Journal of International Technology and Information Management

Volume 28 | Issue 4 Article 3

2020

Machine Learning Stock Market Prediction Studies: Review and Research Directions

Troy J. Strader
Drake University, Troy.Strader@drake.edu

John J. Rozycki
Drake Univ, john.rozycki@drake.edu

THOMAS H. ROOT
DRAKE UNIV, TOM.ROOT@DRAKE.EDU

Yu-Hsiang John Huang
Drake University, yu-hsiang.huang@drake.edu

Recommended Citation

Strader, Troy J.; Rozycki, John J.; ROOT, THOMAS H.; and Huang, Yu-Hsiang John (2020) "Machine Learning Stock Market Prediction Studies: Review and Research Directions," *Journal of International Technology and Information Management*: Vol. 28: Iss. 4, Article 3.
DOI: https://doi.org/10.58729/1941-6679.1435
Available at: https://scholarworks.lib.csusb.edu/jitim/vol28/iss4/3

Machine Learning Stock Market Prediction Studies: Review and Research Directions

Troy J. Strader
John J. Rozycki
Thomas H. Root
Yu-Hsiang (John) Huang
(Drake University)

ABSTRACT

Stock market investment strategies are complex and rely on an evaluation of vast amounts of data. In recent years, machine learning techniques have increasingly been examined to assess whether they can improve market forecasting when compared with traditional approaches. The objective for this study is to identify directions for future machine learning stock market prediction research based upon a review of current literature. A systematic literature review methodology is used to identify relevant peer-reviewed journal articles from the past twenty years and categorize studies that have similar methods and contexts. Four categories emerge: artificial neural network studies, support vector machine studies, studies using genetic algorithms combined with other techniques, and studies using hybrid or other artificial intelligence approaches. Studies in each category are reviewed to identify common findings, unique findings, limitations, and areas that need further investigation. The final section provides overall conclusions and directions for future research.

Keywords: machine learning, stock market prediction, literature review, research taxonomy, artificial neural network, support vector machine, genetic algorithm, investment decision

INTRODUCTION

The world's stock markets encompass enormous wealth. In 2019, the value of global equites surpassed $85 trillion (Pound, 2019). As long as markets have existed, investors have searched for ways to acquire knowledge about the companies listed in the market to improve their investment returns. In the past, investors relied upon their personal experience to identify market patterns, but this

  ISSN: 1941-6679-On-line Copy

is not feasible today due to the size of the markets and the speed at which trades are executed. Simple statistical analysis of financial data provides some insights but, in recent years, investment companies have increasingly used various forms of artificial intelligence (AI) systems to look for patterns in massive amounts of real-time equity and economic data. These systems support human investment decision-making and they have now been used for a sufficiently long period that their features and performance can be reviewed and analyzed to identify which systems improve predictive performance when compared with other techniques.

The objective for this study is to identify directions for future machine learning (ML) stock market prediction research based upon a review of current literature. A systematic literature review methodology is used to identify relevant peer-reviewed journal articles from the past twenty years, evaluate and categorize studies that have similar methods and contexts, and then compare the studies in each category to identify common findings, unique findings, limitations, and areas that need further investigation. This will provide artificial intelligence and finance researchers with directions for future research into the use of ML techniques to predict stock market index values and trends.

A systematic literature review methodology has provided insights into ML applications across a wide range of information technology (IT) and scientific domains. The following are four highly cited articles that used this methodology. Wen et al. (2012) employed a systematic literature review to evaluate empirical studies of ML models for software development effort estimation (SDEE). SDEE is the process for predicting the effort required for a software development project. The selected articles were published between 1991 and 2010. The authors looked at four dimensions for each system: machine learning technique, model estimation accuracy, model components, and estimation context. This literature review methodology provided insights into the current state of SDEE research that was the basis for researcher recommendations and guidelines for practitioners.

In another IT-related context, Malhotra (2015) conducted a review of literature where ML models were used for software fault prediction. Software fault prediction is a process used in the early phases of the software development lifecycle for detecting faulty software modules or classes. The articles that were reviewed were published between 1991 and 2013. Each article was evaluated based on their predictive performance and a comparison with other statistical and machine learning techniques. The articles were ultimately grouped into seven categories and the overall conclusion was that more work was needed to produce generalizable results for ML-based software fault prediction. The study concluded with a set of practitioner guidelines and researcher recommendations.

The systematic literature review methodology has also been used in highly complex scientific prediction domains. One study evaluated machine learning flood prediction systems (Mosavi, Ozturk and Chau, 2018). Machine learning has been used to model complex physical flood processes. These models were intended to aid hydrologists in predicting floods in the short-term and long-term, and identifying cost-effective solutions that minimize risk, loss of human life and property damage. The ML methods described in each article were evaluated based on their robustness, accuracy, effectiveness and speed. Based on the findings from the review, the authors were able to provide guidelines for hydrologists and climate scientists when choosing the best machine learning technique for a prediction task. This review methodology has also been used for analysis of ML-based orthopedics systems (Cabitza, Locoro and Banfi, 2018). Orthopedics is the area of medicine focused on prevention, diagnosis, and treatment of bone and muscle disorders. The authors identified studies of machine learning orthopedics applications that were published over a twenty-year period. They appraised each article based upon its ML technique, orthopedic application, model data, and predictive performance. Given current orthopedics machine learning system capabilities, they concluded that the systems are best used for supporting physician decision making at this time and much more work needs to be done to fully integrate machine learning into orthopedics practice given its complexity and risks.

