictions or integer-based investment selection, resulting in a disconnect between model predictions and actual trading requirements. This paper fills these voids by predicting price deltas, detecting profitable buy/sell points, and using ILP to infer best investment plans.

Experimental results demonstrate that the model attains a Mean Absolute Error (MAE) of 0.54, a Root Mean Squared Error (RMSE) of 0.85, and an R^2 score of 0.67 on the training set. The ILP module effectively identifies top trade pairs subject to budget constraints, presenting investment plans that are pragmatic and actionable. Compared to traditional regression-based or greedy allocation methods, this method offers better interpretability by directly connecting predictions to real investment decisions. Also, the modularity of the suggested system allows for scalability on numerous stocks and future integration with real-time feeds, and it is therefore suitable both for research use and deployment in financial applications.

Chapter 1

Introduction

Financial markets are inherently complex and subject to a broad spectrum of dynamic and often unpredictable influences. Accurate forecasting of the change in stock prices has long been a desire of investors, traders, and researchers. Within the past years, deep learning has brought significant advances to time series forecasting, and mathematical optimization techniques have improved portfolio decisions under real-world constraints.

This research proposes a hybrid model that integrates Long Short-Term Memory (LSTM) neural networks for forecasting price movements and Integer Linear Programming (ILP) for distributing maximum capital. The goal is to provide an intelligent and understandable model capable of predicting stock directions and recommending investment actions that yield the highest profit for a specified budget.

1.1 Problem Statement

Even with growing uses of machine learning for financial forecasting, many existing models either impose limitations on short-term or uninterpretable actions in real-life trading scenarios. Moreover, few methods actually incorporate actual-world investment constraints such as size of the budget or integer levels of shares into the decision. This creates a gap between theoretical forecast models and practical trading methods.

How do we design a system that not only gives correct predictions regarding stock price movements but also does investment optimization to recommend when and how much one should invest in specific assets in the event of budget limitations?

1.2 Project Motivation

The motivation behind this work is two-fold. Firstly, the field of finance offers a high-stakes area of application for artificial intelligence, in which improved prediction can translate to financial reward in the real world. Secondly, combining deep learning and operations research methods offers an opportunity to bridge data-based prediction and prescriptive investment strategies.

Traditional trading systems usually rely on either technical or fundamental analysis. This project, on the other hand, explores a modern approach that integrates temporal sequence learning and profit-maximizing allocation logic and offers a wiser and more adaptable alternative to retail and institutional investors alike.

1.3 Project Objectives

This project aims to develop a strong trading decision-support system on the basis of LSTM neural networks for prediction and ILP for investment decision-making that is optimized. The system will be validated on real stock market data, and its performance will be compared against regular regression measures.

The objectives of this project are:

1. To collect and preprocess historical stock data with added technical indicators.
2. To build a deep learning model based on LSTM for predicting future price deltas.
3. To design a forecasting strategy for simulating multi-day market patterns.
4. To introduce an ILP optimizer that picks the best trades and quantities of shares under budget constraints.
5. To measure the system through conventional metrics like MAE, RMSE, and R^2.
6. To provide a clear output with foreseen trades, profits, and investment distribution.

1.4 Thesis Organization

This thesis is organized as follows:

- Chapter 1: Introduction provides an overview of the project, including the problem statement, motivation, and objectives.
- Chapter 2: Literature Review describes the key concepts of big data analysis and decision-making processes, in particular for how they apply to intelligent financial systems.
- Chapter 3: System Design and Implementation describes the data pipeline, model architecture, feature engineering, ILP formulation, and the integration of all components.
- Chapter 4: Results and Evaluation presents the model performance with varied metrics and demonstrates the ILP-based investment outcomes using examples and visualizations.
- Chapter 5: Conclusion and Future Work summarizes the contributions of the project and suggests avenues for improving the system and extending its scope.

Chapter 2

Background

Organizations today generate a vast amount of data from sources like social media, Internet of Things (IoT) devices, and business transactions. This explosion of Big Data presents both challenges and opportunities for decision-makers. Extracting meaningful insights from such complex datasets is crucial for informed decision-making.

Big Data Analytics (BDA) employs techniques like machine learning, data mining, and statistical analysis to process large datasets. Its integration enhances efficiency, optimizes performance, and provides competitive advantages. However, challenges such as data quality, scalability, and infrastructure requirements must be addressed.

Traditional decision-making, based on intuition and historical data, is becoming less effective. Real-time analytics and predictive modeling offer a more data-driven approach, making BDA a key research area for bridging the gap between raw data and actionable insights.

BDA is widely applied in healthcare, finance, marketing, supply chain management, and cybersecurity. This thesis explores its impact on transforming decision-making across industries.

2.1 The Five Key Dimensions of Big Data

Big Data is defined by five key characteristics, often referred to as the 5 V's: Volume, Velocity, Variety, Veracity, and Value. These dimensions describe the challenges and opportunities associated with managing and analyzing large-scale datasets. Understanding these aspects is essential for leveraging Big Data effectively in decision-making processes [1]. Below is a detailed discussion of each dimension with Figure 2.1 that shows the FiveVS:

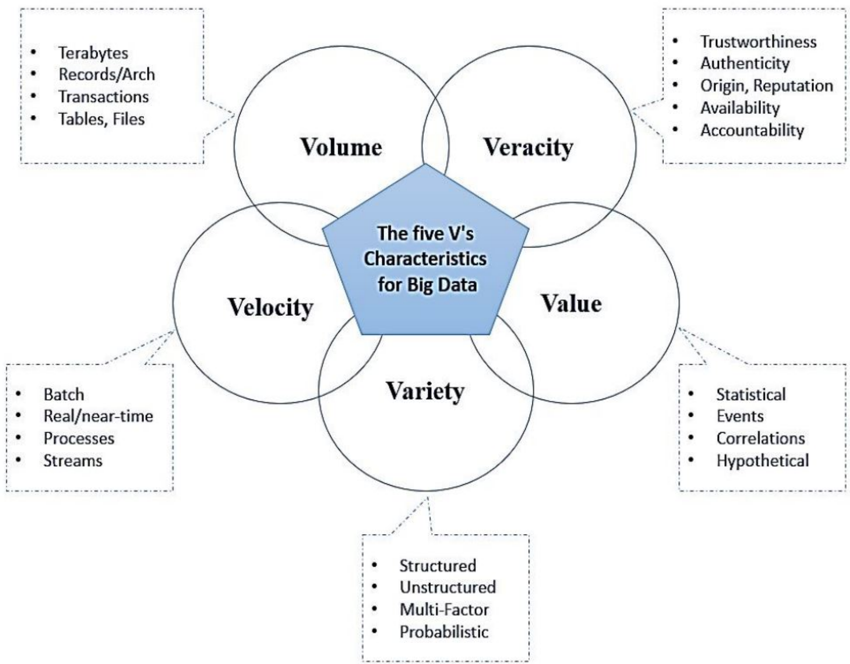

Figure 2.1: FiveVsBig Data Characteristics [1].

2.1.1 Volume

Volume refers to the sheer scale of data generated every second from various sources, including social media, business transactions, and IoT devices. The ability to store, process, and analyze massive datasets is a fundamental challenge in Big Data Analytics. Figure 2.2 below shows the growth of data every year since 2008 in zettabytes.

