Menemui Matematik
(Discovering Mathematics)

journal homepage: https://persama.org.my/dismath/home

Mixed Asset Portfolio Optimization with Machine Learning and Genetic Algorithm

W.T. Foo[1] and L.S. Lee[2*]

[1,2] *Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, MALAYSIA.*

lls@upm.edu.my
*Corresponding author

Received:24 July 2025
Accepted:12 August 2025

ABSTRACT

Real estate investment offers portfolio diversification through direct property ownership or real estate investment trusts (REITs). This study optimizes a mixed-asset portfolio, including stocks, bonds, and REITs from Malaysia and Singapore stock markets using machine learning models for price prediction and Genetic Algorithm (GA) for assets' allocation. Machine learning models, including Ordinary Least Squares Linear Regression (LR), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost), were used for price prediction, with SVR performing best based on root mean squared error, mean absolute error, and mean squared error metrics. Predicted prices from SVR were then used in a GA for portfolio optimization, initially unconstrained and subsequently with risk constraints for practical applicability. The unconstrained GA produced unrealistically high Sharpe ratios, while GA with risk constraints led to more balanced returns. The final portfolio achieved better returns with controlled risk, highlighting the benefits of REITs in portfolio diversification. The proposed approach highlights the synergy between machine learning and GA, providing a framework for constructing better mixed asset portfolios.

Keywords: REITs, Mixed-Asset Portfolio, Machine Learning, Genetic Algorithm

INTRODUCTION

Real estate investment offers an attractive option for portfolio diversification. It acts as a hedge against market volatility and provides a stable stream of income. Investors can participate in real estate directly through property purchases or indirectly via Real Estate Investment Trusts (REITs). REITs offer advantages like professional management and stock-like liquidity, mitigating the high upfront capital requirements typical of direct property investment (Habbab and Kampouridis, 2024).

Accurately predicting future asset prices is crucial for optimizing mixed-asset portfolios that include REITs. This predictive accuracy ensures investments are allocated to maximize returns while minimizing risks. However, the existing literature on price prediction often focuses on a limited range of algorithms and rarely explores the specific context of REITs within portfolio optimization (Behera et al., 2023).

Several researchers have explored using Machine Learning (ML) models to predict REIT prices. Chen et al. (2014) experimented with ML regression algorithms, including Grey Relational

Analysis (GRA) and Artificial Neural Networks (ANN), to predict REIT returns. Li et al. (2017) studied how neural network algorithms predicted prices for both stocks and REITs. They demonstrated that the ML model achieved higher accuracy than the traditional autoregressive integrated moving average (ARIMA) model. Loo (2020) compared predictions from ANN and linear regression (LR) models on Hong Kong-listed REITs. Lian et al. (2021) compared price prediction between ARIMA and a backpropagation neural network on the Vanguard real estate exchange-traded fund (ETF).

In the use of Genetic Algorithm (GA) for portfolio optimization, Li and Wu (2021) presented a GA model that constructs an investment portfolio including real estate assets with reduced risk under uncertainty conditions. Their study proved that GA is effective in finding optimal weights in a portfolio with real estate assets. Adebiyi et al. (2022) also adopted a GA model to optimize a mixed-assets portfolio including real estate assets using historical market data . They suggested that GA could effectively optimize a portfolio, even for mixed-asset portfolios with different asset classes. However, limited insights into the inclusion of real estate investment in an investment portfolio make composing a portfolio with real estate challenging. This is because real estate assets possess distinct risk and return characteristics compared to stocks and bonds, with low volatility being one of the reasons.

Therefore, this study aims to fill this gap in knowledge of real estate assets, empowering individual investors to make more informed portfolio decisions. Real estate assets can also bring more investment opportunities to the public due to their lower entry costs compared to direct property purchases. Habbab and Kampouridis (2024) previously studied this process in the US, UK, and Australian markets, while Marzuki and Newell (2019) focused on the Belgium market. This paper will focus on the Malaysia and Singapore markets.

RESEARCH BACKGROUND

A financial investment portfolio comprises a variety of assets such as stocks, bonds, commodities, cash, and equivalents, including closed-end funds and ETFs. For example, the Singapore stock market includes assets such as Business Trusts, Structured Warrants, and Daily Leverage Certificates, just to name a few (Jayaraman, 2021). In Bursa Malaysia, offerings include fixed income securities like ICULS (Irredeemable Convertible Unsecured Loan Stock) and ETBS (Bonds and Sukuk traded in Bursa Malaysia), as well as the LEAP market (Malaysia's Leading Entrepreneur Accelerator Platform). Asset types can vary by country to meet different market needs.

A fundamental principle in managing an investment portfolio is diversification, or spreading investments across various financial instruments and sectors to mitigate risk. This approach aims to maximize returns by ensuring that different investments react differently to the same event, such as the impact of the COVID-19 pandemic. Regardless of the portfolio's composition, it should always reflect the investor's risk tolerance, return objectives, time horizon, and other constraints like tax position, liquidity needs, legal considerations, and unique circumstances.

Each financial product has its own benefits and risks. A diversified portfolio, or a mixed assets portfolio is crucial as it determines the portfolio's risk/reward profile and long-term performance expectations. Investors often categorize funds based on their holdings, which may focus on a core asset class like equities or fixed income. Other possible investments include commodities or international assets. Portfolio managers use various methodologies to decide the asset mix, with

modern portfolio theory offering a framework for analyzing investments and determining appropriate allocations based on risk preferences and management goals. Asset allocation portfolios combine equity and fixed income classes, with equities historically offering higher returns and higher risks, while fixed income investments tend to provide lower returns with lower risk. Balancing these elements is essential for defining the asset mix of an investment portfolio.

