IADIS International Journal on Computer Science and Information Systems
Vol. 19, No. 2, pp. 115-129
ISSN: 1646-3692

FEATURE SELECTION METHODOLOGY FOR ML STOCK PREDICTIONS USING SET50 OF THE STOCK EXCHANGE OF THAILAND

Gridaphat Sriharee
Department of Computer and Information Science, Faculty of Applied Science
King Mongkut's University of Technology North Bangkok
1518 Pracharat 1, Wonsawang, Bangsur, Bangkok, 10800, Thailand

ABSTRACT

Stock prediction using machine learning is an interesting topic for investors. However, the performance of the prediction depends on different techniques and the data itself. In this paper, a feature selection methodology has been proposed. It consists of filter method and wrapper method. A feature selection experiment was conducted on 50 stocks (SET50) from the Stock Exchange of Thailand (SET). The calculation of feature importance for feature selection was discussed. The feature importance shows how the cohort indicators behave in each wrapping level. Preliminary experiment was conducted to investigate some technical indicators that could be affected by SET50. The basic machine learning models both regression models and classification models were examined to evaluate the performance of the models based on these features. The proposed feature selection methodology was flexible and practical as each stock can be influenced by different features. Based on the measured feature importance, the features can be selected in different ways which can efficiently increase the performance of the machine learning model.

KEYWORDS

Feature Selection, Stock Prediction, Technical Indicator, Machine Learning

1. INTRODUCTION

Stock analysis can be roughly divided into two types: fundamental analysis and technical analysis. However, there are other techniques that are widely used, such as algorithmic trading and quantitative analysis. Quantitative analysis is one of the stock analysis techniques that apply mathematical and statistical principles together. It also includes machine learning or artificial intelligence as a technique for stock prediction. Therefore, these topics are challenging from different investor perspectives. Technical indicators are usually used for buy/sell signals, monitoring stock trends and analyzing movements. Investors can use the indicators to support

their decisions. The technical indicators can be used as features in a machine learning model. However, performance and accuracy can vary depending on the technique. Stock prediction using machine learning can improve algorithmic trading, where bids and offers are executed automatically.

The performance of a machine learning model depends on the selection of features. The model needs features to predict the target, and the features should be independent of each other, but at the same time define the target. The selection of better features leads to a better prediction accuracy and can reduce the computational complexity. Feature selection requires OHLCV information, technical indicators (e.g. RSI, moving average) and economic indicators (e.g. interest rates, consumer price index) (Htun et al., 2023). In stock trading, there are both internal factors (e.g. company, fact sheet, company performance) and external factors (e.g. news, other stock markets, events) that cannot be controlled and affect the stock price. Therefore, investors have carefully selected a risk management analysis.

This paper presents a preliminary experiment to investigate which technical indicators work well for which ML models with SET50 stocks. As a result, a feature selection methodology was proposed. It combines both the filter method and the wrapper method. Some contributions are as follows.

(i) Investigation of technical indicators that may have an impact on SET50 stocks. The preliminary experiment was conducted with three sample stocks: AOT, MINT, and EA. The performance evaluation of the machine learning model was reported and discussed.

(ii) A feature selection methodology that includes multiple groups of technical indicators for stock prediction. The measurement of the importance of features in relation to the wrapping process was presented.

Section 2 contains the background to this work. Section 3 is related work. Section 4 describes a preliminary experiment and correlation analysis using SET50 stocks. Technical indicators for the machine learning model are also presented. Section 5 presents the proposed feature selection methodology. An overview of the processes is described. Section 6 presents the importance of features, a score used to measure the performance of cohort indicators in machine learning models, and Section 7 contains the conclusion of this work.

2. BACKGROUND KNOWLEDGE

2.1 Technical Indicators and Stock Analysis

Stock analysis can be divided into two types: fundamental analysis and technical analysis. However, there are many other types of analysis, such as quantitative analysis and algorithmic analysis, which can be part of quantitative analysis. Quantitative analysis uses mathematical and statistical techniques to model, analyze and predict stock prices. Statistical values such as mean, standard deviation, correlation and probability are factors used to examine the data. This data can also be used for stock prediction. Fundamental analysis focuses on basic company information, e.g. balance sheets, reports and company earnings. This technique relies on public information. Technical analysis is based on technical indicators, e.g. the simple moving average (SMA), the relative strength index (RSI) and the exponential moving average (EMA). These indicators show the development of the stock price over a certain period of time (see Formula 1-3).

$$SMA_n = \sum_{n=1}^{n-1} \frac{P_1 + \ldots + P_n}{n} \quad (1)$$

$$RSI = 100 - \frac{100}{1 + RS} \quad (2)$$

$$EMA_t = (P_t * \frac{Constant}{1+n}) + (EMA_{t-1} * (1 - \frac{Constant}{1+n})) \quad (3)$$

Fundamental data (e.g. OHLC (open, high, low, close), return), financial data (e.g. balance sheet), technical indicators and even external market data (e.g. gold and currencies) can be used to predict stocks. Fundamental data can be return, logarithmic return and percentage change, which are calculated based on the close price. Technical indicators such as SMA, EMA and RSI are calculated from the close price of the stock. An indicator can be calculated from other indicators. Thus, different types of data can be used for stock prediction with a machine learning model. However, the accuracy of the prediction usually depends on the features used.

