Optimizing Lags and Hidden Layers in Hybrid Models for Forecasting Stock Return

Nuttaphat Sukchitt[1], Manad Khamkong[2,*], Lampang Saechan[2], Napon Hongsakulvasu[3]

[1]Program in Applied Statistics, Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

[2]Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

[3]Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand

Received 26 August 2024; received in revised form 24 October 2024; accepted 25 October 2024

DOI: https://doi.org/10.46604/aiti.2024.14192

Abstract

This study aims to minimize the root mean square error for stock return by optimizing lags and hidden layers in a hybrid model. The model combines the autoregressive integrated moving average with the exogenous variables model as linear components. The residuals derived from linear components are used in artificial neural networks and Elman recurrent neural networks as non-linear components. A key feature of this approach is optimizing the selection of hidden layers and lags within the neural network, improving forecasting accuracy. The minimum mean square error forecast expression is derived, and the model is tested on stock price data during the COVID-19 period, marked by significant market shocks. The root mean square error results for the proposed model, traditional hybrid model, and traditional time series model range from 0.0004 to 0.01, 0.0006 to 0.01, and 0.006 to 0.03, respectively. The results show that the proposed model outperforms both traditional models.

Keywords: econometrics, hybrid model, ANN, ARIMAX, ERNN

1. Introduction

In recent years, the financial industry has recognized the paramount importance of forecasting stock market trends. However, the task of predicting stock market movements has become increasingly complex due to the influence of exogenous variables. External events, ranging from geopolitical developments to unforeseen global crises, introduce significant volatility and unpredictability, presenting a formidable challenge to traditional forecasting models.

In response to these challenges, this paper introduces an innovative approach that amalgamates time series analysis with neural network methodologies. By hybridizing these techniques, the study aims to enhance the predictive accuracy of stock market forecasts in the presence of exogenous variables. Central to the approach, the application of the autoregressive integrated moving average (ARIMA) model symbolizes a cornerstone of time series analysis known for its efficacy in capturing temporal dynamics in financial data. To account for the impact of external events, the ARIMA model is extended to an autoregressive integrated moving average with exogenous variables (ARIMAX) framework, which incorporates exogenous variables into the forecasting equation, thereby enriching contextual understanding and predictive capability of the model [1].

Furthermore, the study delves into the realm of neural networks, exploring two pivotal architectures: the artificial neural network (ANN) and the Elman recurrent neural network (ERNN) [2]. These models are celebrated for their ability to learn complex patterns and dependencies from data, enabling these models to render them ideally suitable for forecasting tasks

* Corresponding author. E-mail address: manad.k@cmu.ac.th

English language proofreader: Chih-Wei Chang

where non-linear relationships are prevalent. The research investigates further by optimizing these neural networks, focusing on the identification of the optimal number of lags and the determination of the appropriate number of hidden layers. This optimization process is crucial for developing a hybrid model that synergizes the strengths of time series analysis and neural network architectures, proffering superior forecasting performance by comparing minimum mean square error (MMSE) between the models.

This paper is organized into five main sections. Section 2 is a literature review. Section 3 details the methodology, including descriptions of the data, the experimental environment, and statistical methods. Section 4 reveals the empirical results of each model. The last section, Section 5, concludes with remarks that reflect on the findings and highlight the contributions of the study.

2. Literature Reviews

The application of forecasting models is a crucial aspect of decision-making in diverse fields such as public health, finance, energy, agriculture, and electricity. This section discusses several studies that have employed different forecasting methods to address specific challenges in these areas. The review provides insights into the performance and applicability of various forecasting models including a hybrid ARIMA model and neural network model, in different contexts.

Studies on electrical consumption forecasting have highlighted the effectiveness of hybrid approaches combining traditional time series models with advanced machine learning techniques. Almaleck et al. [3] using a hybrid method combining ANN and ARIMAX models, achieved superior performance in 24-hour-ahead forecasting for sports venues with a mean absolute percentage error (MAPE) of around 9%, outperforming standard machine learning techniques. Pierre et al. [4] employed a combination of ARIMA for trend modeling and deep learning methods (long short-term memory (LSTM) and Gated recurrent unit (GRU)) for capturing fluctuations. Their ARIMA-LSTM hybrid approach outperformed standalone models in predicting peak energy consumption, reaching the root mean square error (RMSE) of 7.35.

Management of natural resources has made significant strides in the use of hybrid forecasting models in recent years. Xu et al. [5] demonstrated that a hybrid ARIMA-LSTM model, based on the standardized precipitation evapotranspiration index, outperforms the conventional ARIMA model, attaining the highest prediction accuracy at 6-month, 12-month, and 24-month scales, indicating the suitability for the forecasting of long-term drought in China. Azad et al. [6] a seasonal autoregressive integrated moving average (SARIMA)-ANN model was applied to predict water levels in India's Red Hills Reservoir, and it was confirmed that the hybrid model outperformed the SARIMA model. In Thailand, Nualtong et al. [7] demonstrated that a SARIMA-ANN model yielded better results than either the SARIMA or standalone ANN model, especially during the wet season.

Furthermore, a study by Shahriar et al. [8] on PM2.5 forecasting in Bangladesh demonstrated that hybrid models such as ARIMA-ANN significantly outperform ARIMA by offering substantial improvements in prediction accuracy. These cases collectively highlight the robust capability of hybrid models to significantly enhance predictive outcomes in managing natural resources, surpassing traditional forecasting approaches. Recent studies highlight the superiority of hybrid forecasting models over conventional methods in various sectors. Fan et al. [9] demonstrated that a hybrid ARIMA-LSTM model significantly improved production forecasting by incorporating the effects of manual operations outperforming the traditional ARIMA model. Similarly, Wang et al. [10] showed that an ARIMA-ERNN model for predicting pertussis in China provided lower error rates than its ARIMA model.

In the financial sector, hybrid forecasting models have demonstrated marked superiority over traditional models in predicting stock market trends. For instance, Shetty and Ismail [2] propose a hybrid non-stationary model using ERNN to predict stock market price indices. This model combines linear and non-linear structures to capture market dynamics

effectively. Besides, this study demonstrates superior forecasting accuracy, emphasizing the potential of neural network-based models in enhancing financial market predictions. Furthermore, Sharma et al. [11] illustrated that integrating LSTM and ARIMAX into forecasting improves accuracy beyond the conventional ARIMA model, which adeptly accommodates external market influences.

Lv et al. [12] advanced this research by integrating complete ensemble empirical mode decomposition with adaptive noise and a hybrid model combining the autoregressive moving average (ARMA) and LSTM models. This approach effectively decomposes the stock index into intrinsic mode functions and applies the ARMA model to stationary series and LSTM to unstable series, refining the prediction process through time series decomposition, and the results exhibit that the prediction of the proposed model is closer to the real value than that of a standalone model like ARIMA, LSTM, GRU models. Alshawarbeh et al. [13] aimed to predict stock market indices using an ARIMA-ANN hybrid model, combining ARIMA and ANN to address the volatility and noise in financial data. By applying this approach to the Nasdaq, Nikkei, and CAC 40 indices, the study finds that the ARIMA-ANN model renders more accurate forecasts than traditional ARIMA and conventional ANN models. Lastly, Singh et al. [14] emphasized the efficacy of combining LSTM methods with ARIMA for forecasting stock prices with high precision, highlighting the robustness of hybrid models in financial forecasting.

The literature review highlights the diverse applications of forecasting models in various fields. These studies collectively provided valuable insights into the evolving landscape of forecasting methods and their impact on decision-making in different sectors.

3. Methodology

The methodology of this study is structured to integrate both linear and nonlinear modeling approaches to enhance the accuracy of stock market forecasting. By focusing on optimizing lags and hidden layers, the hybrid model combines ARIMAX as a linear component and neural networks (ANN/ERNN) as nonlinear components. This section details the experimental environment, the data used, and the statistical methods applied to test the efficacy of the hybrid model.

