A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques

Mehtabhorn Obthong[1][a], Nongnuch Tantisantiwong[2][b], Watthanasak Jeamwatthanachai[1][c], and Gary Wills[1][d]

[1] *School of Electronics and Computer Science, University of Southampton, Southampton, UK*

[2] *Nottingham Business School, Nottingham Trent University, Nottingham, UK*

{*mo1n18, wj1g14, gbw*}*@soton.ac.uk, nuch.tantisantiwong@ntu.ac.uk*

Keywords: machine learning, deep learning, finance, stock price prediction, time series analysis, sentiment analysis

Abstract: Stock market trading is an activity in which investors need fast and accurate information to make effective decisions. Since many stocks are traded on a stock exchange, numerous factors influence the decision-making process. Moreover, the behaviour of stock prices is uncertain and hard to predict. For these reasons, stock price prediction is an important process and a challenging one. This leads to the research of finding the most effective prediction model that generates the most accurate prediction with the lowest error percentage. This paper reviews studies on machine learning techniques and algorithm employed to improve the accuracy of stock price prediction.

1 INTRODUCTION

In financial markets, a machine learning (ML) has become a powerful analytical tool used to help and manage investment efficiently. ML has been widely used in the financial sector to provide a new mechanism that can help investors make better decisions in both investment and management to achieve better performance of their securities investment. Equity securities are one of the most traded securities (Lin et al., 2018) as they have attractive return (He et al., 2015; Chou and Nguyen, 2018) and are a relatively liquid asset given that they can be resold and repurchased through stock exchanges.

Despite the attractive return, equity investment has high risk due to the uncertainty and fluctuation in the stock market (Hyndman and Athanasopoulos, 2018). Investors must, therefore, understand the nature of individual stocks and their dependence factors that effect to stock prices in order to increase their chances of achieving higher returns. But all these, the investors require to make effective investment decisions at the right time (Ijegwa et al., 2014) using an accurate and appropriate amount of information (Nguyen et al., 2015) e.g. investor sentiment and interest rates.

Price prediction based on a few factors would be easy but the result might be inaccurate because some excluded factors may also be important in explaining the movement of stock prices. The prices of individual stocks can be affected by various factors e.g. economic growth (Perwej and Perwej, 2012; Selvin et al., 2017). It is difficult to analyse all factors manually (Nguyen et al., 2015; Sharma et al., 2017), so it would be better if there were tools for supporting the analysis of this data within a timely response.

Making the right decision within timely response has posed a number of challenges as such a large amount of information is required for predicting the movement of the stock market price. These information are important for investors because stock market volatility can lead to a considerable loss of investment. The analysis of this large information is thus useful for investors, and also useful for analysing the direction of stock market indexes (Kim and Kang, 2019).

With the great success of ML in many fields, research on ML in finance has gained more attention and been studied continuously (Nguyen et al., 2015; Attigeri et al., 2015; Kim and Kang, 2019). Thus, this paper will explore the application of machine learning in finance: employed to algorithms and techniques, exclusively focusing on stock prediction.

[a] https://orcid.org/0000-0002-3869-578X

[b] https://orcid.org/0000-0001-5243-2970

[c] https://orcid.org/0000-0002-4622-0493

[d] https://orcid.org/0000-0001-5771-4088

2 FINANCIAL INSTRUMENTS

Financial instrument is a contract of tradable assets (Lehmann, 2017), such as stocks, bonds, bills, currencies, swaps, futures, and options, that gives the right to part- or wholly-own an entity or to claim the assets of the entity (Staszkiewicz and Staszkiewicz, 2014). Financial assets are claims to the income produced by real assets (e.g. selling cocoa beans, letting a building, providing a service).

2.1 Equity

An equity asset, also known as a share, is issued by a public company to represent partial ownership of the company. Individual or group known as the stockholders or shareholders will have the status of company owner. When the company wishes to expand its business, more capital may be needed to finance this plan. To raise this capital, the company can issue new shares, after approval by existing shareholders (because new issues of shares dilutes their ownership), and sell them to investors. The quoted value of the stock will increase if the company is successful. Therefore, the performance of the stock investment relates to both the success and to the real assets of the company (Bodie et al., 2013).

2.1.1 Stock market

A stock market, also known as equity market, is a public market where traders (investors in the financial markets) buy and sell the company's shares and derivatives by exchanging or processing in electronic or in physical form (Preethi and Santhi, 2012; Göçken et al., 2016). Generally, financial instruments are traded in the capital market comprising a primary market and a secondary market. The primary market is the place where securities are distributed for the first time. The initial public offering (IPO) occurs here. The secondary market refers to the market for trading among the investors. Examples are New York Stock Exchange (NYSE), London Stock Exchange (LSE), Japan Exchange Group (JPX), Shanghai Stock Exchange (SSE), and NASDAQ.

2.1.2 Stock index

Stock index is a representative of a group of stocks' prices. This index is computed from the prices of defined stocks and its change can reflect the overall performance of the stocks listed in the index. In particular, a stock index is a weighted average market value of a number of firms compared with the value on the base trading day (Bodie et al., 2013). For example, the Financial Times Stock Exchange 100 Index (FTSE 100) and Standard & Poor's Composite 500 Index (S&P500)[1]

2.1.3 Stock trading

Stock trading is an important challenge for investors because trading decision and stock prices can be affected by the variety and complexity of information including economic conditions, local politics, international politics, and social factors (Hadavandi et al., 2010; Chourmouziadis and Chatzoglou, 2016; Naranjo et al., 2018). Stock trading involves buying and selling shares in companies. Many different trading methods are used by traders, such as day trading, position trading, swing trading, and scalping (Mann and Kutz, 2016).