Combination of technical indicators for strategic planning or the search for buy/sell signals applies multiple indicators together. To find a buy or sell signal, for example, the MACD can be used, which calculates the EMA with a period of 12 and 26 days. Observing the SMA and EMA lines, which can be above or below each other, can provide a buy/sell signal. Using two RSI lines with different time periods can confirm overbought and oversold levels. The technical indicators therefore show the value depending on the time period (short period (sell) and long period (buy)). Figure 1 shows the SMA and EMA lines of the AOT. If the EMA line is above the SMA line, this means a sell signal (short period) and in contrast, a buy signal (long period). The value represented by these lines can be used as a feature for the stock prediction.

Figure 1. AOT, and SMA and EMA line

2.2 Feature Selection

During feature selection, unnecessary features are eliminated in order to reduce the feature dimension. Various technical indicators, fundamental data and economic data (e.g. gold market indicators) can be used in stock analysis. Using a large number of features where there is a lack of methodology to evaluate the importance of the features can lead to poor performance when these features are used in a machine learning model. However, it can be difficult to analyze the important features as they can vary depending on the data used. In addition, the indicator may have different effects for each model and/or stock. Therefore, a good feature that is suitable for some conditions may not be suitable for another condition. Thus, a more refined feature selection process is required to find the good/best features for specific data and specific models.

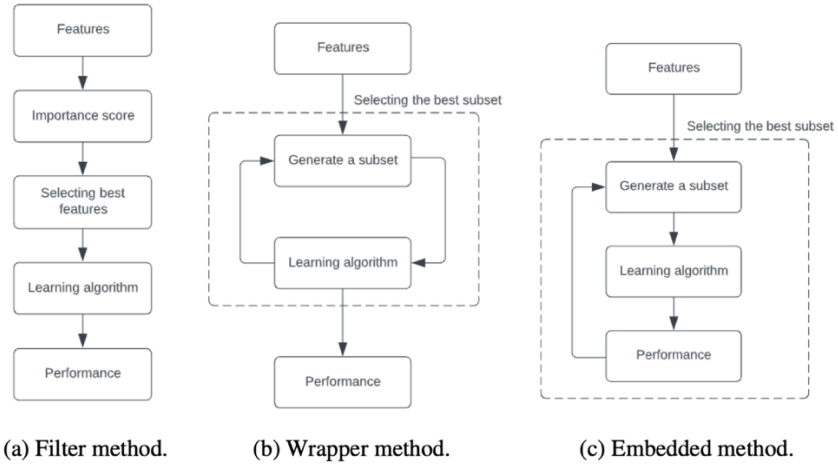

Figure 2. Three types of feature selections (Maguire et.al, 2022)

In general, feature selection can be divided into three techniques (see Figure 2), e.g. the filter method, the wrapper method and the embedded method. The filter method uses statistical methods to select the feature, e.g. correlation analysis, chi-square and even descriptive statistics (e.g. frequency) to find the valuable feature. The wrapper method is used to find the best features that are combined together and have significance to the performance of the machine learning model. The combined features was evaluated according to how well they behave. However, a multiple combination usually requires computing time. Thus, the number of features should be as small as possible, but they have a large impact on the model. The embedded method reduces the computing time if each feature can be applied step by step and the evaluation represents the importance of such a feature. Therefore, the embedded method is in the middle between the two methods where the model only receives the features that are of importance.

3. RELATED WORK

Some research has experimented with SET50. Chaigusin, Chirathamjaree and Clayden (2008) proposed the use of neural networks for the prediction of the Thai stock market (SET). They experimented with the construction of neural network models consisting of three to five layers. For each model, the better model was determined and evaluated using the mean absolute percentage error (MAPE). In their experiments, both internal factors (e.g. the SET index) and external factors (e.g. gold prices, the minimum load ratio (MLR) and other stock indices (Dow Jones, Nikkei, Hang Seng)) were included as features in the constructed model. Sanboon, Keatruangkamala and Jaiyen (2019) experimented with a deep learning model for predicting buy and sell recommendations in SET using long-short term memory. They reported that the LSTM can achieve the highest accuracy among all SET50 stocks. The performance of the proposed model is evaluated and compared with SVM, logistic regression, random forest, decision tree, KNN, and MLP. Rungruang et al. (2019) proposed a prediction accuracy of the direction of SET50 index in Thai stock market by using Support Vector Machines (SVM) model. Inthachot, Boonjing and Intakosum (2016) presented a technique that was a combination of ANN and GA models for predicting the direction of the SET50 stock index. Their experiment showed that the proposed techniques achieve better prediction accuracy than their previous work, that only implemented ANN. Sopipan, Kanjanavajee and Sattayatham (2012) predicted the SET50 index using PCA based multiple regression. Their model included some stock market indices, the gold market and the currency market. They reported the correlation matrix of the SET50 index and the explanatory variables as well as the experiments of the models with different PCAs. Feature selection was not clearly addressed in the above-mentioned papers.

