International Journal of Economics and Financial Issues

ISSN: 2146-4138

available at http: www.econjournals.com

International Journal of Economics and Financial Issues, 2024, 14(2), 5-14.

Forecasting Stock Market Realized Volatility using Random Forest and Artificial Neural Network in South Africa

Lamine Diane, Pradeep Brijlal*

Commerce Faculty, University of Cape Town, South Africa. *Email: pradeep.brijlal@uct.ac.za

Received: 11 October 2023 **Accepted:** 25 January 2024 **DOI:** https://doi.org/10.32479/ijefi.15431

ABSTRACT

Volatility is often used as a key input into several financial models, yet there is still no consensus on the best-performing model in forecasting stock market returns volatility. Conventional time series models such as GARCH are the preferred models in the literature. However, this project aims to first adopt two novel non-linear machine learning algorithms, namely the Random Forest and Artificial Neural Network (ANN). The project then compares the performance of these two models in predicting stock market realized volatility for the JSE Basic Material Index (JBIND) and the JSE Financials Index (JFIN) over a period of 5 years. Based on the results of the project, the Random Forest model outperformed the ANN model for both the JFIN and JBIND index. Lastly, the COVID effect on the model's performance was also considered and the results show that the negative impact of COVID on the model's performance is ambiguous.

Keywords: Forecasting, Realized Volatility, Random Forest, Artificial Neural Network
JEL Classifications: G11, G17

1. INTRODUCTION

Volatility is a measure of the degree of fluctuation in financial return and is referred to as a proxy for risk by many market practitioners when it comes to investment decisions and portfolio creation (Poon and Granger, 2003). Therefore, when volatility is considered as a measure of risk, an appropriate forecasting model is crucial for market practitioners and policymakers. Three reasons were outlined by Poon and Granger (2003) for the importance of an appropriate forecasting model. Firstly, Poon and Granger (2003) noted that volatility becomes a key input to many investment decisions and portfolio creations once it is interpreted as uncertainty. The second reason for the importance of volatility outlined by Poon and Granger (2003) is that when it comes to pricing derivatives securities volatility is the most important factor. Thirdly, after the establishment of the Basle Accord in 1996, financial risk management has taken a dominant role, making appropriate modelling of volatility a necessary risk-management exercise for financial institutions around the world. Moreover, according to Bonga-Bonga (2017), foreign investors and asset managers are increasingly viewing emerging markets such as South Africa markets as a source of potential portfolio diversification, therefore, an accurate volatility forecast enables these investors to accurately assess the risks their investment will be exposed to.

It is very difficult to observe volatility; thus, volatility is usually considered as the standard deviation of asset returns over a given period (Pati et al., 2018). Several statistical and computational models have been developed over the years to forecast the volatility of financial assets. The autoregressive conditional heteroskedasticity (ARCH) model proposed by Engle (1982) and its extended version, namely, the generalized autoregressive conditional heteroskedasticity (GARCH) model proposed by Bollerslev (1986) has been the most prominent models used in literature in the past decades. However, the effect of non-linearity in stocks returns and the effect of complex interactions between stock returns and market variables such as economic conditions or trader's expectations, has led to the development of newer non-parametric machine learning methods over the past few decades including Random forest (RF) and Artificial Neural Networks (ANN) which captures nonlinear behavior and complex interaction between stock returns and market variables (Huang et al., 2005). Random Forest model and ANN have received very little attention in terms of forecasting financial stock market volatility despite their potential predictive accuracy, and their successful application in other fields including the energy field (Ahmad et al., 2014) and water resources field (Maier and Dandy, 2000).

Kumar and Thenmozhi (2006) stated that machine learning algorithm accuracy varies across countries and regions. Since the focus of this project will be on the South African market, this raises the need to test and compare the accuracy of Random Forest and ANN in the South African market context to establish the best-performing model. As of 2020, JSE was the sixteenth largest stock exchange in the world with a total market capitalization of 1.05 trillion US dollars (World Bank, n.d.), yet there is very limited research on forecasting stock market volatility.

Therefore, this research project aims to implement a Random Forest and ANN algorithm and evaluate which model performs the best in predicting realized volatility of returns for the JSE Basic Materials Index (JBIND) and the JSE Financial Index (JFIN) in the South African stock market. Realized volatility was used instead of implied volatility due to the lack of derivatives on the JFIN and JBIND Index. The lack of o derivatives on these two indices makes it infeasible to use implied volatility because implied volatility requires the availability of derivatives on securities (Poon and Granger, 2003). In terms of the number of indices selected, this project uses two indices. This is similar to Alberg et al. (2008) paper which also made use of two indices. Several indices were available on the JSE but the two indices selected were JSE Financials Index (JFIN) and the JSE Basic Material Index (JBIND) since these two industries have a significantly higher constituent on the JSE top 40 than any other industries, accounting for approximately half of the JSE top 40 index which reflects the importance of the two indices on the JSE stock market (FTSE Russell, 2022). The two indices also exhibit very similar trends in realized volatility over the sample period from June 1st, 2017 to June 1st, 2022 as illustrated in Graph 1 below:

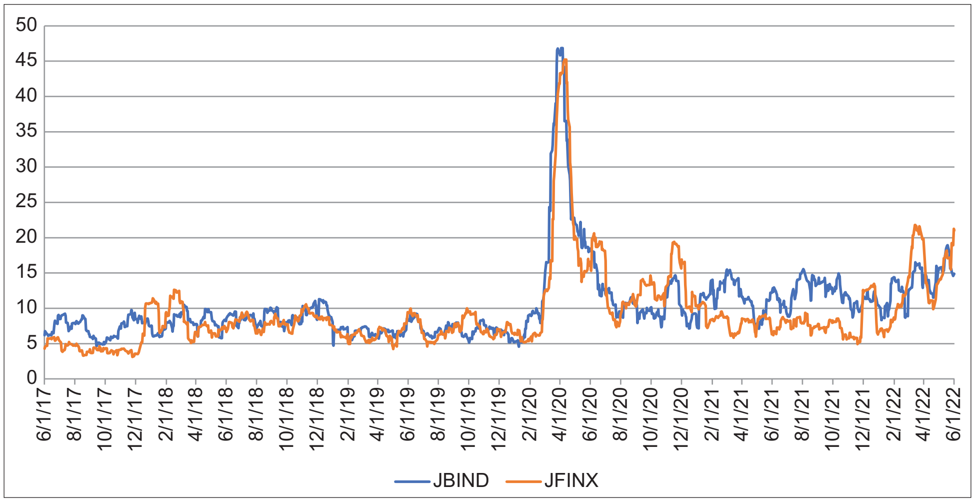
The remainder research project will be structured as follows: Section 2 will review some empirical evidence from past literature regarding stock market volatility forecasting. Section 3 will be the methodology section, where the variable selection process and datasets will be described. Section 4 is where the Random Forest and ANN model used in the project will be described. Section 5 will discuss the results obtained from the models and the possible impacts of the COVID-19 pandemic on the model's performance. Lastly, section 6 will conclude the main findings of the research project which will be followed by a discussion on the potential application and limitations of this research project and recommendations for future research.

2. LITERATURE REVIEW

This section will start by briefly discussing past literature on the traditional ARCH and GARCH models, which will be followed by a discussion on previous literature relating to the applications of ANN and Random Forest for forecasting purposes.

As stated before, the most important development in volatility modelling was the introduction of the ARCH model by Engle (1982). According to Engle (1982), the ARCH model captures one important aspect of returns volatility which is known as volatility persistence. Volatility persistence means that a period of large volatility is likely to be followed by subsequent periods of high volatility (Engle, 1982). The ARCH model was later generalized by Bollerslev (1986), who introduced the GARCH model. The GARCH model has been extensively and successfully implemented in the literature to forecast volatility in stock returns in emerging markets (Emenike, 2010; Cifter, 2012). In South Africa, Samouilhan and Shannon (2008) paper was among the first papers to predict the volatility of stock returns. Samouilhan

Graph 1: Realized volatility of JBIND index and JFINX (from June 1st, 2017 to June 1st)



Source: Own calculations using data from Bloomberg (2022) and Shannon (2008) implemented the ARCH models to forecast volatility in the JSE top 40 from February 1st 2004 to September 28th 2006. The paper found ARCH models to be a good predictor of volatility in the JSE top 40 index.

Nevertheless, the standard GARCH fails to capture the leverage effect which is often present in stock market returns volatility. The leverage effect is when a negative return shock is associated with larger increases in volatility relative to a similar positive return shock (Tripathy and Garg, 2013). The leverage effect was first observed in South Africa by Samouilhan and Shannon (2008) on the JSE top 40. Babikir et al. (2012) highlighted that it is important to account for the leverage forecasting volatility for better predictions. The leverage effect results in an asymmetric distribution of returns volatility, which violates the normality assumption requirement of the traditional GARCH model (Mashamba and Magweva, 2019). In addition to asymmetry returns distribution, the relationship between volatility and several fundamental and technical variables is often nonlinear (Chauduri and Ghosh, 2016). Therefore, nonlinear GARCH was developed in the literature to account for the leverage effect. The most notable nonlinear asymmetric GARCH model in forecasting volatility is the Exponential GARCH model (or EGARCH) (Bollerslev and Mikkelsen, 1996). The EGARCH model was the top-performing model in predicting two major Israeli Tel-Aviv indices from October 20th 1992 to May 31st 2005 result according to findings by Alberg et al. (2008) paper.

Though the presence of leveraging effect and the non-linearity in returns are accounted for by certain non-linear GARCH models such as EGARCH, the complex relationship between stock returns and market variables like changes in economic conditions have led to alternative models being proposed (Huang et al., 2005). One of these models is the Artificial Neural Networks model (ANN). ANN is still not widely used for financial times series forecasting due to its complexity and computational power requirement but the model has been extensively used in other fields such as science or engineering (Abiodun et al., 2018). That being said, in the emerging market context, Chauduri and Ghosh (2016) made use of the ANN based on the backpropagation method to forecast volatility in returns on the Indian stock market (NIFTY Index). Unlike most literature, this paper did not include the lagged values of volatility NIFTY returns to prevent the current value of the dependent variable from being affected by the previous values. Chauduri and Ghosh (2016) results showed that the ANN model outperforms the standard GARCH and EGARCH model. In developed markets, D'Ecclesia and Clementi (2021) also found that the ANN model outperforms the EGARCH model for major indices in China, Australia, Japan, Italy, Germany, the UK and the USA in terms of forecasting implied volatility over the period January 03 2011 to July 30 2018. Hamid and Iqbal (2004), on the other hand, aimed to forecast the volatility of S and P 500 future prices over 10 years from February 1st 1984 to January 31st 1994. Though Hamid and Iqbal (2004) acknowledged that the ANN model outperformed the alternative model, they stated that it was difficult to take advantage of the full potential of ANN in the finance field because model specification in ANN is not a perfect science resulting in several unexplored areas in modelling ANN, especially in terms of the different parameters that go into the ANN model.

On the other hand, Random Forest Models are often used for classification purposes rather than regression even though the model has been proven effective in both cases (Ballings et al., 2015). Similar to ANN, there is limited research on time-series financial forecasting for the Random forest model. Ballings et al. (2015) found that the Random Forest performs better than other machine learning algorithms in predicting the direction of 5767 European companies' stock prices. Furthermore, Khaidem et al. (2016) also found the random forest model to be superior to another popular machine learning algorithm, namely, Support Vector Machines (or SVM) in predicting the stock market direction of US stocks. Similar results were also found in emerging market contexts, Sharma and Juneja (2017) demonstrated higher performance of Random forest with LS-boost relative to SVM in predicting the Indian stock market index from 2006 to 2015. Luong and Dokuchaev (2018) later used the Random forest model to forecast the direction of realized volatility of S and P 200 and found that the Random forest was able to forecast the direction of volatility at 80% accuracy. However, Luong and Dokuchaev (2018) acknowledge that including technical indicators could have improved the performance better. For that reason, technical indicators will be considered as an independent variable in this project in the methodology section.

