Intelligent Factor-Based Stock Selection System

Jingjing Peng

***Abstract*—This project presents an intelligent, data-driven stock selection and portfolio optimization system that integrates financial indicators, macroeconomic variables, and sentiment features extracted from SEC filings. The system applies supervised machine learning model XGBoost to predict stock movement and identify investment candidates based on confidence thresholds. A Markowitz Mean-Variance Optimization framework is then used to allocate capital across selected assets. The dataset spans five innovation-driven sectors and incorporates macro-level factors such as GDP, CPI, and unemployment rate, along with sentiment scores derived from 10-K and 10-Q filings using the Loughran-McDonald dictionary. Backtesting over the 2023–2024 period reveals that although the system demonstrates promising integration of structured and unstructured data, the resulting portfolios underperform benchmark strategies. Limitations are traced to insufficient diversification, weak return magnitude filtering, and high prediction variance. Future enhancements include formulation of the stock selection problem as a discrete optimization problem, expanded stock coverage, advanced textual embeddings, and reinforcement learning for dynamic allocation.**

***Index Terms*—Stock selection, portfolio optimization, machine learning, financial sentiment analysis, macroeconomic indicators, Markowitz optimization, XGBoost, SEC filings.**

I. Introduction

Financial markets are increasingly complex and dynamic, influenced by a wide array of sector-specific fundamentals, macroeconomic conditions, corporate performance metrics, and investor sentiment. Traditional methods of stock selection and portfolio construction often struggle to keep pace with the velocity and volume of information now available. In response, institutional investors have embraced algorithmic and data-driven strategies that combine machine learning, quantitative finance, and natural language processing (NLP) to uncover patterns and signals embedded in both structured and unstructured data sources [1]–[3].

Large financial institutions leverage vast datasets, advanced analytics, and high-performance computing infrastructure to drive informed investment decisions and manage risk. Retail investors and researchers alike are increasingly interested in whether similar techniques can be adapted using publicly available data and open-source tools [4], [5]. Understanding how these institutional-grade systems work—and how to recreate them at a smaller scale—is essential for developing competitive, personalized strategies in today's data-rich environment.

This project aims to develop an intelligent, quantitative investment system that integrates financial indicators, macroeconomic data, and text-based sentiment extracted from corporate disclosures. The primary goal is to outperform traditional equal-weight or passive investment strategies by leveraging machine learning for stock movement prediction and portfolio optimization. In particular, I use sentiment scores derived from SEC 10-K and 10-Q filings as unstructured signals to complement market and macroeconomic data.

The system is built in modular stages: data collection and preprocessing, predictive modeling, stock selection based on upward movement probabilities, and risk-aware portfolio optimization. I focus on five key sectors—AI Healthcare, Fintech, Clean Energy, Cloud and Big Data, and Semiconductors—which provide a diverse testing ground across innovation-driven and capital-intensive industries. To capture broader market dynamics, macroeconomic indicators such as Gross Domestic Product (GDP), the Consumer Price Index (CPI), the unemployment rate, and the federal funds rate are incorporated into the model.

This multifactor approach is designed to increase predictive accuracy, improve risk-adjusted performance, and adapt to volatile conditions such as the COVID-19 pandemic. By combining structured financial signals with document-level sentiment analysis, the system aims to bridge the gap between academic modeling and practical investment strategy.

II. Project Objectives and Workflow

This project aims to develop and evaluate an intelligent stock selection and portfolio optimization system, integrating predictive modeling, feature-driven stock selection, and quantitative portfolio allocation. The overarching goal is to outperform traditional equal-weight or passive investment strategies through the use of advanced analytics and machine learning.

The core objectives of the system are as follows:

- **Predictive Modeling:** Train machine learning models to classify whether stocks from key innovation-driven sectors—including AI Healthcare, Fintech, Clean Energy, Cloud and Big Data, and Semiconductors—are likely to rise, remain stable, or fall over a specified time horizon.
- **Stock Selection:** Identify stocks with strong upward potential based on model-predicted probabilities.
- **Portfolio Optimization:** Allocate capital using Markowitz Mean-Variance Optimization to maximize risk-adjusted returns (Sharpe ratio).
- **Performance Evaluation:** Benchmark system results against traditional investment strategies using historical backtesting.

To achieve these objectives, the system integrates multiple data sources—such as historical stock prices, macroeconomic indicators, and sentiment scores derived from SEC filings—into a unified modeling and decision-making framework.