In terms of feature selection, some related works are as follows. Peng et al. (2021) examined a set of 124 technical analysis indicators of seven stock markets. They proposed feature selection using Lasso, TS and SFFS techniques. They focused on neural networks with different settings of hidden layers and dropout rate. Kumari, Patnaik and Swarnkar (2023) mentioned that four to ten input variables may be required for a suitable model. They suggested that the features can be categorized into fundamental features, technical features and macroeconomic features. The technical features are suitable for analyzing a particular stock. The prediction techniques may vary depending on the features. The most commonly used feature selection methods are PCA, GA and decision trees. Another approach for feature selection is ensemble-based feature selection. Aloraini (2015) proposed the combination of stock features and gold features for stock prediction. Tumay, Aydin, and Kozat (2024) presented hierarchical stacking, where an initial machine learning model is trained with a subset of the features and then the output of the model is updated with another machine learning algorithm that uses the remaining features or a subset of them to adjust the predictions of the first layer while minimizing a user-defined loss. Gunduz, Çataltepe and Yaslan (2017) used different types of feature selection and classification methods together. Logistic regression and gradient boosting machine were performed. Büyükkeçeci and Okur (2022) proposed feature selection based on a stability measure. Selection stability is an important property of feature selection algorithms. The stability of the feature selection algorithm is defined as the variation in the results of the selection algorithm due to small differences in the training set (data). They proposed five feature selection techniques such as filter method, wrapper method, embedded method, hybrid method, and ensemble method. Wah et al. (2018) conducted two experiments to compare the search for significant features. Their study compared filter and wrapper methods to maximize classifier accuracy. They concluded that significant features can be better represented by the wrapper method than by the filter method.

4. THE PRELIMINARY EXPERIMENT

A preliminary experiment was conducted to examine the fundamentals of the SET50 stocks and investigate the performance of the technical indicators on different machine learning models. The results of the experiment was considered in the development of the feature selection methodology.

4.1 Correlation Analysis

The various processes of the preliminary study are shown in Figure 3. SET50 stocks were the input for the correlation analysis. In this step, the fundamental data and some common technical indicators were analyzed. In the exploratory analysis of SET50 stocks and the SET market, their correlations were also evaluated. The evaluation based on machine learning models using all 50 stocks may take some time. Therefore, a preliminary experiment was conducted with three samples. In this experiment, the stocks AOT, MINT and EA were selected based on the correlation values. In the preliminary experiment, linear regression, logistic regression and artificial neural network were selected for the prediction of percentage change; decision tree, random forest and XGBoost were selected for the prediction of buy/sell signals. XGBoost and logistic regression were identified as powerful models of the preliminary experiment. Important indicators were ranked and reported.

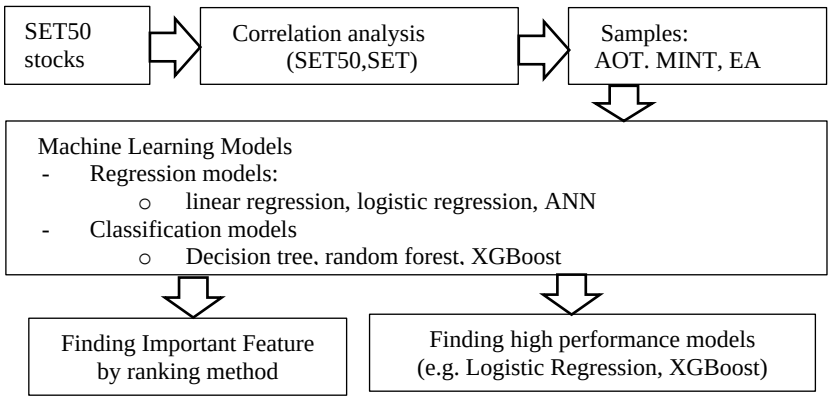

Figure 3. Preliminary experiment

Since the data for the feature selection can be considered in different values and thus the features are not clear which indicators can be selected. Therefore, a preliminary test was conducted considering a small sample. An exploratory analysis was conducted to investigate the relationships between SET50 and SET. The SET50 index available in January 2024 was used for the proposed experiment. The list of SET50 stocks includes ADVANC, AOT, AWC, BANPU, BBL, BDMS, BEM, BGRIM, BH, BTS, CBG, CENTEL, COM7, CPALL, CPF, CPN, CRC, DELTA, EA, EGCO, GLOBAL, GPSC, GULF, HMPRO, INTUCH, IVL, KBANK, KCE, KTB, KTC, LH, MINT, MTC, OR, OSP, PTT, PTTEP, PTTGC, RATCH, SAWAD, SCB, SCC, SCGP, TISCO, TLI.BK, TOP.BK, TRUE, TTB, TU and WHA.

An exploratory analysis was conducted to examine the correlation between the OHLC of SET50 stocks and the SET. It has been found that the correlation values can range from high to low respectively near price, percentage change, volume and volatility (see Table 1). Stocks are ranked from high to low score for each features. The comparison based on open price, high price and low price was omitted as no information is included in the SET. Since the correlation value between SET and SET50 is very low and most of the stocks are not related to SET, the fundamentals of SET were not considered in the feature selection. However, the exploratory analysis provided useful information showing which stocks can influence the SET market.

The experiment to investigate features and their correlation was carried out on 50 shares. The technical indicators – SMA, EMA, RSI and percentage change for the periods 14, 30 and 60 days were used for the correlation analysis. The technical indicator with a similar time period shows higher correlations, e.g. the percentage change over 14 days shows a high correlation with SMA14, RSI14 and EMA14 (see Figure 4 x-axis 6,7). Among the indicators, the SMA14 has a high correlation with the EMA14 (see Figure 4 x-axis 10). The volume has a negative correlation with the percentage change of 1 day (see Figure 4 x-axis 11).

