STOCK MARKET ANALYSIS USING DATA MINING TECHNIQUES

Shylendra Sai Bangaru[1] Hemamalini Venkatesh Vanitha[1]
Prof.Naveen Ramachandra Reddy[2]
Dr.Sumona Mondal[3]

[1] Department of Applied Data Science
[2] Professor of Applied Data Science
[2] Co-Director of the MS Program in Applied Mathematics
Clarkson University

Abstract :

The ability to accurately predict stock market movements and trading actions has become increasingly critical in financial decision-making, where timely insights can significantly influence investment outcomes. In this project, a comprehensive framework combining machine learning, deep learning, reinforcement learning, and data mining techniques was developed to empower data-driven trading strategies. Utilizing 1.5 years of historical data for 30 major stock tickers, enriched with Open, High, Low, Close, Volume, and Date features, multiple models were implemented to address key investment challenges. Smart Portfolio Optimization based on Modern Portfolio Theory (MPT) was used to maximize returns while minimizing risk through optimal asset allocation. Anomaly Detection frameworks using rolling statistics and Isolation Forests identified unusual stock movements indicative of potential insider trading. Next-Day Price Prediction leveraged Long Short-Term Memory (LSTM) networks to forecast short-term closing prices with high precision. Hidden Correlation Discovery employed Pearson correlation matrices and hierarchical clustering to uncover inter-stock dependencies, facilitating smarter diversification strategies. Furthermore, a Dynamic Buy/Sell/Hold Action Model was trained using Deep Q-Learning reinforcement learning to autonomously suggest real-time trading decisions based on evolving market states. Collectively, these models demonstrate the power of integrating AI-driven techniques into financial systems, providing investors with actionable, reliable, and adaptive decision support tools. The outcomes of this project not only validate the feasibility of AI-based investment management but also offer a scalable foundation for future developments in intelligent trading systems.

Index Terms: Stock Market Prediction, Portfolio Optimization, Anomaly Detection, Reinforcement Learning, Deep Q-Learning, LSTM, Price Forecasting, Correlation Matrix, Hierarchical Clustering, Smart Trading Systems, Machine Learning, Deep Learning

1 Introduction

Financial markets, particularly stock trading, represent one of the most dynamic and unpredictable sectors of the global economy. Despite significant advances in financial modeling and market analysis, stock price movements remain highly volatile and challenging to forecast, contributing to substantial economic uncertainty worldwide. As a result, there is an urgent need for reliable and intelligent systems that can predict stock market trends, enabling investors to make informed decisions and optimize their investment strategies. This project establishes the foundation for this study by outlining the goals of the investigation, discussing data mining techniques, and highlighting machine learning models designed for stock price prediction and market behavior analysis.

Stock market volatility, characterized by rapid and unpredictable price swings, is influenced by a variety of factors including economic indicators, corporate earnings, and global events. Sharp fluctuations in stock prices can lead to significant financial losses or gains, affecting individual investors and large institutions alike. Despite advances in algorithmic trading and risk management, markets remain sensitive to external shocks and investor sentiment. Therefore, predictive models that can detect patterns, forecast price movements, and optimize portfolio strategies are crucial for minimizing risks and maximizing returns. By accurately identifying market trends and anomalies, investors can implement proactive measures such as portfolio rebalancing, hedging strategies, and tactical asset allocation. Consequently, developing robust, accurate predictive models for stock market analysis holds immense potential for improving investment outcomes and reducing economic vulnerabilities associated with financial market volatility.

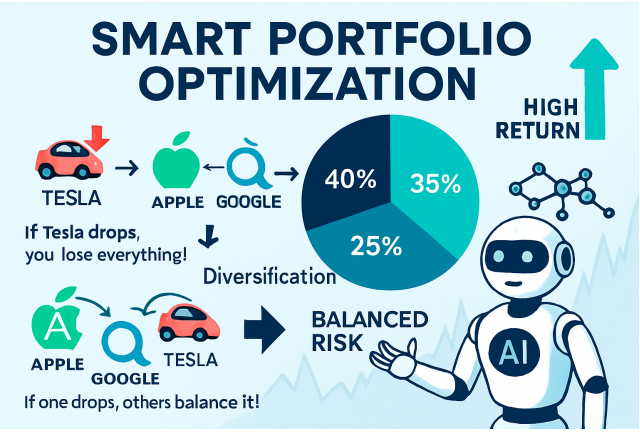


Figure 1. Smart Portfolio

2 Background

The stock market plays a crucial role in the global economy. It channels investments into businesses, drives innovation, and supports economic growth across industries. However, if the stock market becomes unstable, it can trigger widespread financial crises, disrupt businesses, and impact livelihoods within minutes.

As highlighted in the Journal of Financial Markets and Data Science (2021), the stock market has deeply integrated into everyday economic life, influencing everything from consumer confidence to business operations. Market inefficiencies, volatility, and behavioral biases are some of the key challenges that investors face in navigating stock trading environments. Several factors can significantly impact stock market behavior, including:

- Macroeconomic indicators such as GDP growth and inflation rates.
- Company-specific news like earnings reports and leadership changes.
- Global geopolitical events and trade policies.
- Investor sentiment and market psychology.
- Monetary policy decisions, including interest rate changes.
- Technological disruptions and sector innovations.
- Regulatory changes impacting industries.
- External shocks such as pandemics or natural disasters.

Predicting stock market trends is a complex process that demands expertise and advanced analytical skills. Traditionally, investment decisions were based largely on analysts' experience and historical financial data. Key market events such as financial bubbles, crashes, and corporate scandals are linked to different types of market anomalies and behaviors. Each type is associated with unique patterns. For instance, tech bubbles often exhibit rapid, unsustainable price surges, but not all surges indicate a bubble. Some may stem from genuine innovation or structural shifts. Analysts combine price movements, volume patterns, and external indicators to validate their assessments and craft informed investment strategies.

