Research Article

Isaac Kofi Nti*, Adebayo Felix Adekoya, and Benjamin Asubam Weyori

Efficient Stock-Market Prediction Using Ensemble Support Vector Machine

https://doi.org/10.1515/comp-2020-0199
Received Aug 31, 2019; accepted Mar 01, 2020

Abstract: Predicting stock-price remains an important subject of discussion among financial analysts and researchers. However, the advancement in technologies such as artificial intelligence and machine learning techniques has paved the way for better and accurate prediction of stock-price in recent years. Of late, Support Vector Machines (SVM) have earned popularity among Machine Learning (ML) algorithms used for predicting stock price. However, a high percentage of studies in algorithmic investments based on SVM overlooked the overfitting nature of SVM when the input dataset is of high-noise and high-dimension. Therefore, this study proposes a novel homogeneous ensemble classifier called GASVM based on support vector machine enhanced with Genetic Algorithm (GA) for feature-selection and SVM kernel parameter optimisation for predicting the stock market. The GA was introduced in this study to achieve a simultaneous optimal of the diverse design factors of the SVM. Experiments carried out with over eleven (11) years' stock data from the Ghana Stock Exchange (GSE) yielded compelling results. The outcome shows that the proposed model (named GASVM) outperformed other classical ML algorithms (Decision Tree (DT), Random Forest (RF) and Neural Network (NN)) in predicting a 10-day-ahead stock price movement. The proposed (GASVM) showed a better prediction accuracy of 93.7% compared with 82.3% (RF), 75.3% (DT), and 80.1% (NN). It can, therefore, be deduced from the fallouts that the proposed (GASVM) technique puts-up a practical approach feature-selection and parameter optimisation of the different design features of the SVM and thus remove the need for the labour-intensive parameter optimisation.

Keywords: Stock Market, Ensemble Methods, Genetic Algorithm, Ghana-Stock-Exchange, Random Forest, Decision Trees, Support Vector Machine, Neural Networks, Stock Market Prediction, Trading Strategies

1 Introduction

The availability of big data generated from different sources (*e.g.*, government, financial, health, marketing, and social networks) daily, has paved the way for the application of intelligent systems for classification and pattern recognition tasks [1]. The stock market is one of the sectors that enjoy massive data daily from various sources. With many benefits associated with the stock market, it is sometimes used as a primary gauge of a nation's economic power and development [2, 3]. Hence, predicting the stock market is an essential area of interest in the financial world [4, 5].

Though market trend prediction is seen to be a complicated and challenging task, the profit associated with accurate prediction has resulted in an increase in research in this field [6–8]. Lately, researchers are working towards improving the accuracy of stock market predictive models by applying several business analytics techniques, including artificial intelligence, statistical means and soft computing [3]. Techniques such as artificial neural networks, random forest, support vector machine, decision trees, logistic regression and many more have been used in building algorithmic trading systems [3]. Among these techniques, the SVM has gained popularity due to its ability to deal with intricate nonlinear patterns [7, 9].

Usmani *et al.* [10, 11] proposed a hybrid machine learning algorithm based on SVM, Single Layer Perceptron (SLP), Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Autoregressive Integrated Moving Average (ARIMA) for predicting the Karachi Stock Exchange (KSE) index. Similarly, Chen and Hao [12] applied Feature

*Corresponding Author: Isaac Kofi Nti: Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana; Department of Computer Science, Sunyani Technical University, Sunyani, Ghana;
ORCID: 0000-0001-9257-4295
Adebayo Felix Adekoya: Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana; ORCID: 0000-0002-5029-2393
Benjamin Asubam Weyori: Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana; ORCID: 0000-0001-5422-4251

Table 1: A comparison of SVM in building trading systems

Reference	Method	Evaluation Metrics		Feature Selection	Value Predicted
[11]	Ensemble (SLP, MLP and SVM)	Accuracy	Training = 95.7% Testing = 66.6%	No	Stock Prices
[12]	FWSVM and FWKNN	MAPE = 0.27 RMSE = 0.0070		Yes	Market index
[7]	SVM	Accuracy = 89%		No	Stock-price
[13]	SVM, RF, K-Nearest Neighbour (KNN), Naive Bayes, and Softmax	Accuracy: SVM (66.14%), RF (69.73%) F-measure = SVM (0.7836), RF = (0.7987)		Yes	Stock price
[37]	SVM, MLP, RBF, ARIMA, SLP	Accuracy = 77%		No	Market index
[14]	SVM	Accuracy = 81.56% to 83.22% RMSE = 51.9525, MAPE = 0.22656		Yes	Stock price
[38]	SVM	Accuracy = 86.69% –89.33%		No	Stock price
[16]	SVM and LR	Accuracy: LR (82%), SMV (60%)		No	Stock price

SLP = Single Layer Perceptron. MLP = Multi-layer Perceptron, FWSVM = feature weighted support vector machine, FWKNN = feature weighted K-nearest neighbour algorithm, RNN = Recurrent Neural Network, LSTM = Long Short-Term Memory Cells, LR = Logistic Regression

Weighted SVM (FWSVM) and Feature Weighted K-Nearest Neighbour (FWKNN) algorithms for predicting the SSE Composite Index. Also, Mathur, Pathak and Bandil [8] presented a Recurrent Neural Network, Long Short Term Memory (RNN-LSTM) based stock-price predictive model used for portfolio management based on public sentiment. Likewise, Bousono-Calzon *et al.* [7] put forward an SVM predictive model for predicting stock price movement. All the above studies reported satisfactory prediction results by the SVM algorithm.

