Article

Statistical Modeling of High Frequency Datasets Using the ARIMA-ANN Hybrid

Etaf Alshawarbeh [1] , **Alanazi Talal Abdulrahman** [1] and **Eslam Hussam** [2,3,*]

1 Department of Mathematics, College of Science, University of Ha'il, Ha'il P.O. Box 55476, Saudi Arabia; e.alshawarbeh@uoh.edu.sa (E.A.); t.shyman@uoh.edu.sa (A.T.A.)
2 Department of Accounting, College of Business Administration in Hawtat bani Tamim, Prince Sattam bin Abdulaziz University, Hawtat bani Tamim, Saudi Arabia
3 Department of Mathematics, Faculty of Science, Helwan University, Cairo 12613, Egypt
* Correspondence: e.hafez@psau.edu.sa

Abstract: The core objective of this work is to predict stock market indices' using autoregressive integrated moving average (ARIMA), artificial neural network (ANN) and their combination in the form of ARIMA-ANN. Financial data are, in fact, trendy, noisy and highly volatile. To tackle their chaotic nature and forecast the three considered stock markets, namely Nasdaq stock exchange, United States, Nikkei stock exchange, Japan, and France stock exchange data (CAC 40 index), we use novel approaches. The data are taken from the Yahoo Finance website for the period from 4 January 2010 to 20 August 2021. To assess the relative predictive effectiveness of the selected tools, the dataset was divided into two distinct subsets: 75% of the data was allocated for training purposes, while the remaining 25% was reserved for testing. The empirical results suggest that ARIMA-ANN produces more accurate forecasts than the separate components of all stock markets. In light of this, it may be inferred that the combining tool is more effective in analyzing financial data and provides a more accurate comparative prediction.

Keywords: stock markets; machine learning; hybridization; forecasting

MSC: 60E05

1. Introduction

The stock market, or equity market, consists of numerous stock exchanges across the globe. The general public and investors sell and purchase shares, whose prices fluctuate constantly by dint of the law of demand and supply. A stock or share represents partial possession of a company or corporation. Buyers attempt to purchase a share at the lowest feasible price, while sellers attempt to sell it at the highest price [1]. One of the most significant venues for raising capital is the stock market, alongside debt markets, which are more intimidating but not publicly traded. Due to the high liquidity of the stock market, investors can quickly and easily buy and sell securities. A rising stock market and widespread participation in this are the two main indicators of an improving economy.

Stock market fluctuations can have a considerable influence on individuals as well as the whole economy. A dramatic drop in stock prices can be extremely destabilizing for economic activities. For example, the 1929 stock market collapse was the primary cause of the Great Depression in the 1930s [2]. When stock prices are high, a large number of companies are likely to launch an initial public offering (IPO) in order to enhance their capital by transferring ownership of their businesses. During a bull market, mergers and acquisitions are also influential. Due to the greater investment, economic development is accelerated [1].

What if investors could predict when the price of a stock would increase or decrease? They would invest all their funds in that company in order to maximize their profits.

However, it is feasible to estimate the unknown parameters and achieve a forecast for the future based on historical and current data regarding specific shares. This type of analysis refers to technical analysis or machine learning (ML). ML models have shown effectiveness in a variety of financial processes, including portfolio management [3] and bankruptcy forecasting [4].

ML is an AI subfield concerned with developing and testing algorithms with the aid of data. Automation is taking over a lot of industries; using mathematical models, computers make quick decisions about online trade [5]. This generates markets in which the long-term outlook is replaced by short-term fluctuations and sell-offs. The algorithms that are most often used for predicting and analyzing the stock market and future movements are SVM and ANN. Using tick data, these systems achieve up to 99.9% accuracy. Financial forecasting is characterized by data-intensive, non-stationary, noisy, unstructured, and hidden relationships [6].

Ref. [7] utilized neural networks to predict US stock prices and demonstrated that neural networks outperform conventional models such as generalized linear models, main component regressions, and regression trees. Long short-term memory (LSTM) networks were utilized by [8] in order to accurately predict stock trends that attract investor sentiment and report big profits. Ref. [9] utilize neural networks to predict bond excess returns and report large economic gains. The neural network model has also been applied to cryptocurrencies in some of the literature; these studies demonstrate that the approach is more accurate at predicting future price changes [10,11]. Fathali et al. [12] used various neural network techniques, including recurrent neural networks (RNNs), LSTM, and convolutional neural networks (CNNs), for anticipating stock market price movements. They discovered that LSTM is the best model after running numerous experiments with different inputs and epochs. Ref. [13] used random forests to examine how investor confidence affects US monthly aggregate realized stock-market volatility, in addition to a large number of financial and macroeconomic variables. They found that investor confidence, specifically investor confidence uncertainty, predicts overall realized volatility and its "good" and "bad" variants out-of-sample. Ref. [14] introduced an investor attention index that relies on proxies found in the existing literature. Their findings indicate that this index effectively forecasts the stock market risk premium, demonstrating its predictive accuracy in both the sample and post-sample periods. Notably, the individual proxies exhibit a limited predictive ability when considered independently. Ref. [15] carried out the study and showed that the Markov-switching multifractal (MSM) is superior to the dynamic conditional correlation-generalized autoregressive conditional heteroscedasticity (DCC-GARCH) model in terms of predictive accuracy. Ref. [16] predicted three stock market indexes of SAARC countries using the ARIMA model and novel machine-learning techniques including multilayer perceptron and recurrent neural networks. They showed that hybrid models are a viable choice for forecasting financial time-series data. The study carried out by [17] demonstrated that the integration of ARIMA and ANN models yields a superior predictive performance compared to the individual use of either ARIMA or ANN models. To predict stock market movement, Ref. [18] evaluated a variety of ML algorithms for the standard time series model, and it was determined that LSTM accurately predicts stock market data. To address the challenge of predicting stock closing prices, Ref. [19] proposed the Deep Convolutional Generative Adversarial Network (DCGAN) architecture and demonstrated that it outperforms current tools in both single-step and multi-step forecasting, demonstrating that deep learning (and GANs in particular) is a promising tool for financial time series forecasting.

Ref. [20] compared the forecast performance of volatilities using two different hybrid ANN models and GARCH-type models. The results demonstrate notable leverage effects in the Chinese energy market and that the EGARCH-ANN model outperforms other models in predicting the volatilities of log-returns series.

According to [21], the goal of this study is to develop a novel parallel hybrid model in order to provide a comprehensive hybrid framework that can accurately simulate all pure

and mixed linear and/or nonlinear patterns found in real-world time series. The suggested hybrid model performs better than the individual models of ARIMA, MLPNN, RBFNN, and LSTM, as well as the hybrid models of the ARIMA-MLPNN and MLPNN-ARIMA series, and the hybridization of ARIMA and MLP models in parallel.

Numerous time series forecasting techniques that employ linear and nonlinear models, alone or in combination, have been studied by [22]. The research indicates that integrating linear and nonlinear models can enhance forecasting accuracy. Nevertheless, in some circumstances, the performance of those current methods may be limited by specific assumptions that they make. We offer a novel hybrid technique that operates within a broader framework: ARIMA-ANN. We demonstrate that combining our hybrid approach with EMD with any of the other approaches that we employed independently can be a useful strategy to increase the forecasting accuracy attained by conventional hybrid methods.

In the fields of economics and finance, there is a pressing need to enhance the precision of forecasts to the utmost degree. In order to effectively implement strong macroeconomic policies, it is important to engage in empirical analyses and strategic planning that relies on projections pertaining to significant macroeconomic indicators. Consequently, a range of univariate and multivariate methodologies have been devised to effectively manage data noise and enhance the precision of forecasting. However, it is important to acknowledge that real-world phenomena do not strictly adhere to either linear or nonlinear patterns. Consequently, both linear and nonlinear models frequently fall short of accurately representing the underlying trend within the data. This study integrates linear and nonlinear models to develop a hybrid model, specifically ARIMA-ANN, which effectively incorporates both linear and nonlinear components of a series. Consequently, this hybrid model enhances predictive accuracy in comparison to the use of individual linear (ARIMA) or nonlinear (ANN) models alone.

Our research aims to bridge a significant gap in the existing literature by investigating the use of stock market indices within the context of G7 countries. These nations, including the United States, Canada, Japan, Germany, France, the United Kingdom, and Italy, collectively represent some of the world's largest and most influential economies. Despite their critical role in the global financial landscape, there has been a notable scarcity of studies that explore the application of stock market indices in hybrid models within this specific group of countries.

The central objective of our research is to enhance prediction accuracy by integrating both linear and non-linear modeling approaches, specifically by combining the linear (ARIMA) model with a nonlinear (ANN). Thus, our study focuses on analyzing the historical closing prices of key stock indices, namely the Nasdaq stock exchange in the United States, the Nikkei stock exchange in Japan, and the CAC 40 index in France. These indices represent a sample from the G7 countries, and our aim is to evaluate and compare the predictive capabilities of standalone linear and non-linear models against a hybrid model, known as ARIMA-ANN.