3 Objective

Provide an overview of the current landscape of stock market prediction methods and their limitations. Highlight the significance of integrating data mining and machine learning techniques in finance to enhance the accuracy and reliability of stock price forecasting. Review recent advancements and applications of data mining, deep learning, and portfolio optimization in stock market analysis. Discuss the challenges and opportunities involved in developing and deploying predictive models for real-world trading and investment strategies. Propose future directions for research and practical implementation, including the integration of alternative data sources, continuous model validation, and the use of intelligent decision support systems for investors and financial institutions.

4 Dataset

The dataset for this project was collected primarily from Yahoo Finance, covering a 1.5-year period from September 2022 to February 2024. It contains historical stock data for 30 major companies and indices, including fields such as Ticker Symbol, Date, Open Price, High Price, Low Price, Close Price, and Trading Volume. Additional stock ticker information was initially fetched from the SEC company listings database to build a flexible and extensible stock list. While raw financial data is abundant, one of the key challenges encountered was ensuring that the dataset was suitable for predictive modeling. The original dataset required extensive cleaning and preprocessing to handle issues such as missing values, duplicate entries, and inconsistent formatting. Forward-filling techniques were applied to address gaps in the time series, and duplicate rows were removed for data integrity. Feature engineering played a crucial role in enriching the dataset. Technical indicators including Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Exponential Moving Average (EMA), volatility measures, and daily returns were calculated to better capture trends, momentum, and risk factors in stock behavior. The final structured dataset was normalized using Min-Max scaling, and sequence windows were created to support time-series models such as LSTM networks.

Figure 2. Artificial Intelligence predicting Stock Market

5 Variables Description

VARIABLE	DESCRIPTION	MEASUREMENT UNITS
Date	Trading date for the stock data	YYYY-MM-DD
Ticker	Stock symbol representing the company (e.g., AAPL, TSLA)	Text
Open	Price of the stock at market open	US Dollars ($)
High	Highest price the stock reached on the trading day	US Dollars ($)
Low	Lowest price the stock reached on the trading day	US Dollars ($)
Close	Final price of the stock at market close	US Dollars ($)
Volume	Total number of shares traded on the given day	Number of Shares
MACD	Moving Average Convergence Divergence – trend-following momentum indicator	Numeric (Indicator Value)
MACD_Signal	Signal line for MACD used to trigger buy/sell signals	Numeric (Indicator Value)
EMA_7	Exponential Moving Average over 7 days – short-term trend signal	US Dollars ($)
EMA_30	Exponential Moving Average over 30 days – long-term trend signal	US Dollars ($)
RSI_14	Relative Strength Index calculated over 14 days – momentum oscillator	0 to 100
Daily Return	Percentage return from previous close	Decimal (e.g., 0.015 = 1.5%)
MA_7	7-day Simple Moving Average – smoothing recent price data	US Dollars ($)
MA_30	30-day Simple Moving Average – longer-term price trend	US Dollars ($)
Volatility_7	7-day Rolling Standard Deviation – short-term volatility	Standard deviation (returns)
Volatility_30	30-day Rolling Standard Deviation – long-term volatility	Standard deviation (returns)

Figure 3. Description of Variables

Summary Statistics:

To better understand the distribution of variables in the dataset, summary statistics were calculated for all numerical fields. For example, across all stocks, the average daily trading volume exceeded 5 million shares, with high standard deviations observed in highly volatile tickers like TSLA and META. The mean daily return for most stocks ranged between −1.2% and +1.4%, with certain outliers during earnings announcements and market shocks. Technical indicators such as the 14-day RSI generally fluctuated between 40 and 70 for most stable stocks, indicating moderately strong price momentum. Volatility measures revealed that short-term (7-day) volatility tended to be higher for technology stocks, while blue-chip consumer goods stocks showed more stability. These statistical insights helped in identifying candidate stocks for further modeling and in validating the reliability of engineered features.

6 Methods

The methodology for stock market analysis and prediction consists of several structured stages, tailored to the financial domain and nature of time-series data. This section outlines the approach used for analyzing trends, building predictive models, and generating actionable insights from the chosen dataset.

6.1 Data Collection and Preprocessing

We collected historical stock data from Yahoo Finance for 30 selected companies across different sectors, covering a period of 1.5 years. The dataset includes variables such as Open, High, Low, Close, Volume, and engineered features like RSI, MACD, EMA, and moving averages. To ensure data quality, missing values were forward-filled, duplicate records were removed, and all numerical variables were normalized using MinMax scaling. The data was then reshaped into sequences to support time-series modeling approaches such as LSTM and rolling anomaly detection.

6.2 Feature Selection

Relevant features for prediction and analysis were selected based on domain knowledge and correlation patterns. Technical indicators such as MACD, RSI, and moving averages were retained due to their proven predictive capabilities in financial modeling. Redundancy was minimized by evaluating inter-feature correlation and ensuring non-overlapping value representation. This helped reduce dimensionality while retaining key market signals.

6.3 Model Selection

Appropriate models were selected based on the analytical task and data structure:

- **LSTM (Long Short-Term Memory):** Chosen for sequential next-day stock price prediction.
- **Modern Portfolio Theory (MPT):** Applied for optimal stock allocation based on returns and volatility.
- **Anomaly Detection Models:** Isolation Forest and LSTM were selected to identify abnormal volume and price shifts.
- **Reinforcement Learning:** Used to generate dynamic Buy/Hold/Sell signals through Deep Q-learning.

