مجلة التجارة والتمويل

https://caf.journals.ekb.eg/

كلية التجارة – جامعة طنطا

العدد : الأول

مارس ٢٠٢٣

Hybrid deep learning and ARIMA model for prediction Egyptian Stock Exchange

د/ حنان خضاري مهدي محمود

مدرس الاحصاء التطبيقي بمعهد النيل العالي للعلوم التجارية

وتكنولوجيا الحاسب بالمنصورة

Dr. Hanan Khadari Mahdi Mahmoud

Lecturer of applied statistics at the Nile Higher
Institute of Commercial Sciences
And computer technology in Mansoura

Abstract

The Autoregressive Integrated Moving Average (ARIMA) is a flexible, good, and simple linear model for forecasting and time series analysis. Some time series forecasting researches also propose the Artificial Neural Network (ANN) model as a substitute nonlinear forecasting model. The ARIMA model is good at capturing linear patterns, but the ANN model is effective at capturing nonlinear patterns.

ANN and ARIMA models were significantly used in the prediction of the Egyptian stock exchange. Both ANN and ARIMA can also be merged as a hybrid model to capitalize on the capabilities of both models in linear and nonlinear modeling. We use the hybrid model in this research to merge ARIMA models and ANN model (the Deep Neural Network with numerous hidden layers) The Egyptian Stock Exchange is the actual dataset used.

The initial comparison made between the experimented prediction models for the time horizons of 10 days, 20 days, 30 days, 40 days, and 50 days in advance using the datasets in this work. To assess performance, statistical measurements for instance mean squared error (MSE), reveal that the DNN-ARIMA hybrid model outperforms non-hybrid models in predicting the Egyptian Stock Exchange and is particularly effective in improving prediction accuracy.

Keywords*:* ARIMA model, deep learning, forecasting, hybrid, Egyptian Stock Exchange, time series

1. Introduction

With fast economic growth in recent years, the number of financial operations has increased, and their variation trend has gotten more complicated. Understanding the patterns of financial operations and forecasting their growth and changes are major areas of study in both economic and academic circles. A financial data approximate prediction using one or more methodologies can assist in explaining the evolution and changes in the financial market at the macroscopic level and offer a foundation for-profit company and investors to make trading choices and plans at the microscopic level (Zhang et al. 2018), therefore helping them to increase profits. Predicting the trend of the development of financial data becomes particularly challenging because it contains complicated, partial, and unclear information (Oliveira and Meira 2006).

One of the most commonly used time series models is autoregressive integrated moving average (ARIMA). It is commonly used in time series forecasting. ARIMA is a collective term for numerous time series processes, including pure moving average (MA), pure autoregressive (AR), AR and MA composite (ARMA), and ARMA with differencing (ARIMA). Because the ARIMA model is a linear model, the data is supposed to follow a linear pattern. However, data in real-world issues does not necessarily follow a linear pattern. The linear approximation may not always be effective for forecasting with excellent performance (Miftahuddin. et al. 2017)

In time series forecasting, Artificial Neural Network (ANN) is another nonlinear model that has been examined and utilised (Zhang et al., 1998). The capacity of ANN models to do nonlinear modelling is its main

advantage. It is not essential to define a specific model form. The ANN model is created adaptively based on the data features. This data-driven method is suitable for many empirical data sets when no theoretical guidance is given to propose an acceptable data generation procedure (Suhartono et al. 2017).

Neural Networks (NN) (Tkáč and Verner 2016), Support Vector Machines (SVM) (Mishra and Padhy S 2019), metaheuristic algorithms (Sadaei et al. 2016), fuzzy logic networks (Esfahanipour and Aghamiri 2010), and other techniques were used in stock price prediction research. Furthermore, deep learning, the most recent trend in machine learning with the capacity to efficiently map the relations between input and output, has changed the modeling of prediction by allowing trading systems to predict stock prices (Shi et al. 2016) Artificial neural network (ANN) is another nonlinear model that has been used in time series forecasting and is widely examined (Bai et al. 2019). The ability of ANN models to perform nonlinear modeling is the main benefit. It is not required to define a specific model form. The ANN model is created adaptively according to the data attributes. This data-driven technique applies to many empirical sets of data where no theoretic guidance is given to propose a suitable process of data generation (Sheremetov. et al., 2014).

