HYBRID AUTOREGRESSIVE INTEGRATED MOVING AVERAGE-SUPPORT VECTOR REGRESSION FOR STOCK PRICE FORECASTING

Hanan Albarr[1)]
Rosita Kusumawati[2)]
[1,2)]Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
E-mail: rosita_kusumawati@uny.ac.id

ABSTRACT

Stock investment provides high-profit opportunities but also has a high risk of loss. Investors use various decision-making methods to minimize this risk, such as stock price forecasting. This research aims to predict daily closing stock prices using a hybrid Autoregressive Integrated Moving Average (ARIMA)-Support Vector Regression (SVR) model and compare it with the single model of ARIMA and SVR, as well as compiling the R-shiny web for the hybrid ARIMA-SVR model which makes it easier for investors to use the model to support investment decision making. The hybrid ARIMA-SVR model is composed of two components: the linear component from the results of stock price forecasting using the Autoregressive Integrated Moving Average (ARIMA) model and the nonlinear components from the residual forecasting results of the ARIMA model using the Support Vector Regression (SVR) model. The data used was closing stock price data from April 1, 2019, to April 1, 2021, from PT Unilever Indonesia Tbk (UNVR.JK), PT Perusahaan Gas Negara Tbk (PGAS.JK), and PT Telekomunikasi Indonesia Tbk (TLKM.JK), from the Yahoo Finance website. The research results conclude that the hybrid ARIMA-SVR model has excellent capabilities in forecasting stock prices with the MAPE values for UNVR, PGAS, and TLKM stocks, respectively of 0.797%, 2.213%, and 0.993%, which are lower than the MAPE values of ARIMA-GARCH and SVR models. The hybrid model can be an alternative model with excellent capabilities in forecasting stock prices.

Keywords: Stock price forecasting, hybrid model, autoregressive integrated moving average (ARIMA), support vector regression (SVR).

INTRODUCTION

Stock investment is attractive because it provides the opportunity to profit from increases in stock prices and dividend distribution. However, stock investment also risks loss from a decline in stock prices. Stock prices can experience fluctuations in a short time due to several factors, such as company performance, inflation, rising interest rates, exchange rates, and politics (Erkilet et al., 2022; Hajiabadi & Samadi, 2019; Lampart et al., 2023). Stock investments that provide a high rate of return will also have a high risk of loss. The higher the rate of return on investment, the higher the risk obtained (Muslih, 2018; Nukala & Prasada Rao, 2021). It is necessary to consider before investing in stocks to reduce the risk. Various analyses determine investment decisions, such as predicting stock price's future rise and fall.

Time series forecasting is an appropriate analysis for predicting stock prices. Several forecasting methods can be divided into linear and nonlinear approaches. Several forecasting methods using linear models are Autoregressive (WITH), Moving Average (MA), and Autoregressive Integrated Moving Average (ARIMA). In contrast, some forecasting methods use nonlinear models such as Support Vector Regression (SVR), Artificial Neural Network (ANN), and Neural Network (NN) (Kumar & Thenmozhi, 2014). The ARIMA model has excellent capabilities in short-term forecasting with non-stationary linear components (Juberias et al., 1999; Lee & Ko, 2011; Li & Zhang, 2009). The SVR and ANN models have more effective capabilities in forecasting with nonlinear components (Amin & Hoque, 2019; Lu et al., 2004). Forecasting results using the

ARIMA model have a linear pattern of increasing or decreasing, while forecasting using the SVR model produces a fluctuating and varied pattern (Banerjee, 2014; Septiningrum et al., 2015).

The ARIMA model has a very good level of accuracy in short-term forecasting on non-stationary data, but accuracy decreases on time series data containing nonlinear components. Meanwhile, the SVR model is a superior forecasting model on time series data with nonlinear components (Purnama & Setianingsih, 2021). Several studies have been carried out to compare forecasting using the ARIMA model with nonlinear models such as NN and SVR, showing that the ARIMA model has a lower level of accuracy than the nonlinear model because the data contains nonlinear components. However, the ARIMA model can produce good accuracy on the data containing linear components (Dhini et al., 2015; Ho et al., 2002; Tao et al., 2021; Taskaya-Temizel & Ahmad, 2005).

The ARIMA model is a combination of Autoregressive (AR) and Moving Average (MA) with a differencing process by using past and present values for the dependent variable used for prediction or forecasting (Box et al., 2015). The SVR method is part of the Support Vector Machine (SVM) method for regression problems. SVM is a learning system using a linear function hypothesis space in a high-dimensional feature space, which is trained using a learning algorithm with optimization theory that implements learning bias originating from statistical learning theory (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995). SVR aims to find a regression function as a hyperplane by minimizing the slightest possible error (Scholkopf & Smola, 2001).

Considering the strengths and weaknesses of linear and nonlinear models, many studies combine both models to perform forecasting. Kumar and Thenmozhi (2014) used hybrid ARIMA-SVM, ARIMA-ANN, and ARIMA models-random forest to forecast the S&P CNX Nifty index returns. It shows that combined models have better capabilities than single models, and the hybrid ARIMA-SVM model is recommended for stock price forecasting. Pai and Lin (2005) forecast ten stock prices using the hybrid ARIMA-SVR model and state that the hybrid ARIMA-SVR model performs better than the single model. Nie et al. (2012) use the hybrid ARIMA-SVR model to forecast stock prices of power generation companies in Heilongjiang, China, and conclude that the hybrid ARIMA-SVR model has better performance than SVR and ARIMA based on MAPE and RMSE values. This research aims to forecast stock prices using the ARIMA, SVR, and hybrid ARIMA-SVR models and compare the accuracy values of each model and its implementation in R-shiny to make it easier for investors to use the model to forecast stock prices.

METHOD

In this research, the data used was daily closing stock price data from April 1, 2019, to April 1, 2021, from three companies from different sectors: PT Unilever Indonesia Tbk (UNVR.JK) with a total of 497 observations, PT Perusahaan Gas Negara Tbk (PGAS.JK) with 496 observations, and PT Telekomunikasi Indonesia Tbk (TLKM.JK) with 496 observations obtained from the website https://finance.yahoo.com/. The method used in forecasting stock prices was a forecasting method using the ARIMA, SVR, and hybrid ARIMA-SVR models.

