International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 6
ISSN: 2088-5334

Improvement Support Vector Machine Using Genetic Algorithm in Farmers Term of Trade Prediction at Central Java Indonesia

Ifran Lindu Mahargya[a,1], Guruh Fajar Shidik[a,2]

[a] *Faculty of Computer Science, Dian Nuswantoro University, Semarang City, Indonesia*
E-mail: [1]ifran.mahargya@gmail.com; [2]guruh.fajar@research.dinus.ac.id

***Abstract*— The welfare of farmers is a strategic problem in Indonesia. The Farmer term of the trade (FTT) is one indicator to measure the welfare of farmers. FTT is a measurement of the comparison of the price index received by farmers (It) with the price index paid by farmers (Ib). Some models for FTT prediction on previous research are using ANN, SVM, MLR, Markov Chain - Predictive Probabilistic Architecture Modeling Framework (P2AMF), Singular Spectrum Analysis (SSA) - ARIMA and ANN-PSO. Previous FFT research in 2018 used three prediction methods, namely using the ANN, SVM and MLR algorithms with the best RMSE being 0.00098. Then in 2019, FTT research was followed by optimization of the ANN parameters using PSO for weighting (ANN-PSO) and obtaining the best RMSE was 0.00062. This study evaluates the robustness of the prediction models of FTTs in the Central Java region using SVM. Then proceed with increasing SVM prediction accuracy using GA (SVM-GA). SVM-GA has resulted in an increase in FTT prediction accuracy. This study has found that the SVM method has better robustness than the ANN method. The development of research related to the accuracy for the FTT prediction model in the Central Java Province has increased, starting from 2018 with RMSE 0.00098; in 2019 with RMSE 0.00062 and the results of this study resulted in the best RMSE of 0.00037.**

***Keywords*— farmer term of trade; FTT; support vector machine; neural network; multi linear regression; genetic algorithm.**

I. INTRODUCTION

Increasing public welfare is one of the priorities of the Indonesian Government's performance. Indonesia is an agricultural country with a large population that predominantly works in the agricultural sector. Central Java Province is based on business fields in August 2017 and August 2018, the population most work in the agricultural business field as many as 4.32 million people or 25.16 percent in August 2017 and 4.20 million people or 24.38 percent in August 2018 [1]. The economic structure of Central Java Province according to the main business field in 2017, one of the main business fields that are dominant is the agricultural sector by 14.09 percent [1]. One of the indicators to measure the welfare of farmers is the value of the Farmer Term of Trade (FTT) or called NTP (*Nilai Tukar Petani*) in Indonesia.

Central Java Province as one of the provinces of agricultural potential, so the Central Java Provincial Government needs to implement appropriate policies in the agricultural sector so that the increase in FTT for the welfare of farmers is right on target. Several factors can affect the size or size of FTT, such as inflation and the large number of costs needed by farmers to meet their agricultural production needs.

Research for FTT prediction needs to be done to find a good prediction model. Several prediction models were carried out in previous studies, namely using Artificial Neural Network (ANN), Support Vector Machine (SVM), Multi Linear Regression (MLR), Markov Chain - Predictive Probabilistic Architecture Modeling Framework (P2AMF), Singular Spectrum Analysis (SSA) - ARIMA and Artificial Neural Network - Particle Swarm Optimization (ANN-PSO). These models have been used by previous researchers for FTT prediction models in Indonesia with good results.

This research will focus on FTT in Central Java Province. Previous Central Java Province FTT research in 2018 used 3 (three) prediction methods, namely using ANN, SVM and MLR algorithms [2]. Then the research of Central Java Province FTT continued with optimization on ANN parameters using PSO for weighting [3]. The results of research conducted by Rizchi Eka Wahyuni[3] resulted in a better Root Mean Square Error (RMSE) than the research conducted by Ifran et al. [2].

SVM could generalize good data even with a limited dataset, and this is one of its advantages. While ANN with limited dataset conditions will reduce the ability to generalize data, this research will build a robust, robustness prediction model. Previous research in 2019 [3] has produced a good prediction model for FTT in Central Java

Province. This is evidenced by the results of a more RMSE evaluation of ANN, SVM and MLR or can be said to be better than the study in 2018 [2]. But the prediction model built in the 2019 study [3] still has poor reliability. This is shown in the dataset used for the ANN method. ANN will decrease the ability to generalize data to a limited dataset. So, the 2019 study [3] used the data-splitting method with Leave-One-Out Cross-Validation (LOOCV) to overcome the weaknesses of ANN generalization on a limited dataset.

SVM is one of the Machine Learning algorithms, derived from statistical learning techniques discovered by Vapnik and Chervonenkis in 1992. SVM is an efficient classification technique with the ability to solve nonlinear problems. This study will use the SVM algorithm to build reliable prediction models. Previous research has been conducted by using SVM in developing prediction models for FTT in Central Java Province [2]. But the SVM that was built still has several weaknesses that need further research, namely optimization of the SVM method, replacement of the data splitting method and finding the optimal window size. The window size can affect the construction of the prediction model. As in the 2019 study [3], the window size can help improve predictive accuracy. So, the window size needs to be tested to get optimal prediction accuracy. The SVM model was built using windowing with 12 (twelve) window sizes and data splitting with the k-Fold Cross Validation method (k = 10).

This research aims to find out the weaknesses in the previous studies [2], [3]. In the first step, a prediction model was built for FTT with SVM using 2 (two) data splitting methods, namely, Hold Out Cross Validation (HOCV) and Leave-One-Out Cross-Validation (LOOCV). This was done to prove SVM robustness from the results of the study [3], namely ANN-PSO. In this step, the optimal window size search is also performed. Next in the second step is to optimize the SVM method with a metaheuristic approach using Genetic Algorithm (GA) or it can be called SVM-GA to produce optimal RMSE from SVM in the first step while proving from the results of the study [2]. In this step, the optimal window size search is also performed.