Yu and Yan (2019) studied the problems of stock price prediction by taking into account many stock indexes from around the world, including the Nikkei 225, S&P 500, and ChiNext, which were gathered from appropriate sources. Deep NN with lSTM was used by the researchers, and it outperformed ARIMA, deep multilayer perceptions, and others in terms of total prediction accuracy, MSE, MAPE, and correlation coefficients.

In Ho et al. (1988) studied three forecasting machine learning models, ARIMA, NN, and LSTM, and they were used to predict Bursa Malaysia closing stock prices from 1/2/2020 to 1/19/2021. Among these three models, The LSTM model has the best performance in Bursa Malaysia stock price prediction because it has the smallest MAPE and RMSE values.

Puspitasari et al. (2012) used an adaptive neuro-fuzzy inference system (ANFIS) and ARIMA models in another hybrid technique, and the results were excellent. Several studies demonstrated that the deep neural network (DNN) model, also known as the deep learning model, is an ANN with many hidden layers that have been commonly used for predicting tasks and have produced excellent results with high precision.

In the study of Qian et al (2017) the predicting precision of financial time series between traditional time series models ARIMA, and mainstream machine learning models including logistic regression, multiple-layer perceptron, support vector machine along with deep learning model denoising auto-encoder are compared through experiment on real data sets composed of three stock index data including Dow 30, S&P 500 and Nasdaq. The results show that machine learning as a modern method actually far surpasses traditional models in precision.

Khodaee et al. (2022) Constructed a hybrid Convolutional Neural Network and Long Short-Term Memory model forecasting turning points of stock price.

ARIMA and ANN can be merged to create a hybrid model for improved predicting performance. The hybrid technique merges the model

of ARIMA and the model of ANN to get the benefits of both models. Zhang (2003) did comparison research in a different study in which he merged an auto-regressive integrated moving average (ARIMA) model with an artificial neural network (ANN) to predict time series. The results demonstrated that the ANN was superior at processing nonlinear data and analyzing it.

In this research, we use the hybrid technique by merging the ARIMA model and the DNN for forecasting or predicting the Egyptian Stock Exchange. After that, the hybrid model (DNN-ARIMA) is compared to both the ARIMA and DNN models. In addition, the results of performance and accuracy of forecast are studied.

2. Time Series Models

The examination of observational data that occurs in a temporal sequence with a specified time interval known as time series analysis. Time series analysis is a method that used to analyze the data of time series to discover data patterns and features. Time series analysis is used for tasks of forecasting (Wilson 2015). Forecasting is estimating the future value based on previous data value. Many models of time series have evolved. Where ARIMA is the most widely known linear time series model.

Many studies have used linear models because they are simple to comprehend and explain, and they are simple to apply. However, real-world troubles are frequently more complicated, and the data includes nonlinear patterns. When modeling this type of data, sometimes the linear approximation performs poorly. Many studies have recently proposed the ANN model as a

substitutional to time series forecasting. The key strength of ANN is its ability in nonlinear modeling.

2.1. The ARIMA Model

ARIMA processes shape an adaptable class of regular linear processes with a wide variety of uses (William and Wei 2006) The ARIMA model general form is as follows (Teräsvirta et al. 1993)

$$(1-B)^d y_t = \mu + \frac{\emptyset(B)}{\theta(B)} a_t \qquad (1)$$

where

$$\emptyset\,(B) = 1 - \emptyset_1 B - \cdots . - \emptyset_p B^p$$

$$\emptyset\,(B) = 1 - \emptyset_1 B - \cdots . - \emptyset_q B^q$$

Y_t indicates the real value, B indicates the operator of the backshift, and a_t indicates the sequence of white noise which equals zero mean and constant ARIMA, $a_t \sim WN\,(0,\sigma^2)$. Model parameters are $\emptyset_i (i = 1,2,..,p)$, $\theta_j (j = 1,2,...,q)$. The differencing order is indicated by d_i. When $d=0$, ARMA is a subclass of the ARIMA model. The Box-Jenkins method is used in the construction of the ARIMA model. A Box-Jenkins process is an experimental approach required to identify ARIMA(p,d,q) model order, evaluate the model diagnostics, assess the parameters, forecast, and choose the best model.