There are also not many financial papers which compare the Random Forest and ANN models directly. That being said, in the energy sector, Ahmad et al. (2017) found that the ANN performs slightly better than the Random forest model. On the other hand, Sevgen et al. (2019) compared the two models in the field of landslide susceptibility and found the ANN to be superior relative to Random Forest. This shows that there is no clear consensus on which models are the best and the results might vary across different datasets and fields.

The international evidence of successful implementation of both the Random forest and ANN along with the successful application of these methods in other fields such as the energy sector has motivated the need to develop an ANN and Random forest model for predicting volatility in the South African stock market. Currently, there is a scarcity of research on forecasting stock market returns volatility using machine learning algorithms such as ANN and Random Forest. Therefore, the objective of this project is to build an ANN and Random Forest model and evaluate its performance against each other in predicting the volatility of JFIN and JBIND Index according to relevant metrics. Thereby, the project will contribute towards the current literature by determining the most appropriate volatility forecasting model between the ANN and Random Forest in the South African market context.

The next section will consist of justification and explanation for all of the variables selected for the project which will then be followed by the section on ANN and Random Forest model specification.

3. RESEARCH METHODOLOGY

In this section, the methodology used in the project will be discussed. This section will cover the selection process of dependent and independent variables. A brief description of each variable will also be provided.

This project uses daily observations on the JFIN and JBIND index over the sample period starting June 1st 2017 to June 1st 2022. The selected sample period also includes the COVID effect which allows for further analysis in the results section regarding the impact of COVID on the model performance. The pricing data for the indices were obtained from Bloomberg terminals and 5 years was chosen due to the availability of data. Except for the Relative Strength index which will be discussed later, all variables were calculated using Excel.

Historical realized volatility will be the variable of interest and the dependent variable in the models. But, we need to first calculate the returns on the indices to be able to calculate the realized volatility. Since daily returns on assets is the preferred method used in the literature according to Krollner et al. (2010) paper, log daily returns for the two indices were calculated using the following traditional formula:

$$R_t^{index} = log(P_t^{index}) - log(P_{t-1}^{index})$$

Where,

P_t^{index} *is the price of a given index at time t*

P_{t-1}^{index} *is the price of a given index on the previous day*

R_t^{index} *is the daily returns on the index*

Chaudhuri and Ghosh (2016) method of calculating realized volatility was then applied by finding the annualized 20-day rolling standard deviation of daily returns of JFIN and JBIND Index over the sample period from June 1st 2017 to June 1st 2022.

In terms of independent variables, Krollner et al. (2010) survey suggested that lagged values of dependent variables, volatility in all share index and technical indicators are extensively used in the literature as the main predictors of volatility in stock returns. Therefore, similar to Hamid and Iqbal (2004), 1-day lagged values of the dependent variable will be used as an independent variable. Therefore, 1-day lagged realized volatility will be used as an independent variable in this project.

Since there is only one all-share index on the JSE which is the all-share index (ALSI), this will be the second independent variable. The volatility of ALSI was calculated in Excel using the same method which was applied in calculating the volatility of the JFIN and JBIND index. Technical indicators, on the other hand, is a broad term consisting of several indicators. To determine which technical indicators should be selected, this project considered four indicators suggested by Basak et al. (2019). The four indicators are listed and explained below:

i. Relative strength indicator (RSI)

This was introduced by Wilder (1978) and is one of the most popular technical indicators. The values for RSI were obtained directly from Bloomberg. The formula is given below:

$$RSI = 100 - \left[\frac{100}{1 + \frac{Average\ gain\ over\ 14\ days}{Average\ loss\ over\ 14\ days}} \right]$$

The intuition behind the RSI is that it helps traders identify entry and exit points, whereby an RSI of above 70 indicates a "sell" signal and an RSI of below 30 indicates a "buy" signal.

ii. Stochastic oscillator (SO)

Introduced by Lane (1984), similar to RSI, this indicates a buy and sell signal, where a SO >80 indicates a "sell" signal and SO <20 indicates a "buy" signal to traders. It is calculated as follows:

$$SO = \frac{Closing\ price - lowest\ low\ price\ in\ past\ 14\ days}{Highest\ high\ price\ in\ past\ 14\ days - lowest\ low\ price\ in\ past\ 14\ days}$$

iii. Williams percentage range (WPR)

Another technique which was introduced by Williams (1978), is also very similar to SO but the slight difference in WPR is that values range from −100 to 0 where values lower than −80 indicate a "buy" signal and values above −20 indicate a "sell" signal. WPR is calculated as follows,

$$WPR = -100 * \frac{Highest\ high\ price\ in\ past\ 14\ days - lowest\ low\ price\ in\ past\ 14\ days}{Highest\ high\ price\ in\ past\ 14\ days - lowest\ low\ price\ in\ past\ 14\ days}$$

iv. On balance volume (OBV)

This is whereby changes in volume have an impact on volatility in stock prices. OBV was introduced by Granville (2018). This indicator has also been supported in the South African context by the findings of Naik et al. (2018). Naik et al. (2018) identified trading volume on the JSE as one of the factors that can partially explain returns volatility on the JSE. OBV is calculated as follows:

$$OBV_{previous\ day} + OBV \begin{cases} actual\ volume\ at\ time, \\ \ if\ last\ price_t > last\ price_{t-1} \\ 0,\ if\ last\ price_t = last\ price_{t-1} \\ -actual\ volume\ at\ time\ t, \\ \ if\ last\ price_t > last\ price_{t-1} \end{cases}$$

The intuition behind OBV, is that volume traded increases when stock prices rise above the previous trading day. Conversely, volume traded falls when stock prices fall below the previous trading day.

Before specifying the models, it is important to determine if all of the independent variables considered (Lagged volatility, ALSI volatility, RSI, SO, WPR and OBV) display some degree of sensitivity to the output variable because including variables with very low sensitivity to dependent variable can lead to a poorer forecasting model Hamid (2004). JingTao and Tan (2001) also stated that the usage of all the independent variables might not improve forecasting ability due to very low sensitivity between the dependent variables and some independent variables. Following the procedure outlined by Hamid and Iqbal (2004), any independent variable which exhibits a correlation between −5% and 5% will be excluded.