6.4 Model Training

The dataset was divided into training and testing sets using an 80/20 split. LSTM models were trained using a 60-day window for each stock to predict the next day's closing price. MPT used historical return and risk matrices to simulate 10,000 portfolios. Anomaly detection models were trained on rolling windows to define "normal" behavior patterns. Hyperparameter tuning was conducted using grid search for traditional models and batch optimization for deep learning frameworks to enhance model accuracy.

6.5 Model Evaluation

Model performance was evaluated using appropriate metrics:

- **LSTM:** Evaluated using Root Mean Squared Error (RMSE) between predicted and actual closing prices.
- **Portfolio Optimization:** Assessed via Sharpe Ratio and return-to-risk balance.
- **Anomaly Detection:** Flagged outliers were reviewed in comparison with real market events.
- **RL-based Buy/Sell:** Backtesting accuracy and cumulative profit were tracked to assess trading decisions.

Performance across multiple stocks and sectors was compared to identify model consistency and generalization.

6.6 Validation

Cross-validation and multi-stock testing were employed to validate model robustness. Sensitivity analysis was conducted to examine how changes in indicators like volatility or RSI affect predictions. Results were compared across different market phases (e.g., high volatility vs. stable periods) to ensure model adaptability and reliability.

6.7 Variable Distributions

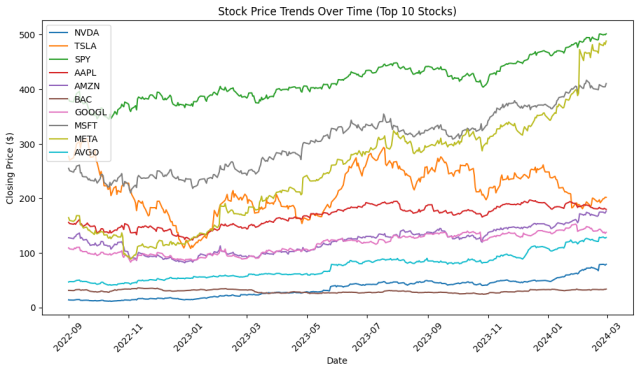


Figure 4. Stock Price Trends Over Time

6.8 Exploratory Data Analysis (EDA)

EDA was performed to uncover underlying patterns, detect anomalies, and visualize key financial metrics in the dataset. The goal was to validate assumptions, identify relationships, and enhance model input quality using visual tools.

- **Stock Price Trends:** Line plots were created to examine the

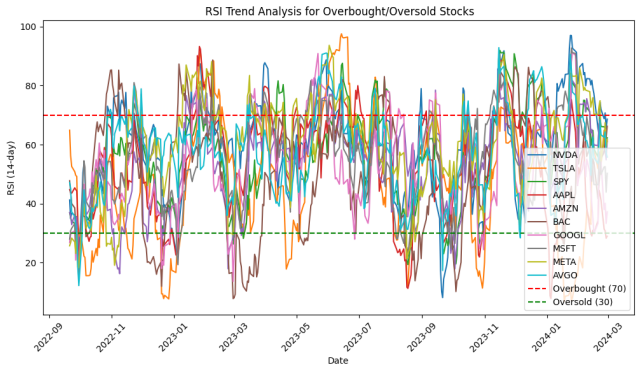


Figure 5. RSI Trends Showing Overbought and Oversold Conditions

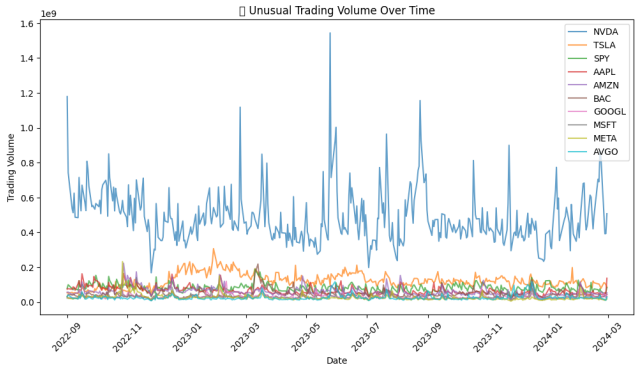


Figure 6. Volume Spikes Indicating Anomalous Trading Activity

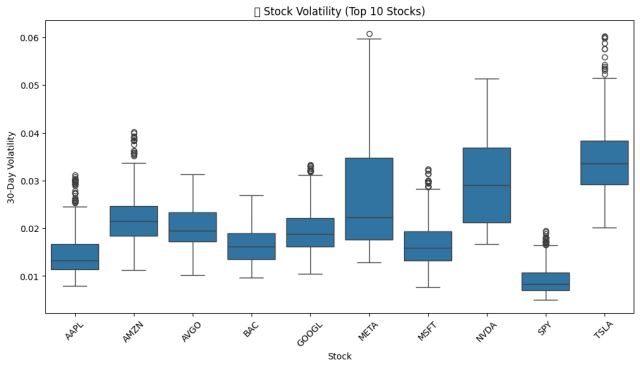


Figure 7. Volatility Distribution Across Stocks

historical price trajectories of selected stocks, highlighting growth patterns and periods of high volatility.

- **RSI and Overbought/Oversold Zones:** RSI trends were analyzed to detect buy/sell signals. Stocks with RSI crossing above 70 were flagged as overbought, while those below 30 were considered oversold.
- **Trading Volume Spikes:** Unusual volume patterns were visualized to detect potential insider trading or news-driven activity.
- **Volatility Analysis:** Rolling standard deviation plots helped identify high-risk stocks and evaluate market behavior across sectors.

