Hybrid Genetic Feature Selection and Support Vector Machine for Prediction LQ45 Index in Indonesia Stock Exchange

Abdul Syukur[1, a)], Deden Istiawan[2, b)], Wellie Sulistijanti[2, c)], Ahmad Ilham[3, d)]

[1]*Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia*
[2]*Department Statistics, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Semarang, Indonesia*
[3]*Department of Informatics, Universitas Muuhammadiyah Semarang, Semarang, Indonesia*

[a)] Corresponding author: abah.syukur01@dsn.dinus.ac.id
[b)] deden.istiawan@itesa.ac.id
[c)] wellie.sulistijanti@itesa.ac.id
[d)] ahmadilham@unimus.ac.id

Abstract. Stock market predictions play a very important role and have attracted a lot of attention, this is because stock price predictions can bring huge profits in the future by making the right decisions. LQ45 index is one of the most popular and influential stock indices on the Indonesia Stock Exchange. LQ45 index is an index that measures the price performance of 45 stocks that have high liquidity and large market capitalization and are supported by good company fundamentals and adjusted every six months at the beginning of February and August. Stocks with declining performance will be excluded from the index. Prediction of stock composition in the LQ45 index is an important issue in investment, always attracts the attention of public investors and academics for research. Prediction LQ45 index will be very useful for investors to be able to see how the prospects for investing in a company's stock in the future. In order to build a better model to predict the composition of the LQ45 index effectively and efficiently, we developed a prediction model with a hybrid approach using genetic algorithms and supporting vector machines to predict which companies will enter and leave the LQ45 index. This proposed algorithm namely GA-SVM. The results show the proposed algorithm yield excellent performance compared with PSO-SVM, FS-SVM and BE-SVM and promising results with the accuracy is 93.49%.

INTRODUCTION

Investment refers to placing some funds at a specific time in order to make profits in the future [1-2]. Stock market predictions play a very important role and have attracted a lot of attention, this is because stock price predictions can bring huge profits in the future by making the right decisions [3]. Liquid stock 45 or better known as LQ45 is index stock market index on the Indonesia Stock Exchange which consists of 45 leading stocks with the largest market capitalization and the highest transaction value. LQ45 index is an imaginary portfolio to measure changes in stock prices in the Indonesian capital market. LQ45 index aims to complement the Composite Stock Price Index. This is because LQ45 index and the composite stock price index have a very high correlation, if LQ45 index rises, it is certain that the Composite Stock Price Index will also rise. Every six months the Indonesia Stock Exchange will hold evaluation of the stocks that are included in the LQ45 index [5]. Apart from having a high level of liquidity, the main characteristic of the stocks included in the LQ45 index is their high market capitalization. Therefore, some investors consider LQ45 index shares to be blue chip stocks.

The LQ45 index is a reliable and objective tool for investment managers, financial analysts, capital market observers and investors to monitor trends in the price of actively traded stocks. The LQ45 index is a stock index indicator of the Indonesian Stock Exchange, which can be used for evaluating the performance of stock trading.

Among the stocks on the Indonesian capital market, the existing LQ45 stocks are the stocks most in demand by investors. This is because LQ45 share capital has a high capital and trading frequency, so the growth prospects and the financial status of the stock are very good. Until now, companies that are included in the LQ45 market index are still among the most attractive company stocks for investors to invest their capital in, but there are still many investors who do not understand how to form the right stock portfolio and analyse which stocks have the possibility of bringing large returns in the future. Prediction LQ45 index will be very useful for investors to be able to see how the prospects for investing in a company's stock in the future.

Stock market prediction became a big challenge for the capital market analyst and investors to get benefit from the money that has been invested, this is because the data is very large, noise and non-stationary [5]. The stock market prediction has become an increasingly important issue in the present time. One of the algorithms employed is technical analysis, but such algorithms do not always yield accurate results. However, it is quite difficult to predict stock market movements. So, it is important to develop algorithms for a more accurate prediction [6].

There are three main stock market prediction methods, including technical analysis (chart), fundamental analysis, and machine learning algorithms [7-8]. Fundamental analysis is an analysis that considers things that can move stock prices, including financial performance, level of business competition, industry potential, market and economic analysis, both macro and micro [9-10]. Fundamental analysis is subjective, because it uses a lot of assumptions and this is what causes the results of the analysis to be different. Technical analysis is an analysis of stock price movements based on past stock price movements. Future stock price movements can be analyzed by looking at past stock price movements [11-13]. In this study we focus on machine learning approaches. Stock market predictions using machine learning have been studied extensively, because they can provide accurate analysis results [14]. For stock market prediction, supervised learning machine learning approach has shown great promise [15].