A comparative study of supervised ML algorithms for stock market trend prediction using time windows of size 1 to 90 was proposed [13]. The study reported that RF outperformed all other algorithms for large datasets, while the Naive Bayesian Classifier was the best for small datasets. An enhancement of SVM with swarm intelligence optimisation technique was proposed for predicting stock price [14]. Nayak, Mishra and Rath [15], Pimprikar, Ramachadran, and Senthilkumar [16], and Stanković, Marković, and Stojanović [17] applied SVM for predicting the stock market and reported moderate prediction accuracies. In another study, Ślepaczuk and Zenkova [9] proposed an investment strategy framework based on the SVM algorithm for the cryptocurrency market and investigating its profitability. Table 1 presents a comparative summary of previous algorithmic trading studies that used SVM as the primary classifier or among other classifiers.

Despite the SVM novelty and better performance reported in previous studies (see Table 1), some criticisms associated with its application for "big data" and "noisy data" analysis were overlooked. These issues are: (i) the classical SVM algorithm assumes that all the features of a sample give the same contribution to the target value. However, this theory is not always valid in many real problems [12]. (ii) SVM predictive models with flexible nonlinear kernels are prone to overfitting when input dataset is of high-noise and high-dimensional [18]. (iii) the practicality of SVM is impacted due to the problems of choosing suitable SVM parameters (C, σandε) as pointed out in the literature [14].

Moreover, Lin *et al.* [19], argued that the SVM could not precisely select a feature subset to contain features that are highly associated with the output, yet uncorrelated with each other [19]. Notwithstanding, according to Gonzalez *et al.* [20], these issues impede the generalisation performance of the SVM [20]. This limitation causes overfitting of the SVM in predicting financial data, due to the higher dimensional, precariousness, and noise associated with financial data.

However, as shown in Table 1, almost (62.5%) of the reviewed works based on SVM paid no attention to these limitations of the SVM. All the same, the widely use of SVM in algorithmic trading demands for an effective technique of building an SVM classifier with high classification capability. Therefore, to scale-down these issues mentioned above, feature optimisation techniques must be applied in predictive frameworks based on SVM where necessary.

In a way to reduce over-fitting and computational time due to high-noise and high-dimensional in financial data, and increase prediction accuracy, two main techniques are used in feature optimisation. Thus, (i) Feature-Extraction (FE) also called Feature-Transform (FT) and (ii) Feature-Selection (FS). FE seeks to construct a new feature variable from the original variables. On the other hand, FS focuses on picking a subset of features from the original variable

that are highly correlated with the expected output and discards the less significant variables.

Several techniques are available for performing dimensionality reduction in machine learning, and some of these techniques include Particle Swarm Optimization (PSO), Principal Component Analysis (PCA), and Genetic Algorithm (GA) [21, 22]. However, the literature shows that GA is an efficient and adaptive technique for FS [2, 23, 24]. Furthermore, the high sensitivity nature of Genetic Algorithm's to noise and their ability to operate without any domain knowledge puts them ahead for optimisation problems.

The introduction of Ensemble Methods (EM) was aimed at boosting the accuracy in predictive models by overcoming variance and stability issues with single classifiers. They are among the most potent predictive analytics algorithms which combine multiple learning algorithms, forming committees [25]. The power in EM has been applied in several areas to improve prediction accuracy. Example, in oil and gas industries to predict future oil price [26], in agriculture to predict crop yields [27], in health for classifying lung cancer severity [28] and in the energy sector for predicting short-term energy consumption [29]. Meanwhile, little of its application is known in the field of financial analysis, in Europe, India, Shanghai, and Bovespa Index [3].

Regardless of the global achievement made by machine learning and soft computing techniques in predicting the stock-market, studies show that over 95% of research works on the Ghana Stock Exchange (GSE) focused on identifying the degree of association between stock price and macroeconomic variable [3, 30].

Given the above discussions, this study seeks to employ the power in EM and the ability of SVM in dealing with intricate nonlinear patterns while overcoming its drawbacks discussed above with GA to predict the stock market.

Precisely, we propose a "homogeneous" ensemble classifier called (GASVM) based on an enhanced SVM with GA for feature selection and kernel parameter optimisation to predict stock market price movement. Our contributions to knowledge are (i) a unique ensemble SVM classifier enhanced with GA to carry out feature-selection and parameters-optimisation framework for predicting stock market price movement. (ii) A source of reference and a standard of comparison for future studies on predicting stock-price movement on the GSE.

The choice of the SVM classifier for this study was motivated by the effectiveness of the SVM algorithms in building algorithmic trading models [7, 9, 12, 15–17]. Furthermore, we used ensemble techniques to highlight the strength of different input features and SVM parameters while watering down their weakness.

The remaining sections of the current study are organised as follows: Section 3 presents details of materials and methods used for building our proposed GASVM predictive model and a brief description of other machine learning methods for algorithmic trading. Section 4 discusses the results of the current study, and Section 5 concludes this study and describes avenues for future research.

2 Material and methods

The details of the proposed predictive model (GASVM) and a brief description of other machine learning techniques used for benchmarking proposed design are presented in this section.

2.1 Ensemble GASVM

The fundamental goal of the current study was to enhance the prediction accuracy of a set of SVM using GA as a feature selection optimiser. In building ensemble classifiers, there is no fast rule for the number of base-classifier to apply; however, for this study, fifteen (15) different SVM classifiers were ensembled. This number was picked based on findings in [3]. Every based classifier was created with different parameters and attributes to increase diversity. The Gaussian Radial Basis Function was used as a kernel-based for our SVM for higher diversity achievement as reported in [20]. Figure 1 shows the proposed framework for predicting stock-market price movement.

