2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA) | 978-1-7281-9437-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/CITISIA50690.2020.9371852

Analysing Stock Market Trend Prediction using Machine & Deep Learning Models: A Comprehensive Review

Doan Yen Nhi Le[1], Angelika Maag[2], Suntharalingam Senthilananthan[3]
[1]Study Group Australia,[2]Study Group Australia, [3]Charles Sturt University, Australia.
alexandrale297@gmail.com, amaag@studygroup.com, ssenthilananthan@csu.edu.au,

***Abstract* - The applications of intelligent financial forecasting play an utmost important role in facilitating the investment decisions activities of many investors. With the right insight information, the investors can tailor their portfolio to maximise return while minimising risks. However, not every investment guarantees a good return, and this is mainly because most investors have limited information and skills to predict the stock trend. Nevertheless, the complex, chaotic and volatile nature of the stock market make any prediction attempts extremely difficult. This paper aims to provide a comprehensive review of the exiting researches which related to the application of Machine Learning and Deep Learning models in financial market forecasting domain. To prepare for this project, more than sixty research papers were analysed in-depth to extract required quantitative information, applications, and results on different methodologies. It is found from this project that Deep Learning outperformed Machine Learning in all the collected research papers, and it is the most suitable methodologies to apply to the stock market forecasting domain.**

Keywords - Machine Learning, Deep Learning, Neural Networks, Support Vector Machine, Stock Market Prediction, Technical Analytic, Trend Forecasting

I. INTRODUCTION

The stock market is widely known for its chaotic, non-stationary and non-linear nature [1]. There have been 24 stock market crashes since 1637, the most significant impact in the history of all crashes is the one in 2007, 2008. The need for forecasting the stock price trend, therefore, becoming vital. Investors require to have the ability to analyse or possess an abundance of insight information on the market trend to tailor their static for monetary mechanisms and maximise the profit [2]. The difficulties in forecasting are due to the unstable and noisy sample due to various unpredictable factors [3]. Countless research papers were proposed to improve the accuracy of the stock prediction models [4]. Methods are varying from fundamental, sentiment analysis to technical analysis [5]. Machine learning methods like Support Vector Machine (SVM) [5], [6] Auto-Regressive Moving Average (ARMA), Least Square (LS) [7], [8] and Deep Learning methods like Convolutional Neural networks (CNN), Multi-Layer Networks [9], Recurrent Neural Networks (RNN), Long-short Terms Memory (LSTM) networks and Artificial Neural Networks [8] are often applied in stock movements prediction [10], [11], [12]. Furthermore, many researchers like Long [4]have proposed several hybrid models to boost accuracy sharply like the Deep Stock Trend Prediction Neural networks (DSPNN) and ARIMA-LS-SVM [7]. This paper aims to provide a review on stock trend forecasting methodologies and show which method is the best to put in practice. This research will not cover the analysis of Fundamental or Statistical Analysis. It will only focus on Technical Analysis, which involves Quantitative Analysis and Behavioural Recognition in financial time series forecasting [13]. After carefully study more than sixty research papers, the following findings were found, firstly, the best method to handle non-linear data is Deep Learning method. These methods are best to discover structures that are intricate in high-dimensional data as according to Orimoloye [11]. Secondly, the SVM in Machine Learning, underperformed in comparing to Deep Learning Neural networks, however, according to Wang [14] SVM is used to deal with non-linear data by transferring the data to the hyperplane and SVM has higher accuracy in comparison to other Machine Learning methods. Lastly, the result when combining multiple models into one outperform the result of a single model [7], [4], [15]

II. LITERATURE REVIEW

Various researchers have attempted to forecast the stock market movements using a pool of sophisticate means. The investors need to predict the future trend of the financial market to minimise risks and maximise the returns on the investment assets [4]. This paper focuses on analysing different machine learning and deep learning models that assist in forecasting the trend of the stock market. Finding from more than sixty papers is sorted and put under analysis. The first section of this paper will discuss the background, the benefit, and limitation of Machine Learning and Deep Learning. The second part of this paper will elaborate on the Data used and the pre-processing of Data. The third section will introduce to the reader how the model implementation is carried out. The verification section is section fourth section and Evaluation Metrics and Result Analysis will be provided in the fifth sections.

A. Machine Learning Prediction Methodologies

1) The common methods

In the Machine Learning arena, the best performer is the Support Vector Machine. As stated Tang [13] SVM methods are based on structural risk minimisation and performs better than traditional statistical prediction methods like Autoregressive Integrated Moving Average (ARIMA), Random Forest (RF) and Linear Regression (LR) and can handle high dimensional data. In this paper, we will focus more on SVM models [16].

SVM model is supervised learning which uses associate algorithms to study data and is used for regression analysis and classification [1]. A study by Yang [6] also shows that SVM can solve the problem of linear constraint quadratic

programming, by applying Kernel function. Moreover, Yuan [8] stated that SVM can solve the indivisibility problems of data, map the dataset to a high-dimensional space and become linear separable and reduce the data complexity (Fig. 1). SVM minimises overfitting and is easy to modify and results are high in accuracy [17] similarly, the direction of NIKKEI 225 index were successfully predicted using SVM in Yuan [8] study. Likewise, Xiao [7] also utilised SVM to predict the long terms stock price movement of the Chinese A-Share market while utilising SVM in their research papers. Long terms dependencies of financial assets' return are collected successfully by Maqsood [10] to optimal investor's stock portfolio. In another study, Li [5] incorporates SVM models with news sentiment to forecast the future movement of stock price and stocks' sell/buy points by Tang [13]. A study by Yang [18] suggest that the ARMA model is introduced to handle high-frequency data due to its ability to deal with the volatility of the stock market. There are many papers that have attempted to combine different machine learning methods, Xiao [7] proposed a model that combined ARMA with Least Square and Support Vector Machine to implement in stock forecasting. According to Zhang [19] and Ramezanian [9] ARMA is an econometric model which can also be used to assist with the decision-making process in situation where a lack exists in technical pattern identification.

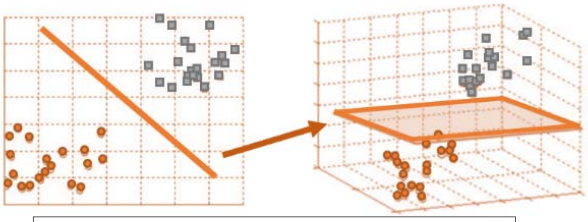

Fig 1. SVM mapping data to higher dimensional space

2) The Gap and Limitations

Computing SVM models is time-consuming and incur computing cost, and this is due to its complexity and its dependence on the kernel function [17]. Similarly, research by Xiao [7] shows similar experiences with the ineffectiveness of the kernel function. Moreover, Xiao [7] suggest that SVM failed to maximise the training period and fail in enhancing the sentiment analysis algorithms. The models failed to consider factors that are used for testing the prediction model, this is including macro and micro factors. Additionally, implementing SVM methodologies can be very tedious due to the number of algorithms involved. Lastly, SVM models based on a small amount of training and testing sample, it is not suitable for financial time-series forecasting on a large scale [20]. With the effectiveness and limitation given for Machine Learning, further exploration in Deep Learning methods is required for better comparison.