Figure 1 illustrates the complete workflow for stock selection and optimization. The process begins with a broad stock universe, which is filtered based on sector relevance and data availability to form a pre-selected stock pool. Machine

learning models are then applied to classify price movement direction. Stocks with sufficiently high predicted upward probability are retained and passed into the portfolio optimization stage, where risk-return profiles are evaluated. The final portfolio is constructed by allocating optimal weights across the selected assets to maximize the Sharpe ratio under predefined constraints.

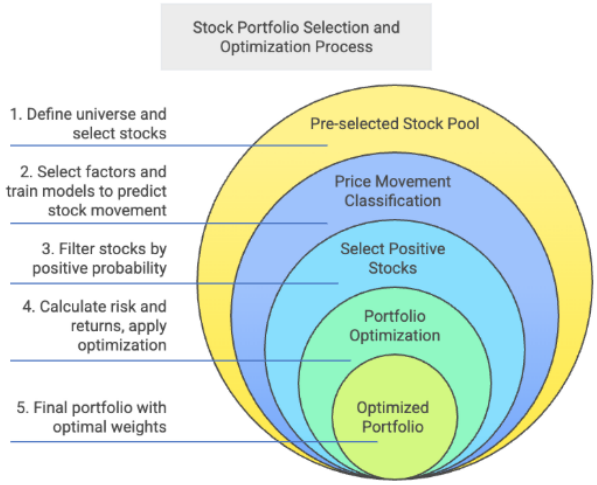

Fig. 1: End-to-End Stock Selection and Optimization Workflow.

III. Data Collection and Preprocessing

A. Data Sources

This study leverages multi-modal data from both financial markets and macroeconomic indicators to build a comprehensive investment analysis dataset.This study leverages multi-modal data from both financial markets and macroeconomic indicators to build a comprehensive investment analysis dataset. The input consists of:

- **Stock-level data**: Daily prices, trading volumes, returns, and volatility for 25 companies across 5 sectors—AI Healthcare, Fintech, Clean Energy, Cloud and Big Data, and Semiconductors. These features are commonly used in quantitative asset pricing and machine learning-based prediction models [1], [4].
- **Macroeconomic indicators**: Daily or monthly data on U.S. interest rates, GDP growth, unemployment rate, and the Purchasing Managers' Index (PMI). Such macroeconomic variables have long been recognized as important drivers of asset returns and are increasingly incorporated into systematic strategies [1], [5].
- **Textual financial reports**: SEC filings (10-K and 10-Q) from 2019 to 2024, with computed sentiment scores based on the Loughran-McDonald financial sentiment dictionary. Prior work has shown that financial text sentiment contains predictive power for stock returns and volatility [2], [3].

B. SEC Filing Retrieval

Company filings are retrieved using the `sec_edgar_downloader` package, which downloads 10-K and 10-Q reports from the SEC EDGAR database. Each company ticker was associated with one of the five selected sectors, and filings were collected for the period between 2019 and 2024.

C. Text Preprocessing

To ensure the quality and usability of textual data for sentiment analysis and language modeling, a multi-stage preprocessing pipeline was applied to the raw SEC filings:

1) **Header Removal**: XML and metadata headers (e.g., `<DOCUMENT>`, `<TEXT>`, `<FILER>`) were stripped using regular expressions.
2) **HTML Cleaning**: All HTML tags and encoded entities (e.g., ` `, `&quot;`) were removed or converted to plain text.
3) **Special Character Filtering**: Irregular or non-informative patterns such as GAAP tags (e.g., `us-gaap:Revenue`) and XBRL markup were filtered out.
4) **De-duplication**: Repetitive sequences and redundant boilerplate phrases commonly found in SEC documents were identified and eliminated.
5) **Normalization**: The final cleaned text was lowercased, stripped of extra whitespace, and filtered for length and relevance.

D. Sentiment Annotation

I implemented a custom version of the Loughran-McDonald sentiment classifier to annotate each cleaned SEC filing with sentiment-based features. The classifier leverages the Loughran-McDonald Master Dictionary, which is tailored specifically for financial language, to identify and count domain-specific terms associated with positive, negative, uncertainty, and litigious sentiment categories.

Sentiment signals were extracted from cleaned 10-K and 10-Q filings by computing the normalized frequency of positive and negative words. For each report, word counts were divided by total word count to generate standardized sentiment scores, ensuring comparability across documents of varying length and format.