3.1. Experimental environment

All experiments were conducted on a Microsoft Surface Pro (5th Generation) equipped with an Intel Core i5-7200U processor, 8 GB of RAM, and a 256 GB SSD. The device operates on Windows 10 Pro. The software environment comprises R version 4.4.1, which runs on RStudio version 2024.04.2+764. The instruments utilized in this section are listed as follows. R packages were employed to conduct the analysis. The tseries package was used for time series modeling and statistical tests. The neural networks in R using the Stuttgart Neural Network Simulator (RSSNS) was utilized to implement ERNN, and the neuralnet package was applied to implement ANN. To ensure consistency and reproducibility of the neural network models, a fixed random seed (set.seed(51237)) was used in all relevant experiments.

3.2. Data description

This research considers secondary data, including the Stock Exchange of Thailand (SET) index across eight industries, as shown in Table 1. These industries encompass the Agro and Food Industry (SETA), Consumer Products (SETC), Financials (SETF), Industrials (SETI), Property and Construction (SETP), Resources (SETR), Services (SETS), and Technology (SETT) [15].

Table 1 The list of industry groups and sectors in the Stock Exchange of Thailand (SET)

Industry	Abbreviation	Sector
Agro and Food Industry	SETA	Agribusiness, Food and Beverage
Consumer Products	SETC	Fashion, Home and Office Products, Personal Products and Pharmaceuticals

Table 1 The list of industry groups and sectors in the Stock Exchange of Thailand (SET) (continued)

Industry	Abbreviation	Sector
Financials	SETF	Banking, Finance and Securities, Insurance
Industrials	SETI	Automotive, Industrial Materials and Machinery, Paper and Printing Materials, Petrochemicals and Chemicals, Packaging, Steel
Property and Construction	SETP	Construction Materials, Construction Services, Property Fund and Real Estate Investment Trusts, Property Development
Resources	SETR	Energy and Utilities, Mining
Services	SETS	Commerce, Health Care Services, Media and Publishing, Professional Services, Tourism and Leisure, Transportation and Logistics
Technology	SETT	Electronic Components, Information and Communication Technology

The study also examined the Google Trends (GGT) index and the exchange rate between the Thai Baht (THB) and the US Dollar. It should be noted that the idea of this section is oriented by Napon's work [16]. The collected data was divided into four periods. The first period is from 2 January 2020 to 1 June 2020. Regarding the second period, it is from 2 November 2020 to 1 March 2021. The third period started from 1 March 2021 to 15 June 2021, and the fourth period is from 16 June 2021 to 30 December 2021. All variables will be converted to the growth rate and return by using the conventional formula.

$$R_t = \ln\left(\frac{p_t}{p_{t-1}}\right) \tag{1}$$

where R_t is the return of the index, p_t is the data for today, p_{t-1} is the data for yesterday. To ensure the robustness of the analysis, all variables underwent stationarity testing utilizing the Phillips-Perron (PP) test. The PP test, introduced by Phillips and Perron (1988), is used to determine whether a time series is stationary. The null hypothesis (H_0) of the PP test states that the time series has a unit root, implying it is non-stationary. The alternative hypothesis (H_1) states that the time series is stationary. Concerning this study, a significant threshold was set at a p-value of less than 0.05, ensuring that the variables are appropriately stationary for accurate modeling and analysis [17].

3.3. Cross-correlation function (CCF)

To detect the relationship between the two series, a cross-correlation function (CCF) is conducted to verify the relationships of the series [18]. The CCF is defined as follows:

$$CCF_{XY}(k) = \frac{C_{XY}(k)}{\sqrt{C_{XX}(0)C_{YY}(0)}} \tag{2}$$

$$C_{XX}(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-k}\left(x_t - \bar{x}\right)\left(y_{t+k} - \bar{y}\right), \ k = 0,1,\ldots,n-1 \\ \frac{1}{n}\sum_{t=1-k}^{n-k}\left(x_t - \bar{x}\right)\left(y_{t+k} - \bar{y}\right), \ k = -1,-2,\ldots,-(n-1) \end{cases} \tag{3}$$

where $C_{XX}(0)$ and $C_{YY}(0)$ are the sample variances of $\{X_t\}$ and $\{Y_t\}$. The CCF calculates the linear correlation between the series. This study focuses on lag 0 to determine the cross-correlation between the series.

3.4. Hybrid modeling

A hybrid model separating linear and non-linear components is a sophisticated method used to capture complex relationships in a dataset by individually addressing the linear and non-linear aspects. This approach enhances the ability to handle data nuances by combining the strengths of both linear and non-linear techniques [19]. The model was expressed as:

$$y_t = L_t + N_t \tag{4}$$

where y_t is the original time series, L_t is the linear term, and N_t is the non-linear term. The linear component is estimated using the ARIMAX model, and the residuals are obtained from this model. Given that the linear term is obtained from the following formula, the residual series N_t from the ARIMAX model is expressed as follows:

$$N_t = y_t - \hat{L}_t \tag{5}$$

where $\hat{L}_t$ denotes the forecasting value for time t of the time series y_t by ARIMAX, N_t represents the non-linear term of the model. Subsequently, the neural network is used to estimate N_t with m input nodes, the neural network (ANN/ERNN) model for the residuals can be formulated as follows:

$$N_t = f\left(N_{t-1}, N_{t-2}, \ldots, N_{t-m}\right) + \varepsilon_t^* \tag{6}$$

where $f(.)$ represents a non-linear function determined by the neural network (ANN/ERNN), ε_t^* is the residual.

3.5. Auto regressive integrated moving average with exogenous variables

The ARIMAX model, with an exogenous variable as an additional predictor, can be used to predict the demand for a product based on its demand and some external factors such as advertising or competitors' prices [20]. The ARIMAX model can be expressed as follows:

$$\nabla^d L_t = C_t + \sum_{i=1}^{p} \phi_i \nabla^d L_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varphi X_t + \varepsilon_t \tag{7}$$

where ∇^d is the degree of differencing, L_t is the time series data at time t, C_t is a constant value, ε_t is a residual, L_{t-i} is the autoregressive, ε_{t-j} is the moving average, X_t is the exogenous variable, ϕ_i, θ_j, and φ are the coefficients of L_{t-i}, ε_{t-j}, and X_t, respectively.

3.6. Artificial neural network (ANN)

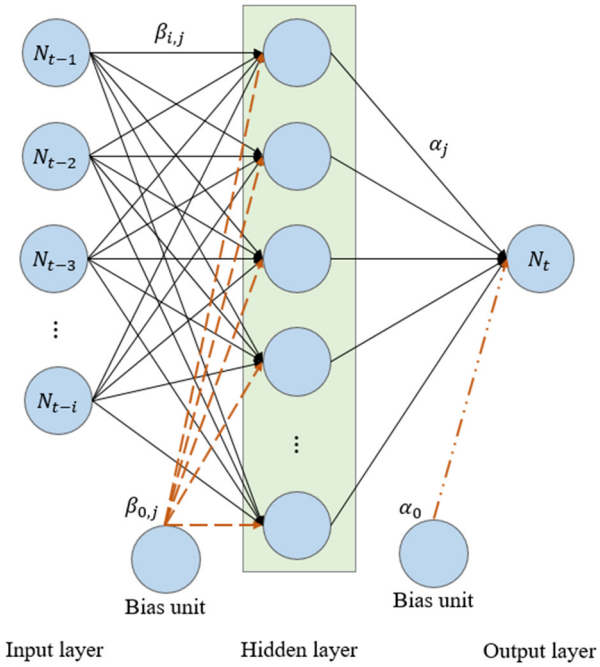

Fig. 1 Structure of an ANN model (modified from [22])

In recent years, neural networks have gained significant attention and are being effectively applied in a wide variety of fields. Neural networks are increasingly used in areas involving prediction, classification, or control tasks. In addition, neural networks can be defined as a network of interconnected simple processing units, modeled after the biological neuron. A

biological neuron is a specialized unit that carries information or knowledge and transmits it to other neurons within the network. Artificial neurons are organized in layers and interconnected, resembling synapses in the brain. The internal layers, called hidden layers, consist of varying numbers of neurons. The final layer, known as the output layer, has several neurons that match the number of outputs. Expanding the number of neurons and layers improves the learning ability of ANN, enabling them to handle more complex data [21]. A single hidden layer feedforward network is the most widely used model form for time series modeling and forecasting. The model is characterized by a network of three layers of simple processing units connected by acyclic links [22]. Fig. 1 shows the structure of an ANN model.