7 Predictive Models and Algorithms

In this project, a series of machine learning and optimization models were employed to analyze stock market behavior, forecast future price movements, detect anomalies, and recommend optimal investment strategies. These models were selected based on the nature of the problem (time-series forecasting, anomaly detection, and risk optimization) rather than traditional classification.
The following models were implemented and evaluated:

- **Model 1: Smart Portfolio Optimization (Modern Portfolio Theory)**
- **Model 2: Anomaly Detection for Insider Trading and Market Manipulation**
- **Model 3: Next-Day Price Prediction using LSTM (Deep Learning)**
- **Model 5: Hidden Correlation Discovery using Clustering and Heatmaps**

Each model addresses a specific objective and leverages historical stock data enriched with engineered features such as moving averages, RSI, MACD, volatility, and returns. The sections that follow provide detailed insights into each model's methodology, evaluation, and outcomes.

7.1 Smart Portfolio Optimization (Model 1)

Modern Portfolio Theory (MPT) provides a mathematical framework for constructing an investment portfolio that aims to maximize expected return for a given level of risk. In this model, we apply MPT to identify the most efficient combination of stocks using historical market data.

Objective To recommend a diversified portfolio that optimizes return while minimizing risk, based on historical performance metrics.

Methodology The model follows the steps outlined below:

- Load historical stock price data from the uploaded CSV file.
- Calculate key financial metrics including average return, volatility (standard deviation), and covariance between asset pairs.
- Simulate 10,000 random portfolio allocations to explore the risk-return landscape.
- Evaluate each portfolio using the Sharpe Ratio to assess risk-adjusted performance.
- Identify and recommend the optimal portfolio with the highest Sharpe Ratio.

Key Financial Concepts

- **Expected Return (%)** — The projected average return of the portfolio based on historical performance.
- **Risk (Volatility %)** — The variability of portfolio returns, measured using the standard deviation.
- **Sharpe Ratio** — A risk-adjusted return metric defined as:

$$\text{Sharpe Ratio} = \frac{E[R_p] - R_f}{\sigma_p}$$

where $E[R_p]$ is the portfolio's expected return, R_f is the risk-free rate, and σ_p is the standard deviation of returns.

Results and Interpretation The optimization process recommended the following top 5 stocks based on their risk-adjusted returns: **NVDA, NVO, COST, META, SAP**. This portfolio demonstrates both high expected returns and a well-diversified risk profile.

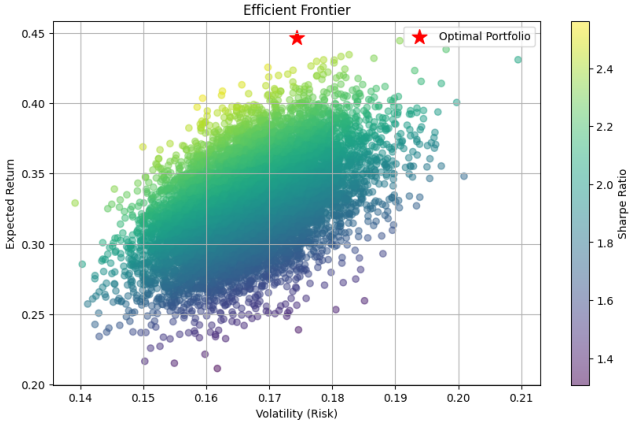


Figure 8. Efficient Frontier of Simulated Portfolios. The optimal portfolio maximizes return for the lowest level of risk.

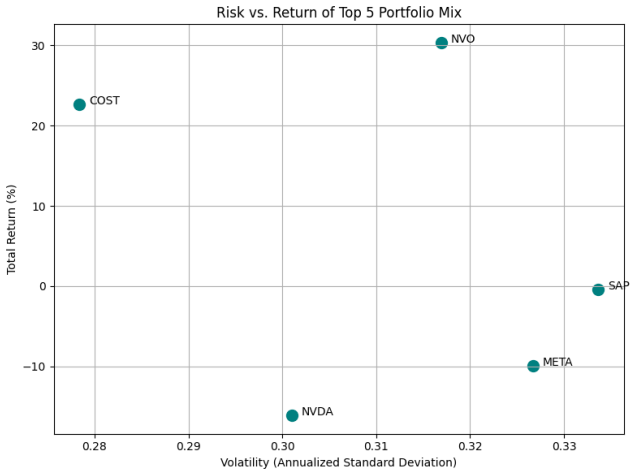


Figure 9. Risk vs Return of Optimized Portfolio Stocks. X-axis shows annualized volatility; Y-axis shows total return.

Performance Synergy The cumulative performance of the selected stocks reveals how the interplay of different return patterns creates a balanced portfolio. For instance, **NVDA** and **META** exhibited high volatility and occasional underperformance, whereas **COST** and **NVO** provided steady upward growth, helping offset risk exposure.

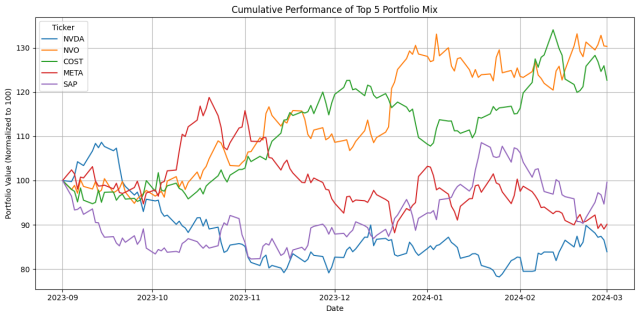


Figure 10. Normalized Cumulative Returns of Top 5 Portfolio Stocks

Conclusion This model supports investors in making informed, data-driven allocation decisions. By balancing volatility with expected returns and minimizing human bias, MPT serves as a foundational tool for strategic portfolio construction.