As shown in Fig. 2, negative sentiment consistently dominates over positive sentiment in both 10-K and 10-Q reports. This asymmetry reflects the typically cautious tone adopted in financial disclosures, particularly in risk sections or forward-looking statements. The variability observed across report types may capture differences in reporting scope, regulatory requirements, or timing within the fiscal year. These sentiment scores were subsequently incorporated as input features for predictive modeling, offering additional signals linked to corporate outlook, risk exposure, and market sentiment. To further investigate sentiment variability across firms, Fig. 3 presents the distribution of positive sentiment scores for each company, separated by report type (10-K and 10-Q). While most

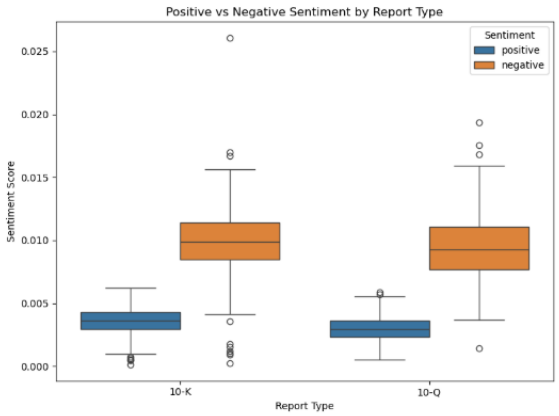

Fig. 2: Distribution of positive and negative sentiment scores from 10-K and 10-Q SEC filings, based on the Loughran-McDonald financial sentiment dictionary. Negative sentiment consistently dominates.

companies cluster around similar sentiment centers (roughly 0.003–0.004), the density and spread vary significantly. For instance, companies like GH and EXAS show narrow distributions skewed toward lower positivity, suggesting a consistently cautious or risk-averse tone. In contrast, TSLA and ENPH display broader distributions, indicating more variation in narrative tone across reports.

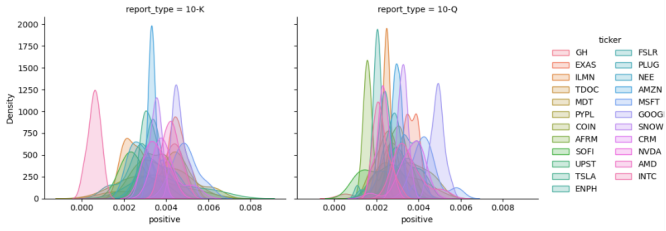

Fig. 3: Distribution of positive sentiment scores across companies by report type. Densities are computed separately for 10-K and 10-Q filings, revealing firm-specific variation in tone.

Interestingly, this variability in sentiment distributions aligns with model performance in several cases. Stocks such as TDOC, CRM, and TSLA—which exhibited broader and more distinguishable sentiment signals—also achieved higher F1 scores in the classification task (see Fig. 6). In contrast, stocks like GH and EXAS, which showed tightly clustered low positivity, corresponded with poor model performance. These observations suggest that sentiment diversity and richness may enhance model learnability, while consistently flat or low sentiment features may provide weaker predictive signals. This relationship highlights the potential of document-level sentiment analysis to support stock movement prediction, particularly when paired with complementary market and technical indicators.

E. Macroeconomic and Market Data Integration

Daily financial metrics—such as adjusted close price, return, and volatility—were collected for each stock using APIs like Yahoo Finance and Alpha Vantage. These were then aligned with macroeconomic indicators sourced from FRED (Federal Reserve Economic Data). All data streams were synchronized on a daily timeline and merged using date-index joins to ensure consistency for model training.

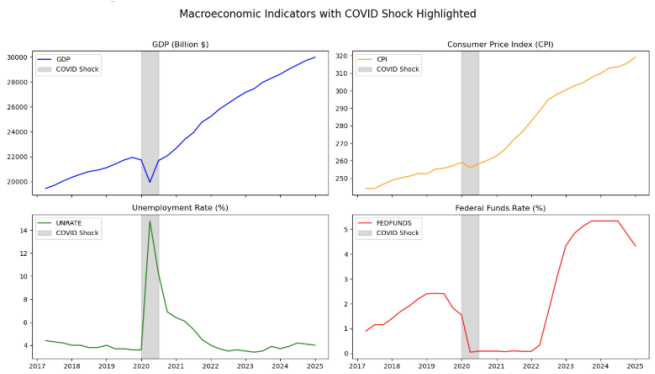

Fig. 4: Macroeconomic indicators with the COVID-19 shock highlighted. The shaded region marks the period of major economic disruption in 2020–2021.