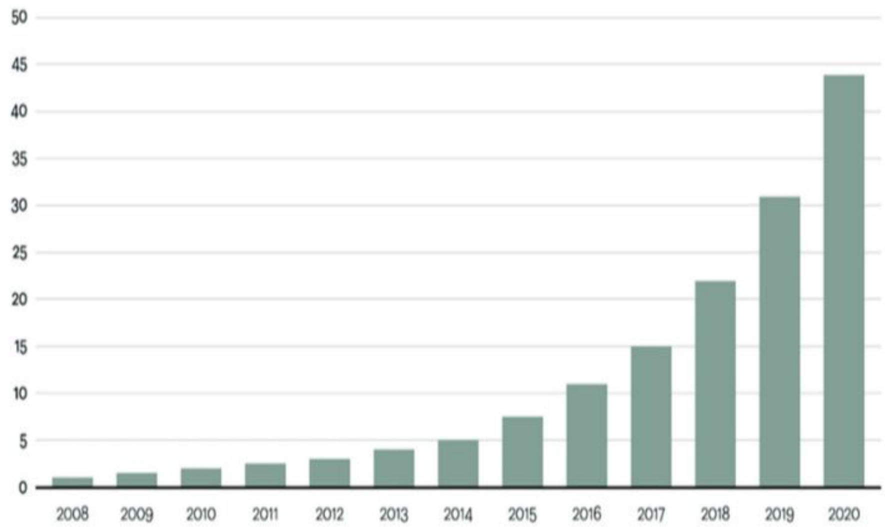

Figure 2.2: Growth of Big data [2].

2.1.2 Velocity

Velocity describes the speed at which data is generated, collected, and processed. Real-time or near real-time processing is crucial for applications such as fraud detection, stock market analysis, and personalized recommendations.

2.1.3 Variety

Variety represents the diverse formats and types of data, including structured, semi-structured, and unstructured data. Organizations must manage data from different sources such as text, images, videos, and sensor readings to gain comprehensive insights.

2.1.4 Veracity

Veracity refers to the quality and reliability of data. Inaccurate or inconsistent data can lead to misleading insights, making data cleansing and validation essential for effective analytics.

2.1.5 Value

Value emphasizes the importance of extracting meaningful insights from data. Simply collecting data is not enough; organizations must analyze it effectively to drive business strategies, optimize processes, and improve decision-making.

2.2 Big Data Applications in Various Fields

Big Data is applied in a variety of fields to aid decision-making and enable organizations to optimize their operations. Some of the key fields where Big Data is widely utilized are:

2.2.1 Healthcare

Big Data in healthcare plays a critical role in patient diagnosis, treatment optimization, and disease outbreak predictive analytics. Through the application of machine learning and advanced analytics, it allows for data-informed decisions among medical professionals to improve patient outcomes and resource optimization. For example, predictive modeling has been used to forecast mortality risk in patients with subarachnoid hemorrhage, showing the possibility of Big Data in optimizing precision medicine and timely disease diagnosis [3]. By analyzing big data sets from electronic health records and imaging, clinicians are able to improve clinical outcomes and advance the limits of personalized treatment strategies.These are different healthcare applications Figure 2.3.

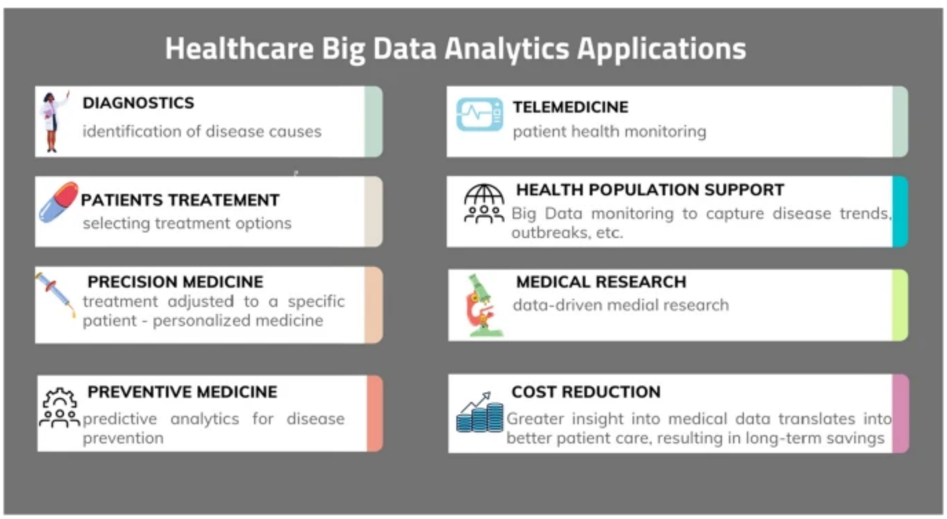

Figure 2.3: Healthcare Big Data Analytics applications [3].

2.2.2 Finance

Big Data in the financial sector helps in fraud detection, risk management, and algorithmic trading. Real-time analytics is used by financial institutions to monitor transactions and stop fraudulent transactions. Data-driven finance improves investment decisions, credit scoring, and customer satisfaction with personalized services [4].

2.2.3 Marketing

Big Data is utilized by marketing professionals to analyze consumer behavior, targeted marketing, and customer segmentation. Companies search through vast data from social media, web use, and purchases to optimize marketing campaigns. Data-driven insights enhance marketing decision-making by maximizing customer interaction and return on investment [5].

2.2.4 Supply Chain Management

Big Data is transforming supply chain management by way of logistics optimization, inventory management, and demand forecasting. Real-time tracking of shipments and predictive analytics allow firms to reduce costs and increase efficiency. Decision-making is data-driven here, and that means better resource allocation and supplier relationship management [6].

2.2.5 Cybersecurity

Big Data analytics plays a critical role in discovering and preventing cyber attacks. Machine learning and predictive analytics are employed by organizations to identify anomalies in network traffic and avoid security incidents. Data-driven decision-making improves cybersecurity initiatives by guaranteeing threat detection, as well as response, in advance [7].

2.2.6 Smart Cities

Big Data is central to smart city building, enhancing urban planning, traffic control, and environmental monitoring. Real-time analytics are utilized by cities to streamline transportation systems, minimize pollution, and improve public services. Smart city decision-making is based on data insights for enhancing the quality of life for residents [8].

2.2.7 Education

For the education sector, Big Data helps in personalized learning, student performance indicators, and enhancing curriculum. Student engagement and learning patterns are examined by institutions to personalize educational content. Evidence-based decision-making enhances teaching methodologies and institutional functions to provide higher academic outcomes [9].

2.3 Big Data Analysis Process

Data Analysis Process involves a series of steps that transform raw data and make it interpretable to inform decisions. It always includes defining the why, data collection, data processing, data analysis, and data visualization. All the stages are crucial to ensure that proper control of high-volume data is maintained so that it provides meaningful insights [10]. Steps of data analysis Figure 2.4.

Figure 2.4: Steps of data analysis [10]

2.3.1 Define the WHY?

The "why" is determined through an understanding of the domain, data available, and problem to be solved. It is reached through analysis of stakeholder needs, identification of key goals, and ensuring that the data meets the desired results. A clear "why" gives well-targeted, relevant, and actionable insights.

2.3.2 Data Collection

Data collection involves the act of acquiring data from various sources to address a research issue, test hypotheses, and measure effects. Data collection is divided into primary and secondary data collection processes. Secondary data refers to information that has previously been published in media like books, newspapers, journals, and websites. Considering that large quantities of data are available in every field of study, the utilization of appropriate criteria to select adequate secondary data becomes crucial in an attempt to render research valid and reliable [10].

Primary data collection is divided into quantitative and qualitative methods. Quantitative methods rely on mathematical calculations and are exemplified by instruments like closed-ended questionnaires, correlation and regression analysis, and statistical measures like mean, mode, and median. These methods are cost-effective, time-efficient, and highly structured, which makes it easier to compare results. Qualitative methods focus on non-numerical material like words, sounds, feelings, and visual components, with greater focus on deeper insights instead of numerical analysis [11]. Methods for data collection Figure 2.5.