2.2. The Deep Feedforward Networks

The model of neural network is effective at modelling data which has nonlinear pattern (Chen. et al. 2017) It is not necessary to make assumptions during the process of model-building. The data features are more important during the model formation. The network of single hidden layer feedforward is one of the most extensively utilized ANN models that is used in time series modelling and forecasting (Liu. et al. 2017).

DNN, or deep feedforward network, is a model of a feedforward neural network that contains multiple hidden layers. It is a fundamental model of deep learning (Zhao. et al. 2017). The feedforward network objective is to approximate some function $f*$. The feedforward network gets the optimal function approximation by learning the values of the parameters from a mapping $y = f(x;\theta)$, as seen in Figure 1. For instance, concerning the time series model, the following equation describes the relationship between the inputs ($Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$) and the output ($Y_t$) in a DNN model which has three hidden layers.

$$Y_t = \sum_{i=1}^{s} \propto_i g\left(\sum_{j=1}^{r} \beta_{ij} g\left(\sum_{k=1}^{q} \gamma_{jk} g\left(\sum_{l=1}^{p} \theta_{kl} Y_{t-l}\right)\right)\right) + \varepsilon_t \qquad (2)$$

where ε_t represents the error term, $\alpha_i (i = 1,2,\ldots, s)$, $\beta_{ij} (i = 1,2,\ldots, s; j = 1,2,\ldots,r)$, $\gamma_{jk} (j = 1,2,\ldots,r; k = 1,2,\ldots,q)$, and $\theta_{kl} (k = 1,2,\ldots,q; l = 1,2,\ldots, p)$ represent parameters of model which is known as connection weights, p indicates the input nodes number, and q,r,s indicates the nodes number in the first, second, and third hidden layers, respectively. The function of hidden layer activation is indicated by the function $g(.)$. As illustrated in Figure 2, the function of Tanh is used widely as

the activation function when it occurs in the range (−1,1). The definition of the function is as follows.

$$g(X) = \frac{e^{X} - e^{-X}}{e^{X} + e^{-X}} \qquad (3)$$

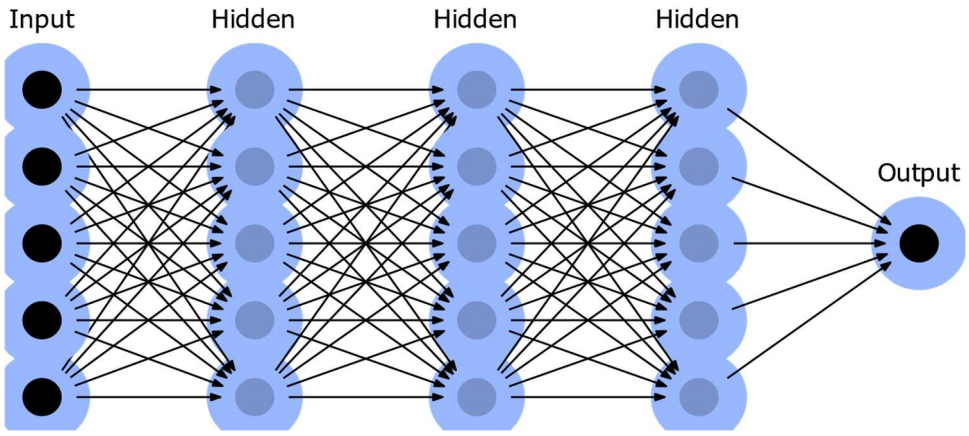


Fig. 1. The DNN neural architecture includes 5 nodes in input layer, 5 nodes in each hidden layer, and 1 node in output layer