Looking at the correlation matrix in Table 1, we observe a very low correlation between the dependent variable (realized volatility) and two independent variables namely, SO and WPR. SO and WPR variables also have a high correlation with the RSI variable, which can lead to a poorer model as a result of the high correlation between independent variables (JingTao and Tan, 2001). Consequently, SO and WPR will be excluded from the project. As a result, only 4 final independent variables will be used in the models: lagged realized volatility, RSI, ALSI volatility and OBV.

The descriptive table shown in Table 2 shows that the financial industry volatility is slightly lower than the basic material industries. Both industries exhibit leptokurtic distribution since the value of the kurtosis is >3, which is the maximum number beyond which a distribution stops following a normal distribution and displays asymmetric features whereby the effect from negative returns announcement is disproportionately greater than the effect of positive returns announcement (Babikir et al., 2012). This confirms Samouilhan and Shannon (2008) and Babikir et al. (2012) findings of the presence of leverage effect in South Africa, which motivates the need for nonlinear models such as Random Forest and ANN in order to predict volatility in stock market returns in South Africa.

3.1. Model Specifications

This section will first explain some important processes involved in machine learning techniques. The section will then describe the Random Forest and ANN model used in the project. This section will conclude by defining the evaluation metrics that will be used to compare the performance of ANN and Random Forest.

The independent variables in machine learning are known as features and the dependent variables are known as target variables. The original data is divided between two groups, in-sample data which is referred to as a training dataset and out-of-sample data

Table 1: Correlation matrix

5-year correlation matrix for JFINX							
Variable	**Realized volatility**	**One-day lagged volatility**	**RSI**	**ALSI Volatility**	**SO**	**WPR**	**OBV**
RV	1.000						
RV_t-1	0.992	1.000					
RSI_14d	−0.093	−0.075	1.000				
ALSI_Vol	0.887	0.876	−0.176	1.000			
SO	0.032	0.042	0.828	0.001	1.000		
WPR	−0.032	−0.042	−0.828	−0.001	−1.000	1.000	
OBV	−0.417	−0.415	0.201	−0.368	0.138	−0.138	1.000
5 years correlation matrix for JBIND							
Variable	**Realized Volatility**	**One-day lagged volatility**	**RSI**	**ALSI Volatility**	**SO**	**WPR**	**OBV**
RV	1.000						
RV_t-1	0.990	1.000					
RSI_14d	−0.138	−0.121	1.000				
ALSI_Vol	0.917	0.914	−0.166	1.000			
SO	0.034	0.050	0.841	0.034	1.000		
WPR	−0.034	−0.050	−0.841	−0.034	−1.000	1.000	
OBV	0.262	0.263	0.118	0.162	0.075	−0.075	1.000

Table 2: Descriptive table

JFINX descriptive analysis					
Variable	**Realized volatility**	**One-day lagged volatility**	**RSI**	**ALSI volatility**	**OBV**
Mean	9.336	9.323	50.873	6.911	1353826630
Standard deviation	5.719	5.711	10.920	3.878	1485354500
Kurtosis	14.529	14.634	0.040	22.356	−0.254
Skewness	3.262	3.274	−0.178	4.094	−0.769
Minimum	3.110	3.110	10.370	2.210	−2.764E+09
Maximum	45.230	45.230	78.050	33.620	4208628511
JBIND descriptive analysis					
Variable	**Realized volatility**	**One-day lagged volatility**	**RSI**	**ALSI volatility**	**OBV**
Mean	10.375	10.369	53.473	6.911	1389415416
Standard Deviation	5.475	5.475	11.933	3.878	668900972
Kurtosis	18.251	18.273	−0.338	22.356	−0.5614706
Skewness	3.656	3.660	−0.330	4.094	−0.6509538
Minimum	4.570	4.570	15.510	2.210	−387079906
Maximum	46.850	46.850	82.970	33.620	2553111542

which is known as a test dataset. Scikit-Learn software was used to split the dataset between training and tests for this project. Since the 80:20 ratio is a rule of thumb in most of the previous literature (Joseph, 2022), the 80:20 ratio was applied in this project as well, where 80% of the total sample was used for training and 20% of the total sample was used for testing. Thereby, the machine learning algorithm will learn from training data and generalize a model which will be used to make predictions using the testing dataset.

An important step in implementing the ANN and Random Forest is the selection of hyperparameters because the performance of the model is dependent on the hyperparameters used. Hyperparameters are the parameters in machine learning models that influence the learning process of the model (Claesen and De Moor, 2015).

The process of finding optimal hyperparameters is known as hyperparameter tuning. The importance of optimizing hyperparameters was highlighted in Claesen and De Moor (2015) paper, where the paper found that hyperparameter tuning optimizes the bias-variance trade-off which is prevalent in machine learning. The bias-variance trade-off states that a very complex model (with complex hyperparameters) might fit the training data very well, however, it might not perform well when dealing with testing data (data that the model has not seen before). This is known as an overfitting problem. On the other hand, an underfitting problem occurs when the model is too simple/general and it fails to capture sufficient information in the training data to accurately predict testing data (Claesen and De Moor, 2015). The traditional Grid Search method will be employed in this project Similar to Ding et al. (2008) paper, though this method suffers from the problem of dimensionality. Dimensionality is when the number of combinations increases exponentially with the number of hyperparameters (Liashchynskyi and Liashchynskyi, 2019). However, for ease of comparability of results with other papers, GridSearch will be used in this project.