In the specific context of G7 countries, numerous prior research endeavours have employed various forecasting techniques, such as AR, ARIMA, ANN, and VAR, among others. However, a notable gap exists in the utilisation of hybrid models for this purpose. As previously discussed, hybrid models are deemed more appropriate for forecasting due to their ability to capture both linear and nonlinear trends in the data. This characteristic ultimately leads to more precise and accurate forecasts. The primary objective of our research is to investigate the efficacy of the hybrid ARIMA-ANN model in comparison to the individual ARIMA and ANN models. This analysis is conducted using a dataset comprising stock market indices.

The remaining sections of the paper are organized as follows. Section 2 discusses the data and the procedures. Section 3 presents the research's empirical findings. The paper arrives at a conclusion in Section 4.

2. Data and Methods

Within this section, we shall provide a comprehensive examination of the stock markets that are the focal point of our inquiry. In this study, we explore the complexities associated with data acquisition and preprocessing methodologies, elucidating the process by which we gathered and prepared the data for subsequent analysis. In addition, we expand on the procedures utilised in the current study, offering a comprehensive description of the strategies and approaches adopted to conduct our research.

2.1. Data

This research uses daily data on the closing prices of three Stock market indexes including Nasdaq stock exchange in the United States, Nikkei Stock exchange in Japan and CAC 40 index (a benchmark France stock market index). The data were taken from the Yahoo Finance website for the period from 4 January 2010 to 20 August 2021. In order to assess the prediction capabilities of the hybrid model in comparison to the individual ARIMA and ANN models, the dataset was divided into two distinct subsets: a training set including 75 percent of the data, and a testing set comprising the remaining 25 percent. The training data were utilized to calibrate the models, whereas the testing data were employed to assess the predictive capability of the underlying tools.

2.2. Methodology
The Linear and Non-Linear Models

The field of time series prediction is experiencing rapid growth and holds significant potential for future improvement. A commonly employed strategy for updating the accuracy of predictions involves the integration of multiple methods. This approach relies on the inherent abilities of various models or methodologies with the aim of constructing a prediction framework that is both more resilient and precise. Extensive research has been conducted in this particular domain, resulting in the proposal of various combinations of approaches, as documented in the existing literature [23–25].

ARIMA: In recent decades, ARIMA has become a popular statistical methodology for forecasting stationary and non-stationary time series data. This model frequently incorporates autoregressive (AR) and moving average (MA) models, as well as a data transformation term called differentiation. Nevertheless, the ARIMA model has certain limitations, such as the assumption of linearity, a condition that is challenging to satisfy in practical scenarios, or relying solely on historical data as input variables. The ARIMA model can be transformed into an AutoRegressive Moving Average (ARMA) model by eliminating the differencing component. In general, the ARMA model can be considered a specific instance of the more comprehensive ARIMA model, and its formulation is represented by Equation (1).

$$X_t = b + \sum_{i=1}^{p} \gamma_i X_{t-i} + \mu_t - \sum_{j=1}^{q} \theta_j \mu_{t-j} \tag{1}$$

The ARMA model is utilized to predict the value of a time series variable (X_t) one step ahead. This prediction is based on the historical values of the time series (X_{t-1}, X_{t-2}, ..., X_{t-p}) and the previous errors (μ_{t-1}, μ_{t-2}, ..., μ_{t-q}). The parameters γ_i and θ_j are of an unknown nature, whereas b represents an intercept term. The stochastic error term μ_t is independently and identically distributed, with a mean of zero and a variance of δ^2. The model incorporates prior values up to orders p and q.

In order to make the preceding formula easier to understand, the backward shift operator (A), which is illustrated as $A^i X_t = X_{t-i}$, is substituted to represent the ordinary algebraic symbols in Equation (1). As a result, the ARMA model can be mathematically represented in the following manner:

$$X_t = b + \sum_{i=1}^{p} \gamma_i X_t A^i + \mu_t - \sum_{j=1}^{q} \theta_j \mu_t A^j \tag{2}$$

Then, after adjusting the terms associated with X_t in Equation (2), we can obtain the following ARMA model:

$$\left(1 - \sum_{i=1}^{p} \gamma_i A^i\right) X_t = b + \left(1 - \sum_{j=1}^{q} \theta_j A^j\right) \mu_t \tag{3}$$

For a simplified version of the expression:

$$\gamma_p(A) X_t = b + \theta_q(A) \mu_t \tag{4}$$

where

$$\gamma_p(A) = 1 - \sum_{i=1}^{p} \gamma_i A^i, \tag{5}$$

and

$$\theta_q(A) = 1 - \sum_{j=1}^{q} \theta_j A^j$$

representing, respectively, the AR operator and MA operator.

Despite the ARMA model's inability to incorporate the unit root impact in time series data, it is necessary to use difference transformation to obtain stationarity and attain accurate findings. The integration term is then adjusted in this manner.

$$\gamma_p(A)(1 - A)^s X_t = b + \theta_n(A) \mu_t \tag{6}$$

The ANNs approach: The relaxation of the linear constraint in the model form leads to a vast range of alternative non-linear structures that can be utilized for the purpose of explaining and predicting a time series. A well-established nonlinear model should be globally adequate to deal with the specific nonlinear structure of the data. For further detail, we refer to [26]. ANNs are specifically designed to effectively approximate nonlinearities present in datasets.

A variety of nonlinear issues can be simulated by ANNs, which are flexible computer frameworks. One primary advantage of ANN models in comparison to other non-linear models is in their capacity to effectively estimate a diverse array of functions [27]. Its strength comes from the simultaneous processing of data. No prior assumptions regarding the model shape are required during the construction process. Instead, the ANN models are primarily specified by the data attributes.

The utilization of a single hidden-layer feed-forward network is a commonly employed functional framework for the purpose of time series prediction [28]. A matrix of three layers of fundamental processing units is defined by cyclical connections. The relationship between the output (Q_m) and the inputs $(Q_{m-1}, Q_{m-2}, \ldots, Q_{m-n})$ is depicted mathematically below.

$$Q_m = \beta_0 + \sum_{l=1}^{k} \beta_l g\left(\alpha_{0l} + \sum_{i=1}^{n} \alpha_{il} Q_{m-i}\right) + e_m \tag{7}$$

β_l $(l = 0, 1, 2, \ldots, k)$ and α_{il} $(i = 0, 1, 2, \ldots, n; l = 0, 1, 2, \ldots, k)$ are the model parameters, also known as connection weights. n is the number of input nodes, while k is the number of hidden nodes. The logistic function is widely utilized as a hidden layer transfer function and can be written as follows.

The model parameters, denoted as β_l $(l = 0, 1, 2, \ldots, k)$ and α_{il} $(i = 0, 1, 2, \ldots, n; l = 0, 1, 2, \ldots, k)$, are commonly referred to as connection weights in the academic literature. The variable n represents the quantity of input nodes, whereas k denotes the quantity of hidden nodes. The logistic function is frequently employed as a transfer function in hidden layers and can be expressed as follows:

$$g(x) = \frac{1}{1 + \exp(-x)} \tag{8}$$

Consequently, the ANN model described in Equation (8) exhibits the ability to execute a non-linear functional mapping. This mapping is achieved by utilizing prior observations $(Q_{m-1}; Q_{m-2}, \ldots, Q_{m-n})$ to predict the future value Q_m.

$$Q_m = f(Q_{m-1}, Q_{m-2}, \ldots, Q_{m-n}, v) + e_m \tag{9}$$

Here, v stands for a vector containing all parameters, and f is a function based on the network structure and connection weight. As a result, the neural networks (NNs) correspond to a nonlinear AR model. One output node in the output layer is used in Equation (9) to produce a one-step ahead prediction.

In terms of prediction, simple ANN algorithms are extremely effective. Time series data are frequently better forecasted by NNs with one or two hidden nodes [28].

