International Journal of Engineering & Technology, 7 (2.33) (2018) 695-702

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Application of machine learning in stock trading: a review

Kok Sheng Tan [1] *, Rajasvaran Logeswaran [2]

[1] Faculty of Computing, Engineering & Technology, Asia Pacific University of Technology and Innovation
57000 Kuala Lumpur, Malaysia
**Corresponding author E-mail: sam-tanks@hotmail.com*

Abstract

The wide adoption of machine learning techniques in predicting stock prices has led to the emergence of many articles on the topic. Howev-er, a systematic review on the topic remains lacking. This paper provides a systematic review of the recent applications of machine learning techniques in the construction of stock prediction models. A framework is designed to classify and evaluate the relevant work in recent arti-cles based on the type of model, type of financial market, type of prediction technique, type of optimization approach, type of indicators, type of performance metrics, type of benchmark models and prediction results. It is observed that financial indicators are the frequently used input variables and different forms of machine learning techniques are integrated to predict the stock prices. There are 4 variables that im-pose significant influence on the prediction model, namely the type of input variables, type of prediction technique, type of optimization approach and number of analysis layer. Thus, the limitations and potential enhancement on the 4 variables are discussed so that optimal combinations will be established in future research efforts.

Keywords: *Fundamental Analysis; Machine Learning; Stock Prediction; Technical Analysis*

1. Introduction

Stock investment is currently a widely studied research field as the digitization of stock trading has generated huge volume of real-time data [26]. In 2016, there are 206.1 trillion shares traded through electronic order book and the total value of the stocks traded globally has reached a staggering figure of USD 86.5 trillion [45]. The availability of large amounts of stock market data encourages researchers and financial experts to draw insights from these to produce lucrative financial returns and hedge against market risk. However, investors face difficulties in discovering new constructive insights from the data in the dynamic market environment. Ravi, Pradeepkumar and Deb [35] proposed that economic, social, industrial and geo-political factors will cause the financial time series data to be uncertain, noisy and incomplete, but asserted that the practical predictive quality of financial time series will generate huge financial gains. The efficient market hypothesis (EMH), a theory in financial economics proposed by [24] states that it is impossible to constantly outperform the overall market as stock prices will reflect all the relevant information in an efficient market. However, Malkiel [25] implied that the fluctuations in the strength of market efficiency would stimulate the short-term existence of predictable patterns in the stock market, which will lead to the partial predictability of the future direction of stock prices. Patel et al. [31] suggested that the perfect information assumption of the strong form EMH unveils the possibility to predict the stock prices by using the data generated from market activities. The authors proposed that the prediction of stock trends or stock index is possible with the combination of efficient pre-processing of information acquired from stock prices and application of appropriate algorithm, which has led to the emergence of many prediction algorithms to predict the behavior of stock prices. The four common methodologies for predicting the direction of stock prices are technical analysis, time series forecasting, data mining and machine learning. Technical analysis involves heavy utilization of charts to determine the trend of the stock [29]. Brockwell and Davis [8] stated that time series forecasting involves fitting models on historical data and predict the future observations with the models. Traditional statistical models such as moving average model (MA), autoregressive model (AR), autoregressive moving average model (ARMA) and autoregressive integrated moving average (ARIMA) are the conventional approaches for time series modelling. The strong potential of the combination of technical analysis is demonstrated in [2] using historical stock data and ARIMA model in predicting short-term market trends. Data mining is the process of discovering useful, previously unknown interesting patterns in large datasets. The four main build blocks of data mining are association pattern mining, data clustering, data classification and outlier detection [1]. According to Awad and Khanna [3], the potential of data mining to decipher complex problems has led to the extensive application of data mining in multiple fields such as medicine, biological science, engineering, social media and business intelligence. Data mining employs core machine learning algorithms for classification, clustering, and dimensionality reduction. Chen and Hao [11] proposed that many different combinations of machine learning methods have been developed to predict the stock market due to the ability of machine learning to handle the random, chaotic and non-linear data of the stock market. Single classifiers models such as support vector machine (SVM), artificial neural networks (ANN), K-nearest neighbor (KNN) and logistic regression are frequently used in predicting stock prices [6]. Ensemble methods that combined multiple classifiers such as random forest, adaptive boosting and kernel factory may be applied to predict the stock market as they are proven to be the top performer in other domains such as credit scoring, customer churn behavior and social media analytics [4], [5], [30]. There are a limited number of comprehensive literature reviews conducted on the subject of stock market prediction over the past few years. Kamley, Jaloree and Thakur