On the other hand, a REIT is a fund or trust that owns and manages income-generating commercial properties such as shopping complexes, hospitals, plantations, industrial properties, hotels, and office buildings. Management companies for REITs can deduct distributions paid to shareholders from their corporate taxable income. To maintain this tax-free status, REITs must have most of their assets and income tied to real estate and must distribute at least 90% of their total income to investors or unit holders annually (Baker and Chinloy, 2014).

REITs offer other advantages. First, REITs are immediate candidates for dividend or yield-oriented investors. Consequently, income-oriented funds have been increasing their portfolio holdings in REITs. Secondly, some REITs are almost mirror image of conventional corporate entities. Other firms hold cash in tax-avoidance strategies because of high and varying corporate tax rates. Investors treat REITs differently from stocks, some view REITs as a separate asset class (Baker and Chinloy, 2014).

MODERN PORTFOLIO THEORY

Modern Portfolio Theory (MPT) provides a framework for addressing asset allocation by assuming that investors prefer to minimize risk while achieving a given level of expected return. The theory suggests that investors will only accept higher risk if it is accompanied by higher anticipated returns. The balance between maximizing returns and minimizing risk depends on individual risk tolerance, and MPT offers a mathematical approach to navigate this trade-off (Habbab and Kampouridis, 2024).

MPT is based on the Efficient Market Hypothesis (EMH), which asserts that a security's price reflects all available information and its true economic value. In an efficient market, prices are solely influenced by available information, not by managerial decisions. This theory is vital for investment decision-making and forecasting market trends that affect asset allocation (Habbab and Kampouridis, 2024).

Effective portfolios, according to MPT, are those that maximize expected return for a given level of risk or minimize expected risk for a given return. The expected return of a portfolio is calculated by considering the historical returns of its assets, weighted according to the proportion of each asset in the portfolio. This can be expressed as:

$$E(r_p) = \sum_{i=1}^{n} w_i E(r_i), \tag{1}$$

where $E(r_p)$ is the expected return of the portfolio, w_i is the weight of the i-th asset, $E(r_i)$ is the expected return of the i-th asset, and n is the number of assets in the portfolio.

The expected risk of a portfolio, however, is not merely the sum of individual asset risks. Instead, it depends on the interdependencies of the assets, captured by their pairwise correlations. The greater the correlations between the assets, the higher the portfolio's expected risk. The portfolio variance, which represents risk, is given by:

$$\sigma_p^2 = \sum_i w_i^2 \, \sigma_i^2 + \sum_i \sum_{j \neq i} w_i \, w_j \, \sigma_i \sigma_j \rho_{i,j} \, , \quad (2)$$

where σ_p^2 is the portfolio variance, σ_i^2 is the variance of the i-th asset, and $\rho_{i,j}$ is the correlation between the i-th and j-th assets.

MPT emphasizes the importance of diversification in managing portfolio risk. While the risk of individual assets within a portfolio is important, so too are the correlations between them. Diversification aims to combine assets with low pairwise correlations, as these assets can offset each other's risk and contribute to a lower overall portfolio risk (Habbab and Kampouridis, 2024).

Correlation measures the strength and direction of the relationship between two assets. Positively correlated assets tend to move in the same direction, while negatively correlated assets move in opposite directions. In MPT, low correlation between assets is desirable as their movements tend to counteract each other, leading to a more stable portfolio with lower overall risk. Conversely, high correlation between assets amplifies risk, as both assets are likely to move in the same direction, potentially leading to significant losses.

A key concept in MPT is managing the correlation structure within a portfolio. By including assets with diverse correlation levels, investors can achieve a more balanced risk profile. Assets with low or negative correlations can help mitigate the risk associated with highly correlated assets. Therefore, MPT advocates for constructing portfolios that optimize expected return for a given level of risk by strategically considering the correlations between assets.

METHODOLOGY

While ANN with multiple variables have been the dominant approach for predicting REIT prices, other ML models remain less explored. This study aims to bridge this gap by comprehensively evaluating four ML models: Ordinary Least Squares Linear Regression (LR), Support Vector Regression (SVR), k-Nearest Neighbors Regression (KNN), and Extreme Gradient Boosting (XGBoost). The predicted results obtained from these models will then be used in a GA to optimize asset weights within an investment portfolio. The GA adopted in this study will find the optimal weights for a given set of assets based on the return and risk parameters derived from MPT.

Data Collection and Preprocessing

This study will conducts all the ML prediction and GA optimization using data from two country: Malaysia and Singapore. 30 assets including 10 from stocks, 10 from bonds, and 10 from real estates, were chosen randomly from the available market. The data used in the ML method was from a 12-month period of December 2023 to November 2024. These data of adjusted closing prices were collected from Yahoo Finance database (url: https://finance.yahoo.com/). The reason of adjusted closing prices were being used is that it allows investors to obtain an accurate record of the stock's performance. It is extremely useful if an analyst is examining the historical returns because it can provide actual representation of the firm's or companies' equity value. The adjusted closing price factors in anything that might bring impact to the assets' price after the market close on each trading day, such as stock splits, right offerings, or dividends. However, The adjusted closing price is excellent for long-term analysis but can hide valuable insights about nominal price levels and short-term price behaviors, such as those caused by stock splits or dividends. Long-term

investing and analysis were assumed in this study. Table 1 shows the list of assets and their respective tickers chosen in this study.