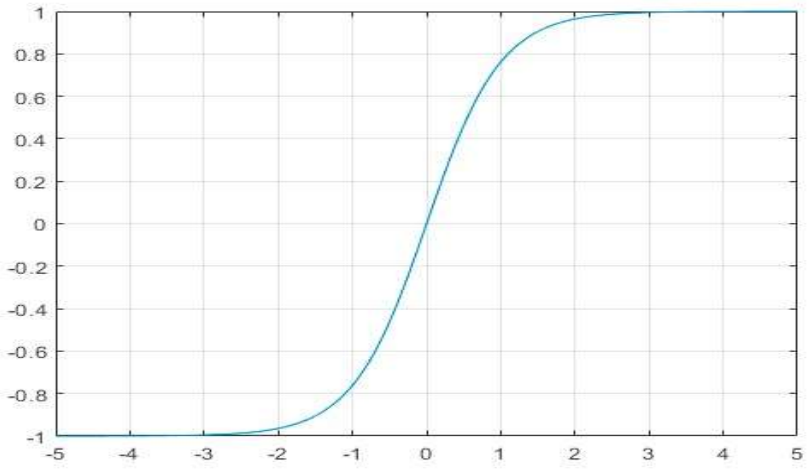


Fig. 2. Tanh function plot

2.3. The Hybrid Model

A hybrid model is a model that combines linear and nonlinear components. The ARIMA method is most effective in modelling data with a linear pattern, but the DNN model is most effective in modelling data with a nonlinear pattern. It is difficult to fully understand the features of the data in real-world problems. As a result, hybrid model, which is capable of both linear and nonlinear modelling, can be a viable strategy for data modelling. The ability of both linear and nonlinear modelling can capture the underlying patterns (Voyant. et al., 2017).

considered building a time series model with a linear component and a nonlinear component. That is,

$$Y_t = L_t + N_t \qquad (4)$$

where the linear component is represented by L_t and the nonlinear component is represented by N_t. Both components are acquired from the data estimation results. If the linear component is acquired from the ARIMA model, then residues from the ARIMA model. will only include the nonlinear component. If e_t represents the residual vector from the linear ARIMA model at the t-th interval, then

$$e_t = Y_t + \hat{L}_{t,} \qquad (5)$$

where $\hat{L}_t$ represents the value of forecast for time t from ARIMA model.

Residuals are essential for determining the linear model sufficiency. Then the residuals are treated as linear pattern-free. Because a linear model is insufficient if the linear correlation structures still exist in the residuals.

As a result, the residuals will just include the nonlinear pattern. The nonlinear relationships can be captured by modelling the residuals with DNN. The DNN model for residuals is presented as follows:

$$e_t = f(e_{t-1,}\, e_{t-2,} \dots, e_{t-n}) + \varepsilon_t, \qquad (6)$$

where DNN-determined nonlinear function is represented by f and the random error is represented by ε_t. If $\hat{N}_t$ represents the result of forecast from equation (6), the combined forecast is as follows.