Copyright © 2018 Kok Sheng Tan, Rajasvaran Logeswaran. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[21] reviewed the application of machine learning techniques in forecasting the performance of the share market, in which they discussed the significance of fundamental and technical indicators in predicting the stock market performance. However, they stated that extensive research on all machine learning algorithm is impractical, thus, they studied a few popular algorithms such as SVM, genetic algorithm (GA), ANN, decision trees and Bayesian network. The publication period of the presented techniques ranged from 2000 to 2015, but they did not describe the potential shift in the implementation of the techniques such as modifications and optimization approaches over the period of the 15 years. Patel et al. [32] presented a comparative study on different machine learning techniques such as random forest, ANN and naïve Bayes, in predicting 2 stock price indices (CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex) and 2 stocks (Infosys Ltd. and Reliance Industries) from the Indian stock markets. Nayak, Pai and Pai [27] analyzed the prediction accuracy of three models, specifically boosted decision tree, logistic regression and SVM in predicting the movement of the Indian stock market. The existing academic studies on machine learning techniques in predicting stock prices aim to answer various research questions but they are incompetent to answer the scarcity of a systematic up-to-date review on the subject. Therefore, a review is undertaken in this article to organize analyses and summarize the state-of-the-art machine learning techniques in predicting the stock prices. The objectives of this review paper are: (1) to create a framework for effective comparison of the wide applications of machine learning techniques for stock market prediction, (2) to propose a systematic review of the existing literature, and (3) to investigate knowledge gaps and propose potential directions of future research. This paper aims to contribute to the existing state of knowledge by identifying, classifying and analyzing the machine learning techniques and input variables employed to predict the stock market. This paper also investigates the major limitations in the existing literature are investigated and a guideline for future research is proposed. The paper is structured as follows: Section 2 presents the research methodology to identify the highly relevant existing researches. Section 3 introduces the framework to classify the machine learning techniques employed to predict the stock market. Section 4 analyses and evaluates the machine learning techniques. Section 5 discusses the limitations in the existing models and provides future scope of study, while section 6 presents the conclusion.

2. Research methodology

There are four stages to systematically review the existing literature, namely research definition, article searching, article evaluation and research analysis. The structure of the systematic review is constructed by assimilating the essence of the systematic review methods proposed by [18], [22] [41].

2.1. Research definition

The application of machine learning techniques in predicting the stock prices is the research interest of the paper. The objective of this review paper is to present a framework to classify and compare the machine learning techniques and a system review of the techniques. The systematic review enables the determination of knowledge gaps and proposal of research scope in the future. The scope of the research focuses on academic articles on the application of machine learning techniques for stock market prediction published with their full text publications from 2012.

2.2. Article search

Broad search terms such as machine learning, technical analysis and fundamental analysis are used to compile a list of articles to determine the highly relevant keywords to increase the rate of discovering relevant articles. The adjustment of search terms subject to the information in the list of primary articles. As a result, the search terms were finalized to K-nearest neighbor, support vector machine, artificial neural networks, logistic regression and random forest. Online databases such as ACM Digital Library, IEEE Computer Society Digital Library, Science Direct, Emerald Insight, ProQuest, Springer Link and Google Scholar are utilized as the source for articles. More than 120 articles were captured from the online databases. The reference information in the articles that fell within the defined research scope was used to expand the coverage of articles search to result in a total of 150 articles are obtained in this step.

2.3. Article evaluation

The articles identified were evaluated based on a set of criteria to determine the highly relevant articles. The defined criteria for selecting articles are the application of machine learning techniques in building stock prediction models and the availability of performance metrics to assess the proposed prediction models. Articles without implementation of machine learning techniques in predicting the stock prices were categorized as unrelated articles and eliminated from the list. The significance of the implementation of machine learning techniques and the quality of performance metrics were used to evaluate the remaining articles. The selected articles were then classified accordingly.

2.4. Research analysis

The selected articles are analyzed from different perspectives to understand the recent development of machine learning techniques in stock market prediction. Research gaps were established to identify the potential directions for future research.

3. Framework classification

The existing research works were classified based on the financial analysis techniques in which a table will be constructed to organize the analysis of the researches. A glossary of the abbreviations used in this paper is provided in the Appendix.

3.1. Financial analysis technique

Suciu [38], Hong, and Wu [17] proposed that there are [2] main stock analysis approaches. These are further described below.
Fundamental Analysis: This form of financial analysis investigates the health and performance of a company by evaluating economic indicators. Fundamental analysis assumes that stock markets are not absolutely efficient, thus the company will not be priced accurately all the time [38]. Graham [14] originally proposed that fundamental analysis will produce a quantitative value to compare with the current price of a security and determine whether the stock is undervalued or overvalued, which will be the basis for trading decisions. In the paper, the three essential elements of fundamental analysis identified were economy analysis, industry analysis and company analysis. Economy analysis studies the effect of macroeconomic factors such as exchange rates, consumer price index, gross domestic product, and money supply. Industry analysis estimates the value of a company by comparing the revenue of the company to the prominent entities in the industry. Company analysis focuses on the value of a company by analysing the financial reports of the company with financial ratios such as working earnings per share, price-earnings ratio and debt-equity ratio.
Technical Analysis: The foundation of technical analysis is constructed on three assumptions which are the market discounts everything, prices move in trends, and history tends to repeat itself. Technical analysis believes that the factors such as the company's fundamentals, broad market factors and market psychology are incorporated into the price of the stock. Technical analysis assumes that stock price is expected to continue a past trend instead

of moving unpredictably. The price movement is repetitive due to predictable market psychology such as fears or excitement. This form of analysis forecasts the direction of prices by examining the data generated from market activities, such as price and volume [47]. In contrast to the approach of fundamental analysis to measure the intrinsic value of a security, Teixeira and Oliveira [39] stated that technical analysis is dependent on the historical stock prices to predict the behaviour of the stocks. Hu et. al [18] identified that technical analysis can be characterized into 8 classes which are sentiment, flow-of-funds, raw data, trend, momentum, volume, cycle and volatility. Sentiment represents the market behaviour, flow-of-funds indicator investigates the strength of buying and selling, raw data explain the price patterns, trend is an indicator that identifies the stock price trends, momentum is an indicator to identify potential trend reversal by examining the velocity of price change, volume-based indicators determines the investing enthusiasm, cycle theories propose the periodic pattern of price movement, and volatility investigates the variation in stock prices.