B. Deep Learning Prediction Methodologies

It has been found that Deep Learning methodologies outperform classical computation intelligent model like SVM According to Niu [20], by focusing on developing deep learning method, the above-mentioned research gap of Machine Learning can be filled. Deep Learning alibility is most suitable for the time series data in comparison to the machine learning models, this is due to the memory ability [1]. This method is different as its networks structure can be customised for specific data formation and objective tasks [3]. A series of unit can be provided by using Deep Learning, such as the LSTM unit, recurrent unit, convolutional unit and gate recurrent unit [1]. According to Long [3], the purpose of these units is for features extraction for samples that have different characteristic. The more the complexity and volatility of the market nature, the more layers of networks need to be incorporated to boost the accuracy rate [2]. The architecture of the Neural Networks comprises of 3 types of layers which are the input layers, the hidden layers, and the output.

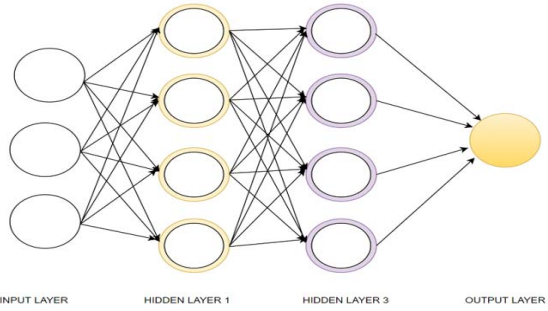

Fig 2. Structure of Neural Networks Based

1) Common Methods

The methods that were used among the best papers are as follow:

Long-Short Terms Memories Networks (LSTM): Niu [20] stated in their paper that the RNN is the best method to use for the time-series data due to its ability to capture long and short terms dependencies in time-series. According to Lei [21] this is a policy-based algorithm trading model. LSTM is a special type of RNN, Wang [14] introduce the concept of LSTM networks to learn the sequential pattern. LSTM is incorporated into the Recursive Neural Networks' hidden layers due to its ability to memorise the data. This is a special Neural Networks which can store and access a bigger range of sequential contextual input information and perform very well at handling vanishing gradient issues. The cell is comprised of input gate, forget gate and output gate, when the input data enter the LSTM networks, the cell structure determined which information can stay and which can be forgotten in the forget gate according to specific rules [20] (Fig. 2).

Deep Neural Networks (DNN): DNN has simple architectures, better to handle time-series data in a chaotic environment, better in predicting the trend even in the stock market crash situation. The study by Ben and Gbenga [22] using (DNN) to predict the stock market intraday price shows that the model and the actual market movement pattern in the market crash in 2007, 2008 is consistent with each other. Therefore, showing that deep learning approaches can be utilised to divert risks, as stated in Moews [23]. Nevertheless, Orimoloye [11] used DNN to predict the price movement in multiple markets, their result shows that DNN also can learn the underlying structure particularly well when given a large dataset and the prediction accuracy exceed sharply than the Machine Learning models and the methods have feature extraction function that can be tailor to fit the data samples [3]

Hybrid Methods: Many studies show that the combination of different methods can help the models perform better than applying one single method. Deep Stock

Trend Prediction Neural Networks (DSPNN) is a combined method developed by [3]. This method combined the Market Information extraction, investor clustering with the DNN along with LSTM networks models and the result shows that the five-day and seven-day stock prices are predicted effectively with high accuracy. A study by Li [5] which incorporate the LSTM Networks with sentiment analysis which extracts data from tops news to predict the stock future trend, their study found that the LSTM Networks significantly outperformed the machine learning models.

2) The Gaps and Limitations

Research conduct by Long [3] suggests that Deep Learning's integration of filters is complex and if done incorrectly will greatly affect the quality and characteristic of the signals and result in the study unable to minimise risks Long [3]. Lei [21] research paper stated a disadvantage of RNN, of which it tends to input all environmental features without perception and presentation. This issue with RNN can be fixed by using DNN but DNN cannot perform feature selection and capture long-term dependencies [21]. Nevertheless, most of the research paper has not fully utilised the ability of the Neural Networks due to limited in the number of hidden layers, the layers should be increased as this will increase the refinement and lead to higher accuracy percentages [8]. Except for the study conduct on combination models, remaining research papers only focus on individual models of which accuracy is not as great as the combination of method, various other external factors can affect the prediction, such as government policies and interest rate and events that have impacted on the financial market. Orimoloye [11] study show that DNN did not perform well with small data size. In light of the mentioned Benefits and Limitation of Deep Learning methodologies, these project research papers are provided with means to compare between Deep Learning methodologies and Machine Learning methodologies.

C. Research Classifications

The current study is conducted using more than sixty papers that are relevant to the field of stock forecasting from the year 2018, 2019 and 2020. A process has been put in to sort out the most related with most up to date technologies and methods and below Table II shows the works of fifteen papers with its state-of-art methodologies which proved to sharply increase the accuracy. This table is an elaboration of table one, this table classified in details the data, methods and evaluation metrics for each paper. The classification categories in Table 2 comprise of the Data Source, Name, Attributes and Coverage along with different algorithms, techniques and tools used to perform and test the state-of-art models. The 15 papers are also classified based on Evaluation Metrics and Primary Output. Most research papers suggest the implementation of an individual model, while some suggest using a combination of models to increase the accuracy or reduce the error rate.

D. Models Implementation Diagram

This section focusses on analysing the prediction process of both Machine Learning and Deep Learning models. The raw data is taken from different Data Source, as stated in Table II and III, these are non-processed, non-linear, and chaotic datasets. As per Fig. 3, the raw data is fed into the Input pre-processing Stage to reduce the noise and artifacts as well as sorting out relevant datasets that are needed for the models. At this state, different denoising method is put to work, such as Wavelet Transform or ARIMA. After data is pre-processed, the output from the Pre-Processing Stage is fed into the next stage where the ML/DL Model Implementation Stage where the Machine Learning and Deep Learning models are implemented to extract the pattern of the stock price, Table III have further elaborated the tools, algorithm and techniques used in different studies. The Evaluation Metrics then applied to the output from the models in Evaluation Stage to evaluate if the model meets the accuracy or error threshold. The highest accuracy models will be used in practical applications where different charts will be constructed based on the output of the models to show visually the trend of the future stock prices intra-day, 3 months to 6 months from the chosen point of time.

