Using Harris hawk optimization towards support vector regression to ozone prediction

Robert Kurniawan1 · I. Nyoman Setiawan2 · Rezzy Eko Caraka3,4 · Bahrul Ilmi Nasution5

Accepted: 13 September 2021 / Published online: 30 January 2022 © The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
As an area experiencing air pollution, especially ozone concentrations that often exceed the threshold or are unhealthy, JABODETABEK (Jakarta, Bogor, Depok, Tangerang, and Bekasi) seeks to prevent and control pollution as well as restore air quality. Therefore, this study aims to build a predictive model of ozone concentration using Harris hawks optimization-support vector regression (HHO-SVR) in 14 sub-districts in JABODETABEK. This goal is achieved by collecting data on ozone concentration as a response variable and meteorological factors as predictor variables from the website that provides the data. Other predictor variables such as time and significant lag detected with partial autocorrelation function of ozone concentration were also used. Then the variables will be selected using the recursive feature elimination-support vector regression (RFE-SVR) to obtain a significant predictor variable that affects the ozone concentration. After that, the prediction model will be built using the HHO-SVR method, support vector regression (SVR) whose parameter values are optimized with the Harris hawks optimization (HHO) algorithm. When the model has been formed, several evaluation metrics used to determine the best model include mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), Coefficient of Determination ($R^2$), Variance Ratio (VR), and Diebold–Mariano test. The results of this study indicate that lag 1, lag 2, air temperature, humidity, and UV index are significant predictor variables of the RFE-SVR results for most sub-districts. In general, the HHO process takes longer than other metaheuristic algorithms. On average, 7 of the 14 sub-districts using the HHO-SVR model yielded the best predictions with MAE below 10, RMSE and MAPE below 20, $R^2$ around 0.97, and VR around 0.98. Then, the results of the Diebold–Mariano test also show that the accuracy of the prediction results and the stability of the performance of the HHO-SVR model is better, especially for the Ciputat and South Bekasi sub-districts. This shows that the two sub-districts are very suitable to use HHO-SVR in predicting ozone concentrations.

Keywords Ozone · SVR · HHO · RFE · JABODETABEK

1 Introduction

Prior to the Republic of Indonesia’s Government Regulation No. 41 Year 1999, air pollution is the entry of substances, energy, and other components into ambient air by human activities, so that the quality of ambient air drops to a certain level which causes ambient air to be unable to fulfill its function. Ambient air is free air that is in the troposphere or the atmosphere closest to the earth’s surface. Currently, poor ambient air quality is a problem that is being faced by various countries in the world.

Good or bad ambient air quality is strongly influenced by human activities. According to the World Health Organization (WHO 2006), human activities that are the
main factors affecting ambient air quality are transportation, industry, agriculture, and energy generation and use. Most of these activities occur in urban areas and produce hazardous waste that can increase the concentration of air pollutants and thus affect ambient air quality (Permadi and Kim Oanh 2008). Several types of air pollutants affect ambient air quality, including particulate matter (PM), ozone (O₃), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂).

One of the most significant pollutants in the atmosphere is ozone (Zhao et al. 2015). Ozone is formed by photochemical reactions in the troposphere. Ozone is a secondary pollutant formed from the reaction between nitrogen oxides (NOx) and volatile organic compounds (VOCs) in the atmosphere with solar irradiation (Zhang et al. 2019). In addition, the decomposition process and ozone concentration are influenced by meteorological factors with very dynamic changes (Wasi’ah and Driejana 2017). Air temperature, solar radiation, and air pressure can increase ozone formation, while air humidity can reduce ozone concentrations (Souza et al. 2018). In addition, the ozone concentration in the future also tends to be correlated or influenced by the ozone concentration in the past.

High ground level ozone concentrations can affect health and the environment (World Bank Group 1998). Exposure to ozone pollutants can cause decreased performance of the human body due to disruption of the respiratory system. Acute diseases that can occur due to this, namely eye and nose irritation, respiratory diseases, and decreased lung function (Zhang et al. 2019). Moreover, the environmental aspects, especially agricultural crops and trees, will also experience growth disturbances. The visible responses of these plants are defoliation and changes in leaf color, which reduces plant productivity.

Air pollution control has actually become a program of the Indonesian government through pollution prevention and control also restoration of air quality. However, these activities must begin with continuous monitoring and research to determine developments in the ambient air condition (Masseran and Safari 2020). The development of the air condition can be seen by observing data on air pollutants, one of which is ozone, from time to time. Data of past, present, and future are types of data that can be used in observations to determine developments in the air condition. Past data can be known through the results of measurements that have been made before, but current and future data can only be known through prediction techniques.

Prediction technique is a technique to predict something that is happening now and in the future. The concentration of air pollutants such as ozone is one of the conditions that can be predicted through prediction techniques. Ozone concentrations can be predicted by constructing models that utilize suitable prediction methods. In this regard, SVR can be applied in this prediction problem because it implements the principle of minimizing structural risk by minimizing the upper limit of generalization errors (Vapnik 1995). SVR is a reliable machine learning with a strong theoretical basis and has excellent performance (Hu et al. 2013). However, SVR performance is greatly influenced by parameter value settings, as well as the type of kernel in it (Cao et al. 2020).

The application of a metaheuristic algorithm can be done to get the optimal SVR parameter value so as to improve generalization performance (Jiang et al. 2013). HHO is a metaheuristic algorithm that can be used to solve optimization problems. The HHO algorithm is inspired by the cooperative behavior and style of the Harris hawks in hunting prey—rabbits. (Heidari et al. 2019). The application of HHO in optimization problems has excellent performance and efficiency in terms of computation (Abbasi et al. 2019).

In addition, the selection of variables at the data preparation stage is an important process to remove noisy and irrelevant variables so that only relevant variables are selected (Elgamal et al. 2020). The application of RFE-SVR in the variable selection process will reduce the use of variables by removing insignificant variables in the ranking process using the SVR. The results of the RFE-SVR used in the HHO-SVR will improve the prediction results of the efficiency of the model building process.

There are several related studies that have existed before. Ismael et al. (2020, 2021) propose HHO to optimize v-SVR hyperparameters by embedding feature selection simultaneously or without separating the process. But in the evaluation of the model using only MSE. Elgamal et al. (2020) used HHO as a comparison algorithm against improvised HHO embedded feature selection simultaneously. HHO is not explained in detail. Cao et al. (2020) used the HHO-SVR hybrid model only as a comparison method from the proposed method and was not explained in detail and without being preceded by the selection of the predictor variables used. Bui et al. (2019) and Sammen et al. (2020) used HHO to optimize the performance of ANN without prior feature selection. Huang et al. (2014) used RFE with SVM ranking method for variable selection. However, this study focuses on the case of classification.

In this study, JABODETABEK was selected, an acronym for Jakarta-Bogor-Depok-Tangerang-Bekasi, an urban area known for its dense transportation and industrial activities. Caraka et al. (2020c) stated that in DKI Jakarta during the COVID-19 pandemic, the density of air
pollution still occurred. This is presumably due to the existence of these activities which are still running as before with large-scale restrictions. So that the amount of waste generated from these activities results in the ozone concentration often exceeding the standard limit (Wasi’ah and Driejana 2017). In addition, the high ozone concentration in Jakarta is also due to its location near the equator, abundant sunlight, and high temperatures which will increase ozone formation (Permadi and Kim Oanh 2008).

As a result, this area experiences problems related to air pollution and needs full attention from the government (Kusumaningtyas et al. 2018). The government must determine the right policy regarding air pollution based on the data. Especially the ozone concentration in the future can be predicted to find out the estimated value.

Thus, this study aims to combine SVR with HHO to predict ozone concentration in JABODETABEK, which is preceded by the process of selecting predictor variables. The selection of predictor variables used RFE-SVR method, followed by the development of the SVR model whose parameters were optimized using HHO (HHO-SVR). HHO-SVR will be compared with several other metaheuristic methods based on the evaluation and test of prediction results. Evaluation and testing of the prediction results are carried out to get the best model. With these results, it is hoped that the government as a policy maker can implement prediction models with better accuracy to observe the development of ozone concentrations. So that the right decisions can be taken according to the circumstances that will occur which of course can minimize the impact of air pollution on human health.

Unlike previous studies, this study uses real-time data from web scraping. An effective method of collecting data from websites that provide data. In addition, Setiawan et al. (2020) states that HHO-SVR is inefficient in terms of time so that the selection of variables is carried out to reduce the use of variables and improve the performance of HHO-SVR in later modeling. The selection of variables will be done first before the HHO-SVR modeling. The evaluation measures used in this study are also more diverse such as accuracy measures consisting of MAE, RMSE, MAPE, R2, and VR followed by testing the difference in the accuracy of the prediction results using the Diebold–Mariano test. This evaluation measure will help determine the best model that is more accurate to be applied and used as a reference in determining policy.