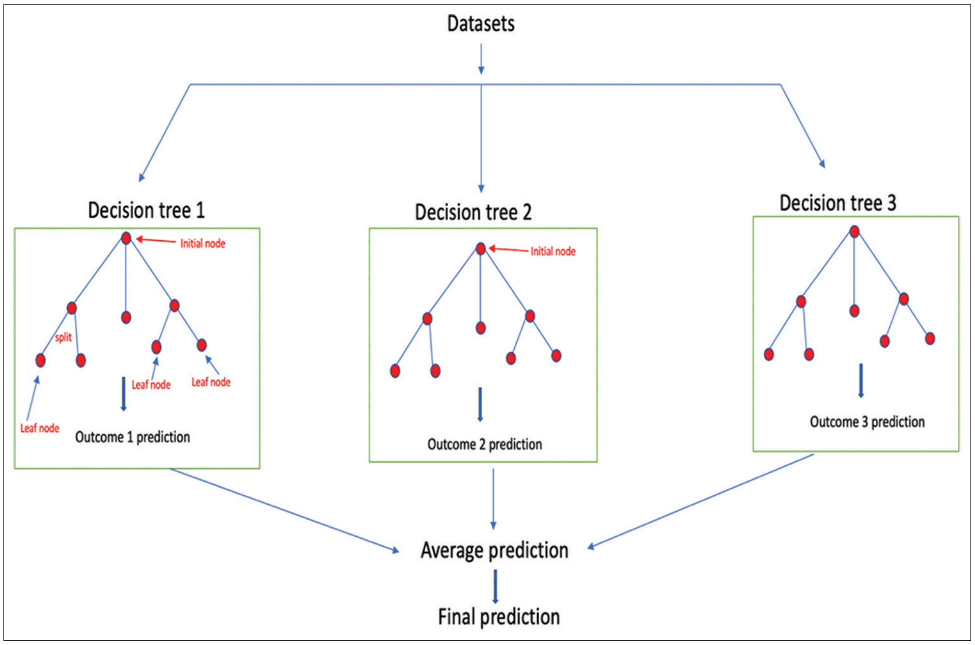
i. Random forest model

Random Forest is an ensemble machine-learning technique first introduced by Breiman (2001). The Random forest uses decision trees for forecasting. An example of a random forest model is illustrated in Graph 2, whereby we start with an initial dataset which will be "bootstrapped." Bootstrap aggregating, also known as bagging, is the process by which random samples are obtained from the initial full datasets for prediction before these random samples are replaced in the original dataset to build even more random samples for further predictions. This is one of the advantages of Random Forest because it reduces the model exposure to overfitting problems discussed earlier. Another benefit of using the Random Forest model is that it does not use all the features/independent variables in each decision tree. Instead, each decision tree will have a different combination of randomly selected independent variables leading to a lower variance and a more stable model (Khaidem, et al. 2016).

As shown in Graph 2, each decision tree begins with a root/initial node and continuously splits the trees based on the condition set by the root/initial node and succeeding nodes until the condition is met. Once the splitting stops, the model selects the leaf/last node which satisfies the final condition set by the model and an outcome prediction is obtained for decision tree 1. This process will be the same for the other decision trees in the "forest". Once the "ensemble" or the group of trees have grown in the forest, a predicted outcome is obtained for all the decision trees in the forest and the average forecast of the combined decision trees will result in the final output prediction (Breiman, 2001).

As explained before hyperparameters tuning is very important for machine learning models to improve forecasting ability. The optimal hyperparameters are summarized in Table 1 in the appendix for the different data sets used in this model. The

Graph 2: A simple illustration of random forest



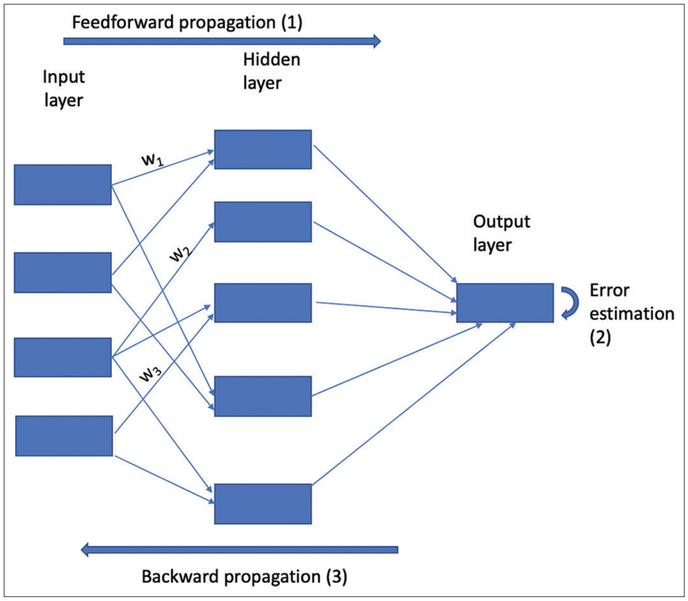
hyperparameter tuning process was done using the GridSearch method using Python programming language.

ii. Artificial neural network (ANN)

An artificial neural network is a computational learning algorithm inspired by the human brain. The model was developed to process information and identify patterns through neurons similar to human brains. This is done through data being processed from an initial layer and transferred between the different layers to obtain a final output layer. As illustrated in Graph 3, each layer contains a set number of neurons (For example, in Graph 3, the input layer contains 4 layers because there are 4 inputs). Neurons are the component inside the layers which learns from the relationship between the independent and dependent variable. As information is being processed and passed between the different layers, weight is calculated from an initial randomized weight and readjusted frequently when the model is being trained through a transfer function which is built into the model. The transfer functions transform the values obtained between the layers into a range of 1 to 0 or −1 to 1 before passing to the next layer. This process ensures the output levels are not extremely high and standardized (Hamid, 2004).

The weight reflects the input's influence on the output variable which is illustrated as w1 and w2 in Graph 1. The process begins with the initial layer, this initial layer contains the input variables values which are multiplied by the initial randomized weight. The value obtained from the calculation in the input layers will be passed to a hidden layer. The hidden layer is "invisible" because it cannot be accessed by the input or output layer. The hidden layer along with the initial layer helps in developing the trained model by identifying trends and relationships in the datasets. The number of hidden layers is at the discretion of the researcher; however, one hidden layer should suffice for a financial dataset as suggested by Hamid (2004). Once the hidden layer has learned from the training dataset, it passes these values to the single neuron's output layer by generating suitable weights and multiplying them by the values received from the input layer (Hamid, 2004).