Figure 2.5: Methods of Data Collection [11]

2.3.3 Data Cleaning

Data cleaning is a crucial data preprocessing method that involves the detection and correction of errors, inconsistencies, and inaccuracies in datasets to improve data quality and reliability. Data cleaning makes data accurate, complete, and consistent, which is essential for effective analysis and decision-making. Poor-quality data has the potential to bring about misleading conclusions, inefficiencies, and costly mistakes, especially in data-intensive applications such as machine learning and business intelligence.

The data cleaning process typically includes detecting and removing duplicate records, handling missing values, correcting structural defects, eliminating outliers, standardizing formats, and validating data accuracy. By performing these measures, data cleaning improves the overall quality of datasets, leading to more precise results and better performance in analytical models [10].

Apache Spark in Data Cleaning

In the context of big data, Apache Spark is a widely used open-source framework with built-in data cleansing capabilities. It provides a powerful and scalable platform for handling massive datasets efficiently through its Spark DataFrame API. Spark enables automated and distributed data cleaning with native functionalities for data filtering, transformation, and aggregation, making it an ideal solution for businesses dealing with high-volume data. Its integration with various storage systems and support for multiple programming languages, such as Python (PySpark) and Scala, further enhances its flexibility in data preprocessing tasks.

2.3.4 Data Analysis

Data analysis refers to the process of investigating, modifying, and modeling data in order to determine significant insights and support decision-making. It forms a significant component of research, business analytics, and data-driven processes such as machine learning and artificial intelligence. Raw data is converted into structured information that reveals patterns, trends, and relationships by employing analytical processes. Effective data analysis enhances accuracy, efficiency, and decision-making by ensuring conclusions derived from good and appropriately processed data [10].

Types of Data Analysis

Data analysis has been classified into several types based on the purpose and approach used. The four broad types of data analysis are [12]:

- Descriptive Analysis: It is all about condensing and structuring data so that the image that is formed is what has occurred. It involves the calculation of such measures as mean, median, standard deviation, and percentages in order to achieve trends and distributions. Aids to visualization, such as histograms, bar charts, and pie charts, are extensively used under this type of analysis.

- Diagnostic Analysis: It extends the description of the data that seeks to determine the causes of past events. It aims to discover patterns, relationships, and correlations between variables in order to account for why certain results have been observed. Regression analysis and hypothesis testing are common methods of conducting diagnostic analysis.
- Predictive Analysis: It utilizes historical facts and statistical models to forecast future occurrences. It relies on machine learning algorithms, time series analysis, and predictive modeling techniques to forecast probabilities and trends. Predictive analytics is used by businesses in demand forecasting, risk analysis, and forecasting customer behavior.
- Prescriptive Analysis: It takes predictive analysis a step further by recommending what to do in order to achieve maximum results. Prescriptive analysis draws upon advanced techniques such as optimization algorithms, decision trees, and artificial intelligence to provide recommendations on the best course of action. Prescriptive analysis is widely used in healthcare, finance, and supply chain management.

The Role of Data Analysis in Big Data

Big data environment requires sophisticated tools and frameworks that are capable of processing large volumes of structured as well as unstructured data. Processing and data analysis of humongous datasets rely heavily on tools such as Apache Spark, Hadoop, and SQL-based data warehouses. Utilization of these tools combined with machine learning and AI enhances the chances of obtaining meaningful insights and better data-driven decisions.

2.3.5 Interpreting Results

Interpretation of results is an essential step in the data analysis process, whereby raw results are translated into useful conclusions. After many analysis techniques have been utilized, one should ensure that the results are not only statistically relevant but also contextually useful for the problem at hand. In this stage, patterns, trends, and correlations in the data are determined to be significant and interpreted based on their implication in the real world.

One of the key aspects of result interpretation is differentiating causation from correlation. Two variables can appear related but are not necessarily causing one another. Analysts should be on the lookout for external influences, biases, and inaccuracies that can skew the findings. The findings must also be cross-checked with existing knowledge, business best practices, or earlier research to verify their validity and utility.

Visualization plays a central role in this phase, where graphs, charts, and dashboards are employed to convey insights more effectively to stakeholders. Interpretation correctly requires not just technical capabilities but also domain knowledge so that conclusions are correct and actionable [10].

By making careful interpretation feasible, analysts can extract important insights that result in actionable conclusions. The true worth of data lies in how insights created guide strategic decisions. We examine how organizations utilize analyzed data to inform data-driven decisions impacting business outcomes and policy decisions in the next section.

2.4 Decision Making Process

Decision-making comes after the Big Data analysis phase, when data-driven insights are converted to action. After cleaning, analyzing, and interpreting data, organizations must use these insights to drive their decisions in a maximally efficient and minimally risky way. Proper decision-making involves balancing the results of data analysis, exploring different possible courses of action, and selecting the most data-informed, evidence-based choice.

This process is essential in any field where decisions must be made with real-world data rather than on intuition alone. With predictive models, trend analysis, and statistical analysis, organizations can predict challenges, identify opportunities, and trigger strategies that produce optimum outcomes. Here is the whole decision-making process Figure 2.6

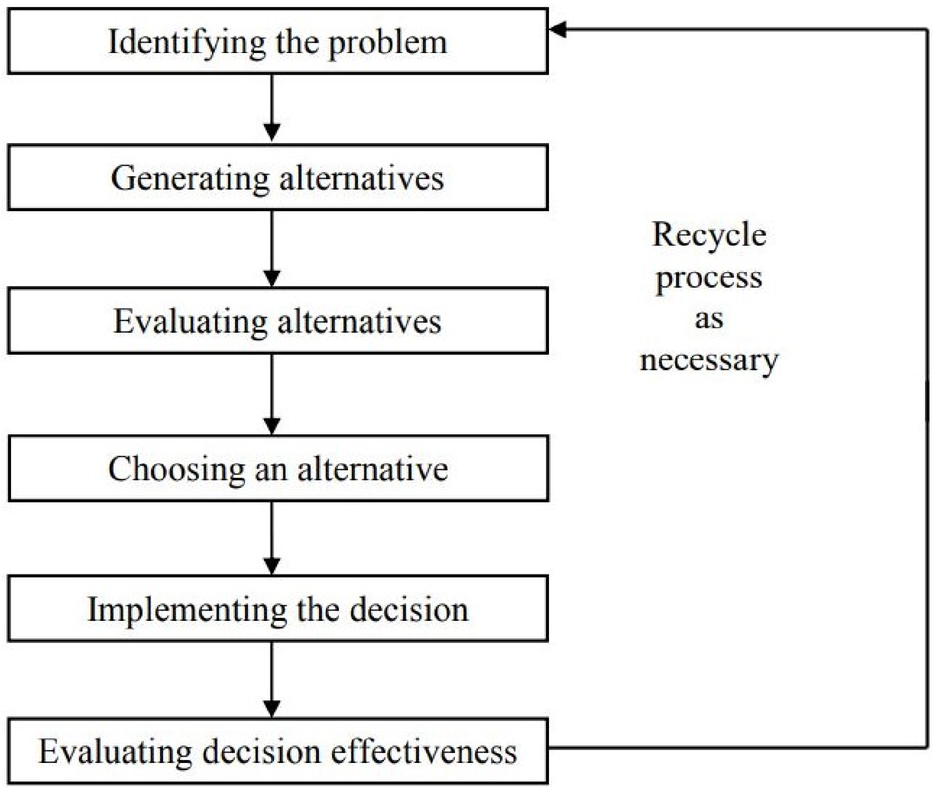

Figure 2.6: The decision-making process [13]

2.4.1 Identifying the Problem

A well-defined problem will have a significant impact on decision-making quality. One of the useful approaches is that sometimes it is simpler to decide what the problem is not, compared to deciding what the problem is. Problems and potential solutions must also be assessed in relation to other problems that exist in order to determine their priority and importance. A key step in that respect is to define cause-and-effect relationships in order to identify the root causes of the problem. In short, this process of analyzing problems involves four general steps [13]:

1. Problem identification.
2. Definition of what the problem is and is not.
3. Prioritizing the problem.
4. Testing for cause-effect relationships.