III. Research Evaluation

In the process of determining which methodology is better in stock market forecasting domain, different models' components are collected and are subjected to validation and evaluation. *Validation* deal with the accuracy of the output, how small the error gap between the output and the actual target stock's trend. Whereas, *evaluation* look into the practicality of the models in the real-life application on the actual stock market. The below table has analysed the 15 chosen papers in details based on the prediction results, algorithms used, data types and data selection methods Table III analysed in-depth the result of each research papers according to mentioned evaluation metrics in Table I, tools and algorithm names are also included for each paper to show the methods that were used to generate the result. The key columns in Table III is the last two, which show details information on the accuracy and performance of the model in a practical environment. This means that how close the model performance is in comparing to the actual movement of the stock price. The expected result is that the performance will closely mimic the price trend of up-trend and down-trend. Regard to the accuracy, the column has been split into Machine Learning side and Deep Learning side, from these two columns plus the additional information from performance column, comparison can be drawn out of which methodologies is most suitable to apply in the forecasting of the stock market. Table IV focus on analysing the dataset and how data is collected. This focus on the coverage of data, name of different data classifier as well as the training and testing dataset and its selection indicators.

IV. Research Verification

To verify the proposed methodology that is best to use in the financial market forecasting, fifteen research papers were collected to further analyse and compare in relation to stock market forecasting domain. As Fig. 4 shows, all fifteen papers discussed data attributes, such as Data Source, name of chosen Dataset, Data Name and the Coverage of Data. In terms of algorithm and methods, 73% of papers discussed different methods to produce the stock prediction result. It is necessary to keep in mind that some attributes are subjected to further elaboration, such as Techniques, Algorithm, Tools and Performance Accuracy, for example, there are two types of techniques that are relevant to this research project that is Machine Learning and Deep Learning, which both included within the 73%.

TABLE I. CLASSIFICATION TABLE (DL: DEEP LEARNING, ML: MACHINE LEARNING)

Ref	Data				Data analysis			Evaluation	Output
	Source	*Dataset Name*	*Attributes*	*Coverage*	*Algorithm*	*Tech*	*Tools*	*Accuracy metrics*	*Primary output*
[4]	Chinese Share A-Market	CITIC Securities GF Securities China Pingan	BP, CP, HP, LP, , TV, RoC,	7 years of stock data	IC, MII, CNN, BILSTM, DSPNN, KG and GE	DL	Python programming language, TensorFlow	Accuracy %	Accuracy % highest for DSPNN
[8]	Wind-Economic database	Chinese A share the dataset	60 stock features are obtained	8 years, 01/01/2010 to 01/01/2018	SVM, RF, ANN, NB, sliding window	ML	MATLAB2016a, LS-SVM lab, Libsvm toolket	Return on Investment Accuracy %	Increase in accuracy
[7]	Chinese A-share market data	367 sets of data	Today's HP, LP, OP, CP, TV	Within the mark of 5 minutes - 102 trading day within 2015 – data collected every single day.	SVM, ARI-MA, ARI-MA-LS-SVM	ML	MATLAB software	MSE, RMSE	High-frequency data is recommended
[24]	UK Stock Exchange 100 Index	This sample chose 100 stock from FTSE 100	Historical series of adjusted: opening price, CP, HP, TV	March 1994 until March 2019, covering 25 years.	Linear Regression SVM LSTM	ML, DL	N/A	RMSE, MAE	Accuracy increase
[12]	Yahoo! Finance S&P 500, DJIA, HSI	Data set of S&P 500, DJIA and HSI	Date, OP, HP, LP, CP, Close Adjust, TV.	S&P 500 and DJIA datasets: 19 years - HIS dataset:17 years	LSTM	ML	Python programming on Tensor Flow	SNR, RMSE, MAE, R2, GRU	Increase in accuracy % Low Error Rate
[5]	HKEx HIS FINET	CKH, SHK PPT, HSBC Holdings	OP, CP, TV	Daily prices from January 2003 to March 2008.	LSTM, SVM, SD, NP and SA, EasyMKL, MKL	ML	Opinion Finder lexicon, Google-Profile of Mood States	Accuracy, Weighted Macro F1-score	High Accuracy Low Error Rate LSTM
[11]	Tick-Write Data Inc	IBEX 35, ISE-100, CAC 40, OSE, ATX, AEX, BEL20	daily , hourly, minutes and tick level	01/02/2008 – 19/02/2014 Exception for Brazilian market: data only for 4 years, 01/02/2010 to 19/02/2014	DNN, A rectifier linear, relative predictive accuracy, DNN, SVM, NN	ML, DL	ANOVA, CIFAR-	Accuracy	Two DNNs are more accurate
[21]	Yahoo! Finance	CTL, HRB, SLB, COF, JCI, APD, NUE	OP, CP, HP, LP and the TV	The range of data is in daily trading	DRL	ML	Technical indicator calculated with talib	Profit curve, Profit, AR, SR, TT	The TFJ-DRL model obtain the highest
[3]	CSI 300	Data from CSI	OP, CP, HP, LP, TV	Training and testing 09/12/2013 to 07/12/2013	MFNN	DL	GoogLeNet, AlexNet	Accuracy %, Profitability %	Deep Learning show high accuracy rate
[9]	Tehran Stock Exchange website	Alborz, Dekosar, Sharak, Vama'aden	FP, MP, MinP, TV	29/50 companies with high trading volumes of at least 200 trading days is selected.	GNP, MLP, Time series models	ML	No tools were mentioned	MAPE,RMSE, MSE, MDA	Improve prediction accuracy
[13]	Shanghai Stock Exchange	Stock Codes: 600736, 600197, 600211, 600694	Open, High, Low, Close	Data are collected from 04/01/2010 to 18/08/2011	SVM, WSVM, PLR	ML	Scikit-learn	Revenue CV	The proposed model outperforms other models
[2]	From different world financial exchange	DJI, FCHI, GDAXI, GSPC, GSPTSE	OP, OP, LP, CP	The timeframe 01/04/2013 to 29/12/2018	MLR, SVR, FFBP	ML	No tools were mentioned	HE – information efficiency	MLR with HE+SE+RE algorithm the result yields the highest in accuracy rates
[22]	S&P 500, Thomson Database	Not specified	Transaction date Transaction time	2011 to Apr 2016, 5 years 1996 to Apr 2016, 20 years.	LR, FANN	ML	AZFinText	Accuracy %	Improvement in accuracy in stock trend prediction
[17]	MSCI, Thompson Database	DAX 30, FTSE 100, BOVESPA, KOREA SE	Days, Annualised Average return, volatility	22 years from Jan 1995 to Dec 2016	LR, RF, SVM	ML	Did not specified which tools were used	Accuracy %,	Higher returns are achieved overall with the SVM model
[1]	N/A Literature Review	N/A Literature Review	N/A Literature Review	N/A Literature Review	NN, GA, Time series model	DL, ML	N/A Literature Review	MAPE, accuracy, RMSE, MSE, MAE,	Deep Learning show better prediction than Machine Learning

Note: Abbreviations refer to Appendix 1.