Several machine learning classification algorithms have been applied for stock market prediction, let in Artificial Neural Network (ANN) [16], Support Vector Machines (SVM) [17-19]. In particular, ANN is often used in the stock market because it is well known that the ability to predict is stronger than other capabilities. However, it is widely reported that artificial neural network models generally require large amounts of labelled training data to estimate the distribution of input patterns and it is difficult to generalize the results [20], because convergence speed is slow, easy to get stuck to local extremes and overfitting nature [19,21]. Support vector machine (SVM) is a supervised learning algorithm based on statistical learning theory [22], the problem of slow convergence speed, easily trapped in local extremes and the overfitting nature of the ANN algorithm can improve the SVM algorithm. In addition, the SVM algorithm has excellent generalizability for small samples [18, 23-24]. Although SVM has excellent generalizability, classification performance is affected by dimensions or number of features [19]. SVM is a powerful prediction tool for stock market predictions, however the classification performance is affected by dimensions or number of features [6]. According to [25], if the SVM algorithm is applied without considering feature selection, the input space size is large and unclean, thereby reducing SVM performance.

Feature selection is indispensable for machine learning when dealing with high-dimensional data and noisy attributes. Feature selection is also the most important part to improve accuracy performance [26]. Feature selection is a technique used to reduce the complexity of the attributes to be managed in processing and analysis. This technique is used to see the most significant feature subset of the data set. The benefits of feature selection are twofold: it improves classification performance and helps reduce domain features, eliminating redundant features [27-28]. Not all of these features are equally important in a specific application problem. Some features can be removed to improve performance. As a result, we propose removing redundant and irrelevant data while retaining the data discriminating power through feature selection. Genetic algorithm is a method that can be used to select features. This algorithm is applied to the classification process to find solutions in the full search space and use a global search ability. The application of the Genetic Algorithm is carried out to perform feature selection to be able to parse the number of features used but still get good results [29-30].

In the research developed, the prediction model is based on a hybrid approach, which combines a Genetic Algorithm (GA) which is applied to select relevant features and Support Vector Machines (SVM) to predict companies that will enter and leave the LQ45 index list. Prediction LQ45 index can help investors and capital market observers in making an initial indication of the company's shares that will be included in the LQ45 index list.

MATERIALS AND METHODS

Data Description

The LQ45 index consists of 45 issuers with high liquidity, which were selected through several selection criteria. In addition to assessing liquidity, the selection of these issuers also considers market capitalization. The Indonesia Stock Exchange (ISE) regularly monitors developments in the performance of issuers included in the LQ45 index calculation. Every three months, an evaluation of the order of the shares is carried out. Issuers that do not meet the LQ45 stock evaluation criteria will be removed from the list. Share replacement will be carried out every six months, namely at the beginning of February and August.

The dataset of this study was issued from (https://www.idx.co.id) with a total of 261 companies approved in the ISE from 2015 to 2018 years. The number of companies included in LQ45 was 180 and 81 non-LQ45. A brief description of the dataset of this study is presented in TABLE 1.

TABLE 1. Dataset description

No	Variable	Description
1	Volume	Volume is a measure of how much of a given financial asset has been traded in a given period of time
2	Value	Value is a range of prices where the majority of trading volume took place on the prior trading day
3	Frequency	Frequency is basically the number of trades executed in a specific time interval
4	Days	The trading day or regular trading hours (RTH) is the time span that a particular stock exchange is open
5	Earnings per Share	Provides the profitability indication of a firm, and can be determined by dividing the firm's net income with its whole number of remaining stocks
6	Book Values per Share	Market value ratio that weighs Stockholders' equity against shares outstanding. In other words, the value of all shares divided by the number of shares issued
7	Debt to Assets Ratio	Leverage ratio that measures the amount of total assets that are financed by creditors instead of investors
8	Debt to Equity Ratio	A financial, liquidity ratio that compares a company's total debt to total equity
9	Return on Assets	This ratio signifies the proportion of earnings a firm earns about the firm's overall assets or resources. Thus, an indication of how profitable a firm is relative to the firm's total resources or assets
10	Return on Equity	This ratio offers an overview of how well the shareholder's funds were used and the gain made out of its investment. When ROE is low, it implies that the shareholder's funds were not used properly
11	Gross Profit Margin	*Gross profit margin* is a ratio that indicates the performance of a company's sales and production
12	Operating Profit Margin	The earnings that a business generates from its operating activities
13	Net Profit Margin	The percentage of revenue left after all expenses have been deducted from sales
14	Pay-out Ratio	Shows the proportion of earnings paid out as dividends to shareholders
15	Yield	Yield refers to the earnings generated and realized on an investment over a particular period of time

Feature Selection

Process of selecting relevant features in a dataset is called feature selection. Feature selection is an important pre-process to apply to a dataset before applying a classification algorithm, because the feature selection method can eliminate redundant and irrelevant features. The term feature selection refers to an algorithm that can generate a subset of a set of feature inputs. Reducing data dimensions can also reduce hypothetical space and allow algorithms to operate more quickly and effectively. Feature selection is an important part of the machine learning process because it defines the quality of the input data. Feature selection is done by selecting the relevant features that affect the classification results. The features that can decrease the accuracy of the prediction model are irrelevant features and redundant features. Therefore, only relevant and non-redundant features will be used as input for the prediction model.