Figure 4 illustrates four key macroeconomic indicators used as input features in the modeling framework. Gross Domestic Product (GDP) reflects overall economic growth, the Consumer Price Index (CPI) tracks inflation levels, the Unemployment Rate (UNRATE) signals labor market conditions, and the Federal Funds Rate (FEDFUNDS) represents central bank policy and the interest rate environment. The highlighted COVID-19 period reveals sharp disruptions across all indicators, illustrating the extent to which macroeconomic shocks can reshape market dynamics. Incorporating these variables enhances the model's ability to capture systemic risk and improve prediction robustness during volatile periods.

F. Final Dataset

The resulting dataset includes structured time-series features for each stock, macroeconomic variables, and document-level sentiment scores. This multi-source, multi-resolution dataset enables training and evaluation of financial models that incorporate both numerical and textual signals.

IV. Methodology

The proposed system consists of three core components: predictive modeling, stock selection, and portfolio optimization. These components are supported by a pipeline of data preprocessing and feature engineering that integrates structured financial data, macroeconomic indicators, and unstructured text data from regulatory filings.

A. Feature Engineering

For each stock, custom features such as daily return, volatility, and momentum were calculated using historical OHLCV

data. Additionally, sentiment signals were extracted from SEC 10-K and 10-Q filings using a custom Loughran-McDonald classifier, which counted domain-specific word frequencies in categories such as positive, negative, uncertainty, and litigious. Sentiment scores were aggregated by ticker and report type and incorporated as features for supervised learning.

B. Predictive Modeling

One supervised learning algorithms was employed to predict future stock movement.

XGBoost Classifier: A multi-class classifier is implemented to predict one-month forward movement. The label is defined as follows:

- `2`: Up (return $> 5\%$)
- `1`: Stable (return between -5% and 5%)
- `0`: Down (return $< -5\%$)

Probabilities from the XGBoost output are used to determine confidence. Stocks with predicted upward movement probabilities above 0.6 are selected for portfolio construction.

C. Stock Selection

Once models are trained and validated, prediction outputs are aggregated. For XGBoost, the average predicted probability of upward movement is computed per stock. Stocks with average upward probability exceeding 60% are shortlisted as investment candidates.

D. Portfolio Optimization

Selected stocks were passed into a portfolio optimizer using the classical Markowitz Mean-Variance Optimization (MVO) framework. The goal is to construct a portfolio that maximizes the Sharpe ratio—a widely used measure of risk-adjusted return—while satisfying realistic investment constraints.

The Sharpe ratio is defined as:

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

where $\mathbb{E}[R_p]$ is the expected portfolio return, R_f is the risk-free rate, and σ_p is the portfolio volatility.

The optimization problem is formally expressed as:

$$\max_{w} \quad \frac{w^\top (R - R_f)}{\sqrt{w^\top \Sigma w}}$$

where:

- w is the vector of portfolio weights,
- R is the vector of expected returns,
- R_f is the risk-free rate (scalar),
- Σ is the covariance matrix of asset returns.

Subject to the following constraints:

- $w_i \geq 0$ (no short-selling),
- $\sum_i w_i = 1$ (full capital allocation),
- Portfolio Volatility $< 10\%$ (risk cap).

As illustrated in Fig. 5, the process begins with defining the optimization objective, followed by selecting appropriate optimization algorithms and enforcing investment constraints.

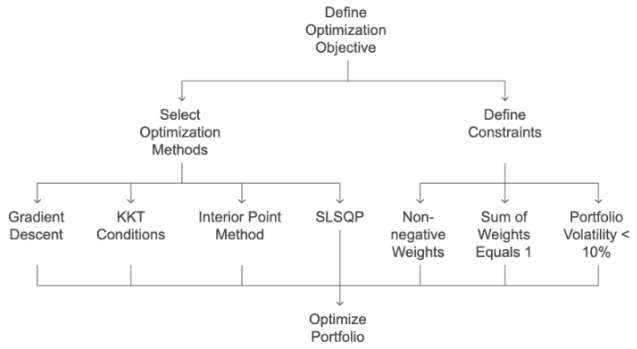

Fig. 5: Portfolio optimization workflow outlining the selection of methods and constraints used to maximize the Sharpe ratio.