ARIMA Modeling

Autoregressive Integrated Moving Average (ARIMA) is a time series analysis method developed by George Box and Gwilym Jenkins in 1970 and is often referred to as the Box Jenkins method (Box et al., 2015). The ARIMA model is a non-seasonal model that combines differences between consecutive observations (process differencing) on the Autoregressive (AR) and Moving Average (MA) models. ARIMA model equations using Backward shift operators can be stated as follows (G. E. P. Box et al., 2015):

$$\varphi(B)z_t = \phi^p(B)\nabla^d z_t = \theta_0 + \theta^q(B)a_t \tag{1}$$

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(1 - B)^d z_t = \theta_0 + \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)a_t \tag{2}$$

It can be written in another equation form as follows (D. Cryer & Chan, 2008):

$$\nabla^d z_t = \theta_0 + \phi_1 \nabla^d z_{t-1} + \cdots + \phi_p \nabla^d z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{3}$$

where $\nabla^d z_t$ is the process of differencing data as much as d in the period t, ϕ_p is the AR model parameter coefficient in the order p, θ_q is the MA model parameter coefficient in the order q, and a_t is the residual in the period t.

In ARIMA modeling, the data used must be stationary. To determine the stationarity of the data, a test can be carried out by Augmented Dickey-Fuller (ADF). Transformation is carried out if the data is not stationary in the variance, and differencing is carried out if the data is not stationary in the average. The ARIMA model order is obtained from the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) values of stationary data. Next, selecting the best model is based on several possible models with significant order values obtained based on ACF and PACF, followed by a significance test of the model parameters. Then, the white noise residual assumption test is carried out using the Ljung-Box test and homogeneity of residual variance using the test Lagrange multiplier for Autoregressive Conditional Heteroscedasticity ARCH effect (Engle, 1982; Hillmer & Wei, 1991; Tsay, 2010). If the residual variance is not homogeneous, then Generalized Autoregressive Conditional Heteroscedasticity (GARCH) modeling needs to be done on ARIMA residuals with the order r and s which is denoted as $GARCH(r,s)$ (Hillmer & Wei, 1991). So, the model formed is $ARIMA(p,d,q)$-$GARCH(r,s)$ with the following equation:

$$\begin{aligned}&ARIMA(p,d,q):\\ &\quad \nabla^d z_t = \theta_0 + \phi_1 \nabla^d z_{t-1} + \cdots + \phi_p \nabla^d z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}\\ &GARCH(r,s):\\ &\qquad \sigma_t^2 = \omega + \alpha_1 \sigma_{t-1}^2 + \cdots + \alpha_r \sigma_{t-r}^2 + \beta_1 a_{t-1}^2 + \cdots + \beta_s a_{t-s}^2\end{aligned} \tag{4}$$

where r and s is the order of the GARCH model, σ_t^2 is the variance of the ARIMA model residuals in the period t, and a_t^2 is the squared residual of the ARIMA model in the period t.

Several selection criteria can be considered to choose the best model: Akaike's information criterion (AIC) and Bayesian information criterion (BIC). The best model selected from these two selection criteria is the model with the smallest AIC and BIC values. Then, stock data forecasting is carried out using the ARIMA model.

SVR Modeling

Support Vector Regression (SVR) is part of the Support Vector Machine (SVM) introduced by Vapnik (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995). SVR is machine learning that implements the idea of mapping the input vector into a high-dimensional feature space through a nonlinear mapping selected in apriori and a hyperplane optimal separator in that feature space.

From train data $(x_1, y_1), \ldots, (x_l, y_l)$ obtained randomly and from functions that are not known independently, SVR can be estimated using the following function form:

$$f(x) = \boldsymbol{w} \cdot \Phi(x) + b \tag{5}$$

with $\boldsymbol{w}$ is a weight vector, b denotes bias, and $\Phi(x)$ denotes a high-dimensional feature space mapped nonlinearly from the input space. The coefficients $\boldsymbol{w}$ and b are estimated by minimizing the risk function as follows (Vapnik, 1995) :

$$\min \frac{1}{2}\left||\boldsymbol{w}|\right|^2 + C\sum_{i=1}^{m} L(|y_i - f(x_i)|) \tag{6}$$

where,

$$L(|y_i - f(x_i)|) = \begin{cases} |y_i - f(x_i)| - \varepsilon, & |y_i - f(x_i)| \geq \varepsilon \\ 0, & \text{others} \end{cases} \quad \text{with obstacles} \quad \begin{matrix} y_i \leq f(x_i) + \varepsilon \\ y_i \geq f(x_i) - \varepsilon \end{matrix} \tag{7}$$

This value $\frac{1}{2}\left||\boldsymbol{w}|\right|^2$ states the flatness of the function, C, and ε states the parameters. C is constantly set to evaluate the trade-off between empirical error and model flatness, and $L(|y_i - f(x_i)|)$ states *ε-insensitive loss function* where the function will be zero if the predicted value and the actual value are smaller than ε.

It is necessary to add variables that are not feasible to overcome the possibility of unnecessary optimization (Ojemakinde, 2006). In this case, two variables of slack are needed for $f(x_i) - y_i > \varepsilon$ and $y_i - f(x_i) > \varepsilon$. The variable of slack is denoted as ξ_i and ξ_i^*, so that the risk function equation is transformed as follows (Scholkopf & Smola, 2001; Vapnik, 1995) :

$$\min \frac{1}{2}\left||\boldsymbol{w}|\right|^2 + C\sum_{i=1}^{m}(\xi_\mathrm{i} + \xi_i^*) \quad \text{with obstacles} \quad \begin{matrix} f(x_i) - y_i \leq \varepsilon + \xi_i \\ y_i - f(x_i) \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, m \end{matrix} \tag{8}$$

where i states the observation index, m states the amount support vector, C is a constant that states the weight of the risk function, $\boldsymbol{w}$ is a term that makes the function as flat as possible, and $\sum_{i=1}^{m}(\xi_\mathrm{i} + \xi_i^*)$ states ε-insensitive loss function.

To produce the best hyperlane, the equation can be expressed using the Lagrangian multipliers principle, which corresponds to quadratic programming optimization as follows (Awad & Khanna, 2015; Scholkopf & Smola, 2001; Vapnik, 1995):

$$\begin{aligned} \min L_p = \frac{1}{2}\left||\boldsymbol{w}|\right|^2 + C\sum_{i=1}^{m}(\xi_\mathrm{i} + \xi_i^*) + \sum_{i=1}^{m} a_i^*(y_i - \boldsymbol{w}\cdot x_i - \varepsilon - \xi_i^*) \\ + \sum_{i=1}^{m} a_i(\boldsymbol{w}\cdot x_i - y_i - \varepsilon - \xi_i) - \sum_{i=1}^{m}(\lambda_i \xi_\mathrm{i} + \lambda_i^* \xi_i^*) \end{aligned} \tag{9}$$