The HHO-SVR data and method are described in Sect. 2, the results and discussion are explained in Sect. 3. Meanwhile, the discussion regarding the results is explained in Sect. 4. Then it is continued by drawing conclusions from this study.

2 Data and research methods

2.1 Research scope and data

This research begins with the collection of research data. Prior to that, the research locus had been determined, namely JABODETABEK (Jakarta, Bogor, Depok, Tangerang, Bekasi). JABODETABEK is divided into 14 regencies/cities or 185 sub-districts. This area is the center of government and industry in the Special Capital Region of Jakarta. JABODETABEK was chosen because it has been around for a long time and is currently experiencing air pollution. In this regard, this research will focus on one of the sub-districts representing each regency/city. The selected sub-district is the administrative center of the regency/city (Fig. 1).

To build an air pollution prediction model, we set some data such as ozone concentration as the pollutant to be predicted and meteorological factors as predictor variables. The data will be used for 4 types of activities including variable selection, model development, model implementation for prediction, and model evaluation. In detail, the data used in this study are presented in Table 1.

The research variable data in Table 2 is determined based on related research. Specifically, the dew point and time are variables that the researcher deliberately added to obtain additional information in the modeling process. The data is obtained from the website https://weather.com/. This website provides data on air pollutants and meteorological factors from many regions around the world. The data provided changes in real time. To collect data, web scraping technique was applied. Scraper is built using the selenium and beautifulSoup libraries in the Python programming language. The flow of data collection can be seen in Fig. 2.

The scraper has been scheduled to run hourly. The data collection period is about 2 months from November 5, 2019 to December 31, 2019. During that time, 1368 records were collected for each sub-district. Descriptive statistics of the research data can be seen in Table 2.

Based on Table 2, ozone concentration data and other variables are separated based on training and testing data. From one of the variables, namely the ozone concentration in the training data, it can be seen that the ozone concentration at the research locus is in the range of values from 1 to 573 g/m³. The average was at 106.46 g/m³. Specifically, the standard deviation of 95.86 indicates that the concentration data varies greatly. Meanwhile, the testing data shows that the ozone concentration at the research locus is in the range of 1 to 386 g/m³. The ozone concentration range is shorter than the training data. The average was at 76.33 g/m³. In particular, the standard deviation is also
quite large, namely 74.65. Likewise with other variables which can be seen in detail through Table 2.

2.2 Research methods

2.2.1 Support vector regression

SVR is a method that can be applied in regression cases that apply the support vector machine concept (Drucker et al. 1997). This method performs well in regression and time series prediction (Müller et al. 1997) because it can handle the problem of overfitting (Smola and Schölkopf 2004). In the linear case, consider the function as follows.

\[ f(x) = \langle w, x \rangle + b \quad \text{with} \quad w \in X, \quad b \in R \]  

(1)

whereas \( w \) is the slope, \( x \) is the feature space, and \( b \) is the intercept. To get the function as flat as possible and minimizing the Euclidean value, thus the equation can be written as follows.

\[
\text{minimum} \quad \frac{1}{2} \|w\|^2 \\
\text{depend on} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\
\langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases}
\]

(2)

SVR aims to obtain a hyperplane by minimizing a margin of error (\( \varepsilon \)), in contrast to SVM which aims to obtain a hyperplane by maximizing margins and has a limit of 1 by following the equation \( y_i(w \cdot x_i - b) \geq 1 \). The difference can be seen in the Fig. 3 below.

From Fig. 3, the difference is clearly visible according to the purpose of the function. SVM which aims for classification, namely separating data by class will maximize the margin so that the differences between classes are more significant. The points are expected to be outside the hyperplane support. In contrast to SVM, SVR aims for regression, which is to obtain the value with the smallest error or minimum margin (2\( \varepsilon \)) so that the point is expected to be within the support hyperplane. When there is a point that is outside the hyperplane support as shown below (Fig. 4).
### Table 1 Research data

| No. | Variable                  | Description                                                                 | Unit            | Related research                                                                 |
|-----|---------------------------|-----------------------------------------------------------------------------|-----------------|----------------------------------------------------------------------------------|
| 1.  | Ozone concentration      | One of the secondary pollutants formed from chemical reactions induced by direct or indirect light (photochemical) which has a negative impact on the environment, nature and humans | μg/m³           | Masmoudi et al. (2020), Wasi’ah and Driejana (2017), Permadi and Kim Oanh (2008) |
| 2.  | Air temperature          | The degree of heat from a place on the earth’s surface that is affected by the sun’s heat | °C              | Masmoudi et al. (2020), Souza et al. (2018), Wasi’ah and Driejana (2017), Permadi and Kim Oanh (2008) |
| 3.  | Wind speed               | The rate of air flow caused by the difference in air pressure from one place to another | km/jam          | Masmoudi et al. (2020), Wasi’ah and Driejana (2017), Permadi and Kim Oanh (2008) |
| 4.  | Air humidity             | The amount of water vapor contained in the air                                | %              | Masmoudi et al. (2020), Souza et al. (2018), Wasi’ah and Driejana (2017), Permadi and Kim Oanh (2008) |
| 5.  | Dew point                | Conditions that cause the process of condensation of water vapor into water droplets because the air can no longer accommodate water vapor | °C              | –                                                                                |
| 6.  | Ultraviolet (UV) Index   | Index that shows the amount of ultraviolet light exposure to the earth’s surface | –              | Masmoudi et al. (2020), Souza et al. (2018), Wasi’ah and Driejana (2017), Permadi and Kim Oanh (2008) |
| 7.  | Air pressure             | The buildup of air masses on the earth’s surface due to the air above pressing the air below | Millibar        | Souza et al. (2018)                                                              |
| 8.  | Visibility               | The maximum distance that an observer can see from an observation point      | km              | Permadi and Kim Oanh (2008)                                                      |
| 9.  | Time                     | A series of times when the process of an event, change or state when an object takes place, a certain length of time to do something | –              | –                                                                                |

### Table 2 Descriptive statistics of research data

| Variable        | Training data | Testing Data |     |     |     |     |     |     |
|-----------------|---------------|--------------|-----|-----|-----|-----|-----|-----|
|                 | Max | Min | Mean | Sd  | Max | Min | Mean | Sd  |
| Ozone Concentration | 573 | 1   | 106.46 | 95.86 | 386 | 1   | 76.33 | 74.65 |
| Air Temperature  | 37  | 22  | 28.32 | 2.84 | 33  | 23  | 27.47 | 2.34 |
| Wind Speed      | 37  | 0   | 8.55  | 5.44 | 34  | 0   | 7.95  | 5.29 |
| Air Humidity    | 98  | 33  | 74.13 | 11.90 | 98  | 55  | 81.28 | 8.85 |
| Dew Point       | 26  | 17  | 22.93 | 1.03 | 26  | 21  | 23.69 | 0.74 |
| UV Index        | 11  | 0   | 2.89  | 4.18 | 11  | 0   | 2.80  | 4.12 |
| Air Pressure    | 1013.9 | 1005.1 | 1009.67 | 1.50 | 1013.6 | 1006.1 | 1009.86 | 1.33 |
| Visibility      | 16.1 | 0.2 | 8.66 | 3.13 | 16.1 | 0.8 | 7.59 | 3.28 |
| Time            | 23  | 0   | 11.44 | 6.92 | 23  | 0   | 11.76 | 6.94 |

Fig. 2 Web scraping flowchart
Then the addition of the slack variable $\xi_i$, $\xi^*_i$ needs to be done to overcome this problem by following the equation below.