In this project, I adopted the Sequential Least Squares Programming (SLSQP) algorithm, which efficiently handles nonlinear objectives and inequality constraints. The optimizer ensures the resulting portfolio is both theoretically optimal and practically viable.

E. Reinforcement Learning Extension

To enhance the adaptability of the system, a Proximal Policy Optimization (PPO)-based reinforcement learning agent is trained. The agent learns a dynamic asset allocation strategy through interaction with a simulated trading environment, which considers price data, position weights, and factor values. The environment rewards the agent based on portfolio returns after transaction costs.

F. Evaluation

The system is evaluated through historical backtesting. Key performance metrics include:

- **Annualized Return**
- **Sharpe Ratio**
- **Maximum Drawdown**
- **Sortino and Calmar Ratios**
- **Win Rate**

Performance is compared against market benchmarks (e.g., S&P 500) and traditional strategies (e.g., equal-weighted portfolios).

In the final dataset preparation, a time-based split was applied to separate the training and validation periods. Specifically, data from January 1, 2019 to December 31, 2022 was used for training and model fitting. The subsequent period, from January 1, 2023 to December 31, 2024, was reserved for validation and performance evaluation. This chronological split ensures the model is trained only on past data and tested on future data, simulating real-world forecasting. The training process was configured with a batch size of 64, a maximum of 100 epochs, and an early stopping criterion set to 10 rounds to prevent overfitting.

V. Results and Analysis

The proposed system was evaluated on a dataset of 25 stocks across five sectors, spanning multiple years of historical data and macroeconomic indicators. Both predictive accuracy and portfolio performance were assessed through statistical metrics and backtesting.

A. Prediction Performance

The classification models demonstrated varying performance across individual stocks, as measured by F1 score—a balanced metric that considers both precision and recall. Figure 6 summarizes the F1 scores for each stock.

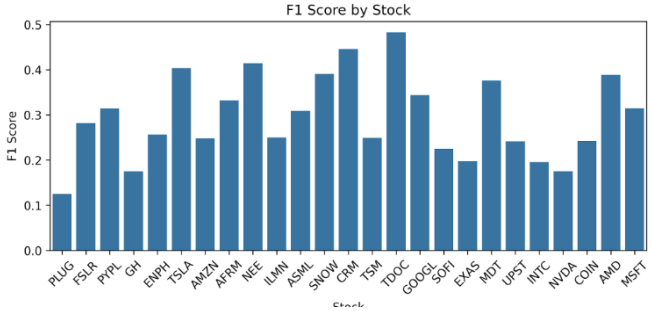

Fig. 6: F1 score by stock, illustrating the classification performance across different tickers. Higher scores indicate better balance between precision and recall.

Top-performing predictions (F1 $\approx$ 0.4–0.48) were achieved on TDOC, CRM, NEE, TSLA, and SNOW. These models consistently produced accurate predictions while maintaining low false positive and false negative rates.

Medium-performing predictions (F1 $\approx$ 0.3–0.4) included FSLR, PYPL, AFRM, GOOGL, MSFT, AMD, and ENPH. While reasonably effective, these models exhibited occasional misclassifications and may benefit from further tuning.

Low-performing predictions (F1 ¡ 0.2) were observed for PLUG, GH, EXAS, and NVDA. These results suggest that the current feature set may not provide strong predictive signals for these stocks, or that model complexity may need adjustment. Possible factors include data quality issues, sector-specific volatility, or insufficient separation between movement classes.

Overall, the results highlight considerable variability in predictive quality across stocks. High-performing models could be prioritized for investment decisions or further analysis, while low-performing cases warrant feature engineering, data augmentation, or alternative model structures. Improving multi-class classification for stock movement prediction remains a key challenge and an area for continued research and system enhancement.

B. Stock Selection Outcome

To identify promising candidates for portfolio construction, XGBoost classification probabilities were analyzed across all stocks. Stocks with an average predicted probability of upward movement greater than 60% were selected as investment candidates. The resulting shortlist included:

- GH (Guardant Health)
- PLUG (Plug Power)
- PYPL (PayPal)
- ENPH (Enphase Energy)
- FSLR (First Solar)
- SNOW (Snowflake)
- TSLA (Tesla)

As shown in Fig. 7, the top-performing candidates—such as GH, PLUG, and PYPL—exhibited consistently high average upward probabilities, ranging from approximately 64% to 95%. These stocks were then passed into the portfolio optimization stage for risk-return analysis and weight allocation.