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t. \qquad (7)$$

3. Dataset and Methodology

3.1. Dataset

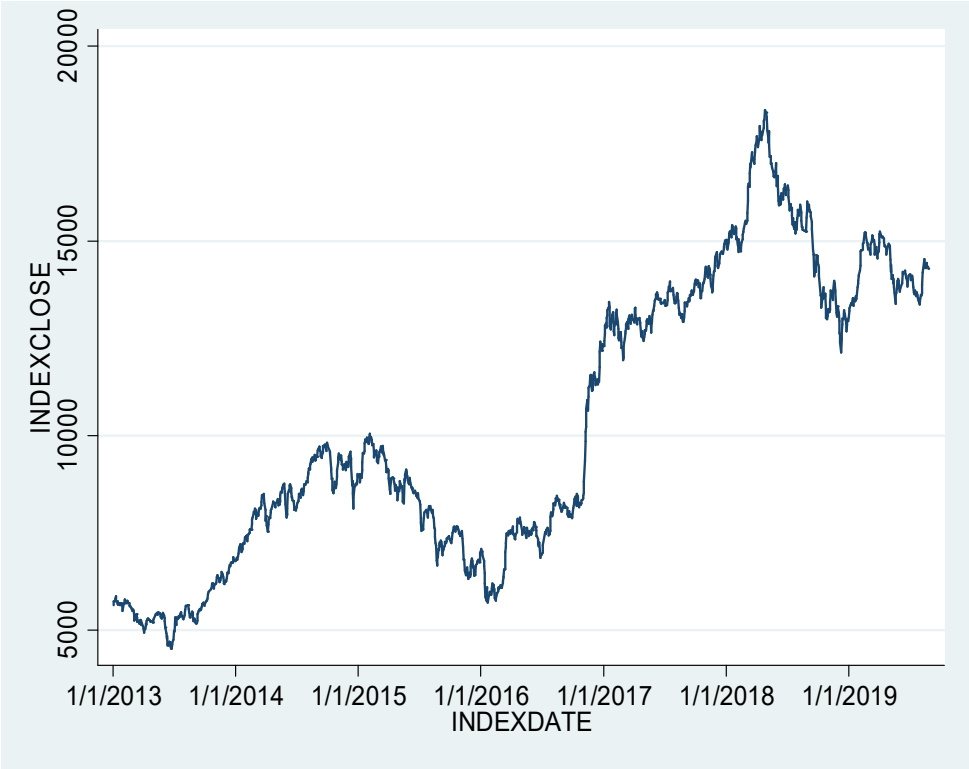
This research uses historical data of indicators from the Egyptian Stock Exchange (EGX) to construct a high-accuracy prediction model. The data for this study were obtained from Egypt for Information Dissemination (EGID), a governmental organization that gives data to EGX. The data include six indices of the stock market; for instance, the EGX-30 index local currency is utilized for interest calculations and is denominated in US dollars. It assesses the top 30 companies in terms of activity and liquidity.

The period of collecting data in this study begins with the EGID registration of these data from 1/2/2013 to 8/27/2019. (1617 views). The choice of data range mainly depend on two factors. One factor is that models need enough amount of training data. The other factor is that we hope to use latest data, i.e. data from recent years The data is categorized into eight categories: date, code, high price, open price, close price, low price, value,

and volume. As indicated in Table 4 and the Time series plot of the data series in Figure4, this study utilizes close price as goal feature.

Table (4): Description of used parameters

Parameter	Description
Open price	The Price in the beginning of daily dealing
High price	The highest price reached by the end of daily dealing
Low price	The lowest price reached by the end of daily dealing
Close price	The price in the end of daily dealing



3.2. Methodology

To build the model and forecast the data, we divided the data into two parts: in-sample and out-of-sample. The data utilized for modelling comes from a single sample. However, the forecast is based on out-of-sample data. The first 1000 data points are used as in-sample data. The remainder of the data is used as out-of-sample data. We employ root mean squared error prediction (RMSEP) and mean square error (MSE) to calculate prediction accuracy. The equation [8] is used to calculate RMSEP.

$$RMSEP = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(Y_{n+1} - \hat{Y}_n(l)^2} \qquad (8)$$

where L indicates the out-of-sample size, Y_{n+l} indicates the l-th actual value of out-of-sample data, and $\hat{Y}_n(l)$ indicates the l-th forecast.

4. Results and Discussion

In the first step, we use the Box-Jenkins approach to construct the ARIMA model of the Egyptian Stock Exchange. Orange is used to implement the model building. To choose the optimum model, we employ a backward elimination approach. A subset ARIMA model with zero mean, which is ARIMA ([1,2,3],1,[1,2]), is the best model for the Egyptian Stock Exchange as this study indicated.

The ARIMA model is then used to build the DNN model. A nonlinearity test, known as the Terasvirta test (Qin. et al. 2017) , is performed beforehand, and it reveals that the data has a substantial nonlinear pattern.