Based on the condition of *Karush-Kuhn-Tucker* (KKT), the result of decreasing the *lagrange* multiplier is obtained in the form of the parameter $\boldsymbol{w}$, which is stated as follows:

$$\boldsymbol{w} = \sum_{i=1}^{m}(a_i^* - a_i)x_i \tag{10}$$

By using the lagrangian multipliers principle and maximizing the quadratic function (L_p), which is a primal problem becomes a dual problem, the following equation is obtained:

$$\max -\varepsilon \sum_{i=1}^{m}(a_i^* + a_i) + \sum_{i=1}^{m} y_i(a_i^* - a_i) - \frac{1}{2}\sum_{i,j=1}^{m}(a_i^* - a_i)(a_j^* - a_j)(a_i^* - a_i)K(x_i, x_j) \quad (11)$$

with obstacle,

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} a_i = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} a_i^*$$
$$0 \le a_i \le C, \quad i = 1, \ldots., m$$
$$0 \le a_i^* \le C, \quad i = 1, \ldots., m$$

where a_i and a_i^* are lagrangian multipliers, so the equation for the SVR model is obtained as follows:

$$f(x) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}(a_i - a_i^*)\, K(x, x_i) + b \quad (12)$$

where $K(x, x_i)$ represents the kernel function, and b states the bias that can be estimated using KTT conditions, and the estimated value b is obtained $b = y_i - \boldsymbol{w} \cdot x_i - b - \varepsilon$ for $0 \le a_i \le C$ and $b = -y_i + \boldsymbol{w} \cdot x_i - \varepsilon$ for $0 \le a_i^* \le C$.

In this research, the kernel function is the radial basis function (RBF). Pain and Lin (2005) suggest using the RBF kernel function because it performs better for estimating the data nonlinear behavior, with Kernel functions which can be stated as follows (Pai & Lin, 2005) :

$$K(|x - x_i|) = \exp\left\{-\frac{|x - x_i|^2}{2\sigma^2}\right\} \quad (13)$$

So, the equation of the SVR model with the RBF kernel function is as follows:

$$f(x) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}(a_i - a_i^*)\exp\left\{-\frac{|x - x_i|^2}{2\sigma^2}\right\} + b \quad (14)$$

with $\sigma > 0$, and there are three parameters, namely σ (sigma), ε (epsilon), and C (cost). Each different parameter value produces different prediction results. To get the best results, it is necessary to estimate optimal parameters. Estimating optimal parameters can be done by tuning or optimizing parameters using optimization methods.

To determine the optimal parameters, tuning parameters is carried out using the grid search CV method on the data train. The cross validation (CV) method used is N-fold CV. Kaneko and Funatsu (2015) suggest using the N-fold CV from a leave-one-out CV to prevent over-fitting. After obtaining the optimal parameters, forecasting stock prices is carried out using the single SVR model and the ARIMA residual forecasting using the SVR model as a nonlinear component of the hybrid ARIMA-SVR model.

Hybrid ARIMA-SVR Modeling

The hybrid ARIMA-SVR model combines the ARIMA model and the SVR model. The hybrid model has linear and nonlinear components which can be used as an alternative in forecasting. ARIMA and SVR have different capabilities in analyzing linear or nonlinear data characteristics, so this hybrid model aims to compile ARIMA and SVR components, which are systematically written as follows:

$$Y_t = Z_t + N_t \tag{15}$$

where Z_t is the linear component and N_t is the nonlinear component of the hybrid model. Linear components are modeled using the ARIMA method. The estimation results of the ARIMA model are referred to as linear components so that residual equations can be formed from the linear components of the ARIMA model.

$$\boldsymbol{\varepsilon_t = Z_t - \widehat{Z}_t} \tag{16}$$

where ε_t is the residual ARIMA model in the period *t*, Z_t states observations in the period *t*, and $\hat{Z}_t$ is the estimation result of the ARIMA model, which is a *linear* component. Then, a nonlinear model is formed using the SVR model with input data in residuals from the ARIMA model, so it can be written as follows.

$$\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-n}) + \Delta_t \tag{17}$$

where *f* is a nonlinear component that is modelled using SVR and $\boldsymbol{\Delta_t}$ is the random error SVR model at the period *t.* So the combination obtained is as follows:

$$\boldsymbol{\widehat{Y}_t = \widehat{Z}_t + \widehat{N}_t} \tag{18}$$

with $\hat{Y}_t$ a hybrid ARIMA-SVR forecasting model, $\hat{Z}_t$ linear component of the ARIMA model forecast results and $\widehat{N}_t$ is a nonlinear component of the SVR forecasting results.

Model Accuracy

A good model is a model that has a small error value; the smaller the error value, the closer the forecasting results are to the actual value. Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAP) can be used to measure the model accuracy level.

The RMSE value is obtained by squaring the error value divided by the number of observations and then rooted. RMSE can be obtained using the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Z_t - \hat{Z}_t\right)^2}{n}} \tag{19}$$

where n $(t = 1,2, \dots, n)$ is the number of observations, Z_t is the observed value in the t-th period, and the $\hat{Z}_t$ estimated value in the t-th period.

Meanwhile, the MAPE value can be obtained by adding up the absolute value of the residual divided by the actual value, then dividing by the number of observations and multiplying by 100%. The MAPE equation can be stated as follows:

$$MAPE = \frac{1}{n}\sum\left|\frac{Z_t - \hat{Z}_t}{Z_t}\right| x100\% \qquad (20)$$

where n $(t = 1,2, \dots, n)$ is the number of observations, Z_t is the observed value in the t-th period, and $\hat{Z}_t$ is the estimated value in the t-th period.

RESULT AND DISCUSSION

The graph of daily closing stock price data from April 1, 2019, to April 1, 2021, from PT Unilever Indonesia Tbk (UNVR.JK), PT Perusahaan Gas Negara Tbk (PGAS.JK), and PT Telekomunikasi Indonesia Tbk (TLKM.JK) is presented in Figure 1.

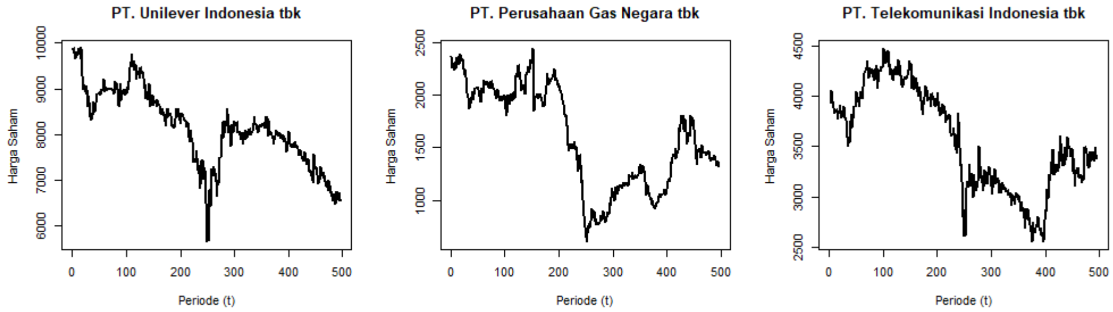


Figure 1. Plot three stock data

ARIMA Modeling

Before performing ARIMA modeling, the data must be stationary in variance and mean. Based on Figure 1, the three-stock data used have non-stationary data in variance and average. If the data is not stationary in variance, it is necessary to carry out Box-Cox transformation. If the data is not stationary in the mean, then it is necessary to conduct differencing. Augmented Dickey-Fuller (ADF) test can be used to ensure data stationarity. The ADF test results for three stock data can be seen in Table 1 as follows:

Table 1. ADF test before and after transformation and differencing

Data	Before		After	
	$p - value$	**Information**	$p - value$	**Information**
UNVR.JK	0,3121	Non-stationary	0,01	Stationary
PGAS.JK	0,6467	Non-stationary	0,01	Stationary
TLKM.JK	0,7302	Non-stationary	0,01	Stationary

Table 1 shows that the ADF test results on data before transformation and differencing produce a p-value less than the 5% significance level, so the data is declared non-stationary. Carrying out transformations and differencing once produces an ADF test with a p-value more than the 5% significance level, so the data is declared stationary in variance and average.

After carrying out the transformation and differencing, the model order of $ARIMA(p,d,q)$ is determined by looking at the ACF and PACF plots. The ACF and PACF plots are obtained for three stock data transformed and differencing once (order = 1) in Figure 2.

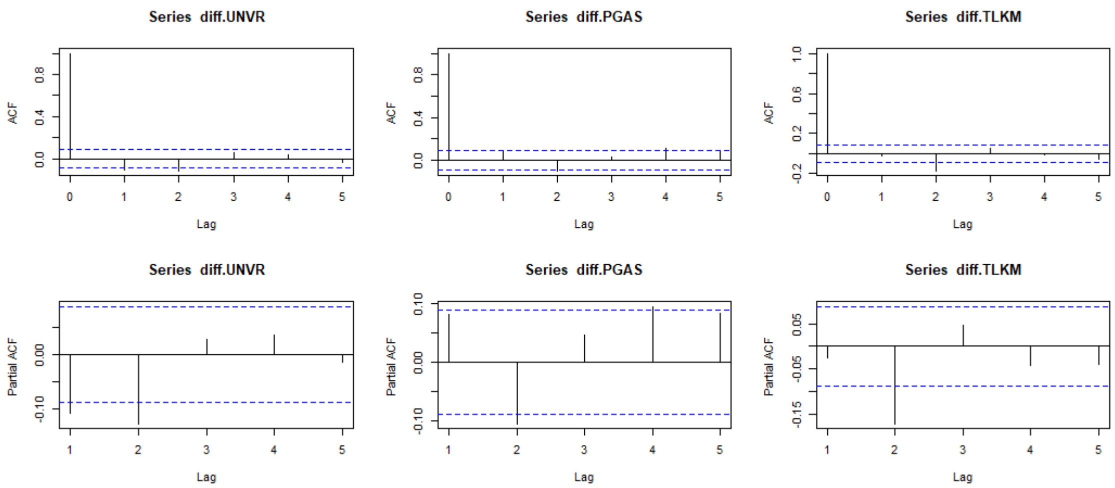


Figure 2. The ACF and PACF plots of stock data in one-time differencing

Based on Figure 2, the ACF plot of UNVR data has a significant lag at lags 0, 1, and 2. And the PACF plot has a significant lag at lags 1 and 2. So the combination of the model of $ARIMA(p,d,q)$ for UNVR data that is formed is $ARIMA(0,1,1)$, $ARIMA(0,1,2)$, $ARIMA(1,1,1)$, $ARIMA(1,1,2)$, $ARIMA(2,1,1)$, and $ARIMA(2,1,2)$. For PGAS data, the ACF plot has a significant lag at 0, 1, and 4. And the PACF plot has a significant lag at lags 2 and 4. So the model of $ARIMA(p,d,q)$ for PGAS data that is formed is ARIMA(0,1,2), ARIMA(0,1,4), ARIMA(2,1,2), ARIMA(2,1,4), ARIMA(4,1,2), and ARIMA(4,1,4). For TLKM data, the ACF plot has a significant lag at lags 0 and 2. The PACF plot has a significant lag at lag 2. So, the model of $ARIMA(p,d,q)$ for TLKM data that is formed is $ARIMA(0,1,2)$ and $ARIMA(2,1,2)$.

Parameter significance tests are then carried out from the combination of ARIMA models formed. The results of the parameter significance test for UNVR data are suitable for proceeding to the following process: $ARIMA(0,1,1)$, $ARIMA(0,1,2)$, $ARIMA(1,1,2)$, $ARIMA(2,1,1)$, and $ARIMA(2,1,2)$. For PGAS data, a combination of the $ARIMA(p,d,q)$ model that is feasible to proceed to the following process is the models of $ARIMA(0,1,2)$, $ARIMA(0,1,4)$, $ARIMA(2,1,2)$, and $ARIMA(4,1,4)$. For TLKM data, a combination of the $ARIMA(p,d,q)$ model that is feasible to proceed with the following process is the $ARIMA(0,1,2)$ model.

The ARIMA model formed must meet the assumptions of white noise. Testing this assumption can be done using the Ljung-Box test. The Ljung-Box test results are presented in Table 2.

Table 2. The Ljung-Box test result for three stock data

Data	Model	$\boldsymbol{p-value}$	Information
UNVR	$ARIMA(0,1,1)$	0,005	Assumptions not met
	$ARIMA(0,1,2)$	0,028	Assumptions not met
	$ARIMA(1,1,2)$	0,046	Assumptions not met
	$ARIMA(2,1,1)$	0,098	Assumptions met
	$ARIMA(2,1,2)$	0,176	Assumptions met
PGAS	$ARIMA(0,1,2)$	0,059	Assumptions met
	$ARIMA(0,1,4)$	0,112	Assumptions met
	$ARIMA(2,1,2)$	0,220	Assumptions met
	$ARIMA(4,1,4)$	0,155	Assumptions met
TLKM	$ARIMA(0,1,2)$	0,280	Assumptions met

In Table 2, the $ARIMA(p,d,q)$ models of UNVR.JK data that meet the assumptions of white noise are the models of $ARIMA(2,1,1)$ and $ARIMA(2,1,2)$, with the p-value less than the 5% significance level. The ARIMA models of PGAS data that meet the assumptions of white noise are the models of $ARIMA(0,1,2)$, $ARIMA(0,1,4)$, $ARIMA(2,1,2)$, and $ARIMA(4,1,4)$. The ARIMA model of TLKM data that meets the assumptions of white noise is the $ARIMA(0,1,2)$.

From several models formed, one model will be selected for each stock data by comparing the smallest values of Akaike's information criterion (AIC) and Bayesian information criterion (BIC) of each model. The results of calculating the AIC and BIC values are displayed in Table 3.