Table 1. Correlations of SET and SET50 index

Stocks	Close	Stock	Change	Stock	Volume	Stock	Volatility
RATCH	0.93802	GULF	0.61281	BBL	0.49435	CPF	0.480143
EA	0.91301	EA	0.59177	TRUE	0.48930	PTT	0.456681
GPSC	0.89368	KTC	0.58778	GULF	0.46083	BEM	0.392176
KTC	0.88344	SAWAD	0.58677	SCC	0.42970	OR	0.372512
SCGP	0.86771	GPSC	0.57184	KBANK	0.39216	TLI	0.322013
HMPRO	0.85355	CPALL	0.55251	PTTEP	0.39198	SCB	0.315708

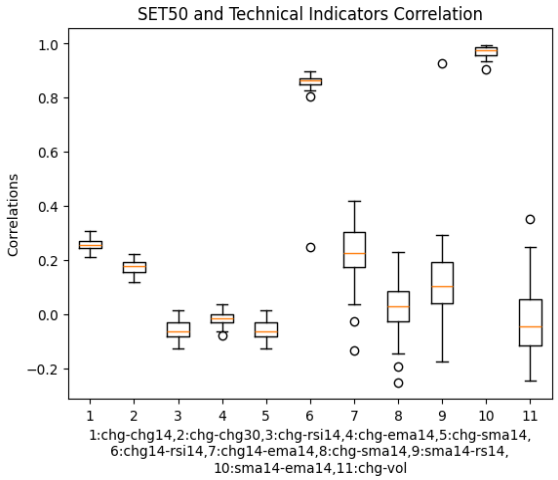

Figure 4. Correlations of technical indicators

Since the RSI is important for the percentage change, the RSI correlation value was a condition for the selection of the sample stocks. The stocks can be considered as stocks with high, neutral and low correlation values and one from each group was selected for the experiment. They were AOT, MINT and EA respectively. These stocks were analyzed using machine learning models.

4.2 The Experiment on Machine Learning Models

The stock prediction to predict the percentage change was carried out using regression models – linear regression, logistic regression and artificial neural network. Table 3 shows the R-squared value of linear regression with one parameter and with multiple parameters. The RSI 14 days has a better performance than other indicators in a model with one parameter. There were 470 and 453 observed values (data from 2022-01-01 – 2023-12-31) for the 14-day and 30-day periods. The R-squared values are very low, which means that the independent parameter may not be able to explain the dependent parameter.

The logistic regression (Table 2) was tested with a maximum depth of 30 and a minimum sample size of 3. MINT and EA have similar results, with the 30-day period performing better than other indicators, but EMA14 performing better on AOT. For the multiparameter model, the 30-day period model performed well. The 30-day period performed better for AOT and EA. ANN can be one of the best performing models when the parameter setting was explored. However, in this experiment, ANN was run under fixed conditions with 1000 epochs, 4 dense layers and ReLU for activation. Thus, the result of training performance was poor.

Table 2. R-squared of logistic regression (train data)

Models	RSI30	RSI14	EMA14	SMA14	Multiple 30	Multiple 14
Linear regression						
AOT	0.001	0.005	0.000	0.000	0.147	0.190
MINT	0.002	0.005	0.000	0.000	0.229	0.244
EA	0.001	0.001	0.004	0.004	0.041	0.097
Logistic regression						
AOT	0.43063	0.37978	0.47935	0.37949	0.63109	0.57445
MINT	0.48032	0.43847	0.40616	0.45016	0.64832	0.64816
EA	0.47842	0.45866	0.45191	0.36394	0.60882	0.58383
Neural Network						
AOT	0.26379	0.20738	0.08915	0.09014	0.69221	0.65010
MINT	0.12065	0.11058	0.09430	0.09763	0.30619	0.92346
EA	0.09621	0.05507	0.07869	0.04347	0.54837	0.52474

For the classification model, the stock prediction of the buy/sell signal strategy was performed. The target was determined by the buy signal. For example, if the percentage change on the next day is greater than 0, the signal is buy (1), otherwise do nothing (0). Decision tree, random forest and XGBoost models were tested. All models were specified with 5 maximum depths. Table 4 shows the results of the experiment. The EMA performed better with the random forest and XGBoost models, but performed poorly with the decision trees. The use of multiple parameters leads to better accuracy than the use of a single parameter. The training data performs better than the test data for all models.