2.2 Other financial instruments

Bonds, also known as debt securities, are issued by an obligated borrower to make the specified coupon payments to the holder, also known as a bondholder, over a specified period. Debt instruments include treasury notes and bonds, municipal bonds, corporate bonds, federal agency debt, and mortgage securities (Bodie et al., 2013). Most of these instruments promise either fixed income streams or income streams that are defined from a specific formula. That is the reason why they are sometimes called fixed-income securities.

Derivatives are securities whose payoffs are based on the value of other assets, so-called underlying assets, for example, stocks, currencies, bonds, commodities, etc. (Bodie et al., 2013). Financial derivatives play an important role in the financial markets because they are used to hedge risks occurring from the operational, financing and investment activities of companies (Lehmann, 2017). Four popular types of derivatives are futures, options, forwards, and swaps.

The foreign exchange rate is the price of one currency in term of another currency. The foreign exchange market is a formal network in which the group of banks and brokers can exchange currencies immediately or enter a contract to exchange currencies in the future at the determined rate (Bodie et al., 2013). The contracts traded in the exchange markets divided into three types: spot, outright forward, and swap (Brown, 2017).

Commodities are goods that are interchangeable with the same type and same grade of commodities,

[1] S&P500 is one of leading indicators and the important benchmark for the 500 top-traded companies (Althelaya et al., 2018b).

usually used as a raw material (cocoa, tea, silver) to produce goods or services. Commodities can be traded based on current prices in the spot market, also known as the cash market, or at a pre-specified price in the futures market (Roncoroni et al., 2015). Some commodities can be underlying assets of *derivatives*. Commodities trading in the spot market are used for immediate delivery, but the futures market is used for trading for delivery at an agreed date in the future (Whalley, 2016).

3 MACHINE LEARNING FOR FINANCIAL INSTRUMENTS

Over the past few years, ML has been applied in many research fields, especially finance and economics (Xu and Wunsch, 2005). Many researchers have used ML algorithms to create tools to analyse historical financial data and other related information (e.g. economic conditions) for supporting decision-making in investment. For example, Jeong et al. (2018) used ML algorithms to support decision-making of stock investment by using financial news data and social media data, while Chou and Nguyen (2018) forecast the stock prices of construction companies in Taiwan using a promising non-linear prediction model.

More importantly, using historical or time series financial data, carefully selecting appropriate models, data, and features are all essential in order to produce accurate results. The accurate results depend on efficient infrastructure, collection of relevant information, and algorithms employed (Alpaydin, 2014). The better quality of data, the more accurate the ML result.

With the great success in ML over the past few years, it has changed the way investors use information and it offers optimal analytic opportunities for of all investing types. Thus, ML is a significant tool to help financial investment. Table 1 summarises ML techniques used and applied to forecasting asset returns or finding the pattern or distribution of asset returns. These techniques include clustering, prediction, classification, and others (e.g. portfolio optimisation), while Table 2 presents the advantages and disadvantages of each ML techniques used in the financial fields.

4 TIME SERIES DATA

Time series data are groups of continuous data that were collected over a period of time (T). The data are collected yearly, monthly, weekly, daily or every hour, minute, or second. Examples are the daily exchange rate of pounds sterling (GBP) against the US dollar (USD) between 1 January 2019 and the 31 December 2019, the monthly UK unemployment rate each year, the daily closing price of stocks, and so on.

Time series data is comprised of four components (Yaffee and McGee, 2000):

Trend or secular trend shows the direction of movement of data in the long term. The tendencies may be stable, increasing, or decreasing, during different time intervals.

Cycle is data movement patterns over periods longer than one year. These fluctuations are usually affected by conditions associated with an economic or business cycle (Hyndman and Athanasopoulos, 2018). Cycle is similar to season, but with longer duration of fluctuations, at least two years. The nature of cyclical variation is periodic and will repeat itself; for example, the rise and fall of the number of batteries sold by National Battery Sales, Inc. from 1984 to 2003.

Table 1: Existing algorithms and techniques applied to financial instruments

Methods[*]	**Type of financial instrument**				
	Stocks	**Bonds**	**Derivatives**	**Foreign Exchange**	**Commodities**
Clustering					
K-Means	✓				
SOM	✓				
Hierarchical Clustering	✓				
Prediction					
RF	✓	✓			✓
SVM	✓	✓			
MLP	✓		✓	✓	✓
LSTM	✓				
RNN	✓	✓	✓	✓	✓
GAs	✓		✓		✓
KNN	✓	✓	✓	✓	
SVR	✓	✓	✓	✓	
MCS	✓	✓	✓	✓	
ANNs	✓	✓	✓		
CART	✓	✓			
GP	✓		✓		
BSM	✓		✓		
GRNN	✓				✓
RBF			✓		
BPNN	✓	✓	✓		
LR	✓		✓		
HMM	✓	✓	✓		
Classification					
SVM	✓	✓			✓
KNN	✓	✓		✓	✓
LR		✓			
ANNs		✓			