This present study utilizes a methodology that is similar to the ones described above because it is also evaluating articles describing ML use for predictions in a highly complex problem domain. In this methodology, relevant articles are selected through a systematic search process and studies using similar ML methods are grouped together. A research framework (taxonomy) is then provided to encompass each of the study categories and provide a description for how the categories differ. The studies in each taxonomy category are then assessed to identify common and unique findings within the set of studies. This provides the basis for making researcher recommendations.

The remainder of the paper is organized in the following sections. First, the process used to identify relevant studies is described. Next, based on an assessment of the studies identified, a research framework (taxonomy) is presented that groups the studies based on the ML technique used to predict stock market index values and trends. The studies in each category are then individually summarized and discussed to identify common findings, unique findings, limitations, and areas where more study is needed. The final section provides some answers related to the overall study objective that focuses on identifying directions for future research and recommendations for improving study generalizability.

METHOD FOR IDENTIFYING RELEVANT STUDIES

Each researcher involved in this study conducted an independent search for peer-reviewed journal articles where some form of machine learning was used to predict

  ISSN: 1941-6679-On-line Copy

a stock market related outcome. Articles were found using Google Scholar, EBSCO, and EconLit. To identify findings that are relevant to today's IT environment, only studies from the past twenty years (1999-2019) were included in the final list. Each study used one, or more, machine learning techniques to predict stock market index values or expectations for whether the future index value will rise or fall.

After removing all of the duplicate articles, an initial set of 41 relevant articles were identified. Several studies were eliminated from the list because they only focused on predicting individual stock values. For example, one study sought to create a genetic algorithm (GA) to predict the value for two individual stocks: Infosys and Tata consultancy services (Kumar, et. al., 2011) rather than a stock index. Twenty-six studies were included in the final list to provide a representative sample of research studies in this area. The list is not intended to be a comprehensive list of all related articles, but it does provide sufficient coverage to draw conclusions and make recommendations. Each researcher then reviewed each paper to identify groups of related studies that used a single machine learning technique, or those that used a hybrid or multi-method approach. The result is the following machine learning stock market research taxonomy. Each of the articles fits into one of the following four categories: (a) artificial neural network studies, (2) support vector machine studies, (3) studies using genetic algorithms with other techniques, and (4) studies using hybrid or other artificial intelligence approaches.

MACHINE LEARNING STOCK MARKET PREDICTION STUDY RESEARCH TAXONOMY

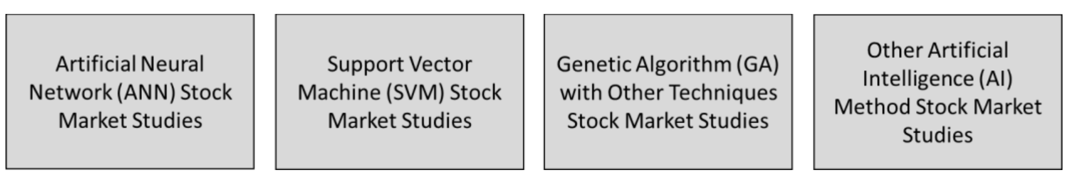

Figure 1. Machine Learning Stock Market Prediction Study Research Taxonomy

In the following section, the individual articles included in each research taxonomy category are summarized focusing on their unique model, dataset and contribution. A complete list of reviewed studies is provided in the Appendix. A brief description of each machine learning approach is also provided prior to describing the related studies.

LITERATURE REVIEW

Studies Using Artificial Neural Networks to Predict Stock Market Values

The first set of articles includes studies that primarily focus on stock market prediction using artificial neural networks (ANNs). ANNs are computational models based on biological neural networks. In the network, sets of nodes are grouped into layers starting with an input layer and ending with an output layer. Signals are transmitted (propagated) through the connected nodes as they learn based on examples and attempt to reduce the level of prediction error. As the system is working to improve its performance, weights are adjusted for the signals between connected nodes. The following provides a brief description of each ANN-related study's unique research focus and findings.

Jasic and Wood (2004) developed an artificial neural network to predict daily stock market index returns using data from several global stock markets. The focus is on trying to support profitable trading. A method is introduced based on univariate neural networks using untransformed data inputs to provide short-term stock market index return predictions. The study uses the daily closing values of the Standard and Poor's 500 Index (S&P 500), the German DAX Index, the Japanese TOPIX index, and London's Financial Times Stock Exchange Index (FTSE All Share). The samples for the S&P 500, DAX and FTSE Index are from January 1, 1965 to November 11, 1999. The sample for TOPIX covers the period from January 1, 1969 to November 11, 1999 since data from earlier years was not available. The prediction performance for the neural network is evaluated against a benchmark linear autoregressive model and prediction improvement is confirmed when applied to the S&P 500 and DAX indices.

