ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 43(2023) 49-62

Deep and Machine Learning Models For Accurate Prediction and Classification of the Selected Online Shopping Stock Market

Godfrey Joseph Saqware and Ismail B.
[Received on January, 2022. Accepted on March, 2023]

ABSTRACT

Prediction and classification of the stock price movement remain crucial to traders and financial analysts. The complexity and fluctuations of the stock market make its prediction and classification a difficult task. In this study, we compared the performance of the machine learning and deep learning models in predicting and classifying the Amazon online stock market Index. The prediction and classification have resulted in high accuracy compared to previous studies. After applying the single-step wrapping method, the second stage resulted in the best six features. The sequential ANN with hard-sigmoid transfer function and Deep learning with Rectified Linear Unit (ReLU) transfer function has correctly predicted and classified the price movements. The experimental accuracy of the prediction models has improved by more than 15% compared to the previous studies. The proposed model may be applied to other stock market indices to evaluate their price movement.

1. Introduction

The technical and fundamental analysis of the stock market movement has become the most important aspect of buying or selling decisions of the stocks. The current progressively international economy, return, and volatility spillover affect markets which are the major macroeconomic drivers for the stock dynamics (Lee & Yoo, 2019). Thus, information regarding foreign markets is one of the essential factors in forecasting stock prices. The stock price prediction movement is mainly affected by the ups and downs in the financial market. Thus,

: Godfrey Joseph Saqware
Email: godjose70@yahoo.com

Extended authors information available after reference list of the article.

the timely and powerful stock price predictive model becomes necessary for more profits and returns. The financial time series movement is commonly noisy, non-linear, and complicated to predict (Kara *et al.*, 2011). Due to frequent macroeconomic, political, and government policy changes, stock market prices fluctuate. These policies significantly impact the stock market investors' life, thus resulting in complex analyses and decisions (Raza, 2017). The studies on machine learning (ML), Artificial Intelligence (AI), and Deep Learning (DL) are mainly based on medical image classification, electronic health record analysis, and Parkinson's disease diagnosis (Tahmassebi *et al.*, 2018). Deep learning has recently proven high performance in solving complex stock market issues (Raşo & Demirci, 2019). Therefore, accurate decisions in the stock market's fundamental and technical analysis using AI, ML, and DL methods have been widely used for predicting non-linear, noisy, and chaotic kinds of data, including the stock market. For years, studies have considered data mining and Neural Networks (NN) methods. These methods had limitations for learning the more extensive non-linear and complex stock data (Kara *et al.*, 2011).

Therefore, the study will identify and select technical indicators using Boruta feature selection techniques and later accurately predict and classify stock price movements.

2. Related Literature

The prediction of the stock price movement based on Deep learning has improved the stock market prediction and classification performance with an accuracy of 5%-6% (Naik & Mohan, 2019). The stock market's nonlinearity and high change rates have become challenging for traders and data scientists. The extensive application of machine learning models provides a better solution to complex stock market problems. The prediction of the intraday forex currency has been effectively performed using a multi-layer algorithm that has successfully demonstrated a predictive relationship between market type and the textual data of news with an accuracy of 83.3%, which broke financial news headlines (Nassirtoussi *et al.*, 2015).

Interestingly, using the support vector machines to build a regression model by transforming the fuzzy particle sequence has decreased the financial time series estimation errors (Gui *et al.*, 2014). The hybrid model performed well in forecasting the stock market movement from the selected important technical indicators features based on kernel and random walk methods (Ince & Trafalis, 2017). Deep learning technique has more enormous potentialities in the modelling and prediction of the noisy and complex financial data based on support vector machine and k-nearest neighbors (Chen *et al.*, 2016). Reducing the noise and nonlinearity using dimension reduction techniques such as fuzzy robust principal component analysis and KPCA helps to construct the stock market forecasting model (Zhong & Enke, 2017). The three feature selections, namely PCA, genetic algorithms and decision trees, have performed well in forecasting stock prices(Tsai & Hsiao, 2010). The prediction capacity of the stock price model based on the integrated Kernel Principal Component Analysis (KPCA) and Support Vector Machines for Regression (SVR) has significantly improved the effectiveness and accuracy of the prediction method (Nahil & Lyhyaoui, 2018).