[*] Definitions of the methods are provided in Appendix section

Table 2: Advantage and disadvantage of each ML algorithms and technique

Methods	**Data**	**Purpose**	**Method**	**Advantages**	**Disadvantages**	**References**
ANNs: Artificial Neural network	Non-time series, Time-series and Financial time series	Classification and Forecasting	Model	+ High ability to tackle complex nonlinear patterns + High accuracy for modelling the relationship in data groups Model can support both linear and non-linear processes + Model is robust and can handle noisy and missing data	- Over fitting - Sensitive to parameter selection - ANNs just give predicted target values for some unknown data without any variance information to assess the prediction	Wang et al. (2011); Göçken et al. (2016); Zhou and Fan (2019)
ARIMA: Autoregressive integrated moving average model	Time-series, Financial time-series	Forecasting and Clustering	Model	+ Works well for linear time series + It is the most effective forecasting technique in social science + For short-run forecasting, it provides more robust and efficient than the relative models with more complex structural	- Does not work well for nonlinear time series - The model determined for one series will not be suitable for another - Requires more data - Takes a long time processing for a large dataset - Requires set parameters and is based on user assumptions that may be false, the resulting clusters being inaccurate - The forecast results are based on past values of the series and previous error terms	Adebiyi et al. (2014); Hyndman and Athanasopoulos (2018); Selvin et al. (2017)
BPNN: Back propagation neural network	Non-time series, Time-series and Financial time series	Forecasting	Model	+ Flexible nonlinear modelling capability + Strong adaptability + Capable of learning and massively parallel computing Popular for predicting complex nonlinear systems + Fast response + High learning accuracy	- Sensitive to noise - Actual performance based on initial values - Slow convergent speed - Easily converging to a local minimum	Zhao et al. (2010); Wang et al. (2015); Singh and Tripathi (2017)
CART: Classification and Regression Trees	Non-time series, Financial time-series	Classification and Forecasting	Model	+ Can model nonlinearity very well + Results are easily interpretable	- Unstable even when the training data are small changed	Pradeepkumar and Ravi (2017)
FCM: Fuzzy c means	Non-time series, Time-series and Financial time series	Clustering	Algorithm	+ Works well for searching spherical-shaped clusters + Work effectively for small to medium datasets	- Sensitive to noise - Has problems with handling high dimensional datasets - The membership of the data point depends directly on the membership values of other cluster centres which may lead to undesirable results	Suganya and Shanthi (2012); Grover (2014)
GAs: Genetic Algorithms	Non-time series, Time-series and Financial time series	Clustering, Classification and Forecasting	Algorithm	+ Can search the clusters with different shapes by using different criteria + One of the best-suited algorithms for learning the time-series datasets Works well for the noisy data + Suitable for peculiarly hard problems when little or no knowledge of the optimal function is given and the search space is very large + Suitable for solving the issue of defining proper parameters for ANNs	- Sensitive to parameter selection	(Alfred et al., 2015; Wang et al., 2011)
GMDH: Group Method of Data Handling	Financial time-series	Forecasting	Algorithm	+ Best ANN for handling the incorrect, noisy, or small datasets + Provides higher accuracy and is an easier structure than traditional ANN models	- Can generate a complicated polynomial even for a simple system - Does not consider the input-output relationship well because of its limited architecture - Inefficient for modelling nonlinear systems that have different characteristics in different environments	Pradeepkumar and Ravi (2017)
GP: Gaussian Processes	Time-series and Financial time-series	Classification and Forecasting	Model	+ Flexible and easy computational implementation + Sufficiently robust to generate the automatic model	- Generates ”black box” models which are difficult to interpret - Can be computationally expensive	Rizvi et al. (2017)
GRNN: Generalized Regression Neural Network	Non-time series, Time-series and Financial time series	Classification and Forecasting	Model	+ Easy to implement because of a much faster training procedure than other ANNs + Useful for performing predictions in real-time + Does not require an iterative training process Can estimate any arbitrary function by adapting the function exactly from the training data + Quick training approach + Provides the high accuracy of both linear and nonlinear functional regressions, based on the kernel estimation theory	- Requires more memory space to store the model - Can be computationally expensive because of its huge size	Pradeepkumar and Ravi (2017); Al-Mahasneh et al. (2018)
Hierarchical Clustering	Non-time series, Time-series	Clustering	Algorithm	+ Does not need to set any parameters, e.g. the number of clusters	- The length of each time series is the same because of the Euclidean distance calculation requirement - Unable to handle long time series effectively because of poor scalability Useful only for small datasets because of its quadratic computational complexity	Wang et al. (2006)

continued on the next page

Table 2: Advantage and disadvantage of each ML algorithms and technique – continued from previous page