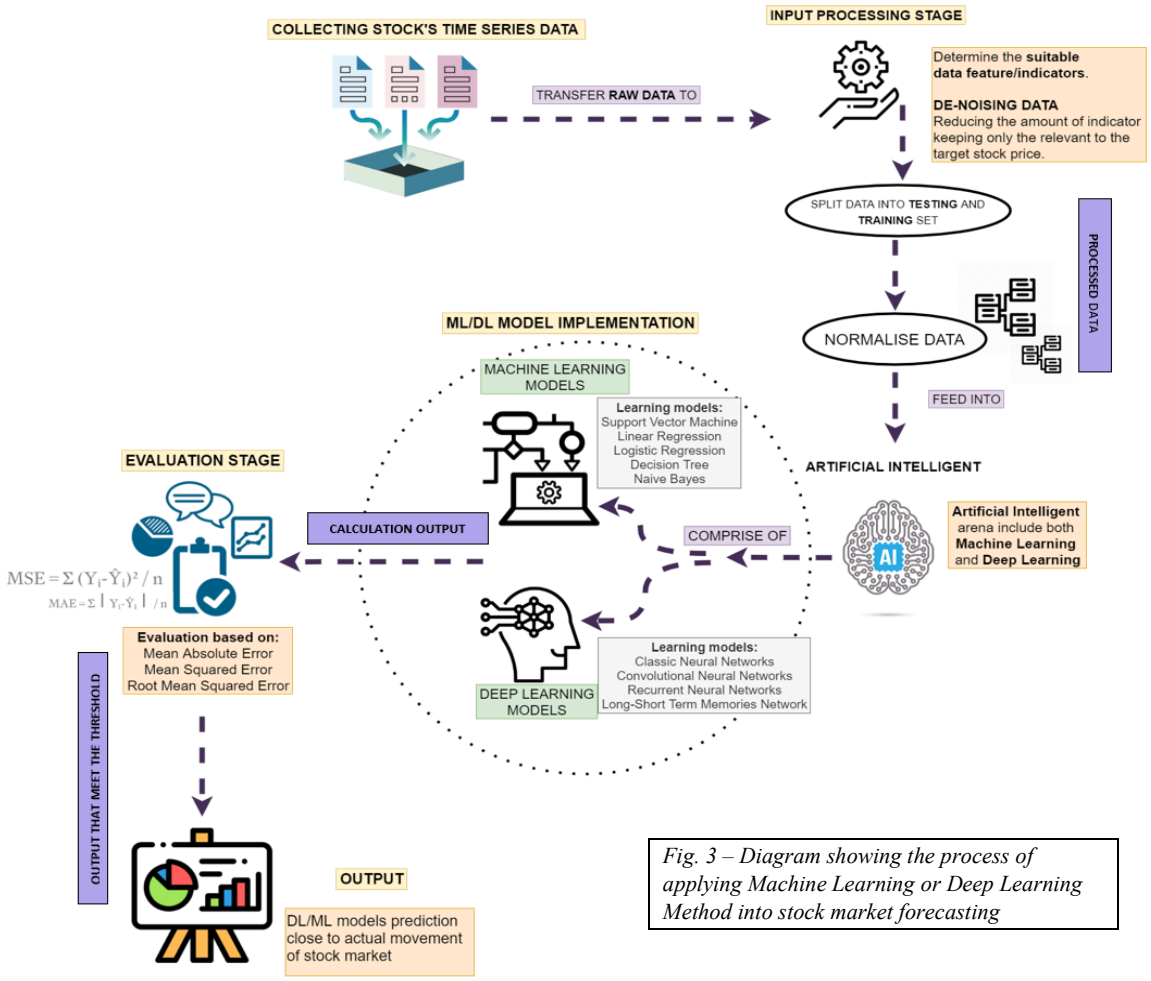

Fig. 3 – Diagram showing the process of applying Machine Learning or Deep Learning Method into stock market forecasting

Primary outputs are given in most paper except for the practical papers. Lastly, only certain papers mentioned the tools used in data pre-processing and method implementation while others do not. Most papers put the output data into graph and chart that mimic the stock market trend while others only discuss the output.

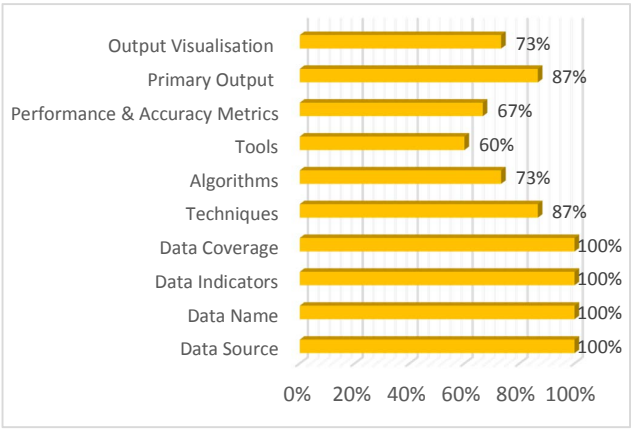

Fig. 1 The Percentage of Components or Classes Which were Described in The Selected Publications

The below Table II show the frequencies of different terms used in 30 papers that are relevant to the research domain. The frequency of the time the word repeats show how relevant the paper is to the domain topic. As can be seen in the below tables, the words related to the domain and the techniques used repeatedly and up to one thousand to three thousand times within thirty research papers. This indicates that the papers collected for this study are highly related to the domain as well as the proposed methods and techniques.