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi^*_i)$$

Fig. 3 SVM (left) and SVR (right)

Fig. 4 SVR with slack variable

whereas the value of $C > 0$ is a penalty coefficient that controls the balance between the complexity of the model and the error rate in training (Liu et al. 2018). To make it easier to the optimization problem solving, the implementation of multiple formulations is applied with a
standard dualization method that utilizes Lagrange multiplication (Smola and Scholkopf 2004).

\[
L = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) - \sum_{i=1}^{N} (\eta_i \xi_i + \eta_i^* \xi_i^*) - \sum_{i=1}^{N} \alpha (\varepsilon + \xi_i^* - y_i + \langle w, x_i \rangle + b) - \sum_{i=1}^{N} \alpha^* (\varepsilon + \xi_i^* - y_i - \langle w, x_i \rangle - b)
\]

whereas \( L \) is a Lagrange function and \( \eta_i, \eta_i^*, \alpha_i, \alpha_i^* \) are Lagrange multipliers. The multiplier variable in the Lagrange function above must meet the following limitations.

\[
\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0
\]

By following the saddle point condition that the partial derivative of \( L \) function to main variables \((w, b, \xi_i, \xi_i^*)\) must be removed or equal to 0 in order to optimise the function (Smola & Scholkopf, 2004). The substitution of the derivative of \( L \) function against the variable \( w \) into a linear function will produce a function of the support vector expansion as follows.

\[
f(x) = \sum_{i=1}^{N} (x_i - x_i^*) \langle x_i, x \rangle + b
\]

where \( \langle x_i, x \rangle \) or \( K(x_i, x) \) are kernel functions which are used to satisfy Mercer’s conditions. Based on research conducted by Qu and Zhang (2016), the radial kernel is one of the most commonly used kernels in SVR because its parameter values can be optimized to improve modelling results (Caraka et al. 2019). The radial kernel can be formulated as follows.

\[
K_R(x_i, x) = e^{-\gamma \|x-x_i\|^2}
\]

whereas \( \gamma \) is a kernel parameter that implicitly defines the high dimensional feature space structure and controls the complexity of the model (Hu et al. 2013). Generally, SVR parameters, namely cost \((C)\), epsilon \((\varepsilon)\) and gamma \((\gamma)\) (Sahoo et al. 2019) must be determined precisely in order to produce a good model (Yasin et al. 2016, 2020; Caraka et al. 2019).

### 2.2.2 Recursive feature elimination-support vector regression (RFE-SVR)

RFE or also known as backward selection is a method of selecting predictor variables that is effective in predictive modelling (Pullanagari et al. 2018) (Caraka et al. 2020a). RFE improves generalization performance because it eliminates unaffected variables so as to minimize errors during training (Chen and Jeong 2007). RFE begins by utilizing all predictor variables then eliminating these variables one by one based on the ranking generated by a ranking method (Mishra and Mishra 2015). One of the ranking methods that can be used in RFE is SVR, thus it is known as the RFE-SVR method. With the pseudocode process in Fig. 5, each research area uses different independent variables.

The selection process is carried out based on the MSE criteria and the number of predictor variables. The selected subset was the subset with the least number of predictor variables and the smallest MSE or not significantly different from the smallest MSE. MSE is said not to differ significantly when the difference between the MSE and the smallest MSE is less than 1.

#### 2.2.3 Harris hawks optimization (HHO)

HHO is a metaheuristic algorithm that is used in parameter optimization problems (Lagunes et al. 2019). HHO is built on the hunting pattern of Harris hawks and prey behaviour which is formulated mathematically (Heidari et al. 2019). Two phases of Harris hawks hunting, namely exploration and exploitation, are phases that are modeled to find optimization results. Prior to this, HHO performance is very good and efficient in terms of computation (Abbasi et al. 2019).

1. **Exploration Phase**

In the exploration phase, the Harris hawk is in a state of waiting, looking for, and observing the situation of the prey. The position of the Harris hawk is random. Two conditions of the Harris hawk affected the probability value of \( q \) ranging from 0 to 1. The mathematical formulation of the position of the Harris hawk follows the Eq. 7.

\[
X(t+1) = \begin{cases} 
X_{\text{random}}(t) - r_1 X_{\text{random}}(t) - 2r_2 X(t) & q \geq 0.5 \\
X_{\text{rabbit}}(t) - X_{\text{avg}}(t) - r_3(UB - LB) + r_4(UB - LB) & q < 0.5 
\end{cases}
\]

where \( X(t+1) \) is the position of the Harris hawk. \( q \) indicates that the position of the Harris hawk is close to the prey. Meanwhile, \( q \geq 0.5 \) indicates that the position between Harris hawk is random. \( X_{\text{random}}(t) \) is the position of Harris hawk which is selected at random from the population. \( r_1, r_2, r_3, \) and \( r_4 \) are random numbers from 0 to 1. \( X_{\text{rabbit}}(t) \) is the position of the prey in the form of a rabbit. \( X_{\text{avg}}(t) \) is the average position of the Harris hawk. UB is the upper limit of the search space while LB is the lower limit. The mathematical formula of \( X_{\text{random}}(t) \) and \( X_{\text{avg}}(t) \) follows Eq. 8:
Determine training set $X$
Determine variable set of training set $F = \{f_1, f_2, \ldots, f_n\}$
Determine ranking method $R(X,F)$
For (each variable) do
  Do the ranking process to $F$
  Update the set variable by removing the last ranking variable the last ranking variable in $S$
End for
Return $S$

Fig. 5 Pseudocode RFE-SVR

$$idx = \lfloor N r_5 + 1 \rfloor$$
$$X_{rand}(t) = X(t)[idx]$$
$$X_m(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t)$$

(8)

where $N$ is the Harris hawk population. $r_5$ is a random number between 0 and 1. $X(t)[idx]$ indicates the position of the-idx hawk. $X_i(t)$ is the position of the-i hawk.

2. Transition between Exploration and Exploitation Phase
The transitional phase is influenced by the prey’s escaping energy. The mathematical formulation of the prey escaping energy ($E$) follows Eq. 9.

$$E_0 = 2r_6 - 1$$
$$E = 2E_0 \left(1 - \frac{t}{T}\right)$$

(9)

where $E_0$ is the initial strength of rabbit which ranges between $-1$ and 1. $r_6$ is a random number between 0 and 1 and $T$ is the maximum iteration. When the value of $E_0$ decreases from 0 to -1 indicates a weakened rabbit strength and an increasing value of $E_0$ from 0 to 1 indicates a strengthening rabbit strength (Heidari et al. 2019). The prey’s escaping energy will decrease with iteration. $|E| \geq 1$ means the hawk is still in the exploration phase while $|E| < 1$ means the hawk has entered the exploitation phase.

3. Exploitation Phase.
In the exploitation phase, the Harris hawk will attack prey suddenly from various directions. The prey has a chance to dodge the hawk’s attack. However, the probability of success or failure indicated by the value of $r$ which ranges between 0 and 1. $r < 0.5$ indicates success while $r \geq 0.5$ indicates failure. The hawk has 4 different strategies to deal with these conditions.

1. Soft besiege is a strategy when the prey still has the power to escape ($|E| \geq 0.5$) but is likely to fail to evade the hawk’s attack ($r > 0.5$). The hawk surrounds the prey continuously until the prey is exhausted. When the prey is tired then that’s when the hawk attacks suddenly. The mathematical formulation of this behavior follows the Eq. 10.

$$X(t+1) = AX(t) - E[JX_{rabbit}(t) - X(t)]$$

(10)

$X(t+1)$ is the position of the hawk. $AX(t)$ is the difference in distance between the hawk and the prey. $J$ is the jump strength of prey.

$$\Delta X(t) = X_{rabbit}(t) - X(t)$$
$$J = 2(1 - r_7)$$

where $r_7$ is a random number ranging between 0 and 1.

2. Hard Besiege is a strategy when the prey lacks of escaping energy ($E \leq 0.5$) and is likely to fail to evade the attack ($r \geq 0.5$). The hawk will attack the prey directly. The mathematical formulation of this behavior follows Eq. 12.

$$X(t+1) = X_{rabbit}(t) - E|\Delta X(t)|$$

(12)

3. Soft besiege with progressive rapid dives is a strategy when the prey still has the escaping energy ($|E| \geq 0.5$) and the probability of success in evading the attack ($r < 0.5$). The pattern of prey in escaping is explained through the concept of levy flight (LF) such as Eq. 13.

$$LF(x) = 0.01 \times \frac{|x| \times \sigma}{|v|^b};$$

where

$$\sigma = \left(\frac{\Gamma(1 + \beta) \sin\left(\frac{\pi \beta}{2}\right)}{\Gamma\left(\frac{1 + \beta}{2}\right) \times \beta \times 2^{\frac{\beta+1}{2}}}\right)^{\frac{1}{b}}$$

(13)

Furthermore, the movement of the hawk will follow Eq. 14.

$$Y = X_{rabbit}(t) - E|JX_{rabbit}(t) - X(t)|$$

(14)

when the movement turns out to be unsuccessful due to the deceptive movement of the rabbit, the hawk will explore the area by following the LF-based irregular pattern formulated through Eq. 15.
\[ Z = Y + S \times LF(D) \]  

where \( D \) is the dimension of the search space. \( S \) is a random vector of size \( 1 \times D \). Next, the hawk will update the position following the Eq. 16.

\[
X(t + 1) = \begin{cases} 
Y & \text{if } F(Y) < F(X(t)) \\
Z & \text{if } F(Z) < F(X(t))
\end{cases}
\]  

4. Hard besiege with progressive rapid dives is a strategy when the prey lacks of escaping energy (\( |E| < 0.5 \)) but is likely to be successful in evading the attack (\( r < 0.5 \)). In this strategy, the hawk will close its distance by sucking through its average position with the prey. This behavior follows the condition of hard besiege which is formulated with the Eq. 17.

\[
X(t + 1) = \begin{cases} 
Y & \text{if } F(Y) < F(X(t)) \\
Z & \text{if } F(Z) < F(X(t))
\end{cases}
\]  

where the values of \( Y \) and \( Z \) are obtained through the Eq. 18.

\[
Y = X_{\text{rabbint}}(t) - E|X_{\text{rabbint}}(t) - X_m(t)|
\]

\[
Z = Y + S \times LF(D)
\]

Based on the phases and strategies of the hawk in hunting rabbits, the HHO flow can be described by following the flowchart in Fig. 6.

