



Stock Selection Using Machine Learning Based on Financial Ratios

by: Pei-Fen, Cheng-Han Gao and Shyan-Ming Yuan

Computer Science Department
National Yang Ming Chiao Tung University – Taiwan

Publisher: MDPI – 2023





Contents

1. Introduction
2. Methodology
3. Results
4. Challenges and Limitations
5. Dataset for Training My Model

Introduction

RF and FNN have better portfolio scores. The GRU and FinGAT models can capture temporal and spatial dependency. Therefore, this study chose these four models.

Random Forest (RF):

The paramount aim of employing Random Forest for stock selection revolves around forecasting stock performance and behavior, guided by a mosaic of characteristics and factors. Via an examination of historical data and patterns, this model can furnish insights into the potential value or trajectory of stocks.

Feedforward Neural Network (FNN):

FNN plays an invaluable role in stock selection by adeptly modeling and scrutinizing non-linear associations between inputs and outputs. Leveraging the backpropagation algorithm, the FNN can be trained to enhance its performance and elevate the precision of stock predictions and investment decisions.

Gated Recurrent Unit (GRU):

GRU, with its adept handling of sequential data, proves highly effective in analyzing historical stock data and uncovering influential patterns and trends that can impact future stock performance. As a result, GRU assumes a pivotal role in stock prediction models, contributing to the precision of stock selection for investment purposes.

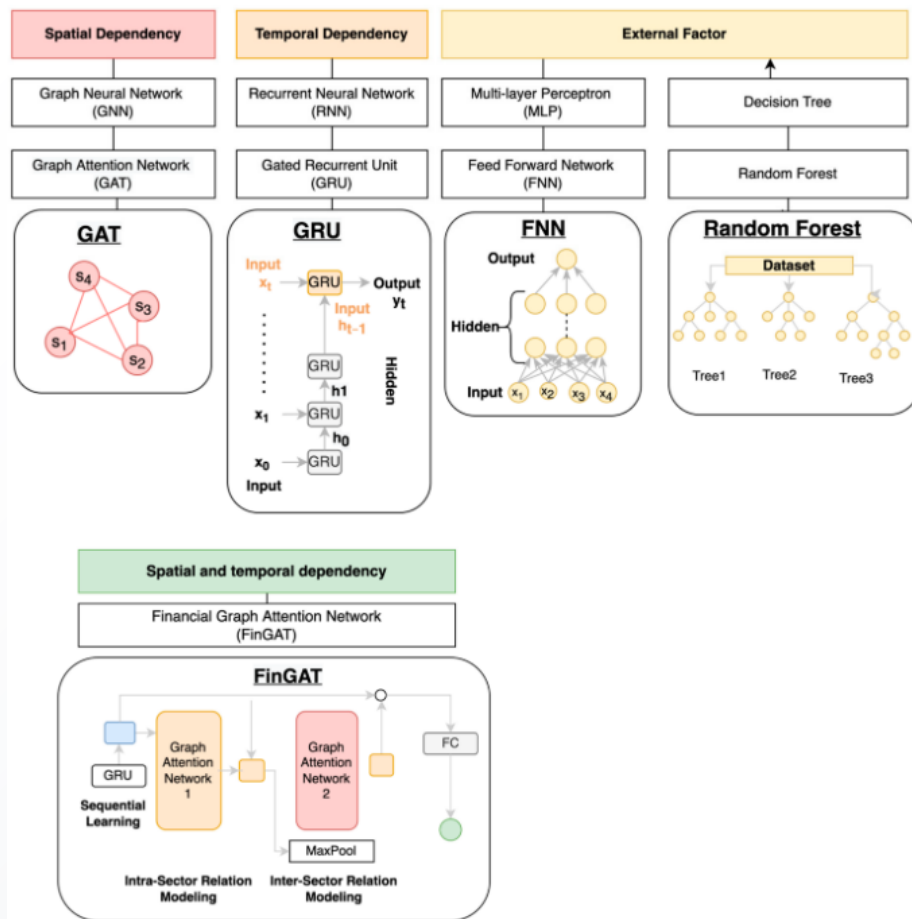
Graph Attention Network (GAT):

The Graph Attention Network (GAT) is a specialized type of Graph Neural Network (GNN), known for its effectiveness in analyzing the complex relationships found in stock markets.

Financial Graph Attention Network (FinGAT):

FinGAT stands as an advanced neural network meticulously designed for the precise prediction of stock returns. It effectively harnesses the power of graph attention mechanisms to grasp the intricate relationships between diverse stocks.