In both the DNN model and the DNN-ARIMA hybrid model, we employ the lag of AR components in the ARIMA([1,2,3],1,[1,2]) model as the input layer. Orange is used to implement DNN modelling. We model using varying numbers of nodes in the hidden layer, ranging from 1 to 10. The superior model is then the one with the lowest RMSEP in the out-of-sample forecast. According to the results, the superior pure DNN model has three nodes in each hidden layer, whereas the superior hybrid model has four nodes in every hidden layer.

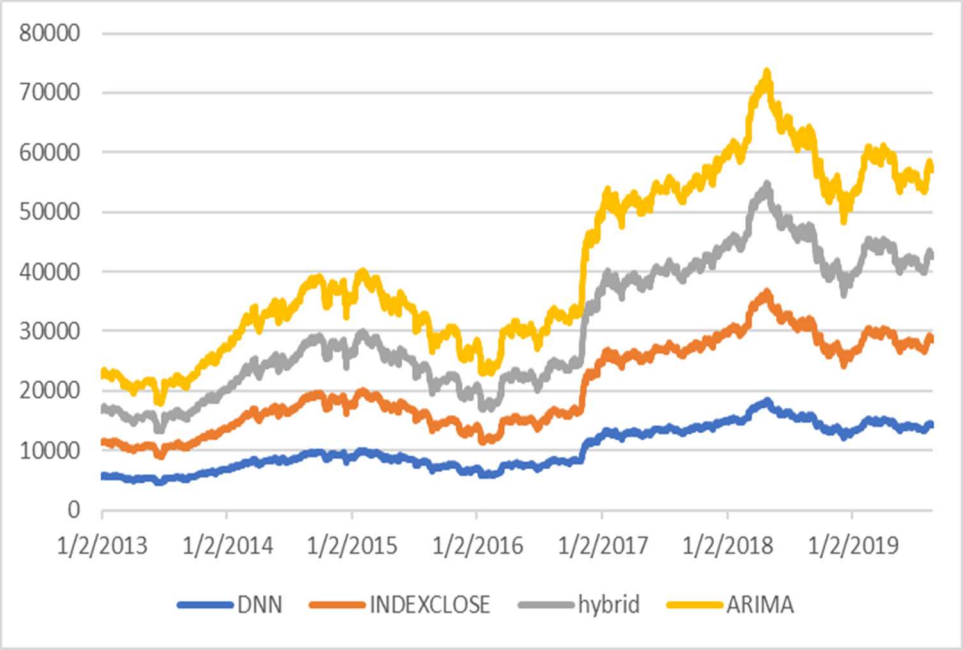


Fig. 5. A comparison of forecast between DNN, ARIMA, and hybrid DNN-ARIMA

Table 1. RMSE comparison between DNN, ARIMA, and DNN-ARIMA hybrid models in different forecast horizons

Forecast Horizon	DNN	ARIMA	Hybrid
10	0.055	0.058	0.052
20	0.114	0.134	0.078
30	0.201	0.217	0.137
40	0.109	0.218	0.015
50	0.239	0.301	0.095

Figure 5 shows that the predictions of the three models may match the real data pattern. However, the ARIMA model does not perform as well as the other two models. The hybrid model and the DNN model perform better in terms of forecasting. Clearly, the hybrid model forecast corrects the DNN model error forecast and enhances the accuracy of forecast. Table 1 indicates that the hybrid forecast model has the minimum RMSEP across all forecast horizons. The hybrid model can minimize the RMSEP of the DNN model by 42.50% in a 250-step at the beginning of forecasting. The hybrid model improves the accuracy of forecast of ARIMA model by reducing RMSEP by 36.14%. These findings suggest that the hybrid model is effective. The hybrid model also improves the ARIMA model's forecast accuracy by reducing RMSEP by 36.14% The hybrid model can greatly increase the accuracy of forecast of both the DNN and ARIMA models as the results indicated. Then, the capacity to capture both linear and nonlinear patterns becomes better.