Table 1: Lists of assets and tickers

MALAYSIA		SINGAPORE	
REITS			
5116.KL	Al-'Aqar Healthcare REIT	A17U.SI	CapitaLand Ascendas REIT
5269.KL	Al-Salam Real Estate Investment Trust	AJBU.SI	Keppel DC REIT
5130.KL	Atrium Real Estate Investment Trust	AU8U.SI	CapitaLand China Trust
5106.KL	Axis Real Estate Investment Trust	D5IU.SI	Lippo Malls Indonesia Retail Trust
5121.KL	Hektar Real Estate Investment Trust	J69U.SI	Frasers Centrepoint Trust
5227.KL	IGB Real Estate Investment Trust	J91U.SI	ESR-REIT
5212.KL	Pavilion Real Estate Investment Trust	ME8U.SI	Mapletree Industrial Trust
5176.KL	Sunway Real Estate Investment Trust	P40U.SI	Starhill Global Real Estate Investment Trust
5110.KL	Uoa Real Estate Investment	SK6U.SI	Paragon REIT
5109.KL	YTL Hospitality REIT	T82U.SI	Suntec Real Estate Investment Trust
BONDS			
0800EA.KL	ABF Malaysia Bond Index	A35.SI	ABF Singapore Bond Index Fd
0820EA.KL	FTSE Bursa Malaysia KLCI ETF	BYJ.SI	Phillip SGX APAC Dividend Leaders REIT ETF
0821EA.KL	Dow Jones Islamic Market Malaysia Titans 25 ETF	CFA.SI	NikkoAM-StraitsTrading Asia ex Japan REIT ETF
0822EA.KL	Principal FTSE ASEAN 40 Malaysia	CLR.SI	Lion-Phillip S-REIT ETF
0824EA.KL	MSCI Malaysia Islamic Dividend ETF	CYC.SI	ICBC CSOP FTSE Chinese Government Bond Index ETF
0828EA.KL	TradePlus Shariah Gold ETF	G3B.SI	Nikko AM Singapore STI ETF
0829EA.KL	TradePlus S&P New China Tracker ETF	MBH.SI	Nikko AM SGD Investment Grd Corp Bd ETF
0836EA.KL	TradePlus DWA Malaysia Momentum Tracker ETF	OVQ.SI	Phillip Sing Income ETF
0837EA.KL	TradePlus MSCI Asia ex Japan REITs Tracker ETF	QK9.SI	iShares MSCI India Climate Transition ETF
0838EA.KL	VP-DJ Shariah China A-Shares 100 ETF	ES3.SI	SPDR Straits Times Index ETF
STOCKS			
1066.KL	RHB Bank Berhad	D05.SI	DBS Group Holdings Ltd
1961.KL	IOI Corporation Berhad	G13.SI	Genting Singapore Limited
2445.KL	Kuala Lumpur Kepong Berhad	OV8.SI	Sheng Siong Group Ltd
3026.KL	Dutch Lady Milk Industries Berhad	MZH.SI	Nanofilm Technologies International Limited
3182.KL	Genting Berhad	O39.SI	Oversea-Chinese Banking Corporation Limited
4197.KL	Sime Darby Berhad	QC7.SI	Q & M Dental Group (Singapore) Limited
4863.KL	Telekom Malaysia Berhad	S08.SI	Singapore Post Limited
5908.KL	DKSH Holdings (Malaysia) Berhad	Z74.SI	Singapore Telecommunications Limited
6012.KL	Maxis Berhad	8K7.SI	UG Healthcare Corporation Limited
6033.KL	PETRONAS Gas Berhad	BS6.SI	Yangzijiang Shipbuilding (Holdings) Ltd.

The datasets obtained were split into 70% for training, 10% for validation, and 20% for testing. However, it cannot be assured that one ML model's will perform well when it encounter new and unseen data. In order to ensure that the ML model is robust and generalizes well to new data, cross validation can be processed and added to prevent overfitting in a ML model. The 5-fold cross validation carried out in this study was by randomly divides the datasets into 5 different sets of data in the above mentioned percentage. The ML method will then use the first set of data for training, second set of data for validation, and the third set of data for testing, for 5 different folds of datasets split by ***sklearn.model_selection***, ***KFold*** in the Python libraries. After the data being split into 5 folds, they can now being proceeded into the next step, which is data preprocessing.

In order to maintain the stationarity of the time series data, differencing and scaling of data preprocessing is required before the data is applied in the ML training process. Differencing the dataset will remove the upward trend and keeping the average constant over time. Stationarity is crucial in time series analysis models such as ARIMA, as this model will assume the data is independent with one another. However, it is impossible for a market price time series to be independence without any data preprocessing (Aderson et al., 2021).

In this process, the following equation will be applied

$$N_t = \frac{D - D_{min}}{D_{max} - D_{min}}, \quad (3)$$

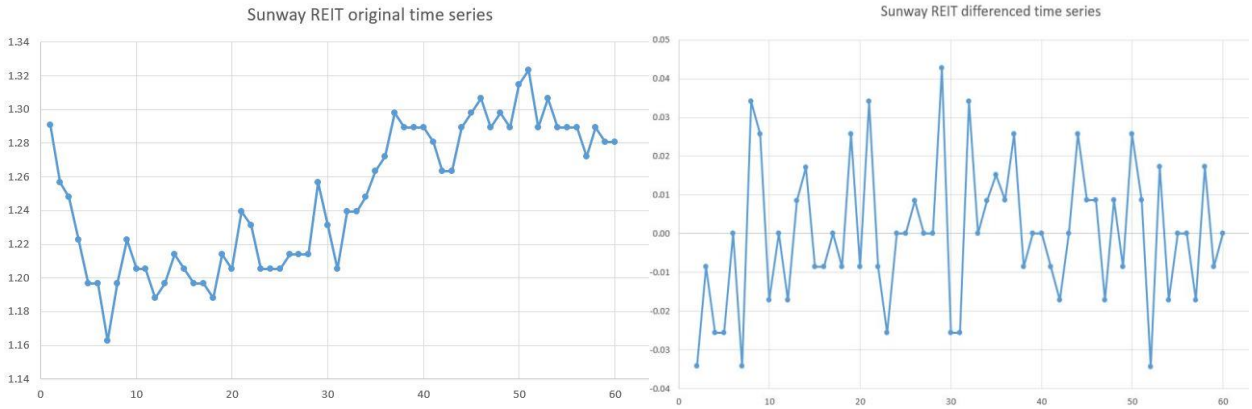
to normalizes the independent variable to a range of values between 0 to 1. The target variable is N_t, the standardized value of each differenced price D, and D_{min} or D_{max} are the minimum and maximum value of D from the whole set of data.