Table 3. Accuracy of decision tree, random forest, and XGBoost (train data)

Models	RSI30	RSI14	SMA14	EMA14	Multiple 30	Multiple 14
Decision Tree						
AOT	0.65745	0.63466	0.63031	0.61968	0.69889	0.65333
MINT	0.64088	0.62133	0.63297	0.66755	0.68508	0.66933
EA	0.65745	0.66133	0.62765	0.65957	0.66574	0.65066
Random forest						
AOT	0.64364	0.69333	0.71542	0.72340	0.79281	0.77066
MINT	0.70441	0.70400	0.71542	0.71809	0.80939	0.82666
EA	0.69613	0.71466	0.69946	0.73670	0.79005	0.79733
XGBoost						
AOT	0.85359	0.85333	0.81383	0.87500	0.95027	0.98133
MINT	0.83425	0.83466	0.82978	0.85638	0.95027	0.97600
EA	0.85635	0.84800	0.82180	0.86170	0.94475	0.95466

The ranking of the technical indicators according to the ML model is shown in Table 4. All indicators were rated at least 2 out of 3 by all stocks. Some ranks cannot be scored (indicated by -). Since the experiment conducted resulted in different ranks for different stocks, each stock was further examined. The result of prediction using SET50 stocks with logistic regression and random forest is shown in Table 5. Each stock in SET50 behaved differently for different indicators. The table shows the number of stocks with the indicators in the first rank. For example, the 30-day RSI is the strongest indicator in the one-parameter model for 16 stocks with logistic regression, while the 14-day EMA performs better for 17 stocks with random forest. Using a multiple parameters model with a 30-day period showed good results for most stocks.

Table 4. Rank of indicators of AOT, MINT, and EA

ML models	Single-parameter model				Multiple-parameter model	
	(1)	(2)	(3)	(4)	(1)	(2)
Linear regression	RSI14	RSI30	-	-	14 days	30 days
Logistic regression	RSI30	-	-	SMA14	30 days	14 days
ANN	RSI30	RSI14	SMA14	EMA14	30 days	14 days
Decision tree	-	-	-	-	30 days	14 days
Random forest	EMA14	SMA14	RSI14	RSI30	14 days	30 days
XGBoost	EMA14	RSI30	RSI14	SMA14	14 days	30 days

Table 5. Rank of indicators of SET50 stocks

ML Model	Single-parameter model				Multiple-parameter model	
	RSI30	RSI14	SMA14	EMA14	30 days	14 days
Logistic regression	16	12	11	11	31	19
Random forest	13	8	12	17	36	14

5. THE PROPOSED FEATURE SELECTION METHODOLOGY

In this work, all features were technical indicators, which were categorized into three groups, described as follows:

(i) Group A - high performance features. The indicators in this group have a high correlation with the objective (target) of the machine learning model used.
(ii) Group B – support features. These are features that can improve ML performance, but their importance is lower than that of Group A. For example, their correlation with the target may be lower than that of the features in group A.
(iii) Group C – specific features. These are features for specific conditions. For example, strategy indicator can be a specific feature for stock prediction.

The proposed feature selection methodology (Figure 5) consists of a filtering process and a wrapping process. The filtering process is the process of selecting indicators based on some criteria, which are explained in more detail below.

- First, the correlation between indicator and target (e.g. percentage change) was examined. The RSI shows a high correlation with the percentage change; therefore, it can be selected for group A. The current price of a stock can be influenced by different time periods. Moreover, using multiple parameters increases the performance of the ML model. Therefore, a selected feature can be specified with different time periods, e.g. RSI14 (14-days), RSI30 (30-days) and RSI60 (60-days). In addition, the importance of a feature can be considered based on a machine learning model to include the feature in this group. For example, the importance is calculated based on the information gain (tree-based model) and the importance is considered based on the coefficient (regression model). The indicators determined in this step belong to group A.
- Second, other indicators that may have lower correlation values can be considered as supportive performance feature. The chi-square test can be used to test whether each indicator is independent of another indicator. For example, EMA and SMA can serve as support feature. The indicators determined in this step belong to group B.
- Third, an extension of the features can be considered. For example, when predicting upward/downward movements of stocks, strategy techniques that predict upward/downward movements can be applied, so that indicators related to such a strategy can be specified as specific features. Different prediction targets require specific indicators. The indicators identified in this step belong to group C.

In this experiment, RSI14, RSI30 and RSI60 were selected for group A. The chi-square test shows that SMA14 and SMA30 are interdependent for 23 stocks. SMA14 and EMA14 were selected as support features for Group B. Specific features are MACD and SMA/EMA strategy. The MACD was calculated from the difference between the EMA14 and the EMA30. The SMA/EMA strategy was calculated by the distance between SMA14 and EMA14. Therefore, all 7 features were used for the next process. The importance of the feature based on the technique used such as information gain (measured by the tree-based model) and coefficient (measured by the regression model) were further conditions to confirm that the features should be defined for the test.

The wrapping process is the process of examining the features in an ML model. The combination of indicators from three groups was defined for wrapper level W1 – W7. For example, Wrapper W1 uses features from group A, Wrapper W4 uses features from groups A and B together, and Wrapper W7 uses all groups together. Various combinations can be

considered, as shown in Figure 5 (bottom left corner). These are represented as multiple subgraphs. For example, W1, W2, W4 and W7 are nodes in a subgraph representing the combination of features in groups A, B and C; W1, W3 and W5 are nodes in a subgraph representing the combination of features in groups A and C.