Figure 3. Deep learning models including spatial dependency: GNN [32] and GAT [31]; temporal dependency: RNN [33] and GRU [26]; external factor: MLP, FNN [7], and Random Forest [5,6]; and spatial and temporal dependency: FinGAT [9].





Methodology



Dataset



**Financial Ratios of 18 Ratios as
Attributes**



Moving Time Period



Evaluation Metrics



**Model Architecture for
Training/Validation/Test**



Dataset

Stock Pool

In the context of the Taiwan stock market, our stock pool consists of a fixed set of 97 stocks. These selections were made based on their capitalization rankings on the Taiwan Stock Exchange as of March 2022 and are immutable once established.

Benchmark Analysis

TW50 index, a stock market index comprising the top 50 companies listed on the Taiwan Stock Exchange, arranged according to market capitalization.

Time Period

The spanning from 2013Q1 to 2020Q2, encompassing a total of 27 quarters' worth of financial data.

Financial Ratios of 18 Ratios as Attributes

Using 18 financial ratios as input for models, which are the same as those used in research. The 18 financial ratios are listed, and the calculation formula is given in Table 1.

There are five main types of financial ratios:

- ☐ Liquidity ratios
- ☐ Leverage ratios
- ☐ Asset efficiency ratios
- ☐ Market value ratios
- ☐ Profitability ratios

Table 1. Financial ratios used for the training model.

Type	Ratio	Calculation
Liquidity Ratios	Current	$\frac{\text{Current assets}}{\text{Current liabilities}}$
Leverage Ratios	Debt to Equality	$\frac{\text{Long Term Debt}}{\text{Shareholders' Equality}}$
	Debt to Capital	$\frac{\text{Long Term Debt}}{\text{Shareholders' Equality} + \text{Long Term Debt}}$
Asset Efficiency Ratios	Asset Turnover	$\frac{\text{Sales}}{\text{Total Assets}}$
	Inventory Turnover	$\frac{\text{Sales}}{\text{Inventories}}$
	Receivable Turnover	$\frac{\text{Net Credit Sales}}{\text{Receivables}}$
	Days Sales in Receivable	$\frac{\text{Receivables}}{\text{Net Credit Sales}} \times 365$
Market Value Ratios	Book Value per Share	$\frac{\text{Shareholders' Equality}}{\text{Shares Outstanding}}$
Probability Ratios	Gross Margin	$\frac{\text{Sales} - \text{Cost of Goods Sold}}{\text{Investment Base}}$
	Operating Margin	$\frac{\text{Operating Income Before Interest and Taxes}}{\text{Revenue}}$
	Pre-tax profit margin	$\frac{\text{Sales} - \text{Cost of Goods Sold}}{\text{Investment Base}}$
	Net Profit Margin	$\frac{\text{Net Operating Income}}{\text{Revenue}}$
	Return on Equality	$\frac{\text{Net Operating Income}}{\text{Shareholder's Equality}}$
	Return on Tangible Equality	$\frac{\text{Net Operating Income}}{\text{Tangible Shareholder's Equality}}$
	Return on assets	$\frac{\text{Net Operating Income}}{\text{Total Assets}}$
	Return on investment	$\frac{\text{Investment Gain}}{\text{Investment Base}}$
	Operating Cash Flow Per Share	$\frac{\text{Operating Cash Flow}}{\text{Shares Outstanding}}$
	Free Cash Flow per Share	$\frac{\text{Operating Cash Flow} - \text{Capital Expenditures}}{\text{Shares Outstanding}}$



Moving Time Period

Data split for RF and FNN training, including the moving time period

Table 2. Data split for RF and FNN training, including the moving time period.

RF/FNN	Training Set (20 Quarters)	Validation Set (6 Quarters)	Test Set (1 Quarter)	Moving Time Period During Training and Validation
Moving time period 1	2013Q1–2017Q4	2018Q1–2019Q2	2019Q3	Iterative training 1 Quarter as training input for each moving time period;
Moving time period 2	2013Q2–2018Q1	2018Q2–2019Q3	2019Q4	
Moving time period 3	2013Q3–2018Q2	2018 Q3–2019 Q4	2020Q1	
.....				
Moving time period 11	2015Q3–2020Q2	2020Q3–2021Q4	2022Q1	



Moving Time Period

Data split for GRU and FinGAT training, including the moving time period

Table 3. Data split for GRU and FinGAT training, including the moving time period.