Enke and Thawornwong (2005) use a machine learning information gain technique to evaluate the predictive relationships for numerous financial and economic variables. By computing the information gain for each model variable, a ranking of the variables is obtained. A threshold is determined to select only the strongest relevant variables to be retained in the forecasting models. Neural network models for level estimation and classification are examined for their ability to provide an effective forecast of future values. A cross-validation technique is also employed to improve the generalizability of several models. The models are compared using S&P data from a 24-year period from March 1976 to December 1999. The results show that the trading strategies guided by the classification models generate higher risk-adjusted profits than the buy-and-hold strategy, the other neural network models, and the linear regression models.

The next study introduces a stochastic time effective neural network model to uncover the predictive relationships of numerous financial and economic variables (Liao and Wang, 2010). It is presumed that investors choose their investment positions by analyzing historical stock market data, and the historical data are given weights based on how near they are to the present. The nearer the historical data time is to the present, the stronger the impact the data have on the predictive model. The model's effectiveness is analyzed using a numerical experiment based on data from each trading day in an 18-year period from December 19, 1990 to June 7, 2008. The data is from several stock markets including the Shanghai and Shenzhen Stock Exchange Stock A Index (SAI), Stock B Index (SBI), and the Hang Seng (HIS), Dow Jones Industrial Average (DJIA), NASDAQ Composite (IXIC) and S&P500. The forecasting performance of the model is assessed using various volatility parameters.

Chavan and Patil (2013) contribute to our understanding of ANN stock market prediction by surveying different model input parameters found in nine published articles. They attempt to find the most important input parameters that produce better model prediction accuracy. Based on their survey, they find that most ML techniques make use of technical variables instead of fundamental variables for a particular stock price prediction, while microeconomic variables are mostly used to predict stock market index values. In addition, hybridized parameters produce better results when compared with the use of only a single input variable type.

Chong, Han and Park (2017) analyze deep learning networks for stock market analysis and prediction. Deep learning networks extract features from a large set of raw data without relying on prior knowledge of predictors which makes it useful for high frequency stock market prediction. They provide an objective assessment of both the advantages and drawbacks of deep learning algorithms for stock market analysis and prediction. Using high-frequency intraday stock returns as input data, they examine the effects of three unsupervised feature extraction methods—principal component analysis, autoencoder, and the restricted Boltzmann machine—on the network's overall ability to predict future market behavior. Testing is done using data from 38 companies listed in the Korean KOSPI stock market from the period January 4, 2010 through December 30, 2014. Empirical results suggest that deep neural networks can extract additional information from the residuals of the autoregressive model and improve predictive performance.

Studies Using Support Vector Machines to Analyze Stock Markets

The second group of articles includes studies primarily using support vector machines (SVMs) to make stock market predictions. SVMs offer an alternative method to ANNs for improving stock market prediction accuracy through example categorization. The technique uses supervised learning. Training examples are identified as being part of one category or another. An SVM model represents the examples as points in a space with the goal of creating a gap between the categories that is as wide as possible. New examples are classified based on the category in which they most likely belong. For example, in the context of stock market prediction, according to Schumaker and Chen (2010), SVM is a machine learning algorithm that can classify a future stock price direction (rise or drop).

Lee (2009) developed a prediction model based on a support vector machine with a hybrid feature selection method to predict the trend of stock markets. This proposed hybrid feature selection method, named F-score and Supported Sequential Forward Search (F_SSFS), combines the advantages of filter methods and wrapper methods to select the optimal feature subset from the original feature set. To evaluate the prediction accuracy of this SVM-based model combined with F_SSFS, performance is compared with a backpropagation neural network (BPNN) along with three commonly used feature selection methods: information gain, symmetrical uncertainty, and correlation-based feature selection via paired t-tests. The study focuses on predicting the direction of the NASDAQ index using commodity, currency, and other financial market index data from November 8, 2001 through November 8, 2007. It is shown that the SVM outperforms BPN for stock trend prediction. In addition, experimental results show that the proposed SVM-based model combined with F_SSFS has the highest level of predictive accuracy and generalization in comparison with the other three feature selection methods.

A unique study by Schumaker and Chen (2009) used an SVM in conjunction with textual analysis looking at the impact of news articles on stock prices. They developed a predictive machine learning approach for financial news article analysis using several different textual representations: Bag of Words, Noun Phrases, and Named Entities. Through this approach, they investigated a large number of financial news articles and stock quotes covering stocks listed on the S&P500 during a five-week period from October 26, 2005 to November 28, 2005. They estimated a discrete stock price twenty minutes after a news article was released. Using an SVM derivative specially tailored for discrete numeric prediction, and models containing different stock-specific variables, they showed that the model containing both article terms and stock price at the time of article release provided the closest estimate to the actual future stock price, the same direction of price movement as the future price, and the highest return using a simulated trading engine.