Classification of Reviewed Stock Prediction Model

The reviewed stock prediction models will be analyzed in which a table will be constructed to organize the results. The procedures are further described as follow.

Overview of the Classification Framework: The framework proposed to classify the 25 selected articles consists of 6 elements, namely model, experiment object, prediction technique, optimization approach, indicator and input variables. Model refers to the name of the proposed method, experiment object refers to the stocks or market indices for testing the model, prediction technique refers to the type of machine learning technique employed for prediction, optimization approach refers to the method to improve the performance of the model, indicator refers to the type of input variables and input variables refer to the variables that will be applied in the model. Table 1 summarizes the classification of the reviewed articles, which comprises 5 variables namely, the analysis model, the experiment object to test the model, the machine learning techniques employed to build the model, the approach to optimize the model and type of financial indicator used in the model.

Description of Reviewed Stock Prediction Models: There were 30 stock prediction models constructed from 9 distinct machine learning techniques identified in Table 1 which are SVM, ANN, naïve Bayes (NB), adaptive boosting (AB), boosted tree decision (BTD), kernel factory (KF), K-nearest neighbour (KNN), logistic regression (LR), random forest (RF). The three most frequently used machine learning techniques are SVM, ANN and RF. 17 out of the 25 articles employ SVM, 11 employ ANN and [5] employ RF to predict the price movement of stock market. The application of SVM reduces the likelihood of overfitting and globalize the optimal solution [11]. Dunis et al. [13] explained that the ability of SVM to overcome some limitations of ANN such as the difficulty to interpret the analysis results has led to the preference of SVM over ANN. Chai et al. [9] stated that SVM outperforms ANN in generalization with structural risk minimization principle and in avoiding the local minima in training. [22] out of the 25 articles reviewed exploit optimization approaches such applying machine learning techniques for feature selection and parameter optimization to increase the performance of the prediction model. Yu, Chen and Zhang [48] employ Principle Component Analysis (PCA) to extract efficient features to improve the accuracy of the SVM model. Inthachot, Boonjing and Intakosum [20] utilize GA to optimize the parameters of ANN. Only 1 of the 30 models was a multi-stage stock prediction model as proposed by [31]. Of the 25 articles, 17 incorporated technical indicators, 4 incorporated fundamental indicators and 2 incorporated technical and fundamental indicators in the prediction model.

Table.1: Classification of Proposed Stock Prediction Model

Model	Market	Financial Indicator	Prediction Technique	Optimization Approach
BNNMAS [15]	DAX Index	F, T	ANN	BA
FWSVM-FWKNN [11]	SHCOMP, SZCOMP	T	SVM, KNN	FW, IG
GA-SVM [13]	FTSE 100, ASE 20	O	SVM	GA
EMD-LSSVM [13]	CSI 300 Index	F	SVM	EMD, SGS, PSO, GA
CEFLANN [12]	BSE SENSEX, S&P 500	T	ANN	EML
NSVM-KNN [28]	BSE Sensex, CNX Nifty	T	KNN	Nil
PCA-SVM [48]	SHASHR	F	SVM	PCA, RBKF, GS
TBSM-SVR [46]	7 stocks (US)	T	SVM	SRA, RBKF
MLP-ANN [37]	Dow 30	T	ANN	MP
GA-ANN [20]	Thailand's SET 50	T	ANN	GA
BPNN [34]	Nikkei 225	F	ANN	BP, GA, SA
SVR, RF, ANFIS [10]	Indian Stock Market	T	SVM, RF, ANN	Nil
GA-ANN [33]	Nikkei 225	T	ANN	GA
LSSVM [16]	S&P 500	T	SVM	FPA, BA, MCS, ABC, PSO
ANN, LR, SVM, KNN, RF, AB, KF [6]	5767 public companies (Europe)	F	ANN, LR, SVM, KNN, RF, AB, KF	Nil
LSSVM [42]	CSI 300	T	SVM	GRBF
SVM [40]	CNX NIFTY	T	SVM	Lin, Poly, HTK, RBF
ANN-RS [7]	Dhaka market	T	ANN	BP
PCA-SVM [43]	KOSPI, HSI	O	SVM	PCA
1st Layer SVR 2nd Layer SVR-ANN, SVR-RF, SVR-SVR [31]	CNX Nifty S&P BSE Sensex	T	SVM, ANN, RF	2 stage fusion approach
ANN, SVM, RF, NB [32]	Reliance, Infosys, CNX Nifty, BSE Sensex	T	ANN, SVM, RF, NB	TDDPP
BDT, LR, SVM [27]	ISM	O	BDT, LR, SVM	Nil
ERF [23]	KOSPI	T, F	RF	Weightage modification
SVM [36]	S&P 500	T	SVM	HTRBF
VWSVM [49]	20 random stocks	T	SVM	Fisher score

4. Model evaluation

A model was established to evaluate the existing researches in which a table is constructed to organize the analysis results. This section describes the evaluation framework and results of the reviewed prediction models.