GRU/FinGAT	Training Set (20 Quarters)	Validation Set (6 Quarters)	Test Set (4 Quarter)	Moving Time Period During Training and Validation
Moving time period 1	2013Q1– 2017Q4	2018Q1–2019Q2	2019Q3– 2020Q2	Iterative training 4 Quarters as training input for each moving time period;
Moving time period 2	2013Q2– 2018Q1	2018Q2–2019Q3	2019Q4– 2020Q3	
Moving time period 3	2013Q3– 2018Q2	2018 Q3–2019 Q4	2020Q1– 2020Q4	
.....				
Moving time period 11	2015Q3– 2020Q2	2020Q3–2021Q4	2022Q1– 2022Q4	



Evaluation Metrics



Excess Return

$$ER_p = \left\{ \prod_{q=1}^{\# \text{ quarters}} 1 + R_p(q) \right\} - \left\{ \prod_{q=1}^{\# \text{ quarters}} 1 + R_{\text{benchmark}}(q) \right\}$$



Top-k Precision

$$\text{Top - k precision} = \frac{|L@K(R_q^S) \cap L@K(\widehat{R}_q^S)|}{K}$$

$$\text{Top - k precision} = \frac{K}{\# \text{ stocks in stock pool}}$$



Portfolio Score

$$\text{Portfolio Score} = \frac{\overline{R}_p}{\sigma_p}$$

$$\overline{R}_p = \frac{1}{\# \text{ quarters}} \sum_{q=1}^{\# \text{ quarters}} R_p(q)$$



Model Architecture for Training/Validation/Test



Data Clean

After data collection, the following steps are 'data standardization' to ensure the mean of the characteristic is at 0 and the standard deviation is at 1 for all features. Data standardization can help create a more consistent and reliable dataset. Moreover, it can help the algorithm to search the effective path for relative global minimum and not be easily trapped by local minimum during model training. We use the z score to standardize the vector of features x with the following formula:

$$z = \frac{x - \mu}{\sigma}$$



Relative Return as Target y in Training

The absolute returns quarterly of stock is defined as follows in Equation

$$\frac{\text{price}(q_{\text{end}})}{\text{price}(q_{\text{start}})} - 1$$

The optimization variable y is the relative quarterly return of a stock as in Equation

$$R(q) - R_{\text{TW}50}(q)$$

$R(q)$ is the quarterly absolute return of stock for quarter q , and the $R_{\text{TW}50}(q)$ is the quarterly return of the TW50 index as a benchmark portfolio.