The framework consists of two sections: phase 1 and phase 2. Phase 1 presents the data pre-processing stage. Our dataset was pre-processed in two primary steps: (i) data cleaning (ii) and (ii) data transformation. The data cleaning stage took care of missing values, noisy-data, and identifying and removing outliers where necessary and resolving of data inconsistency. The next step was data transformation; the max-min function as defined in Eq. (1), was adopted to normalise our dataset within the range [0, 1], to prevent local optima, overcome numerical complications and obtain a better efficacy.

$$u' = \frac{u - u_{\min}}{u_{\max} - u_{\min}} \tag{1}$$

where u' is the normalised value, u = the value to be normalised $u_{\min}$ and u_{max} = the minimal and maximum value of the dataset.

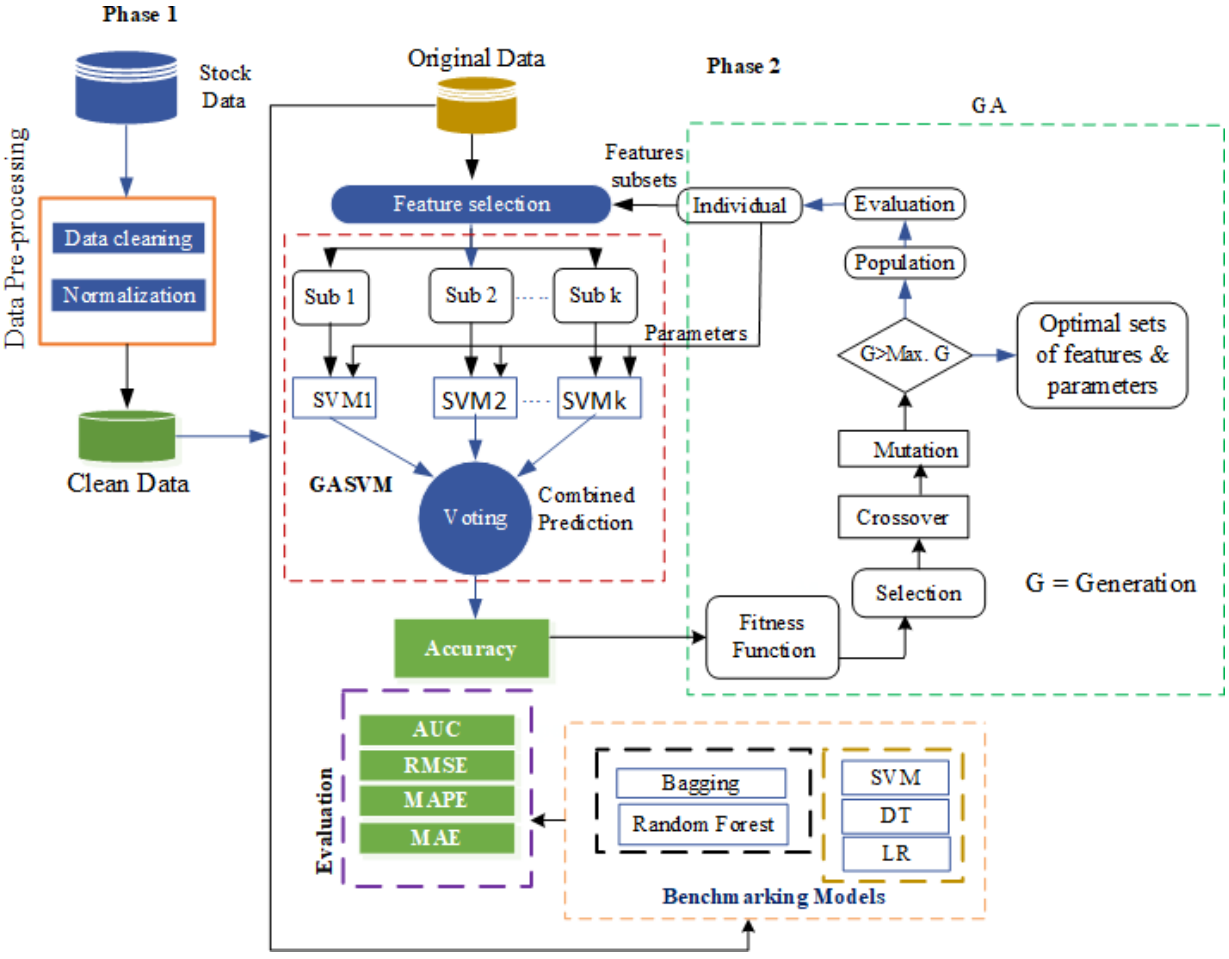

Figure 1: Proposed Framework

Phase 2 accounts for building the proposed GASVM classifier and other selected algorithms for benchmarking. The following section discusses it in detail.

2.1.1 Support Vector Machine (SVM)

The SVM is a supervised ML algorithm used for regression and classification tasks. It serves as a linear separator sandwiched between two data nodes to detect two different classes in the multidimensional environs [31].