Yeh, Huang and Lee (2011) address problems that arise when using support vector regression to forecast stock market values when dealing with kernel function hyperparameters. Typically, a hyperparameter is a parameter whose value is set before the learning process begins. In their system, advantages from different hyperparameter settings can be combined and overall system performance can be improved. They develop a two-stage multiple-kernel learning algorithm by incorporating sequential minimal optimization and the gradient projection method. Experimental results using datasets taken from the Taiwan Capitalization Weighted Stock Index show that the modified method performs better than other methods. The daily stock closing price datasets used for training, validating, and testing the model were from October 2002 through March 2005.

In this next study, the authors recognize that the nature of markets may be different in different regions so their study tests a machine learning model using data from the National Stock Exchange (NSE) of India Limited for the time period from January 1, 2007 to December 31, 2010. Das and Padhy (2012) use two machine learning techniques: backpropagation (BP) and SVM to predict future prices in the Indian stock market. The performance of these techniques are compared and it is observed that SVM provides better results when compared with the results from the BP technique. The implementation is carried out using MATLAB and SVM Tools (LS-SVM Tool Box).

Studies Using Genetic Algorithms with Other Techniques to Analyze Stock Markets

As illustrated in the first two study categories, systems primarily based on ANNs or SVMs have had some success improving stock market value prediction but, over time, there appears to be an increasing interest in trying to further improve results using multi-technique approaches. One alternative machine learning method that has potential to do this is incorporating genetic algorithms (GAs) with either ANNs or SVMs to reduce single technique limitations. A genetic algorithm is a form of evolutionary algorithm (Holland, 1992). The evolutionary process begins with a set of randomly generated problem solutions. In each iterative generation, the fitness of each solution is measured by an objective function. The solutions with higher fitness are retained (survival of the fittest) and combined with other high fitness solutions to create a new generation of solutions. Parent solutions combine to create a new child solution that retains some characteristics from both original solutions. This process continues until a certain number of generations has been created or the population of solutions reaches a satisfactory fitness level. The following studies develop systems that integrate GAs with ANNs and SVMs.

In the first study in this category by Kim and Han (2000), they propose a genetic algorithm approach to feature discretization and the determination of connection weights for artificial neural networks to predict the value of a stock price index. Previous research using the combination of GAs and ANNs have been used for training the network, feature subset selection, and topology optimization. In most of these studies, however, the GA is only used to improve the learning algorithm itself. In this study, the GA is employed not only to improve the learning algorithm, but also to reduce the complexity in the feature space. The research data used in this study is technical indicators and the direction of change in the daily Korea stock price index (KOSPI) from January 1989 to December 1998. Experimental results show that the GA approach incorporated into the feature discretization model outperforms the other two conventional models.

A study by Kim and Lee (2004) is also based on the same two machine learning techniques used in the previous study. They compare a feature transformation method using a GA with two conventional ANN methods. The GA is incorporated to improve the learning and generalizability of ANNs for stock market prediction. The study data includes technical indicators and the direction of change in the daily Korea composite stock price index (KOSPI) for 2,348 trading days from January 1991 to December 1998. Three ANN feature transformation methods are compared. Results achieved by a feature transformation method using the GA are compared against the other two feature transformation methods showing that the performance of the proposed model is better. The authors found that the experimental results indicate that the proposed approach reduces the dimensionality of the feature space and decreases irrelevant factors for stock market prediction.

Kim, Min and Han (2006) develop a unique hybrid system using an ANN and GA to predict stock market index values. Their system is based on the idea of multiple classifier combination where different classifiers attempt to solve the same problem and then their decisions are combined to reduce estimation errors and improve overall classification accuracy. They note that there are limitations when this technique is applied to solving business problems because the problem complexity makes it difficult to completely explain the results provided by ML-driven classifiers. This study proposes an approach that is capable of incorporating the subjective problem-solving knowledge of humans into the results of quantitative models. They use a three-layered backpropagation neural network as the machine-driven classifier. Genetic algorithms are used to combine classifiers stemming from three sources – machine learning, experts, and users. Training, testing and validation data come from the Korea stock price index (KOSPI) over a 572-week period from January 1990 to December 2001.

Kim and Shin (2007) investigate the effectiveness of a hybrid artificial neural network and genetic algorithm method for stock market prediction. The study utilizes two unique networks: adaptive time delay neural networks (ATNNs) and time delay neural networks (TDNNs). To estimate the aspects of the ATNN and TDNN design, a general method based on trial and error along with various heuristics or statistical techniques is proposed. A GA is incorporated into the models to support optimization of the number of time delays and network architectural factors simultaneously for the ATNN and TDNN models. Research data in this study come from the daily Korea Stock Price Index 200 (KOSPI 200) for January 1997 through December 1999. The results show that the proposed integrated approach produces better results than that of the standard ATNN, TDNN, and the recurrent neural network (RNN).