The combination of prediction and classification methods has always been changing. The performance accuracy from the previous studies is not satisfactory. This paper will study and classify the online shopping stock movement's prediction and classification. The study aims to develop the required features to predict and classify the Amazon online shopping stock market index with higher accuracy.

3. Materials and Methods

3.1 Data specification

This paper considers stock data collected from *https://www.investing.com/equities/amazon-com-inc*. The data contain information about stock such as stock day open price, day low price, day high price, day close price, and volume. Amazon online shopping daily stock market data from January 2010 to February 2020 is considered. Moreover, the stock Market contains Open, High, Low, Closing, Adj. Close, and Volume. Comparing the current and previous prices has made a stock tag up or down movement in the online purchase market. The data is divided into 80% training and 20% testing sets.

The study considered 32 combinations of technical indicators and computed them based on formulas by Kara *et al.* (2011), described in Table 1.

Table 1: Formula for computing technical indicators.

Name of Indicator	Calculation	Number of Days
Simple Moving Average (SMA)	$(C_t + C_{t-1} +\ldots+ C_{t-n+1})/n$	5, 10,14, 30, 50, 100, 200
Exponential Moving Average	$(C_t + SMA(n)_{t-1})$ *(2/n+1) +$SMA(n)_{t-1}$	5, 10,14, 30, 50, 100, 200
Momentum indicator	C_t - C_n -9	5, 10,14
Stochastic Oscillator	100*(C_t - L_t(n))/(Ht(n) - L_t(n))	14
Stochastic Oscillator	100*(C_t - L_t(n))/(H_t(n) - L_t(n)))/3	14
Moving Average Convergence Divergence	SMA(n) – SMA(n)	26, 13, 19, 45, 25, 15
Relative Strength Index	100-(100/(1+Avg (Gain)+Avg (Loss))	14, 28
Williams R	(H_n - C_t)/ (H_n- L_n) *-100	14, 28, 50, 100
Accumulation/Distribution Index	((H_t - C_{t-1})/H_t) - L_t ((C_t /L)/(H/C))/(H/L)	14
Commodity Channel Index	(H+L+C)/3–SMA/ (0.015*mean deviation)	14, 50, 100

Source (Naik & Mohan, 2019)

Where, C_t is the closing price, L_t is the low price, H_t is the high price, and Avg is Average.

3.2 The Selection of the Technical Indicators

A random forest algorithm has been applied to find important technical indicator features based on higher mean values. In this algorithm, the Z score threshold value of 0.70 is considered. Any technical indicator features with a threshold value greater than 0.70 is considered for classification. The proposed Boruta feature selection algorithm is stated in Algorithm 1. The Boruta feature selection uses the Boruta package in R programming. Then, the selected technical indicators using the Boruta algorithm were inputted into the prediction model.

Algorithm 1

1. Input 32 technical features F
2. Create duplicate/shadow of technical indicators D.
3. Do a random shuffle on the original technical indicator F and duplicate copies of technical indicator D to remove correlation with the outcome variable.
4. Apply the random forest algorithm to find important technical indicators based on higher mean values.
5. Compute the Z score by using Mean/Standard deviation.
6. Find the maximum Z score on duplicate technical indicator features.
7. Remove the technical indicator feature if the Z score is less than the technical indicator feature.

Fig. (1) represents the overall process flow, which involves inputting data, computation of technical indicators, feature selection, and models used for classification tasks.