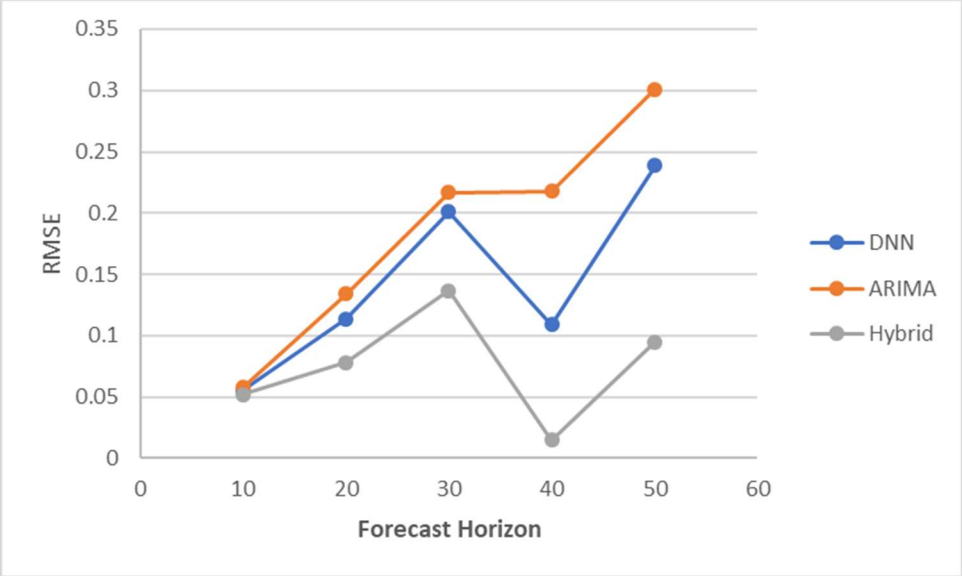


Fig. 6. Plot of RMSEP in several forecast horizons

Figure 6 illustrates how the RMSEP of all models change over various forecast horizons. Consider solely the hybrid model. The 50-step forecast is the best forecast performance. The RMSEP grow in the forecast with 10-step but decline in the forecast with 20-step before increasing again in the forecast with 30-step. In the forecast with 40-step, the declining trend continues. It demonstrates that the RMSEP does not increase monotonically while the forecast horizon becomes bigger. The RMSEP difference between the forecast with 10-step and the forecast with 50-step is 0.04208, but this difference is not adequately important. As a result, the Egyptian Stock Exchange's prediction with 50 steps (approximately 17 seconds) ahead is still rather accurate.

According to (Khodaee et al. 2022) the forecast performance of the model is also affected by the forecast horizon, while the best model

changing as the forecast horizon changes. When compared to our results, it is clear that prediction performance is steady in the order, when the hybrid model outperforms the others across all forecast horizons. Because we only have three models in our comparison, it may be more interesting to add additional models to compare and find out the performance differences existing which depend on the forecast horizon.

5. Conclusions and Future Works

The DNN and ARIMA models are effective at forecasting the Egyptian Stock Exchange. Their forecast can track the real data pattern. The hybrid model outperforms the other two models by having the lowest RMSEP. It demonstrates that the hybrid method has the possibility to greatly improve the accuracy of forecast of both the DNN and ARIMA models. We estimate that this methodology is a good method for enhancing forecast performance, particularly in the Egyptian Stock Exchange study. In addition, many different combinations of linear and nonlinear models can be developed in future research.

References

1) Bai Y, Jin X, Wang X, Su T, Kong J, Lu Y. Compound Autoregressive Network for Prediction of Multivariate Time Series. Complexity. 2019;2019:1-11.
2) Chen Y, He K, Tso GK. Forecasting crude oil prices: a deep learning based model. Procedia computer science. 2017;122:300-7.
3) Esfahanipour A, Aghamiri W. Adapted Neuro-Fuzzy Inference System on indirect approach TSK fuzzy rule base for stock market analysis. Expert Systems with Applications. 2010;37(7):4742-8.
4) Ho, M. K., Hazlina Darman, and Sarah Musa. "Stock price prediction using ARIMA, neural network and LSTM models." Journal of Physics: Conference Series. Vol. 1988. No. 1. IOP Publishing, 2021.
5) Khodaee, Pouya, Akbar Esfahanipour, and Hassan Mehtari Taheri. "Forecasting turning points in stock D15price by applying a novel hybrid CNN-LSTM-ResNet model fed by 2D segmented images." Engineering Applications of Artificial Intelligence 116 (2022): 105464.
6) Liu L, Chen R-C. A novel passenger flow prediction model using deep learning methods. Transportation Research Part C: Emerging Technologies. 2017;84:74-91.
7) Miftahuddin, Helida D, Sofyan H. Periodicity analysis of tourist arrivals to Banda Aceh using smoothing SARIMA approach. AIP Conference Proceedings: Author(s); 2017.
8) Mishra S, Padhy S. An efficient portfolio construction model using stock price predicted by support vector regression. The North American Journal of Economics and Finance. 2019;50:101027.