In this hybrid GA and SVM-related study by Yu, Chen, Wang and Lai (2008), an evolving least squares support vector machine (LSSVM) learning paradigm with a mixed kernel is proposed to explore stock market trends. In the proposed learning paradigm, a genetic algorithm is first used to select input features for LSSVM learning. Then another GA is used for parameter optimization for the LSSVM. Finally, the evolving LSSVM learning paradigm with the best feature subset, optimal parameters, and a mixed kernel is used to predict stock market movement direction in terms of historical data series. For evaluation purposes, testing is done using data from three stock indices – the S&P 500, Dow Jones Industrial Average, and the New York Stock Exchange Index. The entire data set of monthly values covers the period from January 1926 to December 2005 with a total of 960 observations. Experimental results reveal that the proposed evolving LSSVM can produce some forecasting models that are more easily interpreted because they use a smaller number of predictive features and are more efficient than other parameter optimization methods.

A dynamic fuzzy model is proposed by Chiu and Chen (2009) in combination with a SVM to explore stock market dynamics. The fuzzy model integrates input variables using factors with varying degrees of influence. A GA adjusts the influential degree of each input variable dynamically. The SVM is then used to predict stock market dynamics. A multi-period experiment is designed to simulate stock market volatility. The 61 input variables in the study include stock market technical indicators, futures market technical indicators, and macroeconomic variables. To evaluate the performance of the new integrated model, they compare it with traditional forecast methods. Stock and futures market data from January 2003 to December 2004 is taken from the Taiwan Stock Exchange Corporation while macroeconomic variables are from the Ministry of Economic Affairs, ROC. The experimental results show that the model is more accurate when compared with alternative prediction methods.

Studies Using Hybrid or Other AI Techniques to Analyze Stock Markets

ANNs, SVMs, or multi-method GA approaches are some of the most common techniques for tackling the problem of stock market prediction. This final category describes studies that have used other unique, or multi-method, artificial intelligence techniques in this problem domain.

Rule-based expert systems have been used for decades to provide domain-specific knowledge to novice decision makers. Lee and Jo (1999) developed a candlestick chart analysis expert system for predicting the best stock market timing. The expert system includes patterns and rules which can predict future stock price movements. Defined patterns are classified into five forms of price movements: falling, rising, neutral, trend continuation, and trend-reversal patterns. The experimental results revealed that the knowledge base they developed could provide indicators to help investors get higher returns from their stock investments. Through experiments using data from a sample of stocks listed in the Korean stock market from January 1992 to June 1997, it was shown that the developed knowledge base was time and field-independent.

Asset allocation is another important stock-related financial decision, but it has received less attention in machine learning studies. O, Lee, Lee and Zhang (2006) present a new stock trading method that incorporates dynamic asset allocation in a reinforcement-learning framework. The proposed asset allocation strategy, called meta policy (MP), is designed to utilize the temporal information from both stock recommendations and the ratio of the stock fund over the asset. Formulating the MP in the reinforcement learning framework is achieved through an environment and learning agent design. Data sets are used to train the meta policy generator and each local trader. Experimental results using Korean stock market (KOSPI) index data from 1998 to 2003 show that the proposed MP method outperforms other fixed asset-allocation strategies and reduces the risks inherent for local traders.

In this next study several machine learning techniques are combined to be used for stock market prediction. According to the efficient market hypothesis, stock prices should follow a random walk pattern meaning that the market should not be predictable with more than about 50 percent accuracy. Qian and Rasheed (2007) investigated the predictability of the Dow Jones Industrial Average index to show that not all periods are equally random. They used the Hurst exponent to select a period with great predictability and found that the best period for analysis was from June 4, 1969, to June 4, 1973 (1010 trading days). Stock market value predictions were made using three inductive machine learning classifiers—an artificial neural network, decision tree, and k-nearest neighbor. Through appropriate model collaboration, the resulting prediction accuracy was better than random at 65 percent.

The following study involves the broadest range of machine learning techniques. Individual data mining techniques have successfully generated accurate stock price movement forecasts, but, over time, traders have realized that they need to use multiple forecasting methods to gather better information about the future of the stock market. In this paper by Ou and Wang (2009), ten different data mining techniques are discussed and applied to predict price movement in the Hong Kong stock market Hang Seng index. The approaches include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbor classification, naïve Bayes based on kernel estimation, logit model, tree-based classification, neural network, Bayesian classification with Gaussian process, support vector machine, and least squares support vector machine (LS-SVM). They examine the daily change of closing prices in the Hang Seng index based on five predictors using data from January 3, 2000 to December 29, 2006. Experimental results show that the SVM and LS-SVM perform better when compared with the other models. Specifically, SVM is better than LS-SVM for in-sample prediction but LS-SVM is, in turn, better than the SVM for the out-of-sample forecasts in terms of hit rate and error rate criteria.

A number of neural network models and hybrid models have been proposed in an attempt to outperform traditional linear and nonlinear approaches for stock market forecasting, but there are some limitations in most ANN model performance in this domain. Guresen, Kavakutlu and Daim (2011) evaluate the effectiveness of a multi-layer perceptron (MLP), a dynamic artificial neural network (DAN2), and hybrid neural networks which use generalized autoregressive conditional heteroscedasticity (GARCH) to extract new input variables. Comparison of methods is made using daily NASDAQ index values from October 7, 2008 to June 26, 2009. One finding is that the simple MLP seems to be the best and most practical ANN architecture.