Table 3. The results of AIC and BIC value calculation

Data	$\boldsymbol{ARIMA(p,d,q)}$	AIC	BIC
UNVR	$ARIMA(2,1,1)$	6341,896	6358,722
	$\boldsymbol{ARIMA(2,1,2)}$	**6341,161**	**6362,193**
PGAS	$ARIMA(0,1,2)$	5279,393	5292,007
	$ARIMA(0,1,4)$	5279,027	5300,050
	$\boldsymbol{ARIMA(2,1,2)}$	**5276,265**	**5297,288**
	$ARIMA(4,1,4)$	5280,828	5318,669
TLKM	$\boldsymbol{ARIMA(0,1,2)}$	**5642,804**	**5655,417**

Table 3 shows that the model for each stock data with the smallest AIC value is the $ARIMA(2,1,2)$ model for UNVR data, the $ARIMA(2,1,2)$ model for PGAS data, and the $ARIMA(0,1,2)$ model for TLKM data.

Before forecasting, the model must have a homogeneous residual variance or not have an ARCH effect using the LM ARCH test. If the model has an ARCH effect, carrying out ARIMA-ARCH/GARCH modeling is necessary. The results of the LM ARCH test can be seen in Table 4.

Table 4. LM ARCH test results (assuming homogeneity of residual variance)

Data:$ARIMA(p,d,q)$	Lag	LM	$p-value$	Description
UNVR:$ARIMA(2,1,2)$	1	9,880	0,002	ARCH effect
	2	22,018	1,655x10^{-5}	ARCH effect
	3	117,515	2,646x10^{-25}	ARCH effect
	4	118,094	1,363x10^{-24}	ARCH effect
	5	120,166	2,894x10^{-24}	ARCH effect
PGAS:$ARIMA(2,1,2)$	1	34,178	5,030x10^{-9}	ARCH effect
	2	34,526	3,182x10^{-8}	ARCH effect
	3	34,846	1,313x10^{-7}	ARCH effect
	4	34,975	4,701x10^{-7}	ARCH effect
	5	34,915	1,565x10^{-6}	ARCH effect
TLKM:$ARIMA(0,1,2)$	1	14,442	1,446x10^{-4}	ARCH effect
	2	38,162	5,166x10^{-9}	ARCH effect
	3	43,595	1,839x10^{-9}	ARCH effect
	4	56,908	1,294x10^{-11}	ARCH effect
	5	56,956	5,163x10^{-11}	ARCH effect

Based on Table 4, using the decision criteria that p-value < 0.05, the residual has a non-homogeneous variance or an ARCH effect. So, it can be concluded that each model from the data has an ARCH effect, so carrying out ARIMA-ARCH/GARCH modeling is necessary.

Because the ARIMA models from three stock data have residual variances that are not homogeneous, it is necessary to carry out ARIMA-ARCH/GARCH modeling. A simulation is carried out to obtain the optimal GARCH model order by looking for the smallest AIC value of the $GARCH(r,s)$ model with a combination of the order *r* and *s* as many as 30, and the best model obtained for the UNVR.JK data is $ARIMA(2,1,2)$-$GARCH(1,3)$ with an AIC value of 12.436. The best model for PGAS.JK data is $ARIMA(2,1,2)$- $GARCH(1,2)$ with an AIC value of 10.438. The best model for TLKM.JK data is $ARIMA(0,1,2)$- $GARCH(2,4)$ with an AIC value of 11,255. And the best model for PGAS.JK data is - with an AIC value of 11.255. The results of stock price forecasting using the ARIMA-GARCH model for the next five periods are displayed in Table 5.

Table 5. The stock price forecasting of the ARIMA-GARCH model

UNVR.JK		PGAS.JK		TLKM.JK	
t	**Forecasting**	**t**	**Forecasting**	**t**	**Forecasting**
498	6540,734	497	1313,583	497	3421,922
499	6509,254	498	1309,049	498	3416,627
500	6478,613	499	1303,121	499	3416,627
501	6448,095	500	1297,692	500	3416,627
502	6417,438	501	1294,263	501	3416,627

SVR Modeling

In SVR modeling, the data used is stock closing price data with the input variable (x) in the form of stock closing price data based on a significant PACF value lag, and the output variable is actual stock closing price data. In this research, the number of lags used is 2 for each stock data. The data formed is divided into train and test data, with a 70% share and 30% as test data.

After sharing the data, the next step is conducting a simulation to obtain the optimal SVR model by tuning parameters using the grid search CV on train data. In this research, the parameters of the SVR model with the RBF kernel function are presented in Table 6 as follows:

Table 6. Range of SVR model parameter values

Parameter	Value Range
σ	1, 25, 50, 75, dan 100
C	1, 10, 33, 55, 78, dan 100
ε	0,1

From these three parameters, several combinations formed with the number N-forlds is 5. So, the optimal parameters for the SVR model for UNVR data obtained are *sigma* $(\sigma) = 1$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. For PGAS data, it is *sigma* $(\sigma) = 1$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. For TLKM data, it is *sigma* $(\sigma) = 1$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. The results of stock price forecasting for the next five periods are as follows:

Table 7. The stock price forecasting of the SVR model

UNVR.JK		PGAS.JK		TLKM.JK	
t	Forecasting	t	Forecasting	t	Forecasting
498	6656,303	497	1292,162	497	3363,227
499	6766,049	498	1269,384	498	3310,025
500	6907,675	499	1248,478	499	3250,648
501	7031,027	500	1229,674	500	3199,490
502	7091,308	501	1212,931	501	3166,985

Hybrid ARIMA-SVR Modeling

The hybrid ARIMA-SVR model is obtained by combining linear and nonlinear components. The linear component results from stock data forecasting using the ARIMA model, and the nonlinear component results from ARIMA residual forecasting using the SVR model. The ARIMA-SVR hybrid linear component model has been obtained in the discussion of ARIMA modeling, while the nonlinear component model can be obtained by performing SVR modeling on ARIMA residuals.