The relationship between the output, N_t, and the inputs, N_{t-1}, N_{t-2}, ⋯, N_{t-p} can be mathematical represented as:

$$N_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g_H \left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} N_{t-i} \right) + \varepsilon_t \tag{8}$$

where N_t indicates the predicted output of the network at time t, α_0 symbolizes the bias term for the output layer, α_j signifies the weights connecting the hidden layer to the output layer where $j = 1, 2, \cdots, q$, in this case, the logistic function, β_{0j} are the bias terms for each of the hidden nodes $j = 1, 2, \cdots, q$, β_{ij} are the weights connecting the input layer to the hidden layer where $i = 1, 2, \cdots, p$; $j = 1, 2, \cdots, q$, N_{t-1} are the input values at previous time steps (lags in time series terms) $i = 1, 2, \cdots, p$, ε_t is the error term for the prediction at time t. When g_H is the linear function in the hidden layer which is selected as the activation function defined by:

$$g_H(v) = v \tag{9}$$

where v is the input value.

3.7. The Elman recurrent neural network (ERNN)

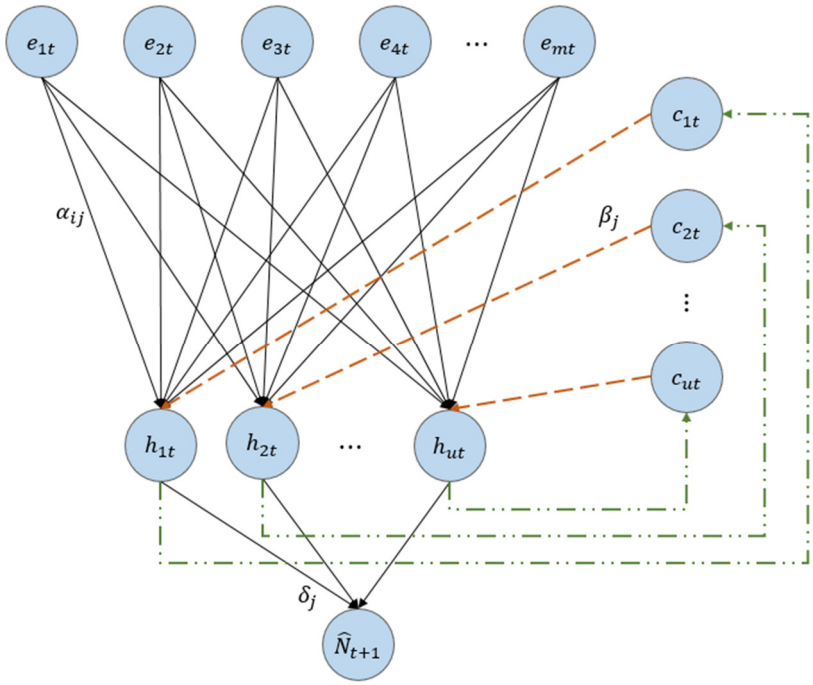

Fig. 2 Structure of an ERNN model (modified from [2])

ERNN, also referred to as a simple recurrent network, is a form of feedforward ANN that incorporates feedback connections, enabling it to handle sequential data and time series [2] which are comprised of input, hidden, output, and recurrent layers [23-24]. In the ERNN, the recurrent layer receives feedback from outputs of the hidden layer, empowering the network to learn, identify, and generate both spatial and temporal patterns. Each neuron in the hidden layer connects to a

corresponding neuron in the recurrent layer with a fixed weight of one. This structure makes the recurrent layer function as a memory of the hidden layer's previous state. The number of neurons in the recurrent layer matches that of the hidden layer. Neurons pass information to subsequent layers by applying a nonlinear function to the weighted sum of their inputs. This design enables the ERNN to effectively analyze and respond to complex patterns [23]. Fig. 2 shows the structure of an ERNN model.

The inputs of the hidden layer are given by the following formula,

$$net_{it}(k) = \sum_{i=1}^{m} \alpha_{ij} e_{it}(k-1) + \sum_{j=1}^{u} \beta_j c_{jt}(k) \tag{10}$$

$$c_{jt}(k) = h_{jt}(k-1),\ i = 1,2,\cdots,m;\ j = 1,2,\cdots,u \tag{11}$$

where net_{it} denote network at time t, m is the number of neurons in the inputs layer, u is the number of neurons in the hidden layer, e_{it} is the set of an input vector of neurons at the time t where i = 1, 2, ⋯, m, h_{jt} is the output of hidden layer neurons at time t where j = 1, 2, ⋯, u, c_{jt} is the context layer neuron at time t where j = 1, 2, ⋯, u, α_{ij} is the weight that connects the node i in the input layer neurons to the node j in the hidden layer, β_j are the weights that connect the node j in the context layer neurons to the node in the hidden layers, and k is the maximum training iteration number and initial connective weights. Additionally, the topology of the network architecture is determined by the number of neural nodes in the hidden layer. As a result of the hidden neurons, the output is as follows:

$$h_{jt}(k) = f_H\left(\sum_{i=1}^{m} \alpha_{ij} e_{it}(k-1) + \sum_{j=1}^{u} \beta_j c_{jt}(k)\right) \tag{12}$$

where the linear function f_H in the hidden layer is selected as the activation function based on the following equation:

$$f_H(v) = v \tag{13}$$

where v is the input value.

The output of the hidden layer is given as follows:

$$N_t = f_T\left(\sum_{j=1}^{m} \delta_j h_{jt}(k)\right) \tag{14}$$

where N_t is the output of the network at time t, δ_j are the weights that connect the node j, in the hidden layer neurons to the node in the output layers, and $f_T(.)$ is an identity map as the activation function.

3.8. The general framework of the hybrid model

The estimation of N_t will result in the estimation of the non-linear component in the time series $\hat{y}_t$. The estimated values of the time series are obtained as follows:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{15}$$

The neural network (ANN/ERNN) can effectively capture the non-linear patterns present in the residuals obtained from the ARIMAX model. By incorporating neural network (ANN/ERNN) into a hybrid model, forecasting that performance can be significantly enhanced [10]. Hence, the problem of spurious regression can be addressed by using hybrid ARIMAX models with neural network (ANN/ERNN) to account for linear and non-linear terms in time series data. These methods furnish a

powerful tool for accurately modeling time series data and can help overcome the limitations of traditional regression models in the presence of spurious regression. Traditionally, the number of hidden layers is selected based on minimizing the MMSE [25].

$$MMSE = \frac{1}{q}\sum_{i=1}^{q}\left(N_{t-i} - \hat{N}_{t-i}\right)^2 \tag{16}$$

3.9. Optimization of the hybrid model

The approach extends this methodology by systematically optimizing both the number of lagged inputs (e_{it}) and the number of hidden layers (denoted as h_{jt}). The optimization is conducted through a hybrid model that aims to minimize the mean squared error, as detailed in the following equation:

$$MMSE = \frac{1}{q}\sum_{i=1}^{q}\left(y_{t-i} - \hat{y}_{t-i}\right)^2 \tag{17}$$

Fig. 3 depicts the enhanced process flow of the hybrid model featuring alternative optimization. The sequence initiates with the input y_t, x_t which enters the ARIMAX model, functioning as the linear estimator to generate $\hat{L}_t$. The residuals from this model are fed into a neural network (ANN/ERNN) thereafter, serving as the nonlinear estimator to compute $\hat{N}_t$. The system subsequently assesses if the combination of $\hat{L}_t$ and $\hat{N}_t$ satisfies the MMSE threshold. If the MMSE criteria are not fulfilled, the process iteratively refines the lags and hidden layers within the neural network to optimize $\hat{N}_t$, thereby enhancing the predictive accuracy. Upon meeting the MMSE threshold, the final output $\hat{y}_t$ is produced.