**2.2.4 Harris hawks optimization-support vector regression (HHO-SVR)**

HHO-SVR is an SVR method whose parameter values are optimized using HHO. The HHO workflow that follows the flowchart in Fig. 6 is run iteratively to get the best parameters. The parameters obtained from the optimization process can generally improve the generalization performance of SVR (Jiang et al. 2013). The model training process using HHO-SVR can be seen through the flowchart in Fig. 7.

Based on Fig. 7, the process begins with data preparation. Then the data is divided into training, validation, and testing data. Training and validation data will be utilized in the optimization process. Prior to that, parameter initialization was carried out based on research from Kaimian et al. (2019; Setiawan et al. 2020). Some parameters that are generally initialized are population size (\( N \)), maximum iteration (\( T \)), dimensions, and parameter search space limits. Population size is the number of Harris hawks that exist and occupy a certain position. Maximum iteration is the maximum stage required by the Harris hawk to be very close to its prey. Dimensions indicate the number of parameters (positioning). The parameter search space limits are the upper and lower limits of the parameters (hunting environment). For population size, the value is in line with the probability of finding a global optimum (Gutowski 2005). However, population size affects computing time and computer memory, so it is necessary to determine the smallest possible size that is effective and efficient (Setiawan et al. 2020) as well as the limits of the parameter search space. In addition, the maximum iteration also affects the computation time, so it is necessary to apply stopping criteria.

Before entering the optimization process, it is necessary to determine a function that minimizes/maximizes a certain indicator so that the optimization objective can be achieved. This function is an objective function. In this study, the objective function is used to improve the evaluation of SVR by minimizing the mean absolute percentage error (MAPE) indicator. A small MAPE value indicates a good model formed. Next, the optimization process can be started.

The optimization process begins with generating parameter pairs with a number of population sizes that
meet the search space constraints. After that, the results will be iterated to get the fitness value using the objective function. The parameter that produces the fitness value will be used as the location of the rabbit. Then, the initial energy values, jump power, and escape energy possessed by the rabbit must be updated. Based on this value, the strategy will be determined, whether Harris hawk uses hard besiege, soft besiege, hard besiege with progressive rapid dives, or soft besiege with progressive rapid dives. The chosen strategy will determine the next location of hawk. When the chosen strategy is able to increase the hawk’s chances of getting its prey, then proceed to the next stage (next iteration). Instead, another strategy will be carried out based on the update of the rabbit’s strength. This process will be carried out continuously until the maximum iteration runs out.

Because in this study the stopping criterion was applied, it is likely that not all iterations will be executed. In this regard, the iteration will stop if the conditions meet the convergent MAPE criteria. The MAPE criteria are said to be convergent if five times in a row, there is no significant change in the resulting MAPE value. Changes in MAPE values are said to be no different if the difference is not more than 0.00001 (10^{-5}). When the stopping criteria are met, the parameter search process with HHO will end and the parameter results in the last iteration are the optimal parameters for SVR modeling. Therefore, the results of these parameters will be used in SVR training which will later produce the HHO-SVR model.

2.3 Evaluation metrics

2.3.1 Accuracy size

In a prediction, the results are not absolutely in accordance with the real situation. Prediction results tend to contain elements of uncertainty and error. However, it does not mean that the prediction results are not appropriate to use because the model used in definite predictions contains the smallest errors. Therefore, before the prediction results are applied, the accuracy must be calculated to determine whether the error magnitude can still be tolerated or not. According to Makridakis and Wheelwright (Makridakis and Wheelwright 1974), there is a standard statistical measure whose calculation results can reflect the accuracy of the prediction results. This research uses five accuracy measures, namely Mean Square Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), R-Square ($R^2$), and Variance Ratio (VR). The formula for this accuracy (Fan et al. 2020; Makridakis and Wheelwright 1974; Mohammadi and Mehdizadeh 2020; Mohammadi et al. 2015) is as follows:
\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |X_i - F_i|
\]

where \( n \) is the amount of data, \( X_i \) is the real data, and \( F_i \) is the predicted data. VR is the variance ratio of the predicted variance (\( \text{Var}_{\text{pred}} \)) and the real data variance (\( \text{Var}_{\text{real}} \)). The high VR value illustrates the stability of the better prediction model (Yang & Wang, 2017). The greater the VR value, the more stable the forecasting results obtained (Du et al. 2020; Xiao et al. 2015). Meanwhile, the MAPE accuracy measure is expected to be as small as possible because it indicates that the prediction error is small (accurate) (Zeinalnezhad et al. 2020). Lewis (1982) classified MAPE scores into four groups that reflect the goodness of the predicted results. Excellent for \( \text{MAPE} < 10\% \); Good
for 10% < MAPE < 20%; Fair for 20% < MAPE < 50%; Poor for MAPE > 50%. $R^2$ has a range between 0 and +1. The value of $R^2$ is close to +1 indicating that there is a linear relationship between the estimated value and the other value, whereas if $R^2$ is close to zero it indicates that there is no linear relationship (Mohammadi et al. 2015).

2.3.2 Diebold–Mariano test (DM)

The DM test is a test that can be used to see if there is a difference in accuracy between the two prediction results. The data used in this test is the error data from the prediction results. This test is classified as a non-parametric test because it is free of assumptions, such as normality, number of samples, zero mean, and correlation (Diebold and Mariano 1995). In this study, it was used to verify whether there was a significant difference between prediction 1 and prediction 2 (Niu and Wang 2019; Sun et al. 2019; Wang et al. 2018). Prediction 1 used is the HHO-SVR model. The test hypothesis in this test, respectively:

$H_0$ Error of prediction results 1 $\geq$ Error of prediction results 2.

$H_1$ Error of prediction results 1 < Error of prediction results 2.

With statistic test at (24):

$$S = \frac{\bar{d}}{\sqrt{\frac{2nf_d(0)}{f}}}$$

(24)

where $\bar{d}$ is the average difference of errors’ sample prediction. $\hat{f}_d(0)$ is a consistent estimate of $\hat{f}_d(0)$ obtained by adding the sum weights to the auto-covariant sample values. $\hat{f}_d(0)$ is the spectral density of the difference in prediction error with a frequency of 0. The Diebold–Mariano test statistic follows the standard normal distribution of N(0,1) (Harvey et al. 1997; Yang and Wang 2017).

2.4 Research flow

This research begins with the selection of significant predictor variables, which are used in the construction of predictive models. Then, the prediction model will be built using the SVR method which is optimized using HHO and several other algorithms as comparison, namely the Antlion Optimizer (ALO), Dragonfly Algorithm (DA), and Gray Wolf Optimizer (GWO). Amroune, et al. (Amroune et al. 2017) said that ALO-SVR is able to predict better voltage stability compared to ANFIS. Rahman and Rashid (2019) said that DA is an algorithm that is easy to implement into any predictive model. Meanwhile, GWO combined with SVR produces a good predictive model for predicting the amount of mold in rice (Sunli et al. 2018). Panahi et al. (2020) stated that the GWO-SVR is the best predictive model for predicting landslide vulnerability. The model that has been formed is then used in prediction and the results will be evaluated and tested. The complete flow of this research can be seen through the flowchart in Fig. 8.