Graph 3: A simple illustration of artificial neural network



Feedforward propagation shown in Graph 3 reflects the connectivity from the input layer to the output layer. Whilst Backward propagation (or backpropagation) is a training method. With backpropagation, once the output layer receives the values from the hidden layer, it multiplies by a random weight to determine an output. This output is then compared with the desired output level and the discrepancy between these two values reflects the estimation error. This error sends a signal into the output layer which is transferred back into the input layer following a reverse path of feedforward propagations. This process continues until the discrepancy between the desired output and output obtained by iterative forward propagations is optimized (Hamid, 2004). Regarding the ANN model used in this project, Chauduri and Ghosh (2016) paper model specifications were followed through the implementation of a feed-forward propagation ANN model with one input layer, one hidden layer and an output layer. The backpropagation method was used to train the model. Due to a lack of proper guidance in the literature, the number of neurons in the input layer and hidden layer was set equal to the number of inputs. While the output layer only has a single neuron.

Using the grid search process, the optimal hyperparameter for the ANN is in Appendix Table 2. The main takeaway is that the optimal hyperparameters selects "Stochastic Gradient descent" (SGD) as the optimizer instead of the standard Adam method which is frequently used as the optimizer in ANN. The use of SGD as an optimizer is supported by Zhou et al (2020), who found the SGD to outperform ADAM as an optimizer.

3.2. Evaluations Metrics

The 5 main evaluation methods which will be used to compare the different models are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination (R-squared). These values were calculated using the ScikitLearn in Python programming language. The formulas for each metric are given below:

MAE

$$\frac{1}{n}\sum_{i=1}^{n} |\, actual\ output_i - predicted\ output_i \,|$$

MSE

$$\frac{1}{n}\sum_{i=n}^{n}(actual\ output - predicted\ output)^2$$

RMSE

$$\frac{1}{n}\sum_{i=n}^{n}\sqrt{(actual\ output - predicted\ output)^2}$$

MAPE

$$\frac{1}{n}\sum_{i=n}^{n}\frac{|\,Actual\ output - predicted\ output\,|}{Actual\ output}$$

R-squared

$$\frac{\sum_{1}^{n}(predicted\ output - average\ output)^{2}}{\sum_{i}^{n}(actual\ output - average\ output)^{2}}$$

4. RESULTS AND ANALYSIS

In this section, results from the Random Forest and ANN model will be discussed. The COVID effect on the models will also be discussed. The main findings in this section show that Random Forest outperforms the ANN model in forecasting the realized volatility of the JFIN and JBIND Index. A total of 4 experiments were conducted for the full sample (For example, a random forest model will be run for JBIND and JFIN separately. Then, an ANN model will be run for JBIND and JFIN separately). These models were run on Python in Google Collab and the results are presented in the table as follows:

The out-of-sample data or the test data inputs were used to predict the output for each model and dataset. These predicted values were compared with the actual test outputs to measure the forecasting error in the model. The four first metrics (MAE, MAPE, MSE and RMSE), all measure the deviation of the predicted values from the actual values and a lower metrics value indicates a better model. A MAE, MAPE, MSE or RMSE of 0 indicates a perfect model with no deviation between the predicted and the actual output values. On the other hand, a higher R-squared is preferred because this implies that more variations in output can be explained by the model, whereby an R-squared of 100% indicates a perfect prediction.

According to the results in Table 3, the Random Forest model performs better in predicting the realized volatility of JFIN and JBIN Index relative to the ANN model concerning R-squared, considering the full sample columns only. The R-squared of at least 97% indicates that the Random Forest model was able to explain volatility in the two indices returns at a satisfactory level. Random forest is also better when it comes to all the other metrics for the full sample datasets. In terms of MAE, the Random forest is just slightly better than the ANN for both indices. The relatively lower RMSE and MAPE also show the Random Forest still has better predictive accuracy compared to the ANN model. Therefore, going back to the main aim of this project, which is to identify the best-performing model in the South African market context, it is clear from the results in Table 3 that the Random Forest model outperforms the ANN considering the full sample dataset.

This result is consistent with results from Hamid and Iqbal (2004) and Chauduri and Ghosh (2016), where the Random Forest model was the preferred model in forecasting stock returns volatility. The possible reasons for the poorer performance of the ANN can be attributed to the relatively smaller dataset (5 years) in this project. The ANN models usually perform better with a larger dataset (Picasso et al., 2019). Another possible reason for the relatively poorer performance of ANN relates to variable selection and hyperparameter tuning. Variable selection and hyperparameters tuning can always be modified to improve forecasting ability but it requires more training time and computational power. Another limitation of machine learning techniques such as ANN was raised by Hamid (2004). Hamid (2004) stated that it is difficult to break down the ANN model network and understand why the model is performing poorly. The introduction of other ANN models such as long short-term neural memory (also known as LSTM) might have produced a better model because those models can understand longer sequences of input unlike the feedforward model used in this project (Nelson, et al. 2017). This means that LSTM can better understand input variable values that are further into the past compared to feedforward models.

Moreover, concerning the ANN model, low and negative R-squared was observed for the JFIN and JBIND indices. Persson and Dabiri (2021) results for the ANN model also displayed negative R-squared. This indicates that the model does not explain sufficient information about the predicted values and the model is very poor in forecasting. On the other hand, despite the

Table 3: Results

Variables	Random forest model			ANN		
	Full sample	Pre-COVID period	Post-COVID period	Full sample	Pre-COVID period	Post-COVID period
	JFIN index	JFI index	JFIN index	JFIN index	JFIN index	JFIN index
MAE	0.438	0.303	0.438	0.601	0.948	0.717
MAPE	0.047	0.047	0.055	2.193	2.154	5.334
MSE	0.693	0.180	1.391	0.84	1.983	0.932
RMSE	0.832	0.424	1.180	0.918	1.408	0.966
R-squared	97.1%	95.2%	98.0%	15.7%	−98.3%	6.8%
Variables	**Full sample**	**Pre-COVID period**	**Post COVID period**	**Full sample**	**Pre-COVID period**	**Post- COVID period**
	JBIND index	**JBIND index**	**JBIND index**	**JBIND index**	**JBIND index**	**JBIND index**
MAE	0.438	0.318	0.438	0.771	1.016	0.608
MAPE	0.042	0.043	0.045	1.537	2.080	1.000
MSE	0.500	0.235	1.512	1.882	2.051	1.000
RMSE	0.707	0.485	1.230	1.372	1.432	1.000
R-squared	97.8%	88.1%	97.9%	−88.2%	−108.0%	0.0%

disappointing MAE, MAPE, MSE and RMSE values for the ANN model in this project in predicting volatility, these values are still better than the results found by D'Ecclesia and Clementi (2021).