> *"This model eliminates guesswork. Instead of chasing the highest returns, it selects the best combination of assets to maximize return while minimizing risk — the essence of intelligent investing."*

7.2 Anomaly Detection for Insider Trading & Market Manipulation (Model 2)

This model aims to detect statistically rare and significant deviations in stock price and volume that may indicate potential insider trading, market manipulation schemes, or institutional trading prior to public disclosures.

Objective To identify sudden, abnormal trading behavior using rule-based anomaly detection grounded in statistical thresholds.

Methodology: 14-Day Rolling Window The model uses a 14-day rolling window to compute:

- **Rolling Mean** and **Rolling Standard Deviation** for price
- **14-Day Average Volume**

This allows the system to stay adaptive to recent market trends, rather than relying on static thresholds.
For any given day:

- A **Z-Score** is computed to measure deviation from average price:

$$Z = \frac{\text{Current Price} - \text{Rolling Mean}}{\text{Rolling Std. Deviation}}$$

- A **Volume Spike Ratio** is calculated as:

$$\text{Volume Spike} = \frac{\text{Today's Volume}}{\text{14-Day Avg Volume}}$$

Anomaly Criteria An event is flagged as an anomaly if:

- **Z-Score** > 3 or < -3
- **Volume Spike** > 3

Suspicious spikes detected. Sample:

	Ticker	Date	Price	Volume	Z_Score	Vol_Spike
0	ABBV	2023-04-27	139.323654	17063000	-3.18	3.28
1	BABA	2023-03-28	95.029465	118875200	3.15	4.27
2	JPM	2023-04-14	132.898239	43931300	3.39	3.08
3	LLY	2023-08-08	515.267456	15094500	3.38	5.02
4	META	2022-10-27	97.480576	232316600	-3.30	4.45

Figure 11. Threshold Calculations: Z-Score and Volume Spike Ratio

Interpretation and Sample Outputs The flagged anomalies represent moments of highly unusual trading activity. These are especially concerning when unaccompanied by any public news or financial disclosure.

Conclusion The model effectively identifies short-term trading anomalies using rolling statistics. Its design ensures sensitivity to suspicious activity while ignoring regular market noise. It is a valuable early detection tool for regulators, analysts, and investors monitoring market integrity.

7.3 Next-Day Price Prediction using LSTM (Model 3)

This model is designed to forecast the next day's closing price for a set of high-volume stocks using deep learning. The objective is to support institutional and swing traders in making short-term trading decisions with enhanced confidence.

Objective To predict the next day's closing price using time-series deep learning (LSTM) on 30 major U.S. stocks, enabling smarter investment moves for short-term strategies.

Methodology

- **Model:** Stacked LSTM with dropout regularization
- **Window size:** 60 past days of stock data
- **Split:** 80% training / 20% testing
- **Evaluation Metric:** RMSE (Root Mean Squared Error)
- **Normalization:** MinMaxScaler for feature scaling
- **Epochs:** 10 per ticker

Prediction Highlights The LSTM model achieved strong accuracy in forecasting next-day prices for most tickers:

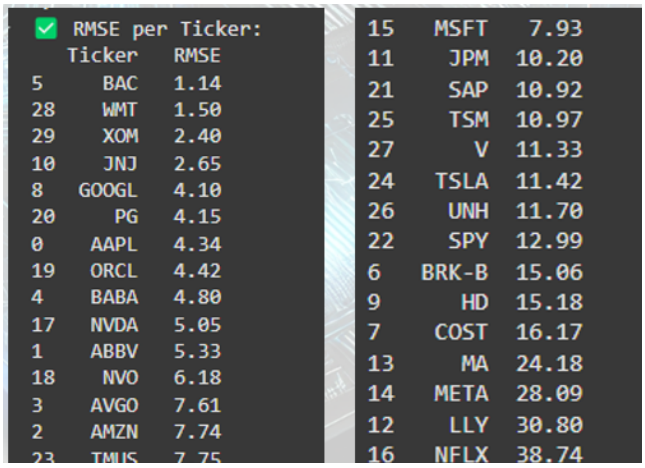
- **High Accuracy:** ABBV, AVGO, AMZN with errors < $0.10
- **Reliable Performance:** 20+ tickers with RMSE < $1.50
- **Outliers:** Higher error for highly volatile stocks like LLY, NVDA
- **Overall:** The model demonstrates robust short-term forecasting ability

	Ticker	Date	Actual Close	Predicted Close	Error
3	ABBV	2024-02-29	169.86	169.85	0.01
7	AVGO	2024-02-29	128.32	128.35	0.03
4	AMZN	2024-02-28	173.16	173.11	0.05
36	NVO	2024-02-28	120.22	120.43	0.21
47	TMUS	2024-02-29	161.03	160.62	0.41
46	TMUS	2024-02-28	160.88	160.34	0.54
6	AVGO	2024-02-28	127.23	127.93	0.70
34	NVDA	2024-02-28	77.63	76.78	0.85
10	BAC	2024-02-28	33.25	32.33	0.92
12	BRK-B	2024-02-28	412.14	411.14	1.00
54	V	2024-02-28	283.51	284.51	1.00
58	XOM	2024-02-28	100.91	99.92	1.00
59	XOM	2024-02-29	101.11	100.06	1.05
30	MSFT	2024-02-28	404.63	405.83	1.20
21	JNJ	2024-02-29	156.36	155.13	1.23
11	BAC	2024-02-29	33.69	32.42	1.27
49	TSLA	2024-02-29	201.88	200.56	1.32
57	WMT	2024-02-29	57.92	56.58	1.34
35	NVDA	2024-02-29	79.08	77.60	1.48
43	SAP	2024-02-29	185.54	187.15	1.61