2.4.2 Checking Alternatives

After the problem has been defined, the second step in the decision-making process is looking for potential solutions. There are three significant steps involved in this stage:

- Generating alternatives.
- Checking alternatives.
- Choosing the best alternative.

Generating Alternatives

This stage involves thinking and generating a number of potential solutions to the problem. The goal is to develop a diverse set of alternatives, including conventional and innovative solutions. In data-driven decision-making, this can include examining historical data, predictive analytics, and subject matter expert inputs to develop workable solutions. Effective alternative generation ensures that decision-makers do not only have a single path to follow but multiple options to consider [14].

Alternatives Analysis

Alternatives, after they are developed, must be evaluated in detail based on pre-established criteria such as feasibility, cost, risks involved, and results expected. Quantitative techniques are helpful in the comparison of the different options in an objective manner, decision matrices and cost benefit analysis. Big data approaches, such as machine learning models and simulations, can also illuminate the probable impact of each alternative.

To make an overall evaluation, the decision-makers have to ask three significant questions:

- Is the option feasible? It examines whether the proposed solution is realistic within the available resources, time, and constraints. Feasibility study considers factors such as budget, technical feasibility, and challenges in implementation.

- Is it a practical alternative? Even if an alternative is feasible, it must also effectively resolve the problem. Decision-makers must ask themselves whether the alternative meets the desired objectives and delivers the expected outcomes. This stage may involve performance analysis, risk assessment, and alignment with organizational goals.

- What will be its effect on people? The final question addresses the human aspect of decision-making. It considers how the alternative is going to affect stakeholders, employees, customers, or society. A desirable alternative must minimize negative impacts and improve benefits to concerned parties.

By sequentially answering these questions, decision-makers can ensure that the selected alternative will not only be feasible but also desirable [13].

Choosing the Best Alternative

After due consideration, the most suitable option is selected based on data-driven conclusions and a strategic agenda. The selection has to be in accordance with the organization's goals and constraints while minimizing risks and enhancing benefits. In some cases, a combination of multiple options is used to derive a more feasible solution [14].

2.4.3 Implementing and Evaluating the Decision

After a decision has been reached, the final step is its implementation, making sure that the selected action is properly implemented. Implementation entails resource allocation, assigning tasks, and setting an execution timeline. Open communication among stakeholders is necessary to make sure that all parties involved know their responsibilities and what the expected results are. Similarly, probable risks and difficulties should be anticipated with contingency plans in place to counter any difficulty that may arise. Evaluation is an ongoing process that analyzes the effectiveness of the decision that was made. This involves quantitative assessments of significant performance indicators (KPIs), feedback gathering, and researching whether or not the decision reduced the problem it was designed to solve. The evaluation process allows for adjustments that are needed so that improvements or other tools can be utilized if the intended result is not observed. A successful evaluation process provides the possibility for continuous improvement such that decisions in the future are informed by lessons acquired. The circular process develops problem-solving ability and enhances the ability to make better and fact-driven decisions with time [13][14].

2.5 Big Data Analytics for Stock Market Analysis

This research aims to analyze stock price movements using big data analytics techniques. Historical stock market data, collected from Yahoo Finance API, are utilized in the study, covering January 1, 2015, to April 30, 2025. The dataset includes stock prices of multiple companies, incorporating key indicators such as opening price, closing price, high, low, and trading volume, providing a comprehensive foundation for analyzing stock market trends.

To predict future stock prices, the research employs Long Short-Term Memory (LSTM) networks [15], a specialized type of Recurrent Neural Network (RNN) Figure 2.7. LSTMs are widely recognized for their ability to capture long-term dependencies in sequential financial data, making them well-suited for stock price forecasting. By leveraging this model, the research aims to predict stock price movements and identify potential investment opportunities. The ultimate objective is to develop a data-driven approach that assists investors in making informed decisions about which stocks to buy, when to buy, and when to sell based on predictive insights. The results of this analysis can help traders, financial analysts, and portfolio managers optimize investment strategies while mitigating risks associated with market volatility.

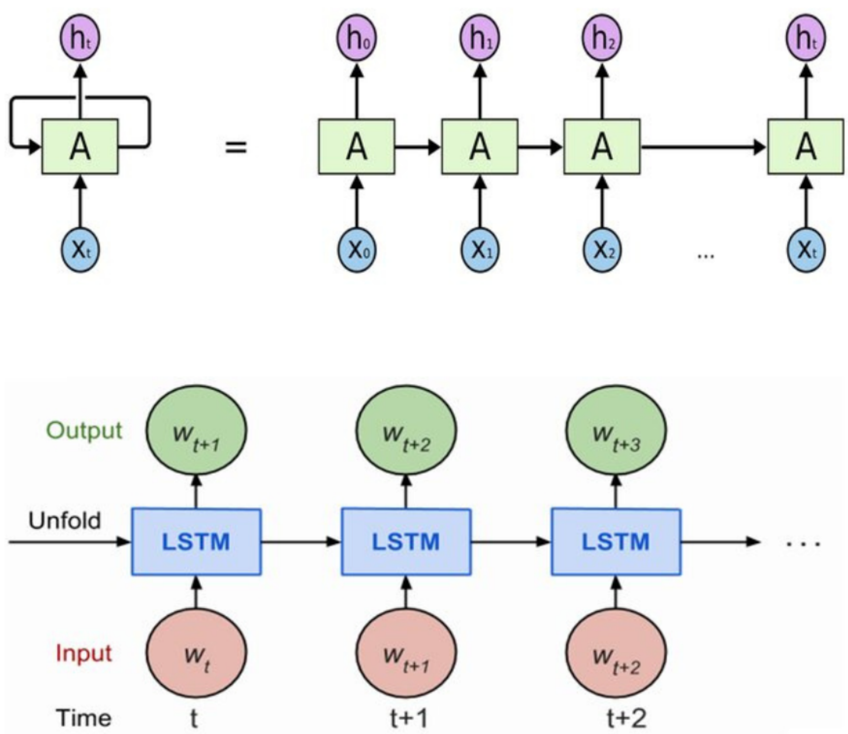

Figure 2.7: The RNN model for stock prices prediction [16].

Chapter 3

Methodology

The research adopts a deep learning approach using Long Short-Term Memory (LSTM) networks for the prediction of stock prices and investment decision-making. The process entails some necessary stages: data collection, pre-processing, model design, training and testing, and investment simulation.

3.1 Data Collection and Feature Engineering

Historical daily stock data was obtained using the `yfinance` API for multiple publicly traded assets, spanning from January 2015 to April 2025. This is the top 5 Rows of the AAPL stock . The dataset includes standard OHLCV (Open, High, Low, Close, Volume) features [17] here is the top 5 rows of AAPL stock Table 3.1.

Table 3.1: Top 5 Rows of the Stock Dataset (AAPL Example)

Date	**Open**	**High**	**Low**	**Close**	**Volume**
2024-01-02	185.22	186.15	182.92	184.25	74890200
2024-01-03	183.59	185.40	182.25	183.31	63140000
2024-01-04	183.90	186.10	183.15	185.92	75122100
2024-01-05	186.45	188.22	185.50	187.94	78561200
2024-01-08	188.20	189.44	186.75	187.22	68947200

Also, There was added some new features as :

- Relative Strength Index (RSI)
- Exponential Moving Averages (EMAF, EMAM, EMAS)

The target variable is defined as the next-day change in closing price as shown in Equation 3.1:

$$\text{Target}_t = \text{Close}_{t+1} - \text{Close}_t \tag{3.1}$$

3.2 Sequence Construction and Feature Scaling

To prepare the input for the LSTM model, we transformed the historical data into a supervised learning format using a sliding window approach. Each sample corresponds to a sequence of backcandles consecutive timesteps (set to 45), representing the historical context for prediction.