SVM-GA is a combination of SVM and GA (hybrid) methods to obtain an optimal data generalization model. In this study, we will focus on SVM-GA which is expected to reduce the RMSE of SVM in the 2018 study [2] and prove the reliability of SVM against ANN-PSO in the 2019 study [3]. Thus it can be produced a good prediction model of FTT in Central Java Province. The formulation of the problem to be raised in this study is the SVM reliability test, SVM-GA optimization test and find the best window size for SVM-GA and SVM.

II. Materials and Method

The role of the Central Java Provincial Government in supporting the agricultural sector is through farmers' assistance policies to improve FTT. Determination of appropriate farmer assistance policies by the Central Java Provincial Government must be known FTT scores first. The Central Java Provincial Government needs tools to find out FTT earlier, so the Central Java Provincial Government can determine policies that are right on target. Related research and FTT research methods will be discussed in the next discussion.

A. Related Works

Several Indonesian FTT studies that have been carried out have shown good results. Zulyadi et al. [4] in 2015 conducted research on FTT prediction in the plantation subsector to assist the Riau Provincial Government of Indonesia. The method used is Markov Chain and Predictive, Probabilistic Architecture Modeling Framework (P2AMF). The study resulted in Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) were 2.3954 and 0.0133. Ifran et al. [2] in 2018 the research conducted on the prediction of FTT in Central Java Province using the method of Artificial Neural Network (ANN), Support Vector Machine (SVM), and Multi Linear Regression (MLR). Producing the best RMSE in MLR is 0.00098, while ANN and SVM produce RMSE of 0.00112. Rizchi [3] in 2019 continued the research of Ifran et al. [2] that is the prediction of FTT in Central Java Province by optimizing ANN using PSO for weighting. That produces the best RMSE of 0.00062.

We continued the research of Ifran et al. [2] by optimizing SVM and Rizchi Eka Wahyuni's research [3] with experiments on data splitting methods on SVM to obtain reliable prediction models. Rizchi Eka Wahyuni's research [3] used 2 (two) data splitting methods (HOCV and LOOCV) to obtain optimal accuracy in ANN while also covering ANN's weaknesses. The hybrid method for FTT prediction is one of the techniques for machine learning optimization to obtain optimal accuracy.

Optimization of the SVM method using the metaheuristic approach using Genetic Algorithm (GA) has been carried out by researchers to determine SVM parameters. Syed Rahat Abbas and Muhammad Arif [5] in 2006 conducted an Electric Load Forecasting research using SVM-GA. Produce MAPE of 1.93 with optimal SVM parameters at C = 0.3050; σ = 0.9901; ε = 0.5. Xiaogang Chen [6] in 2009 conducted the Railway Passenger Volume Forecasting research using SVM-GA. MAPE SVM-GA is better than Radial Basis Function - Neural Network (RBF-NN). SVM-GA with MAPE 1.6891 and RBF-NN with MAPE 4.7196. Jianguo Zhou and Huaitao Liang [7] in 2010 conducted research on Prediction of the NOx Emissions from Thermal Power Plant using SVM-GA. Produce Et error relative and RMS relative error (ERMS) SVM-GA better than SVM and BPNN. ERMS SVM-GA is 1.79 percent, ERMS SVM is 2.61 percent, and ERMS BPNN is 2.66 percent. Gu Jirong et al. [8] conducted a forecasting study on housing prices in 2011 using SVM-GA. Produces MAPE SVM-GA of 1.94 and MAPE Grey Model (GM) of 9.52.

Tao Yerong et al. [9] conducted an intrusion detection on computer networks in 2014 using SVM-GA. SVM-GA accuracy is better than SVM and RBF-NN. SVM-GA accuracy is 91.12 percent, while SVM is 88.79 percent and RBF-NN is 85.63 percent. Phan et al. [10] in 2016 conducted a study of GA-based SVM parameters for classification problems. SVM-GA produces the best accuracy of 99.90. Zhang et al. [11] in 2017 conducted a study on predictions of compound retention in gas chromatography using GA-SVR-PO. GA-SVR-PO produces

an accuracy of 96.03 percent. Shafizadeh et al. [12] in 2017, conducted a study of the estimated growth of urban areas using SVM-GA. GA can improve SVM performance with 8.5 percent linear kernel, 5 percent polynomial kernel, 6.5 percent RBF kernel, and 3 percent sigmoid kernel. Miao et al. [13] in 2017 conducted a study predicting the behavior of landslide movements using SVM-GA, SVM-GA (cpg), SVM-PSO, SVM-Grid Search (SVM-GS). SVM-GA produces better RMSE compared to SVM-PSO and SVM-GS. SVM-GA has RMSE of 12.322 and 19.247. SVM-GA (cpg) has RMSE of 11.298 and 19.073. Alade et al. [14] in 2018 conducted a study on the prediction of oxygen bias and hemoglobin deoxygenation using SVM-GA. SVM-GA produced RMSE of 0,00039 for oxygen data and 0,00039 for hemoglobin deoxygenation data.

B. Decimal Scaling Normalization

Decimal scaling method is by changing the data attribute by moving the decimal value in the desired direction [15]. Decimal scaling normalization can use the following formula:

$$NEWDATA = \frac{DATA}{10^i} \quad (1)$$

Where NEWDATA is normalized data results, while i is the desired scaling value.

C. Sliding Window Algorithm (SWA)

Sliding window or called windowing on a dataset is a method by making a series of temporary data from the results of the division of observation data into several segments based on actual time series data. The window size and segment can be adjusted with the smallest error result. In Figure 1, it can be seen as an example of applying the windowing method [16], [17].

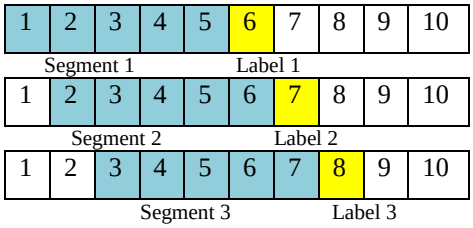

Fig. 1 Example of the windowing method.