Table 2 is an example of differenced data, followed by graph of the original and differenced time series (Figure 1(a) and 1(b)) using the same dataset from Table 2.

Table 2: Example of time series differentiation and feature selection

t	P_t	P_{t-1}	D_t	N_t	N_{t-1}
1	1.290928	–	–	–	–
2	1.256731	1.290928	–0.034197	0.002353	–
3	1.248182	1.256731	–0.008549	0.334901	0.002353
4	1.222534	1.248182	–0.025648	0.113198	0.334901
5	1.196887	1.222534	–0.025647	0.113203	0.113198
6	1.196887	1.196887	0.000000	0.445750	0.113203
7	1.162690	1.196887	–0.034197	0.002349	0.445750
8	1.196887	1.162690	0.034197	0.889150	0.002349
9	1.222534	1.196887	0.025647	0.778297	0.889150
10	1.205436	1.222534	–0.017098	0.224051	0.778297

Table 2 included the column of time steps, t, the adjusted closing price of the asset, P_t, P_{t-1} which is one-lagged value of the adjusted closing price, the differenced D_t values which is calculated as $P_t - P_{t-1}$. Following by the normalized value of D_t which is N_t, and its lagged values with n determined by Akaike Information Criterion (AIC).



(a) Sunway REIT original time series. (b) Sunway REIT differenced time series.

Figure 1: Example of original and differenced time series using Sunway REIT data.

AIC is a widely used metric for model selection (Yamaoka, 1978; Vrieze, 2012) as an estimator of prediction error. The value of n may vary depending on the specific dataset, resulting in a different number of features for each dataset. In this study, the value of n is limited at 3 in all the ML methods to preserve sufficient amount of data for the training, validation, and testing process. After the ML method obtained the predicted time series value, the differencing and scaling process will be done reversely to calculate the performance metrics as decribed in this paper.

Machine Learning Models for Price Prediction

Once the data has been pre-processed and the relevant lagged features have been created, they will then be transferred to ML models to predict the prices of the datasets, which belong to three asset classes: REITs, stocks, and bonds. The price will serve as the target variable for this regression task. This study will evaluate the performance of four widely-used ML models: LR, SVR, XGBoost, and KNN.

To implement these ML models, the following Python libraries will be utilized: ***sklearn***, and ***xgboost***. The specific functions that will be used to fit the models to the training data are:

sklearn.linear_model.LinearRegression
sklearn.svm.SVR
xgboost
sklearn.neighbors.KNeighborsRegressor

To optimize the performances of the ML models, parameters included in functions above were determined using a grid search method. The best parameters were chosen by Grid Search method in Python. The range of parameters used in ML models are as Table 3.

Table 3: ML models and its parameters

Model	Parameter	Value Range
SVR	Kernel function	'linear', 'poly', 'rbf', 'sigmoid'
	Degree of kernel function	1, 2, 3
	Kernel coefficient (gamma)	'scale', 'auto'
	Tolerance of stopping criterion	0.001, 0.01, 0.1
	Epsilon	0.1, 0.5, 0.8
XGBoost	Number of estimators	10, 20, 30
	Maximum depth of a tree	3, 4, 5
	Maximum child weight	1, 5, 10
	Learning rate	0.001, 0.01, 0.1
KNN	Number of neighbors	5, 10, 20
	Weights	'uniform', 'distance'
	Algorithm	'auto', 'ball_tree', 'kd_tree'

LR was not tuned as it lacks parameters to be tuned. The ML models were then fitted to the training data, validated, and lastly applied to the test set using ***predict*** attribute of the respective model. The predicted prices from test sets were recorded for the usage in the GA optimization.

Performance Metrics of Machine Learning Models

The performance metrics that will be evaluated in this study includes:

i) Mean Squared Error (MSE)

The MSE is a cost function commonly used in regression. MSE measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted

values. When a model has no error, the MSE equals zero. As model error increases, its value increases.

$$MSE = \frac{\sum_{t=1}^{T}(P_t - \hat{P}_t)^2}{T}, \quad (4)$$

where T is the number of observations, P_t refers to the actual value of the price, and $\hat{P}_t$ is the predicted value of the price.

ii) Mean Absolute Error (MAE)

The MAE is another common cost function used in regression problems to measure the average difference between the predicted and actual values. It focuses on the absolute value of the errors, giving equal weight to both overestimations and underestimations. The smaller the MAE, the better the model's predictions align with the actual data.

$$MAE = \frac{\sum_{t=1}^{T}|P_t - \hat{P}_t|^2}{T}, \quad (5)$$

where T is the number of observations, P_t refers to the actual value of the price, and $\hat{P}_t$ is the predicted value of the price.

iii) Root Mean Squared Error (RMSE)

The RMSE is one of the main performance indicators for a regression model. It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy). The lower the value of the RMSE, the better the model is.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(P_t - \hat{P}_t)^2}{T}}, \quad (6)$$

where P_t refers to the actual value of the price, $\hat{P}_t$ is its predicted value, and T is the number of observations.