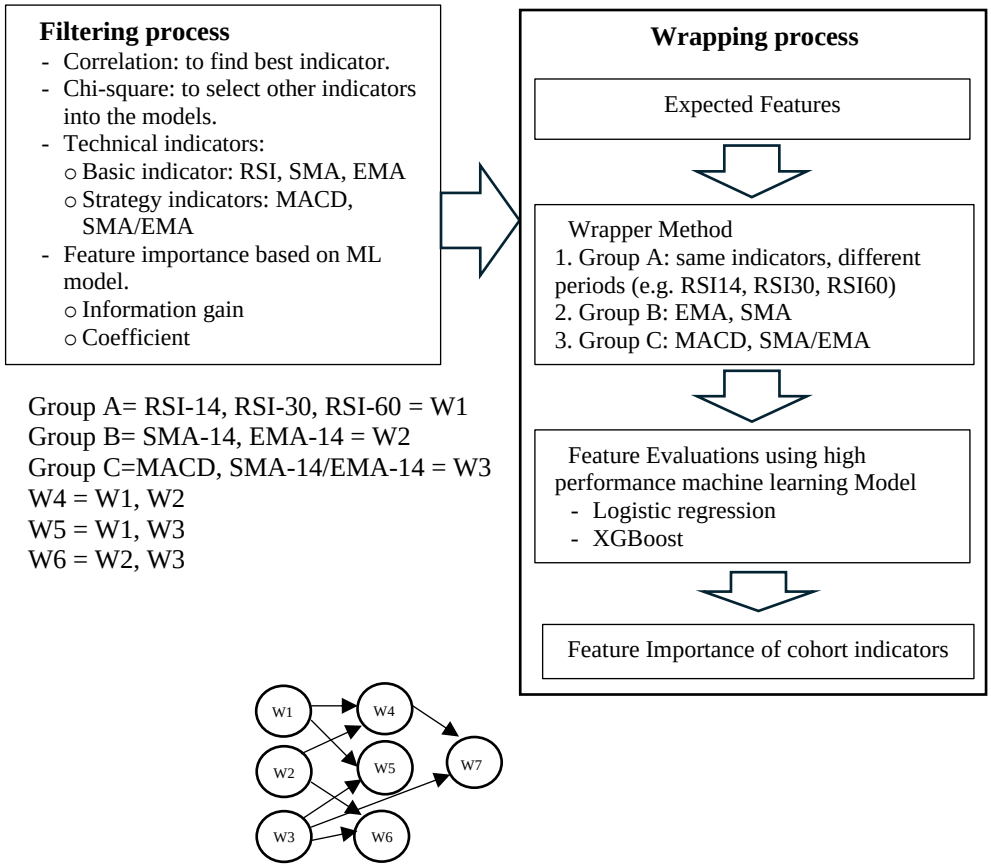

Figure 5. The proposed feature selection methodology

6. THE FEATURE IMPORTANCE

The ranking of the indicators using the feature importance score (calculated by the information gain) was performed with XGBoost with 10 estimators, a maximum depth of 5 and a learning rate of 1.0 to investigate their importance in each stock. The result is shown in Table 6. In wrapper W1 using XGBoost, RSI14 was at the top of 43 stocks. In Wrapper W4, EMA ranked first out of 36 stocks, while RSI30 and RSI60 ranked second. It can be seen that each indicator can have a different importance in the different wrappers. The importance of these features represents the importance of a technique used. However, they are not representative of overall

performance when used in conjunction with the others. To understand the importance of indicators when applied to a model, it is necessary to observe their behaviour when applied with other indicators. Currently, Lasso and Boruta (Kursa and Rudnicki, 2010) are feature importance assessment techniques to reduce the feature, but these have no purpose to observe the performance when cohort indicators are used. In case there are too many indicators, Lasso and Boruta can be used in the filtering process of the proposed feature selection methodology. In this study, the importance of features in the wrapping process was calculated. The feature importance is analyzed for each wrapper level. In this experiment, the R-squared for the logistic regression model and the precision for XGBoost were analyzed. Precision, accuracy, recall and F1 score can serve as a measure for the classification model and calculate the importance of the features.

Table 6. Rank of indicators of SET50 stocks

ML Model	**Feature Importance (XGBoost)**			**Performance (Logistic)**		
	Ranked-1	Ranked-2	Ranked-3	A,B,C	A,C	B,C
W1 (RSI14, RSI30, RSI60)	RSI14 (43)	RSI30 (4)	RSI60 (3)	-	-	-
W2 (SMA14, EMA14)	EMA14(32)	SMA14(18)	-	-	- -	-
W3 (SMA/EMA, MACD)	SMA/EMA (50)	-	-	-	-	-
W4 (W1, W2)	EMA14 (36)	RSI30 (7), RSI60(7)	-	(-11)	-	-
W5 (W1, W3)	SMA/EMA(28)	RSI14(15)	RSI60(5)	-	(- 3)	-
W6 (W2, W3)	MACD (26)	EMA14(24)	-	-	-	(-4)
W7 (W1,W2, W3)	SMA/EMA(30)	RSI30(8)	RSI60(2)	(- 8)	-	-

Here, the importance of the indicators from the performance of a wrapper was calculated using formula (4) as follows.

$$P_{wi} = \sum_{wj=\, wrapper\ before} P_{wj}\ x\ SP_{wi} \quad = \sum_{wj=\, wrapper\ before} P_{wj/wi} \tag{4}$$

Where P_{wi} is the performance of wrapper *wi*, P_{wj} is the performance of the previous wrapper before *wi*, SP_{wi} is a shared performance weighting at *wi*, and $P_{wj/wi}$ is the performance of the pervious wrapper (used as a group indicator) at wrapper *wi*.

