Development of the Support Vector Regression and Genetic Algorithm Hybrid Models for Forecasting of Madura Rice Yield

Devie Rosa Anamisa[1*] [ID], Budi Dwi Satoto[1] [ID], Moch. Kautsar Sophan[1] [ID], Muhammad Yusuf[1] [ID], Yeni Kustiyahningsih[1] [ID], Mohammad Yanuar Hariyawan[2] [ID]

[1] Informatics Engineering Department, University of Trunodjoyo Madura, Bangkalan 69162, Indonesia
[2] Department of Computer Engineering, Institut Teknologi Telkom Surabaya, Surabaya 60231, Indonesia

Corresponding Author Email: devros_gress@trunojoyo.ac.id

ABSTRACT

The combination of population growth and reduced use of agricultural rice land on Madura Island has caused instability in rice production, so more accurate forecasting methods to support food security planning. This study developed a hybrid Support Vector Regression (SVR) model optimized with a Genetic Algorithm (GA) to enhance prediction accuracy. The main contribution of this research lies in the use of GA to optimize SVR parameters, including cost (C), gamma (γ), and epsilon (ε), a strategy that has been empirically proven to overcome the limitations of a single SVR model in finding optimal parameters. A rice production dataset of 420 data points from 2019 to 2024 was used to compare the performance of the standard SVR with that of the hybrid SVR-GA model. The experimental results demonstrated that the hybrid model yielded a significant improvement in accuracy, with a training Mean Absolute Percentage Error (MAPE) of 1.94% and a testing MAPE of 5.35%. It utilized optimal parameters C = 495.52, ε = 0.029, γ = 2.505, radial basis function kernel, mutation probability 0.2, and crossover probability 0.8 in 1000 iterations, resulting in Root Mean Square Error (RMSE) of 171384.7, a Mean Absolute Error (MAE) of 106339.4, and a Coefficient of Determination (R^2) of 0.968. In contrast, a single SVR produced a training MAPE of 3.02% and a testing MAPE of 13.92%, an RMSE of 210546.3, an MAE of 229048.8, and an R^2 of 0.893. These findings prove that GA integration can improve the generalization ability of SVR and significantly reduce forecasting errors, making this hybrid model more reliable for forecasting rice production in Madura.

1. INTRODUCTION

Agriculture is a profession that produces food from crops. Indonesia is a country with a majority of its population as farmers, which is why it is often referred to as an agrarian nation [1]. Agriculture plays a crucial role in national food security, including on Madura Island, where rice is a primary food commodity [2]. However, land use in Madura is increasingly being converted to agricultural purposes, leading to a decline in production due to land shortages. Furthermore, demand for rice as the staple food of the Madurese also increases annually. That solution, the local government has implemented planning strategies, including forecasting. The purpose of this forecasting is to provide information on future production volumes, which can be used as a decision-support system to determine steps to ensure food security for the population.

Forecasting is a process of estimating future values using past data [3]. Forecasting can be used accurately and precisely predict events by minimizing errors. However, agricultural data, especially in dry agro-climate regions like Madura has characterized by nonlinear patterns. Therefore, to predict this agricultural data, the forecasting system requires a forecasting

method. Selecting of the proper forecasting method will yield high accuracy and minimal error, namely the time-series method. Conventional time-series models are often unable to capture these complex relationships. Various time-series-based forecasting methods have been applied in the agricultural sector, including Support Vector Regression (SVR). In this study, the forecasting model used was SVR to predict the Madura rice harvest due to its ability to model nonlinear relationships through kernel functions. SVR is a regression-based prediction model that is a development of the Support Vector Machine (SVM) [4]. Research by Leksakul and Sopadang [5] found that SVR can predict the off-season supply of Thai longans by comparing it with the backpropagation neural network method. The results of this study have shown that SVR with a linear kernel is superior to a neural network. The results of this study show that SVR with a linear kernel performs better than a neural network. In addition, research by Suranart et al. [6] found that SVR can predict gold prices based on time-series data, outperforming other methods such as radial basis function and neural network (RBF-NN). Several studies have shown that SVR performs better than other methods, such as neural networks, RBF-NNs, and traditional statistical models [7]. However, the literature