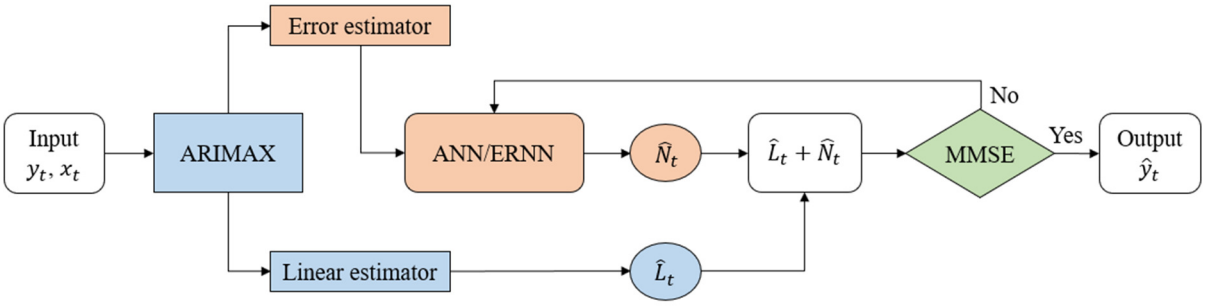

Fig. 3 Conceptual framework of alternative optimization of hybrid model

4. Results and Discussion

This section presents the findings from the CCF analysis, and the statistical tests conducted to evaluate its performance. The results comprise a detailed comparison of model performance and highlight key findings from the analysis across each period.

4.1 Result from cross-correlation function (CCF)

This section presents the results of the CCF analysis, a crucial tool for detecting relationships between two-time series. The analysis focuses on the correlations between various indices of the SET as endogenous variables with GGT and the THB across four distinct periods from Table 2 to Table 5.

Table 2 Cross-correlation analysis for the first period

Variable	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
GGT	–0.2957*	–0.3769*	–0.3041*	–0.3056 *	–0.3289*	–0.2268*	–0.2846*	–0.2456*
THB	–0.2609*	–0.2384*	–0.2161*	–0.1970*	–0.2429*	–0.1721	–0.2790*	–0.2228*

*95% significant level at confidence interval between [-0.1803, 0.1803]

Table 3 Cross-correlation analysis for the second period

Variable	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
GGT	0.0519	0.1187	–0.0901	–0.0798	0.0270	–0.0353	–0.0439	0.0108
THB	–0.2422*	0.1656	–0.3154*	–0.2757*	–0.3241*	–0.3106*	–0.3262*	–0.1332

*95% significant level at confidence interval between [-0.1825, 0.1825]

Table 4 Cross-correlation analysis for the third period

Variable	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
GGT	0.0120	0.1542	–0.2905*	–0.0019	–0.2717*	–0.0575	–0.2533*	–0.1090
THB	–0.1063	–0.0282	0.0357	0.0775	0.0112	–0.0612	–0.0275	–0.0926

*95% significant level at confidence interval between [-0.2340, 0.2340]

Table 5 Cross-correlation analysis for the fourth period

Variable	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
GGT	–0.0243	–0.0669	0.0369	–0.0458	0.0092	0.0251	0.0399	–0.1299
THB	–0.1446	0.008	–0.2250*	–0.1362	–0.2671*	–0.2166*	–0.2699*	–0.0260

*95% significant level at confidence interval between [-0.1543, 0.1543]

The CCF analysis presented in the tables reveals significant insights into the relationships between indices of the SET with GGT and the THB across four distinct periods. During the first period, GGT and THB exhibited significant negative correlations with all sectors except for the relationship between THB and SETR. In the second period, significant correlations of GGT were absent, while THB demonstrated significant negative correlations with all sectors except for SETC and SETT. The third-period analysis indicated that GGT had significant negative correlations with SETF, SETP, and SETS, while THB did not show significant correlations during this period. Finally, in the fourth period, GGT did not exhibit significant correlations, while THB exhibited significant negative correlations with SETF, SETP, SETR, and SETS.

4.2. Empirical application

This section evaluates the performance of the optimized hybrid ARIMAX models with neural network (ANN/ERNN) by applying the SET index data across each sector with exogenous variables for each period. The performance of the optimized ARIMAX models with neural network (ANN/ERNN) is benchmarked against traditional ARIMAX models and traditional ARIMAX models with neural network (ANN/ERNN), thereby highlighting keys from the analysis.

4.2.1 Result from the first period with GGT

Table 6 presents evidence that the ARIMAX model from the first period with GGT as an exogenous variable exhibits significant results in all sectors including negative coefficients. Furthermore, Table 7 displays the lowest Akaike Information Criterion (AIC) values and p-value of the Ljung-Box test supporting the null hypothesis and implying that residuals of the model are white noise. Table 8 details the chosen lag and hidden layer of each model. Fig. 4 presents the RMSE value for each model demonstrating that the hybrid model outperforms the ARIMAX model. Moreover, every optimized hybrid model yields better performance compared to both the traditional hybrid model and the traditional ARIMAX model.

Table 6 Coefficients and p-values of the first period ARIMAX model with Google Trends index

Variable	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
AR1	0.8771*	0.7522*	1.3238*	0.7791*	1.4713*	–0.5070	–1.2318*	–0.7022*
AR2	–0.9002*	-	–0.58801*	-	–0.8341*	0.3895	–0.3124*	–0.8062*
AR3	-	-	-	-	-	0.2865*	-	-
MA1	–1.0171*	–0.9111*	–1.5068*	–0.9304*	–1.6086*	-	0.9999*	0.6983*
MA2	0.9999*	0.2680	0.8323*	0.2153*	0.9999*	0.3859	-	0.8503*
MA3	-	-	-	-	-	–0.4460	-	–0.1719
GGT	–0.0307*	–0.0165*	–0.0388*	–0.0385*	–0.0348*	–0.0319*	–0.0316*	–0.0189*

*Significant at 0.05

Table 7 Akaike Information Criterion (AIC) values and Ljung-Box test p-values for ARIMAX models with Google Trends index in the first period

Technique	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
AIC	–590.294	–777.892	–540.771	–522.827	–591.543	–506.495	–593.228	–629.452
Ljung-Box	0.2553	0.84331	0.9837	0.9813	0.462	0.9996	0.9707	0.8894

Table 8 Configuration of lags and hidden layers in neural network models with Google Trends index for the first period

Technique	Lags and hidden layers	SETA	SETC	SETF	SETI	SETP	SETR	SETS	SETT
Traditional ANN	Lags	1	1	1	1	1	1	1	1
	Hidden layers	23	7	32	17	5	17	91	7
Optimized ANN	Lags	4	1	1	4	1	3	2	3
	Hidden layers	6	71	70	30	45	52	10	42
Traditional ERNN	Lags	1	1	1	1	1	1	1	1
	Hidden layers	59	88	93	90	80	82	93	59
Optimized ERNN	Lags	1	1	2	1	4	5	2	2
	Hidden layers	46	40	63	52	97	58	43	89