D. Hold-Out Cross-Validation

Hold-Out Cross-Validation (HOCV) is a method that is widely used and popular because of its efficiency and convenience. Hold-Out Cross-Validation is a good data splitting method for building time-series data prediction models [18]. Split on HOCV is divided using ratios, for example, 60:40 or 60% so 60% of the dataset will be used as a training subset, and 40% will be used as a testing subset.

E. Leave-One-Out Cross-Validation

Leave-One-Out Cross-Validation (LOOCV) is another method of k-Fold Cross-Validation by making testing subset in all segments one by one. Its strength is that it can overcome a limited dataset problem, and its weakness is having a poor computational performance time when used in large numbers of datasets [3]. LOOCV has the performance of dividing subset in a dataset with k-1 as training subset, and 1 subset will be used as a testing subset. The divided subset is repeated up to iteration = k. With k is the number of datasets.

F. Support Vector Machine

The method used by SVM is mapping data into low-dimensional space into high-dimensional space using the kernel function. SVM can solve nonlinear classification problems when the classification of nonlinear data in low-dimensional space is not resolved, so SVM resolves the problem using the help of the kernel function to solve the problem [2]. The problems that are solved by SVM always use kernel nonlinear programming functions. In general, there are 4 (four) types of kernels that are often used [10]–[14], namely as follows:

Linear Kernel:

$$K(x, x_k) = x_k^T x \quad (2)$$

Polynomial Kernel:

$$K(x, x_k) = (x_k^T x + 1)^d \quad (3)$$

Radial Basis Function (RBF) Kernel:

$$K(x, x_k) = \exp\{-\| x - x_k \|_2^2 / \sigma^2\} \quad (4)$$

Sigmoid Kernel:

$$K(x, x_k) = \tanh[K x_k^T x + \theta] \quad (5)$$

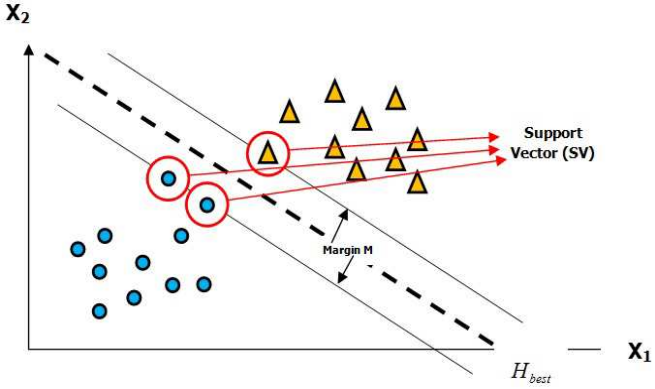

Fig. 2 SVM method [16].

SVR method is estimating a function based on training data mapped from input to real amount. SVR is similar to SVM classification, namely the method of maximizing margins and having kernel tricks on nonlinear data. Training data in SVM regression, namely dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ where x_i is an n-dimensional vector and y is the real number every x_i. So the SVM regression has the procedure of finding the x_i function with y_i using the linear equation as follows:

$$y_i = f(x) = w.x + b \quad (6)$$

Where w = vector weight and b = bias. The two parameters must be determined to get the best function model from input to output data. Furthermore, for nonlinear problems, the function is as follows:

$$y_i = f(x) = w.\phi(x) + b \quad (7)$$

evaluation using the insensitive loss (ε) function as follows:

$$L_\varepsilon(y, f(x)) = \begin{cases} 0 \text{ for } |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon \text{ and vice versa} \end{cases} \quad (8)$$

The variable ε must be regulated and determined because to determine the limit of the difference in output/target with the results of estimates/predictions [11]–[14]. Then the procedure is arranged using the slack variable ξ, ξ* to describe the deviation from the training data outside the zone ε.

In addition to minimizing empirical errors with equation (8), we must also make a minimum ||w|| and will be related to the ability to generalize SVR based on training or learning outcomes. The goal is to get the maximum margin hyperplane. So that the equation to solve the regression problem can be used as the following quadratic optimization problem:

$$L(w, \xi) = \frac{1}{2}\|w\|^2 + c\sum_i (\xi_{2i}, \xi'_{2i}), c \ge 1$$

$$\text{Subject to } \begin{cases} y_i - w * \phi(x_i) - b \le \varepsilon + \xi_i \\ w * \phi(x_i) + b - y_i \le \varepsilon - \xi_i^* \\ \Xi_i, \xi'_i \ge 0 \end{cases} \quad (9)$$

With C it is a penalty coefficient which functions to control optimization between margins and misclassification ξ. The value of this variable C needs to be determined [10]–[14]. Equation (9) is also called the concept of Soft Margin. Then from equation (9) to produce a classification, using the Dual Lagrangian equation [10], as follows:

$$f(x_i) = w.\phi(x_i) + b = \sum_{j=1}^{n} \alpha_i K(x_i, x_j) + b \quad (10)$$

With $K(x_i, x_j)$ is the specified kernel function. Furthermore, the value of the variable σ needs to be determined [10][11][12][13][14]. This variable is very useful for controlling the speed of learning.

G. Genetic Algorithm

Genetic Algorithm (GA) is a computational algorithm that adopts the evolutionary theory of Charles Darwin to find a solution to a problem. GA is used to find a combination of values in variables to produce optimal solution values for a problem that has many possible solutions to problem-solving. GA is one of the optimization algorithms with a metaheuristic approach in the field of Artificial Intelligence. The steps of the GA method are starting from population initialization; calculate the chromosome value; selection, crossover and mutation; output results [7]. There are three key operators in GA, namely Selection, Crossover and Mutation [10]–[14]. Pseudocode for GA is as follows:

Step 1 - Initializing the population

Step 2 - Initialization of chromosomes

Step 3 - Do

- Chromosome evaluation;
- Chromosome selection;
- Chromosome Crossover;
- Chromosome Mutation;
- IF the chromosome criteria have been found THEN stop ELSE next iteration until it reaches the max generation or the chromosome condition of the gene cannot improve.