The residuals from each stock data will be formed into new data with a dependent variable in the form of a residual and an independent variable in the form of a residual lag based on the MA order of the ARIMA model. By tuning parameters using the grid search CV method on train data, the optimal parameters obtained for the residual SVR model for UNVR data are *sigma* $(\sigma) =$

100, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. Untuk data PGAS adalah *sigma* $(\sigma) = 1$, *cost* $(C) = 1$, dan *epsilon* $(\varepsilon) = 0{,}1$. Moreover, for TLKM data, it is *sigma* $(\sigma) = 100$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. The residual forecasting results for the next five periods are obtained as follows:

Table 8. Residual forecasting (nonlinear components)

UNVR.JK		PGAS.JK		TLKM.JK	
t	**Forecasting**	**t**	**Forecasting**	**t**	**Forecasting**
498	-61,622	497	-6,182	497	-9,824
499	84,080	498	7,357	498	17,135
500	-27,674	499	3,572	499	19,476
501	43,250	500	-3,697	500	27,905
502	-52,603	501	-1,028	501	8,410

The combination of the hybrid ARIMA-SVR model for UNVR stocks that is formed is the model $ARIMA(2{,}1{,}2)$ model with the $GARCH(1{,}3)$ effect and the SVR model with parameters of *sigma* $(\sigma) = 25$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. The combination of the hybrid ARIMA-SVR model for PGAS stocks is the $ARIMA(2{,}1{,}2)$ model with the $GARCH(1{,}2)$ effect and the SVR model with parameters of *sigma* $(\sigma) = 1$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. The combination of the hybrid ARIMA-SVR model for TLKM stocks is the $ARIMA(0{,}1{,}2)$ model with the $GARCH(2{,}4)$ effect and the SVR model with parameters of *sigma* $(\sigma) = 100$, *cost* $(C) = 1$, and *epsilon* $(\varepsilon) = 0{,}1$. The results of stock price forecasting using the hybrid ARIMA-SVR model for the next five periods are as follows:

Table 9. The stock price forecasting of the hybrid ARIMA-SVR model

UNVR		PGAS		TLKM	
t	**Forecasting**	**t**	**Forecasting**	**t**	**Forecasting**
498	6479,112	497	1307,401	497	3412,099
499	6593,335	498	1316,407	498	3433,762
500	6450,939	499	1306,693	499	3436,104
501	6491,345	500	1293,995	500	3444,532
502	6364,835	501	1293,235	501	3425,037

RMSE and MAPE Values

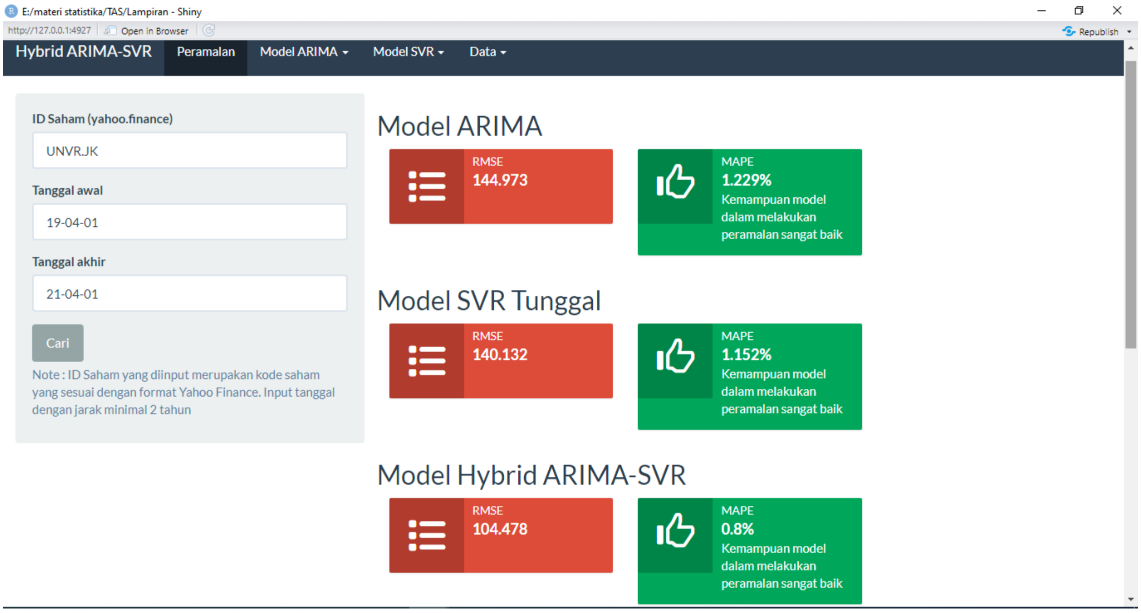
Based on the research results, RMSE and MAPE values are obtained for the ARIMA-GARCH, SVR, and hybrid ARIMA-SVR models in Table 10.

Table 10. RMSE and MAPE values of ARIMA-GARCH, SVR, and hybrid ARIMA-SVR models

Data	**ARIMA-GARCH**		**SVR**		**Hybrid ARIMA-SVR**	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
UNVR	144,973	1,229%	142,600	1,285%	104,478	0,797%
PGAS	49,508	2,308%	49,756	2,519%	47,693	2,213%
TLKM	71,889	1,522%	94,569	1,915%	56,634	0,993%

Table 10 presents the MAPE values obtained below 10% for each forecasting model formed, which states that the model has very good capabilities. By comparing the MAPE values obtained, the hybrid ARIMA-SVR model has the smallest MAPE value, so this hybrid model is more effective with smaller error values than the single model.

In this research, the model obtained is implemented as an R-shiny web. A simulation is carried out using UNVR.JK stock data from April 1, 2019, to April 1, 2021, to test whether the results displayed on the R-shiny website follow the previous discussion. The simulation results are obtained in Figure 3.



ID Saham (yahoo.finance)

UNVR.JK

Tanggal awal

19-04-01

Tanggal akhir

21-04-01

Note : ID Saham yang diinput merupakan kode saham yang sesuai dengan format Yahoo Finance. Input tanggal dengan jarak minimal 2 tahun