Let (D_{set}), defined in Eq. (2), represent the training dataset, where (D_{set}) is a pair (x_i, y_i) of an n-dimensional feature vector, (y_i) = label of (x_i) *and* i *=1,2,3,…,n*. For every expected outputand inputs feature (x_i) in (D_{set}), we applied a hyperplane for separating the binary-decision classes in the 2-attributes, using the formula defined in Eq. (3). Where (w_i) characterise the weight-values of the hyperplane to be memorised by the algorithm. With this, we describe the maximum margin-hyperplane as defined in Eq. (4), where $x_{(i)}$ is the support-vector, (y_i) is the class-value of training examples $x_{(i)}$, x is the test instance. determine the hyperplane parameters. Eq. (5) represents a higher-dimensional version of Eq. (3), for nonlinearly separable instances. Finally, the SVM transforms the inputs-features (x_i), into a high dimensional feature space as expressed in Eq. (6) and Eq. (7).

$$D_{set} = \left\{ \left(x_i, y_i, \ldots, (x_n, y_n) \right) \right\} \in {}^{n} \times -1, 1 \tag{2}$$

$$y = w_o + w_1 x_1 + w_2 x_2 \tag{3}$$

$$y = b + \sum_{i=0} \left(\alpha_i y_i x(i) \right).x \tag{4}$$

$$y = b + \sum_{i=0} \alpha_i y_i K\left(x_i, x_j \right) \tag{5}$$

$$\min_{d,b\varepsilon,} \frac{1}{2} W^T W + C \sum_{i=1}^{n} \varepsilon i \tag{6}$$

$$subject\ to\ y_i\left(W^T\emptyset\left(x_i+b\right)\geq 1-\varepsilon_i\right) \quad (7)$$
$$\varepsilon_i > 0$$

When $\propto_i$ is within the lower boundary, almost or equal 0, then a detachable situation is obtained. For non-separable conditions, an upper-boundary C is placed on $\propto_i$ in addition to the lower boundary to generalise the SVM as pointed out in [19]. C is also (penalty-factor regulates) the trade-off between achieving a low error rate on the training data and complexity of the model. The value of C is of high importance in the model training because a small value of *(C)* causes an increase in training errors, due to the bigger margin of separating hyperplane it generates. While a substantial value of C leads to the lesser-margin hyperplane, leading to a more severe punishment of non-separable points.

The SVM was optimised with the algorithm as expressed in Eq. (6) and Eq. (7). With the function $\emptyset$, the vectors x_i (training dataset) are mapped into a dimension of higher space. In this dimension, SVM finds a linear separating hyperplane with the best margin. For every SVM, there are two values from Eq. (5) and Eq. (6) (*i.e.* $\propto$ and C) to be optimised by the genetic algorithm. The Radial Basis Function (RBF) expressed in Eq. (8) was adopted as the kernel function for our SVM ensemble, since RBF tends to give a decent performance under all-purpose smoothness guess.

$$RBF: K\left(x_i, x_j\right) = \exp\left(-y||x_i - x_j||^2\right), y > 0 \quad (8)$$

where $\left(x_i - x_j\right)$ is the Euclidian distance between two data point.

The majority voting ensemble technique defined in Eq. (9), is adopted for combining the outputs of individual SVM classifiers for the final prediction, due to its simplicity, yet compelling technologically.

$$P_k(x) = \begin{cases} 1 & if \\ 0 & otherwise \end{cases} \sum_{i=1}^{n} p_{i,k}(x) > n \quad (9)$$

where $P_k(x)$ is the prediction of ensemble on *(x)* for category k, $p_{i,k}(x)$ is the prediction of a specific SVM *(i)* on category *(k)* for review *(x)*, and n is the number of SVM ensembled.

2.1.2 Genetic Algorithm (GA)

GA is an evolutionary algorithm first introduced by John Holland [32]. GAs are adaptive-optimisation search techniques based on a direct similarity to Darwinian natural selection and genetics in biological systems. They can deal with bigger search spaces effectively and efficiently. GAs work with a set of an appropriate solution, called population *(p)*, in which every individual (chromosome) stands for a feasible solution to the given problem. Each *(pi)* value is accessed (where i = {1,2,3,4… p}) through a fitness function *(f)*, to determine how good a chromosome is in the possible solutions of the optimisation task. Based on the fitness value of (p_i), it is selected and passed through some genetic operations, namely mutation *(m)*, and crossover *(c)* with definite probabilities, forming new ones. It continues from the first (p) to the $\left(i^{th}\right)$ p in a loop; each loop referred to as a generation. As the cycle goes on, a new population evolves, and acceptable results are obtained, the cycle ceases when the stopping criteria (*sc*) are realised. The list of all chromosomes found within all generations becomes the GAs answer to the given problem. The following steps show the implementation of the GA feature selection process in this study.

Step 1: An initial population of chromosomes was generated, which were bit strings of arbitrarily created binary values. The size of the chromosome and population used in this study was 85 and 250, respectively. The chromosomes (C_h) value was calculated as expressed by Eq. (10). Where signifies the number of features in the dataset (D_{set}) and denotes the number of classifiers.

$$C_h = (F_e \times C_l) + 2C_l \quad (10)$$

Step 2: The C_h were decoded (bit strings) to discover which input variables must be chosen.

Step 3: We ran an ensemble SVM model to predict a 10-day stock price movement.

Step 4: The prediction accuracy by individual C_h from the ensembled SVM was used to determine how good a C_h is in the possible solutions of the optimisation task. The accuracy values were calculated based on Eq. (12).

Step 5: A decision to exit or continue the loop is taken at this stage. The stopping gauge was set to the number of generations, fifty (50) for this study.

Step 6: The tournament selection technique was used to select C_h to cross over. A tournament selection consists of running several tournaments on a small number of C_h chosen at random from the p. The victor for each tournament is selected for crossover.