Approaches for feature selections can be categorized into two models, namely a filter method and a wrapper method [31-32]. FIGURE 1, gives the outline of the two algorithms: "a" as filter algorithm and "b" as wrapper algorithm. Filter method uses the feature ranking technique as the basic criterion for the feature selection sequence. The ranking method is used because of the simplicity of the process in selecting features. Generally, this method makes use of the threshold value to determine how many minimal features will be used as input to the prediction model [33]. The minimum number is sorted by feature rating. The feature with the highest score will be the candidate for predictive modeling input [34].

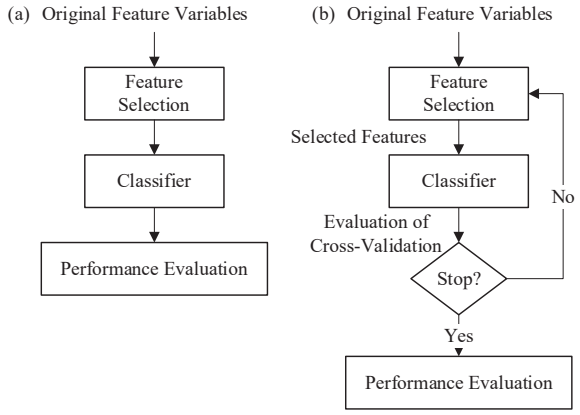

FIGURE 1. The comparison in (a) filter algorithms, and (b) wrapper algorithms

Wrapper method is a method that uses a classification algorithm as a black box to find the best feature subset. The basic idea of this method is to search through a subset of the feature space by inputting all features from the dataset. Then evaluate each candidate feature subset using the selected Feature Selection criteria. This process is repeated until the best suitable feature subset is found based on the feature evaluation criteria. Wrapper approach divides the task into three stages: feature search, classification algorithm, and feature subset evaluation [35]. Forward Selection and Backward Elimination is wrapper method can play different roles with different learning algorithms for different datasets and that no learning scheme dominates, i.e., always outperforms the others for all datasets. This means the different learning schemes for different datasets should be chosen, and consequently, the evaluation and decision process are important. In fact, many experimental results report that the wrapper method is better than the filter method [24].

Because of the feature selection techniques use local search throughout the entire process, optimal solutions are quite difficult to be achieved. Metaheuristic optimization can find a solution in the full search space and by using a global search ability, the ability of finding high quality solutions within a reasonable period of time can be significantly increased [29]. Cano et al. (2003) have shown that better results can be obtained with the metaheuristic optimization method as compared to many traditional and non-evolutionary feature selection methods in terms of higher classification accuracy [36]. The metaheuristic optimization methods have been applied for feature selection.

Among the different approaches to feature selection, Genetic Algorithm (GA) is one of the most popular methods when faced with datasets that have high feature space dimensions [37]. In literature, some feature selection algorithms based on genetic algorithm (GA) were proposed [38-39]. In this study, we combine Genetic Algorithm for selecting

features and Genetic Algorithm (GA) to improving accuracy of Support Vector Machines (SVM) for stock market prediction.

Support Vector Machine

Support Vector Machine is a technique that is relatively new compared to other techniques, but has better performance in various fields. Support vector machine is a learning system that uses a hypothetical space in the form of linear functions in a high-dimensional feature space, trained with a learning algorithm based on optimization theory by implementing biased learning derived from statistical learning theory [40].