Hybrid model: In a nutshell, the process of developing a hybrid model involves two distinct stages. In the initial stage, the ARIMA model is employed to examine the linear aspect of the data. In the second stage, the residuals recovered from the estimated ARIMA model are used to build a neural network. The residuals of the ARIMA model include significant information pertaining to nonlinearities, as the ARIMA model is unable to effectively represent the nonlinear pattern present in the data. The ANNs' algorithm can be used to forecast the residuals of an ARIMA model. The hybrid model uses the distinct traits and strengths of the ANN and ARIMA models to identify alternative structures. Linear and non-linear patterns can be adequately described using multiple models, and their predictions can be combined to improve overall modelling and predictability [28]. Figure 1 shows the steps followed in this study.
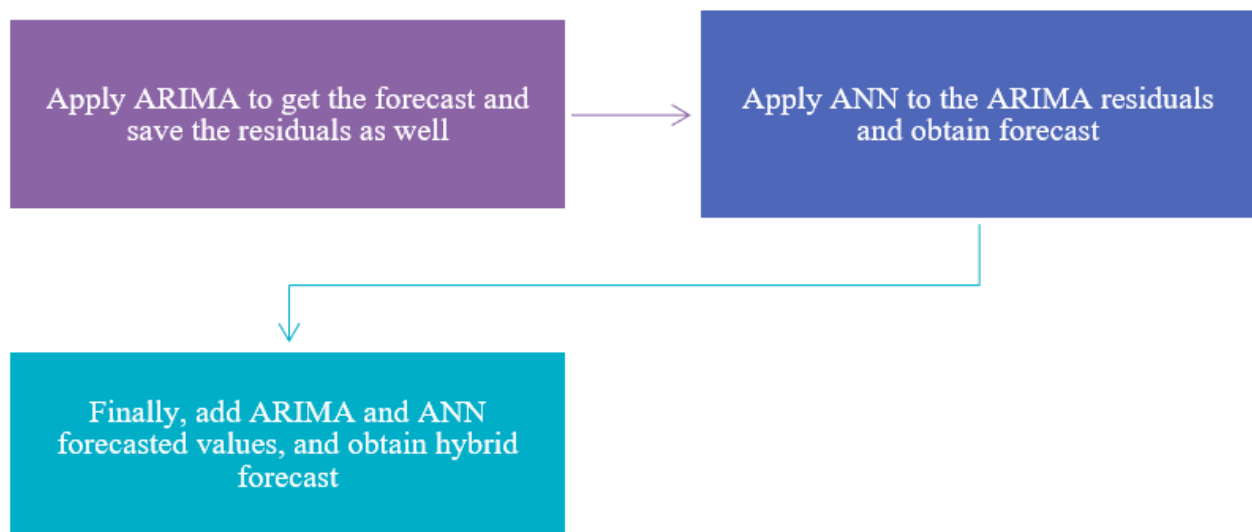

Figure 1. Flowchart of hybrid model. **Noted:** In the initial stage, the ARIMA model is employed to examine the linear aspect of the data. The residuals recovered from an estimated ARIMA model are used to build a neural network in the second stage. Finally, to make a hybrid, the forecasted values of ANN and ARIMA are added.

3. Empirical Results

This section provides a thorough analysis and graphical representation of the three stock markets.

3.1. Nasdaq USA Stock Market

In Figure 2a, the original series is shown to increase over time, which shows that the underlying series is non-stationary. More specifically, the statistical characteristics exhibit temporal variability. To achieve smoothness and eliminate fluctuations from the data, we initially transform the series by taking the natural logarithm and then perform

the first difference to achieve stationarity. Figure 2b portrays the graph of the transformed time series, which manifests that the series is difference stationary. In Figure 3a, the ACF plot is steadily declining. This is another indication of a unit root. As Figure 3c shows, as we performed the transformation, the ACF plot is very quickly declines, which suggests a differenced stationary series. Thus, we can proceed with the stationary series. Certain patterns in the ACF and PACF plots correspond to specific orders of q and p.

| (a) | (b) |
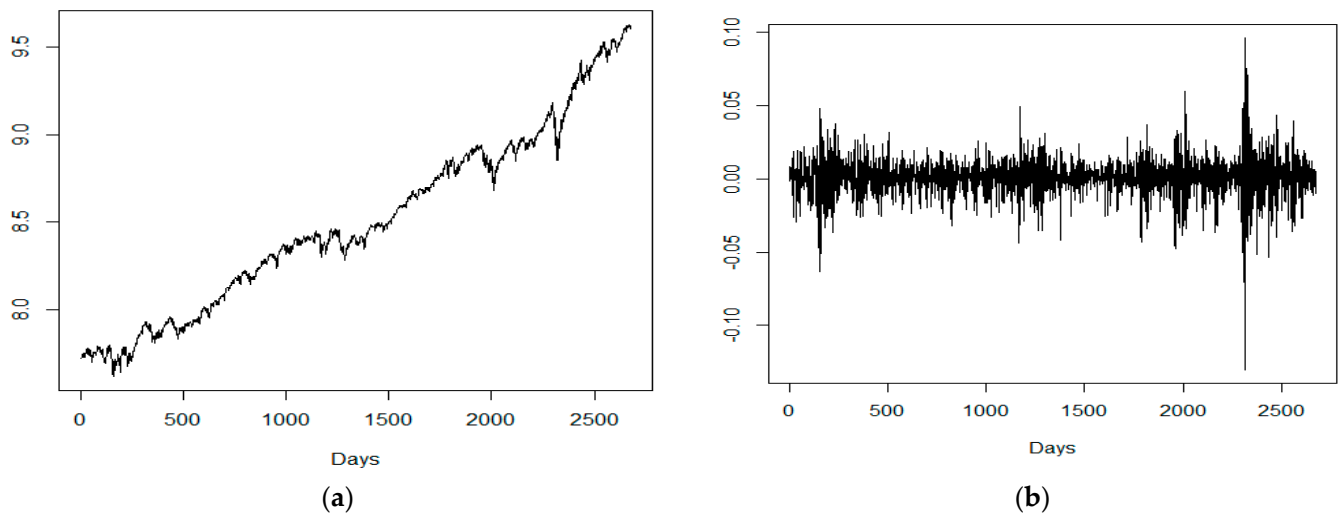
Figure 2. Level and first difference of USA stock market. Figure (**a**) shows that the series increasing over time, but as we take the first difference, indicated by figure (**b**), then the series is mean stationary. **Noted:** Level and first difference of USA stock market, where the series at level shows an increasing trend and achieves smoothness after difference transformation.

There are a few ways in which we can observe the residuals' randomness in the estimated model. We adopt a graphical approach, as well as a statistical approach, in Figure 4. The residuals' ACF reveals no serious autocorrelations. The last plot on the bottom provides *p*-values for the Ljung–Box statistic for each lag up to 10. These tests consider the accumulated residual autocorrelation from lag 1. The dashed blue line indicates a 5 percent level of significance, and it can be observed that all *p*-values (denoted by circles) are above this. Thus, we can conclude that residuals are purely random. Hence, this model is suitable for prediction.

Post ARIMA modeling, we utilize another approach for forecasting, known as ANN. ANN is considered the most well-known machine learning technique for forecasting. Therefore, this study adopts this technique to capture the complex behavior of the Nasdaq US stock market and resultantly achieve a better forecast. The process of configuring the ANN is comprehensively elucidated in Section 2.2. In the ANN model fitting, we employ an iterative approach, utilizing a trial-and-error method to determine the optimal number of hidden layers. To elucidate this, we commence with a single hidden layer and individually increment the layer count until we achieve the most precise outcome. During this progression, it was observed that the minimum test error was attained when employing three hidden layers and five input layers.
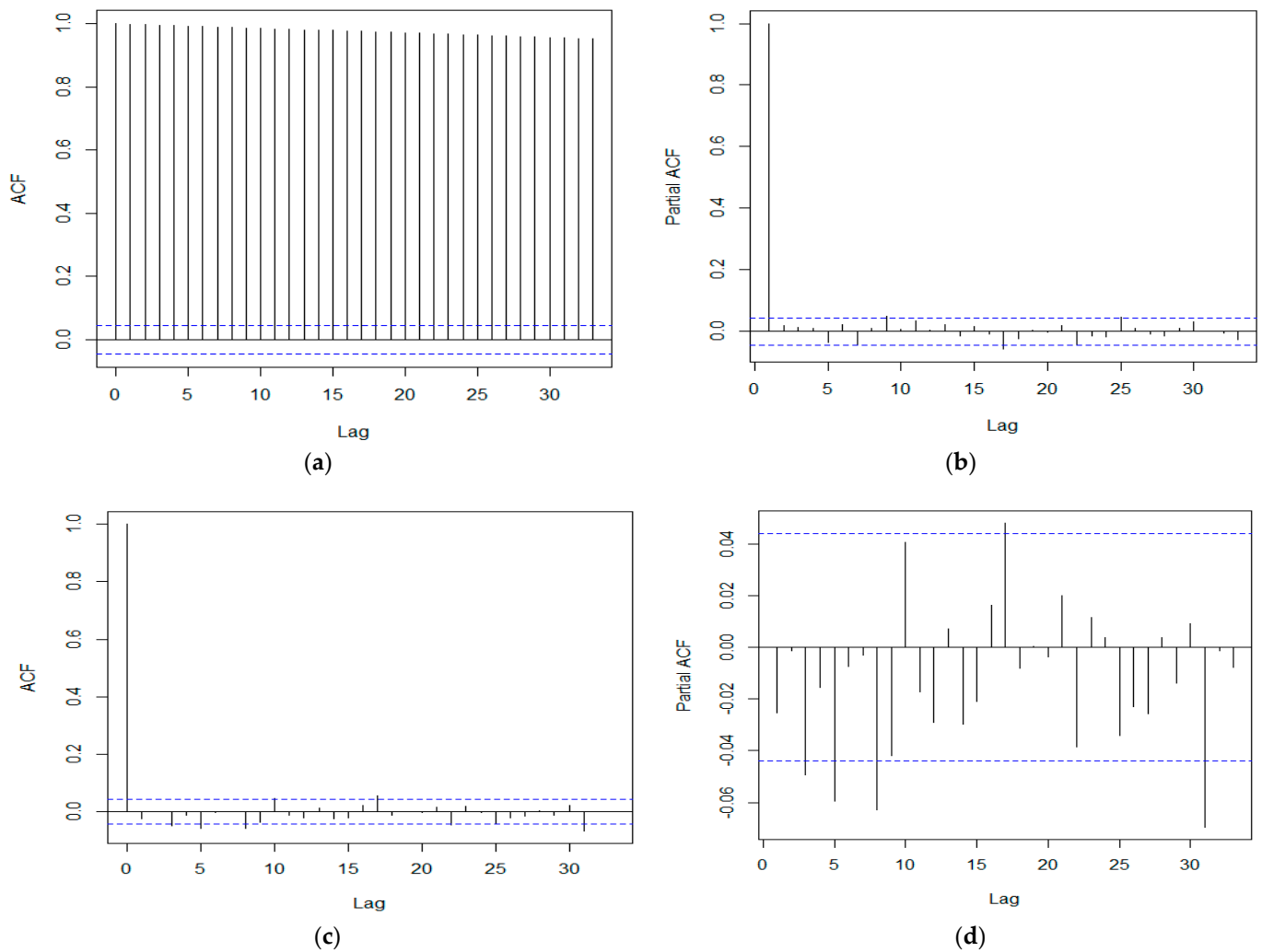
(a)