Model ARIMA

RMSE 144.973

MAPE 1.229% Kemampuan model dalam melakukan peramalan sangat baik

Model SVR Tunggal

RMSE 140.132

MAPE 1.152% Kemampuan model dalam melakukan peramalan sangat baik

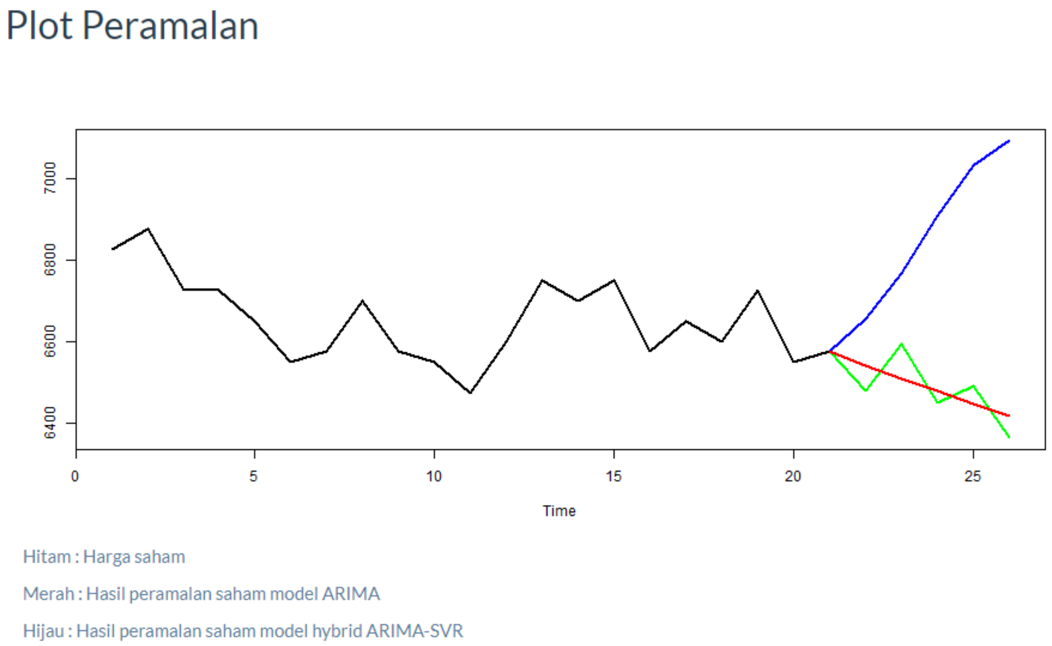
Model Hybrid ARIMA-SVR

RMSE 104.478

MAPE 0.8% Kemampuan model dalam melakukan peramalan sangat baik

Figure 3. UNVR data simulation results on R-shiny

Figure 3 shows that the hybrid ARIMA-SVR model has smaller RMSE and MAPE values than the single ARIMA and SVR models, so the hybrid ARIMA-SVR model has better abilities than the single ARIMA and SVR models in predicting UNVR stock prices.



Plot Peramalan

Hitam : Harga saham

Merah : Hasil peramalan saham model ARIMA

Hijau : Hasil peramalan saham model hybrid ARIMA-SVR

Figure 4. UNVR stock price forecasting plot

In Figure 4, the black line represents the actual stock price, the red line represents the stock price forecasting results of the ARIMA model, the blue line represents the stock price forecasting results of the single SVR model, and the green line represents the stock price forecasting results of the hybrid ARIMA-SVR model. Moreover, the UNVR data simulation output on the R-shiny web obtains the same results as the modeling in the previous discussion. The results of implementing the hybrid ARIMA-SVR model on the R shiny web can be accessed via the website https://project-tas.shinyapps.io/Hybrid_ARIMA_SVR/.

Based on the results of the discussion, the hybrid ARIMA-SVR model has smaller RMSE and MAPE values than the ARIMA-GARCH and SVR models, so the hybrid ARIMA-SVR model has better capabilities than the ARIMA and SVR models in forecasting stock prices. These results follow Pai and Lin (2005) and Hongzhan Nie et al. (2012), which state that the hybrid ARIMA-SVR model performs better than the ARIMA and SVR models.

In this research, the R-shiny web formed can make it easier for investors to forecast stock prices. The application of the R-shiny web has been carried out in various fields, such as health and education, where this application makes it easier for users to carry out analysis without having to re-form commands in the analysis software (Gibranda et al., 2017; Owen et al., 2019; Potter et al., 2016; Wojciechowski et al., 2015). By entering the stock code, this website can display graphs or forecast plots for the next five periods using the ARIMA, SVR, and hybrid ARIMA-SVR models and display the level of accuracy for each model so that it can be a consideration in investing without needing to understand the forecasting model material.

Combining the ARIMA and SVR models produces a more powerful and adaptive model for forecasting stock prices. The stock market often experiences nonlinear behavior, which causes fluctuations in stock prices (Chai et al., 2022; Teplova & Gurov, 2022). The ARIMA model has weaknesses in predicting data that contains nonlinear patterns, while the SVR model can overcome these weaknesses (Shrivastav & Kumar, 2019; Sinay et al., 2022). Moreover, in this research, the hybrid ARIMA-SVR model produces more flexible forecasting of stock price fluctuations than the ARIMA and SVR models with small error values and can adapt to changes in stock market behavior. So, this hybrid model can be an alternative model for forecasting stock prices.

CONCLUSION

Stock price forecasting using the hybrid ARIMA-SVR model on daily closing stock price data from April 1, 2019, to April 1, 2021, from PT Unilever Indonesia Tbk (UNVR.JK), PT Perusahaan Gas Negara Tbk (PGAS.JK), and PT Telekomunikasi Indonesia Tbk (TLKM.JK), presents results with the smallest MAPE values compared to the ARIMA and SVR models. So, it can be concluded that the hybrid ARIMA-SVR model has the ability and a better level of accuracy to reduce the error rate in forecasting stock prices. Moreover, implementing the model on the R-shiny website can make it easier for users to forecast the desired stock prices by entering the stock code and time range without re-forming commands in the analysis software.

REFERENCES

Amin, M. A. Al, & Hoque, M. A. (2019). Comparison of ARIMA and SVM for Short-Term Load Forecasting. *IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference*, 205–210. https://doi.org/10.1109/IEMECONX.2019.8877077

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engimeers and System Designers.* Apress Media.

Banerjee, D. (2014). Forecasting of Indian stock market using time-series ARIMA model. *2014 2nd International Conference on Business and Information Management, ICBIM 2014*, 131–135. https://doi.org/10.1109/ICBIM.2014.6970973

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control, 5th Edition*. John Wiley & Sons In.

Chai, S., Chu, W., Zhang, Z., Li, Z., & Abedin, M. Z. (2022). Dynamic Nonlinear Connectedness Between The Green Bonds, Clean Energy, and Stock Price: The Impact of The COVID-19 Pandemic. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04452-y

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

D. Cryer, J., & Chan, K.-S. (2008). *Time Series Analysis With Applications in R Second Edition* (2nd ed.). Springer. https://doi.org/https://doi.org/10.1007/978-0-387-75959-3

Dhini, A., Surjandari, I., Riefqi, M., & Puspasari, M. A. (2015). Forecasting Analysis of Consumer Goods Demand Using Neural Networks and ARIMA. *International Journal of Technology*, *6*(5), 872–880. https://doi.org/10.14716/ijtech.v6i5.1882

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, *50*(4), 987-1007. https://doi.org/10.2307/1912773

Erkilet, G., Janke, G., & Kasperzak, R. (2022). How Valuation Approach Choice Affects Financial Analysts' Target Price Accuracy. *Journal of Business Economics, 9*(5). https://doi.org/10.1007/s11573-021-01061-w

Gibranda, Ramdani, F., & Aknuranda, I. (2017). Pengembangan WebGIS untuk Analisis dan Pemodelan Data Menggunakan Teknik Regresi Spasial dan R-Shiny Web Framework (Studi Kasus: Data Kemiskinan dan Zakat Jawa Timur). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2*(3), 1290–1298. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1132/426