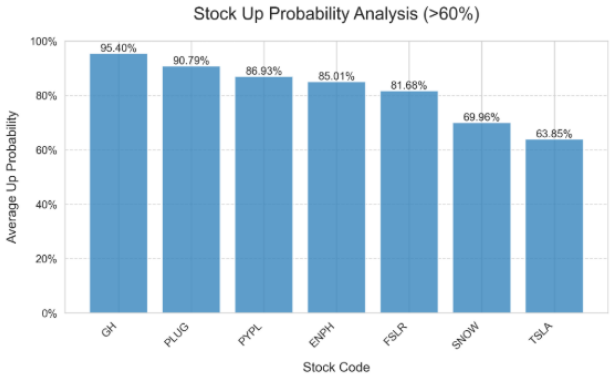

Fig. 7: Average upward probability for selected stocks exceeding the 60% threshold. These stocks were shortlisted for portfolio optimization.

C. Portfolio Optimization Results

The selected stocks were optimized using the Markowitz Mean-Variance Optimization (MVO) framework. However, the resulting portfolios failed to outperform traditional single-sector benchmarks. Across most portfolio configurations, the Sharpe ratios were negative, indicating suboptimal reward-to-risk performance. Several factors may have contributed to this outcome:

- A limited number of selected stocks, leading to insufficient diversification.
- A stock selection criterion based solely on the probability of upward movement, without consideration of return magnitude or volatility.
- A relatively high false-positive rate in price movement predictions.

Figure 8 illustrates the return, volatility, and Sharpe ratio across six industry portfolio configurations. Notably, all industries—including AI Healthcare, Fintech, Clean Energy, Cloud and Big Data, Semiconductors, and the customized portfolio—exhibited negative Sharpe ratios. This suggests either returns were consistently below the risk-free rate or the portfolios were exposed to high levels of risk with inadequate compensation. These patterns highlight the limitations of the current selection criteria, particularly the lack of focus on return magnitude and volatility profiles.

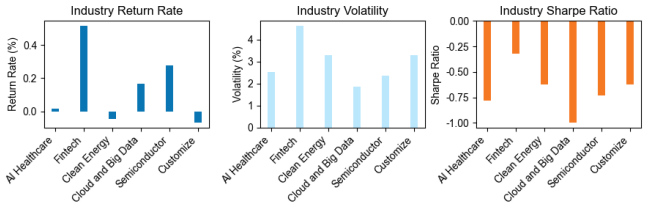

Fig. 8: Industry-level return rate, volatility, and Sharpe ratio for different portfolio configurations. While Fintech and Cloud sectors show higher returns, all sectors exhibit negative Sharpe ratios, indicating suboptimal reward-to-risk profiles.

To further analyze the impact of optimization methods on performance, I compared four algorithms—gradient descent, interior point method, KKT-based optimization, and SLSQP (via SciPy)—on sector-specific portfolios.

Fig. 9: Comparison of optimization methods for the Cloud and Big Data portfolio during the testing phase. While all methods yielded positive returns, differences in volatility and Sharpe ratio suggest that method selection impacts risk-adjusted performance.

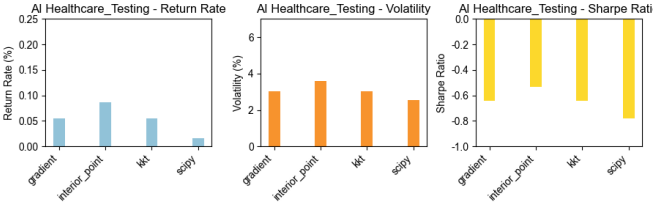

Fig. 10: Optimization method comparison for AI Healthcare portfolio in testing. Despite moderate returns, all methods yielded negative Sharpe ratios, indicating poor risk-adjusted performance in this sector.

Figures 9 and 10 demonstrate the variation in portfolio outcomes across optimization strategies. For the Cloud and Big Data sector (Fig. 9), all methods generated comparable returns but exhibited noticeable differences in volatility and Sharpe ratio, indicating that optimizer choice affects risk-adjusted outcomes. In contrast, for AI Healthcare (Fig. 10), all optimization methods failed to produce a positive Sharpe ratio, reinforcing the difficulty of extracting consistent alpha from this sector under current model constraints.