also confirms that SVR performance is highly sensitive to the selection of key parameters, namely cost (C), gamma (γ), and epsilon (ε). These parameters greatly determine the model's accuracy, so their selection cannot be arbitrary. Most previous studies only report the success of SVR without discussing the reasons for its improved performance or how limitations in parameter selection can lead to overfitting, underfitting, or decreased generalizability. This shows that there are research gaps that have not been resolved. Therefore, several studies have combined methods. Purnama [8] used a combination of SVR and the Autoregressive Integrated Moving Average (ARIMA) to increase profits from stock trading transactions. The results show that the Mean Absolute Percentage Error (MAPE) value for the ARIMA-SVR hybrid model with the RBF kernel is 4.001% lower than that of the ARIMA model alone. In research by Umiyati et al. [9], comparisons were also made between hybrid Particle Swarm Optimization (PSO) and SVR, and improved PSO-SVR, for coal price forecasting. The results showed that the PSO-SVR method achieved an MAPE of 3.911%, better than the improved IPSO-SVR method, which achieved a MAPE of 3.916%. In 2023, one of the studies conducted by Borrero et al. [10], which produced a smaller Mean Absolute Deviation value for the SVR, namely 3.67, compared to neural networks, was 4.16. In research conducted by Qiao et al. [11] regarding fertility prediction using the PSO-SVR model, the highest accuracy level was SVM-PSO with an RBF kernel of 89%, then SVM at 88%.

In addition, several studies are continuing to develop to achieve a minimum error rate by combining prediction parameter optimization methods, such as Genetic Algorithm (GA), such as the research by Ji et al. [12] to improve the estimation of above ground forest biomass using synthetic aperture radar data, the GA-SVR model is used to improve the accuracy of forest Above Ground Biomass estimation through the identification of optimal synthetic aperture radar features and the simultaneous selection of SVR model parameters. The results of this study show that the proposed GA-SVR model can improve the accuracy of forest Above Ground Biomass estimation with a cross-validation coefficient of 80.21%. This shows that the GA-SVR model has shown better performance compared to the combination of GA models accompanied by grid search for SVR (GA-grid SVR). However, these studies were conducted in regional contexts or data types, with data dynamics characteristics that are not similar to rice production in Madura. No previous studies have specifically applied the SVR-GA approach to rice production data in Madura, which has unique seasonal patterns, extreme weather variability, and relatively dry soil conditions. This research gap is crucial to address, as the high variability of Madura rice production requires predictive models capable not only of learning nonlinear patterns but also of adapting through truly optimal parameter selection. Using GA to optimize SVR parameters has the potential to produce more stable and accurate models than standard SVR, which relies solely on manual parameter selection or grid search.

Based on this background, this study contributes by applying a hybrid SVR-GA model to a critical analysis of the effect of SVR parameter optimization on improving prediction accuracy based on Madura's rice harvest data of characteristics. In addition, a comparative evaluation of the performance between standard SVR and the hybrid SVR-GA model using data from 2019–2024 are also conducted. This study aims to obtain optimal SVR parameters through GA to produce more accurate rice production forecasts [13]. And GA

also has complexity and performance that can later optimise the prediction model [14]. Meanwhile, SVR can overcome nonlinear data problems by using SVR parameters to map low-dimensional nonlinear data and transform it into a higher-dimensional space, so that it can provide accurate predictions [15]. The research results are expected to inform decision-making for local governments and stakeholders in formulating more adaptive food security strategies by providing a more accurate forecasting system that is tailored to Madura's agroclimatic characteristics.

2. BACKGROUND AND RELATED WORK

2.1 Literature review

The literature on time series methods is extensive, and many previous studies have combined forecasting models to systematically and pragmatically predict future events from relevant past data. However, producing a precise and accurate forecasting model requires in-depth analysis. Therefore, research on time series forecasting, such as Box-Jenkins ARIMA neural network [16], exponential smoothing neural network [17], double exponential smoothing and neural network, and ARIMAX [18], has shown that hybrid approaches can improve prediction accuracy. However, these statistical and neural network models are still limited in capturing the nonlinearity and complex dynamics often found in agricultural data, especially in areas with high climate variability. This underscores the need for methods that can handle nonlinear relationships more flexibly.

However, these models are not yet fully capable of identifying the characteristics of time series. Therefore, further literature review is underway. In 2024, a study was conducted on rainfall forecasting in South Bangka Regency using SVR and Seasonal Autoregressive Integrated Moving Average (SARIMA) [19]. In the research, the Root Mean Square Error (RMSE) was used to evaluate both methods [20], where SVR is a popular method due to its ability to model nonlinear relationships via kernel functions. And the results of this study showed the smallest RMSE value of 0.039 with γ parameters of 0.0005, C of 0.0001, and ε of 0.00001. In this context, SVR is widely used for its ability to model nonlinear patterns via kernel functions. Recent studies have shown that SVR can achieve low RMSE values across various time-series applications. However, the literature also underlines that SVR performance is highly dependent on the selection of parameters C, γ, and ε. Without an optimisation mechanism, SVR is prone to producing suboptimal solutions and inconsistent accuracy. Thus, improving SVR performance is closely related to the success of parameter optimization [20]. Without an optimisation mechanism, SVR performance can fluctuate and is prone to suboptimal solutions. To overcome these limitations, a hybrid approach of SVR with optimisation algorithms has been developed. The need for parameter optimisation is the background to the development of a hybrid approach of SVR with optimisation algorithms such as GA. GA offers global search capabilities that can explore the parameter space more widely than traditional approaches, such as grid search. Various studies have reported that the Integration of SVR and GA has proven effective across a range of applications. In 2025, Anshori et al. [21] developed methods to optimise hotel-sector occupancy using SVR and GA models. This model can deliver superior predictive