Before sequence construction, all features were normalized using MinMax scaling to the $[0, 1]$ range, ensuring stable gradient flow during training. The scaler was fit on the entire feature set to preserve relative scaling between indicators.

For each feature f_j (where $j = 1, \ldots, d$ features), we constructed a feature matrix X such that each sample X_i contains the last T timesteps of f_j, where T is the backcandles window size. In cases where fewer than T historical points were available (i.e., at the start of the dataset), sequences were padded with zeros to maintain consistent input shape.

Formally, for a given time i, the sequence for feature j is defined as in Equation 3.2:

$$X_i^{(j)} = \begin{cases} [0, \ldots, 0, f_0^{(j)}, \ldots, f_{i-1}^{(j)}], & \text{if } i < T \\ [f_{i-T}^{(j)}, \ldots, f_{i-1}^{(j)}], & \text{otherwise} \end{cases} \tag{3.2}$$

Once sequences were generated, we reshaped the data into the 3D format required for LSTM training: $(\text{samples}, \text{timesteps}, \text{features})$. The target variable was defined as the third-last column in the feature set, corresponding to the pre-engineered price change label. The dataset was finally split into training and testing sets using an 80/20 temporal split to prevent data leakage.

3.3 Model Architecture and Training Procedure

The complete pipeline for the stock market prediction and decision-making system is illustrated in Figure 3.1. The flow consists of the following sequential steps:

1. Historical Stock Data Collection: The process begins by acquiring historical stock data, including open, high, low, close, and volume (OHLCV) information. This data serves as the foundation for time-series forecasting.

2. Feature Engineering: Technical indicators such as RSI and EMA are computed from the raw OHLCV values. These features provide the model with domain-relevant signals that enhance prediction accuracy.

3. LSTM Model Prediction: The engineered features are input into a long-short-term memory (LSTM) model trained to predict the next-day delta, the change between the closing and opening prices ($\text{Close}_{t+1} - \text{Open}_{t+1}$).

4. Multi-Day Forecasting: Using recursive inference, the LSTM model forecasts the expected price changes over a configurable horizon (e.g., the next 10 trading days), simulating how prices might evolve.

5. Buy/Sell Opportunity Extraction: For each predicted price sequence, the system identifies the optimal buy and sell days that would yield the maximum profit per share.

6. ILP Optimization: An Integer Linear Programming (ILP) model receives the candidate trades and selects the combination of trades that maximizes total profit while respecting a user-defined investment budget constraint.

7. Final Portfolio Plan: The result of the ILP solver is a complete portfolio strategy that specifies how many shares to buy, on which day to buy and sell each stock, and the expected return for each trade.

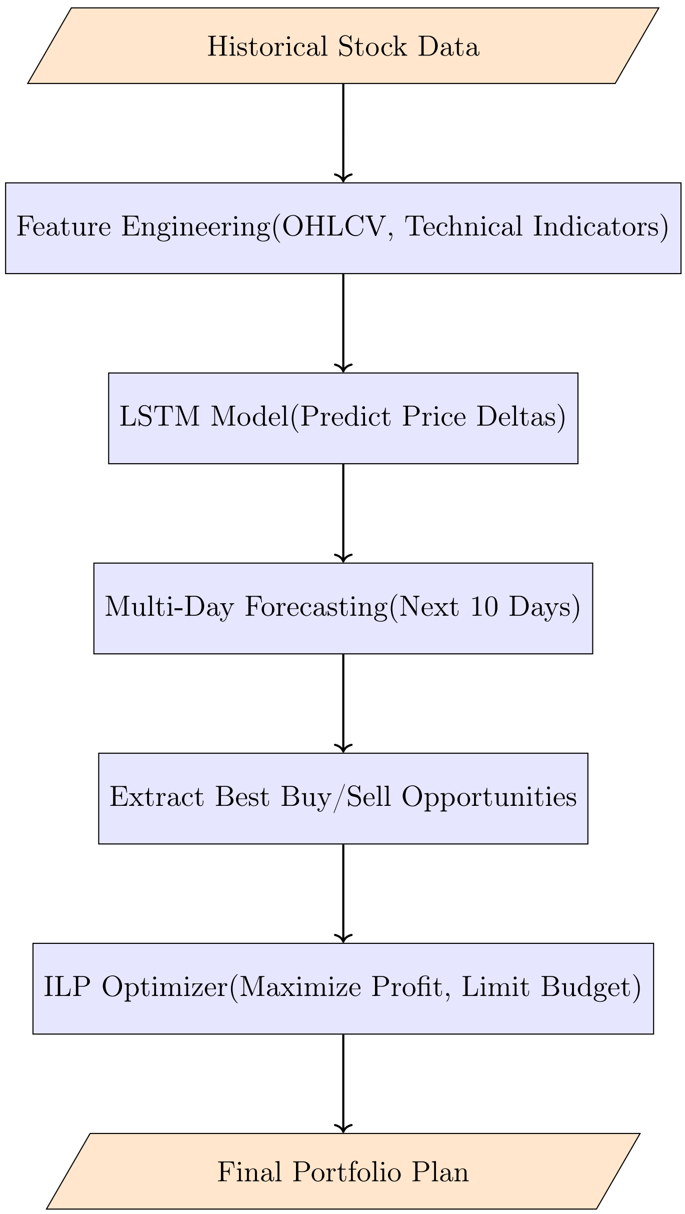

Figure 3.1: System pipeline from time series modeling to portfolio decision-making.

The predictive model used in this study is based on a unidirectional Long Short-Term Memory (LSTM) neural network designed for time series regression. The architecture consists of a single LSTM layer followed by a fully connected output layer. The network is defined as follows Figure 3.2:

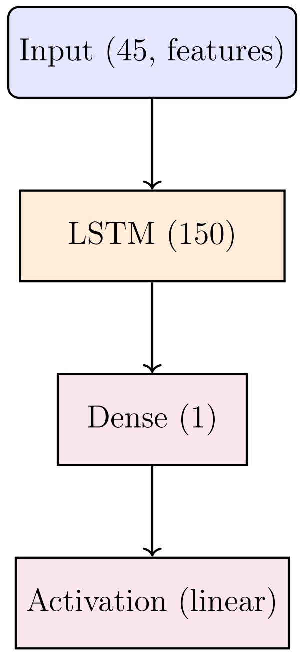

Figure 3.2: LSTM model architecture used for predicting stock price deltas.

The model is trained using the Adam optimizer [18] with a learning rate of 0.001, and the Mean Squared Error (MSE) loss function which is defined in this equation Equation 3.3, which is well-suited for continuous regression tasks. Formally, the MSE is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.3}$$

where:

- n is the number of samples
- y_i is the true value for sample i
- $\hat{y}_i$ is the predicted value for sample i

The model is compiled and trained using the following parameters:

- Epochs: 100
- Batch size: 32
- Validation split: 10% of the training data
- verbose: 1

The use of MSE encourages the model to penalize large deviations from the ground truth, while the Adam optimizer ensures adaptive learning rate adjustments for efficient convergence. Model training is monitored with validation loss to ensure generalization and prevent overfitting.

3.4 Model Evaluation and Multi-Step Forecasting

To evaluate the predictive performance of the LSTM model, we report the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2) for both training and testing datasets. Table 3.2 summarizes the results.