Step 7: An arithmetic crossover operator which expresses a linear amalgamation of two (2) C_h was applied.

Step 8: We introduce new genes into the p with uniform m operator and generate an arbitrary slot number of the c as and flip the binary value in that slot.

Step 9: We replace old C_h with two best offspring of C_h for the next generation.

Step 10: Go to Step 2.

After manual adjustment of the GA coding, the main parameters were set, as shown in Table 2.

Table 2: Parameters used in the proposed GA feature Selection

Parameters	Value
Population Size	250
Number of Generations	50
Genome length	100
Probability of Crossover	85%
Probability of Mutation	10%
Type of Mutation	Uniform Mutation
Elite Count	2
Type of Selection	Tournament

2.2 Benchmarked Machine Learning Algorithms

This section presents a short description of selected algorithms for benchmarking proposed model. The three models were selected based on a survey report on the most common machine learning algorithm used in stock market prediction presented by Nti *et al.* [3].

Neural networks (NN):
The NN is among the soft computing techniques that are commonly found in algorithmic investments [3]. A feedforward single-layer artificial neural network of ten (10) hidden neuron and a "*relu*" activation function with a maximum iteration of four thousand (4000) was used in this study. The weights were chosen randomly from the start with an initial range parameter set to 0.1, while entropy constraint was set to use the topmost condition's likelihood. Weight decay was used to circumvent overfitting.

Decision Tree (DT):
A DT is a form of the flow-chart-like tree structure that uses a branch-off method to spell out every distinct likely result of a decision. DT represents a set of conditions, organised hierarchically and serially applied from root to a leaf of the tree or terminal node [33]. Every distinct node within the tree personifies a test on a precise variable, and each branch is the outcome of that test. An information-gain approach was used to decide the appropriate property for each node of a generated tree.

Random forest (RF):
RF combines the performance of (*N)* number of DT algorithms to make prediction [33]. When an RF receives an input of (x), where x is a vector made up of a variety of different evidential features examined for a given training area, the RF builds (N) DTs and averages their results as the final predicted output. So that for N tress , the FR predictor is formulated as expressed by Eq. (11). For this study, the value of N was fifty (50) based on findings in [3].

$$f_{rf}^{N}(x) = \frac{1}{N}\sum_{N=1}^{N} T(x) \tag{11}$$

2.3 Evaluation Metrics

In ascertaining the performance of the proposed model, four (4) evaluation metrics among the lot discussed in [3] were used, namely, (1) accuracy, (2) area under the receiver operating characteristic curve (AUC), (3) root mean squared error (RMSE) and (4) mean absolute error (MAE) expressed in Eq. (12)-(15), respectively.

$$Accuracy(\%) = \left(\frac{(TP + TN)}{(TP + FN + FP + TN)}\right) \times 100 \tag{12}$$

$$AUC = \int_0^1 \left(\left(\frac{TP}{TP + FN}\right) d\left(\frac{FP}{FP + TN}\right)\right) = \int_0^1 \left(\frac{TP}{P} d\frac{FP}{N}\right) \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}}\left(\sum_{i=1}^{n}(t_i - y_i)\right) \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{t_i - y_i}{t_i}\right| \tag{15}$$

where TN = True Negative, TP = True Positive, P = positive event, N = negative event, t_i = actual value and y_i = predicted value

2.4 Study Dataset

We downloaded the stock-dataset for this study from the GSE official website (https://gse.com.gh), from June 25^{th}, 2007 to August 27^{th}, 2019. The GSE is a developing market, where most companies listed were from 2010 upwards. However, to obtain bigger datasets in this study, two companies from the banking and petroleum sectors, among the few companies listed before 2005 that had fewer missing values in their dataset were selected for this study. Nine (9) appropriate technical indicators were calculated and added to the opening-price, closing-price, year-lowest-price, year-highest-price, the total stock-traded-volume, as initial feature sets. Since technical indicators are reported to be useful tools for characterising the real-market situation in financial time-series prediction and also, they conceal more information than actual stock prices [3, 9, 13, 14]. The indicators calculated include Simple-Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence/ Divergence Rules (MACD), Relative-Strength Index (RSI), On-Balance-Volume (OBV), stochastic %K (%K), stochastic %D (%D), accumulative ratio (AR) and volume ratio (VR). Thus, a total of fourteen (14) initially selected features (Table 3). A comprehensive approach for calculating these features is provided in Nti *et al.* [3].

Table 3: Initially selected features

S/N	Features
1.	Opening-Price
2.	Closing-Price
3.	Year-Lowest-Price
4.	Year-Highest-Price
5.	The Total Stock-Traded-Volume
6.	Simple-Moving Average (SMA)
7.	Exponential Moving Average (EMA)
8.	Moving Average Convergence/Divergence Rules (MACD)
9.	Relative-Strength Index (RSI)
10.	On-Balance-Volume (OBV)
11.	Stochastic %K (%K)
12.	Stochastic %D (%D)
13.	Accumulative Ratio (AR)
14.	Volume Ratio (VR

Let n be the number of proposed inputs (features) and (m) number of estimated rows, represented by the matrix A defined in Eq. (16), of size In this study n was 14.

$$\mathrm{A} = \begin{pmatrix} x_{11} & x_{12}, \ldots, & x_{1n} \\ x_{21} & x_{22}, \ldots, & x_{2n} \\ \vdots & \vdots & \vdots \\ x_{m1} & x_{m2}, \ldots, & x_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n} \tag{16}$$

The downloaded dataset had no class label; hence, a new attribute was introduced by researchers to represent the class label, as expressed by Eq. (17) [12].

$$\Delta C_{st}^{i} = S_{i,close} - S_{(i+n),close} \, (\forall_i = 1, 2, \ldots, n) \tag{17}$$

where ΔC_{st}^{i} represent the change of closing stock price between the n^{th} day and the $(x+n)^{th}$ day, n is the step size of time horizon. Two targets label such as down or up were formulated, thus, if $\Delta C_{st}^{i} \geq 0$ indicates an up else a down. If a label is an up, then were assigned the target variable $(Y_m) = \pm 1$, and if a label is a down, we assigned $Y_m = -1$.