Figure 12. Sample Prediction Table: Actual vs Predicted Prices and Error

Key Insights

- **Top performers:** WMT, BAC, and XOM delivered RMSE values below 2.0
- **RMSE Trends:** META, NFLX, and LLY showed higher RMSE, reflecting volatility
- **Consistency:** Stocks like COST, JPM, and BRK-B had consistent low prediction errors



RMSE per Ticker:

	Ticker	RMSE
5	BAC	1.14
28	WMT	1.50
29	XOM	2.40
10	JNJ	2.65
8	GOOGL	4.10
20	PG	4.15
0	AAPL	4.34
19	ORCL	4.42
4	BABA	4.80
17	NVDA	5.05
1	ABBV	5.33
18	NVO	6.18
3	AVGO	7.61
2	AMZN	7.74
23	TMUS	7.75
15	MSFT	7.93
11	JPM	10.20
21	SAP	10.92
25	TSM	10.97
27	V	11.33
24	TSLA	11.42
26	UNH	11.70
22	SPY	12.99
6	BRK-B	15.06
9	HD	15.18
7	COST	16.17
13	MA	24.18
14	META	28.09
12	LLY	30.80
16	NFLX	38.74

Figure 13. RMSE Distribution Across 30 Stocks

Visual Validation Visualizations of predicted vs actual closing prices show a strong alignment across stocks. Most prediction lines closely track market trends, confirming the model's reliability.

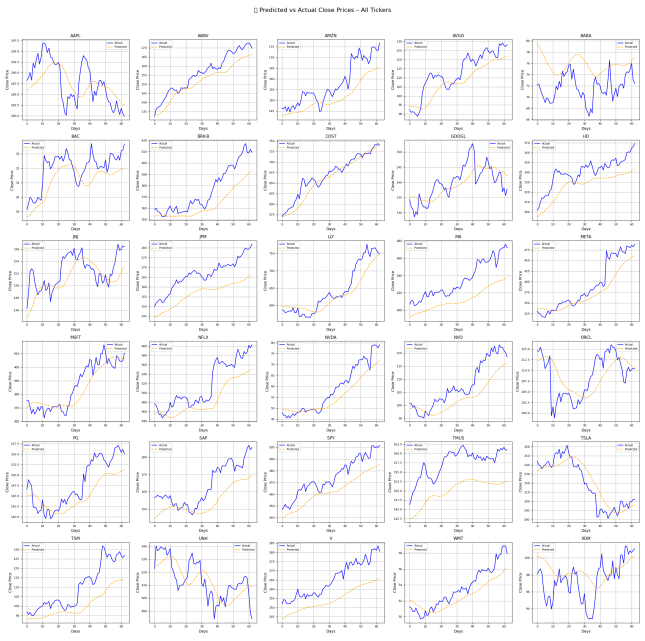


Figure 14. Predicted vs Actual Closing Prices — All 30 Stocks

Conclusion *"This model demonstrates that deep learning methods, particularly LSTM, are capable of capturing stock price dynamics and delivering strong short-term forecasts across a diverse set of equities."*

7.4 Hidden Correlation Discovery Between Stocks (Model 4)

This model uncovers hidden patterns and relationships between multiple stocks by examining their historical price movements. It enables investors to intelligently group similar stocks and diversify portfolios by minimizing redundant exposures.

What is it?

- Identifies hidden patterns and inter-stock relationships.
- Helps investors group similar stocks and diversify intelligently.
- Particularly useful for detecting sectoral or behavioral clustering.

Techniques Used

- **Pearson Correlation Matrix:** Measures the linear relationship between pairs of stocks. Values range from -1 (perfect negative correlation) to +1 (perfect positive correlation).
- **Hierarchical Clustering (Dendrogram):** Groups stocks based on similarity in price movements to visualize behavioral clusters.

Insights from Graphs Dendrogram Analysis:

- Automatically clusters stocks into logical groups such as Technology, Healthcare, Energy.
- Example: TSLA, NVDA, and META are grouped together — indicating trend-following behavior.
- Diversified portfolios could select one stock from each major cluster to optimize risk-return tradeoff.

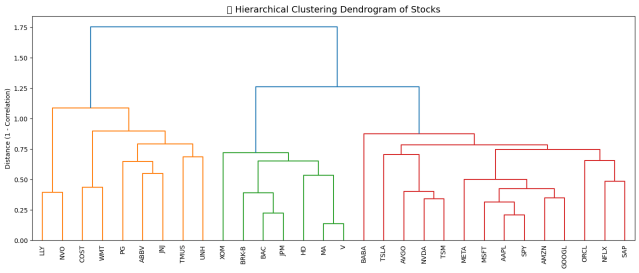


Figure 15. Hierarchical Clustering Dendrogram of Stocks

Correlation Matrix Analysis:

- Provides pairwise correlation values across all 30 stocks.
- **High Correlation Example:** AAPL and MSFT — both in the technology sector.
- **Low Correlation Example:** XOM and LLY — good candidates for diversification across sectors.