Overview of Evaluation Framework: There are [2] types of articles selected in this paper, which will be labelled as Class 1 (17 articles) and Class 2 (9 articles). Class 1 consists of articles that proposed a prediction model and measure the performance against

other proposed benchmark models. Class 2 consists of articles that compare the performance of different models in predicting the stock prices. The performance of the proposed prediction model in the article will be measured with predefined performance metrics. There are 2 types of performance metrics used in all the reviewed articles to measure the performance of the prediction model, namely, profitability and prediction accuracy of the model. The measurement of profitability includes annualized return, rate of return and Sharpe ratio. The measurement of prediction accuracy includes accuracy, MAE, MSE, RMSE, NMSE, MAPE and F-measure. MAPE, MSE, hit ratio and accuracy are the most frequently used indicator to evaluate the prediction accuracy of the model in the 25 articles. However, only [5] articles used profitability to evaluate the performance of the prediction model. Hu et al. (2015) emphasized on the importance of using profitability to measure the performance of a prediction model as profit is the main goal of an investor. Table 2 provides a summary of the evaluation findings of the 25 reviewed articles.

Table.2: Evaluation of Proposed Stock Prediction Model

Model	Performance Metrics	Benchmark	Results
BNNMAS [15]	MAPE	GANN, GRNN	BNNMAS has the lowest MAPE
FWSVM-FWKNN [11]	MAPE, RMSE	SVM-KNN	The proposed model has lower MAPE, RMSE and outperforms SVM-KNN in the medium and long term
GA-SVM [13]	IR, AR (including costs), MD, CDI	HONN, NBC, AR-MA, MACD, BHS	GA-SVM outperforms all benchmark models in terms of AR, CDI, and IR after transaction costs.
EMD-LSSVM [13]	NMSE, MAPE, HR	WD-LSSVM	EMD-LSSVM (parameters selected by GS) has the smallest NMSE, MAPE and highest HR.
CEFLANN [12]	Profitability	SVM, NB, KNN, DT	CEFLANN generates the highest profit
NSVM-KNN [28]	MSE	FLIT2FNS CEFLANN	NSVM-KNN outperforms all benchmark models.
PCA-SVM [48]	Prediction Accuracy	Nil	SVM classification method is accurate and efficient
TBSM-SVR [46]	Profitability	PLR, PLR-SVR, PLR-BPN, SM	TBSM-SVR outperforms all benchmark models
MLP-ANN [37]	Prediction Accuracy	BHS	MLP-ANN produced comparable results without parameter optimization.
GA-ANN [20]	Prediction Accuracy	ANN-SVM	GA-ANN performs better than ANN-SVM, but the prediction accuracy is not high.
BPNN [34]	MSE	CBPM	The proposed model performs better with parameter training.
SVR, RF, ANFIS [10]	MSE, MAPE	CBPM	ANFIS outperforms the other proposed models
GA-ANN [33]	HR	CBPM	GA-ANN Type 2 performs better than GA-ANN Type 1
LSSVM [16]	RMSE	SLS-SVM ANN	The hybrid model, FPA-LV-SVM produced the lowest RMSE.
ANN, LR, SVM, KNN, RF, AB, KF [6]	AUC, Prediction Accuracy, Profitability	CBPM	RF outperforms the other proposed models
LSSVM [42]	Prediction Accuracy	PNN, LDA, QDA	LSSVM outperforms the benchmark models.
SVM [40]	Prediction Accuracy	CBPM	SVM is most suitable method for stock market forecasting problem.
ANN-RS [7]	Prediction Accuracy	CBPM	ANN-RS has 97% accuracy which outperformed the single ANN, RS models
PCA-SVM [43]	Hit Ratio	SVM, ANN, PCA-ANN	The PCA-SVM slightly outperformed PCA-AAN and both models outperformed single SVM, ANN models.
1st Layer SVR 2nd Layer SVR-ANN, SVR-RF, SVR-SVR [31]	MAPE, MAE, rRMSE, MSE	CBPM	The SVR-ANN outperformed the other proposed models. Multi-stage analysis shows significant improvement in prediction accuracy.
ANN, SVM, RF, NB [32]	Prediction Accuracy, F-measure	CBPM	NB with 90.19% accuracy outperformed other proposed models.
BDT, LR, SVM [27]	Prediction Accuracy	CBPM	BDT outperformed other proposed models.
ERF [23]	Prediction Accuracy	RF	ERF achieved higher average prediction accuracy, outperformed the RF
SVM [36]	AR, MD, SD, SR	BHS	The SVM model outperformed BHS in downtrend than uptrend in S&P 500
VWSVM [49]	RoR, MD	BHS, SVM	VWSVM outperformed BHS and SVM.

4.1. Review result

The systematic study of all the proposed prediction models presented in the 25 articles led to the discovery of 4 essential characteristics of a stock prediction model which are the type of input variables, type of prediction technique, type of optimization approach and number of analysis layer.

Type of Input Variables: Most of the studies employ technical indicators as the inputs of the prediction model. Hu et al. (2015) identified that the wide access to data required to form the technical indicators and strong profitability of technical analysis are the prominent drivers. The combination of various technical variables managed to produce different magnitudes of accuracy. Qiu, Li and Song [33] tested the hybrid Genetic Algorithm-Artificial Neural Network (GA-ANN) model with 2 different sets of technical variables. The authors concluded that GA-ANN Type 2 outperformed GA-ANN Type 1 by achieving a hit ratio of 86.39%, thus, the combination of different technical indicators does produce significant impacts on the performance of the prediction model. However, the quantity of indicators does not secure better prediction accuracy of the model. Ballings et al. [6] incorporated more than 80 fundamental input variables in testing the single classifier and ensemble techniques, in which the authors concluded that RF was the top performer among the other proposed benchmark models. However, their best RF model failed to generate significant improvement compared to the principle component analysis-support vector machine (PCA-SVM) model proposed by [48] that incorporated only seven fundamental input variables and the empirical model decomposition-least square support vector machine (EMD-LSSVM) model proposed by [9] that incorporated ten fundamental input variables. According to Zbikowski [49], the inclusion of irrelevant variables may even influence the prediction model negatively.