In selection there are many methods that can be used, including Rank-based Fitness Assignment, Roulette Wheel Selection (RWS), Stochastic Universal Sampling, Tournament Selection, Boltzman Selection and so on. This research will use RWS.

H. Root Mean Square Error

Root Mean Square Error (RMSE) is a value used to measure the deviation between the predicted value of a model and the actual value [2][11][12]. The RMSE formula can be seen as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}} \quad (11)$$

With X_{obs} is the observation value or the actual value and X_{model} is the value of the results of the prediction model. RMSE value that is close to 0 (zero) is the best value [2].

I. Paired T-Test

Paired T-Test or can be called testing T significance is used to measure 2 samples of paired data whether there is a difference or not between before and after the changes in conditions have been made [3] [19] [20]. The purpose of this test is to find out the difference by comparing the mean of 2 (two) paired samples. Paired samples are pairs of samples or a group of samples that experience different treatments or measurements but use the same experimental subjects.

To do testing t (T-Test) can use the formula as follows:

$$t = \frac{D}{\left(\frac{SD}{\sqrt{n}}\right)} \quad (12)$$

Where:

t = t table;

D = the average of the paired sample differences;

SD = standard deviation of the difference in paired samples;

n = data on the sample.

This test will be used to measure the significance of whether GA can change or improve the SVM method of the RMSE produced.

J. Research Method

Optimization in this study is a systematic step in finding the minimum value or maximum value of a function model that is built. Optimization will be carried out using the metaheuristic approach method. This study will use the Genetic Algorithm (GA) for optimal SVM parameter searches. The algorithm for the NTP prediction model used is the SVM algorithm. Next will be added with the GA (Proposed Method) optimization algorithm for SVM parameter optimization. The hyperplane parameter in SVM Regression is ε, σ and C. The results of this optimization are

to improve the accuracy of predictions or decrease the value of RMSE.

In Figure 4 we present the development of the prediction model of the Central Java FTT that had been carried out prior research based on the RMSE results obtained. Ifran and Rizchi's research [2] in 2018 produced the best RMSE in the MLR algorithm of 0,00098. For Rizchi's research [3] in 2019, the best RMSE was 0.00062 with the ANN-PSO algorithm using LOOCV data splitting and window size 6. Until the time this research was conducted, research [3] in 2019 was the best RMSE achievement.

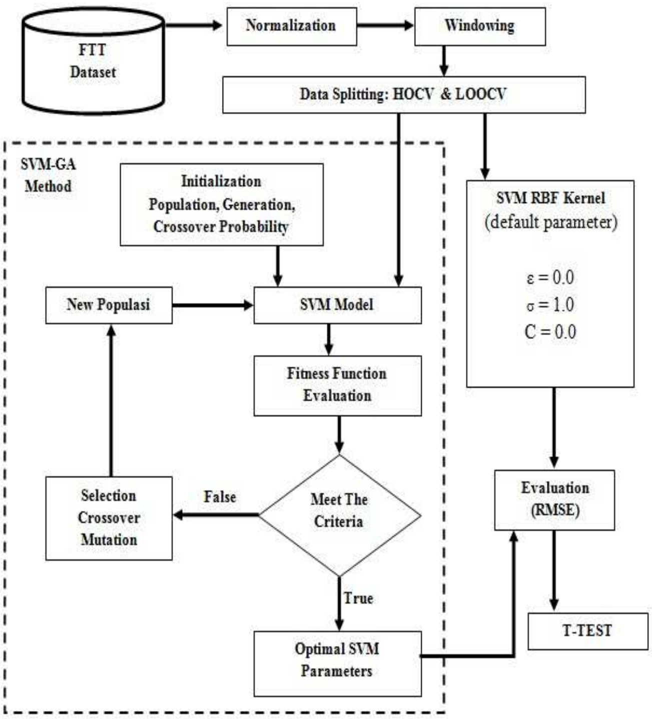

Fig. 3 Proposed method.

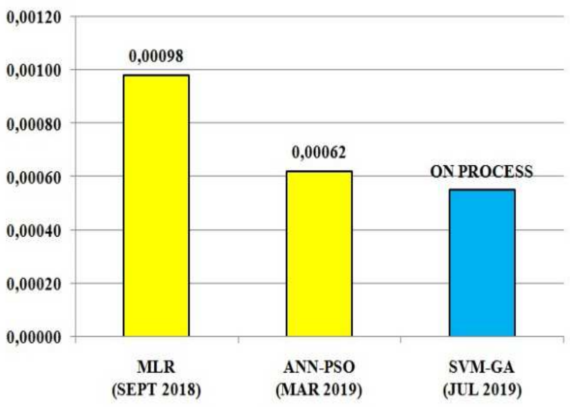

Fig. 4 Increased RMSE prediction model for FTT.

III. Results and Discussion

A. Dataset

The dataset in this study is using the time series FTT dataset in Central Java Province. The data is obtained from official statistical news or called BRS in Indonesia published by the Central Java Provincial Statistics Agency (BPS). FTT data in the BRS are published monthly by the Central Java Provincial BPS. The dataset that will be used is from January 2008 to August 2017 [1]–[3].

B. Decimal Scaling Normalization

Some examples of NTP datasets after decimal scale normalization are provided in table II. Thus the entire NTP dataset is entered in the interval [0,1].