3 Results

3.1 Initial setup

In this study, the data were divided into three sets, respectively the training set, validation, and testing. Distribution was carried out by a comparison of 60% training set, 20% validation set, and 20% test set (Kaimian...
| Sub-district         | Metaheuristic algorithm | Iteration (3) | Time (4) | Average time for each iteration (5) | Cost (6)  | Gamma (7) | Epsilon (8) |
|---------------------|-------------------------|---------------|----------|-------------------------------------|-----------|-----------|-------------|
| Kembangan           | ALO                     | 5             | 00:03:16 | 00:00:39                            | 423.2400  | 0.3422    | 0.0104      |
|                     | DA                      | 18            | 00:04:27 | 00:00:15                            | 1024.0000 | 0.0308    | 0.0039      |
|                     | GWO                     | 15            | 00:06:47 | 00:00:27                            | 915.8583  | 0.0316    | 0.0039      |
|                     | HHO                     | 12            | 00:03:37 | 00:00:18                            | 44.5637   | 0.0594    | 0.0039      |
| Menteng             | ALO                     | 5             | 00:06:00 | 00:01:12                            | 425.2400  | 0.3500    | 0.0039      |
|                     | DA                      | 21            | 00:09:02 | 00:00:26                            | 504.8412  | 0.0717    | 0.0039      |
|                     | GWO                     | 11            | 00:06:35 | 00:00:36                            | 294.9462  | 0.0821    | 0.0039      |
|                     | HHO                     | 12            | 00:07:06 | 00:00:36                            | 148.6889  | 0.0945    | 0.0039      |
| Kebayoran Baru      | ALO                     | 5             | 00:04:47 | 00:00:57                            | 424.2400  | 0.3461    | 0.0143      |
|                     | DA                      | 5             | 00:01:07 | 00:00:13                            | 424.2400  | 0.3461    | 0.0143      |
|                     | GWO                     | 8             | 00:08:30 | 00:01:04                            | 270.9313  | 0.1831    | 0.0039      |
|                     | HHO                     | 19            | 00:49:27 | 00:02:36                            | 383.1371  | 0.1676    | 0.0047      |
| Cakung              | ALO                     | 6             | 00:04:43 | 00:00:47                            | 425.2400  | 0.0078    | 0.0039      |
|                     | DA                      | 16            | 00:04:58 | 00:00:19                            | 65.3659   | 0.1000    | 0.0039      |
|                     | GWO                     | 7             | 00:02:07 | 00:00:18                            | 27.5240   | 0.1099    | 0.0051      |
|                     | HHO                     | 9             | 00:12:03 | 00:01:20                            | 93.5102   | 0.0774    | 0.0039      |
| Koja                | ALO                     | 5             | 00:03:15 | 00:00:39                            | 423.2400  | 0.3422    | 0.0104      |
|                     | DA                      | 9             | 00:02:05 | 00:00:14                            | 436.9753  | 0.3380    | 0.0101      |
|                     | GWO                     | 12            | 00:03:01 | 00:00:15                            | 26.7147   | 0.1578    | 0.0123      |
|                     | HHO                     | 9             | 00:09:11 | 00:01:01                            | 27.4987   | 0.1568    | 0.0039      |
| North Thousand Islands | ALO                  | 5             | 00:01:58 | 00:00:24                            | 424.2400  | 0.8480    | 0.0039      |
|                     | DA                      | 13            | 00:02:09 | 00:00:10                            | 613.4428  | 0.7821    | 0.0039      |
|                     | GWO                     | 6             | 00:03:29 | 00:00:35                            | 623.6411  | 0.7765    | 0.0039      |
|                     | HHO                     | 8             | 00:11:47 | 00:01:28                            | 827.2026  | 0.7257    | 0.0039      |
| Central Bogor       | ALO                     | 5             | 00:02:58 | 00:00:36                            | 506.6380  | 0.0084    | 0.0039      |
|                     | DA                      | 7             | 00:01:41 | 00:00:14                            | 511.6969  | 0.0376    | 0.0039      |
|                     | GWO                     | 6             | 00:11:05 | 00:01:51                            | 981.0511  | 0.0492    | 0.0039      |
|                     | HHO                     | 9             | 00:07:58 | 00:00:53                            | 1024.0000 | 0.0071    | 0.0039      |
| Cibinong            | ALO                     | 6             | 00:03:36 | 00:00:36                            | 1024.0000 | 0.0971    | 0.0039      |
|                     | DA                      | 11            | 00:02:36 | 00:00:14                            | 733.0358  | 0.1068    | 0.0039      |
|                     | GWO                     | 19            | 00:12:37 | 00:00:40                            | 625.2154  | 0.1143    | 0.0039      |
|                     | HHO                     | 9             | 00:04:40 | 00:00:31                            | 348.8484  | 0.1239    | 0.0039      |
| Pancoran Mas        | ALO                     | 6             | 00:04:58 | 00:00:50                            | 425.2400  | 0.8519    | 0.0039      |
|                     | DA                      | 13            | 00:04:29 | 00:00:21                            | 755.6371  | 0.7260    | 0.0039      |
|                     | GWO                     | 9             | 00:08:52 | 00:00:59                            | 344.2201  | 1.0000    | 0.0049      |
|                     | HHO                     | 10            | 00:21:04 | 00:02:06                            | 880.3684  | 0.7510    | 0.0039      |
| Tangerang           | ALO                     | 43            | 00:12:11 | 00:00:17                            | 489.8787  | 0.0619    | 0.0046      |
|                     | DA                      | 5             | 00:00:56 | 00:00:11                            | 424.2400  | 0.3461    | 0.0143      |
|                     | GWO                     | 19            | 00:07:52 | 00:00:25                            | 964.0312  | 0.0406    | 0.0039      |
|                     | HHO                     | 6             | 00:01:52 | 00:00:19                            | 77.6844   | 0.0319    | 0.0039      |
| Tigaraksa           | ALO                     | 9             | 00:07:39 | 00:00:51                            | 512.5000  | 0.0117    | 0.0039      |
|                     | DA                      | 5             | 00:01:14 | 00:00:15                            | 424.2400  | 0.3461    | 0.0143      |
|                     | GWO                     | 12            | 00:04:55 | 00:00:25                            | 1024.0000 | 0.0109    | 0.0045      |
|                     | HHO                     | 9             | 00:03:27 | 00:00:23                            | 183.6585  | 0.0133    | 0.0039      |
et al. 2019). Regarding parameter values, the range of SVR parameter values to be optimized is $C = [2^0, 2^{10}]$, $\gamma = [2^{-5}, 2^5]$, and $\epsilon = [2^{-5}, 2^5]$ (Liu et al. 2018). Furthermore, the value of other parameters such as the number of Harris hawk population is determined as much as 50 and a maximum of 100 iterations.

3.2 Variable selection

The selection of variables using the RFE-SVR utilizes several data, such as time in hours, air temperature, wind speed, humidity, UV index, air pressure, and significant lag of ozone which will be a predictor variable and ozone as a response variable (Sect. 2.2). Due to the significant lag of ozone, initial identification was made to determine this through the partial autocorrelation function (PACF) (Suhartono 2011). The consideration of using the lag variable from the dependent variable is part of a strategy to obtain a better predictive estimate (Wilkins 2018). Caraka et al. (2020b) conducted data pre-possessing with lag data from the dependent variable used in the independent variable. Based on Fig. 9, the amount of significant lag data from ozone to be selected is two lag data. Here is an example of a significant initial identification results of ozone lag data for Menteng and Central Bogor sub-districts.

Based on the picture above, the biggest lag partial correlation to ozone is lag 1 and 2. This can be seen from the correlogram bars which are higher than the others and pass through the light blue area. Correlogram bars that stick up indicate positive correlation and downward ones indicate negative correlation. In general, each sub-district has the same significant lag, namely lag 1 and 2. Furthermore, the data will be used in the RFE-SVR to obtain the following results.

According to Fig. 10, the RFE-SVR results prove that not all predictor variables are used to produce good model accuracy (Pullanagari et al. 2018). In general, lag 1 and 2 were significant predictor variables for each district. This shows that the use of past data for prediction has a fairly good effect on modelling (Wilkins 2018). Moreover, predictor variables in the form of air temperature, humidity, and UV index were also significant for most districts. This is consistent with the research of de Souza et al. (2018) which states that meteorological factors affect ground ozone concentration. However, wind speed, air pressure and time were only significant for a small proportion of the sub-districts.

3.3 Modelling and prediction

Modelling was drawn using SVR radial kernel. The parameters in the SVR will be adjusted to the results of optimization using HHO and other metaheuristic algorithms, namely ALO, DA, and GWO to see how well the HHO performs compared to these algorithms. In the optimization process, the time and number of iterations needed will be calculated until the MAPE value converges and the optimal cost, gamma, and epsilon parameter values.