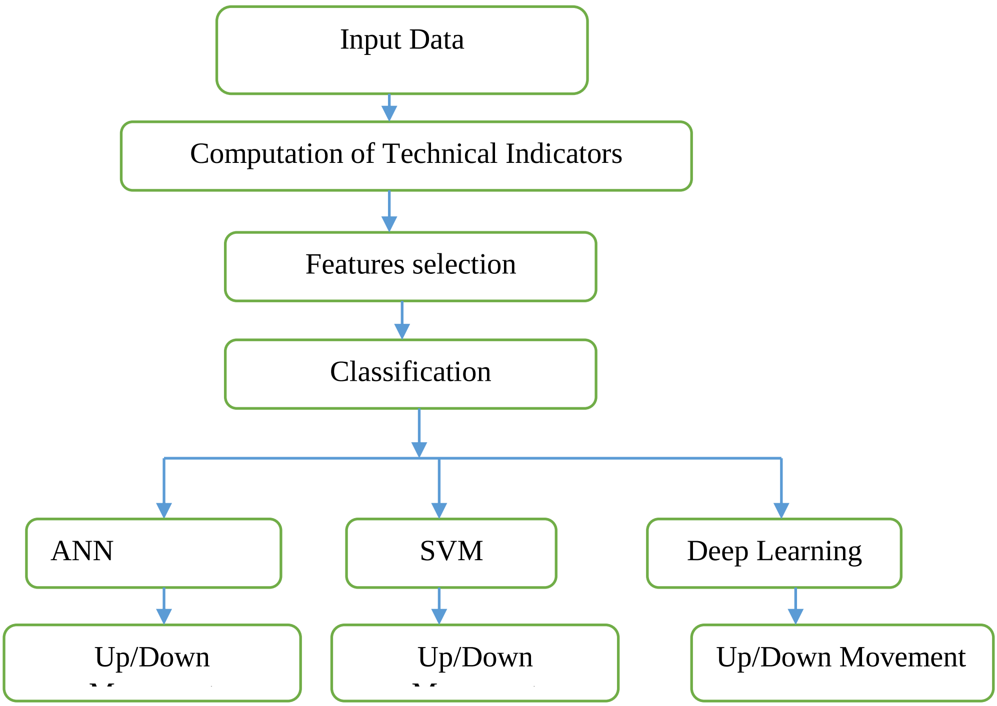

Figure 1: Overall process flow.

3.3 Model Prediction

3.3.1 Deep Learning Model

In recent computer application errors and technological advancements, deep learning models have become famous in different areas, such as image recognition, classification, and electronic health record analysis (Ravì *et al.*, 2016). Three layers of artificial neural network and Deep learning framework are proposed for the stock prediction (Kara *et al.*, 2011). The H2O and Boruta were used in the task implementation and features selection, respectively. The deep learning model classifies the ups and downs movement of the stock price. The model consists of five interconnected layers of neuron units that transform the data. The selected technical indicators act as the input layer t_i and W_i weights of the neurons. The weight adjustments are made by applying stochastic gradient descent and back-propagation. The bias input is assigned in each layer except in the output layer. The objective function $L(W, Bias\,|\,j)$ reduces the classification error in the stock data. The linear combination of the inputs and weight is given in Eqn. (1.1).

$$\psi = \left(\sum_{i=1}^{n} W_i\, t_i + Bias \right) \tag{1.1}$$

The regularization function is given in Eqn. (1.2) below helps to avoid overfitting in the deep learning model.

$$L(W, Bias\,|\,j) = L(W, Bias\,|\,j) + \lambda_1 R_1(W, Bias\,|\,j) + \lambda_2 R_2(W, Bias\,|\,j) \tag{1.2}$$

Where L_1 and L_2 are the regularizers.

3.3.2 Artificial Neural Network (ANN) Model

The selected technical indicator based on the Boruta algorithm was used as the ANN model input. The ANN is used to classify the stock price movement in this work. The proposed ANN has three interconnected hidden layers. The selected technical indicators will act as the input layer t_i and W_i weights of the neurons. The sigmoid activation function g incorporated into the ANN model squishes the effects of the minimal and maximum values. We consider the gradient descent momentum parameters to determine the appropriate weights and reduce the global minimum. Eqn. (1.3) shows the ANN model representation.

$$f(t) = g\left(\theta_j + \sum_{i=1}^{p} W_{ij} t_i\right) \tag{1.3}$$

3.3.3 Support Vector Machines (SVM)

The SVM has based on Vapnik–Chervonenkis (VC) learning theory, under which the major components were developed (Boser *et al.*, 1992; Vapnik & Chervonenkis, 1974). The SVM shows strong performance in predicting and classifying real-world problems. In this prediction model, the SVM is used to predict and classify two classes, namely up and down price movements. Moreover, the SVM hyperplane separates the two input vectors. Stock market data are non-linear separable datasets; thus, SVM can be more productive. The stock data prediction used the SVM model with polynomial or radial basis kernel activation function. Eqn. (1.4) and (1.5) shows the Polynomial and Radial Basis Activation functions, respectively.