These findings indicate a need to refine the current stock selection strategy. While the system successfully identified stocks with a high probability of upward movement, it failed to account for the magnitude and consistency of those returns—critical factors in portfolio optimization. Moreover, the limited stock pool likely hindered effective diversification, reducing the system's ability to mitigate risk across sectors.

Future work will focus on enhancing the selection framework to incorporate not only directional trends but also return magnitude, volatility profiles, and inter-stock correlations. A more comprehensive and data-rich approach will support the construction of portfolios that offer better balance between return and risk, improving the system's ability to outperform conventional benchmarks.

D. Backtesting and Evaluation Metrics

Portfolio performance was evaluated using historical backtesting over the validation period (2023–2024). Table I summarizes key metrics—expected return, volatility, Sharpe ratio, and risk-adjusted return—for each industry-specific portfolio.

TABLE I: Industry Backtesting Summary (2023–2024)

Sector	Return (%)	Volatility (%)	Sharpe Ratio	Risk-Adj. Return
AI Healthcare	0.02	2.54	-0.78	0.007
Fintech	0.52	4.61	-0.32	0.112
Clean Energy	-0.05	3.29	-0.62	-0.014
Cloud & Big Data	0.17	1.84	-1.00	0.090
Semiconductor	0.28	2.36	-0.73	0.117

Across all portfolios, the Sharpe ratios were negative, indicating that the excess returns did not sufficiently compensate for volatility. Fintech and Semiconductor portfolios showed relatively better outcomes in terms of return and risk-adjusted metrics but still failed to exceed the performance of baseline benchmarks such as the S&P 500. The weak performance reflects limited diversification, low return magnitude, and possible model misclassifications.

E. Summary of Insights

Although the final portfolio strategies did not outperform standard market benchmarks, the system achieved several key milestones.

- Demonstrated integration of structured (technical, macroeconomic) and unstructured (textual sentiment) data in a unified modeling pipeline.
- Applied supervised learning models to forecast short- and mid-term stock movement, capturing directional signals with moderate accuracy.
- Constructed and evaluated optimized portfolios using Sharpe ratio maximization under real-world constraints.

These accomplishments provide a strong foundation for ongoing development. The results also highlighted gaps—such as insufficient return magnitude filtering, limited stock pool size, and high prediction noise—that must be addressed in future iterations.

VI. Conclusion and Future Work

This project developed and evaluated an intelligent, factor-based stock selection and portfolio optimization system focused on high-potential sectors such as AI Healthcare. The

framework integrated machine learning models, financial time-series, macroeconomic variables, and sentiment scores derived from SEC filings to support data-driven investment decision-making.

While the system did not outperform traditional market benchmarks, it served as a valuable research prototype. The modular pipeline—from feature extraction to model training and portfolio construction—demonstrated the viability of using open-source tools and publicly available data to simulate institutional-grade strategies at a smaller scale.

Future enhancements will target key limitations and expand the system's capabilities:

- **Enhanced stock selection:** Apply discrete optimization method in the stock selection.
- **Broader Stock Universe:** Include more sectors and tickers to improve diversification and risk spreading.
- **Feature Enrichment:** Incorporate more granular textual features (e.g., BERT-based embeddings, earnings call transcripts) and high-frequency market data.
- **Real-Time Responsiveness:** Integrate real-time news sentiment and social media trends using streaming data platforms.
- **Improved Filtering:** Adjust stock selection criteria to consider not only directional confidence but also expected return magnitude, risk, and factor exposure.
- **Reinforcement Learning Extension:** Train agents to dynamically rebalance portfolios and adapt strategies based on market regimes.

These future directions aim to enabling a more robust and adaptive framework for intelligent asset management. The code repository is available at: https://github.com/rspeggy/AI$_Stock_Selection/tree/main$.

REFERENCES

[1] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.

[2] S. Kogan, D. Levin, B. Routledge, J. Sagi, and N. Smith, "Predicting risk and return of equity portfolios using textual analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 272–280.

[3] R. Zhang, W. Li, X. Zhang, and Q. Xu, "Natural language processing for financial forecasting: A survey," *ACM Computing Surveys*, 2023, forthcoming.

[4] G. Feng, H. He, and N. G. Polson, "Deep learning for financial applications: A survey," *Journal of Financial Data Science*, vol. 1, no. 4, pp. 10–29, 2019.

[5] R. Li, Y. Zhang, and J. Zhang, "Can investor sentiment predict stock returns? a machine learning perspective," *Quantitative Finance*, vol. 22, no. 5, pp. 767–785, 2022.