TABLE II. TERM FREQUENCY IN 30 RESEARCH PAPERS

Terms	Frequency	Terms	Frequency	Terms	Frequency
Stock	3171	Machine	523	Decision Making	87
Data	1832	Training	488	Historical	85
Market	1717	Vector	475	Activation	84
Neural	845	Networks	451	Time-series	82
Forecasting	797	Indicators	447	ReLu	79
Financial	789	Features	409	Share	64
Trading	778	Random	297	Algorithmic	47
Accuracy	627	Error	259	Movements	45
Deep	595	Sample	195	Non-linear	40
Trend	576	Testing	146	Pre-processing	7

TABLE III. ANALYSING STOCK MARKET PREDICTION RESULTS

Ref	Data source	Tools	Algorithm	Evaluation metrics	Accuracy results		How suitable data is collected
					Machine learning result	*Deep learning result*	
[4]	Chinese Share A-Market Target Stock: CITIC Securities	Python programming language, TensorFlow	CNN, BILSTM, DSPNN, AdaBoost, RF	Accuracy %	CITIC Stock RF 48.87% AdaBoost 49.45%	DSPNN - CITIC Stock 73.59% GF Securities 70.76% China Pingan 72.82% LSTM - CITIC Stock 70.59%	The relevant to stocks to the target stock are selected → method: KG, GE Techniques
[8]	Wind-Economic database Of Chinese A-Share market	MATLAB2016a, LS-SVM lab, Libsvm toolkit	SVM, RF, ANN, NB,	Accuracy % *none = none feature selection	SVM - *None* 51.7287%/ *RFE* 51.7710%/ *RF* 51.82% RF - *None* 52.9331%/ *RFE* 52.7187%/ *RF* 52.7804%	ANN – *None* 52.3216%/ *RFE* 52.3497%/ *RF* 52.3204%	Feature Selection Methods: 1. RFE 2. RF
[7]	Chinese A-share market data	MATLAB software, LS-SVM and Libsvm tool kits	SVM, ARI-MA, ARI-MA-LS-SVM	MSE, RMSE	MSE - *ARIMA-LS-SVM*: 0.015/ *LS-SVM*: 0.0174/*RS-SVM*: 0.0342/ RMSE - *ARIMA-LS-SVM*: 0.108/*LS-SVM*: 0.0234/*RS-SVM*: 0.0342	N/A	Various indicator of Autodesk (002227) is selected. Based on the initial metrics. All data used must be dispersed
[24]	UK Stock Exchange 100 Index (FTSE 100)	N/A	Linear Regression SVM LSTM	MSE, RMSE, MAPE, MAE, R^2	SVM - MSE 0.0047/ RMSE 0.0649/ MAPE 106.93/ MAE 0.0361/ R^2 0.1958 ARIMA – MSE 25.49/RMSE 3.817/ MAPE 2560/ MAE 3.1913/ R^2 0.0699	LSTM - MSE 0.0033/RMSE 0.0543/MAPE 91.44/MAE 0.0303/R^2 0.2621 DNN - MSE 0.008/RMSE 0.0852/MAPE 161.58/MAE 0.0431/R^2 0.1886 RAF - MSE 0.64/RMSE 0.0723/MAPE 137.53/MAE 0.0432/R^2 0.1518	Feature selection using the RFE + logistic regression algorithm Select the feature with ratio > 0.3, 20 important indicator then selected
[12]	Yahoo! Finance S&P 500, DJIA, HSI	Python programming on Tensor Flow	LSTM	SNR, RMSE, MAE, R2, GRU	N/A	WLSM + Attention model: MSE0.2337/MAE0.1971/RMSE0.3429/$R^2$0.8783	Wavelet Transform: Coif3 wavelet function
[5]	HKEx, HIS, FINET - Assessed Stock: 0005.HK	OFL, Google-Profile, SenticNet 5	LSTM, SVM	Accuracy, Weighted F1-score	SVM - Acc%: 40%/ F1: 0.381 MKL -Acc%:61.7%/F1: 0.550	LSTM - Acc%60.1%/F10.579	Cross-validation is conducted on the training set to select optimal hyperparameters for models.
[11]	Tick-Write Data Inc	ANOVA, CIFAR-	DNN, RNN SVM, NN	Mean	SVM - Daily Data 0.578/Hour Data 0.557/Minute Data 0.558/Tick Data 0.585	DNN (ReLU) - Daily Data 0.489/Hour Data 0.549/Minute Data 0.645/Tick Data 0.607 NN - Daily Data 0.522/Hour Data 0.514/Minute Data 0.547/Tick Data 0.573 RNN -Daily Data 0.555/Hour Data 0.585/Minute Data 0.552/Tick Data 0.591	Data is select in day ranges with same period for as much market possible
[21]	Yahoo! Finance	talib	TFJ-DRL	AR%, SR, TT	N/A	AR% 20.92%/ SR 1.09/TT 609	No specific requirement
[3]	CSI 300	GoogLeNet, AlexNet	MFNN	Accuracy %	SVM41.38%/26.21 41.04%/0.75 - 40.83%/1254 times -3.94%	MFNN 47.27%/RNN 44.85%/CNN 46.93%/LSTM 44.45%	Multi-filter was used to stacking convolutional and recurrent filter
[9]	Tehran Stock Exchange website	No tools were mentioned	GNP-ARMA, MLP	MSE, MAE	N/A	N/A	Rules is extraction methodologies
[13]	Shanghai Stock Exchange in China	Scikit-learn	SVM, WSVM, PLR	Coefficient of Variation (CV)	PLR-WSVM 4.78/PLR-WSVM 0.95/IPLR-WSVM 2.63/PLR-ANN 1.81/BHS 2.24	N/A	Used KDJ method to get the indicators
[2]	Source of data is from different world financial exchange website	No tools mentioned	MLR, SVR, FFBP	HE efficiency	HE+SE+RE indices yield higher in accuracy in compare to the others, rate in MLR, SVR, FFBP are all high.	N/A	Feature selection based on max-min redundancy
[22]	S&P 500	AZFinText	LR, FANN	Accuracy %	ML models (Baseline) 50%	One day trading: 56.02% One-hour gradient interval: 53.95%	Using linear regression to build feature to manual selecting feature.
[17]	MSCI, Thompson Reuters,	Did not specified	LR, RF, SVM	Accuracy %	LR 65.80%/RS 66.20%/SVM 71.5%	N/A	Stock selected base on the MSCI Pearson correlation and VAR model
[1]	N/A Literature Review	N/A Literature Review	Soft Computing Algorithm, NN, Back Propagation, GA, Time series model		Literatures Review: MAPE, accuracy %, RMSE, MSE, MAE, MPE, MRAE, ARV	Literatures Review	N/A Literature Review

Note: Abbreviations refer to Appendix I.