For example (see Figure 6), the performance at W4 is 0.674. The shared performance at this wrapper is SP_{W4}, which is calculated as 0.674/(0.622 + 0.601) = 0.551. The performance of W1 using Group A as indicators in wrapper W4 ($P_{W1/W4}$) is 0.343 (from 0.622 x 0.551), and the performance of W2 using Group B as indicators in wrapper W4 ($P_{W2/W4}$) is 0.331 (from 0.601 x 0.551). In this example, the significance of Group A at wrapper W4 is higher than that of Group B. The calculation of the performance at wrapper W7 can be done in the same way. The performance of a group at the next wrapper can be lower or higher than at the previous wrapper. For example, the performance of group A used together with B is 0.343 at wrapper W4, but rises to 0.417 at wrapper W7 (A, B and C were indicators). However, the performance was a shared value of group A and B, which means that each individual performance was 0.208. However, by combining several groups of indicators, the performance was increased to 0.711 for wrapper W7.

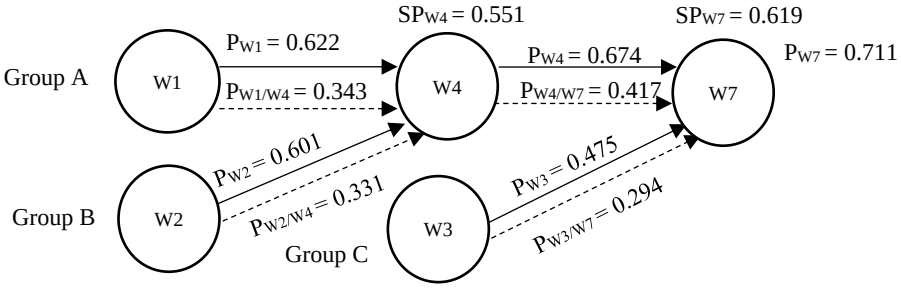

Figure 6. Wrapper and performance

In the logistic regression, the performance of the next wrapper was higher or lower than that of the previous wrapper. For example (see Table 7), when the indicators of groups A, B and C were combined at wrapper W4, W5, W6 and W7, there were 11 stocks with lower performance at wrapper W4 and 8 stocks at wrapper W7. Table 7 shows the performance when using XGBoost on the training data. The feature importance value of group A and B was given for wrapper W4 (importance column). Group A is significantly more important than Group B due to the feature importance for wrapper W4. There were stocks with overfitting model, and the best performance (underlined number) was varied in each wrapper. Table 8 shows the results of the logistic regression. Most stocks have higher performance in wrapper W7, and it is possible to find optimal features for the classification model. In contrast, the regression model may need more features as the performance was increased for more complex wrappers.

The advantages of the proposed feature selection with the wrapper method are as follows.

- It supports the search for the optimal features. The optimal features were in a wrapper layer with fewer features but high performance. For example, MINT (see Table 8) had a similar score in wrappers W5 and W7, but fewer features were used in W5, therefore the optimal features are in group A (RSI14, RSI30, RSI60) and C (SMA/EMA, MACD).
- It represents a significance of a group of features. For example (see Table 8), in wrapper W4, group A has a greater significance than group B in most stocks, except for BBL.
- It can be used to calculate stability. The stability of a feature selection algorithm refers to the robustness of its feature preferences with respect to small changes in the data (Hamer and Dupont, 2021). The stability value of a feature can be calculated by the average of the performance when this feature has been used in the combination wrapper (Büyükkeçeci and Okur, 2022). For example, when using group A in AOT, the stability weight can be 0.9949325 calculated from the average values of columns W4 and W7.

Table 7. Performance and feature importance using XGBoost

Stocks	Performance XGBoost (Precision)							Importance	
	W1	W2	W3	W4	W5	W6	W7	A/W4	B/W4
ADVANC	0.993243	0.963526	0.883281	0.989865	1.000000	0.974763	0.989865	0.502449	0.487416
AOT	0.989865	0.969605	0.864353	0.993243	0.996622	0.968454	0.996622	0.501756	0.491487
AWC	0.993243	0.948328	0.870662	1.000000	1.000000	0.974763	0.993243	0.511567	0.488433
BANPU	0.972973	0.939210	0.854890	0.983108	1.000000	0.958991	0.989865	0.500233	0.482875
BBL	0.949324	0.951368	0.876972	0.979730	0.993243	0.977918	0.972973	0.489338	0.490392
MINT	0.979730	0.975684	0.880126	0.993243	0.996622	0.974763	0.996622	0.497649	0.495594
MTC	0.966216	0.942249	0.870662	0.983108	0.976351	0.987382	0.986486	0.497727	0.485381
OR	0.983108	0.969605	0.889590	1.000000	0.989865	0.981073	1.000000	0.503458	0.496542
OSP	0.976351	0.963526	0.949527	0.996622	0.996622	1.000000	0.996622	0.501605	0.495016
EA	0.983108	0.963526	0.924290	0.983108	0.996622	0.977918	0.996622	0.496499	0.486609