TABLE I
Sample Dataset

Year	JAN	FEB	...	OCT	NOV	DEC
2008	106.69	105.41	...	102.35	101.65	102.7
2009	98.27	98.38	...	99.21	99.81	100.03
...	...	...	...	...	...	...
2015	101.18	101.48	...	101.5	102.07	102.03
2016	101.52	100.53	...	100.15	95.55	99.35
2017	98.98	98.02	...	-	-	-

TABLE II
Example of a Normalized Dataset

Year	Before Normalization			After Normalization		
	Jan	Feb	Mar	Jan	Feb	Mar
2008	106.69	105.41	103.17	0.12	0.11	0.10
2009	98.27	98.38	98.00	0.10	0.10	0.10
2010	100.62	100.23	100.22	0.10	0.10	0.10
2011	102.92	103.58	102.84	0.10	0.10	0.10
2012	106.56	105.42	104.51	0.11	0.11	0.10
2013	106.45	105.70	104.59	0.11	0.11	0.10
2014	106.69	105.41	103.17	0.10	0.10	0.10

C. Windowing

This data windowing method is to retrieve time series data and convert it to a "Cross Sectional" format like the previous discussion.

TABLE III
Sample Dataset in Window Size 5

x1	x2	x3	x4	x5	y
0.11	0.11	0.10	0.10	0.10	0.10
0.11	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10
0.10	0.10	0.10	0.10	0.10	0.10

D. SVM Prediction Model

We provide a research scheme for building SVM prediction models as shown in Figure 5. In Figure 5, the initialization of the hyperplane parameter SVM is $\varepsilon = 0$, $\sigma = 1$ and $C = 0$. As in the previous discussion, the parameters affect the learning outcomes of the SVM prediction model that was built. The parameter ε is the size of the hyperplane line with boundaries, while parameter C is the penalty value for data that crosses the boundary and the parameter σ is the

variable used by the RBF kernel to assist in the learning process of the model. At this stage the default SVM parameters are used with the RBF kernel, namely ε = 0, σ = 1 and C = 0. The results of this stage experiment are provided in the following Table IV.

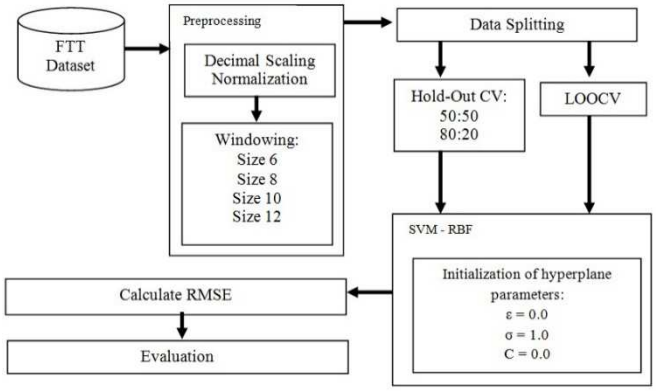

Fig. 5 SVM prediction model scheme.

TABLE IV
SVM PREDICTION RESULTS

No	Window Size				Data Splitting			RMSE
	6	8	10	12	50:50	80:20	LOOCV	
1	√				√			0.00044
2	√					√		0.00044
3		√			√			0.00044
4		√				√		0.00044
5			√		√			0.00044
6			√			√		0.00044
7				√	√			0.00045
8				√		√		0.00045
9	√						√	0.00044
10		√					√	0.00044
11			√				√	0.00044
12				√			√	0.00045

Based on table IV, it can be seen that SVM is not affected by the type of data splitting method used, namely HOCV and LOOCV. As previously discussed, LOOCV is used to increase the amount of training data and test data but will require more time compared to HOCV. LOOCV will be useful to help to learn ANN models if the dataset used is limited, but in this experiment, the learning of SVM models is not affected by the type of data splitting method used in either HOCV or LOOCV.

E. SVM-GA Prediction Model

This stage is an experiment to improve accuracy on SVM prediction models that have been built. SVM prediction model that has been done in the previous experiment has obtained the best RMSE of 0.00044.

As in the previous discussion, the determination of the values in the parameters ε, σ, and C can produce different RMSE. These parameters are the main parameters in the regression SVM (SVR) with the RBF kernel. The SVM method optimization in this study is the search for optimal values for parameters ε, σ and C using the GA optimization method. We provide a research scheme for building SVM-GA prediction models, as shown in Figure 6. SVM-GA experiment results are presented in Table V.

F. Discussion

1) SVM reliability

The implementation of research in SVM reliability results in good predictive models. This is indicated by the RMSE generated by the SVM prediction model as in table VI. SVM shows a more stable prediction model compared to ANN and ANN-PSO prediction models.

In the results of this study, SVM is more reliable than ANN and ANN-PSO. SVM is stable even though it uses two types of data splitting. SVM's RMSE is far superior to ANN and ANN-PSO. So the advantages of SVM are to generalize good data even though limited datasets are indeed proven. More information on the results of this experiment can be seen in Table VI.

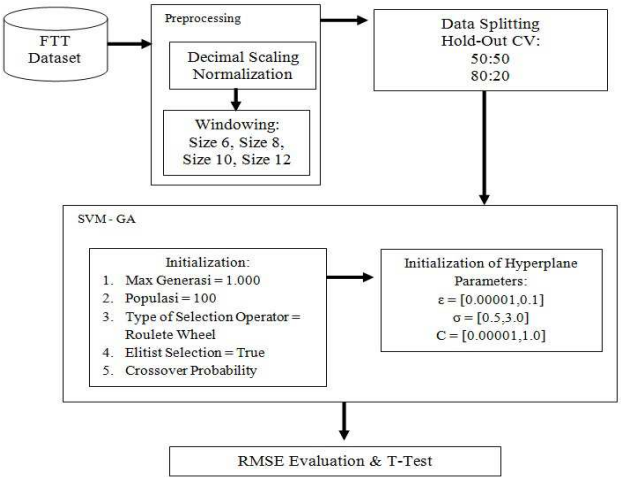

Fig. 6 SVM-GA prediction model scheme.