Methods	Data	Purpose	Method	Advantages	Disadvantages	References
HMM: Hidden Markov Model	Non-time series, Time-series and Financial time series	Clustering, Classification and Forecasting	Model	+ Strong statistical foundation + Able to model high level information (language model, or syntactical rules)	- Requires parameters to be set and is based on user assumptions that may be false with the result that clusters would be inaccurate - Takes a long time processing for a large dataset	Aghabozorgi et al. (2015); Belgacem et al. (2017)
k-Means	Non-time series, Time-series and Financial time series	Clustering	Algorithm	+ Works well for searching spherical-shaped clusters + Works effectively for small to medium datasets + Faster than hierarchical clustering	- The number of clusters must be specified in advance - Sensitive to noise - Only spherical shapes can be determined as clusters - The quality of clustering is highly dependent on the selection of initial centres The length of each time series is the same because of the Euclidean distance calculation requirement - Unable to handle long time series effectively because of poor scalability	Wang et al. (2006); Boomija and Phil (2008)
k-Medoids (PAM)	Non-time series and Time-series	Clustering	Algorithm	+ Works well for searching spherical-shaped clusters + Works effectively for small to medium datasets + More robust to noisy data and outliers than k-means	- The number of clusters must be specified in advance - Only spherical shapes can be determined as clusters - Does not scale well for large datasets	Boomija and Phil (2008); Aghabozorgi et al. (2015)
KNN: K Nearest Neighbour	Non-time series, Time-series and Financial time series	Classification and Forecasting	Algorithm	+ Robust to noisy training data + Very efficient if the training datasets are large	- The number of nearest neighbours must first be determined - Can be computationally expensive - Memory limitation - Sensitive to the local structure of the data	Archana and Elangovan (2014); Jadhav and Channe (2016)
LR: Logistic Regression	Financial time-series	Classification and Forecasting	Model	+ High ability to tackle complex nonlinear patterns	- Sensitive to outliers - Strong assumptions	Wu and Li (2018)
LSTM: Long Short-Term Memory	Non-time series, Time-series and Financial Time Series	Classification and Forecasting	Model	+ Capable of analysing and exploiting the interactions and patterns existing in data through a self-learning process + Makes good predictions because it analyses the interactions and hidden patterns within the data + Good in remembering information for long time	- Lacks a mechanism to index the memory while writing and reading the data The number of memory cells is linked to the size of the recurrent weight matrices	Selvin et al. (2017); Kumar et al. (2018)
MCS: Monte Carlo Simulation	Financial time-series	Forecasting	Model	+ Very flexible and virtually no limit for analysis + Can model complex systems + All kinds of probability distributions can be modelled + Time to results quite short + Easily understood by non-mathematicians + Easy to see which inputs had the biggest effect on the results	- No interactive link between data and parameters - Unidirectional - Does not allow "backward reasoning"	Smid et al. (2010)
MLP: Multilayer Perceptron	Non-time series, Time-series, Financial time series	Forecasting, Classification	Model	+ Can yield accurate predictions for challenging problems	- Convergence is quite slow - Local minima can affect the training process - Hard to scale	Pradeepkumar and Ravi (2017)
PSO: Particle Swarm Optimization	Non-time series, Time-series and Financial time series	Forecasting	Algorithm	+ Easy to implement + Very few parameters to tweak	- Lacks a solid mathematical foundation for analysing future development of relevant theories	Pradeepkumar and Ravi (2017)
RBF: Radial Basis Function Neural Networks	Non-time series, Time-series and Financial time series	Classification and Forecasting	Model	+ Robust to noisy input + The training is faster than perceptron since there is no back propagation learning involved + Very stable, and a generalization capability + Good comprehensive adaptive and learning abilities + Powerful technique for improvement in multi-dimensional space + Quicker in convergence and more accurate in the model than the Back Propagation Neural Network + Does not suffer from local minima in the same way as the multilayer perceptron + Only has one hidden layer making faster learning than MLP	- Classification process is slower than MLP	Markopoulos et al. (2016)
RF: Random Forest	Non-time series, Time-series and Financial time series	Classification and Forecasting	Algorithm	+ Robust method for forecasting and classification problems since its design that is filled with various decision trees, and the feature space is modelled randomly + Automatically handles missing values + Works well with both discrete and continuous variables	- Requires more computational power and resources because it creates a lot of trees - Requires more time to train than decision trees	Pradeepkumar and Ravi (2017)

continued on the next page

Table 2: Advantage and disadvantage of each ML algorithms and technique – continued from previous page

Methods	Data	Purpose	Method	Advantages	Disadvantages	References
RNN: Recurrent neural networks	Non-time series, Time-series and Financial Time series	Classification and Forecasting	Model	+ Very useful where for showing the time relationships that occur between the inputs and outputs in the neural network	- Difficult to train	Moreno (2011); Bai et al. (2018)
SOM: Self organizing maps	Non-time series, Time-series and Financial time series	Clustering and Classification	Algorithm	+ Robust to parameter selection Yields a good clustering result Excellent data-exploring tool	- Does not work well for time series of unequal length because of the difficulty involved in determining the scale of weight vectors - Sensitive to outliers	Aghabozorgi et al. (2015)
SVM: Support Vector Machine	Non-time series, Time-series and Financial time series	Classification and Forecasting	Algorithm	+ Can provide the optimal global solution and has excellent predictive accuracy capability + Works well on a range of classification problems, such as those with high dimensions	- Sensitive to outliers - Sensitive to parameter selection	Wang et al. (2011)
SVR: Support Vector Regression	Time-series and Financial Time Series	Forecasting	Model	+ Powerful for financial time series prediction + Particularly suited to handle multiple inputs + Provides high prediction accuracy + Ability to tackle the overfitting problem	- Sensitive to users' defined free parameters	Nava et al. (2018)

Seasonality, also known as seasonal variation, seasonal fluctuation or seasonal effect, is movement of data caused by the influence of an annual season or specific period that will be repeated at the same time of the year such as month effects and quarter effects. The influence may be driven by natural conditions, business procedures, social and cultural behaviour.

Irregularity or irregular variation is short-period irregular movements in the time series data possibly due to disasters, wars, or strikes. This variation usually affects business activity in a short term.

Time series analysis has been applied in economics and to finance research (Sharma et al., 2017) such as in economic forecasting, sales forecasting, stock market analysis, and yield projection (Montgomery et al., 2015). As a matter of fact, many ML applications have both been proposed and been adopted to swiftly cope with, and solve problems in time series analysis (Siami-Namini and Namin, 2018).

5 MACHINE LEARNING FOR STOCK PRICE PREDICTION

Stock price prediction has played in the very important role of investments as efficient stock price predictions can provide suggestions on trading strategies. However, there is no guarantee that the stock price prediction using historical data will be 100% accurate due to the uncertainty in the future. For example, stock price can fluctuate (Selvin et al., 2017) depending on political and economic conditions. Thus, investors have used fundamental and technical analysis simultaneously for the stock price prediction (Beyaz et al., 2018).