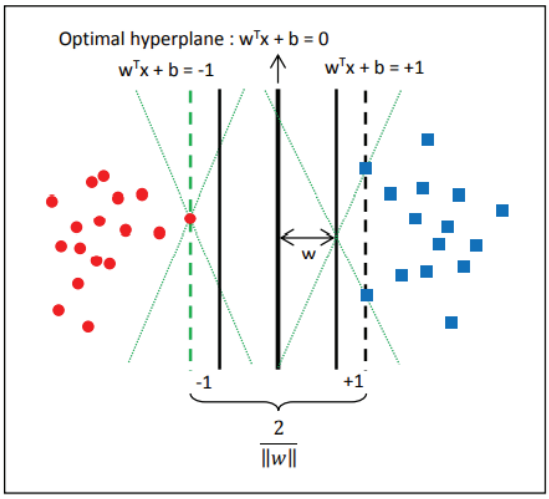

FIGURE 2. Illustration of finding the optimal separator function

The concept of SVM can be explained simply as an effort to find the best hyperplane that functions as a separator of two classes in the input space. Patterns are members of two classes: +1 and -1 and share an alternative dividing line [41]. Binary support vector machine is one of the most powerful machine learning algorithms in the past 15 years [42]. Margin is the distance between the hyperplane and the closest pattern from each class. This closest pattern is known as a support vector. The effort to find the location of this hyperplane is at the core of the learning process on the Support vector machine [40]. The hyperplane found by SVM is illustrated as in Figure 2 its position is in the middle between the two classes, meaning that the distance between the hyperplane and data objects is different from the adjacent (outermost) class which is marked with a red circle and a blue box. In SVM the outermost data object closest to the hyperplane is called a support vector. FIGURE 2 shown an illustration of the idea optimal hyperplane for linearly separable patterns. This training set can be separated by the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ where **w** is the weight vector and b are the bias.

Available data is denoted as $\overrightarrow{X_l}\epsilon\mathcal{R}^d$ while the respective labels are denoted $y_i\epsilon\{-1, +1\}$ for $i = 1,2, \dots, l$ where l is the amount of data. It is assumed that both classes 1- and +1 can be completely separated by a hyperplane with d dimension which is defined:

$$\vec{w}.\vec{x} + b = 0 \tag{1}$$

Pattern $\overrightarrow{x_i}$ which belongs to class -1 (negative sample) satisfy the following inequalities

$$\vec{w}.\vec{x} + b \leq -1 \tag{2}$$

Meanwhile, Pattern $\overrightarrow{x_i}$ which belongs to class +1 (positive sample) satisfy the following inequalities.

$$\vec{w}.\vec{x} + b \leq +1 \tag{3}$$

The largest margin can be found by maximizing the value of the distance between the hyperplane and its closest point, which is $1/\|\vec{w}\|$. This can be formulated as a quadratic programming (QP) problem, which is looking for the minimum point of Eq. 4 by taking into account the constraints of Eq. 8

$$\min \tau\,(w) = \frac{1}{2}\|\vec{w}\|^2 \tag{4}$$

$$y_i(\overrightarrow{x_1.}\vec{w} + b) - 1 \geq 0, \forall_i \tag{5}$$

This problem can be solved with various computational techniques, including the Lagrange multiplier.

$$L(w, b, a) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i(y_i\big((x_i.\, w_i + b) - 1\big)\big) \tag{6}$$

a_i is Lagrange multipliers, which are zero or positive ($a_i \geq 0$) The optimal value of equation 6 can be calculated by minimizing L with respect to $\vec{w}$ and b and maximizing L against a_i Taking into account the nature that the optimal point of gradient is $L = 0$ equation 6 can modified as a maximization problem that only contains a_i as Eq. 7 as follows

$$L(w, b, a) = \frac{1}{2}\,\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i(y_i\big((x_i.\, w_i + b) - 1\big)\big) \tag{7}$$

With Constraint

$$a_i \geq 0(i = 1,2,3, \dots, l)\sum_{i-1}^{l} a_i y_i = 0 \tag{8}$$

From the results of this calculation a_i is obtained most of which are positive. Data correlated with positive a_i is called the support vector. The explanation above is based on the assumption that the two classes can be perfectly separated by a hyperplane. However, in general, the two classes in the input space cannot be separated completely. This causes the constraints in Eq. 8 cannot be fulfilled, so that the optimization cannot be done. To solve this problem, SVM was reformulated by introducing the soft margin technique. In soft margin, Eq. 5 provides by entering slack variable $\xi(\xi > 0)$ as follows:

$$y_i(x_i.\, w + b) \geq 1 - \xi_i \forall_i \tag{9}$$

Thus Eq. 4 is change to:

$$\min\tau(w, \xi) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l} \xi_i \tag{10}$$

The parameter C is selected to control the trade-off between margin and classification error ξ. A large C value means that it will provide a greater penalty for the classification error.

Genetic Algorithm

Genetic algorithm was first discovered by John Holland in 1975. The concept of a genetic algorithm is based on the theory of evolution with the principle of natural selection developed by Darwin [43]. The basic idea of genetic algorithm is the process of evolution. Genetic algorithms follow procedures that resemble the evolutionary process, namely the selection, crossover and mutation processes. Genetic algorithm is a technique to find the optimal solution of a problem that has many solutions. This technique will search from several solutions obtained to get the best solution according to predetermined criteria or what is known as the fitness function [44]. The process of the genetic algorithm begins with the process of forming the initial population. The initial population consists of a set of chromosomes formed using the Greedy algorithm. The number of chromosomes in the initial population is limited by the number of points visited. The next stage is analogous to the natural evolution process, namely selection, crossover and mutation.