$$Polynomial\ \ Function: K\left(f_i, f_j\right) = \left(f_i * f_j + 1\right)^d \tag{1.4}$$

$$Radial\ \ Basis\ \ Function: K\left(f_i, f_j\right) = exp\left(\gamma \left\| f_i - f_j \right\|^2\right) \tag{1.5}$$

The Performance Metrics

Sensitivity, Specificity, and Accuracy are used to qualify the performance of the classifiers in discriminating between upward and downward stock price movement. Stock class tag up and down was obtained by comparing the current and previous stock prices. The model accuracy was used as the performance evaluation criteria for deep and Machine learning models. The Sensitivity, Specificity, and Accuracy formulas are given in Eqn. (1.6), (1.7), and (1.8).

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{1.4}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{1.5}$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{1.6}$$

True Positive (TP) is the number of upward movements classified as upwards movement; the False Positive (FP) is the number of upward stock movements classified as a downward stock movement. True Negative (TN) is the number of downward stock movements classified as a downward stock movement. False Negative (FN) is the number of downward movements classified as an upward stock movement.

4. Empirical Results and Discussion

The Boruta features selection results indicated that out of 32 created features from the Amazon online shopping stock market index, 26 features were essential for the upward and downward movement prediction and classification. The features such as 30 and 100 days Simple Moving Average (SMA); and 14, 30, 100, and 200 days exponential moving average were found with decision tentative after the maximum iteration of 99 and hence unimportant for further model classification and prediction. Table 2 shows summary statistics for the Boruta features selection.

Table 2: Summary statistics for the selected important features.

Indicator	Days	Mean	Median	Min	Max	NormHits	Decision
Simple Moving Average (SMA)	5	3.8935	3.8572	1.9402	6.0224	0.8586	Confirmed
	10	3.2564	3.2465	0.7749	5.5592	0.6768	Confirmed
	14	3.1060	3.1830	0.7067	4.8019	0.6970	Confirmed
	30	2.9789	2.9222	0.1547	5.5907	0.6263	Tentative
	50	3.2174	3.3428	0.0219	4.9796	0.7071	Confirmed
	100	3.0558	3.0641	0.6347	5.8723	0.6465	Tentative
	200	3.1218	3.1401	0.6446	4.8949	0.6768	Confirmed
Exponential Moving Average (EMA)	5	3.5403	3.4540	1.2286	5.5245	0.7980	Confirmed
	10	3.1810	3.2016	1.3086	5.4588	0.7172	Confirmed
	14	3.1690	3.1807	0.2449	4.9262	0.6566	Tentative
	30	2.9514	3.0098	1.0503	4.9893	0.6566	Tentative

Indicator	Days	Mean	Median	Min	Max	NormHits	Decision
	50	3.0600	3.0115	0.7463	5.9091	0.6768	Confirmed
	100	2.9116	2.8391	0.8405	5.2682	0.5960	Tentative
	200	2.9221	2.9989	0.8747	5.4741	0.6061	Tentative
Momentum indicator (MOM)	5	13.2190	13.1571	11.5346	15.3096	1.0000	Confirmed
	10	10.4507	10.4449	8.7093	11.8275	1.0000	Confirmed
	14	8.2265	8.1871	6.5575	9.9767	1.0000	Confirmed
Stochastic Oscillator (K %)	14	30.9279	30.9174	28.4457	33.6591	1.0000	Confirmed
Stochastic Oscillator (D %)	14	26.9166	26.8783	24.3462	29.5670	1.0000	Confirmed
Moving Average Convergence / Divergence (MADC)	13,26	6.4127	6.5192	3.4938	8.5758	1.0000	Confirmed
	19,45	4.4181	4.4148	2.2782	7.7054	0.9394	Confirmed
	25,15	5.5329	5.6268	3.0448	7.9822	0.9899	Confirmed
Relative Strength Index (RSI)	14	11.6055	11.6662	9.3751	13.2195	1.0000	Confirmed
	28	5.8302	5.8856	3.8643	7.8526	1.0000	Confirmed
William %R	14	30.8342	30.8285	28.4237	32.8114	1.0000	Confirmed
	28	20.6999	20.6405	18.9707	22.8284	1.0000	Confirmed
	50	16.7309	16.7629	14.6329	18.3863	1.0000	Confirmed
	100	6.0363	6.0790	4.0598	7.8041	1.0000	Confirmed
Accumulation/Distribution Index(A/D)	14	4.5455	4.5949	2.4574	6.7073	0.9394	Confirmed
Commodity Channel Index (CCI)	14	12.9168	12.9215	10.8466	15.1426	1.0000	Confirmed
	50	9.7846	9.7901	7.4785	11.8110	1.0000	Confirmed
	100	4.5417	4.4770	2.1939	7.2130	0.9192	Confirmed