The window size used by SVM is 6, 8, 10, 12. The SVM prediction model is sequential based on the window size resulting in RMSE of 0.00044 for window size 6; 0.00044 for window size 8; 0.00044 for window size 10; 0.00045 for window size 12. In this experiment, the optimal window size for the SVM prediction model is 6, 8, 10 with RMSE 0,00044. The window size used by ANN [3] is 6, 12, 18. ANN prediction models with data splitting HOCV have RMSE of 0.00096 for window size 6; 0.00094 for window size 12; 0.00087 for window size 18. ANN prediction model with data splitting LOOCV has RMSE of 0.00066 for window size 6; 0.00071 for window size 12; 0.00074 for window size 18. Based on Table VI, it can be seen that the ANN prediction model is affected by the data splitting method. RMSE ANN is better when using LOOCV data splitting. As was done in the study [3], LOOCV was used to cover the weaknesses of ANN generalization to limited datasets. Thus the prediction model for ANN is unstable in this experimental dataset. The window size used by ANN-PSO [3] is 6, 12, 18. ANN-PSO prediction model with CV Hold-Out splitting data has RMSE of 0.00086 for window size 6; 0.00073 for window size 12; 0.00084 for window size 18. ANN-PSO prediction model with data splitting LOOCV has RMSE of 0.00062 for window size 6; 0.00065 for window size 12; 0.00063 for window size 18. From the results of the experiment, the ANN-PSO prediction model is affected by the data splitting method. RMSE ANN-PSO is better when using LOOCV data splitting. Thus, the ANN-

PSO prediction model is unstable in this experimental dataset. From the results of the discussion, it can be concluded that SVM is more reliable than ANN and ANN-PSO for this experiment. Besides that, the RMSE produced by SVM is better than ANN and ANN-PSO.

TABLE V
SVM-GA PREDICTION RESULTS

NO	Window Size				Data Splitting		Crossover Probability									RMSE
	6	8	10	12	50:50	80:20	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
1	√				√		√									0,00040
2	√				√			√								0,00043
3	√				√				√							0,00044
4	√				√					√						0,00043
5	√				√						√					0,00039
6	√				√							√				0,00045
7	√				√								√			0,00039
8	√				√									√		0,00043
9	√				√										√	0,00047
10	√					√	√									0,00042
11	√					√		√								0,00042
12	√					√			√							0,00040
13	√					√				√						0,00038
14	√					√					√					0,00046
15	√					√						√				0,00043
16	√					√							√			0,00037
17	√					√								√		0,00046
18	√					√									√	0,00043
19		√			√		√									0,00039
20		√			√			√								0,00047
21		√			√				√							0,00041
22		√			√					√						0,00045
23		√			√						√					0,00040
24		√			√							√				0,00038
25		√			√								√			0,00043
26		√			√									√		0,00042
27		√			√										√	0,00047
28		√				√	√									0,00041
29		√				√		√								0,00037
30		√				√			√							0,00038
31		√				√				√						0,00042
32		√				√					√					0,00039
33		√				√						√				0,00041
34		√				√							√			0,00043
35		√				√								√		0,00042
36		√				√									√	0,00037
37			√		√		√									0,00047
38			√		√			√								0,00047
39			√		√				√							0,00043
40			√		√					√						0,00045
41			√		√						√					0,00045
42			√		√							√				0,00044
43			√		√								√			0,00051
44			√		√									√		0,00053
45			√		√										√	0,00045
46			√			√	√									0,00045
47			√			√		√								0,00049
48			√			√			√							0,00045
49			√			√				√						0,00048
50			√			√					√					0,00046
51			√			√						√				0,00047
52			√			√							√			0,00047
53			√			√								√		0,00047
54			√			√									√	0,00046
55				√	√		√									0,00042
56				√	√			√								0,00043
57				√	√				√							0,00046
58				√	√					√						0,00042
59				√	√						√					0,00041
60				√	√							√				0,00046
61				√	√								√			0,00044
62				√	√									√		0,00049
63				√	√										√	0,00043
64				√		√	√									0,00044
65				√		√		√								0,00041
66				√		√			√							0,00043
67				√		√				√						0,00049
68				√		√					√					0,00048
69				√		√						√				0,00040
70				√		√							√			0,00047
71				√		√								√		0,00050
72				√		√									√	0,00046

TABLE VI
COMPARISON OF RMSE PREDICTION OF SVM WITH ANN AND ANN-PSO

Algo-rithm	Data Splitting	Window Size				
		6	8	10	12	18
SVM-RBF	HOCV	0,00044	0,00044	0,00044	0,00045	-
	LOOCV	0,00044	0,00044	0,00044	0,00045	-
ANN	HOCV	0,00096	-	-	0,00094	0,00087
	LOOCV	0,00066	-	-	0,00071	0,00074
ANN-PSO	HOCV	0,00086	-	-	0,00073	0,00084
	LOOCV	0,00062	-	-	0,00065	0,00063

2) Results of SVM Optimization

The optimization of the SVM method in this study is to find the optimal value of the parameters ε, σ, C in SVM using the GA search optimization method. The three SVM parameters, namely ε, σ, C are used to construct the SVM prediction model as described in the previous discussion. In SVM reliability research has been obtained for the best RMSE SVM prediction model of 0.00044. Whereas in research [2] in 2018 the SVM prediction model produced RMSE of 0.00112. Both experiments use the same dataset, but the data splitting method and the kernel used are different so as to produce different prediction models. Furthermore, for optimization efforts, it is expected to increase the prediction accuracy of the SVM model that was built.