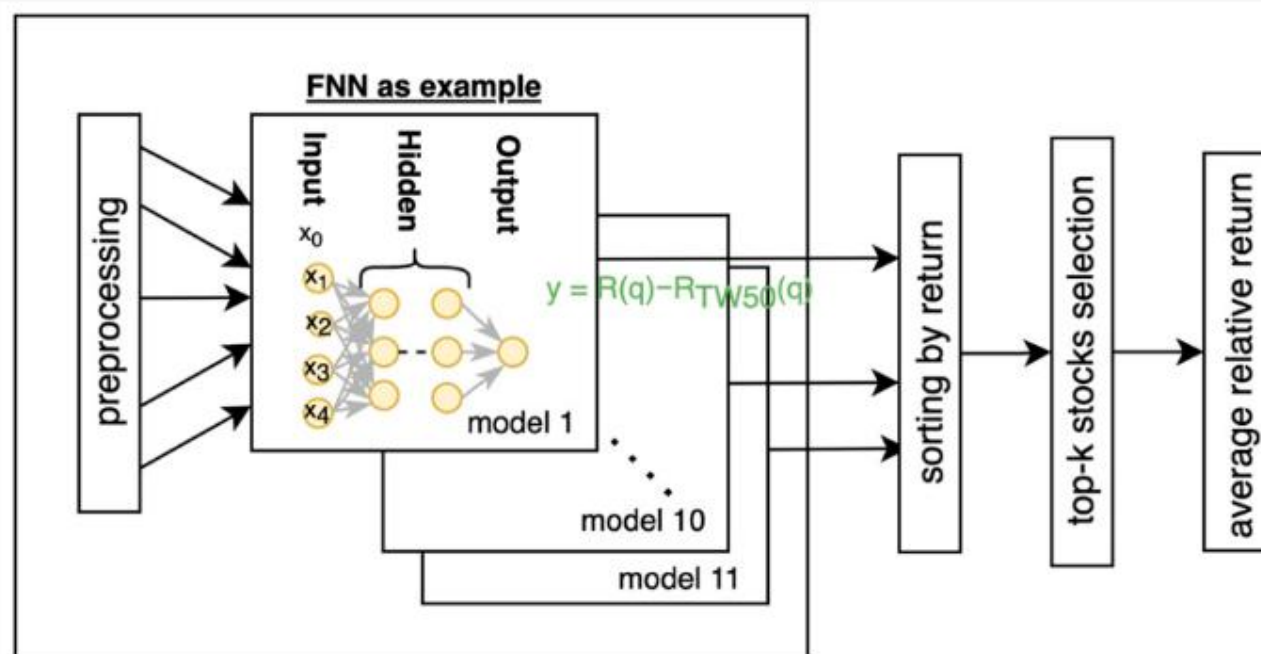


Model Architecture for Training/Validation/Test



Training Procedure

Figure 4. Model training architecture for RF, FNN, GRU, and FinGAT, with FNN as an example in the figure.





Results

Figure 6. Relative return with different portfolios of (a) RF model, (b) FNN model, (c) GRU model, and (d) FinGAT model.

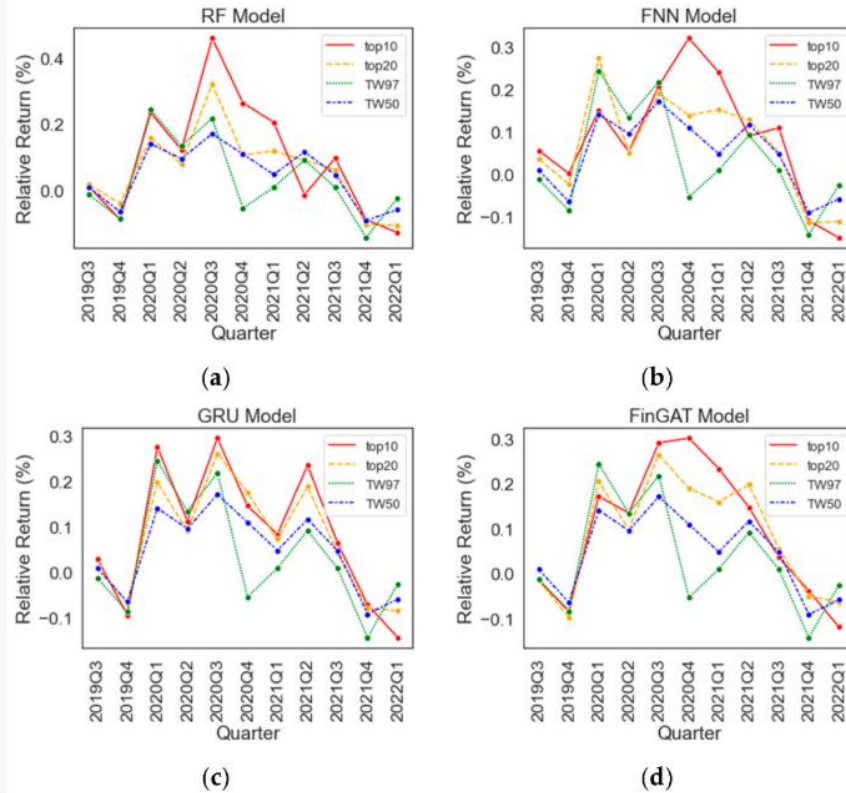


Table 8. Portfolio score, excess return, average return of a portfolio, top-k precision, and STD of portfolio. These indexes are used to compare different portfolio models.

Portfolios	Random Forest		FNN		GRU		FinGAT		TW 97	TW 50
	Top 10	Top 20	Top 10	Top 20	Top 10	Top 20	Top 10	Top 20		
Portfolio Score	0.54	0.53	0.62 (3)	0.58	0.58	0.61	0.65 (2)	0.68 (1)	0.54	0.29
Excess return to the TW50 index	109.2% (2)	50.1%	95.8% (3)	60.1%	65%	79%	114.5% (1)	94.5%	23.5%	Baseline 0%
Average Return of Portfolio	9.8% (1)	6.5%	8.8% (3)	7.0%	8.5%	7.39%	9.67% (2)	8.61%	4.8%	3.5%
Top-k Precision	16.4% ¹	26.8% ² (3)	19.1% ¹	27.7% ² (1)	18.2% ¹	29.5% ²	21.8% ¹	27.3% ² (2)	NA	NA
STD of portfolio	18.1%	12.3%	14.2%	12.2%	14.8%	12.12%	14.87%	12.67%	8.9%	12.3%

¹ The benchmark for top 10 precision in the stock pool is 10.3% (calculated as 10/97). ² The benchmark for top 20 in the stock pool is 20.6% (calculated as 20/97).