Type of Prediction Technique: The collection of a diversified set of input variables, on the other hand, is essential as Weng, Ahmed and Megahed [44] showed that the combination of heterogeneous financial data from multiple sources enhanced the performance of the stock prediction model. The type of underlying prediction technique is an essential element in building an accurate prediction model. Ballings et al. [6] examined the prediction accuracy of benchmark single classifier techniques such as SVM, ANN, LR

and KNN and ensemble techniques such as AB, RF and KF. The findings were that the RF model generated the highest accuracy followed by SVM, KF and AB, proposing that ensemble techniques are generally superior in predicting the stock prices. Patel et al. [32] demonstrated that RF prediction model without optimization generated the highest accuracy and outperformed ANN, SVM and NB. The performance of the prediction model can be enhanced with various optimization methods such as knowledge discovery process, feature selection and parameter optimization. Patel et al. [32] improved the efficiency of the proposed models with trend deterministic data preparation procedure to extract potential trends from the technical indicators. Instead of direct deployment of the technical indicators into the prediction model, the trends information was used for prediction. The procedure enhanced the prediction accuracy of the proposed models in which NB with an accuracy of 90.19% outperformed RF with an accuracy of 89.98%.

Type of Optimization Approach: The authors utilized the new insights discovered were utilized in the reviewed work as the new input variables in the stock prediction model, which led to significant positive results. Zbikowski [49] proved that the practice of feature selection will significantly enhances the accuracy of the prediction model. The author ranked the features by assigning Fisher score values to each feature. Information gain ratio is used in [11] to estimate the significance of the selected features and weighted each feature to reduce the prominence of insignificant features. The formation of hybrid models by combining multiple machine learning techniques through parameter optimization produces a higher prediction accuracy than a single machine learning technique model. The SVM model proposed by [36] performed better than buy and hold (B&H) strategy when there was a downtrend in the S&P 500 stock exchange. [21] of the 25 articles proposed hybrid prediction models and all the proposed models outperformed the single technique benchmark models. The hybrid prediction model, artificial neural networks-rough set (ANN-RS) proposed by [7] achieved an accuracy of 97% and outperformed the single RS forecasting model and ANN forecasting model. However, the success of the proposed prediction model which able to outperform the benchmark models does not ensure a high prediction accuracy. For example, the GA-ANN hybrid model proposed by [20] outperformed the benchmark model, GA-SVM but only achieved an average prediction accuracy of 63.60%, which is the lowest among all the hybrid models with the same performance metrics.

Number of Analysis Layer: Patel et al. [31] proposed a 2-stage hybrid prediction model that consists of [2] layers of analysis. The authors used support vector regression (SVR) in the first layer of analysis. In the second analysis layer, a hybrid approach combining [3] machine learning techniques namely ANN, RF and SVR created 3 hybrid prediction models which are SVR-SVR, SVR-ANN and SVR-RF. The future value of statistical parameters predicted in layer 1 are incorporated into the prediction models in layer 2. SVR-ANN outperformed the other benchmark models. However, there is only 1 out of the 25 reviewed articles that proposed a multi-stage hybrid prediction model. 24 out of 25 reviewed articles proposed single layer consists of a machine learning technique or a combination of machine learning techniques. The model proposed by [31] analysed more information than a single layer model as different types of information are analysed in each layer, which contributed significantly in the prediction accuracy of the model. The multifunctional characteristic of the proposed 2-stage prediction model enabled the application of the underlying approach in multiple domains, such as forecasting the gross domestic product (GDP), energy consumption and weather.

5. Limitation and future work

There are 3 categories of limitations identified based on the analysis presented in the previous section namely limitations in the data pre-processing process, limitations in the number of analysis layers in the prediction model and limitations in the selection of performance metrics. The limitations are further described below.

Limitations in Data Pre-Processing: The number of relevant input variables is a critical element in building a prediction model. Thus, a comprehensive set of variables is necessary so that more highly relevant variables can be identified through the data pre-processing process. Aggarwal [1] asserted that the data pre-processing phase is the most essential step in developing a robust analytics model. However, it is observed that this is often not given much attention in the 25 reviewed articles, as the authors positioned the analytical aspect as the focal point in their proposed prediction models. Data pre-processing affects the prediction accuracy of the model. As an example, Qiu, Song and Akagi [34] emphasized that ANN is practical in predicting the stock market but was limited by the amount of noise in the financial data. Consequently, Qiu, Li and Song [33] suggested that a data pre-processing step was necessary to improve the accuracy of the prediction model. Feature selection is very crucial in stock market prediction as different sets of input variables generate different levels of prediction accuracy. However, the emphasis on feature selection is light in most of the reviewed prediction models, as most of the 25 reviewed articles focused on parameter optimization in the machine learning techniques. The ANN-based hybrid prediction model proposed by [15] proved the significance of parameter optimization in improving the prediction accuracy. However, [19] emphasized that removal of noisy input data will improve the performance of ANN-based prediction model. Thus, the implementation of data pre-processing improves the accuracy of the prediction model as the influence of irrelevant data will be reduced.