Fundamental analysis is a method to estimate intrinsic value of a stock by analyzing various internal and external factors that could have effects on the value of stock or company (Selvin et al., 2017). The fundamental factors include business environment, financial performance, economic data, and social and political behaviour (Beyaz et al., 2018).

Technical analysis is a method to predict the future stock prices (Selvin et al., 2017) by using historical data. This method focuses on an analysis of trends of securities' prices such as daily opening, high, low, and closing prices as illustrated in Figure 1. In addition, other features may be considered and used in the technical analysis for increasing an accuracy in the prediction, for example, volume and relative strength index (RSI).

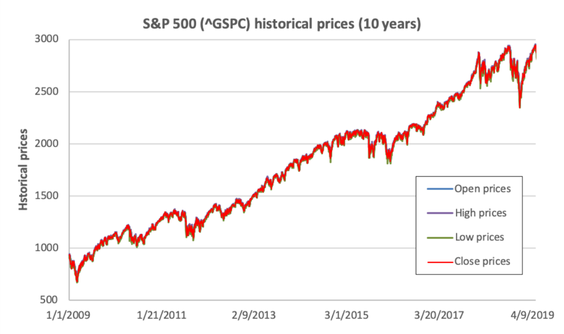

Figure 1: S&P 500 stock index from 1 January 2009 to 31 May 2019

- *Opening Price* is the first price of any listed stock at the beginning of an exchange on a trading day.
- *High* and *Low Prices* are the highest and lowest price of stock on that day. Generally, these data are used by traders to measure volatility of stock.

- *Closing Price* is a price of the stock at the close of the trading day.
- *Volume* is the number of stocks or contracts traded for a security in all the markets during a given time period.
- *Adjusted Closed Prices* is considered as the true price of that stock, and shows the stock's value after distributing dividends.

According to the literature, many algorithms and techniques have been proposed for stock price prediction, where some of them are shown in Section 3. Table 3 summarize the performance of ML algorithms and techniques (accuracy and error percentages) reported in the literature. The comparison shows that many deep learning performed well in term of producing low error percentages (such as ANN, RNN, LSTM, stacked long short-term memory (SLSTM), and bidirectional long short-term memory (BLSTM)) while the mixture of historical daily stock prices and social media data can produce the accuracy up to 70% (Attigeri et al., 2015).

6 SENTIMENT ANALYSIS

Sentiment analysis (SA), also known as opinion mining, is a study in natural language processing to analyse people's opinions, sentiments, appraisals, attitudes, evaluation, and emotions towards subjects such as organisations, products, services, issues, topics, individuals, events, and their attributes (Medhat et al., 2014). SA has played an important role in many applications, especially in finance; news may have impacts on stock markets (Nguyen et al., 2015). For example, on 16th July 2019, President Trump tweeted that the U.S. might impose a boycott on an additional $325 billion worth of imported Chinese goods. CNBC News reported the following day on their website that the stock market faced slight losses after Trump's tweets[2] about the US–China trade war. The Wall Street Journal also reported the progress of the trade deal with China in the same way. Both comments affected stocks, which closed at their lows of the day. For example, the DJIA dropped 0.42%, and S&P500 dropped 0.65%. These samples show that social media can move the stock market.

For these reasons, many researchers (Bollen et al., 2011; Li et al., 2014) have attempted to apply SA to stock prediction in order to increase prediction accuracy by using financial news (Dang and Duong, 2016; Jeong et al., 2018; Nam and Seong, 2019), and social media (Mittal and Goel, 2012; Nguyen et al., 2015; Jeong et al., 2018).

However, the financial news and social media exhibit different features. News is relatively objective due to its reliance on facts, but social data can be based on facts or rumours created by other investors; thus social data is quite subjective (Jeong et al., 2018). In stock price prediction, in addition to historical prices, research have proven that financial news and social media can improve the accuracy of stock price prediction (Zhang et al., 2018; Shah et al., 2018). SA becomes an important process used to extract facts or moods from media and many SA techniques are used for the stock price prediction in existing research, summarised below.

Bollen et al. (2011) used *OpinionFinder* (OF) tool for making binary distinctions between the positive and negative sentiment of daily Twitter feeds. They also applied a mood analysis tool to analyse the text content of tweets, which was called Google-Profile of mood states (GPOMS), that can measure human mood state across six dimensions: Calm, Alert, Sure, Vital, Kind, and Happy.

Li et al. (2014) constructed sentiment dictionaries, using the Harvard IV-4 sentiment dictionary (HVD) and the Loughran McDonald financial sentiment dictionary (LMD), to extract news data from the FINET financial news website of Hong Kong.

Nguyen et al. (2015) proposed a novel feature 'topic-sentiment', called the Joint sentiment/topic method (JST). They applied two methods to capture the topic sentiment associations, JST-based method and the Aspect-based sentiment method, to get the mood information of the stocks from Yahoo Finance Message Board.

Some studies apply information (especially qualitative data such as tweets or investors' sentiment) from the elite, the top-10 largest shareholders, or investors to predict stock prices. As a matter of fact, stocks-related sentiments provided online e.g. social media and financial news may have impact the stock prediction (Li et al., 2014).

