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

Hanan Albarr¹)
Rosita Kusumawati²)
¹,²)Statistika, FMIPA, Universitas Negeri Yogyakarta
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 (Juberías 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
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):

\[
\phi(B)z_t = \theta_0 + \theta(B)\alpha_t \\
(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d z_t = \theta_0 + (1 - \theta_1 B - \cdots - \theta_q B^q)\alpha_t
\]

(1)

(2)

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

\[
\nabla^d z_t = \theta_0 + \phi_1 \nabla^d z_{t-1} + \cdots + \phi_p \nabla^d z_{t-p} + \alpha_t - \theta_1 \alpha_{t-1} - \cdots - \theta_q \alpha_{t-q}
\]

(3)

where \(\nabla^d z_t\) is the process of differencing data as much as \(d\) in the period \(t\), \(\phi_p\) is the AR model parameter coefficient in the order \(p\), \(\theta_q\) is the MA model parameter coefficient in the order \(q\), and \(\alpha_t\) 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:

\[
ARIMA(p, d, q) : \\
\nabla^d z_t = \theta_0 + \phi_1 \nabla^d z_{t-1} + \cdots + \phi_p \nabla^d z_{t-p} + \alpha_t - \theta_1 \alpha_{t-1} - \cdots - \theta_q \alpha_{t-q}
\]

\[
GARCH(r, s) : \\
\sigma^2_t = \omega + \alpha_1 \sigma^2_{t-1} + \cdots + \alpha_r \sigma^2_{t-r} + \beta_1 \alpha^2_{t-1} + \cdots + \beta_s \alpha^2_{t-s}
\]

(4)

where \(r\) and \(s\) is the order of the GARCH model, \(\sigma^2_t\) is the variance of the ARIMA model residuals in the period \(t\), and \(\alpha^2_t\) 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:

\[
f(x) = w \cdot \Phi(x) + b
\]  

(5)

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

\[
\min \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} L(|y_i - f(x_i)|)
\]

(6)

where,

\[
L(|y_i - f(x_i)|) = \begin{cases} 
|y_i - f(x_i)| - \varepsilon, & |y_i - f(x_i)| \geq \varepsilon \\
0, & \text{others} 
\end{cases}
\]

with obstacles

\[
y_i \leq f(x_i) + \varepsilon \\
y_i \geq f(x_i) - \varepsilon
\]

(7)

\[
\frac{1}{2}||w||^2
\]

states the flatness of the function, \(C\), and \(\varepsilon\) 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 \(\varepsilon\)-insensitive loss function where the function will be zero if the predicted value and the actual value are smaller than \(\varepsilon\).

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 \(\xi_i\) and \(\xi_i^*\), so that the risk function equation is transformed as follows (Scholkopf & Smola, 2001; Vapnik, 1995):

\[
\min \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)
\]

(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, \(w\) is a term that makes the function as flat as possible, and \(\sum_{i=1}^{m} (\xi_i + \xi_i^*)\) states \(\varepsilon\)-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):
Based on the condition of Karush-Kuhn-Tucker (KKT), the result of decreasing the lagrange multiplier is obtained in the form of the parameter $\mathbf{w}$, which is stated as follows:

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

(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:

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

with obstacle,

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

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

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

(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 $b = y_i - \mathbf{w} \cdot x_i - b - \varepsilon$ for $0 \leq a_i \leq C$ and $b = -y_i + \mathbf{w} \cdot x_i - \varepsilon$ for $0 \leq a_i^* \leq C$.

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):

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

(13)

So, the equation of the SVR model with the RBF kernel function is as follows:
with $\sigma > 0$, and there are three parameters, namely $\sigma$ (sigma), $\varepsilon$ (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.

\[
\varepsilon_t = Z_t - \hat{Z}_t \tag{16}
\]

where $\varepsilon_t$ 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.

\[
\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-n}) + \Delta_t \tag{17}
\]

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

\[
\hat{Y}_t = \hat{Z}_t + \hat{N}_t \tag{18}
\]
with $\hat{Y}_t$, a hybrid ARIMA-SVR forecasting model, $\hat{Z}_t$, linear component of the ARIMA model forecast results and $\tilde{N}_t$ 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 (MAPE) 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:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (Z_t - \hat{Z}_t)^2}$$  \hspace{1cm} (19)

where $n \ (t = 1, 2, ..., n)$ 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:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{Z_t - \hat{Z}_t}{Z_t} \right| \times 100\%$$  \hspace{1cm} (20)

where $n \ (t = 1, 2, ..., n)$ 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