(b)

(c)

(d)

Figure 3. ACF and PACF plots. **Noted:** ACF and PACF for level (**a**,**b**), where the ACF is steadily declining, which ensures the unit root problem, and differenced data (**c**,**d**), where the ACF plot is declining very fast, which is evidence of stationarity.

The same methodological approach was replicated in the construction of the hybrid model. Here, the task was to identify the ideal configuration of the hybrid model. The iterative process led to the selection of two hidden layers and four input layers as the configuration that yielded the most favorable results.

Figure 5 shows our comparison of different time series and machine learning models. This shows how well the predictions worked visually, with the height of each bar showing the extent to which the predicted values differed from the actual values. A lower bar height is indicative of a smaller margin of error, reflecting a higher level of accuracy in the prediction.

Upon a detailed examination of Figure 5, several key observations and insights come to the fore. First and foremost, it is evident that the ANN model exhibits a commendable ability to capture the directional movements of the Nasdaq US stock market. This implies that, when using the ANN model in isolation, it can offer a relatively accurate forecast. This is a testament to the power of neural networks to uncover complex patterns and relationships within financial time series data.
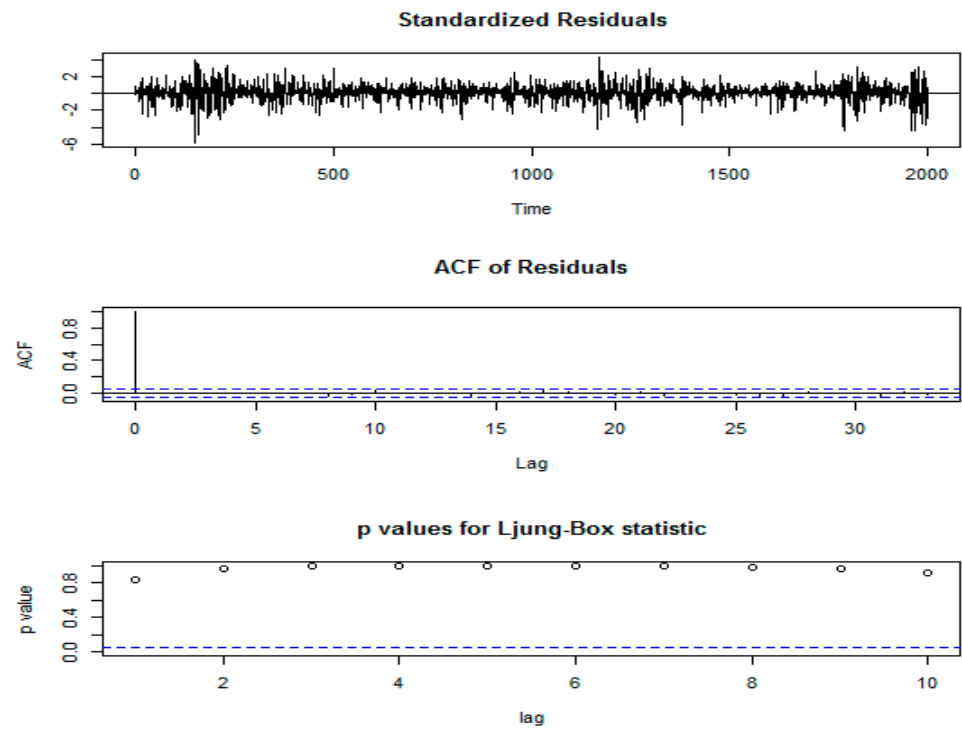
Standardized Residuals

ACF of Residuals

p values for Ljung-Box statistic

Figure 4. Diagnostic check. **Noted:** The ACF of the residuals shows no significant autocorrelations. The dashed blue line indicates a 5 percent significance level, and it can be observed that all *p*-values (denoted by circles) are above this, which ensures the randomness of residuals.
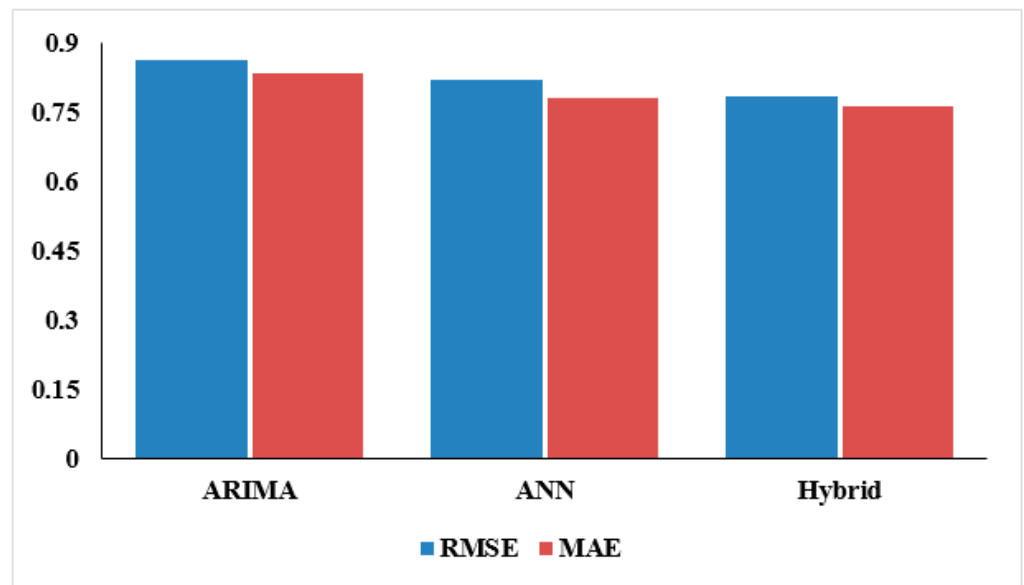
Figure 5. Forecast comparison across several models. **Noted:** This presents a comparison of time series and machine learning models. The smaller height of a bar is evidence of an accurate prediction. Herein, the hybrid model outperforms the rival models.

However, the most intriguing findings emerge when we turn our attention to the hybrid model, specifically the ARIMA-ANN combination. When compared to both the standalone ARIMA and ANN models in this situation, it is clear that the forecast errors produced by the ARIMA-ANN hybrid model are significantly lower. This reduction in forecast errors signifies a higher level of predictive accuracy when utilizing the hybrid approach.

The observed improvement in forecast accuracy achieved with the ARIMA-ANN hybrid model can be attributed to its unique ability to combine the strengths of two distinct forecasting methodologies. The ARIMA component excels in modeling linear trends and capturing seasonality, while the ANN component is adept at handling complex, nonlinear relationships in the data. By integrating these two approaches, the hybrid model leverages their complementary strengths, resulting in a more precise forecast.