7 CONCLUSION

Stock investments have been of interest to many investors around the world. However, making decision is a difficult and complex task as numerous factors are involved. For successful investment, investors are keen to forecast the future situation of the stock market. Even small improvements of predictive efficiency

[2] Trump's tweets swing stock market amid trade deal uncertainty – https://www.cnbc.com/2019/05/10/trumps-tweets-swing-stock-market-amid-trade-deal-uncertainty.html

Table 3: Comparison of ML algorithms and techniques in financial stock price prediction

Paper	**Prediction Techniques**	**Stocks/Index**	**Input Data**	**Accuracy (%)**	**Error (%)**
Hegazy et al. (2014)	PSO, LS-SVM, ANN	S&P 500	Historical daily stock prices	N/A	LS-SVM: 0.1147 PSO: 0.7417 ANN: 1.7212; Note: average of 13 companies which cover all stock sectors in S&P 500 stock market
Adebiyi et al. (2014)	ARIMA, ANN	Dell index	Historical daily stock prices	N/A	ARIMA: 0.608 ANN: 0.8614; Note: average of one month prediction
Nguyen et al. (2015)	SVM	AAPL, AMZN, BA, BAC, CSCO, DELL, EBAY,ETFC, GOOG, IBM, INTC, KO, MSFT, NVDA, ORCL, T, XOM, YHOO	Historical daily stock prices and mood information	54.41 (average) 60.00 (few stocks)	N/A
Patel et al. (2015)	ANN, SVM, RF, Naïve-Bayes	CNX nifty index, S&P Bombay Stock Exchange (BSE) Sensex index, Infosys Ltd., Reliance Industries	Historical daily stock prices	Naïve-Bayes: 90.19 RF: 89.98 SVM: 89.33 ANN: 86.69	N/A
Attigeri et al. (2015)	LR	Stock market price of two companies	Historical daily stock prices, news articles, and social media data (twitter)	LR: 70	N/A
Dang and Duong (2016)	SVM	VN30 Index: EIB, MSN, STB, VIC, VNM	News relating to companies in the VN30 Index	SVM: 73	N/A
Selvin et al. (2017)	LSTM, RNN, CNN, ARIMA	NIFTY-IT index (Infosys, TCS), NIFTY-Pharma index (Cipla)	Minute by minute stock prices (day stamp, time stamp, transaction id, stock price, and volume traded)	N/A	**Infosys**: CNN: 2.36/ RNN: 3.9/ LSTM: 4.18 ARIMA: 31.91 **TCS**: CNN: 8.96/ RNN: 7.65/ LSTM: 7.82 ARIMA: 21.16 **Cipla**: CNN: 3.63/ RNN: 3.83/ LSTM: 3.94 ARIMA: 36.53
Roncoroni et al. (2015)	LSTM	NIFTY 50	Historical daily stock prices	N/A	0.00859
Khare et al. (2017)	LSTM, MLP	10 unique stocks on New York Stock Exchange	Minute by minute stock prices	N/A	MLP: 0.0025 LSTM: 0.048
Althelaya et al. (2018a)	MLP, LSTM, SLSTM, BLSTM	S&P 500	Historical daily stock prices (closing price)	N/A	**Short-term**: BLSTM: 0.00947 SLSTM: 0.01248 LSTM: 0.01582 MLP: 0.03875 **Long-term**: BLSTM: 0.06055 SLSTM: 0.06637 LSTM: 0.08371 MLP: 0.09369

can be very profitable. A good prediction system will help investors make investment more accurate and more profitable by providing supportive information such as the future direction of stock prices.

This paper reviewed and compared the state-of-the-art of ML algorithms and techniques that have been used in finance, especially the stock price prediction. The number of ML algorithms and techniques have been discussed in terms of types of input, purposes, advantages and disadvantages. For stock price prediction, some of ML algorithms and techniques have been popularly selected as to their characteristics, accuracy and error acquired.

In addition to historical prices, other related information could also effect the stock prices such as politics, economic growth, financial news and the mood from social media. Many studies have proven that the sentiment (mood) analysis has a high impact to the future prices. Thus, a mix of technical and fundamental analyses can produce a high efficient prediction.

REFERENCES

Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Al-Mahasneh, A. J., Anavatti, S. G., and Garratt, M. A. (2018). Review of Applications of Generalized Regression Neural Networks in Identification and Control of Dynamic Systems. *arXiv preprint arXiv:1805.11236*.

Alfred, R. et al. (2015). A genetic-based backpropagation neural network for forecasting in time-series data. In *2015 International Conference on Science in Information Technology (ICSITech)*, pages 158–163. IEEE.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Althelaya, K. A., El-Alfy, E.-S. M., and Mohammed, S. (2018a). Evaluation of bidirectional LSTM for short-and long-term stock market prediction. In *2018 9th International Conference on Information and Communication Systems (ICICS)*, pages 151–156. IEEE.

Althelaya, K. A., El-Alfy, E.-S. M., and Mohammed, S. (2018b). Stock Market Forecast Using Multivariate Analysis with Bidirectional and Stacked (LSTM, GRU). In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–7. IEEE.

Archana, S. and Elangovan, K. (2014). Survey of classification techniques in data mining. *International Journal of Computer Science and Mobile Applications*, 2(2):65–71.

Attigeri, G. V., MM, M. P., Pai, R. M., and Nayak, A. (2015). Stock market prediction: A big data approach. In *TENCON 2015-2015 IEEE Region 10 Conference*, pages 1–5. IEEE.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Belgacem, S., Chatelain, C., and Paquet, T. (2017). Gesture sequence recognition with one shot learned CRF/HMM hybrid model. *Image and Vision Computing*, 61:12–21.