3 Empirical Analysis and Discussion

We present the results and discussion of the proposed stock market predictive model in this section. All experiments were carried out on a Samsung laptop with Windows 10 OS, RAM- 6GB, Intel (R) Core™ i5-2410M CPU @ 2.30 GHz. We coded all predictive models in this study using the Scikit learn library [34] and Python.

3.1 Visualisation of the Datasets

Figure 2 and Figure 3 show the visualisation of the opening and closing price of the downloaded dataset from the GSE. It was observed that the change between the opening and closing stock price on the GSE is stable. Thus, there is not much difference between the opening price and the closing price of stocks. However, we observed that the stock price in the petroleum sector increased and dropped remarkably from June 2013 to August 2019. On the other hand, the banking sector stock saw a rise in price in 2018 and fell in 2019. We attributed the increase in the banking sectors stock price to the banking sector reformation in 2018, and in the second and third quarter of 2019 by the government of the day [35]. The reforms removed the banking sector from a significant state of distress [35].

Figure 2: Opening and closing price of the downloaded dataset (Banking)

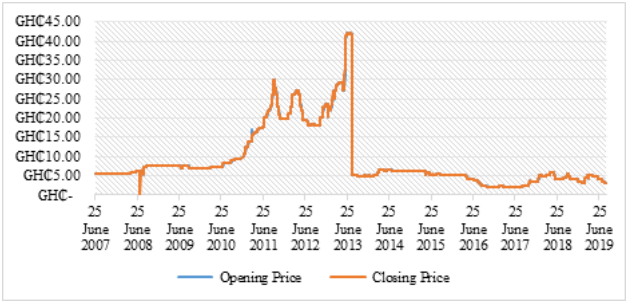

Figure 3: Opening and closing price of the downloaded dataset (Petroleum)

3.2 Model Performance

An efficient ensemble SVM with GA as a feature selection model (GASVM) was proposed for predicting stock price movement on the GSE. The empirical analysis compared the performance of the proposed GASVM model with RF, DT NN based predictive models for predicting a 10-day-ahead stock price movement. The dataset was partitioned into two sets, 80% for training and 20% for testing. 10-fold cross-validation was adopted and applied to attain an enhanced valuation of training accuracy. The results are presented and discussed in the following sections.

Predicting the stock market with high accuracy is an area of importance in business [3]. In this paper, we attempted to optimise the various parameters of the SVM using a genetic algorithm to increase its prediction accuracy and reduce the computational time. The cost associated with the wrong prediction in the financial sector is very high. Therefore, the significant contribution of the current study is the maximisation of investor profit while minimising loss, by proposing an enhancement technique for the traditional SVM and validate it experimentally through real-world stock dataset from the GSE. In this paper, the RBF kernel was adopted for the SVM; we aimed at optimising kernel parameter (C, σ) using GA.

Parameter C (penalty factor) is the cost of $C-SVM$ and the parameter σ is the gamma in kernel function. It is vital in any regularisation strategy that fair value is selected as a penalty factor. A higher value leads to a steep penalty for non-separable points, which might lead to storing several support vectors and overfitting, while too small might lead to underfitting. The obtained results show that with the optimal choice of (ε), the value of the regularisation parameter C had a small consequence on the generalisation performance. Table 4 shows the values of the SVM parameter optimisation ranges.

Table 4: SVM Kernel parameters

Parameter	Value
C	[0.01-1000]
σ	0.0001-10

The computational time of the proposed model (GASVM) against DT, NN, and RF are as shown in Table 5. It was observed that the DT model resulted in the fastest training time of (0.09 sec) followed by the NN (0.118 sec), the RF (1.808 sec). Even though the DT was the quickest in respect to training times, when it comes to testing time, DT and NN took almost the same time (0.006 sec). On the other hand, it was observed that the GASVM training time was high (18091.2 sec), which can be attributed to the several different combinations (2^n, where n represent the number of input features) in the feature selection processes.

Table 5: Training & Testing Time

Prediction Model	Training (sec)	Testing (sec)
GASVM	18091.2	12.81
ESVM	1.91	0.042
DT	0.09	0.006
NN	0.118	0.0064
RF	1.808	0.042

Consequently, the computation time (18091.2 sec) and resource required by the GASVM technique was observed to be higher compared with conventional RF, DT, NN, and ESVM. Thus, finding the optimal solution by the genetic algorithm is an iterative process, which takes time. Nevertheless, compensation was achieved through the increase in the accuracy score (93.7%) compared with literature (Table 1) of accuracy measure between 66% and 83% [13, 14, 17].