In table V, the results of the implementation of the SVM-GA study produce good RMSE. Based on the results in table V for comparison purposes, SVM-GA and SVM will determine the criteria that can be compared. The comparison criteria between SVM-GA and SVM will be used, namely window size (6, 8, 10, 12) and HOCV data splitting (50:50, 80:20). For the results of the SVM experiment with LOOCV data splitting, it will not be included for comparison because the SVM-GA did not experiment with the LOOCV data splitting method. Furthermore, based on tables IV and V, we collect some of the best RMSE values from SVM-GA and SVM that can be compared. For a more detailed comparison, we provide in figure 7, 8, 9. From figure 7, 8, 9, it can be seen that the RMSE SVM-GA is better compared to SVM. GA optimization fails during experiment number 6 (figure 9), this happens because of the stages in the GA method that use random generation of values so that there is a bad possibility that will appear.

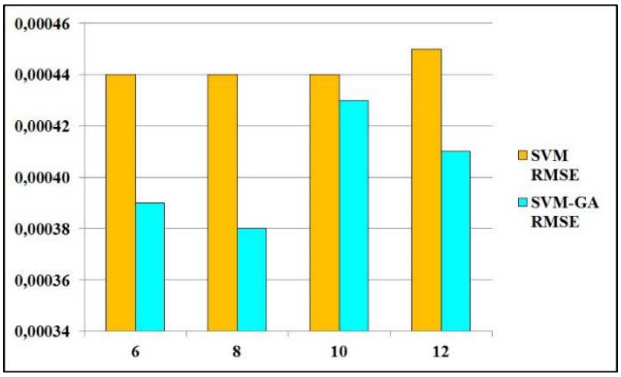

Fig. 7 Comparison of RMSE with HOCV 50:50

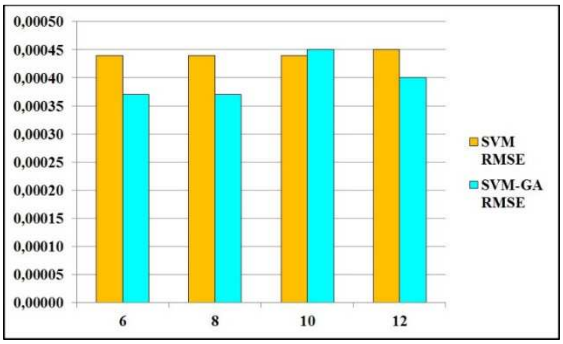

Fig. 8 Comparison of RMSE with HOCV 80:20

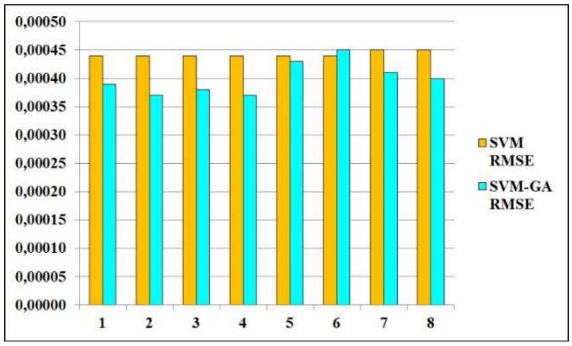

Fig. 9 The best comparison of the overall RMSE SVM-GA and SVM

Based on table figure 7, 8, 9 it can be seen that the best RMSE is in the SVM-GA prediction model, namely with the RMSE value of 0.00037. For the previous study SVM produced RMSE 0.00044 and in the study [2] RMSE SVM was 0.00112.

The results of the RMSE SVM-GA is better than the SVM significantly, it is necessary to test significance with the T-Test. For the statistical test hypothesis, this research is as follows:

- H_0 = RMSE prediction model before optimization and after optimization are the same.
- H_1 = RMSE prediction model before optimization and after optimization is different.

$$t = \frac{D}{\left(\frac{SD}{\sqrt{n}}\right)} = \frac{0{,}0000425}{\left(\frac{0{,}0000287}{\sqrt{8}}\right)} = 4{,}1942$$

The test uses the following rules:

- If t > T_Tabel then H_0 is rejected.
- If t < T_Table then H_0 is accepted.

Based on the above calculations using the significant test rules, t-test is obtained:

t = 4,1942

T_Table = 2.365

So the rule condition t> T_Tabel is true, which is 4.1942> 2.365. So that hypothesis H_0 is rejected, and the H_1 hypothesis is accepted. Based on this, it can be proved significantly by α = 5% that the RMSE prediction model before optimization and after optimization is different from the confidence level used is 95%.

Based on the results of this study, it can be the newest thing related to the prediction of FTT in Central Java Province. In this study, the best RMSE was produced at 0.00037. In the study [2] conducted in 2018 produced the best RMSE of 0.00098. In the study [3] conducted in 2019 produced the best RMSE of 0.00062. A summary of the

development of the RMSE prediction model of the Central Java Province FTT can be seen in Figure 10.

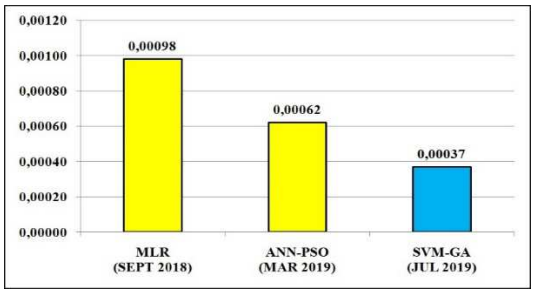

Fig. 10 Increase in RMSE for the current FTT Province Central Java prediction model

IV. CONCLUSIONS

This study seeks reliable predictive models, increases predictive accuracy throgh RMSE evaluation and looks for the optimal window size for SVM and SVM-GA prediction methods. The first conclusion is the SVM prediction model has better robustness than the prediction models of previous studies (ANN and ANN-PSO). Based on the size of the RMSE, SVM outperformed far from ANN and ANN-PSO. The best RMSE obtained by SVM is 0.00044; The best RMSE obtained by ANN is 0.00066, and the best RMSE obtained by ANN-PSO is 0.00062.