Table 3.2: Training vs Testing Performance Metrics

Metric	**Training Set**	**Testing Set**
Mean Absolute Error (MAE)	0.5417	1.2794
Root Mean Squared Error (RMSE)	0.8508	2.1539
Coefficient of Determination (R^2)	0.6701	0.4382

3.4.1 Metric Definitions

1. Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.4}$$

Where:

- n is the number of trading days in the test set,
- y_i is the actual value of the target delta on day i (i.e., $\text{Close}_i - \text{Open}_i$),
- $\hat{y}_i$ is the predicted value of the target delta on day i.

A lower MAE indicates that the model is consistently making predictions that are close to the actual intraday price movements.

2. Root Mean Squared Error (RMSE) is the square root of the average squared differences between predicted and actual values:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.5}$$

Where:

- y_i and $\hat{y}_i$ are the same as above,
- RMSE is especially useful in financial forecasting because larger errors may correspond to significant financial risk.

A lower RMSE indicates that the model avoids large deviations, which is important in portfolio planning.

3. Coefficient of Determination (R^2) evaluates how well the predicted values approximate the actual data points:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.6}$$

Where:

- $\bar{y}$ is the mean of the actual target deltas over the test period,
- The numerator represents the model's residual sum of squares,
- The denominator represents the total sum of squares,
- An R^2 score close to 1 means the model can explain most of the market movement variance.

In financial applications, a higher R^2 indicates that the model captures key trends in price movement and is more reliable for decision-making.

3.5 Inverse Scaling and Forecasting Strategy

After generating predictions, the output was transformed back to its original scale using the inverse of the previously applied MinMax normalization. This step ensures that the error metrics reflect true price differences rather than normalized values.

To support real-world trading decisions, the trained model is applied recursively to predict the next x future days (e.g., $x = 10$). These multi-step forecasts are then passed to the investment decision engine to identify the optimal buy and sell windows using integer linear programming (ILP) to know more about it, check [19].

3.6 Optimal Trade Selection and Capital Allocation via ILP

Once the LSTM model generates multi-day forecasts for each stock, we identify profitable trading opportunities and allocate capital optimally using Integer Linear Programming (ILP). This two-stage process consists of (1) determining the best buy/sell window for each stock, and (2) solving an ILP problem to maximize total expected profit under a fixed budget constraint.

Step 1: Identifying Best Trade Opportunities

For each stock, we examine the predicted price sequence and compute the pair of days $(t_{\text{buy}}, t_{\text{sell}})$ that yield the highest profit per share:

$$\text{profit}_i = \max_{j>k}(\text{Price}_j - \text{Price}_k) \tag{3.7}$$

This is implemented in the function, which loops through each stock's forecast, tracks the lowest historical price seen so far, and records the maximum achievable profit with corresponding buy and sell days. The output is a list of trade options, each described by its stock ticker, buy/sell day, and profit per share.

Step 2: Investment Optimization Using ILP

To determine how many shares of each stock to purchase under a total budget constraint, we formulate the following ILP:

$$\text{maximize} \quad \sum_{i=1}^{n} x_i \cdot (s_i - b_i) \tag{3.8}$$

$$\text{subject to} \quad \sum_{i=1}^{n} x_i \cdot b_i \leq \text{Budget} \tag{3.9}$$

$$x_i \in \mathbb{Z}^+, \quad \forall i \in \{1, \ldots, n\} \tag{3.10}$$

Where:

- x_i is the number of shares to buy of stock i
- b_i and s_i are the predicted buy and sell prices respectively
- n is the number of candidate stocks (optionally limited to the top N by profit-per-share)

This formulation ensures that:

- No fractional shares are purchased (integer constraint)
- The total investment does not exceed the available budget
- Profit is maximized by allocating capital to the most profitable trades

The final plan includes selected stocks, optimal share counts, total invested capital, expected returns, and net profits.

Chapter 4

Experimental Results

This chapter presents a comprehensive review of deep learning models in stock market forecasting. Two models were implemented and compared: one individual Long Short-Term Memory (LSTM) network by itself and a Convolutional Neural Network with LSTM (CNN+LSTM) hybrid [20]. The purpose is to compare their ability to forecast the change in stock prices from historical market activity and technical indicators.

4.1 Experimental Setup

The models were trained on historical stock data from Yahoo Finance, covering the period from January 2015 to April 2024. The dataset included engineered features such as RSI, EMAF, EMAM, and EMAS. The target variable was defined as the next day's price change (i.e., Close_{t+1} - Open_{t+1}).

A sliding window of 45 back-candles (past days) was used to generate sequences for model input. The dataset was split 80/20 into training and testing sets. All input features were normalized using MinMaxScaler, and inverse transformation was applied to evaluate predictions in the original scale.

4.2 Model Architectures

This section discusses the neural network models employed for stock price change prediction. Two models have been designed and validated: the LSTM model and the hybrid CNN+LSTM model. Both models try to capture temporal patterns from financial history, but with different strategies towards feature extraction and sequence modeling.

4.2.1 LSTM Model

The LSTM model consisted of a single LSTM layer with 150 units, followed by a dense layer with a linear activation function. This model is designed to capture long-term temporal dependencies in the time series.

4.2.2 CNN+LSTM Model

The hybrid CNN+LSTM model begins with a 1D convolutional layer to extract local patterns in the time series, followed by max pooling and an LSTM layer with 150 units. This structure combines spatial feature extraction with sequential learning as in the Figure 4.1.

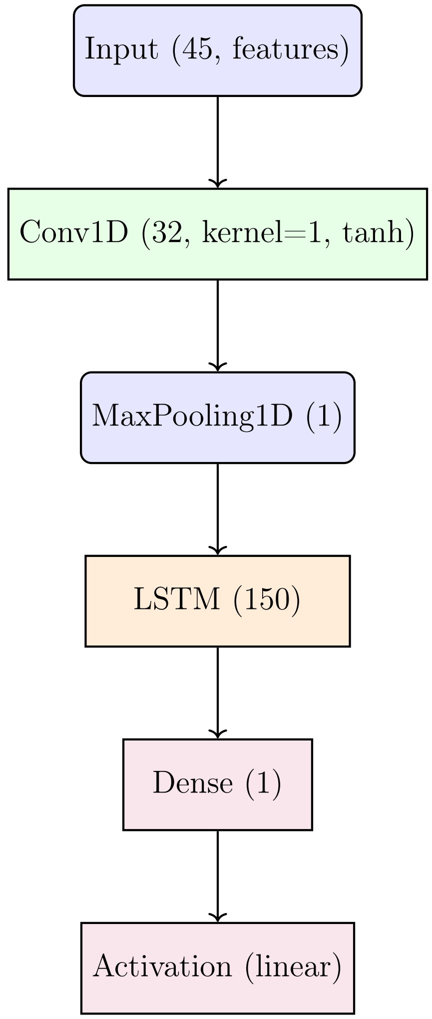

Figure 4.1: CNN+LSTM model architecture used for predicting stock price deltas.

4.3 Evaluation Metrics

The models were evaluated using the following regression metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of errors without considering their direction.
- Root Mean Squared Error (RMSE): Emphasizes larger errors more due to squaring.
- R^2 Score: Indicates the proportion of variance explained by the model (1.0 is perfect, 0.0 means no improvement over mean prediction).

4.4 Comparative Analysis

Table 4.1 summarizes the test performance of both models.