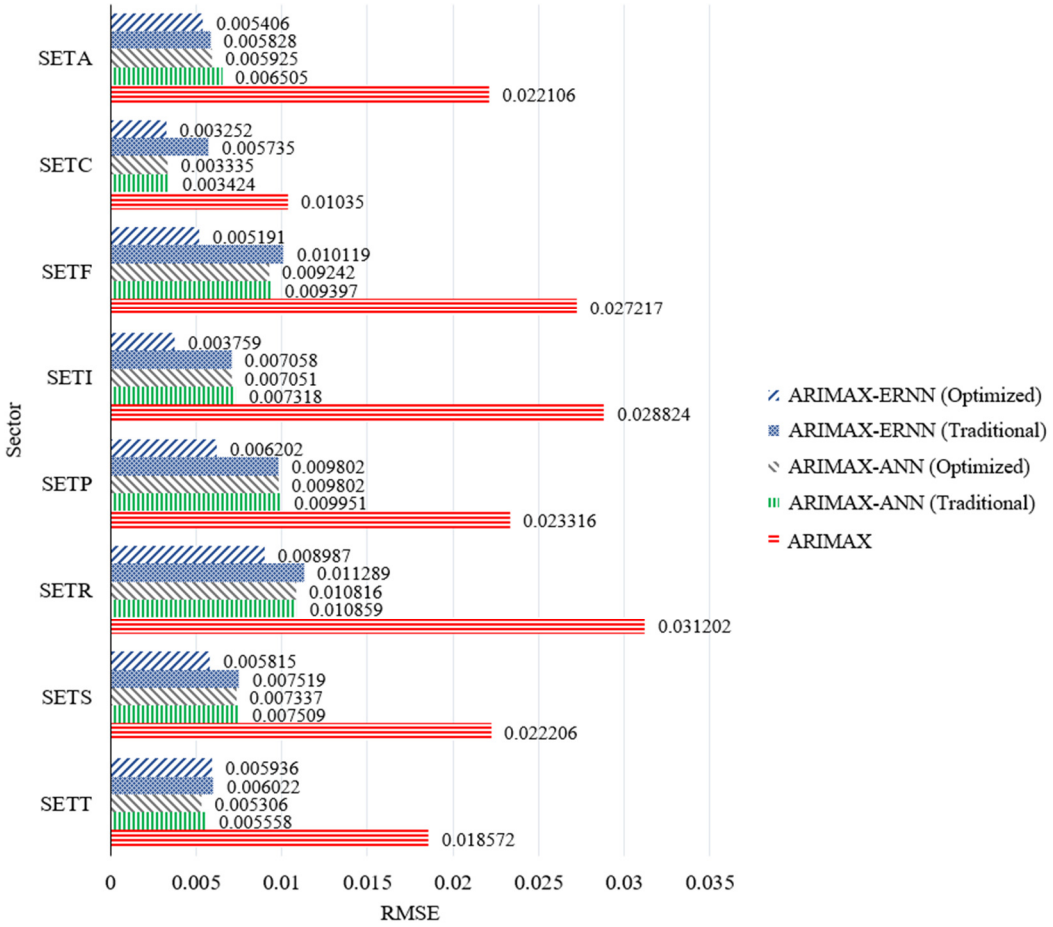

Fig. 4 Performance comparison of models using the Google Trends index as an exogenous variable in the first period

4.2.2 Result from the first period with THB

Table 9 shows that except for SETR the ARIMAX model in the first period with THB as an exogenous variable is significant across all sectors. Table 10 highlights the model with the lowest AIC, along with the p-value of the Ljung-Box test, confirming white noise in the residuals. Table 11 presents the hybrid ARIMAX-ERNN model for SETA and SETP, which has identical lag and hidden layer configurations in both traditional and optimized modes. The results in Fig. 5 demonstrate that the ARIMAX-ERNN model for SETA and SETP yields identical performance in both traditional and optimized configurations. Additionally, the hybrid model consistently outperforms the ARIMAX model. The optimized hybrid model surpasses both the traditional hybrid model and the traditional ARIMAX model in all optimized configurations.

Table 9 Coefficients and p-values of the first period ARIMAX model with Thai Bath (THB)

Variable	SETA	SETC	SETF	SETI	SETP	SETS	SETT
AR1	−0.9607*	1.0241*	0.6415*	−1.2464*	−1.7152*	−1.1349*	−0.9164*
AR2	−0.1618	−0.7139*	0.1651*	−0.9359*	−0.8275*	−0.2893*	−0.8016*
MA1	0.8590*	−1.1832*	−0.7330*	1.1882*	1.6998*	0.9230*	0.9902*
MA2	-	0.9999*	-	0.9999*	0.7553*	-	0.9999*
THB	−1.6612*	−0.5689*	−1.4272*	−1.4624*	−1.6896*	−1.8257*	−1.1215*

*95% significant level

Table 10 Akaike Information Criterion (AIC) values and Ljung-Box test p-values for ARIMAX models with Thai Bath in the first period

Technique	SETA	SETC	SETF	SETI	SETP	SETS	SETT
AIC	−584.175	−770.001	−529.157	−523.496	−588.622	−590.41	−628.809
Ljung-Box	0.874535	0.791028	0.915481	0.271637	0.981394	0.782526	0.696605

Table 11 Configuration of lags and hidden layers in neural network models with Thai Bath for the first period

Technique	Lags and hidden layers	SETA	SETC	SETF	SETI	SETP	SETS	SETT
Traditional ANN	Lags	1	1	1	1	1	1	1
	Hidden layers	8	16	7	47	5	91	7
Optimized ANN	Lags	1	4	2	2	1	3	2
	Hidden layers	17	61	10	85	51	75	96
Traditional ERNN	Lags	1	1	1	1	1	1	1
	Hidden layers	59	88	95	85	59	93	59
Optimized ERNN	Lags	1	3	1	2	1	1	2
	Hidden layers	59	95	52	63	59	94	89

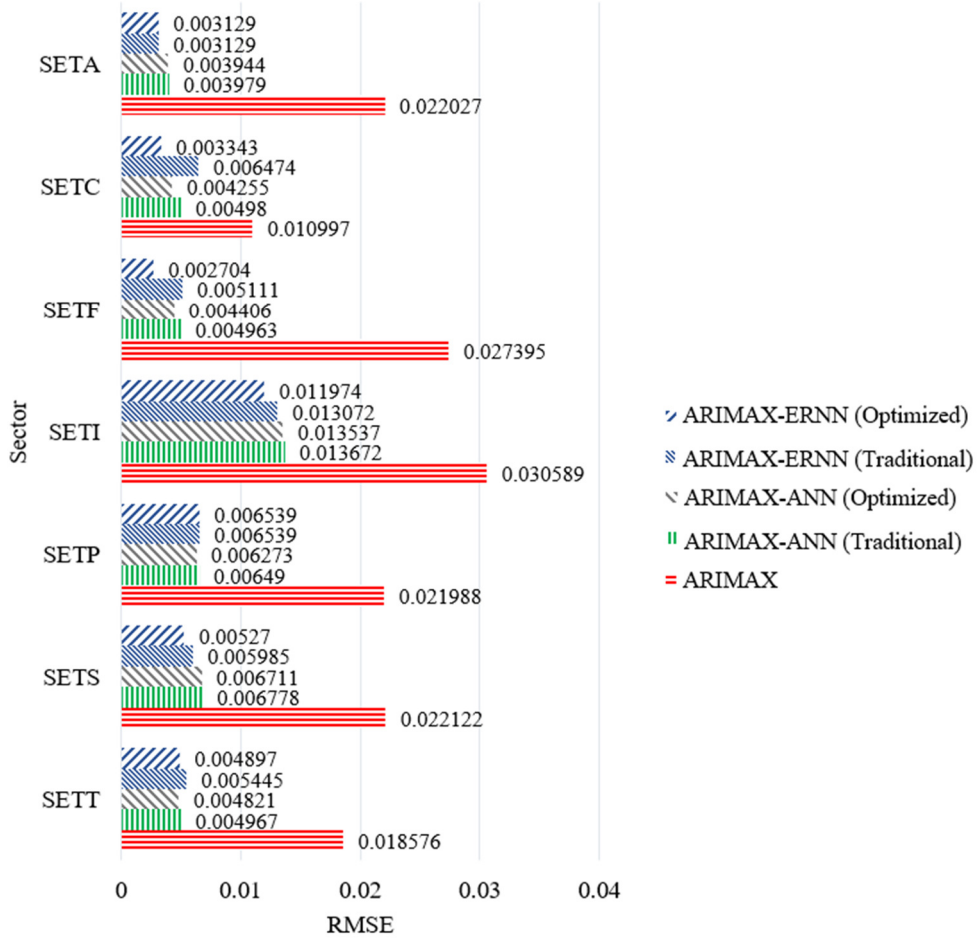

Fig. 5 Performance comparison of models using Thai Bath as an exogenous variable in the first period