The differenced and scaled values are reverted back to their original price values, to calculate the cost function, which is why P_t acts as the target variable in Eq. (4) to (6). This is because all the cost functions measure the error between predicted and actual prices. Without reverting, these errors would be based on the scaled units, not reflecting the true difference in real prices. Additionally, the ML model evaluation aims to understand performance on the original price scale. Reverting ensures the cost function operates on original price values, providing a more meaningful measure of prediction accuracy.

Genetic Algorithm for Portfolio Optimization

After the price prediction process is completed by the ML models, a GA will then be used to optimize the weights of assets in the portfolio. GAs have been previously successfully applied into financial portfolio optimization (Habbab and Kampouridis, 2024). Metrics derived from MPT will be present as follows:

a. Sharpe ratio, which is computed as the ratio of the difference between the mean return and the risk-free rate to the standard deviation of the returns, that is

$$S = \frac{r - r_f}{\sigma_r}, \tag{7}$$

where r is the average return of the investment, r_f is the risk-free rate, and σ_r is the standard deviation of the returns. Generally, the higher the Sharpe ratio, the more attractive the risk-adjusted return.

b. The average return of each asset is calculated as the simple average of the returns of that asset, that is

$$r = \frac{\sum_{t=1}^{T} r_t}{T}, \tag{8}$$

where r_t is the return observed for each time point t and T is the number of observations.

c. To calculate the return r_t, the predicted price time series needs to transformed to returns through the following formula:

$$r_t = \frac{\hat{P}_t - \hat{P}_{t-1}}{\hat{P}_{t-1}}. \tag{9}$$

d. The standard deviation of returns is calculated as the square root of the average of the squared differences between the average return and each observed return follows:

$$\sigma_r = \sqrt{\frac{\sum_{t=1}^{T} (r - r_t)^2}{T}}. \tag{10}$$

Among the formulas above, Eq. (7) will be used as the fitness function in all the GA optimization.

The predicted prices from the best performing ML model, will be used as the main input in the GA implementation. Firstly, daily return of each asset will be calculated using Eq. (8). Mean assets return will then computed by calculating the mean of daily assets return. In this study, each chromosome of the GA having genes of 60, representing the 60 assets chosen at the beginning of the algorithm. A chromosome is represented as:

$$\boldsymbol{c} = [w_1, w_2, w_3, \ldots, w_{60}], \tag{11}$$

where w_i is the weight of portfolio assigned to each gene, and the sum of all genes are always normalized to 1, as $\sum_{i=1}^{60} w_i$ and $0 \leq w_i \leq 1$, which represents total weight of a chromosome or a portfolio always equals to 100%.

The initialization of the weights to each gene was done by assigning a random number to it. This is to ensure diversity and avoids premature convergence to suboptimal solutions in the investment portfolio. In this study, the size of population is 50. The following Figure 2 represents a chromosome with random weights assigned:

0.00353757	0.02392344	0.02838240	...	0.00483185	0.00235717	0.01056850

Figure 2: Example of a chromosome with randomly assigned weights.

Figure 3 showcased the algorithm of the complete GA process.

S1: [**Start**] Generate an initial population P_{50}, of x chromosomes.
S2: [**Fitness**] Evaluate the fitness (Sharpe ratio) of each chromosome x in the population.
S3: [**Elitism**] Select the top 10% best-performing chromosomes.
S4: [**New Population**] Create a new population by repeating the following steps until the new population is complete.
i. [**Selection**] Select 2 parent chromosomes from the elite chromosomes.
ii. [**Crossover**] With a crossover probability $p_c = 0.80$, cross over the parents to form 2 new offspring (children). One-point crossover is performed at a random location within the chromosome.
iii. [**Mutation**] With a mutation probability $p_m = 0.20$, mutate new offspring at a random gene in the chromosome.
iv. [**Replace**] Place new offspring in the new population.
S5: [**Fitness**] Evaluate the fitness $g(x')$ (Sharpe ratio) of each chromosome x' in the new population.
S6: [**Test**] If the number of generations met, **STOP**, and return the fittest solution found; otherwise, go to **S4**.

Figure 3: Algorithm of proposed GA.

RESULTS AND DISCUSSION

Price Prediction by Machine Learning

Table 4 shows the numerical results obtained from ML price prediction in terms of RMSE, MAE and MSE.

Table 4: Numerical results of performance metrics of the ML models

ML Model	Validation			Test		
	RMSE	MAE	MSE	RMSE	MAE	MSE
KNN						
All assets	0.375457	0.310458	1.036949	0.170986	0.137835	0.218166
REITs only	0.098747	0.080446	0.014256	0.046547	0.039218	0.003367
LR						
All assets	0.386247	0.318532	1.124226	0.105330	**0.071834**	0.070605
REITs only	0.099014	0.080186	0.014531	0.028230	0.020610	0.001112
SVR						
All assets	0.400317	0.327846	1.184027	**0.096812**	0.089691	**0.046982**
REITs only	0.099445	0.080218	0.014612	**0.026321**	**0.001657**	**0.000981**
XGBoost						
All assets	0.375441	0.311331	1.027224	0.161823	0.130576	0.194712
REITs only	0.098792	0.079994	0.014298	0.043333	0.035610	0.002859

Among the ML methods studied, SVR is ranked as top in terms of average RMSE, MAE, and MSE of all 30 assets, and also of all REITs. It then followed by LR, XGBoost, and lastly KNN.

In order to present the significant of the differences of the ML models, Friedman non-parametric test had also been performed. In this test, lower rank represents better model's

performance, which represents having lower value of performance metrics in predicting the prices of assets. The test was carried out to compare the RMSE, MAE, and MSE for all four ML models, and the results are presented in Table 5.