Limitations in the Number of Analysis Layer: Of the 25 reviewed articles, only 1 multi-stage prediction model. In [31], the hybrid 2-stage stock prediction model generated significant enhancement in prediction accuracy, in comparison with the single stage prediction model. The concept of multi-stage analysis proposed by the authors in 2014 provides new insights to improve the prediction model, but the implementation of the multi-stage analysis concept is limited in the prediction models reviewed from 2014 to 2017. The scarce adoption of multi-stage analysis concept in the recent development of machine learning driven prediction models will impede the advancement of the model to achieve higher prediction accuracy.

Limitations in the Selection of Performance Metrics: The wide variation in the performance metrics used to measure the prediction model prevent the comparison of the different prediction model. All the 25 reviewed articles utilized different performance metrics to measure the accuracy of the model. Thus, the availability of a fixed and comprehensive set of performance metrics is required for systematic comparison between different prediction models. In addition, the reliability of the accuracy of a prediction model will be validated by the diversity of performance metrics employed to measure the model. Unfortunately, 21 of the 25 articles used less than [3] performance metrics, and only [6] combined profitability and accuracy factors to measure the performance of the model.

Future Directions: The detailed examination of the 25 articles enabled the identification of the shortcomings in each model, which led to great research potential in the construction of an accurate and veracious prediction model. The development of potential methods to overcome the [4] limitations discussed in the previous section will be a significant research scope in the future. The adoption of data pre-processing steps as discussed in [1] to process the raw data of the prediction model will increase the quality of the input variables. There are [5] crucial data pre-processing steps, namely, data cleaning, data integration, data transformation, data reduction and data discretization. Data cleaning involves imputation of missing values, smoothing the noisy data and reducing inconsistencies in the data. The integration of different representations of data is required to form an aggregated set of data to be transformed, reduced and discretized. The employment of data

pre-processing in, prediction model will facilitate the identification of highly relevant input variables.

The development of multiple analysis layers in the prediction model is a feasible approach to enhance the performance of the prediction model as more information can be analyzed. Patel et al. [31] proved that the capability of the two-stage prediction model to analyze different in each stage of the model contributed significantly to the accuracy of the model. Thus, the incorporation of multi-stage analysis concept in the research of stock prediction model will lead to the discovery of new insights as compared to the conventional single stage model configuration.

The future research of a fixed and comprehensive set of performance metrics in measuring the performance of the prediction model in essential to examine the performance of different prediction models. The adoption of a set of standard performance metrics will facilitate the comparison between models.

6. Conclusion

The application of machine learning techniques in predicting stock prices is a crucial element in finance. This paper aims to study the recent development of machine learning techniques in the construc-tion of a stock model from 2012 to 2017. A classification frame-work is established to organize the 25 reviewed articles to identify the characteristics of each model. There are [4] identified crucial vari-ables that influence the performance of a prediction model namely the type of input variables, type of prediction technique, type of optimization approach and number of analysis layer. The limitations and potential improvements on the [four] variables are discussed. Alt-hough the review in this paper cannot claim to be comprehensive, the paper will contribute to the state of knowledge to facilitate future research on the subject. However, the work can be expanded to include more than [7] online databases to search for the articles and publications in languages other than English.

Appendix

AB	Ada Boost
ANFIS	Adaptive network-based fuzzy inference system
ANN	Artificial neural network
ANN-RS	Artificial neural network & rough set theory
BDT	Boosted decision degree
BNNMAS	Bat-neural network multi-agent system
BPNN	Backpropagation neural network
CEFLANN	Computational efficient functional link artificial neural network
EMD-LSSVM	Empirical mode decomposing least square support vector machine
ERF	Enhanced random forest
FWSVM-FWKNN	Feature weighted support vector machine & feature weighted K-nearest neighbor algorithm
GA-ANN	Genetic algorithm & artificial neural network
GA-SVM	Genetic algorithm & support vector machine
KF	Kernel factory
KNN	K-nearest neighbor
LR	Logistic regression
LSSVM	Least square support vector machine
MLP-ANN	Multilayer perceptron & artificial neural network
NSVM-KNN	Naïve support vector machine & K-nearest neighbor
NB	Naïve Bayes
PCA-SVM	Principal component analysis & support vector machine
RF	Random forest
SVM	Support vector machine
SVR	Support vector regression
SVR-ANN	Support vector regression & artificial neural network
SVR-RF	Support vector regression & random forest
SVR-SVR	Support vector regressions & support vector regression
TBSM-SVR	Trend-based segmentation method & support vector regression
VWSVM	Volume weighted support vector machine

Financial Indicator

F	Fundamental
T	Technical
O	Others

Optimization Approach

ABC	Artificial bee colony
BA	Bat algorithm
BP	Backpropagation
EMD	Empirical mode decomposition
EML	Extreme machine learning
FPA	Flower pollination algorithm
FW	Feature weighting
GA	Genetic algorithm
GRBF	Gaussian radial basis function
GS	Grid search
HTK	Hyperbolic tangent kernel
HTRBF	Heavy tailed radian basis function
IG	Information gain
Lin	Linear
MCS	Modified cuckoo search
MP	Multilayer perceptron
PCA	Principal component analysis
Poly	Polynomial
PSO	Particle swarm optimization
RBF	Radial basis function
RBKF	Radial based kernel function
SA	Simulated annealing
SGS	Simplex grid search
SRA	Stepwise regression analysis
TDDPP	Trend deterministic data preparation process