9) Oliveira ALI, Meira SRL. Detecting novelties in time series through neural networks forecasting with robust confidence intervals. Neurocomputing. 2006;70(1-3):79-92.

10) Puspita+D18sari I, Suhartono, Akbar MS, Lee MH. Two-level seasonal model based on hybrid ARIMA-ANFIS for forecasting short-term electricity load in Indonesia. 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE); 2012/09: IEEE; 2012.

11) Qian, Xin-Yao, and Shan Gao. "Financial series prediction: Comparison between precision of time series models and machine learning methods." arXiv preprint arXiv:1706.00948 (2017): 1-9.

12) Qin M, Li Z, Du Z. Red tide time series forecasting by combining ARIMA and deep belief network. Knowledge-Based Systems. 2017;125:39-52.

13) Sadaei HJ, Enayatifar R, Lee MH, Mahmud M. A hybrid model based on differential fuzzy logic relationships and imperialist competitive algorithm for stock market forecasting. Applied Soft Computing. 2016;40:132-49.

14) Sheremetov L, Cosultchi A, Martínez-Muñoz J, Gonzalez-Sánchez A, Jiménez-Aquino MA. Data-driven forecasting of naturally fractured reservoirs based on nonlinear autoregressive neural networks with exogenous input. Journal of Petroleum Science and Engineering. 2014;123:106-19.

15) Shi L, Teng Z, Wang L, Zhang Y, Binder A. DeepClue: visual interpretation of text-based deep stock prediction. IEEE Transactions on Knowledge and Data Engineering. 2018;31(6):1094-108.

16) Suhartono, Saputri PD, Amalia FF, Prastyo DD, Ulama BSS. Model Selection in Feedforward Neural Networks for Forecasting Inflow and

Outflow in Indonesia. Communications in Computer and Information Science: Springer Singapore; 2017. p. 95-105.

17) Teräsvirta T, Lin C-F, Granger CWJ. POWER OF THE NEURAL NETWORK LINEARITY TEST. Journal of Time Series Analysis. 1993;14(2):209-20.

18) Tkáč M, Verner R. Artificial neural networks in business: Two decades of research. Applied Soft Computing. 2016;38:788-804.

19) Voyant C, Notton G, Kalogirou S, Nivet M-L, Paoli C, Motte F, et al. Machine learning methods for solar radiation forecasting: A review. Renewable Energy. 2017;105:569-82.

20) William W, Wei S. Time series analysis: univariate and multivariate methods. USA, Pearson Addison Wesley, Segunda edicion Cap. 2006;10:212-35.

21) Wilson GT. Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. Journal of Time Series Analysis. 2016;37(5):709-11.

22) Yu P, Yan X. Stock price prediction based on deep neural networks. Neural Computing and Applications. 2019;32(6):1609-28.

23) Zhang G, Eddy Patuwo B, Y. Hu M. Forecasting with artificial neural networks. International Journal of Forecasting. 1998;14(1):35-62.

24) Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159-75.

25) Zhang L, Wang F, Xu B, Chi W, Wang Q, Sun T. Prediction of stock prices based on LM-BP neural network and the estimation of overfitting point by RDCI. Neural Computing and Applications. 2018;30(5):1425-44.

26) Zhao Y, Li J, Yu L. A deep learning ensemble approach for crude oil price forecasting. Energy Economics. 2017;66:9-16.