Results



High Portfolio Scores

Table 9. Portfolio score of Top-20 with RF and FNN models in US S&P stock market and TW stock market.

Top-20	Research of US S&P [2]	Ours of TW Stock
	Portfolio Score	Portfolio Score
RF	0.414	0.58
FNN	0.202	0.53

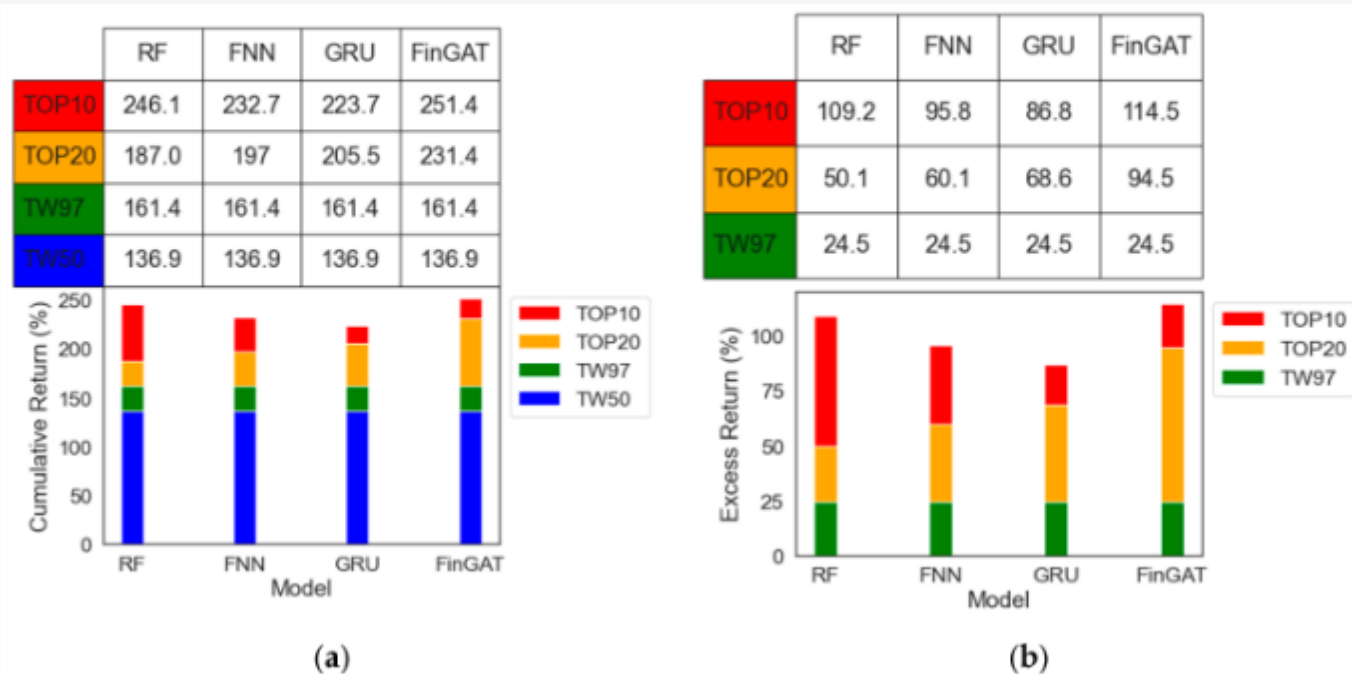


Results



High Excess Return in Top-10 and Top-20 in Test Data for Four Models for Investment Gain

Figure 7. The relative return for 4 models with 4 portfolios from 2019Q3 to 2022Q1, a total of 11 quarters: (a) cumulative return of top 10, top 20, TW97, and TW50 sorted from low to high; (b) excess return of top-10, top-20, and TW97 relative to TW50 index to compare the selected stocks of top-10 and top-20 relative to TW50 index.