Table 5: Statistical test of performance metrics distribution of "all assets" and "REITs only" based on Friedman non-parametric test

ML model	Average Rank	
	All assets	REITs only
KNN	3.60	4.00
LR	2.00	2.00
SVR	1.00	1.00
XGBoost	3.40	3.00

The results obtained from Friedman non-parametric test had shown similar results in terms of all three performance metrics. Apart from proving the significant difference between the ML models, a post-hoc test have also been performed to compare the models pairwise. To discover if the rank differences obtained from the Friedman test are significant, Nemenyi test was performed. The Nemenyi test works by computing average rank of each models and taking their difference. If the average rank differences are larger than or equal to the critical difference (CD) computed, it could be conclude that the performances of the two models corresponding to these differences are significantly different from one another. The CD was computed using this formula,

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}, \tag{12}$$

where k is the number of models, n is the number of datasets, and q_α is the critical value from the Nemenyi table based on the number of models and significance level, 0.05 (Japkowicz and Shah, 2011). After performing the Nemenyi test with $q_\alpha = 2.728$, a critical difference of $CD = 2.2274$ was obtained.

Table 6: Statistical test of performance metrics distribution of "all assets" and "REITs only" based on Nemenyi post-hoc test

Pairs of ML Models	All Assets		REITs only	
	Difference	Significant	Difference	Significant
KNN vs. LR	1.60	NO	2.00	NO
KNN vs. SVR	2.60	YES	3.00	YES
KNN vs. XGBoost	0.20	NO	1.00	NO
LR vs. SVR	1.00	NO	1.00	NO
LR vs. XGBoost	1.40	NO	1.00	NO
SVR vs. XGBoost	2.40	YES	2.00	NO

From the performance metrics and test analysis carried out, it could be concluded that SVR demonstrates superior performance in terms of RMSE, MSE, and MAE. This owed to its capability to effectively model complex, non-linear relationships through the use of kernel functions like the RBF kernel. These kernels enable SVR to capture intricate patterns in the data that simpler methods, such as LR, often overlook. Furthermore, SVR's hyperparameters, including the regularization parameter (C), kernel coefficient (γ), and epsilon (ϵ), offer precise control over the model's bias-variance tradeoff and its tolerance for minor errors. By prioritizing the minimization

of significant deviations while disregarding smaller ones, SVR adeptly manages data variability more efficiently and robustly compared to alternatives like KNN or XGBoost.

On the other hand, the reason LR ranked second among the four ML models, despite lacking parameter tuning, may stem from its simplicity and efficiency when handling relatively well-balanced datasets. LR excels particularly well when the relationships between the features and the target variable are predominantly linear. In cases where the dataset exhibits minimal noise, well-scaled features, and a lack of strong non-linearities, LR can deliver competitive results without the complexity associated with additional hyperparameter tuning. This simplicity allows it to avoid overfitting, which can occasionally be a concern for more advanced models like XGBoost or KNN when not properly tuned.

Portfolio Optimization by Genetic Algorithm

Figure 4 shown the graph of evolution of value of Sharpe ratio and expected risk over 300 generations of GA.

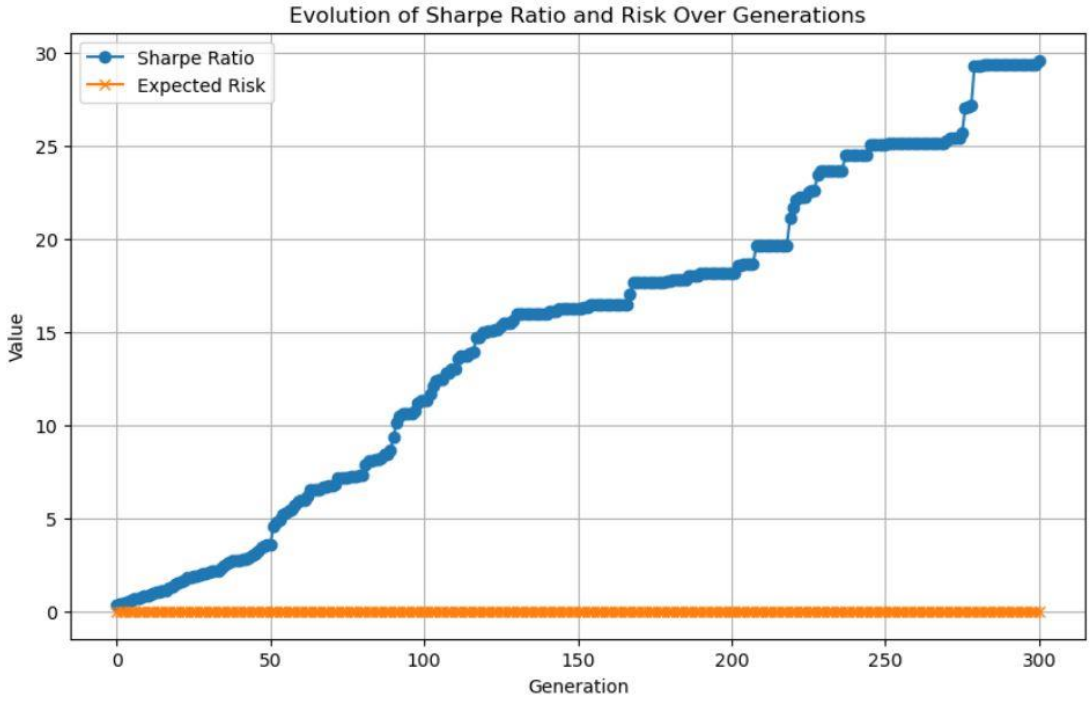


Figure 4: Graph of Sharpe ratio and expected risk of best chromosome of 300 Generations

From Figure 4, the Sharpe ratio at generation 300 (29.601678) is around 87.94 times higher than the Sharpe ratio at generation 0 (0.336626). The Sharpe ratio exhibits a steady and significant increase over the 300 generations. This indicates that the GA effectively identified better solutions with higher risk-adjusted returns as generations progressed. Notable jumps in the Sharpe ratio occur in the early generations around 275, followed by more incremental improvements as the algorithm approaches convergence. While the expected risk decrease constantly across all generations. This stability suggests that while the GA focused on improving the Sharpe ratio, it maintained a slight decreasing level of risk, aligning with the objective of risk control.