TABLE IV. DATASET ANALYSIS

Ref	Dataset used	Data coverage	Classifier		Testing/training data		Selection indicators
			Available	*Name*	*Testing Data*	*Trading Data*	
[4]	CITIC Securities GF Securities China Pingan	7 years of stock data	✓	Direction classification Trend classification	Not clearly defined	Not clearly defined	BP, CP, HP, LP, daily change, TV, turnover, , RoC, amplitude
[8]	Chinese A share the dataset	8 years, 01/01/2010 to 01/01/2018	✓	Recursive feature elimination (SVM-RFE), Random Forest	Jan 2011 to Feb 2011	Jan 2010 to Jan 2011	60 features in below factors: VF, GF, FQF, LF, SF, MoM, VoF, LF, TF
[7]	367 sets of data	102 trading day of 2015	✓	RBF Kernel function	100 sets of data	144 sets of data	Today's HP, LP, OP, CP, TV 244 sets of data selected: each set subject to 20 sets
[24]	This sample chose 100 stock from FTSE 100	March 1994 until March 2019 - 25 years.	**X**	N/A	22 overlapping training and testing datasets	22 overlapping training and testing datasets	Historical series of adjusted: OP, CP, HP, TV, SR, RSI, MoM, TR, ATR, PSAR
[12]	Data set of S&P 500, DJIA and HSI	19 years & 17 years	✓	Attention Mechanism is used on the RNN to classify image	17/05/2019 to 01/07/2019 and 02/01/2002 to 16/05/2019	03/01/2000 to 16/05/2019 and 17/05/2019 – 01/07/2019	Date, OP, HP, LP, CP, Close Adjust, TV.
[5]	CKH, SHK PPT, HK & China Gas, HSBC Holdings	January 2003 to March 2008.	✓	Cross Validation	March 2007 to March 2008	March 2003 to March 2007 LSTM March 2007 to September 2007 and September 2007 to March 2008	OP, CP, TV
[11]	IBEX 35, ISE-100, CAC 40, OSE, ATX, AEX, BEL20	01/02/2008 – 19/02/2014	✓	SVN classify object into two classes and scalar product to define linear classifier	Not mentioned	First 50% (750) of trading days to train the forecasting models with alternative meta-parameter setting and assess their accuracy on the following 375 trading days	daily , hourly, minutes and tick level
[21]	CTL, HRB, SLB, COF, JCI, APD, NUE	The range of data is in daily trading	**X**	N/A	Large scale dataset S&P 500 and Small-scale dataset containing Last 3000 days	Mini batch training with batch size is 32 Large scale dataset S&P 500 and Small-scale dataset containing, first 2000 days	OP, CP, HP, LP and the TV
[3]	Data from CSI	Training and testing 09/12/2013 to 07/12/2013	✓	Stacking convolutional filters method	Market data of CSI 300 in 1-min frequency from December 9th, 2013 to December 7th, 2016 are used for training and testing. The proportion is 7:3 7 for training and 3 for testing		OP, CP, HP, LP, TV
[9]	Alborz, Dekosar, Sharak, Vama'aden	200 trading days is selected. Two main classes, training: 250 days, testing: 100 days	✓	RL + MLP to classify data SARSA RLearning data classification, extract rules in each sub period	After training come the testing of the network 100 testing data is used and compute output for each of 5 classes.	250 training data NeurophStudio software is used to implement and construct Training process and to determine the weights of neural network.	Final price Maximum price Minimum price Trading volume
[13]	Stock Codes: 600736, 600197, 600211, 600694	Data are collected from 04/01/2010 to 18/08/2011	✓	Relative Strength Index	Did not clearly mentioned		Open, High, Low, Close
[2]	DJI, FCHI, GDAXI, GSPC, GSPTSE	The timeframe 01/04/2013 to 29/12/2018	✓	mRmR feature selection	HE + SE + RE test dataset (1251 × 150)	HE + SE + RE training dataset (11161 × 150)	OP, OP, LP, CP
[22]	Not specified	5 years and 20 years.	✓	Linear regression, Cross Validation	Deep Learning Model 25% use for later validation	Deep Learning Model 80% of dataset for the respective fold	Transaction time
[17]	DAX 30, FTSE 100, BOVESPA, KOREA SE	22 years from Jan 1995 to Dec 2016	✓	LR, SVM, RF techniques	N/A	1995 to 2004	Days, Annualised Average return, volatility
[1]	N/A Literature Review	N/A Literature Review	**X**	N/A Literature Review	N/A Literature Review	N/A Literature Review	N/A Literature Review

Note: Abbreviations refer to Appendix 1

V. DISCUSSION

A. Data

1) Data Collection

From Table 4, the data used for the research papers that are collected to conduct this project is historical stock data extracted from various stock market around the globe. The data cover a different time frame, from the shortest time frame of 102 trading day [7] to the longest time frame of 19 years on the S&P 500 and DIJA stock market [12]. Many indicators and features are considered during the extraction of data, this is during Data Pre-Processing Stage (Fig. 3), for SVM model in Yuan [8] the features are selected, and rank based on the important in descending orders and only the highest 80% of features are selected. On the other hand, 14/15 paper only chose the indicators as Opening, Closing, Highest, Lowest Price and Volume.

2) Data Pre-Processing

Stock Data is extremely noisy and chaotic [25], due to this reason, different Data Pre-processing Techniques have applied to denoise the input dataset. At Data Pre-Processing Stage (Fig. 3), Wavelet Transform techniques were used by Qiu [12] to utilise its ability to decompose and reconstruct the time-series data to help denoising their own method of LSTMAttention-based models. Knowledge Graph Embedded Techniques with node2vec algorithm was incorporate with the DSPNN model in Long [3] to assist in extracting the relevant stock to use as input which have result in very high accuracy as data has been processes carefully. Meanwhile, Machine Learning methods like SVM Lee [17] using Pearson correlation and VAR model, to find the linkage between countries and ARIMA Patel [26] and Least Square are also used to reduce the number of attributes and train the sample data to be fed into the next Method Implementation Stage [7]

B. Model Implementation And Results Analysis

1) Machine Learning Implementation

As can be seen in Fig. 3, raw data is collected from Data Source then fed into the Data Pre-Processing State to reduce the amount of attribute, methods like ARIMA and Least Square are normally used to perform these tasks. A kernel function then selected to map the data to infinite-dimensional space [7], [8]. The most popular kernel that was applied in different research papers is the Linear Functions, Polynomial Functions, used by Lei [21] and Gaussian Kernel Function. According to Yuan [8], the kernel function is used to calculate the product of two vectors in low dimensional space and map the low dimensional space to high dimensional space and reducing the complexity of the data at the same time. In the Model Implementation Stage, the parameter optimization then applied, the best parameter will be used for training [7]. The output data then fed into the Evaluation Stage to assess the error rate.