Beyaz, E., Tekiner, F., Zeng, X.-j., and Keane, J. (2018). Comparing technical and fundamental indicators in stock price forecasting. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1607–1613. IEEE.

Bodie, Z., Kane, A., and Marcus, A. J. (2013). *Investments and portfolio management*. McGraw Hill Education (India) Private Limited.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Boomija, M. and Phil, M. (2008). Comparison of partition based clustering algorithms. *Journal of Computer Applications*, 1(4):18–21.

Brown, B. (2017). *The forward market in foreign exchange: a study in market-making, arbitrage and speculation*. Routledge.

Chou, J.-S. and Nguyen, T.-K. (2018). Forward Forecast of Stock Price Using Sliding-Window Metaheuristic-Optimized Machine-Learning Regression. *IEEE Transactions on Industrial Informatics*, 14(7):3132–3142.

Chourmouziadis, K. and Chatzoglou, P. D. (2016). An intelligent short term stock trading fuzzy system for assisting investors in portfolio management. *Expert Systems with Applications*, 43:298–311.

Dang, M. and Duong, D. (2016). Improvement methods for stock market prediction using financial news articles. In *2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, pages 125–129. IEEE.

Göçken, M., Özçalıcı, M., Boru, A., and Dosdoğru, A. T. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications*, 44:320–331.

Grover, N. (2014). A study of various Fuzzy Clustering Algorithms. In 3, editor, *International Journal of Engineering Research*, volume 3, pages 177–181.

Hadavandi, E., Shavandi, H., and Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8):800–808.

He, J., Cai, L., Cheng, P., and Fan, J. (2015). Optimal investment for retail company in electricity market. *IEEE Transactions on Industrial Informatics*, 11(5):1210–1219.

Hegazy, O., Soliman, O. S., and Salam, M. A. (2014). A machine learning model for stock market prediction. *arXiv preprint arXiv:1402.7351*.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Ijegwa, A. D., Rebecca, V. O., Olusegun, F., and Isaac, O. O. (2014). A predictive stock market technical analysis using fuzzy logic. *Computer and information science*, 7(3):1.

Jadhav, S. D. and Channe, H. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, 5(1):1842–1845.

Jeong, Y., Kim, S., and Yoon, B. (2018). An Algorithm for Supporting Decision Making in Stock Investment through Opinion Mining and Machine Learning. In *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*, pages 1–10. IEEE.

Khare, K., Darekar, O., Gupta, P., and Attar, V. (2017). Short term stock price prediction using deep learning. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 482–486. IEEE.

Kim, S. and Kang, M. (2019). Financial series prediction using Attention LSTM. *arXiv preprint arXiv:1902.10877*.

Kumar, J., Goomer, R., and Singh, A. K. (2018). Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters. *Procedia Computer Science*, 125:676–682.

Lehmann, M. (2017). Financial instruments. In *Encyclopedia of Private International Law*, pages 739–747. Edward Elgar Publishing Limited.

Li, X., Xie, H., Chen, L., Wang, J., and Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Lin, F.-L., Yang, S.-Y., Marsh, T., and Chen, Y.-F. (2018). Stock and bond return relations and stock market uncertainty: evidence from wavelet analysis. *International Review of Economics & Finance*, 55:285–294.

Mann, J. and Kutz, J. N. (2016). Dynamic mode decomposition for financial trading strategies. *Quantitative Finance*, 16(11):1643–1655.

Markopoulos, A. P., Georgiopoulos, S., and Manolakos, D. E. (2016). On the use of back propagation and radial basis function neural networks in surface roughness prediction. *Journal of Industrial Engineering International*, 12(3):389–400.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mittal, A. and Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf)*, 15.

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

Moreno, J. J. M. (2011). Artificial neural networks applied to forecasting time series. *Psicothema*, 23(2):322–329.

Nam, K. and Seong, N. (2019). Financial news-based stock movement prediction using causality analysis of influence in the korean stock market. *Decision Support Systems*, 117:100–112.

Naranjo, R., Arroyo, J., and Santos, M. (2018). Fuzzy modeling of stock trading with fuzzy candlesticks. *Expert Systems with Applications*, 93:15–27.

Nava, N., Di Matteo, T., and Aste, T. (2018). Financial time series forecasting using empirical mode decomposition and support vector regression. *Risks*, 6(1):7.

Nguyen, T. H., Shirai, K., and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611.

Patel, J., Shah, S., Thakkar, P., and 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.

Perwej, Y. and Perwej, A. (2012). Prediction of the Bombay Stock Exchange (BSE) market returns using artificial neural network and genetic algorithm. *Journal of Intelligent Learning Systems and Applications*, 4(02):108.

Pradeepkumar, D. and Ravi, V. (2017). Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network. *Applied Soft Computing*, 58:35–52.

Preethi, G. and Santhi, B. (2012). STOCK MARKET FORECASTING TECHNIQUES: A SURVEY. *Journal of Theoretical & Applied Information Technology*, 46(1).

Rizvi, S. A. A., Roberts, S. J., Osborne, M. A., and Nyikosa, F. (2017). A Novel Approach to Forecasting Financial Volatility with Gaussian Process Envelopes. *arXiv preprint arXiv:1705.00891*.

Roncoroni, A., Fusai, G., and Cummins, M. (2015). *Handbook of multi-commodity markets and products: structuring, trading and risk management*. John Wiley & Sons.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., and Soman, K. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1643–1647. IEEE.

Shah, D., Isah, H., and Zulkernine, F. (2018). Predicting the Effects of News Sentiments on the Stock Market. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4705–4708. IEEE.