3.2. Nikkei Japan Stock Market

Figure 6a demonstrates a clear upward trend in the series at a certain level, indicating that the underlying series is non-stationary. In order to achieve flatness and remove fluctuations from an underlying series, researchers commonly employ a logarithm transformation, followed by the application of the first difference to establish stationarity. Figure 6b displays a plot of the converted series, which exhibits a difference stationarity. Figure 7a demonstrates a consistent decrease in the autocorrelation function (ACF) plot, which serves as additional evidence of the presence of a unit root. Figure 7c exhibits a distinct decline in the autocorrelation function (ACF) plot after undergoing transformation, indicating the achievement of stationarity. The arrangement of q and p in a certain sequence correlates with a distinct pattern observed in the plots of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), respectively.
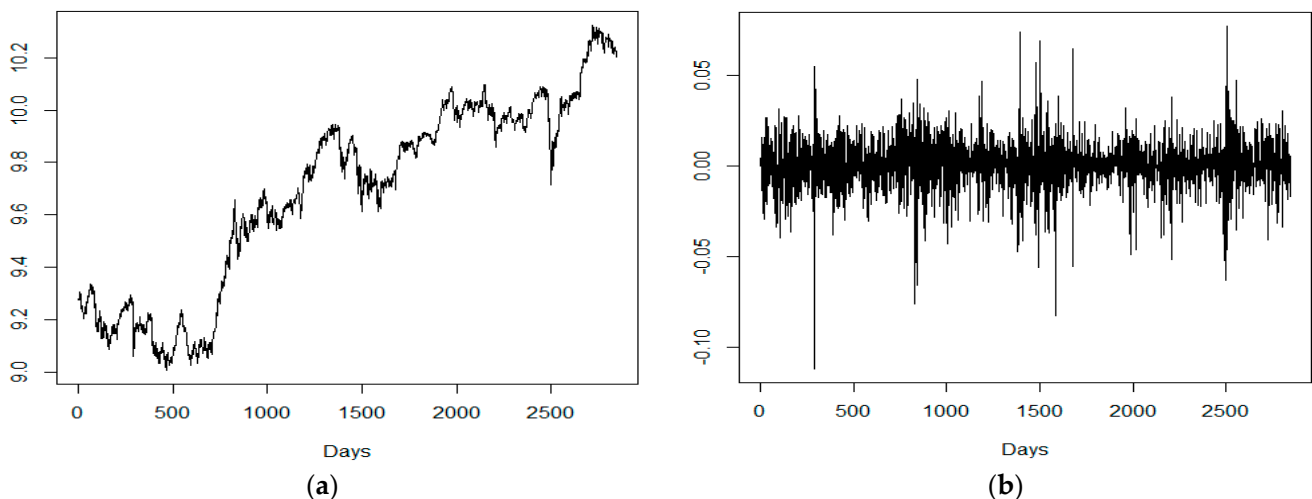

(**a**)   (**b**)

Figure 6. Level and first difference of Japanese stock market. Figure (**a**) demonstrates increasing trend, while figure (**b**) mean stationary. **Noted:** Level and first difference of Japanese stock market, where the series at level shows an upward trend and achieves smoothness after difference transformation.

In Figure 8, the ACF or autocorrelation coefficient of the residuals of fitted ARIMA for lag 1–30 is within the limits. Moreover, the Ljung–Box test also supports this result. Thus, we can conclude that residuals are purely random. Hence, this approach can be applied to forecasting. Post ARIMA prediction, we utilized the ANN algorithm and then a hybrid of both. We used an iterative process to fit the ANN model, determining the ideal number of hidden layers through trial and error. We started with one hidden layer and progressively added layers individually until we reached the most accurate result. It was discovered that using two hidden layers and three input layers resulted in the lowest test error.
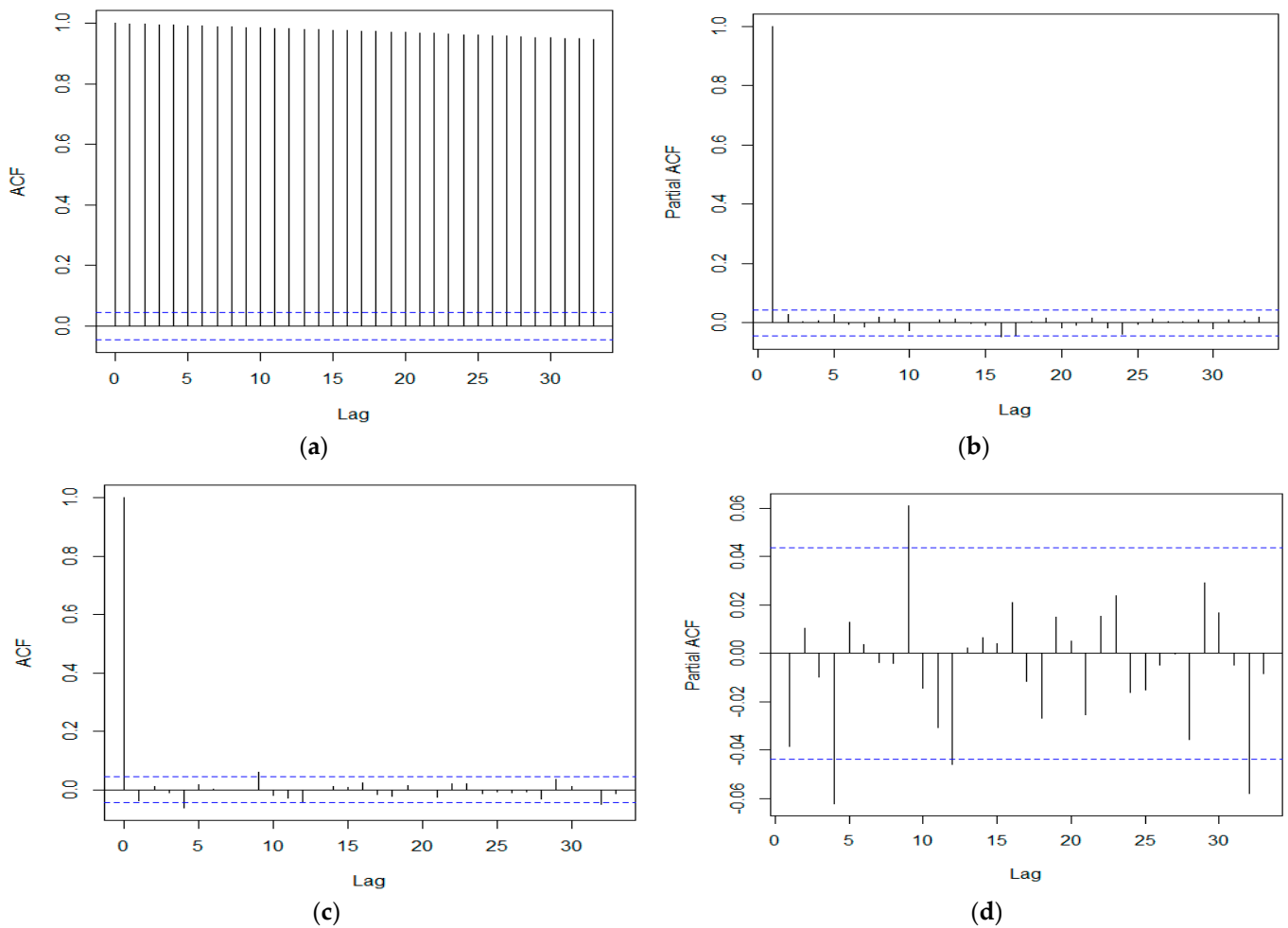
Figure 7. ACF and PACF plots. **Noted:** ACF and PACF for level (**a**,**b**), where the ACF is steadily declining, which ensures the unit root problem, and differenced data (**c**,**d**), where the ACF plot is declining very fast, which is evidence of a stationary series.

The insights drawn from Figure 9 are particularly illuminating, shedding light on the performance of various forecasting models in the context of the Nikkei Japan stock market. This visual representation allows for us to discern and interpret the relative accuracy of these models by observing the heights of the bars, where lower heights signify smaller forecast errors and, consequently, a higher degree of predictive precision.

Upon a closer examination of Figure 9, it becomes evident that the ANN algorithm displays a commendable capacity to capture the overarching trend of the Nikkei Japan stock market. This indicates that, when utilized as a standalone model, the ANN is adept at providing forecasts that align well with the actual market movements. This observation underscores the ability of neural networks to uncover and incorporate intricate patterns and nuances within the time series data of the Nikkei index, contributing to its strong forecasting performance.