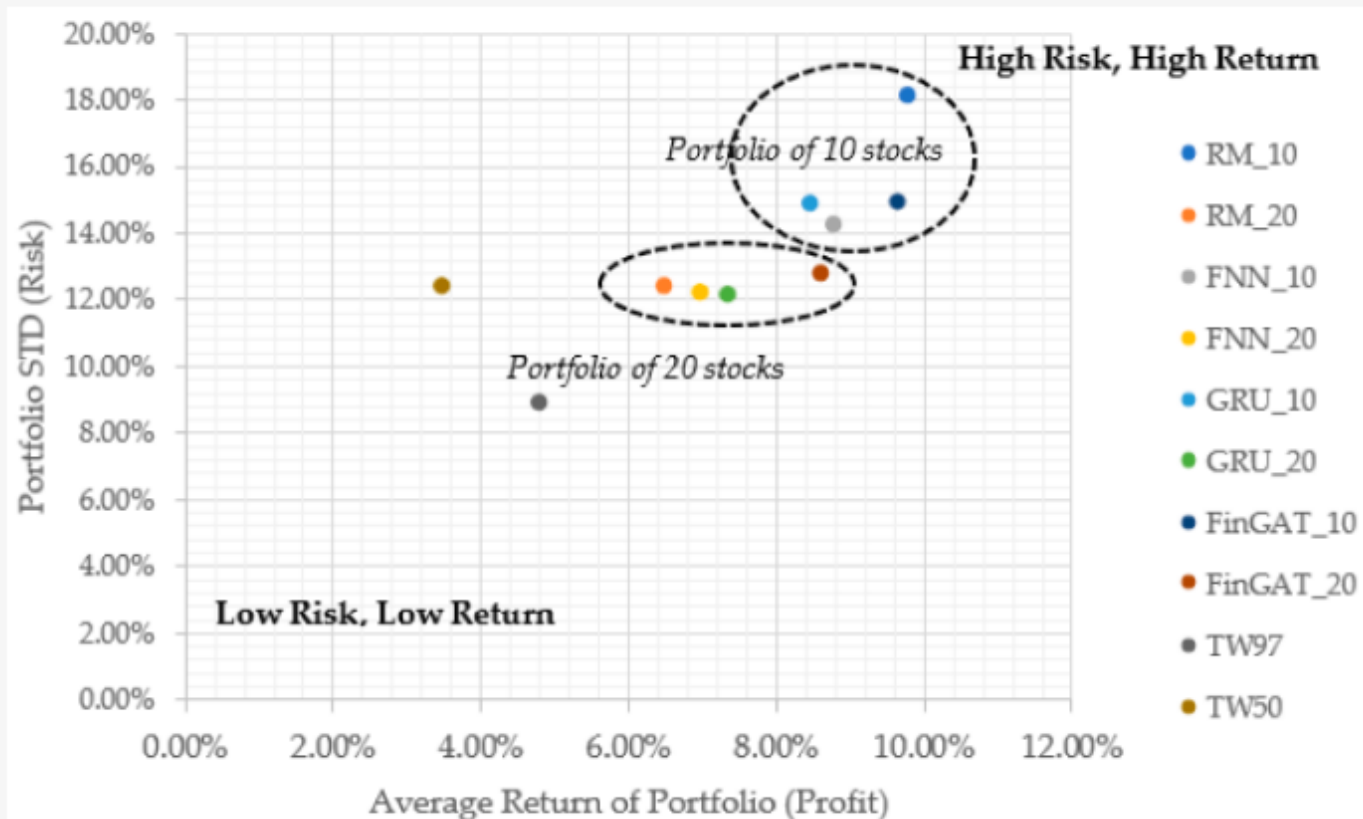


Results



Low-Risk Investment and High Return Rate

Figure 8. Average portfolio return (profit) vs. portfolio STD (risk) of different models.





Challenges and Limitations

Challenges

Financial report timing mismatches can reduce prediction accuracy

May not always deliver top-tier predictions; there's still some uncertainty

Limitations

The accuracy in selecting the absolute best 10 or 20 stocks isn't always high

Models may perform worse in markets where fiscal years of companies don't align

These models rely on past quarterly financials. If data is delayed, inconsistent, or outdated, prediction quality may drop



Dataset for Training My Model

Data	Source
Financial Statements	Tadawul, Company pages
Stock Price History	Yahoo Finance, Tadawul
Macroeconomic Data	Saudi Central Bank, GASTAT
Sentiment Data	News

Thank You