4.2.3 Result from the second period with THB

Table 12 depicts that all sectors, except SETC and SETT, have negative coefficients in the ARIMAX model during the second period with THB as an exogenous variable. Table 13 supports the null hypothesis that the residuals are white noise and highlights the models with the lowest AIC values. Table 14 lists the opted hidden layers and lag configurations, with the hybrid ARIMAX-ERNN model for SETF denoting identical setups in both traditional and optimized models. Fig. 6 indicates that the ARIMAX-ERNN model for SETF yields identical results in both traditional and optimized models. The optimized hybrid model outperforms both the traditional hybrid and ARIMAX models.

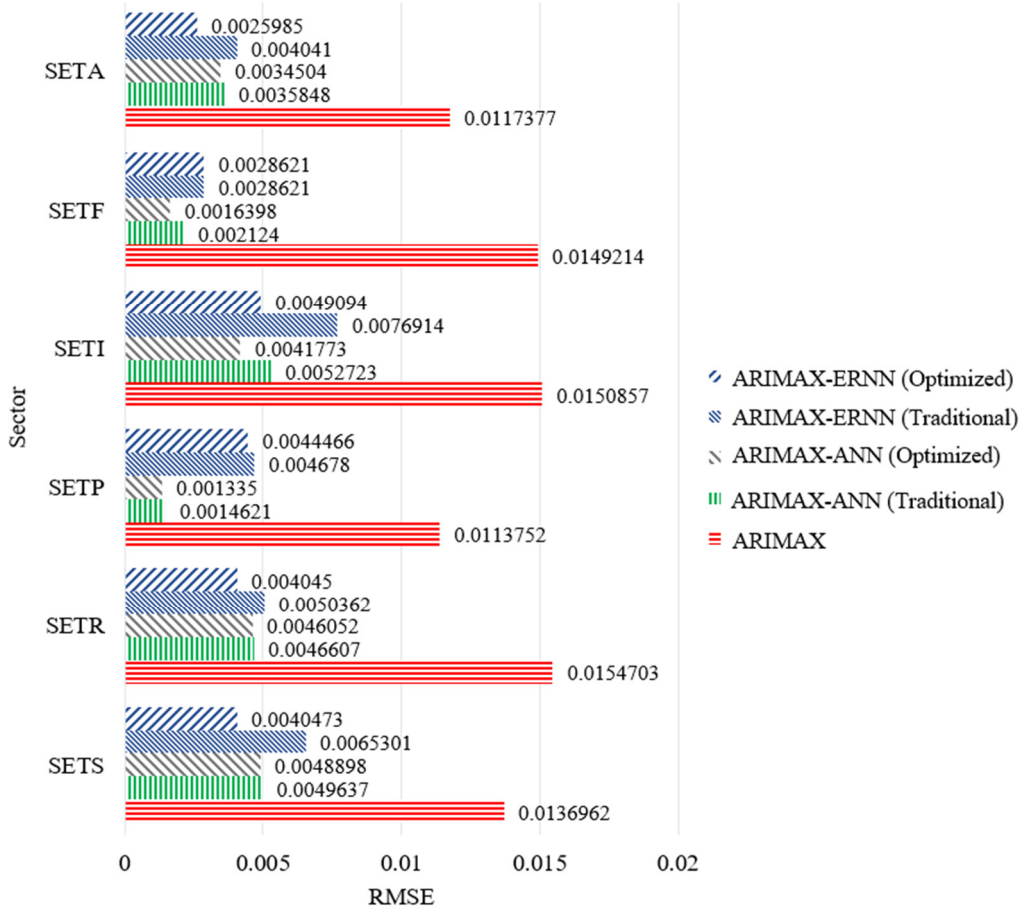

Fig. 6 Performance comparison of models using Thai Bath as an exogenous variable in the second period

Table 12 Coefficients and p-values of the second period ARIMAX model with Thai Bath (THB)

Variable	SETA	SETF	SETI	SETP	SETR	SETS
AR1	0.9561*	–0.8463*	–0.1079	0.8856*	1.4443*	–0.8714*
AR2	-	–0.1112*	–0.7787*	-	–0.6594*	–0.9951*
AR3	-	-	-	-	–0.2047*	-
MA1	–1.1374*	0.7463*	–0.0188	–0.8402*	–1.6374*	0.8762*
MA2	0.1374	-	0.9999*	-	0.9994*	0.9771*
MA3	-	-	-0.0217	-	-	-
THB	–0.8201*	–1.5790*	–1.1882*	–1.1531*	–1.85461*	–1.3732*

*Significant at 0.05

Table 13 Akaike Information Criterion (AIC) values and Ljung-Box test p-values for ARIMAX models with Thai Bath in the second period

Technique	SETA	SETF	SETI	SETP	SETR	SETS
AIC	–728.935	–659.234	–671.951	–727.489	–657.02	–699.31
Ljung-Box	0.9661	0.9995	0.8892	0.32160	0.9164	0.9556

Table 14 Configuration of lags and hidden layers in neural network models with Thai Bath for the second period

Technique	Lags and hidden layers	SETA	SETF	SETI	SETP	SETR	SETS
Traditional ANN	Lags	1	1	1	1	1	1
	Hidden layers	26	80	30	7	47	21
Optimized ANN	Lags	5	2	3	5	5	4
	Hidden layers	52	73	87	11	79	75
Traditional ERNN	Lags	1	1	1	1	1	1
	Hidden layers	40	85	85	40	46	40
Optimized ERNN	Lags	2	1	1	4	3	1
	Hidden layers	46	85	40	86	89	85

4.2.4 Result from the third period with GGT index

Table 15 indicates that SETF and SETP have negative and significant coefficients in the ARIMAX model with GGT as an exogenous variable during the third period. Table 16 highlights the model with the lowest AIC and a p-value from the Ljung-Box test, supporting the null hypothesis and suggesting that the residuals are noise. Table 17 outlines the opted hidden layers and lag configurations, indicating that the hybrid ARIMAX-ERNN model for SETF offers the same configurations in both traditional and optimized modes. Fig. 7 shows that the ARIMAX-ERNN model for SETF produces the same results in both traditional and optimized modes. Moreover, such a finding confirms that the optimized hybrid model outperforms both the traditional hybrid and ARIMAX models.