<table>
<thead>
<tr>
<th>Data</th>
<th>Before p-value</th>
<th>Information</th>
<th>After p-value</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNVR.JK</td>
<td>0.3121</td>
<td>Non-stationary</td>
<td>0.01</td>
<td>Stationary</td>
</tr>
<tr>
<td>PGAS.JK</td>
<td>0.6467</td>
<td>Non-stationary</td>
<td>0.01</td>
<td>Stationary</td>
</tr>
<tr>
<td>TLKM.JK</td>
<td>0.7302</td>
<td>Non-stationary</td>
<td>0.01</td>
<td>Stationary</td>
</tr>
</tbody>
</table>

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](image)

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>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>$p$ value</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNVR</td>
<td>$ARIMA(0,1,1)$</td>
<td>0.005</td>
<td>Assumptions not met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(0,1,2)$</td>
<td>0.028</td>
<td>Assumptions not met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(1,1,2)$</td>
<td>0.046</td>
<td>Assumptions not met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(2,1,1)$</td>
<td>0.098</td>
<td>Assumptions met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(2,1,2)$</td>
<td>0.176</td>
<td>Assumptions met</td>
</tr>
<tr>
<td>PGAS</td>
<td>$ARIMA(0,1,2)$</td>
<td>0.059</td>
<td>Assumptions met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(0,1,4)$</td>
<td>0.112</td>
<td>Assumptions met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(2,1,2)$</td>
<td>0.220</td>
<td>Assumptions met</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(4,1,4)$</td>
<td>0.155</td>
<td>Assumptions met</td>
</tr>
<tr>
<td>TLKM</td>
<td>$ARIMA(0,1,2)$</td>
<td>0.280</td>
<td>Assumptions met</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Data</th>
<th>$ARIMA(p, d, q)$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNVR</td>
<td>$ARIMA(2,1,1)$</td>
<td>6341,896</td>
<td>6358,722</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(2,1,2)$</td>
<td>6341,161</td>
<td>6362,193</td>
</tr>
<tr>
<td>PGAS</td>
<td>$ARIMA(0,1,2)$</td>
<td>5279,393</td>
<td>5292,007</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(0,1,4)$</td>
<td>5279,027</td>
<td>5300,050</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(2,1,2)$</td>
<td>5276,265</td>
<td>5297,288</td>
</tr>
<tr>
<td></td>
<td>$ARIMA(4,1,4)$</td>
<td>5280,828</td>
<td>5318,669</td>
</tr>
<tr>
<td>TLKM</td>
<td>$ARIMA(0,1,2)$</td>
<td>5642,804</td>
<td>5655,417</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Data</th>
<th>$ARIMA(p,d,q)$</th>
<th>Lag</th>
<th>LM</th>
<th>$p$-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNVR:</td>
<td>$ARIMA(2,1,2)$</td>
<td>1</td>
<td>9,880</td>
<td>0,002</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>22,018</td>
<td>1,655x10^-5</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>117,515</td>
<td>2,646x10^-25</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>118,094</td>
<td>1,363x10^-24</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>120,166</td>
<td>2,894x10^-24</td>
<td>ARCH effect</td>
</tr>
<tr>
<td>PGAS:</td>
<td>$ARIMA(2,1,2)$</td>
<td>1</td>
<td>34,178</td>
<td>5,030x10^-9</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>34,526</td>
<td>3,182x10^-8</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>34,846</td>
<td>1,313x10^-7</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>34,975</td>
<td>4,701x10^-7</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>34,915</td>
<td>1,565x10^-6</td>
<td>ARCH effect</td>
</tr>
<tr>
<td>TLKM:</td>
<td>$ARIMA(0,1,2)$</td>
<td>1</td>
<td>14,442</td>
<td>1,446x10^-4</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>38,162</td>
<td>5,166x10^-9</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>43,595</td>
<td>1,839x10^-9</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>56,908</td>
<td>1,294x10^-11</td>
<td>ARCH effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>56,956</td>
<td>5,163x10^-11</td>
<td>ARCH effect</td>
</tr>
</tbody>
</table>

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.
Albarr & Kusumawati. Hybrid Autoregressive Integrated Moving Average-Support...