4.1. COVID Effect Analysis

According to Heymans and Camara (2013), similar to developed economies, emerging markets have been impacted by major financial crises over the years from the 2007 to 2008 financial crises to the 2010–2011 European debt crises. Demirer et al. (2020) noted volatility tends to be higher during periods of major economic crises as observed during the 2008 financial crisis which might adversely impact the model's effectiveness. In this project, the full sample datasets were broken up into two subsamples pre-COVID and post-COVID to determine whether the pandemic shock which increased the volatility in both sectors significantly had an impact on the different models (Graph 1 in section 1). Therefore, the pre-COVID sample consisted of 717 daily volatility which ended on March 1st 2020. Whilst, the post-COVID sample consists of 588 observations from March 1 2020 to June 30 2022. 4 additional experiments were run using Python programming language for these new subsamples. A similar analysis was conducted in the Morema and Bonga-Bonga (2020) paper, which also included the COVID crises in their volatility model to capture the effects of those crises and evaluate the effectiveness of the volatility model during the crisis.

According to the results in pre-COVID and post-COVID columns in Table 3, it is ambiguous whether these machine learning models were affected by the COVID shock since in the Random Forest model, the model accuracy decreased because of higher metrics values of MAE, MAPE, MSE and RMSE in post-COVID relative to Pre-COVID sample. The only exception was the R-squared metrics, which improved for the post-COVID sample. On the other hand, the post-COVID sample data outperformed the pre-COVID Sample in the ANN model due to lower MAE, MSE and RMSE along with higher R-squared for the post-COVID sample. The only exception here was the metrics value of MAPE of the JFIN Index, which was higher. Consequently, there is uncertainty in terms of whether the COVID impact worsens machine learning forecasting ability since a poorer was observed for Random Forest for the post-COVID sample whilst the opposite effect occurred for the ANN model. The result of this project is contrary to Yong et al. (2021) paper, which found the COVID shock did not impact their E-GARCH model in forecasting stock market volatility in Malaysia.

5. CONCLUSION AND RECOMMENDATION

This project aimed to extend the literature on stock market returns volatility by implementing two non-linear machine learning models, namely Random Forest and ANN, instead of the traditional time series models such as GARCH which is the most popular model in the literature. The ANN performance was later compared to the Random Forest model to identify the top-performing model in the South African market context by forecasting realized volatility for the JFIN and JBIND Index. The results in this project show that the Random forest model outperforms the ANN model for the JSE Financial index and JSE Basic Materials index in terms of all metrics. Furthermore, the initial dataset was broken down into two sub-samples pre-COVID (before the March 1st 2020) and post-COVID (after the March 1st 2020). The models were re-run to detect if the COVID pandemic had a negative impact on machine learning model performance but no appropriate conclusion could be reached.

This research project may be useful for portfolio or asset managers who may evaluate the effectiveness of Random Forest and ANN in terms of volatility predictions for risk-management purposes. The paper may also guide policymakers to better identify which machine learning model is most appropriate in forecasting volatility in the South African market context.

On the other hand, it is important to note this project comes with some limitations such as a relatively smaller sample period of 5 years which might have limited the performance of ANN which usually requires a large dataset to produce an adequate (Picasso et al., 2019). Moreover, the independent variable selection and optimization of hyperparameters which is required in machine learning is not a well-defined process in the literature which results in subjective judgements. Improvement in machine learning models is not a perfect science and requires several trials and errors along with significant complexity and computational requirements. Therefore, future researchers can focus on understanding the input selection and hyperparameter optimization to produce the best model possible. Attention should also be paid to the ability to debug the performance of machine learning which is a daunting task.

REFERENCES

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H. (2018), State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11), e00938.

Ahmad, A.S., Hassan, M.Y., Abdullah, M.P., Rahman, H.A., Hussin, F., Abdullah, H., Saidur, R. (2014), A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renewable and Sustainable Energy Reviews, 33, 102-109.

Ahmad, M.W., Mourshed, M., Rezgui, Y. (2017), Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77-89.

Alberg, D., Shalit, H., Yosef, R. (2008), Estimating stock market volatility using asymmetric GARCH models. Applied Financial Economics, 18(15), 1201-1208.

Babikir, A., Gupta, R., Mwabutwa, C., Owusu-Sekyere, E. (2012), Structural breaks and GARCH models of stock return volatility: The case of South Africa. Economic Modelling, 29(6), 2435-2443.

Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R. (2015), Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20), 7046-7056.

Basak, S., Kar, S., Saha, S., Khaidem, L., Dey, S.R. (2019), Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552-567.

Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

Bollerslev, T., Mikkelsen, H.O. (1996), Modelling and pricing long memory in stock market volatility. Journal of Econometrics, 73(1), 151-184.

Bonga-Bonga, L. (2017), Assessing the readiness of the BRICS grouping for mutually beneficial financial integration. Review of Development Economics, 21(4), e204-e219.

Breiman, L. (2001), Random forests. Machine Learning, 45(1), 5-32.

Chaudhuri, T.D., Ghosh, I. (2016), Forecasting Volatility in the Indian Stock Market using Artificial Neural Network with Multiple Inputs and Outputs. arXiv Preprint arXiv:1604.05008.

Cifter, A. (2012), Volatility forecasting with asymmetric normal mixture GARCH model: Evidence from South Africa. Journal for Economic Forecasting, 2, 127-142.

Claesen, M., De Moor, B. (2015), Hyperparameter Search in Machine Learning. arXiv Preprint arXiv:1502.02127.