Performance Metrics

AR	Annualized return
AUC	Operating characteristics curve
CDI	Correct directional change
HR	Hit ratio
IR	Information ratio
MAPE	Mean average percentage error
MD	Maximum drawdown
MSE	Mean square error
NMSE	Normalized mean square error
RMSE	Root mean square error
RoR	Rate of return
rRMSE	Relative root mean square error
SD	Standard deviation
SR	Sharpe ratio

Benchmark

ANN	Artificial neural network
ANN-SVM	Artificial neural network & support vector machine
ARMA	Autoregressive moving average model
BHS	Buy and hold strategy
CEFLANN	Computational efficient functional link artificial neural network
DT	Decision tree
FLIT2FNS	Functional link net and interval type-2 fuzzy logic system
GANN	Genetic algorithm neural network
GRNN	Generalized regression neural network
HONN	Higher-order neural network
KNN	K-nearest neighbor
LDA	Linear discriminant analysis
MACD	Moving average convergence divergence
NBC	Naïve Bayesian classifier
NB	Naïve Bayes
PCA-ANN	Principal component analysis & artificial neural network
PLR	Piecewise linear representation
PLR-BPN	Piecewise linear representation & backpropagation neural network
PLR-SVR	Piecewise linear representation & support vector regression
PNN	Probabilistic neural network
QDA	Quadratic discriminant analysis
RF	Random forest
SLS-SVM	Standard least square support vector machine
SM	Statistical model
SVM	Support vector machine
SVM-KNN	Support vector machine & K-nearest neighbor
WD-LSSVM	Wavelet denoising least squares support machine

References

[1] C. Aggarwal, *Data Mining: The Textbook*. 1st ed., New York: Springer, 2015.

[2] M. C., Angadi and A. P., Kulkarni, "Time series data analysis for stock market prediction using data mining techniques with R," *International Journal of Advanced Research in Computer Science*, vol. 6, no. 6, pp. 105-109, 2015.

[3] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts and Applications for Engineers and System Designers*. 1st ed. New York: ApressOpen, 2015.

[4] M. Ballings and D. Van den Poel, "CRM in Social Media: Predicting increases in Facebook usage frequency," *European Journal of Operational Research*, vol. 244, no. 1, pp. 248-260, 2015.

[5] M. Ballings and D. Van den Poel, "Customer event history for churn prediction: How long is long enough?," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13517-13522, 2012.

[6] M. Ballings, D. Van den Poel, N. Hespeels and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Systems with Applications*, vol. 43, no. 20, pp. 7046-7056, 2015.

[7] S. Banik, A. F. M. K. Khan and M. Anwer, "Hybrid machine learning technique for forecasting Dhaka stock market timing decisions," *Computational Intelligence and Neuroscience: CIN*, vol. 2014. pp. 1-6, 2014.

[8] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. 3rd ed., New York: Springer, 2016.

[9] J. Chai, J. Du, K. K. Lai and Y. P. Lee, "A hybrid least square support vector machine model with parameters optimization for stock forecasting," *Mathematical Problems in Engineering*, vol. 2015, pp. 1-7, 2015.

[10] T. D. Chaudhuri, I. Ghosh, and S. Singh, "Application of machine learning tools in predictive modelling of pairs trade in Indian stock market," *IUP Journal of Applied Finance*, vol. 23, no. 1, pp. 5-25, 2017.

[11] Y. Chen and Y. Hao, "A feature weighted support vector machine and k-nearest neighbour algorithm for stock market indices prediction," *Expert Systems with Applications*, vol. 80. P. 340-355, 2017.

[12] R. Dash and P. K. Dash, "A hybrid stock trading framework integrating technical analysis with machine learning techniques," *The Journal of Finance and Data Science*, vol. 2, no. 1, pp. 42-57, 2016.

[13] C. L. Dunis, S. D. Likothanassis, A. S. Karathanasopoulos and G. S. Sermpinis, "A hybrid genetic algorithm–support vector machine approach in the task of forecasting and trading," *Journal of Asset Management*, vol. 14, no. 1, pp. 52-71, 2013.

[14] B. Graham, *The Intelligent Investor*. 1st ed., New York: Harper & Brothers, 1949.

[15] R. Hafezi, J. Shahrabi and E. Hadavandi, E., "A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price," *Applied Soft Computing*, vol. 29, pp. 196-210, 2015.

[16] S. Hegazy, O. S. Soliman and M. A. Salam, "Comparative study between FPA, BA, MCS, ABC, and PSO algorithms in training and optimizing of LS-SVM for stock market prediction," *International Journal of Advanced Computer Research*, vol. 5, no. 18, pp. 35-45, 2015.

[17] K. Hong and E. Wu, "The Roles of Past Returns and Firm Fundamentals in Driving Us Stock Price Movements," *International Review of Financial Analysis*, vol. 43, pp. 62-75, 2016.

[18] Y. Hu, K. Liu, X. Zhang, L. Su, E. W. T. Ngai and M. Liu, "Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review," *Applied Soft Computing*, vol. 36, pp. 534-551, 2015.

[19] P. A. Idowu, C. Osakwe, A. K. Anderonke and E. R. Adagunodo, "Prediction of stock market in Nigeria using artificial neural network," *International Journal of Intelligent Systems and Applications*, vol. 4, no. 11, pp. 68-74, 2012.