4.1 The prediction and classification of the selected features

The number of the chosen features from 6 to 26 at an interval of four has been tested to see the accuracy. The accuracy of the ANN prediction and classification model is unstable even with an increasing number of features. The final accuracy at 26 features was lower than 22, which gave an accuracy of 79.38%. The SVM has shown an increasing accuracy as the number of features increased and attained the highest accuracy of 76.44% for all 26 features. The Deep Learning model using H2O has the highest prediction and classification accuracy of 79.76% for all 26 selected features. The highest accuracy has performed well but will lower percent compared to the study done by (Kara *et al.* 2011; Masoud, 2014; Naik & Mohan, 2019; Tsai & Hsiao, 2010). Thus, the study has to analyze further to develop the highest prediction and classification accuracy. Table 3 shows the results of the subgroups of the 26 features selected using the Boruta method.

Table 3: Summary of the subsets for the selected features.

Classifiers	Performance	Number of Features					
		6	10	14	18	22	26
ANN	Sensitivity (%)	58.61	69.17	68.06	68.61	89.44	0.00
	Specificity (%)	77.59	61.49	65.23	64.08	95.96	100.00
	Accuracy (%)	67.94	65.40	66.67	66.38	79.38	49.79
SVM	Sensitivity (%)	58.06	63.34	63.93	64.52	78.89	80.94
	Specificity (%)	53.76	53.76	49.46	62.37	72.58	72.31
	Accuracy (%)	55.82	58.35	56.38	63.39	75.60	76.44
Deep Learning	Sensitivity (%)	100.00	95.87	91.46	93.38	86.23	89.53
	Specificity (%)	0.00	21.28	27.96	18.84	61.09	69.00
	Accuracy (%)	52.45	60.40	61.27	57.95	78.61	79.76

4.1 The extension of the Boruta features selection method.

Classifiers	Performance	epochs		
		300	600	900
ANN(Sequential)	Sensitivity (%)	59.17	90.56	100
	Specificity (%)	72.13	91.38	100
	Accuracy (%)	65.54	90.96	100
Deep Learning (Sequential)	Performance	epochs		
		100	200	300
	Sensitivity (%)	61.67	100	100
	Specificity (%)	77.30	100	100
	Accuracy (%)	69.35	100	100

Further analysis was done using the step forward wrapper method to achieve higher prediction and classification accuracy (Mlambo *et al.*, 2016). The methods identified six features after removing the highly correlated features from the 26 selected using the Boruta methods. The Sequential ANN model (7, 16, 1) with the hard sigmoid transfer function has correctly predicted and classified the upward and downward Amazon online shopping stock price using 900 epochs with 100% accuracy. The results have shown an exciting improvement compared to the study on the Libyan stock market using ANN with an accuracy of 91% (Masoud, 2014). Its counterpart, the sequential Deep learning model with the Rectified Linear Unit (ReLU) with four Deep layers of (7, 200, 200, 200, 200, 1), has correctly predicted and classified the Amazon online shopping stock market price movements with only 200 epochs. The results have shown the most significant prediction and classification results compared to the previous studies (Kara *et al.*, 2011; Naik & Mohan, 2019) with more than 15% accuracy. The results for the ANN with a single proposed layer and Deep learning sequential model have also performed better than the combination of multiple models (Tsai & Hsiao, 2010). Moreover, we found the best results compared to (Huynh *et al.*, 2017), which also proposed a Deep learning model. Table 4 shows the proposed model summary at different epochs.