Hajiabadi, M. E., & Samadi, M. (2019). Locational Marginal Price Share: A New Structural Market Power Index. *Journal of Modern Power Systems and Clean Energy*, *7*(6), 1709–1720. https://doi.org/10.1007/s40565-019-0532-7

Hillmer, S. C., & Wei, W. W. S. (1991). Time Series Analysis: Univariate and Multivariate Methods. *Journal of the American Statistical Association*, *86*(413), 245-246. https://doi.org/10.2307/2289741

Ho, S. L., Xie, M., & Goh, T. N. (2002). A Comparative Study of Neural Network and Box-Jenkins ARIMA Modeling in Time Series Prediction. *Computers and Industrial Engineering*, *42*(2–4), 371–375. https://doi.org/10.1016/S0360-8352(02)00036-0

Juberias, G., Yunta, R., Garcia Moreno, J., & Mendivil, C. (1999). New Arima Model for Hourly Load Forecasting. *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*, *1*, 314–319. https://doi.org/10.1109/tdc.1999.755371

Kaneko, H., & Funatsu, K. (2015). Fast Optimization of Hyperparameters for Support Vector Regression Models with Highly Predictive Ability. *Chemometrics and Intelligent Laboratory Systems*, *142*, 64–69. https://doi.org/10.1016/J.CHEMOLAB.2015.01.001

Kumar, M., & Thenmozhi, M. (2014). Forecasting Stock Index Returns Using ARIMA-SVM, ARIMA-ANN, and ARIMA-Random Forest Hybrid Models. *International Journal of Banking, Accounting and Finance*, *5*(3), 284–308. https://doi.org/10.1504/IJBAAF.2014.064307

Lampart, M., Lampartová, A., & Orlando, G. (2023). On Risk and Market Sentiments Driving Financial Share Price Dynamics. *Nonlinear Dynamics*, *111*(17), 16585–16604. https://doi.org/10.1007/s11071-023-08702-5

Lee, C. M., & Ko, C. N. (2011). Short-Term Load Forecasting Using Lifting Scheme and ARIMA Models. *Expert Systems with Applications*, *38*(5), 5902–5911. https://doi.org/10.1016/j.eswa.2010.11.033

Li, W., & Zhang, Z. G. (2009). Based on Time Sequence of ARIMA Model in The Application of Short-Term Electricity Load Forecasting. *ICRCCS 2009 - 2009 International Conference on Research Challenges in Computer Science*, 11–14. https://doi.org/10.1109/ICRCCS.2009.12

Lu, J.-C., Niu, D.-X., & Jia, Z.-Y. (2004). A Study of Short-Term Load Forecasting Based on ARIMA-ANN. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*, *5*, 3183–3187.

Muslih. (2018). Pengaruh Likuiditas Saham dan Risiko Sistematis terhadap Tingkat Pengembalian Saham Perusahaan Plastik dan Kemasan yang Go Public di Bursa Efek Indonesia. *The National Conferences Management and Business (NCMAB) 2018*, 36–45.

Nie, H., Liu, G., Liu, X., & Wang, Y. (2012). Hybrid of ARIMA and Svms for Short-Term Load Forecasting. *Energy Procedia*, *16*, 1455–1460. https://doi.org/10.1016/j.egypro.2012.01.229

Nukala, V. B., & Prasada Rao, S. S. (2021). Role of Debt-to-Equity Ratio in Project Investment Valuation, Assessing Risk and Return in Capital Markets. *Future Business Journal*, *7*(1), 1–23. https://doi.org/10.1186/s43093-021-00058-9

Ojemakinde, B. T. (2006). *Support Vector Regression for Non-Stationary Time Series*. University of Tennessee. http://trace.tennessee.edu/utk_gradthes/1756

Owen, R. K., Bradbury, N., Xin, Y., Cooper, N., Sutton, A., & Rhiannon Owen, C. K. (2019). Metainsight: An Interactive Web-Based Tool for Analyzing, Interrogating, and Visualizing Network Meta-Analyses Using R-Shiny and Netmeta. *Research Synthesis Methods, 10*(4), 569-581. https://doi.org/10.1002/jrsm.1373

Pai, P. F., & Lin, C. S. (2005). A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting. *Omega*, *33*(6), 497–505. https://doi.org/10.1016/j.omega.2004.07.024

Doi, J., Potter, G., Wong, J., Alcaraz, I., Chi, P (2016). Web Application Teaching Tools for Statistics Using R and Shiny. *Technology Innovations in Statistics Education*, *9*(1). https://doi.org/10.5070/T591027492

Purnama, D. I., & Setianingsih, S. (2021). Peramalan Harga Emas Saat Pandemi Covid-19 Menggunakan Model Hybrid Autoregressive Integrated Moving Average - Support Vector Regression. *Jambura Journal of Mathematics*, *3*(1), 52–65. https://doi.org/10.34312/jjom.v3i1.8430

Scholkopf, B., & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Septiningrum, L., Yasin, H., & Sugito. (2015). Prediksi Indeks Harga Saham Gabungan Menggunakan Support Vector Regression (SVR) dengan Algoritma Grid Search. *Jurnal Gaussian*, *4*(2), 315–321.

Shrivastav, L. K., & Kumar, R. (2019). An Empirical Analysis of Stock Market Price Prediction using ARIMA and SVM. *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, 173–178.

Sinay, L. J., Jaariyah, M., Lewaherilla, N., & Lesnussa, Y. A. (2022). Forecasting The Composite Stock Price Index Using Autoregressive Integrated Moving Average Hybrid Model Artificial Neural Network. *Pattimura International Journal of Mathematics (PIJMath)*, *1*(2), 89–100. https://doi.org/10.30598/pijmathvol1iss2pp89-100

Tao, T., Shi, P., Wang, H., Yuan, L., & Wang, S. (2021). Performance Evaluation of Linear and Nonlinear Models for Short-Term Forecasting of Tropical-Storm Winds. *Applied Sciences (Switzerland)*, *11*(20). https://doi.org/10.3390/app11209441

Taskaya-Temizel, T., & Ahmad, K. (2005). Are ARIMA Neural Network Hybrids Better Than Single Models?. *Proceedings of the International Joint Conference on Neural Networks*, *5*, 3192–3197. https://doi.org/10.1109/IJCNN.2005.1556438

Teplova, T., & Gurov, S. (2022). Nonlinear Intraday Trading Invariance in The Russian Stock Market. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04683-7

Tsay, R. S. (2010). *Analysis of Financial Time Series*. A John Willey & Sons Inc. https://doi.org/10.1002/9780470644560

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Wojciechowski, J., Hopkins, A. M., & Upton, R. N. (2015). Interactive Pharmacometric Applications Using R and The Shiny Package. *CPT: Pharmacometrics and Systems Pharmacology*, *4*(3), 146–159. https://doi.org/10.1002/psp4.21