Table 4.1: Performance comparison between LSTM and CNN+LSTM models

Metric	**Dataset**	**LSTM**	**CNN+LSTM**	**Better Model**
MAE	Train	0.5417	0.5368	CNN+LSTM
	Test	1.2794	1.2656	CNN+LSTM
RMSE	Train	0.8508	0.8548	LSTM
	Test	2.1539	2.1962	LSTM
R^2	Train	0.6701	0.6665	LSTM
	Test	0.4382	0.4187	LSTM

As shown in Table 4.1, the LSTM model is better than the CNN+LSTM hybrid model in most of the test set evaluation metrics, including MAE, RMSE, and R^2. Although the performance differences are not large, they indicate that the CNN layer helps extract local temporal features that complement the LSTM's sequential modeling capabilities. This suggests that combining convolutional layers with LSTM may offer a slight advantage in modeling stock price dynamics.

4.5 Visual Analysis

To further understand model behavior, Figures 4.2 and 4.3 show the predicted vs. actual values for both the target delta and the reconstructed close prices, while Figures 4.4 and 4.5 show the predicted vs. actual values for the delta change between the close and open.

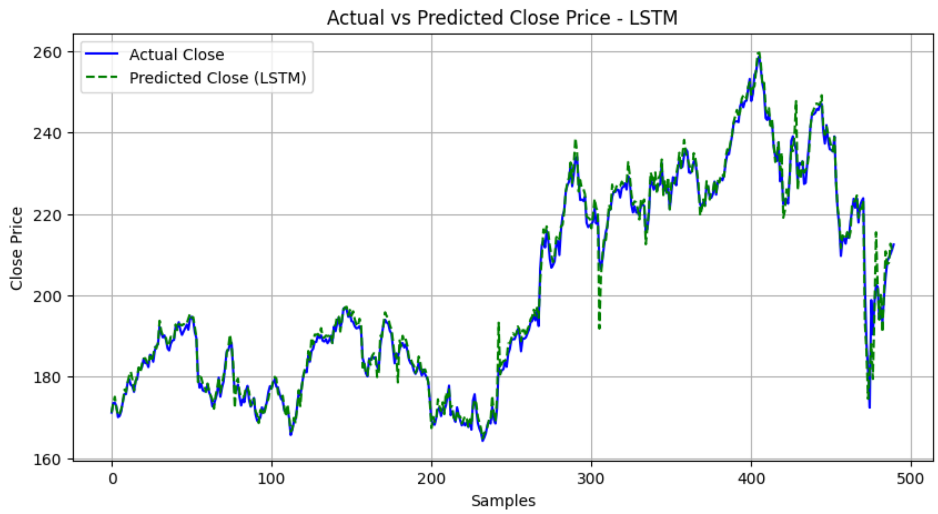

Figure 4.2: LSTM model: Actual vs Predicted Close prices

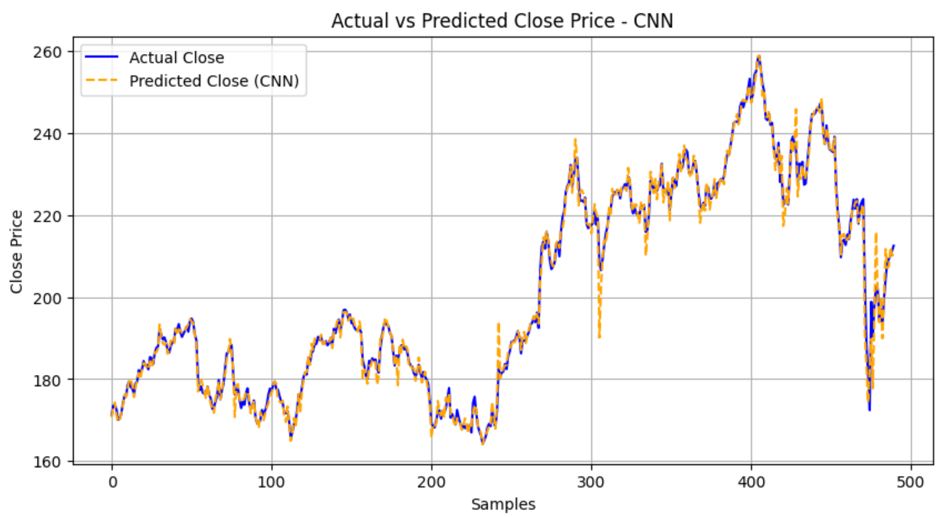

Figure 4.3: CNN+LSTM model: Actual vs Predicted Close prices

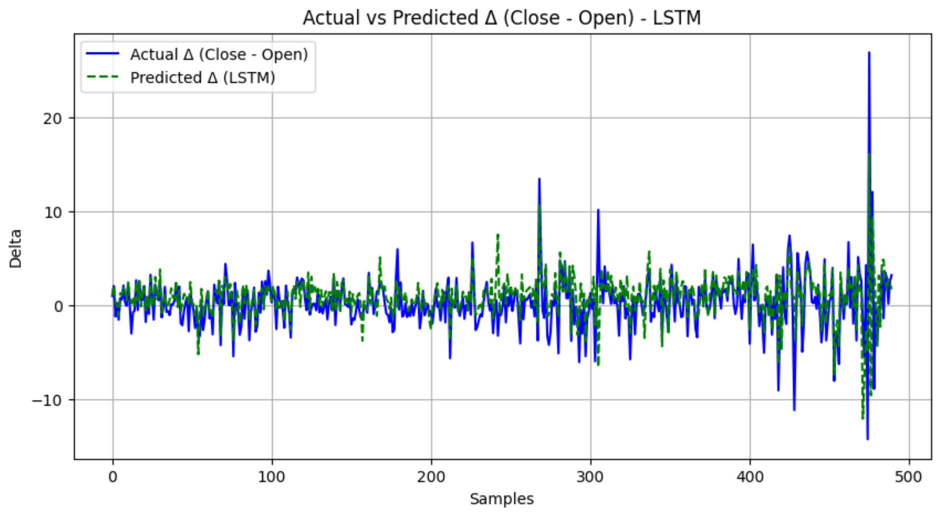

Figure 4.4: LSTM model: Actual vs Predicted (Close - Open)

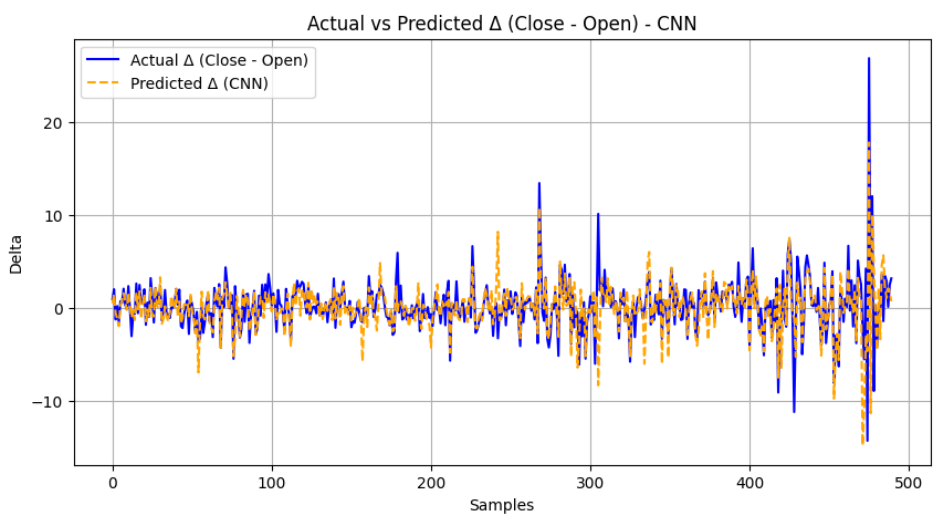

Figure 4.5: CNN+LSTM model: Actual vs Predicted (Close - Open)

Chapter 5

Conclusion and Future Work

This chapter summarizes the key findings of the research and outlines a potential approach for future research. The overall aim of this research was to create and implement a predictive trading system on the basis of deep learning for price forecasting and integer linear programming (ILP) for optimal investment allocation. By integrating time series processing and discrete optimization, the platform intends to support decision-making in algorithmic trading with explainable and actionable insights.