However, the most striking findings emerge when we shift our focus to the hybrid model, specifically the ARIMA-ANN combination. In this context, it becomes readily apparent that the forecast errors generated by the ARIMA-ANN hybrid model are notably reduced when compared to the separate ARIMA and ANN models. This reduction in forecast errors is a clear manifestation of the heightened predictive accuracy that the hybrid approach offers.
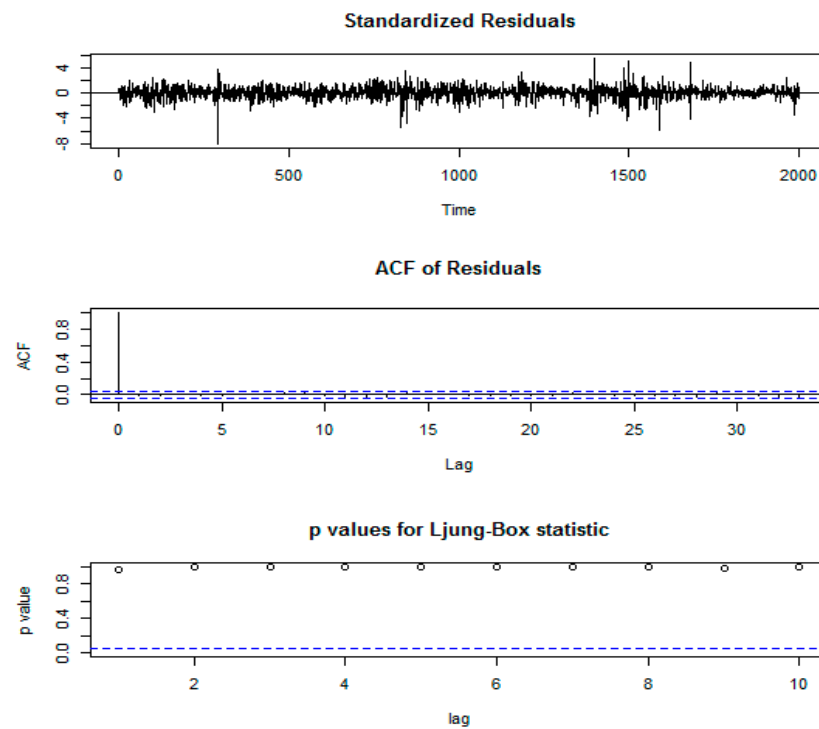
Figure 8. Diagnostic check. **Noted:** The ACF of the residuals does not exhibit any statistically significant autocorrelations. The dashed blue line represents the 5 percent significance level. It is evident that all *p*-values, indicated by circles, are higher than this threshold, indicating that the residuals exhibit randomness.
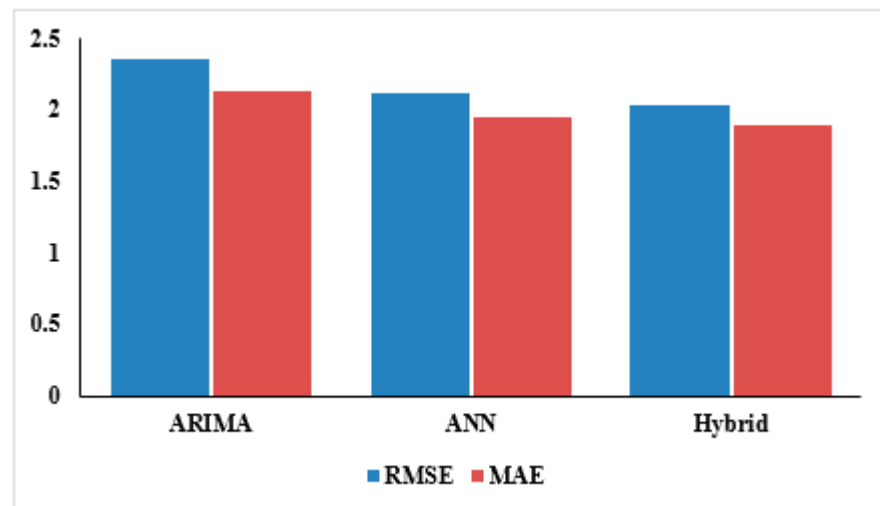

Figure 9. Forecast comparison across several models. **Noted:** A comparison between ML models and time series is presented. An accurate prediction is demonstrated by a shorter bar. The hybrid model outperformed the other models in the present case.

The unique ability of the ARIMA-ANN hybrid model to combine the best features of two different modelling approaches is what makes it better at making predictions. The ARIMA component excels in capturing linear trends, and it effectively addresses issues related to seasonality. Meanwhile, the ANN component demonstrates its prowess in dealing with the complexity of non-linear relationships within the data. By integrating these two approaches, the hybrid model capitalizes on their complementary strengths, culminating in a more precise and reliable forecast.

3.3. France Stock Market (CAC 40 Index)

We can see in Figure 10a that the original stock market time series is increasing over time, which shows that the series is suffering from a unit root problem. To achieve flatness and remove fluctuations in the data, the logarithm transformation is implemented, and difference transformation is performed to obtain a stationary series. Figure 11 represents the transformed series, which ensures stationarity. In Figure 11a, a gradual decrease in the ACF plot is further evidence of a unit root. Following transformation, in Figure 11c, we can notice a sharp fall in the ACF plot. This confirms that the series is a first difference stationary series. Certain orders of q and p are connected to a specific pattern in the ACF and PACF plots, respectively.
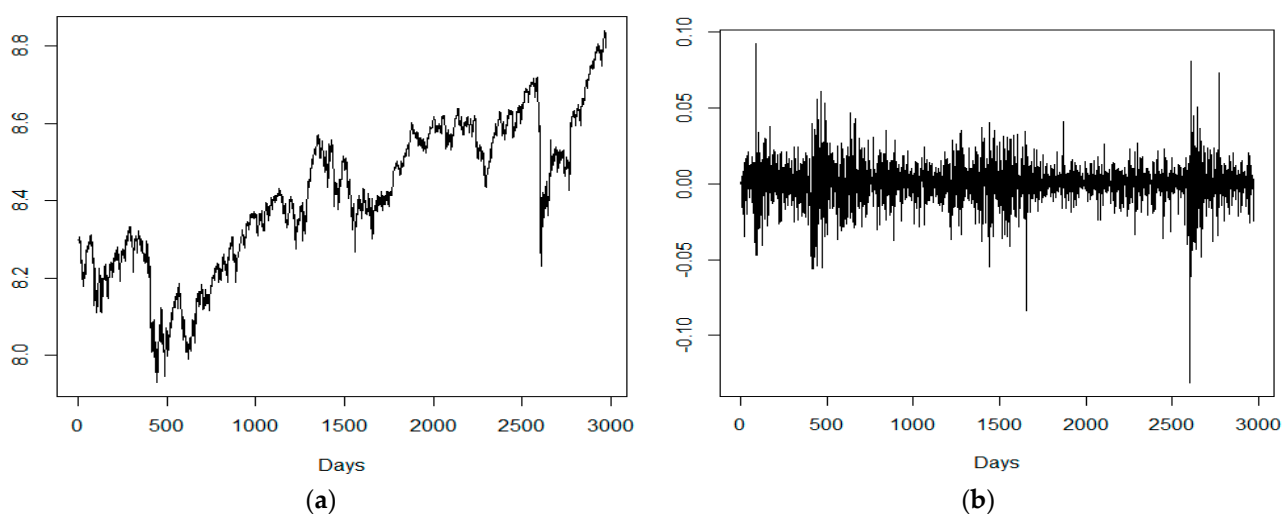
(**a**)

(**b**)

Figure 10. Level and first difference in French stock market. Figure (**a**) is showing increasing trend, while figure (**b**) mean stationary. **Noted:** Level and first difference in French stock market, where the series at level shows an upward trend and achieves smoothness after difference transformation.

Looking at the residuals correlogram and the Ljung–Box test shown in Figure 12, it is clear that there is no noticeable spike, and the *p*-values from the Box–Ljung test are higher than the 5% significance level. The results of this study offer support for the null hypothesis, indicating that the residuals have a random pattern. Therefore, it can be inferred that residuals exhibit characteristics of white noise. Therefore, this model has the potential to be utilised for t making predictions. After ARIMA prediction, the subsequent step employs the ANN technique. Subsequently, a combination of both ARIMA and ANN strategies is utilised. We fit the ANN model iteratively, exploring until we found the optimal number of hidden layers. We began with a single hidden layer and worked our way up to the most accurate outcome, layer by layer. Along the way, it was found that the lowest test error was achieved with three input levels and three hidden layers.

The insights derived from Figure 13 offer a compelling perspective of the performance of various forecasting models within the intricate landscape of the French stock market. This visual representation provides a clear means of gauging the relative accuracy of these models, with lower bar heights indicating smaller forecast errors and, by extension, a higher level of predictive accuracy.