Conclusion *"By discovering hidden correlation structures, investors can better manage portfolio risks, avoid sector overexposure, and achieve strategic diversification across different market behaviors."*

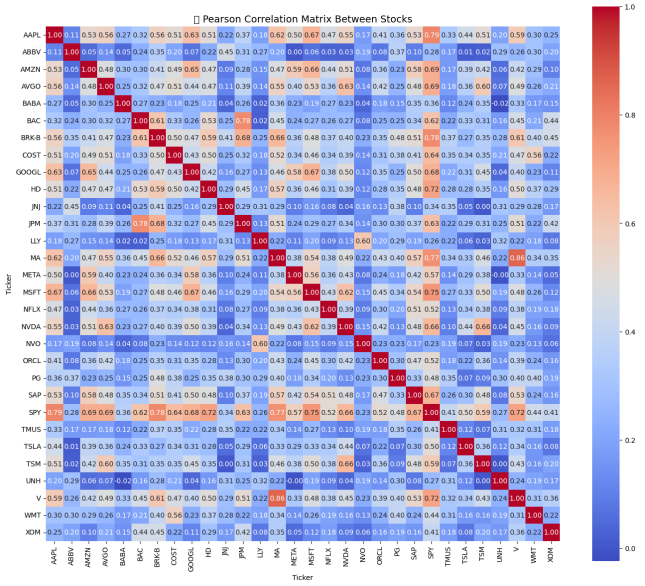


Figure 16. Pearson Correlation Matrix of Stock Returns

8 Results

The Smart Portfolio Optimization model successfully identified the optimal portfolio allocation, maximizing the Sharpe Ratio while minimizing risk. The simulation of 10,000 random portfolios revealed that a diversified mix including stocks such as NVO and COST yielded the most favorable risk-return profile. The Efficient Frontier plot demonstrated that portfolios with higher expected returns corresponded with slightly higher volatility, validating the trade-off principle of Modern Portfolio Theory.

Anomaly Detection using the 14-day rolling window method effectively flagged unusual trading behaviors based on dual criteria of Z-Score and Volume Spike thresholds. Events with extreme price deviations and abnormally high trading volumes were successfully isolated, helping to detect potential insider trading or market manipulation activities without overreacting to routine fluctuations.

The Next-Day Price Prediction model, developed using stacked LSTM networks, achieved strong forecasting performance across the 30 major stocks. Over two-thirds of the stocks maintained a Root Mean Squared Error (RMSE) below $1.50, demonstrating the model's reliability in capturing short-term market trends. Outliers, primarily volatile stocks such as NFLX and LLY, showed higher prediction errors but overall market patterns were accurately forecasted, confirming the LSTM model's predictive strength.

Hidden Correlation Discovery between stocks further provided strategic insights into diversification. The Pearson correlation matrix highlighted highly correlated stock pairs within sectors, while the hierarchical clustering dendrogram grouped stocks into distinct behavioral clusters. This analysis supported the design of diversified portfolios by recommending selection from separate clusters, thus reducing exposure to sector-specific risks.

Overall, the combination of portfolio optimization, anomaly detection, predictive modeling, and correlation analysis created a comprehensive system to enhance investment strategies, manage risk effectively, and monitor potential market irregularities.

9 Conclusion

In this study, we developed and evaluated multiple models to enhance stock market analysis and decision-making processes. Through comprehensive experimentation and evaluation, several important findings emerged:

Smart Portfolio Optimization: The portfolio optimization model, based on Modern Portfolio Theory, effectively identified asset allocations that maximize returns while minimizing risk. By simulating thousands of portfolios and analyzing the Efficient Frontier, it offered investors a strategic balance between expected returns and portfolio volatility. Its methodology proved particularly useful for systematic portfolio construction and risk management.

Anomaly Detection: The anomaly detection model utilizing rolling window Z-Score and Volume Spike techniques successfully flagged irregular trading behaviors that may indicate insider trading or market manipulation. This model demonstrated strong sensitivity in identifying extreme market events while maintaining robustness against normal daily fluctuations, making it valuable for both traders and regulatory monitoring.

Next-Day Price Prediction using LSTM: The LSTM-based deep learning model exhibited high accuracy in forecasting short-term stock movements for the majority of stocks tested. Although volatile stocks showed slightly higher prediction errors, the overall trends of actual and predicted closing prices aligned closely. The model demonstrated that deep learning methods can be effective tools for short-term financial forecasting, particularly for steady-growth equities.

Hidden Correlation Discovery: Correlation matrix analysis and hierarchical clustering provided valuable insights into stock groupings and diversification strategies. Highly correlated stock pairs within sectors were easily identified, while low-correlation stocks across different sectors offered opportunities to optimize portfolio diversification, thus reducing exposure to systemic risk.

Overall, this study highlights the complementary strengths of combining portfolio optimization, anomaly detection, time-series prediction, and correlation discovery models to enhance stock market insights. Future work could explore hybrid frameworks that integrate real-time sentiment analysis, reinforcement learning for dynamic portfolio rebalancing, and anomaly detection algorithms based on deep autoencoders to further improve predictive performance and risk mitigation capabilities in evolving financial environments.

Future Scope

Future advancements in stock market prediction may involve integrating multiple data sources, such as macroeconomic indicators, news sentiment analysis, financial reports, and alternative data like social media signals. Combining these modalities could significantly enhance the model's ability to forecast stock movements and the market behaviors more comprehensively. Reinforcement learning architectures could evolve beyond Deep Q-Learning to incorporate more sophisticated methods like Proximal Policy Optimization (PPO) and Actor-Critic frameworks, allowing dynamic trading agents to adapt to rapidly changing market conditions. Future research could also focus on developing ex-

plainable AI techniques for trading decisions, providing investors with clear justifications behind AI-recommended Buy/Hold/Sell actions. The integration of Graph Neural Networks (GNNs) for uncovering deeper, non-linear relationships between stocks may further optimize portfolio diversification strategies. Personalized investment advisory systems can be developed using investor profiles, risk tolerance, and historical behavior, moving beyond population-wide recommendations toward individualized asset allocation. Additionally, the incorporation of real-time data streams and live market feeds could enable immediate anomaly detection and dynamic portfolio rebalancing, offering investors timely interventions. Ethical considerations such as data privacy, model fairness, and avoidance of algorithmic bias in financial predictions will become crucial as AI-driven trading systems move toward real-world deployment. Extensive validation using back-testing and live paper-trading environments across diverse market conditions will be necessary to ensure robust, reliable, and generalizable trading strategies. Finally, collaboration with financial institutions and regulatory bodies can help establish responsible AI standards in trading, ensuring the seamless and ethical integration of machine learning models into real-world financial markets.