The Sharpe ratio shows rapid improvement during the first 250 generations, suggesting that the algorithm quickly identifies promising solutions. However, after approximately 200 generations, the increment Sharpe ratio begins to slow down, indicating that the GA has likely

converged to an optimal or near-optimal solution. This stagnation suggests that the GA has exhausted its capacity to further improve performance within the given constraints.

From Figure 4, it was observed that extremely high Sharpe ratios were achieved, this is because of the portfolio's exploiting negligible risk values. This phenomenon arises from the mathematical structure of the Sharpe ratio:

$$S = \frac{r - r_f}{\sigma_r}. \tag{13}$$

As the denominator (risk, σ_r) approaches zero, even small returns result in disproportionately high Sharpe ratios, which are not reflective of practical portfolio performance. This necessitated the introduction of a risk constraint to ensure realistic and interpretable outcomes. The risk constraint was implemented by introducing a risk floor. In the fitness function, the adjusted risk was calculated as:

$$\text{Adjusted Risk} = \max(\text{Portfolio Risk}, \text{Risk Floor}). \tag{14}$$

Experiments with different values of risk were carried out and it was found that GA optimization with risk value of 0.002 provided a reality level of Sharpe ratio, which should be ranged around 2 to 3. From Figure 5, it could be observed that the Sharpe ratio are now of value 2.897166, a reality level of good Sharpe ratio, with expected return of 0.006318. In this portfolio, the expected return is about 3.16 times of the risk level, where it represents a balanced risk-return profile. The slightly higher risk provides a realistic return potential while maintaining excellent risk management.

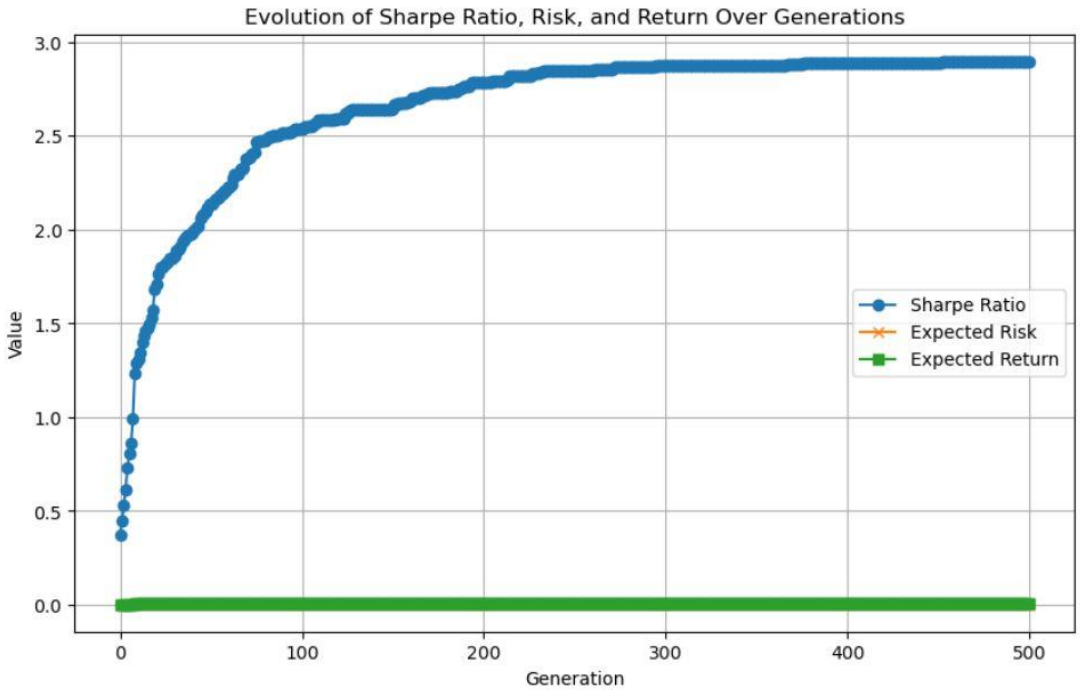


Figure 5: Graph of Sharpe ratio and expected risk of best chromosome of 500 Generations (with risk of 0.002)

Figure 6 shown the pie chart of optimized portfolio obtained from the best chromosome of 500 generations with the adjusted risk level of 0.002, bringing the Sharpe ratio of 2.897166 and expected return of 0.006318.

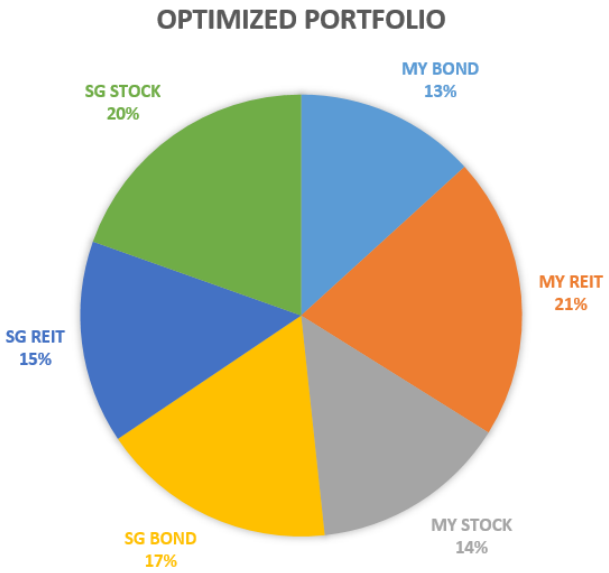


Figure 6: Final optimized portfolio obtained from GA

It can be concluded that, if an investor would like to invest in Malaysia and Singapore stock market, he/she can invest for 13% for Malaysia Bonds, 21% of Malaysia REITs, 14% of Malaysia stocks, 17% for Singapore Bonds, 15% for Singapore REITs, and 20% of Singapore stocks based on 60 assets experimented in this study. This allocations can help in giving the best performance of a portfolio, and is aligned with findings from previous studies (Habbab and Kampouridis, 2024). The results prove that the inclusion of REITs into one's investment portfolio is significant. Apart from bringing lower risks to the portfolio, it also brought better risk-return adjusted portfolio due to its diversification potential.