Chromosomes from a population are taken and used to form a new population. The main goal of the genetic algorithm is to get a new population that is better than the previous population. In terms of optimization, Genetic algorithms are able to handle complex and irregular solution spaces and Genetic algorithms have been applied for

many difficult optimization problems. In addition, Genetic Algorithms can handle high dimensional, nonlinear optimization problems [13,30].

Proposed Method

The proposed algorithm of feature selection for SVM based on genetic algorithm is shown in FIGURE 3 for feature selection of SVM is used. Genetic algorithm has been developed for the process of finding a random solution by developing a population that is considered a set of solutions. The search process is carried out based on the theory of genetic evolution through selection, crossover and gene mutations to find the best individuals. In this research, the genetic algorithm uses a feature selection process to get features to remove irrelevant features so that it can improve the performance of the support vector machine algorithm. As shown in FIGURE 3, the input data set includes training dataset and testing dataset.

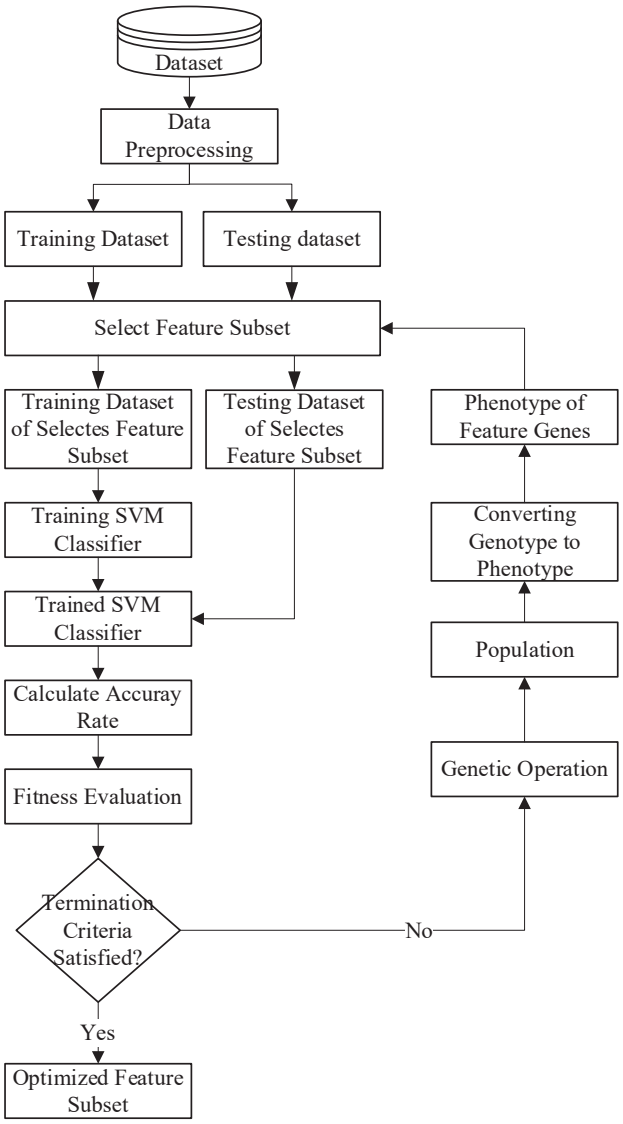

FIGURE 3. Activity diagram of GA-SVM method

The data pre-processing stage, we used to replace missing value by average. The steps taken in optimization using the genetic algorithm are as follows.

1. Converting genotype to phenotype for each feature value.

2. Feature subset, after genetic operation by changing each feature chromosome part from genotype to phenotype then feature results can be obtained.
3. Fitness function, each chromosome states the features selected from the dataset. Then calculated the performance of the classifier to evaluate each chromosome to evaluate the fitness function.
4. Determine the Number of Iterations. Iteration is done to evaluate one generation to produce a fitness value.
5. Genetic Algorithm Operation Process, 1-point crossover and mutation were performed in selected individuals. The best individual scores are obtained when the fitness is in an optimal position from the first to the final generation. Then the fitness is taken in a convergent (stable) condition and the best individual values in the form of chromosomes obtained are then decoded into real values. To keep the high fitness value individual from being lost during evolution, it is necessary to make one or more copies. This procedure is known as elitism.