D'Ecclesia, R.L., Clementi, D. (2021), Volatility in the stock market: ANN versus parametric models. Annals of Operations Research, 299(1), 1101-1127.

Demirer, R., Gupta, R., Pierdzioch, C. (2020), Forecasting Realized Stock-market Volatility: Do Industry Returns Have Predictive Value? SSRN Journal, 2020. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3744537.

Ding, Y., Song, X., Zen, Y. (2008), Forecasting financial condition of Chinese listed companies based on support vector machine. Expert Systems with Applications, 34(4), 3081-3089.

Emenike, K.O. (2010), Modelling Stock Returns Volatility in Nigeria Using GARCH Models. Ebitimi Banigo Auditorium, University of Port Harcourt, Nigeria.

Engle, R.F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society, 50, 987-1007.

FTSE Russell, (2022), FTSE/JSE Top 40 Index. United Kingdom: FTSE Russell, p1-3.

Granville, J.E. (2018), Granvilles's New Key to Stock Market Profits. New Zealand: Papamoa Press.

Hamid, S.A. (2004), Primer on Using Neural Networks for Forecasting Market Variables. CFS Working Papers Series.

Hamid, S.A., Iqbal, Z. (2004), Using neural networks for forecasting volatility of S&P 500 Index futures prices. Journal of Business Research, 57(10), 1116-1125.

Heymans, A., Da Camara, R. (2013), Measuring spill-over effects of foreign markets on the JSE before, during and after international financial crises. South African Journal of Economic and Management Sciences, 16(4), 418-434.

Huang, W., Nakamori, Y., Wang, S.Y. (2005), Forecasting stock market movement direction with support vector machine. Computers and Operations Research, 32(10), 2513-2522.

JingTao, Y.A.O., Tan, C.L. (2001), Guidelines for Financial Forecasting with Neural Networks. In: International Conference on Neural Information Processing, Shanghai, China.

Joseph, V.R. (2022), Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15, 531-538.

Khaidem, L., Saha, S., Dey, S.R. (2016), Predicting the Direction of Stock Market Prices Using Random Forest. arXiv Preprint arXiv:1605.00003.

Krollner, B., Vanstone, B.J., Finnie, G.R. (2010), Financial time series forecasting with machine learning techniques: A survey. In: ESANN 2010, 18th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 28-30, 2010, Proceedings.

Kumar, M., Thenmozhi, M. (2006), Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest. In: Indian Institute of Capital Markets 9th Capital Markets Conference Paper.

Lane, G. (1984) Lanes Stochastics. Second Issue of Technical Analysis of Stocks and Commodities Magazine, p87-90.

Liashchynskyi, P., Liashchynskyi, P. (2019), Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. arXiv Preprint arXiv:1912.06059.

Luong, C., Dokuchaev, N. (2018), Forecasting of realised volatility with the random forests algorithm. Journal of Risk and Financial Management, 11(4), 61.

Maier, H.R., Dandy, G.C. (2000), Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environmental Modelling and Software, 15(1), 101-124.

Mashamba, T., Magweva, R. (2019), Dynamic volatility behaviour of stock markets in Southern Africa. Journal of Economic and Financial Sciences, 12(1), 1-8.

Morema, K., Bonga-Bonga, L. (2020), The impact of oil and gold price fluctuations on the South African equity market: Volatility spillovers and financial policy implications. Resources Policy, 68, 101740.

Naik, P.K., Gupta, R., Padhi, P. (2018), The relationship between stock market volatility and trading volume: Evidence from South Africa. The Journal of Developing Areas, 52(1), 99-114.

Nelson, D.M., Pereira, A.C., De Oliveira, R.A. (2017), Stock Market's Price Movement Prediction with LSTM Neural Networks. In: 2017 International Joint Conference on Neural Networks (IJCNN). United States: IEEE, p1419-1426.

Pati, P.C., Barai, P., Rajib, P. (2018), Forecasting stock market volatility and information content of implied volatility index. Applied Economics, 50(23), 2552-2568.

Persson, M., Dabiri, A. (2021), Comparing Machine Learning Models for Predicting Stock Market Volatility Using Social Media Sentiment: A Comparison of the Predictive Power of the Artificial Neural Network, Support Vector Machine and Decision Trees Models on Price Volatility Using Social Media Sentiment. Sweden: KTH Royal Institute of Technology.

Picasso, A., Merello, S., Ma, Y., Oneto, L., Cambria, E. (2019), Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications, 135, 60-70.

Poon, S.H., Granger, C.W. (2003), Forecasting volatility in financial markets: A review. Journal of Economic Literature, 41(2), 478-539.

Samouilhan, N.L., Shannon, G. (2008), Forecasting volatility on the JSE. Investment Analysts Journal, 37(67), 19-28.

Sevgen, E., Kocaman, S., Nefeslioglu, H.A., Gokceoglu, C. (2019), A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression, ANN and random forest. Sensors, 19(18), 3940.

Sharma, N., Juneja, A. (2017), Combining Random Forest Estimates Using LSboost for Stock Market Index Prediction. In: 2017 2nd International Conference for Convergence in Technology (I2CT). United States: IEEE, p1199-1202.

Tripathy, N., Garg, A. (2013), Forecasting stock market volatility: Evidence from six emerging markets. Journal of International Business and Economy, 14(2), 69-93.

Wilder, J.W. (1978), New Concepts in Technical Trading Systems. North Carolina: Trend Research Ltd.

Williams, L.R. (1978), The Secret of Selecting Stocks for Immediate and Substantial Gains. 1st ed. United Kingdom: Windsor Books.

World Bank. (n.d.), Market Capitalization of Listed Domestic Companies (Current US$)-South Africa. Available from: https://data.worldbank.org/indicator/cm.mkt.lcap.cd?locations=za

Yong, J.N.C., Ziaei, S.M., Szulczyk, K.R. (2021), The impact of COVID-19 pandemic on stock market return volatility: Evidence from Malaysia and Singapore. Asian Economic and Financial Review, 11(3), 191-204.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H. (2020), Towards theoretically understanding why SGD generalizes better than Adam in deep learning. Advances in Neural Information Processing Systems, 33, 21285-21296.