The focus for this next study is on prediction for the highly dynamic and variable Asian stock markets. In a study by Dai, Wu and Lu (2012), a time series prediction model that combines nonlinear independent component analysis (NLICA) and neural networks is proposed for forecasting Asian stock markets. NLICA is a novel feature extraction technique to find independent sources from observed nonlinear mixture data where no relevant data mixing mechanisms are available. In the proposed method, they first use NLICA to transform the input space composed of original time series data into the feature space consisting of independent components representing underlying information from the original data. Then, the independent components are used as the input variables for the neural network to build the prediction model. To evaluate the performance of the proposed approach, data from the Nikkei 225 closing index and Shanghai B-share closing index from February 2, 2004 through March 3, 2009 are used as illustrative examples. Experimental results show that the proposed forecasting model not only improves the prediction accuracy of the neural network approach but also outperforms the three comparison methods.

The study by Patel, Shah, Thakkar and Kotecha (2015) compares four Indian stock market prediction models: ANN, SVM, random forest, and naive-Bayes with two approaches for model input. The first approach for input data involves computation of ten technical parameters using stock trading data (open, high, low and close prices), while the second approach focuses on representing these technical parameters as trend deterministic data. They assess the accuracy of each of the prediction models for each of the two input approaches. Evaluation is based on data from two stocks and two stock price indices - CNX Nifty (50 of the largest stocks found on the National Stock Exchange (NSE) of India) and the S&P Bombay Stock Exchange (BSE) Sensex for the period January 2003 to December 2012. The experimental results suggest that, for the first input data approach, random forest outperforms the other three prediction models. They also find that the performance of all of the prediction models improves when these technical parameters are represented as trend deterministic data.

Dash and Dash (2016) introduce a novel decision support system using a computationally efficient functional link artificial neural network (CEFLANN) and a rule set to more effectively generate trading decisions. They view the stock trading decision as a classification problem with three possible values – buy, hold or sell. The CEFLANN network used in the decision support system produces a set of continuous trading signals by analyzing the nonlinear relationship that exists between some popular technical indicators. The output trading signals are also used to track trends and to produce trading decisions based on that trend using trading rules. This is a novel approach focused on profitable stock trading decisions through integration of the learning ability of the CEFLANN neural network with the technical analysis rules. The model is compared against other machine learning techniques such as a SVM, a naive Bayesian model, a K nearest neighbor model, and a decision tree. Model training and testing are done using five years (January 4, 2010 to December 31, 2014) of historical stock index price values from the S&P Bombay Stock Exchange Sensitive Index (BSE SENSEX) and S&P 500. The overall study concludes that it is more profitable to make trading decisions based on a combination of technical indicators with computational intelligence tools.

Li et. al. (2016) present the design and architecture for a trading signal mining platform that employs an extreme learning machine (ELM) to make stock price predictions based on two data sources concurrently. Experimental comparisons between ELM and support vector machines and backpropagation neural networks (BPNNs) are made based on the intra-day data of the H-share market (shares of companies incorporated in mainland China that are traded on the Hong Kong Stock Exchange) and contemporaneous news archives. The results show that (1) both RBF ELM and RBF SVM achieve higher prediction accuracy and faster prediction speed than BPNN, (2) the RBF ELM achieves similar accuracy with the RBF SVM, and (3) the RBF ELM has faster prediction speed than the RBF SVM. Simulations of a preliminary trading strategy with the signals is made using data from two sources – a market news archive from Caihua, and H-share market stock prices from 2001. Results show that the strategy with more accurate signals will be more profitable with less risk.

Pierdzioch and Risse (2018) use a ML algorithm known as boosted regression trees (BRT) to implement an orthogonality test of the rationality of aggregate stock market forecasts. The BRT algorithm endogenously selects the predictor variables used to proxy the information set of forecasters so as to maximize the predictive power for the forecast error. The BRT algorithm also accounts for a potential non-linear dependence of the forecast error on the predictor variables and for interdependencies between the predictor variables. Study data includes S&P 500 index forecasts from three groups over the time period 1992 to 2014. Their main finding is that, given the set of predictor variables used in this study, the rational expectations hypothesis (REH) cannot be rejected for short-term forecasts and that there is evidence against the REH for longer term forecasts. Results for three different groups of forecasters corroborate the main finding.

Zhong and Enke (2019) present a process to predict the daily return direction of a set of stocks. Deep neural networks (DNNs) and traditional ANNs are deployed over the entire preprocessed, but untransformed, dataset along with two datasets transformed via principal component analysis (PCA) to predict the daily direction of future stock market index returns. While controlling for overfitting, a pattern for the classification accuracy of the DNNs is detected and demonstrated as the number of the hidden layers increases gradually from 12 to 1000. Simulation results show that the DNNs using two PCA-represented datasets give significantly higher classification accuracy than those using the entire untransformed dataset or other hybrid machine learning algorithms. The trading strategies guided by the DNN classification process based on PCA-represented data perform slightly better than the others tested, including a comparison against two standard benchmarks. The dataset utilized in this study includes the daily direction (up or down) of the closing price of the SPDR S&P 500 ETF as the output, along with 60 financial and economic factors as input features. The daily data is from 2518 trading days between June 1, 2003 and May 31, 2013.