We use 10-fold cross validation as a model validation. 10-fold cross validation is one of the K folds cross validations that is recommended for choosing the best model because it tends to provide less biased accuracy estimates compared to ordinary cross validation, leave-one-out cross validation and bootstrap. In 10-fold cross validation, the data is divided into 10 folds of roughly equal size, so that we have 10 data subsets to evaluate the performance of the model or algorithm. For each of the 10 data subsets, cross validation will use 9-fold for training and 1-fold for testing [45]. This 10-fold cross validation has become the standard of recent research, and several studies have also found that the use of stratification can improve a more diverse range of results [46].

Evaluation of the experimental result is a measuring tool that can be used to assess or measure how well the proposed algorithm against others algorithm and whether the proposed algorithm has a significant difference in the results of other models. In this study, we apply AUC (area under curve) as an accuracy indicator in our experiments that can be used as performance of classifiers. The accuracy value is calculated by taking the correct prediction percentage from the whole data. Sensitivity or Recall (RC) in the field of information search measures the proportion of original positives that are correctly predicted as positive. Recall is related to the ability of testing to identify positive results from a number of data that are actually positive. Precision (PR) or also called positive prediction value is a matrix to measure system performance in getting relevant data. The precision is amount of data that is true positive divided by the amount of data that is recognized as positive. Classification performance measurement is done by using a confusion matrix. Confusion matrix is a table for recording the results of classification work. Confusion matrix for two classes (binary) can be seen in the TABLE 2.

TABLE 2. Confusion matrix

Predicted	Actual	
	LQ45	**Non-LQ45**
LQ45	TP	FP
Non-LQ45	FN	TN

Recall (RC) : It is the ratio of true positive predictions to overall true positive data
RC = TP / (TP + FN)

Precision (PR) : It is the ratio of a positive true prediction compared to the overall positive predicted
PR = TP / (TP + FP)

Accuracy (ACC) : Is the ratio of the true prediction (positive and negative) to the overall data
ACC = (TP + TN) / (TP + TN + FP + FN)

AUC value provides an overview of the overall measurement of the suitability of the model used. The larger the area under the curve, the better the variables under study at predicting events [47]. In TABLE 3, be delineated the AUC value and its meaning.

TABLE 3. AUC value and meaning

AUC	Meaning	Symbol
0,90 – 1,00	Excellent classification	a.
0,80 – 0,90	Good classification	b.
0,70 – 0,80	Fair classification	c.
0,60 – 0,70	Poor classification	d.
< 0,60	Failure	e.

RESULTS AND DISCUSSION

The experiments were conducted using a computing platform based on Intel Core i5 1.6 GHz CPU, 8 GB RAM, and Microsoft Windows 10 Home 64-bit. The development environment is RapidMiner 9.5 library. We used the default parameter settings provide my RapidMiner 9.5 library. For support vector machine we tested the kernel type: dot, kernel cache: 200, C: 0, convergence epsilon: 0.001, max iteration: 100000, L Pos: 1.0, L Neg: 1.0 and epsilon: 0.0. Since the prediction results of these classifiers are sensitive to data split, we used cross-validation to judge the ability of generalization. For genetic algorithm we tested the population size: 500, maximum number generation: 30, p initialize: 0.5, p mutation: -1 and p crossover: 0.5. RapidMiner 9.5 will produce a classification performance model as a result of calculations such as the confusion matrix and AUC.

First of all, first, we conducted an experiment using the LQ45 dataset using the Support vector machine algorithm without genetic algorithm feature selection. The results of the SVM experiment without GA can be seen in the TABLE 4 and 5. From the confusion matrix as shown in TABLE 5, area under curve (AUC), accuracy (ACC), precision (PR) and recall (RC) of the model are 0.926, 89.29, 91.51 and 72.92 respectively. This prediction accuracy is good since it is attained 90% average and has excellent AUC since it above 90%. Area under the Curve (AUC) is a measure of the ability to classify to distinguish between classes. The higher the AUC, the better the performance model in predicting LQ45 and Non LQ45 issuers. When the AUC value = 1, the classifier can perfectly distinguish LQ45 and Non LQ45 issuers.

TABLE 4. Confusion matrix of SVM without GA

Predicted	Actual	
	LQ45	**Non-LQ45**
LQ45	174	22
Non-LQ45	6	59

TABLE 5. Summary of SVM evaluation result without GA

Dataset	Model Evaluation			
	AUC	**ACC**	**PR**	**RC**
LQ45	0.926[a]	89.29	91.51	72.92

[a.] excellent

In the second experiment, we implemented GA to choose the relevant feature. Then, the filtered data feed to SVM as main prediction algorithm. The experimental result is shown in TABLE 6 and 7. From the confusion matrix as shown in TABLE 7, area under curve (AUC), accuracy (ACC), precision (PR) and recall (RC) of the model are 0.928, 93.89, 94.82 and 85.28 respectively. This prediction accuracy is good since it's attained 90% average and has excellent AUC since it above 90%. We can see, the accuracy is an increase compared to the first experiment.