Upon a detailed examination of Figure 13, a notable observation comes to the forefront: the ANN algorithm demonstrates a strong ability to capture the underlying trends of the French stock market. This implies that, when employed as a standalone model, the ANN excels at providing forecasts that closely align with actual market behavior. This finding underscores the capacity of neural networks to uncover and incorporate the subtleties and intricacies within the time series data of the French stock market, contributing to its robust forecasting performance.
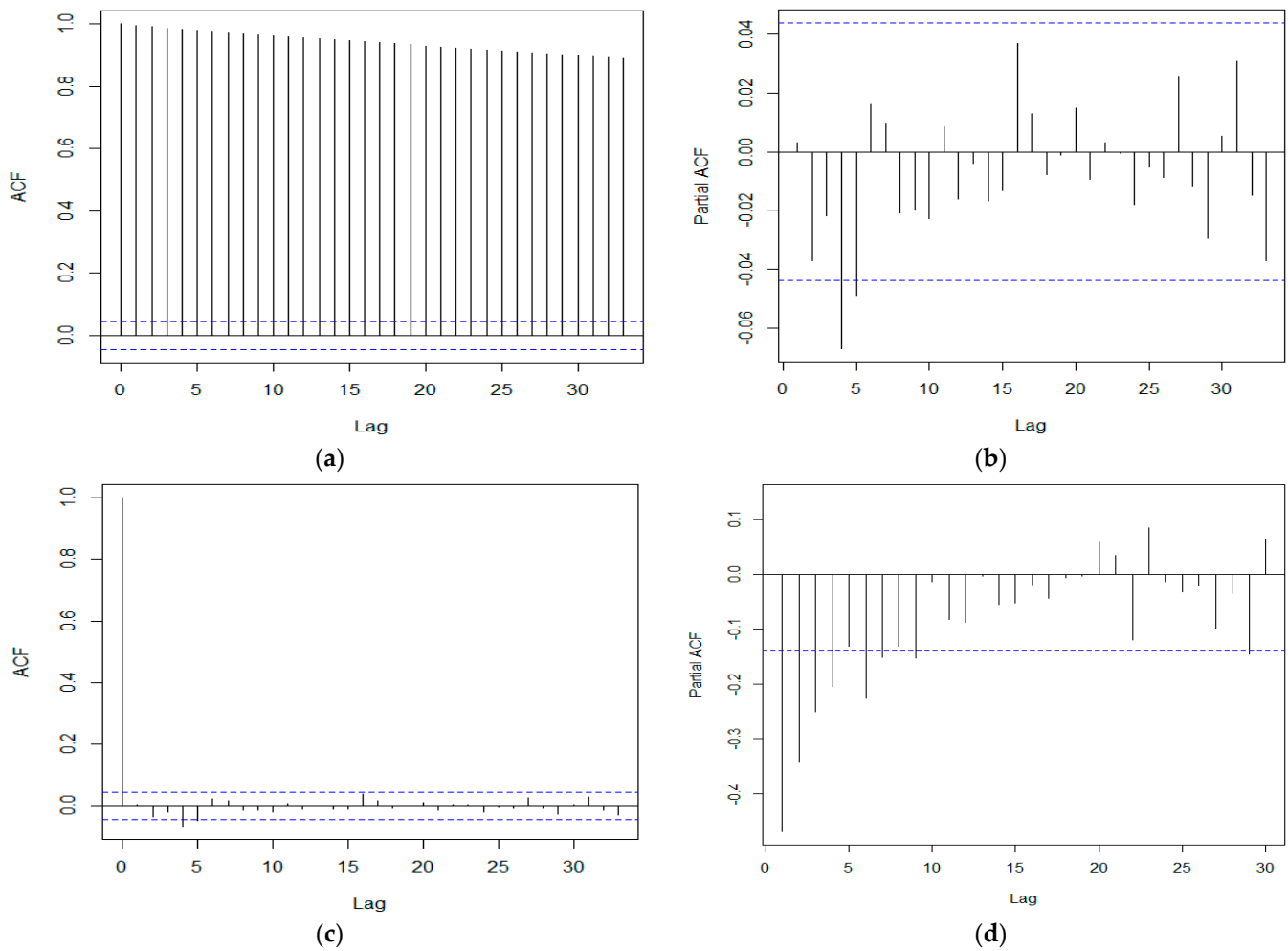
Figure 11. ACF and PACF plots. **Noted:** ACF and PACF for level (**a**,**b**), where the ACF is steadily declining, which ensures the unit root problem, and differenced data (**c**,**d**), where the ACF plot is declining very fast, which confirms stationarity.

However, the most remarkable findings are unveiled as we shift our focus towards the hybrid model, specifically the fusion of ARIMA and ANN. When compared to the individual ARIMA and ANN models, it is clear that the ARIMA-ANN hybrid model significantly reduces the forecast errors. This substantial reduction in forecast errors reflects a higher degree of predictive accuracy, affirming the superior forecasting capability of the hybrid approach.

The improvement in forecasting precision obtained with the ARIMA-ANN hybrid model is a direct consequence of its unique ability to harness the strengths of two distinct modeling methodologies. The ARIMA component effectively captures linear trends and addresses seasonality in the data, while the ANN component excels at managing the complexities of non-linear relationships. By seamlessly integrating these two approaches, the hybrid model optimally leverages their complementary strengths, culminating in a forecast that is both accurate and robust.
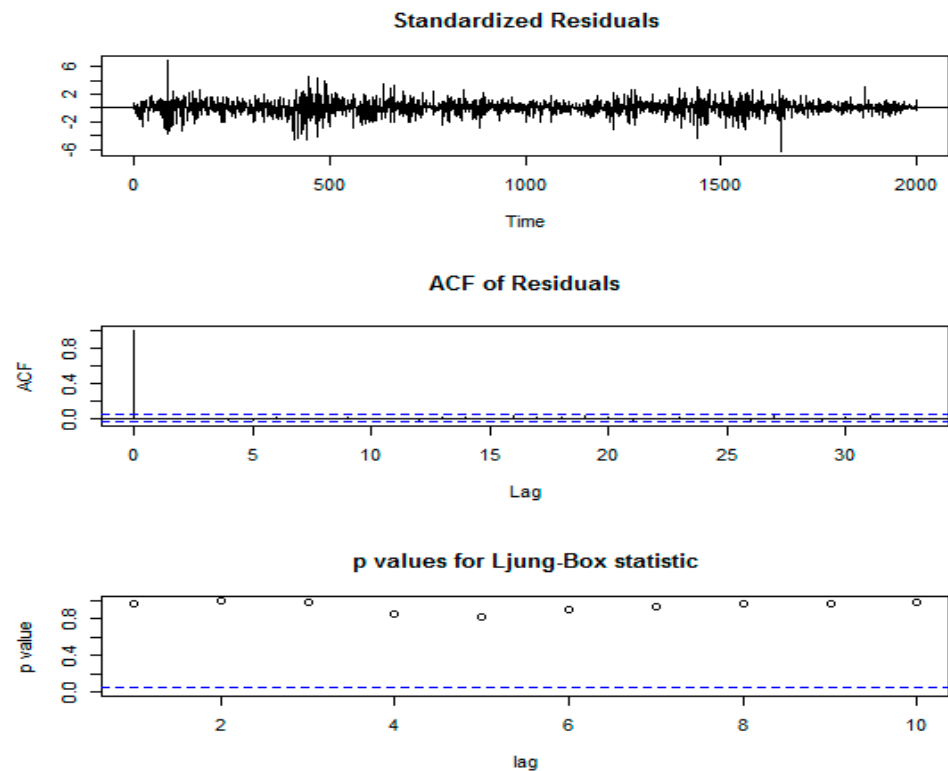
Standardized Residuals

ACF of Residuals

p values for Ljung-Box statistic

Figure 12. Diagnostic check. **Noted:** The ACF of the residuals shows no significant autocorrelations. The dashed blue line indicates 5 percent significance level, and it can be observed that all *p*-values (denoted by circles) are above this, which ensures the randomness of residuals.
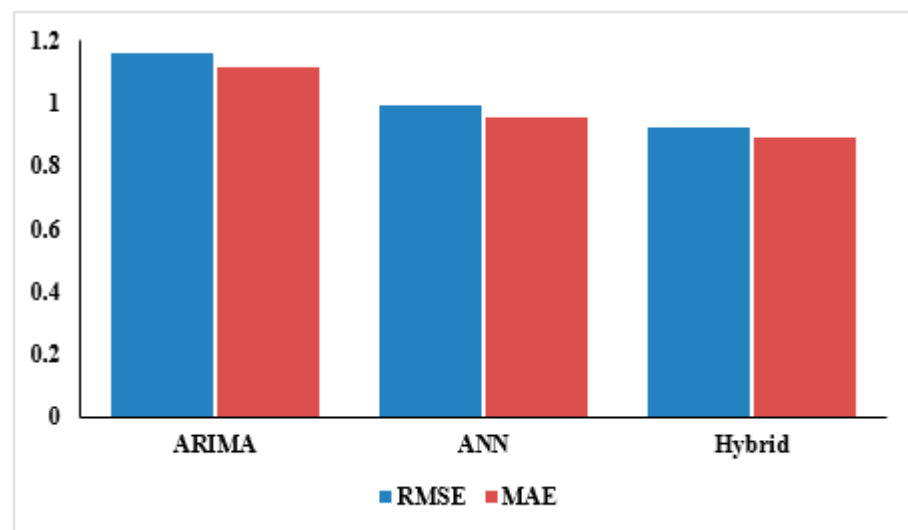
Figure 13. Forecast comparison across several models. **Noted:** A comparison of time series and ML models is produced. The shorter bar denotes accurate prediction. Herein, the hybrid model outperforms the rival models.