Then the second conclusion is the optimization of the SVM method using GA successfully increasing the prediction accuracy of the SVM prediction model without optimization. GA in searching the optimal parameter value ε, σ, C has found optimal values for the prediction of Central Java Province NTP well. The optimal parameter value produced by GA is ε = 0.00001; σ = 1.8832631609640127; C = 0.5004387976587574. The parameter value is used for window size 6 with Hold-Out CV 80:20. The RMSE SVM-GA with the optimal parameter value, produces the best RMSE of 0.00037. The TMS test has been carried out, and the result is a significant change from before and after being optimized.

The third conclusion is the optimal window size for SVM-GA is size 6 and size 8 with the RMSE value obtained at 0.00037 while the optimal window size for SVM is size 6, size 8, and size 10 with the RMSE value obtained at 0.00044.

The fourth conclusion is the development of RMSE for the Central Java FTT prediction model from previous studies also yields better accuracy. In the research [2] in 2018, it produced the best RMSE of 0.00098 using the Multi Linear Regression algorithm method. Continued by research [3] in 2019, it produced the best RMSE of 0.00062 with the ANN-PSO algorithm method. Currently, the most recent Central Java FTT prediction research, namely in 2019, produces the best RMSE of 0.00037 using the SVM-GA algorithm method. Suggestions for further research are using more Central Java Province NTP datasets, comparing SVM-PSO with ANN-PSO or SVM-GA with ANN-GA, looking for other optimization methods and compared with what has been done in the prediction of Central Java Province NTP prediction.

REFERENCES

[1] Website BPS Provinsi Jawa Tengah. https://jateng.bps.go.id.

[2] I.L. Mahargya, R.E. Wahyuni, G. F. Shidik. 2018. "Evaluation Forecasting Method of Farmers Terms of Trade Indonesian Agriculture", 2018 International Seminar on Application for Technology of Information and Communication (iSemantic).

[3] R. E. Wahyuni. 2019. "*Optimasi Metode Neural Network Menggunakan Particle Swarm Optimization untuk Memprediksi Nilai Tukar Petani*". Thesis. Dian Nuswantoro University (UDINUS) Semarang.

[4] Zulyadi, J. Sembiring. 2015. "Design of FTTS Forecasting Model using Markov Chain and P2AMF Framework Case Study: Farmer's Terms of Trade of Smallholders Estate Crops Subsector in Riau". 2015 International Conference on Information Technology Systems and Innovation (ICITSI) Bandung – Bali, November 16 – 19.

[5] Syed Rahat Abbas dan Muhammad Arif, "Electric Load Forecasting Using Support Vector Machines Optimized by Genetic Algorithm". IEEE 2006.

[6] Xiaogang Chen , "Railway Passenger Volume Forecasting Based on Support Vector Machine and Genetic Algorithm". ETP International Conference on Future Computer and Communication, IEEE 2009.

[7] Jianguo Zhou dan Huaitao Liang, "Prediction of the NOx Emissions from Thermal Power Plant Based on Support Vector Machine Optimized by Genetic Algorithm". IEEE 2010.

[8] Gu Jirong, Zhu Mingcang dan Jiang Liuguangyan, "Housing price forecasting based on genetic algorithm and support vector machine". Expert Systems with Applications 38 (2011) 3383–3386, Science Direct (elsevier).

[9] Tao Yerong, Sui Sai, Xie Ke dan Liu Zhe,"Intrusion Detection based on Support Vector Machine using Heuristic Genetic Algorithm". Fourth International Conference on Communication Systems and Network Technologies, IEEE 2014.

[10] A. V. Phan, M. L. Nguyen, L. T. Bui. 2016. "Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems". DOI 10.1007/s10489-016-0843-6. ©Springer Science+Business Media New York 2016

[11] J. Zhang , C.H. Zheng , Y. Xia , B. Wang , P. Chen. 2017. "Optimization Enhanced Genetic Algorithm-Support Vector Regression for the Prediction of Compound Retention Indices in Gas Chromatography". Neurocomputing (2017), doi: 10.1016/j.neucom.2016.11.070.

[12] H. Shafizadeh-Moghadama, A. Tayyebi, M. Ahmadlou, M. R. Delavar, M. Hasanlou. 2017. "Integration of genetic algorithm and multiple kernel support vector regression for modeling urban growth". Computers, Environment and Urban Systems 65 (2017) 28–40.

[13] F. Miao, Y. Wu, Y. Xie, Y. Li. 2017. "Prediction of landslide displacement with step-like behavior based on multialgorithm optimization and a support vector regression model". Landslides DOI 10.1007/s10346-017-0883-y. ©Springer-Verlag GmbH Germany 2017.

[14] I. O. Alade, A. Bagudu, T. A. Oyehan , M. A. A. Rahman , T. A. Saleh, S. O. Olatunji. 2018. "Estimating the Refractive Index of Oxygenated and Deoxygenated Hemoglobin using Genetic Algorithm - Support Vector Machine Approach". Computer Methods and Programs in Biomedicine (2018), doi: 10.1016/j.cmpb.2018.05.029.

[15] J. Han, M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2001, USA.

[16] H.S. Hota, R. Handa, A.K. Shrivas. "Time Series Data Prediction Using Sliding Window Based RBF Neural Network", International Journal of Computational Intelligence Research 2017.

[17] Y. B. Yahmed, A. A. Bakar, A. R. Hamdan, A. Ahmed, S. M. S. Abdullah. "Adaptive sliding window algorithm for weather data segmentation". Journal of Theoretical and Applied Information Technology 80 (2) 2015.

[18] Reitermanova, Z. "Data Splitting". WDS"10 Proceedings of Contributed Paper Part 1, 31-36, 2010.

[19] Montgomery, Douglas C. "Applied Statistics and Probability for Engineers Third Edition". John Wiley & Sons, Inc.2002.

[20] Berger, Paul D & Mike Fritz. "Improving the User Experience through Practical Data Analytics pp.71-89". 2015.