5.1 Key Contributions

The main contributions of this research are as follows:

1. Proposed a hybrid model combining a Long Short-Term Memory (LSTM) neural network for forecasting future price movements and Integer Linear Programming (ILP) for maximizing budget-constrained investments.

2. Constructed a sequence processing pipeline that maps past stock history data into a supervised learning format with the help of a sliding window and normalized feature representation, including technical indicators.

3. Used a unidirectional LSTM model implementation and training in predicting the closing price delta rather than absolute price to improve the stability of learning and predictive quality.

4. Utilized a recursive prediction methodology to generate x-day ahead future price sequences, allowing simulation of true market conditions.

5. Implemented an ILP optimization core to determine the optimal number of shares to buy from the highest profit transactions given a maximum total budget, with integer-only division of capital.

6. Conducted a quantitative evaluation of the model using standard regression metrics (MAE, RMSE, and R^2), demonstrating good predictive accuracy on training and test sets.

5.2 Future Work

While the proposed framework achieved promising results, there are several directions in which this research can be extended and refined:

1. Enlarge the model through cross-validation on bigger and more diverse datasets (e.g., forex, cryptocurrency) to verify the applicability of the model to different markets.

2. Examine reinforcement learning methods to express trading as a sequential decision problem with changing environments.

3. Conduct a back-testing test using past stock price history data to study the profitability of the investment approach based on real-world market conditions.

4. Investigate other feature engineering techniques, such as lag-based statistical features, wavelet transforms, or sentiment analysis of social media sites and financial news, to enhance the input representations of the model.

5. Hyperparameter tuning of the LSTM model, including number of layers, hidden units, dropout rates, learning rate, and sequence length, using techniques such as grid search, random search, or Bayesian optimization.

References

[1] H. J. Hadi, A. H. Shnain, S. Hadishaheed, and A. Haji, "Big data and five v's characteristics," *International Journal of Advances in Electronics and Computer Science*, vol. 2, no. 1, pp. 16–23, 2015. [Online]. Available: https://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf.

[2] L. Qi Zian, Y. Jaya Kumar, and N. Z. Zulkarnain, "Tools for big data and analytics," *Tools for Big Data and Analytics*, Sep. 2021. [Online]. Available: https://www.researchgate.net/publication/358486832_Tools_for_Big_Data_and_Analytics.

[3] K. Batko and A. Ślęzak, "The use of big data analytics in healthcare," *Journal of Big Data*, vol. 9, 2022. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00553-4.

[4] E. Al-Lozi, A. Alfityani, A. Alsmadi, A. Al-hazimeh, and J. Al-Gasawneh, "The role of big data in financial sector: A review paper," *International Journal of Data and Network Science*, vol. 6, Jun. 2022. [Online]. Available: https://www.researchgate.net/publication/361227210_The_role_of_big_data_in_financial_sector_A_review_paper.

[5] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," *Journal of Business Research*, vol. 69, no. 2, pp. 897–904, 2016, ISSN: 0148-2963. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296315002842.

[6] A. Gunasekaran, T. Papadopoulos, R. Dubey, *et al.*, "Big data and predictive analytics for supply chain and organizational performance," *Journal of Business*

Research, vol. 70, pp. 308–317, 2017, ISSN: 0148-2963. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014829631630491X.

[7] D. A. Bajpai and A. Arya, "Big data analytics in cyber security," *International Journal of Computer Science and Engineering (JCSE)*, vol. 6, no. 7, 2018. [Online]. Available: https://www.ijcseonline.org/pdf_paper_view.php?paper_id=2502&116-IJCSE-04289.pdf.

[8] A. M. S. Osman and A. Elragal, "Smart cities and big data analytics: A data-driven decision-making use case," *Smart Cities*, vol. 4, no. 1, pp. 286–313, 2021. [Online]. Available: https://www.mdpi.com/2624-6511/4/1/18.

[9] A. Alkhalil, M. Abdallah, A. Alogali, and A. Aljaloud, "Applying big data analytics in higher education: A systematic mapping study," *International Journal of Information and Communication Technology Education*, vol. 17, Jan. 2021. [Online]. Available: https://www.researchgate.net/publication/352750566_Applying_Big_Data_Analytics_in_Higher_Education_A_Systematic_Mapping_Study.

[10] M. Islam, "Data analysis: Types, process, methods, techniques and tools," *International Journal on Data Science and Technology*, vol. Vol. 6, No. 1, 2020, pp. 10–15, Jan. 2020. [Online]. Available: https://www.researchgate.net/publication/338528149_Data_Analysis_Types_Process_Methods_Techniques_and_Tools.

[11] S. A. Mazhar, "Methods of data collection: A fundamental tool of research," *Journal of Integrated Community Health*, vol. 10, pp. 6–10, Jun. 2021. [Online]. Available: https://www.researchgate.net/publication/356269289_Methods_of_Data_Collection_A_Fundamental_Tool_of_Research.

[12] C. Cote, *4 types of data analytics to improve decision-making*, 2021. [Online]. Available: https://online.hbs.edu/blog/post/types-of-data-analysis.

[13] F. C. Lunenburg, "The decision making process," *National Forum of Educational Administration and Supervision Journal*, vol. 27, no. 4, 2010, Accessed: 15 March 2025. [Online]. Available: https://www.nationalforum.com/Electronic%20Journal%20Volumes/Lunenburg,%20Fred%20C.%20The%20Decision%20Making%20Process%20NFEASJ%20V27%20N4%202010.pdf.

[14] R. Gupta, "Understanding decision making process," *UNDERSTANDING DECISION MAKING PROCESS*, Dec. 2022. [Online]. Available: https://www.researchgate.net/publication/365892363_UNDERSTANDING_DECISION_MAKING_PROCESS.

[15] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, Dec. 2020. [Online]. Available: https://www.researchgate.net/publication/340493274_A_Review_on_the_Long_Short-Term_Memory_Model.

[16] A. H. Manurung, W. Budiharto, and H. Prabowo, "Algorithm and modeling of stock prices forecasting based on long short-term memory (lstm)," *ICIC Express Letters*, vol. 12, pp. 1277–1283, Dec. 2018. [Online]. Available: https://www.researchgate.net/publication/331838603_Algorithm_and_modeling_of_stock_prices_forecasting_based_on_long_short-term_memory_LSTM.

[17] P. Kavya, S. Saagarika, R. Subavarsshini, C. Nivetheni, M. Muthuvel, and P. Velvadivu, "Stock market analysis," p. 7, Jul. 2021. [Online]. Available: https://www.researchgate.net/publication/356776423_STOCK_MARKET_ANALYSIS.

[18] M. Reyad, A. Sarhan, and M. Arafa, "A modified adam algorithm for deep neural network optimization," *Neural Computing and Applications*, vol. 35, pp. 1–18, Apr. 2023. [Online]. Available: https://www.researchgate.net/publication/370260831_A_modified_Adam_algorithm_for_deep_neural_network_optimization.

[19] G. Lancia and P. Serafini, "Integer linear programming," pp. 43–66, Jan. 2018. [Online]. Available: https://www.researchgate.net/publication/319449795_Integer_Linear_Programming.

[20] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A cnn-lstm-based model to forecast stock prices," *Complexity*, vol. 2020, pp. 1–10, Nov. 2020. [Online]. Available: https://www.researchgate.net/publication/347147048_A_CNN-LSTM-based_model_to_forecast_stock_prices.