In a way to examine the effect of the GA feature selection technique, we ensemble, fifteen (15) SVM (ESVM) based on majority voting without any feature selection and applied the 14 initially selected features for prediction. Table 6 shows comparable results of RMSE, MAE and AUC between GASVM, DT, NN, RF and ESVM on GSE in a 10-day-ahead prediction. From the outcome (Table 6), we observed that the GASVM performed well in terms of RMSE and MAE compared with benchmarked models. Again, from Table 6, the GASVM recorded the highest AUC (closer to 1) compared with its counterparts' models. This outcome shows that the GASVM would rank an arbitrarily chosen positive observation in our dataset higher than an arbitrarily chosen negative representation in the dataset.

Table 6: Statistical Analysis of models

	DT	NN	RF	ESVM	GASVM
Accuracy	0.753	0.801	0.923	0.908	0.937
AUC	0.764	0.865	0.905	0.806	0.973
Precision	0.721	0.799	0.913	0.817	0.919
RMSE	0.403	0.255	0.117	0.131	0.042
MAE	0.234	0.165	0.014	0.110	0.09
SD	0.0699	0.0572	0.0678	0.0527	0.0484

The results show that the GAs optimisation of the SVM parameters has contributed to an enhancement in prediction accuracy. Thus, their offers a promising prospect of a hybridised feature selection and ensemble learning in the financial market. This enhancement of efficiency after feature selection shows that some features indeed cause noise, which intend reduces the performance of the machine learning model. The results further show that to maximise the performance of SVM, proper setting of design factors such as the right feature subset, the appropriate kernel, and its parameters are of crucial importance.

Furthermore, it was observed that all the ensemble techniques (RF, ESVM, and GASVM) outperformed the single classifiers (DT and NN). Thus, an affirmation of the argument that ensemble techniques are superior to individual classifiers, as reported in [25, 36]. Concerning the values obtained for RMSE and MAE by this study, the best model was GASVM, followed by RF, ESVM, and NN. The proposed model (GASVM) achieved a smaller difference between MAE and RMSE. As stated in [2, 23], the minor difference stuck between MAE and RMSE implies a lower variance in the individual errors. The Standard Deviation (SD) of 0.0484 by the proposed model shows that the GASVM is more stable compared with the NN, DT, RF and ESVM.

The proposed model (GASVM) recorded a training accuracy of 94.6% compared to 93.7% for testing. Thus, the introduction of GA as a feature selection and parameter optimisation technique offered SVM protection against overfitting, which was identified in the literature as a drawback of the SVM for a high-dimension dataset. Figure 4 shows the True Positive Rate (TPR) and False Positive Rate (FPR) of the models. The results show that the GASVM, RF, and ESVM predicted true positives as compared to the DT and NN. The values show excellent performance of the proposed GASVM predictive model for stock price movement prediction.

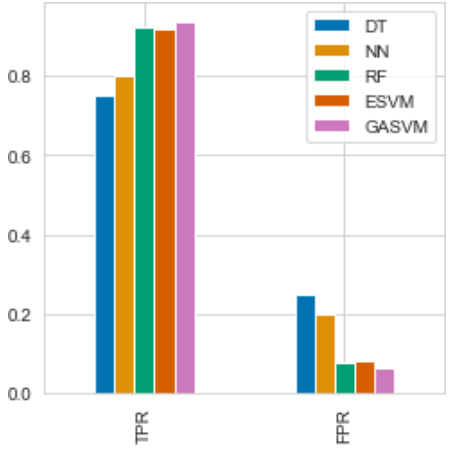

Figure 4: True-positive rate (TPR) and false-positive rate (FPR) of models

4 Conclusion

In this study, we introduce a novel "homogeneous" ensemble classifier (GASVM) based on Genetic Algorithm (GA) for feature-selection and optimisation of SVM parameters for predicting 10-day-ahead price movement on the Ghana stock exchange (GSE). Accuracy metrics such as RMSE, MAE, AUC, Accuracy, Recall were compared, between proposed model (GASVM) and other state-of-the-art predictive models (DT, RF and NN). The GASVM showed a higher prediction accuracy of the GSE stock-price movement as compared with DT, RF and NN. The primary input of this study is the introduction of a GA as a feature selection mechanism to optimise the various design factors of the SVM simultaneously. This yielded evidence in the results obtained from the proposed model compared with the conventional SVM ensemble, random forest, decision trees, and neural network.

Finally, an enormous search space was achieved by performing feature selection and parameter optimisation simultaneously in the current study through the number of features combination and the moderate range of parameter values. We hope that the accuracy of 93.7% achieved by the proposed model would help maximise the profit of investors on the Ghana stock exchange. We, therefore, deduced from the results that the proposed (GASVM) technique puts forward a practical approach feature selection and parameter optimised of the distinct design features of the SVM and thus removes the need for the manual optimisation of the SVM.

The current study adopted only a genetic algorithm for features and parameter optimisation based on findings in the previous studies without experimenting with other available optimisation techniques. Hence, further investigation should look at employing other alternatives. Again, stock price movement depends not only on historical stock data but also on fundamental data such as the satisfaction of the customers with the company's market and web news [3, 8]. So, future research work could also investigate the effects of user sentiment and web financial news to predict stock price movement.

Funding: Authors did not receive any funding to support this study.

Conflict of Interests: The authors declare that they have no conflict of interest.