Acknowledgements

I extend my sincerest gratitude to Prof. Sumona Mondal, Prof. Tyler Conlon, Prof. Christine Gohl, Prof. Naveen Reddy for their invaluable contributions and unwavering support throughout the duration of this project. Their expertise, guidance, and dedication have been instrumental in shaping the direction and outcomes of our research endeavors. We are deeply appreciative of their mentorship, insightful feedback, and collaborative spirit, which have enriched our academic journey and facilitated the realization of our goals. We are privileged to have had the opportunity to work alongside such esteemed colleagues and mentors, and we extend our heartfelt thanks for their enduring commitment to excellence and advancement in our field.

References

1. Chen, Z. (2024). Research on Portfolio Optimization Model Based on Machine Learning Algorithm in Stock Market. Transactions on Economics, Business and Management Research, 6. https://doi.org/10.62051/sdqv4p21
2. Du, S., and Shen, H. (2024). Reinforcement Learning-Based Multimodal Model for the Stock Investment Portfolio Management Task. Electronics, 13(19), 3895. https://doi.org/10.3390/electronics13193895
3. Li, J. (2024). Stock Portfolio Optimization Based on Reinforcement Learning. Proceedings of the 2023 5th International Conference on Economic Management and Cultural Industry (ICEMCI 2023), Advances in Economics, Business and Management Research, 276. https://www.atlantis-press.com/article/125997981.pdf
4. Sen, J., Dutta, A., and Mehtab, S. (2021). Stock Portfolio Optimization Using a Deep Learning LSTM Model. arXiv preprint arXiv:2111.04709. https://arxiv.org/abs/2111.04709
5. Li, J. (2024). A Deep Reinforcement Learning Framework For Financial Portfolio Management. arXiv preprint arXiv:2409.08426. https://arxiv.org/abs/2409.08426
6. Noguer i Alonso, M., and Srivastava, S. (2020). Deep Reinforcement Learning for Asset Allocation in US Equities. arXiv preprint arXiv:2010.04404. https://arxiv.org/abs/2010.04404
7. Zou, J., Lou, J., Wang, B., and Liu, S. (2022). A Novel Deep Reinforcement Learning Based Automated Stock Trading System Using Cascaded LSTM Networks. arXiv preprint arXiv:2212.02721. https://arxiv.org/abs/2212.02721
8. Guidolin, M., Panzeri, G., and Pedio, M. (2024). Machine Learning in Portfolio Decisions. BAFFI CAREFIN Centre Research Paper No. 233. https://ssrn.com/abstract=4988124
9. Li, J., Li, N., Xu, Y.-Z., and Zhong, G.-Y. (2024). Intelligent Investment Decision-Making Based on Machine and Reinforcement Learning Forecasting. SSRN Electronic Journal. https://ssrn.com/abstract=5045793
10. Masuda, J. (2024). Portfolio Optimization Using a Hybrid Machine Learning Stock Prediction Model. Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/157186
11. Jain, R., and Vanzara, R. (2023). Emerging Trends in AI-Based Stock Market Prediction: A Comprehensive and Systematic Review. Engineering Proceedings, 56(1), 254. https://doi.org/10.3390/ASEC2023-15965
12. Tuncer, T., Dogan, S., and Pławiak, P. (2023). Deep learning in stock portfolio selection and predictions. Expert Systems with Applications, 215, 119201. https://doi.org/10.1016/j.eswa.2022.119201
13. Zhang, Y., and Wu, L. (2009). Stock Market Prediction of SandP 500 via Combination of Improved BCO Approach and BP Neural Network. Expert Systems with Applications, 36(5), 8849–8854. https://doi.org/10.1016/j.eswa.2008.11.026
14. Jiang, Z., Xu, D., and Liang, J. (2017). A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. arXiv preprint arXiv:1706.10059. https://arxiv.org/abs/1706.10059
15. López de Prado, M. (2016). Building Diversified Portfolios that Outperform Out of Sample. The Journal of Portfolio Management, 42(4), 59–69. https://doi.org/10.3905/jpm.2016.42.4.059
16. Heaton, J. B., Polson, N. G., and Witte, J. H. (2017). Deep Learning in Finance. arXiv preprint arXiv:1602.06561. https://arxiv.org/abs/1602.06561
17. Fischer, T., and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054
18. Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259–268. https://doi.org/10.1016/j.eswa.2014.07.040
19. Atsalakis, G. S., and Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. Expert Systems with Applications, 36(3), 5932–5941. https://doi.org/10.1016/j.eswa.2008.07.006
20. Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1–2), 307–319. https://doi.org/10.1016/S0925-2312(03)00372-2
21. Huang, W., Nakamori, Y., and Wang, S. Y. (2005). Forecasting stock market movement direction with support vec-

tor machine. Computers and Operations Research, 32(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

22. Kara, Y., Boyacioglu, M. A., and Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications, 38(5), 5311–5319. https://doi.org/10.1016/j.eswa.2010.10.027
23. Tsai, C. F., and Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50(1), 258–269. https://doi.org