Table 3 is the result of SVR parameter optimization using the four metaheuristic algorithms previously mentioned. Based on Table 3, there are quite varied results from the number of iterations, time, and optimal parameters obtained due to the influence of the type of optimization algorithm used. In general, HHO requires a relatively long time to search for optimal SVR parameters in the case of ozone concentration prediction modelling compared to
Table 4: The results of the evaluation using MAE, RMSE, and MAPE

| Sub-district       | Model       | Training MAE (1) | Training RMSE (2) | Training MAPE (3) | Test MAE (5) | Test RMSE (6) | Test MAPE (7) |
|-------------------|-------------|------------------|-------------------|------------------|-------------|---------------|---------------|
|                   |             | MAE (4) | RMSE (5) | MAPE (6) | MAE (7) | RMSE (8) | MAPE (9) |
| Kembangan         | ALO-SVR     | 5.2914 | 10.7730 | 8.1027 | 9.3175 | 17.2796 | 15.8704 |
|                   | DA-SVR      | 7.2557 | 13.8478 | 9.5405 | 7.7864 | 12.7201 | 14.0976 |
|                   | GWO-SVR     | 7.2543 | 13.8461 | 9.5432 | 7.7866 | 12.6953 | 14.0948 |
|                   | HHO-SVR     | 7.3221 | 13.8446 | 9.6312 | 7.7494 | 12.5356 | 14.1448 |
| Menteng           | ALO-SVR     | 4.0348 | 10.4083 | 8.8457 | 14.6883 | 30.1129 | 16.5053 |
|                   | DA-SVR      | 6.4475 | 13.6066 | 9.5891 | 12.5415 | 22.4830 | 16.6123 |
|                   | GWO-SVR     | 6.4216 | 13.5805 | 9.5891 | 12.5415 | 22.4830 | 16.6123 |
|                   | HHO-SVR     | 6.4429 | 13.6415 | 9.7375 | 9.4141 | 16.2255 | 16.6123 |
| Kebayoran Baru    | ALO-SVR     | 4.3023 | 9.6932  | 9.5891 | 12.5415 | 22.4830 | 16.6123 |
|                   | DA-SVR      | 6.4475 | 13.6066 | 9.5891 | 12.5415 | 22.4830 | 16.6123 |
|                   | GWO-SVR     | 6.4216 | 13.5805 | 9.5891 | 12.5415 | 22.4830 | 16.6123 |
|                   | HHO-SVR     | 6.4429 | 13.6415 | 9.7375 | 9.4141 | 16.2255 | 16.6123 |
| Cakung            | ALO-SVR     | 4.8255 | 10.5008 | 8.7495 | 8.0526  | 16.2118 | 13.8328 |
|                   | DA-SVR      | 4.8323 | 10.5241 | 6.2785 | 8.6266  | 17.1066 | 12.2086 |
|                   | GWO-SVR     | 6.2532 | 10.6757 | 9.6654 | 9.4754  | 16.2822 | 16.7167 |
|                   | HHO-SVR     | 6.5532 | 10.9907 | 7.3096 | 8.2574  | 13.0409 | 13.4381 |
| Koja              | ALO-SVR     | 4.8255 | 10.5008 | 8.7495 | 8.0526  | 16.2118 | 13.8328 |
|                   | DA-SVR      | 4.8323 | 10.5241 | 6.2785 | 8.6266  | 17.1066 | 12.2086 |
|                   | GWO-SVR     | 6.2532 | 10.6757 | 9.6654 | 9.4754  | 16.2822 | 16.7167 |
|                   | HHO-SVR     | 6.5532 | 10.9907 | 7.3096 | 8.2574  | 13.0409 | 13.4381 |
| North Thousand Islands | ALO-SVR     | 4.9075 | 11.2435 | 4.3916 | 3.0032  | 5.7569  | 4.3524 |
|                   | DA-SVR      | 4.9106 | 11.2581 | 4.3866 | 2.9592  | 5.7113  | 4.2635 |
|                   | GWO-SVR     | 4.8113 | 11.2569 | 4.3905 | 2.9742  | 5.7196  | 4.2997 |
|                   | HHO-SVR     | 4.8155 | 11.2681 | 4.3913 | 2.9592  | 5.6995  | 4.2700 |
| Central Bogor     | ALO-SVR     | 5.7810 | 10.7041 | 12.1166 | 6.2478 | 10.2258 | 15.8929 |
|                   | DA-SVR      | 4.8374 | 8.9101  | 10.7906 | 6.6115 | 11.1947 | 15.0112 |
|                   | GWO-SVR     | 3.7080 | 8.0739  | 9.8102 | 7.3733  | 12.3152 | 15.6830 |
|                   | HHO-SVR     | 5.7481 | 10.6587 | 12.1076 | 6.2669 | 10.2410 | 15.7519 |
| Clibinong         | ALO-SVR     | 8.3231 | 16.0574 | 21.3344 | 8.0667 | 14.3722 | 20.6413 |
|                   | DA-SVR      | 8.3122 | 16.0201 | 21.1645 | 8.0651 | 14.3961 | 20.5157 |
|                   | GWO-SVR     | 8.2839 | 15.9428 | 20.9129 | 8.0332 | 14.3770 | 20.1301 |
|                   | HHO-SVR     | 8.3287 | 16.0903 | 21.1599 | 8.1089 | 14.4361 | 20.5762 |
| Pancoran Mas      | ALO-SVR     | 6.0867 | 13.5747 | 23.8343 | 10.4548 | 28.1467 | 23.6148 |
|                   | DA-SVR      | 6.2118 | 13.6298 | 23.6386 | 10.5847 | 29.1693 | 23.6061 |
|                   | GWO-SVR     | 5.8503 | 13.2998 | 23.8647 | 10.7510 | 29.1245 | 23.9000 |
|                   | HHO-SVR     | 6.1109 | 13.5202 | 23.4747 | 10.7361 | 30.1391 | 23.6922 |
| Tangerang         | ALO-SVR     | 6.7808 | 11.9114 | 8.7512  | 7.4428 | 12.0701 | 13.6102 |
|                   | DA-SVR      | 4.8350 | 9.7954  | 7.1156  | 8.8736 | 14.9054 | 15.7767 |
|                   | GWO-SVR     | 6.4645 | 12.3794 | 8.5280  | 6.4189 | 10.1794 | 12.4010 |
|                   | HHO-SVR     | 7.1660 | 13.4689 | 8.8960  | 6.4355 | 10.5223 | 12.3587 |
| Tigaraksa         | ALO-SVR     | 6.3821 | 11.6606 | 8.9372  | 5.2927 | 8.3243  | 13.5201 |
|                   | DA-SVR      | 3.3725 | 6.6176  | 6.3227  | 11.5291 | 20.9583 | 24.2082 |
|                   | GWO-SVR     | 6.3266 | 11.5929 | 8.8641  | 5.2129 | 8.2879  | 13.2639 |
|                   | HHO-SVR     | 6.4577 | 11.8330 | 8.9629  | 5.3653 | 8.3445  | 13.7711 |
other optimization algorithms. This shows that the chances of finding the optimal parameter are greater due to the more complex processes that are run in the HHO algorithm. Furthermore, the modelling results will be used for prediction. The test set will be used to obtain the prediction results and its accuracy.

### 3.4 Evaluation of predicted results

To find out the errors contained in the prediction results, several measures of accuracy were calculated. The following are the results of predictive evaluations that have been carried out using the accuracy measures of MAE, RMSE, MAPE, $R^2$, and VR. Based on the evaluation results in Table 4, it can be seen that the accuracy of the prediction results between these methods is different from one another. The difference between the accuracy of training and testing is also not much different, which indicates that the model is fit. Judging from the value highlighted in yellow, these values are the best MAE, RMSE, and MAPE values in each sub-district. It is said to be the best because the value is closest to 1 so that it shows the better the model in predicting (Du et al. 2020; Mohammadi et al. 2015; Yang and Wang 2017). A total of 7 out of 14 sub-districts using the HHO-SVR model produced a higher $R^2$ than other methods, namely Kembangan, Menteng, Kebayoran Baru, North Thousand Islands, Tangerang, Ciputat, and South Bekasi. By using VR, there are also 6 areas that use HHO-SVR that are superior to other methods. This shows that HHO-SVR is able to predict Ozone better than other optimization methods. This is in line with Cao et al. (2020) who say that HHO-SVR has good performance in predicting various types of data (Table 5).