Sharma, A., Bhuriya, D., and Singh, U. (2017). Survey of stock market prediction using machine learning approach. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 2, pages 506–509. IEEE.

Siami-Namini, S. and Namin, A. S. (2018). Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv preprint arXiv:1803.06386*.

Singh, J. and Tripathi, P. (2017). Time Series Forecasting Using Back Propagation Neural Network with ADE Algorithm. *International Journal of Engineering and Technical Research*, 7(5).

Smid, J., Verloo, D., Barker, G., and Havelaar, A. (2010). Strengths and weaknesses of Monte Carlo simulation models and Bayesian belief networks in microbial risk assessment. *International Journal of Food Microbiology*, 139:S57–S63.

Staszkiewicz, P. and Staszkiewicz, L. (2014). *Finance: A Quantitative Introduction*. Academic Press.

Suganya, R. and Shanthi, R. (2012). Fuzzy c-means algorithm-a review. *International Journal of Scientific and Research Publications*, 2(11):1.

Wang, L., Zeng, Y., and Chen, T. (2015). Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. *Expert Systems with Applications*, 42(2):855–863.

Wang, S., Yu, L., Tang, L., and Wang, S. (2011). A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China. *Energy*, 36(11):6542–6554.

Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data mining and knowledge Discovery*, 13(3):335–364.

Whalley, J. (2016). *Developing Countries and the Global Trading System: Volume 1 Thematic Studies from a Ford Foundation Project*. Springer.

Wu, L. and Li, M. (2018). Applying the CG-logistic Regression Method to Predict the Customer Churn Problem. In *2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)*, pages 1–5. IEEE.

Xu, R. and Wunsch, D. C. (2005). Survey of clustering algorithms.

Yaffee, R. A. and McGee, M. (2000). *An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®*. Elsevier.

Zhang, J., Cui, S., Xu, Y., Li, Q., and Li, T. (2018). A novel data-driven stock price trend prediction system. *Expert Systems with Applications*, 97:60–69.

Zhao, D., Chen, J., Han, Y., Song, C., and Liu, Z. (2010). Temperature compensation of FOG scale factor based on CPSO-BPNN. In *2010 Chinese Control and Decision Conference*, pages 2898–2901. IEEE.

Zhou, J. and Fan, P. (2019). Modulation format/bit rate recognition based on principal component analysis (PCA) and artificial neural networks (ANNs). *OSA Continuum*, 2(3):923–937.

APPENDIX: MACHINE LEARNING TECHNIQUES AND DEFINITIONS

ANNs (Artificial Neural Networks): Computing systems with a set of models and algorithms that are brain-inspired, intended to imitate human learning.

BPNN (Back propagation neural network): An iterative process of tuning the weights in neural network models from errors provided by a loss function.

BSM (Black Scholes Model): A mathematical model for calculating price deviation over a period of time for financial instruments.

CART (Classification and Regression Trees): A binary decision tree model performed through a sequence of questions, the answers to which are input to the next question, and each terminal node is the result of prediction.

FCM (Fuzzy c means): A clustering technique where an object can belong to more than one cluster with different probabilities.

GAs (Genetic Algorithms): A search technique that imitates the process of natural selection to find optimal solutions.

GMDH (Group method of data handling): A type of inductive algorithm performed by a mathematical model that provides possible automatic discovery of relations in the data.

GP (Gaussian Processes): A collection of random variables that have a normal distribution.

GRNN (Generalized Regression Neural Network): A basic neural network-based function estimation algorithm used for classification, regression, and prediction.

HMM (Hidden Markov Model): A statistical model used for machine learning with hidden states, as opposed to a standard Markov chain where all states are disclosed.

Hierarchical clustering: Builds a multilevel hierarchy of clusters by creating a cluster tree.

KNN (K Nearest Neighbour): A supervised learning technique use for classification, where an object is considered as a member of a cluster with a plurality voting of its neighbours.

k-means: A clustering technique where an object can belong to only one cluster and a cluster must contain at least one object, in which centroids of clusters are means of objects in each cluster.

k-medoids (PAM): A clustering technique where an object can belong to only one cluster and a cluster must contain at least one object, in which centroids of clusters are the most central object in each cluster.

LR (Logistic Regression): A machine learning algorithm for statistical classification, where an outcome is determined by one or more independent variables.

LSTM (Long Short-Term Memory): One of the most successful RNN architectures that introduces the memory cell, a unit of computation that replaces traditional artificial neurons in the hidden layer of the network. These memory cells make networks effectively associate memories with input remote in time.

MCS (Monte Carlo Simulation): A technique used to model the probability of variant outcomes in hard to predict cases because of the interference of random variables.

MLP (Multilayer Perceptron): A type of ANN consisting of at least three layers: input layer, hidden layer, and output layer.

PSO (Particle Swarm Optimization): A technique performing optimal solution searching, inspired by the behaviour of social creatures in groups.

QRNN (Quantile Regression Neural Network): A QR-based hybrid and a feedforward neural network that can be used to estimate the nonlinear models at different quantiles.

RBF (Radial Basis Function Neural Networks): A type of ANN performed by a linear model.

RF (Random Forest): An ensemble method used to construct a prediction model of both classification and regression problems.

RNN (Recurrent neural networks): A powerful model for processing sequential data such as sound, time series data or written natural language. Some designs of RNN are used to predict the stock market.

SOM (Self organizing maps): A type of ANN using unsupervised learning in training to build a two-dimensional map of the problem space.

SVM (Support Vector Machine): A supervised machine learning technique for classification and regression analysis.

SVR (Support Vector Regression): A supervised machine learning technique for classification and regression analysis.