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

The objective for this study is to identify directions for future machine learning stock market prediction research based upon a review of current literature. Given the ML-related systems, problem contexts, and findings described in each selected article, and the taxonomy categories presented earlier, several conclusions can be made about our current knowledge in this research area. First, there is a strong link between ML methods and the prediction problems they are associated with. This is analogous to task-technology fit (Goodhue and Thompson, 1995) where system performance is determined by the appropriate match between tasks and technologies. Artificial neural networks are best used for predicting numerical stock market index values. Support vector machines best fit classification problems such as determining whether the overall stock market index is forecast to rise or fall. Genetic algorithms use an evolutionary problem-solving approach to identify higher quality system inputs, or predict which stocks to include in a portfolio, to produce the best returns. While each study did illustrate that the methods can be effectively applied, the single method applications do have limitations. Hybrid machine learning techniques are one solution that can mitigate some of these limitations. The problem is that, at some point, the systems become so complex that they are not useful in practice. This is a theoretical and practical problem that can be addressed in future studies.

The second conclusion from this review of past studies is that generalizability of findings needs to be improved. Most studies evaluate their ML system using one market and/or one time period without considering whether the system will be effective in other situations. Three enhancements can be made for the experimental system assessment. First, many of the studies are based on results from Asian stock markets. These systems could also be tested in the same time period for US or European markets. Second, the systems could be evaluated using data from times where markets are rising or when markets are declining to assess how they perform in different market environments. For example, would an approach accurately predict market values in the US during the financial crisis of 2008-2009 and also during the recent market growth period from 2018-2019? If systems are able to predict market growth, are they also able to predict market contraction? Finally, proposed methods could be used to evaluate predictive performance for stock market indices that include only small firms vs. only large firms. Are systems effective under different risk and volatility environments? Any of these experimental method enhancements will provide a stronger research and practice contribution.

The final set of conclusions was also apparent after reflection. Financial investment theory needs to be a stronger driver underlying the ML systems' inputs, algorithms, and performance measures. If this is not the case then results may just be random and not have any practical use. Too many studies use techniques without consideration of the vast amount of financial theory that has been developed over the past centuries. Reporting failures where techniques do not improve predictive performance would also be informative. At this point this rarely occurs so it is impossible to find patterns where there is a mismatch between a particular stock market prediction problem and a machine learning technique. Finally, the irony in this research area is that it is a zero-sum game for investors. If the best machine learning stock market prediction technique is found, and all investors adopt this system, the result is that no one is better off. Large investment firms researching the best machine learning methods have no incentive to share this information with others.

REFERENCES

Cabitza, F., Locoro, A., & Banfi, G. (2018). Machine learning in orthopedics: a literature review. *Frontiers in Bioengineering and Biotechnology*, *6*, 75.

Chavan, P. S., & Patil, S. T. (2013). Parameters for stock market prediction. *International Journal of Computer Technology and Applications, 4*(2), 337.

Chiu, D. Y., & Chen, P. J. (2009). Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. *Expert Systems with Applications*, *36*(2), 1240-1248.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83*, 187-205.

Dai, W., Wu, J. Y., & Lu, C. J. (2012). Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes. *Expert systems with applications, 39*(4), 4444-4452.

Das, S. P., & Padhy, S. (2012). Support vector machines for prediction of futures prices in Indian stock market. *International Journal of Computer Applications, 41*(3).

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science*, *2*(1), 42-57.

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, *29*(4), 927-940.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS quarterly*, 213-236.

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications, 38*(8), 10389-10397.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Jasic, T., & Wood, D. (2004). The profitability of daily stock market indices trades based on neural network predictions: Case study for the S&P 500, the DAX, the TOPIX and the FTSE in the period 1965–1999. *Applied Financial Economics*, *14*(4), 285-297.

Kim, H. J., & Shin, K. S. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Applied Soft Computing*, *7*(2), 569-576.

Kim, K. J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications*, *19*(2), 125-132.

Kim, K. J., & Lee, W. B. (2004). Stock market prediction using artificial neural networks with optimal feature transformation. *Neural computing & applications*, *13*(3), 255-260.

Kim, M. J., Min, S. H., & Han, I. (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems with Applications*, *31*(2), 241-247.

Kumar, L., Pandey, A., Srivastava, S., & Darbari, M. (2011). A hybrid machine learning system for stock market forecasting. *Journal of International Technology and Information Management*, *20*(1), 3.

Lee, K. H., & Jo, G. S. (1999). Expert system for predicting stock market timing using a candlestick chart. *Expert systems with applications*, *16*(4), 357-364.

Lee, M. C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, *36*(8), 10896-10904.

Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., ... & Deng, X. (2016). Empirical analysis: stock market prediction via extreme learning machine. *Neural Computing and Applications*, *27*(1), 67-78.

Liao, Z., & Wang, J. (2010). Forecasting model of global stock index by stochastic time effective neural network. *Expert Systems with Applications*, *37*(1), 834-841.

Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing*, *27*, 504-518.

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10*(11), 1536.

O, J., Lee, J., Lee, J. W., & Zhang, B-T. (2006). Adaptive stock trading with dynamic asset allocation using reinforcement learning. *Information Sciences, 176*(15), 2121-2147.

Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science, 3*(12), 28-42.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications, 42*(1), 259-268.

Pierdzioch, C., & Risse, M. (2018). A machine-learning analysis of the rationality of aggregate stock market forecasts. *International Journal of Finance & Economics*, *23*(4), 642-654.

Pound, J. (2019, December 24). Global stock markets gained $17 trillion in value in 2019. Retrieved from https://www.cnbc.com/2019/12/24/global-stock-markets-gained-17-trillion-in-value-in-2019.html.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. Applied Intelligence, 26(1), 25-33.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS), 27*(2), 12.

Schumaker, R. P., & Chen, H. (2010). A discrete stock price prediction engine based on financial news. *Computer, 43*(1), 51-56.

Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*, *54*(1), 41-59.

Yeh, C. Y., Huang, C. W., & Lee, S. J. (2011). A multiple-kernel support vector regression approach for stock market price forecasting. *Expert Systems with Applications*, *38*(3), 2177-2186.

Yu, L., Chen, H., Wang, S., & Lai, K. K. (2008). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on evolutionary computation*, *13*(1), 87-102.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, *5*(1), 4.

APPENDIX
List of Reviewed Machine Learning Stock Market Prediction Articles

Author(s) [publication year]	Machine Learning Method(s)	Primary Market(s) Studied and Data Time Period (Years)
Jasic and Wood [2004]	Artificial neural network	S&P 500, DAX, TOPIX and FTSE 1965-1999
Enke and Thawornwong [2005]	Artificial neural network	S&P 500 1976-1999
Liao and Wang [2010]	Artificial neural network	Shanghai and Shenzhen Stock Exchange 1990-2008
Chavan and Patil [2013]	Artificial neural network	Study reviews nine ANN studies
Chong, Han and Park [2017]	Artificial neural network	Korean KOSPI stock market 2010-2014
Lee (2009)	Support vector machine	NASDAQ 2001-2007
Schumaker and Chen [2009]	Support vector machine	Companies listed in the S&P500 in 2005
Yeh, Huang and Lee [2011]	Support vector machine	Taiwan Capitalization Weighted Stock Index 2002-2005
Das and Padhy [2012]	Support vector machine	National Stock Exchange (NSE) of India Limited 2007-2010
Kim and Han [2000]	Genetic algorithm with artificial neural network	Korea stock price index (KOSPI) 1989-1998
Kim and Lee (2004)	Genetic algorithm with artificial neural network	Korea composite stock price index (KOSPI) 1991-1998

  ISSN: 1941-6679-On-line Copy

Kim, Min and Han [2006]	Genetic algorithm with artificial neural network	Korea stock price index (KOSPI) 1990-2001
Kim and Shin [2007]	Genetic algorithm with artificial neural network	Korea Stock Price Index 200 (KOSPI 200) 1997-1999
Yu, Chen, Wang and Lai [2008]	Genetic algorithm with support vector machine	S&P 500, DJIA and NYSE Index 1926-2005
Chiu and Chen [2009]	Genetic algorithm with support vector machine	Taiwan Stock Exchange 2003-2004
Lee and Jo [1999]	Expert system	Sample of stocks in the Korean stock market 1992-1997
Jangmin, Lee, Lee and Zhang [2006]	Meta policy	Korean stock market (KOSPI) 1998-2003
Qian and Rasheed [2007]	Artificial neural network with decision tree and k-nearest neighbor	DJIA 1969-1973
Ou and Wang [2009]	Linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbor classification, naïve Bayes based on kernel estimation, logit model, tree based classification, neural network, Bayesian classification with Gaussian process, support vector machine, and least squares support vector machine (LS-SVM)	Hang Seng index 2000-2006
Guresen, Kavakutlu and Daim [2011]	Multi-layer perceptron (MLP), dynamic artificial neural network (DAN2), and hybrid neural networks which use generalized autoregressive conditional heteroscedasticity (GARCH)	NASDAQ 2008-2009
Dai, Wu and Lu [2012]	Nonlinear independent component analysis (NLICA) with artificial neural network	Nikkei 225 and Shanghai B-share closing index 2004-2009
Patel, Shah, Thakkar and Kotecha [2015]	Artificial neural network, support vector machine, random forest, and naïve Bayes with trend deterministic data	CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex 2003-2012
Dash and Dash [2016]	Computationally efficient functional link artificial neural network (CEFLANN) and rule set	BSE SENSEX and S&P500 2010-2014
Li, et. al. [2016]	Extreme learning machine (ELM)	Caihua market news archive and H-share market stock prices from 2001
Pierdzioch and Risse [2018]	Boosted regression trees (BRT)	S&P 500 index forecasts from three groups 1992-2014

Zhong and Enke [2019]	Deep neural networks (DNNs) and traditional artificial neural networks	S&P 500 2003-2013