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

[21] S. Kamley, S. Jaloree and R. S. Thakur, "Performance forecasting of share market using machine learning techniques: A review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, pp. 3196-3204, 2016.

[22] Keele University, "Guidelines for performing systematic literature reviews in software engineering," *Keele University*, 2006. [Online]. Available: https://userpages.uni-koblenz.de/~laemmel/esecourse/slides/slr.pdf [Accessed: Oct. 27, 2017].

[23] H. Kim and S. T. Han, "The enhanced classification for the stock index prediction," *Procedia Computer Science*, vol. 91, pp. 284-286, 2016.

[24] B. G. Malkiel and E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383-417, 1970.

[25] B. G. Malkiel, "The efficient market hypothesis and its critics," *Journal of Economic Perspectives*, vol. 17, no. 1, pp. 59-82, 2003.

[26] R. K. Narang, Inside the Black Box: *The Simple Truth About Quantitative Trading*. 1st ed., New Jersey: John Wiley & Sons, 2009.

[27] A. Nayak, M. M. M. Pai and R. M. Pai, "Prediction models for Indian stock market," *Procedia Computer Science*, vol. 89, pp. 441-449, 2016.

[28] R. K. Nayak, D. Mishra and A. K. Rath, "A naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices," *Applied Soft Computing*, vol. 35, pp. 670-680, 2015.

[29] R. T. F. Nazario, J. L. Silva, V. A. Sobreiro and H. Kimura, "A literature review of technical analysis on stock markets," *The Quarterly Review of Economics and Finance*, vol. 66, pp. 115-126, 2017.

[30] G. Paleologo, A. Elisseeff and G. Antonini, "Subagging for credit scoring models," *European Journal of Operational Research*, vol. 201 no. 2, pp. 490-499, 2010.

[31] J. Patel, S. Shah, P. Thakkar and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2162-2172, 2014.

[32] J. Patel, S. Shah, P. Thakkar and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259-268, 2014.

[33] M. Qiu, C. Li and Y. Song "Application of the artificial neural network in predicting the direction of stock market index," In Proc. 2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), 2016, pp. 219-223.

[34] M. Qiu, Y. Song and F. Akagi, "Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market," *Chaos, Solitons & Fractals*, vol. 85, pp. 1-7, 2016.

[35] V. Ravi, D. Pradeepkumar and K. Deb "Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms," *Swarm and Evolutionary Computation*, vol. 36. pp. 136-149, 2017.

[36] R. Rosillo, J. Giner, D. D. Fuente and R. Pino, "Trading system based on support vector machines in the S&P 500 Index," In Proc. International Conference on Artificial Intelligence (ICAI), 2012, pp. 1-5.

[37] O. B. Sezer, A. M. Ozbayoglu, E. Dogdu "An Artificial Neural Network-based Stock Trading System Using Technical Analysis and Big Data Framework," In Proc. ACMSE 2017 The Annual ACM Southeast Conference Featuring Multidisciplinary and Interdisciplinary Computing, 2017, pp. 223-226.

[38] T. Suciu, "Elements of Stock Market Analysis," *Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences*, vol. 6, no. 2, pp. 153-160, 2013.

[39] L. A. Teixeira and A. L. Oliveira, "A method for automatic stock trading combining technical analysis and nearest neighbour classification," *Expert Systems with Applications*, vol. 37, no. 10, pp. 6885-6890, 2010.

[40] V. P. Upadhyay, S. Panwar, R. Merugu and R. Panchariya, "Forecasting stock market movements using various kernel functions in support vector machine," In Proc. International Conference on Advances in Information Communication Technology & Computing, 2016, pp. 1-5.

[41] J. M. Verner, O. P. Bereton, B. A. Kitchenham, M. Turner and M. Niazi, "Systematic literature reviews in global software development: A tertiary study," *IET Conference Proceedings*, vol. 2012, pp. 1-10, 2012.

[42] S. Wang and W. Shang, "Forecasting direction of China Security Index 300 movement with least squares support vector machine," *Procedia Computer Science*, vol. 31, pp. 869-874, 2014.

[43] Y. Wang, "Stock price direction prediction by directly using prices data: An empirical study on the KOSPI and HIS. Int," *J. Business Intelligence and Data Mining*, vol. 9, no. 2, pp. 145-160, 2014.

[44] B. Weng, M. A. Ahmed and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Systems with Applications*, vol. 79, pp. 153-163, 2017.
[45] World Federation of Exchange, "WFE FY 2016 Market Highlights," *World Federation of Exchange*, 2016. [Online]. Available: https://www.world-exchanges.org/home/index.php/statistics/market-highlights. [Accessed: Oct. 12, 2017].
[46] J. L. Wu and P. C. Chang, "A trend-based segmentation method and the support vector regression for financial time series forecasting," *Mathematical Problems in Engineering*, vol. 2012. pp. 1-20, 2012.
[47] R. Yamamoto "Intraday technical analysis of individual stocks on the Tokyo Stock Exchange," *Journal of Banking & Finance*. vol. 36, no. 11, pp. 3033-3047, 2012.
[48] H. Yu, R. Chen and G. Zhang "A SVM stock selection model within PCA," *Procedia Computer Science*, vol. 31, pp. 407-412, 2014.
[49] K. Zbikowski, "Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1797-1805, 2014.