### 3.5 Prediction result testing

To find out whether the accuracy of the prediction results is statistically different, it is necessary to do one test, namely the DM test. The following are the test results with the HHO-SVR prediction results as a comparison. According to the results of the DM test in Table 6, with a significance level of 10%, most of the test results show the accuracy of the HHO-SVR prediction model is better than other models. Of the 14 regions modelled, 8 regions modelled with HHO-SVR produced better accuracy than ALO-SVR, 9 regions modelled with HHO-SVR produced better accuracy than DA-SVR, and 6 regions modelled with HHO-SVR provides better accuracy than GWO-SVR.

### 4 Discussions

The literature that implements HHO in SVR optimization is still sparse because HHO is a metaheuristic algorithm recently designed by Heidari et al. (2019). Based on the

### Table 4 (continued)

| Sub-district | Model  | Training MAE | RMSE | MAPE | Test MAE | RMSE | MAPE |
|--------------|--------|--------------|------|------|----------|------|------|
| (1)          | (2)    | (3)          | (4)  | (5)  | (6)      | (7)  | (8)  |
| Ciputat      | ALO-SVR | 3.8172       | 9.7248 | 10.6004 | 12.1429 | 22.0601 | 24.4188 |
|              | DA-SVR  | 7.7853       | 14.4656 | 14.0519 | 7.4201  | 12.7418 | 17.1620 |
|              | GWO-SVR | 7.8086       | 14.5078 | 14.0274 | 7.3996  | 12.7125 | 16.8982 |
|              | HHO-SVR | 7.4060       | 13.9598 | 13.6100 | 7.1641  | 12.3950 | 17.8710 |
| South Bekasi | ALO-SVR | 6.7232       | 12.9864 | 8.9698  | 7.7912  | 13.7529 | 12.3903 |
|              | DA-SVR  | 6.7256       | 12.9968 | 8.9643  | 7.7794  | 13.7241 | 12.3614 |
|              | GWO-SVR | 6.7894       | 13.0630 | 8.9943  | 7.7574  | 13.4720 | 12.7105 |
|              | HHO-SVR | 7.9507       | 14.6241 | 10.2255 | 7.6057  | 12.6481 | 12.1553 |
| Central Cikarang | ALO-SVR | 5.8243       | 10.3448 | 11.8884 | 7.0909  | 10.2859 | 18.1418 |
|              | DA-SVR  | 4.7988       | 9.3083  | 10.6533 | 7.4732  | 11.1613 | 19.3587 |
|              | GWO-SVR | 4.8574       | 9.3812  | 10.5582 | 7.3663  | 11.0528 | 18.8495 |
|              | HHO-SVR | 5.0870       | 9.4838  | 11.2709 | 7.2826  | 10.7823 | 18.7812 |
application of HHO with SVR in the construction of prediction models with case studies of ozone concentration, most of the prediction accuracy generated by adjusting parameter values using HHO is better than other meta-heuristic algorithms, namely ALO, DA, and GWO. The finding in this study is that there is a process that must be carried out first, namely the RFE-SVR which helps to select the appropriate independent variables in each region (Chen and Jeong 2007). It turns out from this research that it is evident that each region has different independent variables that affect the ozone concentration. However, it should also be noted about including the lag of the dependent variable, making the prediction model estimation for each region better (Wilkins 2018). As in the North Thousand Islands, only the dependent variable lag (lag 1 and lag 2) affects the ozone variable, and there is no other independent variable that affects it.

The importance of predicting current and future ozone concentrations makes prediction models to be built for this need. In accordance with the results of the evaluation and DM test, the HHO-SVR is a model that can be chosen because its performance is quite good in producing accuracy that is close to the actual situation so that the observations are accurate or can be said to be stable. This is in line with research from Cao et al. (2020) which states that HHO-SVR is a model with good predictive accuracy and stable prediction performance (Setiawan et al. 2020). Although in some areas using the SVR model with optimization ALO, DA and GWO are better.

The research of Ismael et al. (2020, 2021) proposes HHO to optimize v-SVR hyperparameters by embedding feature selection simultaneously or without separating the process. This study also uses only a few chemical datasets to see the performance of the proposed method. In addition, the drawback of this study is that it only uses MSE in the evaluation. Whereas MSE has low reliability because the results can be different depending on different data fractions (Shcherbakov et al. 2013). Elgamal et al. (2020) used HHO as a comparison algorithm against HHO improvisation on KNN embedded with feature selection simultaneously. This study only uses medical datasets from the UCI machine learning repository to test the performance of the proposed method. The evaluation metrics used include the number of selected features, classification accuracy, fitness value, Wilcoxon statistical test, and convergence curve. The study did not explain in depth about HHO. In addition, the application of HHO in the study is for classification. Cao et al. (2020) used the HHO-SVR hybrid model only as a comparison method from the proposed method and was not explained in detail and without being preceded by the selection of the predictor variables used. The evaluation metrics used include RCRank, RCrave, Wilcoxon’s statistical test, and CCRank. Sammen et al. (2020) used HHO to optimize the performance of the ANN without prior feature selection. The data used in this study is data that has been used in related research. The evaluation metrics used include MAE, RMSE, MAPE, coefficient of correlation (CC), Willmott index (WI). In contrast to the research previously mentioned, this study proposes HHO to

| Sub-District                      | R²  | ALO-SVR | DA-SVR | GWO-SVR | HHO-SVR | VR   |
|----------------------------------|-----|---------|--------|---------|---------|------|
|                                  | (1) | (2)     | (3)    | (4)     | (5)     | (6)  | (7)  | (8)  | (9)  |
| Kembangan                        |     | 0.9507  | 0.9728 | 0.9729  | 0.9735  | 0.9902| 0.9951| 0.9955| 0.9988|
| Menteng                          |     | 0.8883  | 0.9616 | 0.9613  | 0.9624  | 0.9855| 0.9720| 0.9735| 0.9742|
| Kebayoran Baru                   |     | 0.9332  | 0.9332 | 0.9637  | 0.9641  | 0.9987| 0.9987| 0.9630| 0.9640|
| Cakung                           |     | 0.9742  | 0.9693 | 0.9706  | 0.9704  | 0.9420| 0.9510| 0.9495| 0.9512|
| Koja                             |     | 0.9340  | 0.9536 | 0.9722  | 0.9716  | 0.9720| 0.9720| 0.9736| 0.9772|
| North Thousand Islands           |     | 0.9708  | 0.9712 | 0.9711  | 0.9713  | 0.9854| 0.9873| 0.9857| 0.9861|
| Central Bogor                    |     | 0.9763  | 0.9717 | 0.9658  | 0.9762  | 0.9961| 0.9989| 0.9935| 0.9947|
| Cibinong                         |     | 0.9705  | 0.9705 | 0.9705  | 0.9703  | 0.9946| 0.9963| 0.9963| 0.9990|
| Pancoran Mas                     |     | 0.9196  | 0.9149 | 0.9150  | 0.9102  | 0.8373| 0.8272| 0.8293| 0.8194|
| Tangerang                        |     | 0.9699  | 0.9552 | 0.9782  | 0.9766  | 0.9912| 0.9670| 0.9894| 0.9974|
| Tigaraksa                        |     | 0.9796  | 0.8789 | 0.9796  | 0.9797  | 0.9890| 0.9899| 0.9111| 0.9874|
| Ciputat                          |     | 0.9283  | 0.9761 | 0.9762  | 0.9776  | 0.9204| 0.9762| 0.9774| 0.9759|
| South Bekasi                     |     | 0.9680  | 0.9682 | 0.9694  | 0.9729  | 0.9956| 0.9953| 0.9973| 0.9792|
| Central Cikarang                 |     | 0.9756  | 0.9715 | 0.9720  | 0.9732  | 0.9534| 0.9858| 0.9855| 0.9930|
This study has several shortcomings, both technical and assumptions used. Some of these shortcomings include the data used are still in a limited range. The data used is only in the span of 2 months or about 1368 records per area. The research data provided by the website is assumed to be valid. No field checks were carried out to ensure that the data provided by the web matched the actual situation. In addition, there are some limitations of tools or needs in web scraping such as the internet and computers. This causes some data to not be successfully retrieved so that there is missing data. Missing data must be treated properly so that it is not filled with wrong data. Therefore, a data imputation method is applied, namely linear interpolation with the assumption that the data between hours does not change far enough from the value before or after it. Then, the SVR kernel used in this study is only a radial kernel based on a literature study (Qu and Zhang 2016) which states that the radial kernel is one of the suitable kernels to use in SVR because its parameters can be optimized so as to improve prediction results. Whereas there are other kernels such as sigmoid, polynomial, and linear. In addition, the determination of the range of the parameter search space in the optimization process using HHO only uses the parameter range from related studies (Liu et al. 2018). It should be possible to determine the parameter range based on the simulation so that better results are obtained.