3.4. Difference among the Three Datasets Results

This study presents a novel approach that combines the ARIMA and ANN models and is then applied to three financial markets within the G7. The findings of this study demonstrate that the hybridization of these models yields highly beneficial results in terms of predicting. It is worth noting that, in the realm of financial markets, the hybrid approach exhibits a notably low level of forecast inaccuracy when applied to the Nasdaq USA stock market as compared to other financial markets under consideration.

Specifically, in the case of the Nikkei Japan stock market, there is a particularly significant degree of forecasting error.

4. Conclusions

Almost all financial decision-makers, such as investors, money managers, hedge funds, and investment banks, needed to forecast financial asset prices such as exchange rates, options, bonds, interest rates, and stocks, among other things, with the aim of making productive decisions. Therefore, to date, the modification and development of new models have not stopped in research on the management of financial markets. According to previous research, prediction plays a key role in financial markets; however, this is a difficult task. Thus, financial stakeholders face many difficulties in achieving accurate forecasts. In the forecasting literature, merging multiple models is one of the most popular ways to gain additional accuracy in comparison with individual models. The literature has put forth a number of methods for dealing with the limitations of the separate approaches and generating more trustworthy results. A combining approach that decomposes a time series into two parts, linear and non-linear, is the most popular approach, and has been theoretically as well as empirically accepted to be more successful than an individual model. These models have advantages in terms of linearity and nonlinearity in the time series nexus.

The current study compares the predictive power of a hybrid of linear/nonlinear (i.e., ARIMA/ANN), such as ARIMA-ANN, with their components using the data of three stock market indices from G7 countries. Empirical research based on three popular real datasets of stock prices from the three stock market indexes, namely the Nasdaq stock exchange, United States, Nikkei Stock exchange, Japan, and France stock exchange, demonstrates that using a hybrid model yields a more accurate forecast than using separate components. It is generally believed that a hybrid model can deliver results that are, to some extent, better than those obtained by individual models. Based on an analysis of real data, the findings revealed that the hybrid ARIMA-ANN is overall superior to individual ANN and ARIMA models. For all the considered stock exchange indexes, the RMSE and MAE values observed in the hybrid model exhibited a significant reduction in comparison to the individual models.

The scope of this study primarily centres on univariate analysis, wherein forecasting models are built solely on historical data related to the stock market indices under consideration. Numerous external variables, including economic indicators, political events, and global trends, can have a profound impact on market movements. Incorporating external economic and financial indicators, such as geopolitical events or macroeconomic data, into the forecasting models can enhance their predictive power. Future studies could explore the impact of exogenous variables on model accuracy. A combination of LSTM and ANN can be utilised for the prediction of complex stock market data.

References

1.   Chhajer, P.; Shah, M.; Kshirsagar, A. The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction. *Decis. Anal. J.* **2022**, *2*, 100015. [CrossRef]
2.   Pettinger, T. UK Wage Growth. Economics Help. 2019. Available online: https://www.economicshelp.org/blog/6994/economics/uk-wage-growth/ (accessed on 8 October 2019).
3.   Yun, H.; Lee, M.; Kang, Y.S.; Seok, J. Portfolio management via two-stage deep learning with a joint cost. *Expert Syst. Appl.* **2020**, *143*, 113041. [CrossRef]
4.   Kou, G.; Xu, Y.; Peng, Y.; Shen, F.; Chen, Y.; Chang, K.; Kou, S. Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decis. Support Syst.* **2021**, *140*, 113429. [CrossRef]
5.   Kshirsagar, A.; Shah, M. Anatomized study of security solutions for multimedia: Deep learning-enabled authentication, cryptography and information hiding. In *Advanced Security Solutions for Multimedia*; IOP Publishing: Bristol, UK, 2021.
6.   Solanki, P.; Baldaniya, D.; Jogani, D.; Chaudhary, B.; Shah, M.; Kshirsagar, A. Artificial intelligence: New age of transformation in petroleum upstream. *Pet. Res.* **2022**, *7*, 106–114. [CrossRef]
7.   Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]
8.   Zhang, Y.; Chu, G.; Shen, D. The role of investor attention in predicting stock prices: The long short-term memory networks perspective. *Financ. Res. Lett.* **2021**, *38*, 101484. [CrossRef]
9.   Bianchi, D.; Büchner, M.; Hoogteijling, T.; Tamoni, A. Corrigendum: Bond risk premiums with machine learning. *Rev. Financ. Stud.* **2021**, *34*, 1090–1103. [CrossRef]
10. Anghel, D.G. A reality check on trading rule performance in the cryptocurrency market: Machine learning vs. technical analysis. *Financ. Res. Lett.* **2021**, *39*, 101655. [CrossRef]
11. Liu, M.; Li, G.; Li, J.; Zhu, X.; Yao, Y. Forecasting the price of Bitcoin using deep learning. *Financ. Res. Lett.* **2021**, *40*, 101755. [CrossRef]
12. Fathali, Z.; Kodia, Z.; Ben Said, L. Stock market prediction of Nifty 50 index applying machine learning techniques. *Appl. Artif. Intell.* **2022**, *36*, 2111134. [CrossRef]
13. Gupta, R.; Nel, J.; Pierdzioch, C. Investor confidence and forecastability of US stock market realized volatility: Evidence from machine learning. *J. Behav. Financ.* **2023**, *24*, 111–122. [CrossRef]
14. Chen, X.; Wu, C. Retail investor attention and information asymmetry: Evidence from China. *Pac.-Basin Financ. J.* **2022**, *75*, 101847. [CrossRef]
15. Liu, R.; Gupta, R. Investors' uncertainty and forecasting stock market volatility. *J. Behav. Financ.* **2022**, *23*, 327–337. [CrossRef]
16. Peng, Z.; Khan, F.U.; Khan, F.; Shaikh, P.A.; Yonghong, D.; Ullah, I.; Ullah, F. An Application of Hybrid Models for Weekly Stock Market Index Prediction: Empirical Evidence from SAARC Countries. *Complexity* **2021**, *2021*, 5663302. [CrossRef]
17. Khan, F.; Urooj, A.; Muhammadullah, S. An ARIMA-ANN hybrid model for monthly gold price forecasting: Empirical evidence from Pakistan. *Pak. Econ. Rev.* **2021**, *4*, 61–75.
18. Majumder, A.; Rahman, M.M.; Biswas, A.A.; Zulfiker, M.S.; Basak, S. Stock Market Prediction: A Time Series Analysis. In *Smart Systems: Innovations in Computing: Proceedings of SSIC 2021*; Springer: Singapore, 2022; pp. 389–401.
19. Staffini, A. Stock price forecasting by a deep convolutional generative adversarial network. *Front. Artif. Intell.* **2022**, *5*, 837596. [CrossRef]
20. Lu, X.; Que, D.; Cao, G. Volatility forecast based on the hybrid artificial neural network and GARCH-type models. *Procedia Comput. Sci.* **2016**, *91*, 1044–1049. [CrossRef]
21. Hajirahimi, Z.; Khashei, M. A novel parallel hybrid model based on series hybrid models of ARIMA and ANN models. *Neural Process. Lett.* **2022**, *54*, 2319–2337. [CrossRef]
22. Rudin, C.; Ertekin, Ş. Learning customized and optimized lists of rules with mathematical programming. *Math. Program. Comput.* **2018**, *10*, 659–702. [CrossRef]
23. Armstrong, J.S. Combining forecasts. In *Principles of Forecasting*; Armstrong, J.S., Ed.; Kluwer Academic Publishers: Norwell, MA, USA, 2001.
24. Zhang, G.P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **2003**, *50*, 159–175. [CrossRef]
25. Armstrong, J.S. Findings from evidence-based forecasting: Methods for reducing forecast error. *Int. J. Forecast.* **2006**, *22*, 583–598. [CrossRef]
26. De Gooijer, J.G.; Kumar, K. Some recent developments in non-linear time series modelling, testing, and forecasting. *Int. J. Forecast.* **1992**, *8*, 135–156. [CrossRef]
27. Khashei, M.; Bijari, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Appl. Soft Comput.* **2011**, *11*, 2664–2675. [CrossRef]
28. Zhang, G.; Patuwo, B.E.; Hu, M.Y. Forecasting with artificial neural networks: The state of the art. *Int. J. Forecast.* **1998**, *14*, 35–62. [CrossRef]