CONCLUSIONS

This study began with price prediction of assets using four ML models: KNN, LR, SVR, and XGBoost. ML model that produced the best performance metrics in terms of RMSE, MAE, and MSE is identified as SVR. Statistical tests, including Friedman and Nemenyi post-hoc analysis, also validated the superiority of SVR in both general assets and REIT-specific datasets. Knowing that the high accuracy of the SVR among four ML models, the predicted prices from this model is then used as the first step in GA, in order to optimize the relevant mixed-assets portfolio. By using Sharpe ratio as the fitness function of the GA, the best weight of assets in the portfolio were determined.

Along with the better accuracy in price prediction, SVR also happened along with longer training and computational times. Single round of model training and price predicting of an asset will typically took around 3.5 to 4.5 minutes. Meanwhile, other ML models (LR, KNN and XGBoost) usually completed the training and predicting phases within the range of 5 to 60 seconds.

The main finding of this research is that it highlights the importance of robust ML models like SVR in predicting prices in a financial setting. Better accuracy of prediction is essential for

the GA optimization in order to return a more reliable portfolio. The inclusion of REITs in a mixed-assets portfolio can enhanced the diversification and provide stable returns due to their low volatility and correlation with other type of assets, such as stocks or bonds.

However, this study only utilized historical data from Malaysia and Singapores stock market. Broader datasets, such as those from the European countries could enhanced the generalization of prediction and optimization. Following that, researchers could possibly experiment on the differences between predicting adjusted closing prices and closing prices. As focusing on the adjusted closing price might obscure nominal price trends and short-term behaviours of an asset.

Researchers could also test on different crossover and mutation parameter to further improve the GA's performance. Hybridizing ML models with other optimization methods such as particle swarm optimization or simulated annealing would also be a good direction to experiment with. Moreover, researchers could also start to incorporate real-world constraints, such as taxes and liquidity to ensure a more practical and realistic portfolio's weight recommendations.

As summary, this study demonstrates the potential of teaming ML models and GA optimization in portfolio management. By addressing prediction accuracy and balancing risk-and-return, the proposed methodology provides a framework for constructing efficient mixed-asset portfolios especially in Malaysia and Singapore stock markets, helping investors to achieve their financial goals in a more assured way.

REFERENCES

Adebiyi, S. O., Ogunbiyi, O. O. & Amole, B. B. (2022), Artificial intelligence model for building investment portfolio optimization mix using historical stock prices data, *Rajagiri Management Journal*, **16(1)**: 36–62.

Anderson, J., Anderson, R., Guirguis, H. S., Proppe, S. & Seiler, M. J. (2021), Time-varying correlations of REITs and implications for portfolio management, *Journal of Real Estate Research*, **43(3)**: 317–334.

Baker, H. K. & Chinloy, P. (2014), Public Real Estate Markets and Investments, Oxford: Oxford University Press.

Behera, J., Pasayat, A. K., Behera, H. & Kumar, P. (2023), Prediction-based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multinational stock markets, *Engineering Applications of Artificial Intelligence*, **120**: 105843.

Chen, J. H., Chang, T. T., Ho, C. R. & Diaz, J. F. (2014), Grey relational analysis and neural network forecasting of REIT returns, *Quantitative Finance*, **14(11)**: 2033–2044.

Habbab, F. Z. & Kampouridis, M. (2024), An in-depth investigation of five machine learning algorithms for optimizing mixed-asset portfolios including REITs, *Expert Systems with Applications*, **235**: 121102.

Jayaraman, B. (2021), Building Wealth Through REITs, Singapore: Marshall Cavendish International Asia Pte Ltd.

Li, M. & Wu, Y. (2021), Dynamic decision model of real estate investment portfolio based on wireless network communication and ant colony algorithm, *Wireless Communications and Mobile Computing*, **2021(1)**: 9261312.

Li, R. Y. M., Fong, S. & Chong, K. W. S. (2017), Forecasting the REITs and stock indices: group method of data handling neural network approach, *Pacific Rim Property Research Journal*, **23(2)**: 123–160.

Lian, Y. M., Li, C. H. & Wei, Y. H. (2021), Machine learning and time series models for VNQ market predictions, *Journal of Applied Finance and Banking*, **11(5)**: 29–44.

Loo, W. K. (2020), Predictability of HK-REITs returns using artificial neural network, *Journal of Property Investment and Finance*, **38(4)**: 291–307.

Marzuki, M. J. & Newell, G. (2019), The evolution of Belgium REITs, *Journal of Property Investment & Finance*, **37(4)**: 345–362.

Vrieze, S. I. (2012), Model selection and psychological theory: A discussion of the differences between the akaike information criterion (AIC) and the bayesian information criterion (BIC), *Psychological Methods*, **17(2)**: 228.

Yamaoka, K., Nakagawa, T. & Uno, T. (1978), Application of akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations, *Journal of Pharmacokinetics and Biopharmaceutics*, **6(2)**: 165–175.