TABLE 6. Confusion matrix of GA-SVM algorithm

Predicted	Actual	
	LQ45	**Non-LQ45**
LQ45	176	12
Non-LQ45	4	69

TABLE 7. Summary of evaluation result GA-SVM

Dataset	Model Evaluation			
	AUC	**ACC**	**PR**	**RC**
LQ45	0.928[a.]	93.89	94.82	85.28

[a.] excellent

Finally, we compare the proposed algorithm with SVM standard algorithm that combined with feature selection algorithm such as Particle Swarm Optimization (PSO), Forward Selection (FS), Backward Elimination (BE). Table 8 shows the comparison between the proposed algorithm and other comparison algorithms with the same LQ45 dataset.

TABLE 8. Comparison results of all algorithm

Model	Model Evaluation			
	AUC	ACC	PR	RC
SVM	0.926[a.]	89.29	91.51	72.29
GA-SVM	0.928[a.]	**93.89**	**94.82**	**85.28**
PSO-SVM	0.934[a.]	90.44	93.89	75.28
FS-SVM	**0.938**[a.]	91.17	92.90	81.39
BE-SVM	0.926[a.]	89.27	92.17	71.53

[a.] excellent

The proposed algorithm shows yield high accuracy for LQ45 index prediction in Indonesia stock exchange and out of performs all comparison algorithm. As shown in TABLE 7, Results obtained using developed GA-SVM approach with and without feature selection were better than those of PSO-SVM, FS-SVM and BE-SVM. Implementation of feature selection with algorithms that can improve the accuracy of vector support algorithms. Based on the experimental results, it shows that a good classification performance can also be obtained from a model with few features. This explains that certain irrelevant and redundant features can affect the classification results. Therefore, the LQ45 index prediction can be an initial screening in investing to select stocks that have the potential to provide profit in the future. Investors can make it a reference for investing in shares in the Indonesian capital market.

CONCLUSIONS

This study presents a Genetic Algorithm-based feature selection approach, which is able to eliminate irrelevant and redundant features to obtain relevant feature subsets. This subset of relevant features is then applied in data training and data testing to obtain optimal results. Based on the experimental results, a combination of genetic algorithms for feature selection with supporting vector algorithms, feature selection that is able to give better results than without feature selection for future LQ45 index predictions. This is because selection removes irrelevant features. Thus, the supporting vector algorithm model based on Genetic Algorithm is the best algorithm model in this study and can provide the best results in testing and predicting the LQ45 index. The comparison of the results obtained with the feature selection approach shows that the GA-SVM approach developed has a better classification performance approaching other approaches for predicting the LQ45 index. We believe that our findings obtained from a real application would contribute to facilitate the investors in determining the stock option. The results of this study are not only beneficial to the literature but also have significant influence in the stock market prediction in terms of the ability to predict LQ45 index. Future research will be concerned with benchmarking the proposed algorithm with other metaheuristic optimization techniques such as bee colony or ant colony optimization, and other meta-learning techniques.

REFERENCES

1. Tealab, H. Hefny, and A. Badr, Int. J. Intell. Eng. Syst. **11**, 3, 49–58 (2018).
2. D. Devianto, Maiyastri, Randy, M. Hamidi, S. Maryati, and A. Wirahadi Ahmad, Proceedings of ICAITI 2018, 78-83 (2019).
3. I. M. Yassin, M. F. Abdul Khalid, S. H. Herman, I. Pasya, N. Ab Wahab, and Z. Awang, Int. J. Adv. Sci. Eng. Inf. Technol. **7**, 3, 1098–1103 (2017).
4. A. Syukur and D. Istiawan, Int. J. Intell. Eng. Syst. **14**, 1, 453–463 (2020).
5. D. P. Gandhmal and K. Kumar, Comput. Sci. Rev. **34**, 100190 (2019).
6. K. H. Sadia, A. Sharma, A. Paul, S. Padhi, and S. Sanyal, Int. J. Eng. Adv. Technol. **8**, 4, 25-31 (2019).
7. M. Dunne, *Stock Market Prediction Declaration of Originality* (University College Cork, Cork, 2015).
8. R. Toríbio, F. Nazário, J. Lima, V. Amorim, and H. Kimura, Q. Rev. Econ. Finance **66**, 115-126 (2017).
9. T. Anbalagan and S. U. Maheswari, Procedia Comput. Sci. **47**, 214–221 (2015).
10. P. Agarwal, S. Bajpai, A. Pathak, and R. Angira, Int. J. Res. Appl. Sci. Eng. Technol. **5**, 5, 1673-1676 (2017).
11. L. Wei, T. Chen, and T. Ho, Expert Syst. Appl. **38**, 11,13625-13631(2011).
12. R. Yamamoto, J Bank Financ. **36**,11 3033-3047 (2012).
13. E. Ahmadi, M. Jasemi, L. Monplaisir, M. A. Nabavi, A. Mahmoodi, and P. Amini Jam, Expert Syst. Appl. **94**, 21-31 (2018).