2) Deep Learning Implementation

Deep Learning Neural Networks Structure has 3 types of layers, which are: input layer, hidden layers, and output layer. The data is pre-processed and fed into the Model Implementation Stage this is when different Deep Learning methodologies are applied. In brief, the pre-processed data is fed into a neural networks structure. Firstly, the data will go through the input layer, the hidden layers then perform necessary algorithms and the final output went through the output layer. After the data passed through the output layer, Error backpropagation process is performed, according to Yuan [8], this algorithm is to update the weight on each node and reduce the error. This process can be achieved using the Gradient Descent Method [8]. The data went through the output gate and enter the Evaluation Stage and being tested to see whether the data meet the threshold.

3) Results Analysis

It can be seen from the chosen research paper that Deep Learning performs better than Machine Learning when it comes to stock prediction domain. Long [4]'s study on Chinese Stock Market shows that DSPNN, which is a hybrid Deep Learning model performed significantly better than RF and AdaBoost Machine Learning models. The DSPNN model achieved the accuracy of 73.59% this is the highest accuracy rate among Deep Learning models in this research project. The pre-processing data stage (Fig. 3) was done very carefully where the investor behaviour is extracted using clustering techniques, and sentimental analysis is also applied in the paper to extract market information. Nevertheless, the author combined both DNN and Bi-LSTM in one model which boost the prediction process and hike up the accuracy rate. In the same paper, the LSTM models is the second-highest in accuracy rate, 70.59%. On the other hand, the highest accuracy rate of Machine Learning models is SVM models, with the highest number come from Lee [17] study, 71.50%. The author collected stock index from 10 different countries to construct two global volatility networks indicator data processing techniques. This result achieved through implementing a different method from the traditional data processing techniques of other paper. However, this rate is still lower than the aforementioned, Deep Learning DSPNN's rate

Nevertheless, in the study by Yuan [8] on the Chinese A-Share Market pointed out that the Deep Learning model achieved a higher accuracy rate of 52.3204% while SVM Machine Learning was only 51.82%. Likewise, in another study on CSI 300 stock market by Long [3] used MFNN model and the accuracy rate was 47.27% while SVM model was only 41.38%. Similarly, Li [5] study on the Hong Kong stock exchange shows the same pattern, where LSTM Deep Learning model reached 60.1% while Machine Learning model, SVM, only achieved 40%.The issue cause by the sentiment analysis in the pre-processing data state which was not optimised.

When it comes to error rate, the lowest error rate is the application of SVM on UK Stock Exchange by Wang [14] of only 0.0047 as measure using MSE. However, this rate is still higher than the LSTM Deep Learning model which was only 0.0033.

C. Limitations

The limitation of this study is that, firstly, most papers did not consider the risk aspect of the investment and only focus on profit. Secondly, even though the Deep Learning models outperformed conventional Machine Learning model like SVM, in certain research the performance of Deep Learning only slightly higher than Machine Learning. Thirdly, among

different research papers, 80% only look at one market to apply their models and did not consider the application on an international scale, hence, the set up and testing and training environment or even the pre-processing of data only focus on one certain market environment. Lastly, certain paper collected in this research project only focus on one Methodologies of either Machine Learning and Deep Learning, not both in the same market environment and dataset.

VI. CONCLUSION

This paper study the most attractive topic in the financial market, which is the forecasting of the stock price. Due to many external factors ranging from predictable to unpredictable, it is nearly impossible to perform forecasting the future price movements. Fortunately, in recent years, the field of predictive analytic, more specifically, the field of Machine Learning and Deep Learning is exploding with many new models and techniques that can be used in predicting the stock market trend. It is found that in Machine Learning arena, the best runner is SVM due to its ability to turn the low dimensional space data to high dimensional space and become linearly separable. As studies show that stock dataset is noisy and non-linear, SVM is the best solution even though the architecture and algorithm can be complex and time-consuming. In recent years, in the field of Deep Learning, various models develop based on the structure of the Neural Networks have been introducing into the domain of the financial market. Models like RNN, CNN, LSTM and DNN is widely known among every financial market analysts. These neural networks models were proved to have better performance and prediction accuracy. The demonstration of the output on different charts shows the prediction resembles the movement of the target stock with a very low error rate. There are a few limitations that are needed to be taken into consideration, the risk aspect were not considered among the research papers, moreover, it was found that certain Deep Learning model achieved the accuracy rate only slightly higher than Machine Learning one. Additionally, most of the paper only consider the model in one market environment and disregard the application in international markets and certain papers did not compare both machine learning and deep learning methodologies in the same market with same dataset.

After weighing the benefit and drawbacks of both methodologies, a conclusion can be drawn that Deep Learning has more advantages in comparison to Machine Learning models. Moreover, until this point in time, the development of methodologies to achieve accuracy of 80% or 90% is still a myth, this is not only due to the non-stationary, non-linear data set with chaotic and complex nature but also due to the influence of external factors like the national monetary, fiscal policies and many unpredictable and unforeseeable external factors [7]

In the future, more data need to be collected from more sources need to be collected where different models in both Machine Learning and Deep Learning are compared. Moreover, with the increase in the advancement of technologies, Deep Learning might become the most favourable method to use when it comes to this domain, future research can focus on exploring other Deep Learning methodologies that were not mentioned in this research paper. Additionally, sentiment analysis and investor behaviour analysis need to be added in the pre-processing stage of the architecture to produce better input data for the model. Finally, risk analyse also needed to be considered as all investment come with risks, by this, the accuracy will significantly increase while risks are minimising

REFERENCES

[1] D. P. Gandhmal and K. Kumar, "Systematic analysis and review of stock market prediction techniques," *Computer Science Review,* vol. 34, 2019, doi: 10.1016/j.cosrev.2019.08.001.

[2] Y. Karaca, Y.-D. Zhang, and K. Muhammad, "Characterizing Complexity and Self-Similarity Based on Fractal and Entropy Analyses for Stock Market Forecast Modelling," *Expert Systems with Applications,* vol. 144, 2020, doi: 10.1016/j.eswa.2019.113098.

[3] W. Long, Z. Lu, and L. Cui, "Deep learning-based feature engineering for stock price movement prediction," *Knowledge-Based Systems,* vol. 164, pp. 163-173, 2019, doi: 10.1016/j.knosys.2018.10.034.

[4] J. Long, Z. Chen, W. He, T. Wu, and J. Ren, "An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market," *Applied Soft Computing,* vol. 91, 2020, doi: 10.1016/j.asoc.2020.106205.

[5] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong," *Information Processing & Management,* 2020, doi: 10.1016/j.ipm.2020.102212.