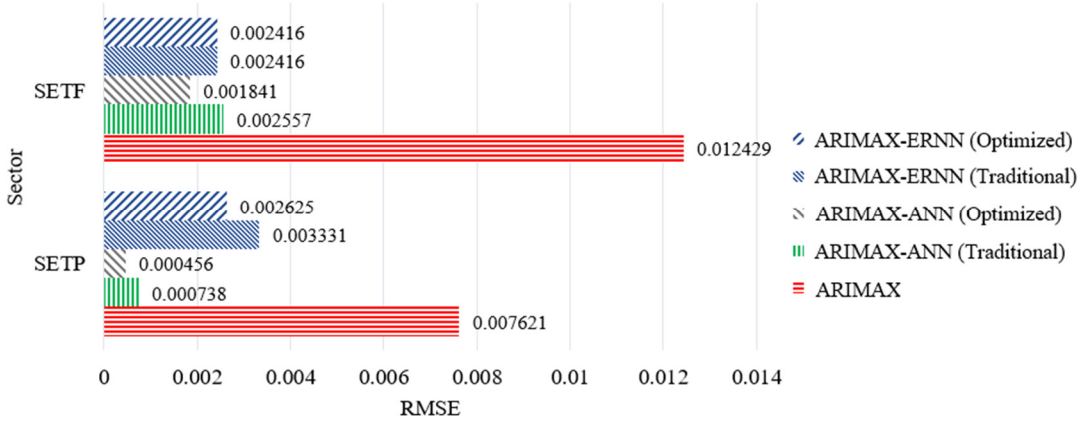

Fig. 7 Performance comparison of models using the Google Trends index as an exogenous variable in the third period

Table 15 Coefficients and p-values of the third period ARIMAX model with Google Trends (GGT) index

Variable	SETF	SETP
AR1	–0.56539	0.99661*
MA1	0.715767*	–0.92404*
MA2	-	–0.06181
GGT	–0.01542*	–0.01042*

*Significant at 0.05

Table 16 Akaike Information Criterion (AIC) values and Ljung-Box test p-values for ARIMAX models with Google Trends Index in the third period

Technique	SETF	SETP
AIC	–431.606	–493.575
Ljung-Box	0.775444	0.857292

Table 17 Configuration of lags and hidden layers in neural network models with Google Trends Index for the third period

Technique	Lags and hidden layers	SETF	SETP
Traditional ANN	Lags	1	1
	Hidden layers	62	30
Optimized ANN	Lags	2	2
	Hidden layers	32	35

Table 17 Configuration of lags and hidden layers in neural network models with Google Trends Index for the third period (continued)

Technique	Lags and hidden layers	SETF	SETP
Traditional ERNN	Lags	1	2
	Hidden layers	57	46
Optimized ERNN	Lags	1	1
	Hidden layers	57	72

4.2.5 Result from the fourth period with THB

Table 18 demonstrates that the ARIMAX model with THB as an exogenous variable in the fourth period is statistically significant for SETF, SETP, SETR, and SETS with all models exhibiting negative coefficients. In Table 19, the ARIMAX model is selected by its lowest AIC values while p-values of the Ljung-Box test confirm that the residuals are white noise supporting the null hypothesis. Table 20 renders a detailed comparison of the selected lags and hidden layers for each model. Notably, SETP and SETS exhibit identical lags and hidden layers in their traditional and optimized ARIMAX-ERNN models. The performance presented in Fig. 8 reveals that the ARIMAX-ERNN model yields consistent results for SETP and SETS in both traditional and optimized approaches. The optimized hybrid model excels both the traditional hybrid and ARIMAX models in performance.

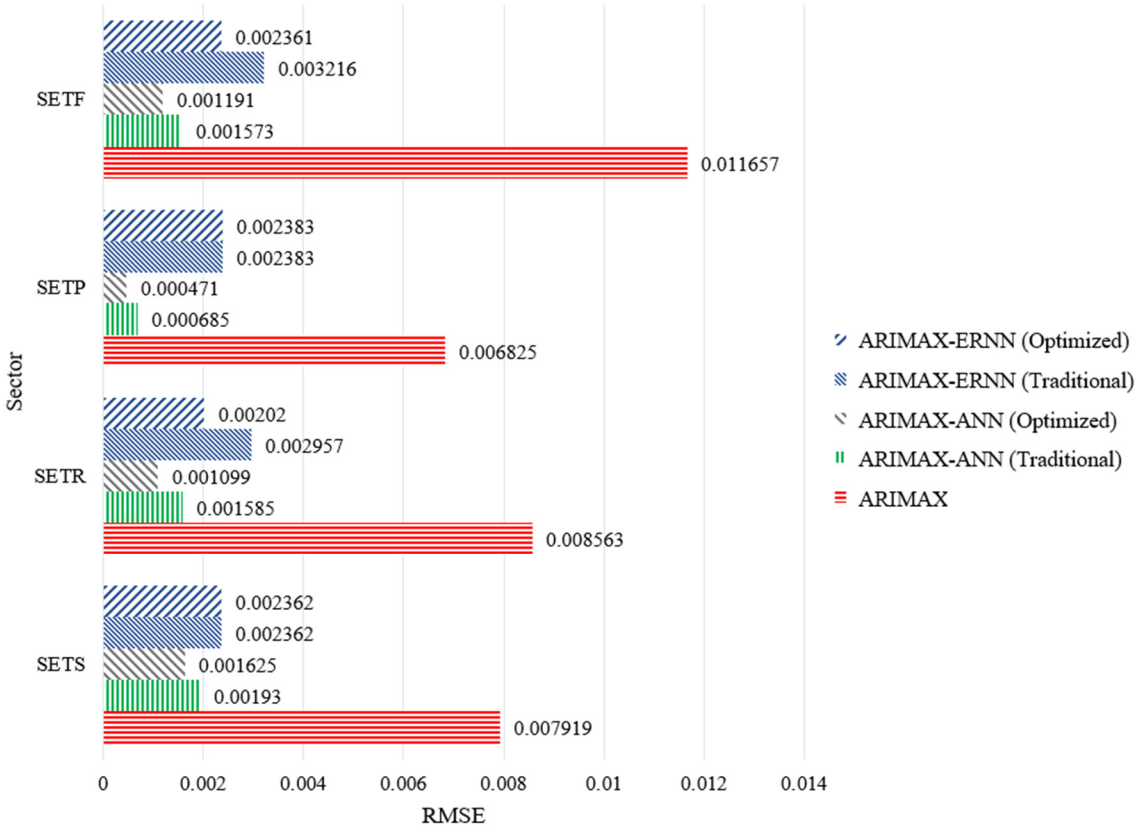

Fig. 8 Performance comparison of models using Thai Bath as an exogenous variable in the fourth period

Table 18 Coefficients and p-values of the fourth period ARIMAX model with Thai Bath (THB)

Variable	SETF	SETP	SETR	SETS
AR1	0.5920*	–0.5989*	–0.6401*	–0.4478
AR2	-	-	-	–0.1642*
MA1	–0.6814*	0.6631*	0.7828*	0.4225
THB	–0.7917*	–0.4754*	–0.3997*	–0.6529

*Significant at 0.05

Table 19 Akaike Information Criterion (AIC) values and Ljung-Box test p-values for ARIMAX models with Thai Bath in the fourth period

Technique	SETF	SETP	SETR	SETS
AIC	−1016.12650	−1193.54	−1121.17	−1153.28
Ljung-Box	0.638347241	0.379886	0.5806	0.978849

Table 20 Configuration of lags and hidden layers in neural network models with Thai Bath for the fourth period

Technique	Lags and hidden layers	SETF	SETP	SETR	SETS
Traditional ANN	Lags	1	1	1	1
	Hidden layers	7	30	34	30
Optimized ANN	Lags	5	3	4	5
	Hidden layers	19	47	85	53
Traditional ERNN	Lags	1	1	1	1
	Hidden layers	80	40	46	99
Optimized ERNN	Lags	1	1	1	1
	Hidden layers	46	40	57	99

5. Conclusion

This study has developed a hybrid forecasting model that integrates ARIMAX with neural networks (ANN/ERNN) using an alternative optimization process. In contrast to the traditional hybrid approach, where the RMSE of the nonlinear component is calculated iteratively until the lowest value is obtained and subsequently added to the linear component, the proposed method calculates the RMSE after both linear and nonlinear components are combined. This RMSE is subsequently refined through iterative looping until the least RMSE is achieved, ensuring a more accurate optimization process for the hybrid model.