14. S. Shen, H. Jiang, and T. Zhang, *Stock Market Forecasting Using Machine Learning Algorithms,* (Department of Electrical Engineering, Stanford University, 2012).
15. D. Shah, H. Isah, and F. Zulkernine, Int. J. Financ. Stud. **7**, 2 (2019).
16. S. H. Kim and S. H. Chun, Int. J. Forecast. **14**, 3, 323–337 (1998).
17. W. Huang, Y. Nakamori, and S. Y. Wang, Comput. Oper. Res. **32,** 10, 2513–2522 (2005).
18. K. Kim, Neurocomputing **55**, 1–2, 307–319 (2003).
19. M. C. Lee, Expert Syst. Appl. **36**, 8, 10896–10904 (2009).
20. I. L. Mahargya and G. F. Shidik, Int. J. Adv. Sci. Eng. Inf. Technol. **10**, 6, 2261 (2020).
21. R. Capparuccia, R. De Leone, and E. Marchitto, Neural Networks **20**, 5, 590–597 (2007).
22. V. Vapnik, *The nature of statistical learning theory (Information Science and Statistics)* (Springer Verlag, New York, 1995).
23. T. N. Cristianini and J. Shawe, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, Cambridge, 2000).
24. J. H. Min and Y. C., Expert Syst. Appl. **28**, 4, 603–614 (2005).
25. S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, Expert Syst. Appl. **35**, 4, 1817-1824 (2008).
26. R. S. Wahono and N. S. Herman, Adv. Sci. Lett. **20**, 1, 239-244 (2014).
27. I. Guyon and A. Elisseeff, J. Mach. Learn. Res. **703**, 2, 152-162 (2003).
28. S. Saha, R. Spandana, A. Ekbal, and S. Bandyopadhyay, Appl. Soft. Comput. **29**, 479-486 (2015).
29. S. C. Yusta, Pattern Recognit. Lett. **30**, 5, 525-534 (2009).
30. M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, Expert Syst. Appl. **38,** 5, 5197-5204 (2011).
31. M. Dash, K. Choi, P. Scheuermann, and H. Liu, Proc. - IEEE Int. Conf. Data Min. ICDM., 115-122 (2002).
32. C. C. Aggarwal, *Data Mining* (Cham: Springer International Publishing, Switzerland, 2015).
33. C. H. Chen, Appl. Soft. Comput. **20**, 4-14 (2014).
34. R. C. de Amorim, J. Classif. **33**,2, 210-242 (2016).
35. G. Chandrashekar and F. Sahin, Comput. Electr. Eng. **40**, 1, 16-28 (2014).
36. B. Zheng, S. W. Yoon, and S. S. Lam, Expert Syst. Appl. **41**, 4, 1476-1482 (2014).
37. J. R. Cano, F. Herrera, and M. Lozano, IEEE Trans. Evol. Comput. **7**, 6, 561-575 (2003).
38. M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, IEEE Trans. Evol. Comput. **4**,2, 164-171 (2000).
39. J. Yang and V. Honavar, IEEE intell. syst. their appl. **13**, 2, 44-48 (1998).
40. G. Wang and J. Ma, Expert Syst. Appl. **39**, 5, 5325-5331 (2012).
41. X. Yu, S. Guo, J. Guo, and X. Huang, Expert Syst. Appl. **38**, 3, 425-1430 (2011).
42. J. Xu, Neurocomputing, **74**, 17, 3114-3124 (2011).
43. K. Manimala, K. Selvi, and R. Ahila, Appl. Soft Comput. **11**, 8, 5485-5497 (2011).
44. T. Nadu, J Comput Sci Technol. **11**, 2 (2011).
45. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. (Morgan Kaufmann Publishers, Waltham, 2012).
46. R. S. Wahono, N. S. Herman, and S. Ahmad, Adv. Sci. Lett. **20**, 10, 1945–1950 (2014).
47. F. Gorunescu, *Data Mining: Concepts, Models and Techniques* (Springer Berlin Heidelberg, India, 2011).