[6] R. Yang, "Big data analytics for financial Market volatility forecast based on support vector machine," *International Journal of Information Management,* vol. 50, pp. 452-462, 2020, doi: 10.1016/j.ijinfomgt.2019.05.027.

[7] C. Xiao, W. Xia, and J. Jiang, "Stock price forecast based on combined model of ARI-MA-LS-SVM," *Neural Computing and Applications,* vol. 32, no. 10, pp. 5379-5388, 2020, doi: 10.1007/s00521-019-04698-5.

[8] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, "Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market," *IEEE Access,* vol. 8, pp. 22672-22685, 2020, doi: 10.1109/access.2020.2969293.

[9] R. Ramezanian, A. Peymanfar, and S. B. Ebrahimi, "An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in Tehran stock exchange market," *Applied Soft Computing,* vol. 82, 2019, doi: 10.1016/j.asoc.2019.105551.

[10] H. Maqsood, "A local and global event sentiment based efficient stock exchange forecasting using deep learning," *International Journal of Information Management,* vol. 50, pp. 432-451, 2020, doi: 10.1016/j.ijinfomgt.2019.07.011.

[11] L. O. Orimoloye, M.-C. Sung, T. Ma, and J. E. V. Johnson, "Comparing the effectiveness of deep feedforward neural networks and shallow architectures for predicting stock price indices," *Expert Systems with Applications,* vol. 139, 2020, doi: 10.1016/j.eswa.2019.112828.

[12] J. Qiu, B. Wang, and C. Zhou, "Forecasting stock prices with long-short term memory neural network based on attention mechanism," *PLoS One,* vol. 15, no. 1, p. e0227222, 2020, doi: 10.1371/journal.pone.0227222.

[13] H. Tang, P. Dong, and Y. Shi, "A new approach of integrating piecewise linear representation and weighted support vector machine for forecasting stock turning points," *Applied Soft Computing,* vol. 78, pp. 685-696, 2019, doi: 10.1016/j.asoc.2019.02.039.

[14] H. Wang, S. Lu, and J. Zhao, "Aggregating multiple types of complex data in stock market prediction: A model-independent framework," *Knowledge-Based Systems,* vol. 164, pp. 193-204, 2019, doi: 10.1016/j.knosys.2018.10.035.

[15] J. M. Calabuig, H. Falciani, and E. A. Sánchez-Pérez, "Dreaming machine learning: Lipschitz extensions for reinforcement learning on financial markets," *Neurocomputing,* 2020, doi: 10.1016/j.neucom.2020.02.052.

[16] E. Ahmadi, M. Jasemi, L. Monplaisir, M. A. Nabavi, A. Mahmoodi, and P. Amini Jam, "New efficient hybrid candlestick technical analysis model for stock market timing on the basis of

the Support Vector Machine and Heuristic Algorithms of Imperialist Competition and Genetic," *Expert Systems with Applications,* vol. 94, pp. 21-31, 2018, doi: 10.1016/j.eswa.2017.10.023.

[17] T. K. Lee, J. H. Cho, D. S. Kwon, and S. Y. Sohn, "Global stock market investment strategies based on financial network indicators using machine learning techniques," *Expert Systems with Applications,* vol. 117, pp. 228-242, 2019, doi: 10.1016/j.eswa.2018.09.005.

[18] F. Yang, Z. Chen, J. Li, and L. Tang, "A novel hybrid stock selection method with stock prediction," *Applied Soft Computing,* vol. 80, pp. 820-831, 2019, doi: 10.1016/j.asoc.2019.03.028.

[19] J. Zhang, S. Cui, Y. Xu, Q. Li, and T. Li, "A novel data-driven stock price trend prediction system," *Expert Systems with Applications,* vol. 97, pp. 60-69, 2018, doi: 10.1016/j.eswa.2017.12.026.

[20] T. Niu, J. Wang, H. Lu, W. Yang, and P. Du, "Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting," *Expert Systems with Applications,* vol. 148, 2020, doi: 10.1016/j.eswa.2020.113237.

[21] K. Lei, B. Zhang, Y. Li, M. Yang, and Y. Shen, "Time-driven feature-aware jointly deep reinforcement learning for financial signal representation and algorithmic trading," *Expert Systems with Applications,* vol. 140, 2020, doi: 10.1016/j.eswa.2019.112872.

[22] M. Ben and I. Gbenga, "Predictive intraday correlations in stable and volatile market environments," 2020.

[23] B. Moews, J. M. Herrmann, and G. Ibikunle, "Lagged correlation-based deep learning for directional trend change prediction in financial time series," *Expert Systems with Applications,* vol. 120, pp. 197-206, 2019, doi: 10.1016/j.eswa.2018.11.027.

[24] W. Wang, W. Li, N. Zhang, and K. Liu, "Portfolio formation with preselection using deep learning from long-term financial data," *Expert Systems with Applications,* vol. 143, 2020, doi: 10.1016/j.eswa.2019.113042.

[25] L. Lei, "Wavelet Neural Network Prediction Method of Stock Price Trend Based on Rough Set Attribute Reduction," *Applied Soft Computing,* vol. 62, pp. 923-932, 2018, doi: 10.1016/j.asoc.2017.09.029.

[26] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Systems with Applications,* vol. 42, no. 1, pp. 259-268, 2015, doi: 10.1016/j.eswa.2014.07.040.

APPENDIX

Appendix 1

ABBREVIATION TABLE		
DATA ATTRIBUTES	ALGORITHM	FEATURES/INDICATOR COLLECTION METHODS
ART: Average True Range BP: Buying Price CP: Closing Price DC: Daily Change FQF: Financial quality factors GF: Growth Factors HP: High Price LF: Liquidity factors LF: Leverage factors LP: Low Price MoM: Momentum Index MF: Momentum factors OP: Open Price PSAR: Parabolic SAR RoC: Rate of Change RSI: Relative Strength Index SR: Simple return, SF: size factors TF: Technical factors TR: True Range TV: Transaction/TV VF: Valuation factors VoF: Volatility factors	ANN: Artificial Neural Network ARIMA-LS-SVM Bi-LSTM: Long Short Terms Memories Network CNN: Convoluted Neural Network DSPNN FANN: Feedforward Artificial Neural Network GE: Graph Embedding Techniques IC: Investor Clustering KG: Knowledge Grap LR: Linear regression MII: Market Information Iteration MLP: Multilayer Perceptron NP: News Processing PRL: Piecewise Linear Representation OFL: Opinion Finder lexicon RF: Random Forest SA: Sentimental Analysis SD: Sentimental Dictionary SVM: Support Vector Machine WSVM: Weighted Support Vector Machine	RFE: Recursive Feature Elimination