Table 4: The summary of the proposed model at different epochs

5. Conclusion

Due to its unpredictable behaviours, stock market prediction and classification have always been difficult for researchers and financial analysts. The study focused on predicting stock price movement and classifying Amazon's online daily stock price. The Boruta features selection followed by the single forward wrapper method has improved accuracy in predicting and classifying the stock price movement. The ANN sequential model with hard-sigmoid transfer function and the Deep learning sequential model with ReLU transfer function at 900 and 200 epochs, respectively, has shown the best performance. Eventually, the study demonstrated that the Boruta features selection could not stand alone to develop the best prediction and classification model in terms of accuracy. The study is based on the Amazon online stock market. The method may be applied to the other stock market to see the consistency of the results.

References

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992): *A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory; Pittsburgh, Pennsylvania, USA. 130401: ACM*.

Chen, J.F., Chen, W.L., Huang, C.P., Huang, S.H., and Chen, A.P. (2016): Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, **87–92**. https://doi.org/10.1109/CCBD.2016.027

Gui, B., Wei, X., Shen, Q., Qi, J., and Guo, L. (2014): Financial time series forecasting using support vector machine. *2014 Tenth International Conference on Computational Intelligence and Security*, 39–43.

Huynh, H. D., Dang, L. M., and Duong, D. (2017): A new model for stock price movements prediction using deep neural network. *Proceedings of the Eighth International Symposium on Information and Communication Technology*,57-62.

Ince, H., and Trafalis, T. B. (2017): A hybrid forecasting model for stock market prediction. *Economic Computation & Economic Cybernetics Studies & Research*, ***51*(3)**.

Kara, Y., Boyacioglu, M. A., and Baykan, Ö. K. (2011): Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, ***38*(5)**, 5311–5319.

Lee, S. I., and Yoo, S. J. (2019): Multimodal deep learning for finance: Integrating and forecasting international stock markets. *The Journal of Supercomputing*, 1-19.

Masoud, N. (2014): Predicting direction of stock prices index movement using artificial neural networks: The case of Libyan financial market. *Journal of Economics, Management and Trade*, 597–619.

Mlambo, N., Cheruiyot, W. K., and Kimwele, M. W. (2016): A survey and comparative study of filter and wrapper feature selection techniques. *Int. J. Eng. Sci*, ***5*(8)**, 57–67.

Nahil, A., and Lyhyaoui, A. (2018): Short-term stock price forecasting using kernel principal component analysis and support vector machines: The case of Casablanca stock exchange. *Procedia Computer Science*, ***127***, 161–169.

Naik, N., and Mohan, B. R. (2019): Stock Price Movements Classification Using Machine and Deep Learning Techniques-The Case Study of Indian Stock Market. In J. Macintyre, L. Iliadis, I. Maglogiannis, and C. Jayne (Eds.), *Engineering Applications of Neural Networks* **1000**, 445–452. Springer International Publishing. https://doi.org/10.1007/978-3-030-20257-6_38.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2015): Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, ***42*(1),** 306–324.

Raşo, H., and Demirci, M. (2019): Predicting the Turkish Stock Market BIST 30 Index using Deep Learning. *Uluslararası Muhendislik Arastirma ve Gelistirme Dergisi*, 253–265. https://doi.org/10.29137/umagd.425560.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. (2016): Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, ***21*(1)**, 4–21.

Raza, K. (2017): Prediction of Stock Market performance by using machine learning techniques. *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, 1–1. https://doi.org/10.1109/ICIEECT.2017.7916583.

Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Goudriaan, A. E., and Meyer-Baese, A. (2018): Deep learning in medical imaging: FMRI big data analysis via convolutional neural networks. In *Proceedings of the Practice and Experience on Advanced Research Computing* (pp. 1–4).

Tsai, C.F., and Hsiao, Y.C. (2010): Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, ***50*(1)**, 258–269.

Vapnik, V., and Chervonenkis, A. (1974): *Theory of pattern recognition*. Nauka, Moscow.

Zhong, X., and Enke, D. (2017): Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, ***67***, 126–139.

Authors and Affiliations

Godfrey Joseph Saqware[1,3*] and Ismail B.[2]

Ismail B.
prof.ismailb@gmail.com

[1]Department of Statistics, Mangalore University, Karnataka, India.
[2]Department of Statistics, Yenepoya (Deemed to be University), Karnataka, India.
[3]University of Dar es Salaam, Tanzania