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

<table>
<thead>
<tr>
<th></th>
<th>UNVR.JK</th>
<th>PGAS.JK</th>
<th>TLKM.JK</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t )</td>
<td>Forecasting</td>
<td>( t )</td>
<td>Forecasting</td>
</tr>
<tr>
<td>498</td>
<td>6540,734</td>
<td>497</td>
<td>1313,583</td>
</tr>
<tr>
<td>499</td>
<td>6509,254</td>
<td>498</td>
<td>1309,049</td>
</tr>
<tr>
<td>500</td>
<td>6478,613</td>
<td>499</td>
<td>1303,121</td>
</tr>
<tr>
<td>501</td>
<td>6448,095</td>
<td>500</td>
<td>1297,692</td>
</tr>
<tr>
<td>502</td>
<td>6417,438</td>
<td>501</td>
<td>1294,263</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma )</td>
<td>1, 25, 50, 75, dan 100</td>
</tr>
<tr>
<td>( C )</td>
<td>1, 10, 33, 55, 78, dan 100</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>0,1</td>
</tr>
</tbody>
</table>

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

### Table 7. The stock price forecasting of the SVR model

<table>
<thead>
<tr>
<th></th>
<th>UNVR.JK</th>
<th>PGAS.JK</th>
<th>TLKM.JK</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t )</td>
<td>Forecasting</td>
<td>( t )</td>
<td>Forecasting</td>
</tr>
<tr>
<td>498</td>
<td>6656,303</td>
<td>497</td>
<td>1292,162</td>
</tr>
<tr>
<td>499</td>
<td>6766,049</td>
<td>498</td>
<td>1269,384</td>
</tr>
<tr>
<td>500</td>
<td>6907,675</td>
<td>499</td>
<td>1248,478</td>
</tr>
<tr>
<td>501</td>
<td>7031,027</td>
<td>500</td>
<td>1229,674</td>
</tr>
<tr>
<td>502</td>
<td>7091,308</td>
<td>501</td>
<td>1212,931</td>
</tr>
</tbody>
</table>

### 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 $\sigma = 100$, $C = 1$, and $\epsilon = 0.1$. For data PGAS, $\sigma = 1$, $C = 1$, and $\epsilon = 0.1$. Moreover, for TLKM data, $\sigma = 100$, $C = 1$, and $\epsilon = 0.1$. The residual forecasting results for the next five periods are obtained as follows:

<table>
<thead>
<tr>
<th>t</th>
<th>Forecasting</th>
<th>t</th>
<th>Forecasting</th>
<th>t</th>
<th>Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>498</td>
<td>-61,622</td>
<td>497</td>
<td>-6,182</td>
<td>497</td>
<td>-9,824</td>
</tr>
<tr>
<td>499</td>
<td>84,080</td>
<td>498</td>
<td>7,357</td>
<td>498</td>
<td>17,135</td>
</tr>
<tr>
<td>500</td>
<td>-27,674</td>
<td>499</td>
<td>3,572</td>
<td>499</td>
<td>19,476</td>
</tr>
<tr>
<td>501</td>
<td>43,250</td>
<td>500</td>
<td>-3,697</td>
<td>500</td>
<td>27,905</td>
</tr>
<tr>
<td>502</td>
<td>-52,603</td>
<td>501</td>
<td>-1,028</td>
<td>501</td>
<td>8,410</td>
</tr>
</tbody>
</table>

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 $\sigma = 25$, $C = 1$, and $\epsilon = 0.1$. 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 $\sigma = 1$, $C = 1$, and $\epsilon = 0.1$. 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 $\sigma = 100$, $C = 1$, and $\epsilon = 0.1$. The results of stock price forecasting using the hybrid ARIMA-SVR model for the next five periods are as follows:

<table>
<thead>
<tr>
<th>t</th>
<th>Forecasting</th>
<th>t</th>
<th>Forecasting</th>
<th>t</th>
<th>Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>498</td>
<td>6479,112</td>
<td>497</td>
<td>1307,401</td>
<td>497</td>
<td>3412,099</td>
</tr>
<tr>
<td>499</td>
<td>6593,335</td>
<td>498</td>
<td>1316,407</td>
<td>498</td>
<td>3433,762</td>
</tr>
<tr>
<td>500</td>
<td>6450,939</td>
<td>499</td>
<td>1306,693</td>
<td>499</td>
<td>3436,104</td>
</tr>
<tr>
<td>501</td>
<td>6491,345</td>
<td>500</td>
<td>1293,995</td>
<td>500</td>
<td>3444,532</td>
</tr>
<tr>
<td>502</td>
<td>6364,835</td>
<td>501</td>
<td>1293,235</td>
<td>501</td>
<td>3425,037</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Data</th>
<th>ARIMA-GARCH</th>
<th>SVR</th>
<th>Hybrid ARIMA-SVR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAPE</td>
<td>RMSE</td>
</tr>
<tr>
<td>UNVR</td>
<td>144,973</td>
<td>1,229%</td>
<td>142,600</td>
</tr>
<tr>
<td>PGAS</td>
<td>49,508</td>
<td>2,308%</td>
<td>49,756</td>
</tr>
<tr>
<td>TLKM</td>
<td>71,889</td>
<td>1,522%</td>
<td>94,569</td>
</tr>
</tbody>
</table>

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.

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.

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-](https://project-)
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