The proposed methodology demonstrates significant improvements in predictive accuracy by effectively capturing both linear and nonlinear components in the SET index across eight industries during the COVID-19 period. According to the results, the optimized hybrid models achieved the lowest RMSE. These findings confirm that the proposed optimization process in hybrid ARIMAX-ANN and ARIMAX-ERNN models outperforms the traditional hybrid model and traditional ARIMAX model, showing its effectiveness in enhancing forecasting model accuracy.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] K. A. Ababio, “Comparative Study of Stock Price Forecasting Using ARIMA and ARIMAX Models,” Ph.D. dissertation, Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, 2012.

[2] D. K. Shetty and B. Ismail, “Forecasting Stock Prices Using Hybrid Non-Stationary Time Series Model With ERNN,” Communications in Statistics-Simulation and Computation, vol. 52, no. 3, pp. 1026-1040, 2023.

[3] P. Almaleck, S. Massucco, G. Mosaico, M. Saviozzi, P. Serra, and F. Silvestro, “Electrical Consumption Forecasting in Sports Venues: A Proposed Approach Based on Neural Networks and ARIMAX Models,” Sustainable Cities and Society, vol. 100, article no. 105019, 2024.

[4] A. A. Pierre, S. A. Akim, A. K. Semenyo, and B. Babiga, “Peak Electrical Energy Consumption Prediction by ARIMA, LSTM, GRU, ARIMA-LSTM and ARIMA-GRU Approaches,” Energies, vol. 16, no. 12, article no. 4739, 2023.

[5] D. Xu, Q. Zhang, Y. Ding, and D. Zhang, “Application of a Hybrid ARIMA-LSTM Model Based on the SPEI for Drought Forecasting,” Environmental Science and Pollution Research, vol. 29, no. 3, pp. 4128-4144, 2022.

[6] A. S. Azad, R. Sokkalingam, H. Daud, S. K. Adhikary, H. Khurshid, S. N. A. Mazlan, et al., “Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study,” Sustainability, vol. 14, no. 3, article no. 1843, 2022.

[7] K. Nualtong, T. Panityakul, P. Khwanmuang, R. Chinram, and S. Kirtsaeng, "A Hybrid Seasonal Box Jenkins-ANN Approach for Water Level Forecasting in Thailand," Environment and Ecology Research, vol. 9, no. 3, pp. 93-106, 2021.
[8] S. A. Shahriar, I. Kayes, K. Hasan, M. Hasan, R. Islam, N. R. Awang, et al., "Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for Atmospheric PM2.5 Forecasting in Bangladesh," Atmosphere, vol. 12, no. 1, article no.. 100, 2021.
[9] D. Fan, H. Sun, J. Yao, K. Zhang, X. Yan, and Z. Sun, "Well Production Forecasting Based on ARIMA-LSTM Model Considering Manual Operations," Energy, vol. 220, article no. 119708, 2021.
[10] M. Wang, J. Pan, X. Li, M. Li, Z. Liu, Q. Zhao, et al., "ARIMA and ARIMA-ERNN Models for Prediction of Pertussis Incidence in Mainland China From 2004 to 2021," BMC Public Health, vol. 22, no. 1, article no. 1447, 2022.
[11] A. Sharma, P. Tiwari, A. Gupta, and P. Garg, "Use of LSTM and ARIMAX Algorithms to Analyze Impact of Sentiment Analysis in Stock Market Prediction," Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020, pp. 377-394, 2021.
[12] P. Lv, Q. Wu, J. Xu, and Y. Shu, "Stock Index Prediction Based on Time Series Decomposition and Hybrid Model," Entropy, vol. 24, no. 2, article no. 146, 2022.
[13] E. Alshawarbeh, A. T. Abdulrahman, and E. Hussam, "Statistical Modeling of High Frequency Datasets Using the ARIMA-ANN Hybrid," Mathematics, vol. 11, no. 22, article no. 4594, 2023.
[14] B. Singh, S. K. Henge, S. K. Mandal, M. K. Yadav, P. T. Yadav, A. Upadhyay, et al., "Auto-Regressive Integrated Moving Average Threshold Influence Techniques for Stock Data Analysis," International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, pp. 446-455, 2023.
[15] "SET Industry Group and Sector Classification," https://www.set.or.th/en/market/index/set/industry-sector-profile, accessed on 2021.
[16] N. Hongsakulvasu, C. Khiewngamdee, and A. Liammukda, "Does COVID-19 Crisis Affects the Spillover of Oil Market's Return and Risk on Thailand's Sectoral Stock Return?: Evidence from Bivariate DCC GARCH-in-Mean Model," International Energy Journal, vol. 20, no. 4, pp. 647-662, 2020.
[17] P. C. B. Phillips and P. Perron, "Testing for a Unit Root in Time Series Regression," Biometrika, vol. 75, no. 2, pp. 335-346, 1988.
[18] D. F. Chang, C. C. Chen, and A. Chang, "Forecasting With ARIMAX Models for Participating STEM Programs," ICIC Express Letters. Part B, Applications, vol. 11, no. 2, pp. 121-128, 2020.
[19] C. H. Aladag, E. Egrioglu, and C. Kadilar, "Forecasting Nonlinear Time Series With a Hybrid Methodology," Applied Mathematics Letters, vol. 22, no. 9, pp. 1467-1470, 2009.
[20] J. D. Cryer and K. S. Chan, Time Series Analysis: With Applications in R, 2nd ed., New York: Springer, 2008.
[21] T. Guillod, P. Papamanolis, and J. W. Kolar, "Artificial Neural Network (ANN) Based Fast and Accurate Inductor Modeling and Design," IEEE Open Journal of Power Electronics, vol. 1, pp. 284-299, 2020.
[22] S. Stephen, O. Argwings, and K. Julius, "Application of ARIMA, Hybrid ARIMA and Artificial Neural Network Models in Predicting and Forecasting Tuberculosis Incidences Among Children in Homa Bay and Turkana Counties, Kenya," in press. https://doi.org/10.1101/2022.07.07.22277378
[23] J. Wang, J. Wang, W. Fang, and H. Niu, "Financial Time Series Prediction Using Elman Recurrent Random Neural Networks," Computational Intelligence and Neuroscience, vol. 2016, no. 1, article no. 4742515, 2016.
[24] D. Zhang, W. Li, X. Han, B. Lu, Q. Zhang, and C. Bo, "Evolving Elman Neural Networks Based State-of-Health Estimation for Satellite Lithium-Ion Batteries," Journal of Energy Storage, vol. 59, article no. 106571, 2023.
[25] K. Sako, B. N. Mpinda, and P. C. Rodrigues, "Neural Networks for Financial Time Series Forecasting," Entropy, vol. 24, no. 5, article no. 657, 2022.

Appendix

Abbreviation	Definition
ANN	Artificial neural network
ARIMAX	Autoregressive integrated moving average with exogenous variables

ERNN	Elman recurrent neural network
ARIMA	Autoregressive integrated moving average
MAPE	Mean absolute percentage error
LSTM	Long short-term memory
GRU	Gated recurrent unit
RMSE	Root mean square error
SARIMA	Seasonal autoregressive integrated moving average
ARMA	Autoregressive moving average
SET	Stock Exchange of Thailand
SETA	Stock Exchange of Thailand in Agro and Food Industry Sector
SETC	Stock Exchange of Thailand in Consumer Products Sector
SETF	Stock Exchange of Thailand in Financials Sector
SETI	Stock Exchange of Thailand in Industrials Sector
SETP	Stock Exchange of Thailand in Property and Construction Sector
SETR	Stock Exchange of Thailand in Resources Sector
SETS	Stock Exchange of Thailand in Services Sector
SETT	Stock Exchange of Thailand in Technology Sector
THB	Thai Baht
GGT	Google Trends
PP	Phillips-Perron
CCF	Cross-correlation function
MMSE	Minimum mean square error
AIC	Akaike Information Criterion