**5 Conclusions**

Based on the results and discussion, there are several things that can be concluded that in relation to the selection of variables, most of the results of the RFE-SVR show that the variables lag 1, lag 2, air temperature, humidity, and UV index are significant variables in modeling predictions of ozone concentration. HHO takes longer time than other optimization algorithms in finding optimal SVR parameters. On average, 7 of the 14 sub-districts modeled produced the best predictions with the HHO-SVR model (Setiawan et al. 2020) with MAE evaluation results below 10, RMSE and MAPE below 20, R2 around 0.97, and VR around 0.98. The resulting MAPE value is classified as a good prediction result. Then, the results of the Diebold–Mariano test also show that the accuracy of the prediction results and the stability of the performance of the HHO-SVR model is better, especially for the Ciputat and South Bekasi sub-districts. This shows that the two sub-districts are very suitable to use HHO-SVR in predicting ozone concentrations.

Based on the limitations, future research should be able to increase the time span of data collection and validate data from web scraping in the field. Regarding the handling of missing data on pollutant data and others, it may be possible to use other imputation methods such as the EM Algorithm to get more accurate results (Junger and Ponce...
de Leon (2015). In addition, predictor variables can also be added using precursor concentration data including CH4, CO, NMHC, NO, NO2, THC (Wasi’ah and Driejana 2017) which can add information in predictions. Then, the selection of variables using RFE can be compared with new methods such as Ensemble of Regressor Chains-guided Feature Ranking (Masmoudi et al. 2020) to find out whether RFE’s performance is the same or inferior to that method. In addition, HHO can also be compared with new metaheuristic algorithms such as the Equilibrium Optimizer (Faramarzi et al. 2020), Archimedes Optimization (Hashim et al. 2021), and Water Streams Optimization (Majani and Nasri 2021). In addition, simulations related to the range of the parameter search space can be carried out to obtain a shorter range so as to save optimization time.

Acknowledgements This study is fully supported by the Polytechnic Statistics STIS, Jakarta, Indonesia.

Declarations

Conflict of interest The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication, double publication, submission, and redundancy have been completely witnessed by the authors.

References

Abbasi A, Firouzi B, Sendur P (2019) On the application of Harris hawks optimization (HHO) algorithm to the design of microchannel heat sinks. Eng Comput. https://doi.org/10.1007/s00366-019-00892-0

Amroune M, Musirin I, Bouktir T, Othman MM (2017) The amalgamation of SVR and ANFIS models with synchronized phasor measurements for on-line voltage stability assessment. Energies. https://doi.org/10.3390/en10111693

Bui DT, Moayedi H, Kalantar B, Osooli A, Pradhan B, Nguyen H, Rashi ASA (2019) A novel swarm intelligence—Harris hawks. Sensors 19(16):3590

Cao W, Liu X, Ni J (2020) Parameter optimization of support vector regression using Henry gas solubility optimization algorithm. IEEE Access 8:88633–88642. https://doi.org/10.1109/ACCESS.2020.2993267

Caraka RE, Bakar SA, Tahmid M, Yasin H, Kurniawan ID (2019) Neurocomputing fundamental climate analysis. Telkomnika 17(4):1818–1827. https://doi.org/10.12928/TELOKNIKA.v17i4.11788

Caraka RE, Chen RC, Bakar SA, Tahmid M, Toharudin T, Pardamean B (2020a) Employing best input SVR robust lost function with nature-inspired metaheuristics in wind speed energy forecasting. IAENG Int J Comput Sci 47(3):1–13

Caraka RE, Chen RC, Toharudin T, Tahmid M, Pardamean B, Putra RM (2020b) Evaluation performance of SVR genetic algorithm and hybrid PSO in rainfall forecasting. ICIC Express Lett Part B Appl 11(7):631–639. https://doi.org/10.24507/icicellb.11.07.631

Caraka RE, Lee Y, Kurniawan R, Herliansyah R, Kaban PA, Nasution BI, Gio PU, Chen RC, Toharudin T, Pardamean B (2020c) Impact of COVID-19 large scale restriction on environment and economy in Indonesia. Glob J Environ Sci Manag 6(Special Issue):65–84. https://doi.org/10.22034/GIJESM.2019.06.SI.07

Chen X, Jeong JC (2007) Enhanced recursive feature elimination. In: Sixth international conference on machine learning and applications enhanced, pp 429–435. https://doi.org/10.1109/ICMLA.2007.35

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263. https://doi.org/10.1080/07350015.1995.10524599

Drucker H, Surges CJ, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. Adv Neural Inf Process Syst 1:155–161

Du P, Wang J, Hao Y, Niu T, Yang W (2020) A novel hybrid model based on multi-objective Harris hawks optimization algorithm for daily PM2.5 and PM10 forecasting. Appl Soft Comput J 96:106620. https://doi.org/10.1016/j.asoc.2020.106620

Elgamal ZM, Yasin NBM, Tubishat M, Alsawtaiti M, Mirjalili S (2020) An improved harris hawks optimization algorithm with simulated annealing for feature selection in the medical field. IEEE Access 8:186638–186652. https://doi.org/10.1109/ACCESS.2020.3029728

Fan J, Wu L, Ma X, Zhou H, Zhang F (2020) Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renew Energy 145(2):2034–2045. https://doi.org/10.1016/j.renene.2019.07.104

Faramarzi A, Heidarinejad M, Stephens B, Mirjalili S (2020) Equilibrium optimizer: a novel optimization algorithm. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.105190

Gutowski MW (2005) Biology, physics, small worlds and genetic algorithms. In: Shannon S (ed) Leading edge computer science research. Nova Science Publishers, New York

Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared errors. Int J Forecast 13(2):281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Hashim FA, Hussain K, Houssein EH, Mabrouk MS, Al-Atabany W (2021) Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems. Appl Intell 51(3):1531–1551. https://doi.org/10.1007/s10489-020-18893-z

Heidari AA, Mirjalili S, Faris H, Aljaraah I, Mafarja M, Chen H (2019) Harris hawks optimization: algorithm and applications. Future Gener Comput Syst 97:849–872. https://doi.org/10.1016/j.future.2019.02.028

Hu Z, Bao Y, Xiong T (2013) Electricity load forecasting using support vector regression with memetic algorithms Zhongyi. Sci World J. https://doi.org/10.1155/2013/292575

Huang ML, Hung YH, Lee WM, Li RK, Jiang BR (2014) SVM-RFE based feature selection and taguchi parameters optimization for multiclass SVM classifier. Sci World J. https://doi.org/10.1155/2014/795624

Ismael OM, Qasim OS, Algalam ZY (2020) Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in v-support vector regression based on opposition-based learning. J Chemom 34(11):1–14. https://doi.org/10.1002/cem.3311

Ismael OM, Qasim OS, Algalam ZY (2021) A new adaptive algorithm for v-support vector regression with feature selection using Harris hawks optimization algorithm. J Phys Conf Ser. https://doi.org/10.1088/1742-6596/1897/1/012057

Jiang M, Jiang S, Zhu L, Wang Y, Huang W, Zhang H (2013) Study on parameter optimization for support vector regression in solving the inverse ECG problem. Comput Math Methods Med. https://doi.org/10.1155/2013/158025

Junger WL, Ponce de Leon A (2015) Imputation of missing data in time series for air pollutants. Atmos Environ 102:96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049
concentration prediction. Environ Res 158:105–117. https://doi.org/10.1016/j.envres.2017.06.002

Yasin H, Caraka RE, Hoyyi A (2016) Prediction of crude oil prices using support vector regression (SVR) with grid search: cross validation algorithm. Glob J Pure Appl Math 12(4):3009–3020

Yasin H, Caraka RE, Hoyyi A, Sugito (2020) Stock price modeling using localized multiple kernel learning support vector machine. ICIC Express Lett Part B Appl 11(4):333–339. https://doi.org/10.24507/icicelb.11.04.333

Zeinalnezhad M, Chofreh AG, Goni FA, Klemič JJ (2020) Air pollution prediction using semi-experimental regression model and adaptive neuro-fuzzy inference system. J Clean Prod. https://doi.org/10.1016/j.jclepro.2020.121218

Zhang JJ, Wei Y, Fang Z (2019) Ozone pollution: a major health hazard worldwide. Front Immunol 10(OCT):1–10. https://doi.org/10.3389/fimmu.2019.02518

Zhao H, Wang S, Wang W, Liu R, Zhou B (2015) Investigation of ground-level ozone and high-pollution episodes in a megacity of eastern China. PLoS ONE 10(6):1–20. https://doi.org/10.1371/journal.pone.0131878

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.