Table 8. Performance and feature importance using logistic regression

Stocks	Performance Logistic Regression (R-squared)							Importance	
	W1	W2	W3	W4	W5	W6	W7	A/W4	B/W4
ADVANC	0.621801	0.600750	0.474679	0.673652	0.655776	0.598592	0.710985	0.342626	0.331026
AOT	0.619637	0.554233	0.473068	0.663663	0.665687	0.622672	0.712094	0.350320	0.313343
AWC	0.607105	0.575186	0.449674	0.641791	0.659524	0.612950	0.663201	0.329559	0.312232
BANPU	0.612984	0.579350	0.444715	0.677255	0.699408	0.627885	0.706065	0.348180	0.329075
BBL.BK	0.605216	0.573987	0.454679	0.615159	0.644432	0.637861	0.652292	0.315725	0.299434
MINT	0.663881	0.584541	0.449702	0.679770	0.684057	0.658959	0.688632	0.361485	0.318284
MTC	0.647361	0.587527	0.457776	0.654449	0.669543	0.576704	0.654726	0.343079	0.311370
OR	0.579583	0.558223	0.433182	0.591719	0.624886	0.596942	0.620821	0.301414	0.290305
OSP	0.544171	0.559304	0.411926	0.592114	0.659859	0.577698	0.647592	0.291997	0.300117
EA	0.617099	0.578732	0.472827	0.650719	0.657720	0.622951	0.687551	0.335798	0.314920

7. CONCLUSION

Feature selection is important for the performance of a machine learning model. The proposed feature selection methodology has been shown to be a promising technique as the combination of filter and wrapper method refines the feature selection process. The proposed feature importance of indicators is an insightful value for assessment, that can increase the performance of the model and reduce unnecessary wrappers. The proposed feature importance represents how important the features are to the training model if they are cohort indicators in each wrapper. The proposed wrapper method can represent more significant features better than using only the filtering method (Wah et al., 2018). The experiment conducted has shown that each stock may have different significant data that may respond to different indicators. Therefore, the proposed methodology is a flexible and practical approach for stock prediction. In this work, the features were considered as cohort features. In this way, the method is more flexible to handle than observing each individual. However, an importance score can be derived for each feature.

XGBoost and logistic regression were used in the experiments, but other machine learning models or even deep learning models can also be tried out. For machine learning models, the parameters need to be adjusted to increase the performance and reduce the losses of the model. This research focused on feature selection, but not on selecting the best machine learning model. On the contrary, parameter tuning can take advantage of using a fixed set of selected features (best feature).

REFERENCES

Aloraini, A., 2015. Penalized ensemble feature selection methods for hidden associations in time series environments case study: equities companies in saudi stock exchange market. *Evolving Systems*, Vol. 6, No. 2, pp. 93-100.

Büyükkeçeci, M. and Okur, M. C., 2022. A comprehensive review of feature selection and feature selection stability in machine learning. *Gazi University Journal of Science*, Vol. 36, No. 4, pp. 1506-1520.

Chaigusin, S., Chirathamjaree, C. and Clayden, J., 2008. The use of neural networks in the prediction of the stock exchange of Thailand (SET) Index. *Proceedings of 2008 International Conference on Computational Intelligence for Modelling Control & Automation*. Vienna, Austria, pp. 670-673.

Gunduz, H., Çataltepe, Z. and Yaslan, Y., 2017. Stock daily return prediction using expanded features and feature selection. *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol. 25, No. 6, Article 32.

Hamer, V. and Dupont, P., 2021. An importance weighted feature selection stability measure. *Journal of Machine Learning Research*, Vol. 22, No. 116, pp. 1-57.

Htun, H. H., Biehl, M. and Petkov, N., 2023. Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, Vol. 9, No. 26.

Inthachot, M., Boonjing, V. and Intakosum, S., 2016. Artificial neural network and genetic algorithm hybrid intelligence for predicting Thai stock price index trend. *Computational Intelligence and Neuroscience*, Vol. 2016, pp. 1-8.

Kumari, B., Patnaik, S. and Swarnkar, T., 2023. Feature selection for stock price prediction: a critical review. *International Journal of Intelligent Enterprise*, Vol. 10, No. 1, pp. 48-72.

Kursa, M. and Rudnicki, W., 2010. Feature Selection with Boruta Package. *Journal of Statistical Software*, Vol. 36, pp. 1-13.

Maguire, T., Manuel, L., Smedinga, R. A. and Biehl, M., 2022. A review of feature selection and ranking methods. *19th SC@ RUG 2021-2022*, 15.

Peng, Y., Albuquerque, P. H., Kimura, H. and Saavedra, C. A., 2021. Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators. *Machine Learning with Applications*, Vol. 5, 100060.

Rungruang, C., Srichaikul, W., Chanaim, S. and Sriboonchitta, S., 2019. Prediction the direction of SET50 index using support vector machines. *Thai Journal of Mathematics*, pp. 153-165.

Sanboon, T., Keatruangkamala, K. and Jaiyen, S., 2019. A deep learning model for predicting buy and sell recommendations in stock exchange of thailand using long short-term memory. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 757-760.

Sopipan, N., Kanjanavajee, W. and Sattayatham, P., 2012. Forecasting SET50 index with multiple regression based on principal component analysis. *Journal of Applied Finance and Banking*, SCIENPRESS Ltd, Vol. 2, No. 3, pp. 1-10.

Tumay, A., Aydin, M. E. and Kozat, S. S., 2024. Hierarchical Ensemble-Based Feature Selection for Time Series Forecasting. *arXiv preprint arXiv:2310.17544*.

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S. and Fong, S., 2018. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, Vol. 26, No. 1, pp. 329-340.