References

[1] Oussous, A. et al., Big Data Technologies: A Survey, Journal of King Saud University - Computer and Information Sciences, 2017

[2] Dosdoğru, A. T. et al., Assessment of Hybrid Artificial Neural Networks and Metaheuristics for Stock Market Forecasting, Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi, 2018, (24) 63–78

[3] Nti, I. K., Adekoya, A. F., Weyori, B. A., A systematic review of fundamental and technical analysis of stock market predictions, Artificial Intelligence Review, 2019

[4] Zhou, X. et al., Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets, Mathematical Problems in Engineering, 2018

[5] Thanh, D. V., Minh H. N., Hieu, D. D., Building unconditional forecast model of Stock Market Indexes using combined leading indicators and principal components: application to Vietnamese Stock Market, Indian Journal of Science and Technology, 2018, (11)

[6] Lin Z., Modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models, Future Generation Computer Systems, 2018, (79) 960–972

[7] Bousono-Calzon, C. et al., On the Economic Significance of Stock Market Prediction and the No Free Lunch Theorem, IEEE Access, 2019, (7) 75177–75188

[8] Pawar, K., Jalem, R. S., Tiwari, V., Stock Market Price Prediction Using LSTM RNN, Rathore V., Worring M., Mishra D., Joshi A., Maheshwari S. (eds) Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing, 2019, (841), 493-503, Springer, Singapore

[9] Ślepaczuk, R, Zenkova M., Robustness of Support Vector Machines in Algorithmic Trading on Cryptocurrency Market, Central European Economic Journal, 2019, (5), 186–205

[10] Usmani, M. et al., Predicting Market Performance with Hybrid Model, 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), IEEE, 2018

[11] Usmani, M. et al., Stock market prediction using machine learning techniques, 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), IEEE, 2016, 322–327

[12] Chen Y., Hao Y., A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction, Expert Systems with Applications, 2017, (80) 340–355

[13] Kumar I. et al., A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction, 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, 2018, 1003–1007.

[14] Devi, K. N., Bhaskaran, V. M., Kumar, G. P., Cuckoo optimized SVM for stock market prediction, 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE, 2015

[15] Nayak, R. K., Mishra, D., Rath, A. K., A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices, Applied Soft Computing Journal, 2015 (35), 670–680

[16] Pimprikar, R., Ramachadran, S., Senthilkumar, K., Use of machine learning algorithms and twitter sentiment analysis for stock market prediction, International Journal of Pure and Applied Mathematics, 2017, (115), 521–526

[17] Stanković, J., Marković, I., Stojanović, M., Investment Strategy Optimization Using Technical Analysis and Predictive Modeling in Emerging Markets, Procedia Economics and Finance, 2015, (19), 51–62

[18] Kim, K. J., Lee, K., Ahn, H., Predicting corporate financial sustainability using Novel Business Analytics, Sustainability (Switzerland), 2018, (11)

[19] Lin, Y., Guo, H., Hu, J., An SVM-based approach for stock market trend prediction, Proceedings of the International Joint Conference on Neural Networks, 2013

[20] Gonzalez, T. R., Padilha, A. C., Couto, A. D., Ensemble system based on genetic algorithm for stock market forecasting, 2015 IEEE Congress on Evolutionary Computation (CEC), 2015, 3102–3108

[21] Gurav, U., Sidnal, N., Predict Stock Market Behavior: Role of Machine Learning Algorithms, Advances in Intelligent Systems and Computing, 2018, (673), 383–394.

[22] Zhang, X. et al., A causal feature selection algorithm for stock prediction modeling, Neurocomputing, 2014, (142), 48–59

[23] Göçken, M. et al., Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction, Expert Systems with Applications, 2016, (44), 320–331

[24] Inthachot, M., Boonjing, V. and Intakosum, S., Artificial Neural Network and Genetic Algorithm Hybrid Intelligence for Predicting Thai Stock Price Index Trend, Computational Intelligence and

Neuroscience, 2016
[25] Ballings, M. et al., Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications, 2015, (42) 7046–7056
[26] Zhao, Y., Li, J., Yu, L, A deep learning ensemble approach for crude oil price forecasting, Energy Economics, 2017, (66), 9–16
[27] Priya, P., Muthaiah, U., Balamurugan, M., Predicting Yield of the Crop Using Machine Learning Algorithm, International Journal of Engineering Sciences & Research Technology, 2018, (7)
[28] Bergquist, S. L. et al., Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data, The 2nd Machine Learning for Healthcare Conference, 2017, 25–38
[29] Khairalla, M. A. et al., Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model, Energies, 2018, (11)
[30] Nti, I. K., Adekoya, A. F., Weyori, B. A., Random Forest Based Feature Selection of Macroeconomic Variables for Stock Market Prediction, American Journal of Applied Sciences, 2019, (16), 200–212
[31] Vaghela, C., Bhatt, N., Patel, P. U, A Survey on Various Classification Techniques for Clinical Decision Support System, International Journal of Computer Applications, 2015, (116), 975–8887
[32] Nuwan, I. S, Genetic Algorithms: The Crossover-Mutation Debate, University of Colombo, 2005
[33] Rodriguez-Galiano, V. et al., Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, Ore Geology Reviews, 2015, (71), 804–818
[34] Pedregosa, F. et al., Scikit-learn, Journal of Machine Learning Research, 2011, (12), 2825–2830
[35] Bank of Ghana, Update on Banking Sector Reforms, 2019, 3–19. https://www.bog.gov.gh/privatecontent/Public_Notices/UPDATE ON BANKING SECTOR REFORMS.pdf.
[36] Creamer G., Freund Y., Predicting performance and quantifying corporate governance risk for Latin American ADRs and banks, Financial Engineering and Applications. Cambridge: MIT, 2004
[37] Raza K., Prediction of Stock Market performance by using machine learning techniques, 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), IEEE, 2017
[38] Patel, J. et al., Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques, Expert Systems with Applications, 2015, (42), 259–268