performance by effectively capturing complex and specific occupancy patterns in the hotel industry in Surabaya. The combined SVR-GA method successfully minimised prediction errors, yielding a very low RMSE of 0.0186. In addition, research conducted by Xia et al. [22] on the prediction of damping of a cantilever beam with particle dampers using the GA-SVR method, where SVR is used to predict the damping ratio of a cantilever beam with particle dampers, and GA is used to select optimal variables to improve the predictive ability of the model. The results of this study show that GA-SVR achieved higher predictive accuracy (96.65%) than Cross Validation-SVR (96.16%). In addition to GA, other optimisation algorithms, such as PSO-SVR [23] and IPSO-SVR, have been applied to commodity price forecasting; however, both exhibit a tendency toward premature convergence and unstable performance in some domains. From an analytical perspective, GA is superior to PSO and deep learning models such as SARIMA [24]. PSO [23] tends to get stuck in local solutions due to its sensitivity to inertial parameters, whereas LSTM [25] requires large datasets and high computational cost, making it less suitable for the relatively small Madura rice production dataset. In contrast, GA [26, 27] offers the flexibility of global search and can explore a wider SVR parameter space, increasing the

chance of finding the optimal parameter combination [21].

Based on the literature review, the SVR-GA model consistently improves accuracy across various domains and is particularly relevant for agricultural data with nonlinear patterns. Therefore, this research employs SVR-GA as the primary method for forecasting Madura rice production. The MAPE approach is used for evaluating accuracy. Predictions with low error rates can serve as a reference for decision-making [28]. Several previous studies have employed various methods to predict crop yields and related agricultural data. However, there is significant variation in application domains, model complexity, and predictive performance (MAPE or RMSE), as seen in Table 1. Although other optimisation alternatives, such as PSO or IPSO, have been reported to tend toward local solutions and unstable performance. While deep learning models, such as LSTM models, require large datasets and substantial computational resources, they are less suitable for the relatively small Madura rice production dataset. Therefore, GA is a more relevant candidate optimisation algorithm for this case. This research gap underscores the need to develop a hybrid SVR-GA model to predict rice production in Madura. In this study, MAPE, RMSE, Mean Absolute Error (MAE), and the Coefficient of Determination (R²) were used to evaluate the accuracy of the predictions.

Table 1. Comparison of agricultural prediction methods in previous research

Studies	Methods	Agricultural Domain	Technical Excellence	Technical Weaknesses	MAPE	RMSE
[19]	SVR	Rainfall forecasting	Good at capturing nonlinear patterns; leveraging kernels.	Very sensitive parameters cost (C), gamma (γ), and epsilon (ε), and performance can fluctuate.	-	0.039
[21]	SVR-GA	Hotel occupancy prediction	GA is capable of global exploration, and SVR is stable on small datasets.	Higher computing time.	-	0.0186
[23]	PSO-SVR	Monthly precipitation forecasts	PSO has fast convergence and efficient optimization.	Easily trapped in a local optimum; unstable results on fluctuating data.	7.84%	0.625
[25]	LSTM	Enhanced multivariate time series analysis	Selecting the appropriate normalization strategy to enhance the precision of the LSTM.	Normalization in analyzing multivariate time series of data preprocessing in deep learning models.	9.17%	0.766
[27]	GA-Holt-winters method	Rice supply prediction	Able to optimize forecasting errors by determining the optimal smoothing coefficient.	Less than optimal for highly nonlinear patterns.	10.34%	0.953

Note: SVR: Support Vector Regression; GA: Genetic Algorithm; LSTM: Long Short-Term Memory, PSO: Particle Swarm Optimization; KNN: K-Nearest Neighbors; MAPE: Mean Absolute Percentage Error; RMSE: Root Mean Square Error

2.2 Data collection

This research uses data on rice harvests in Madura from 2019 to 2024, collected through observations by the Department of Agriculture and Food Security in Pamekasan, Madura. The dataset includes 420 data points with six key features used as variables and one target feature, as shown in Table 2. The six features selected—year, location, land area, planted area, harvested area, and productivity—were chosen because they directly influence rice yields. Initial correlation analysis and data exploration showed that these six variables have a significant nonlinear relationship with the target variable (yield). This relationship cannot be accurately modeled by traditional linear techniques, requiring a model like SVR-GA that can capture nonlinear patterns. Basic variables such as weather (rainfall, temperature), soil characteristics, or satellite imagery of Normalized Difference Vegetation Index are excluded from the rice prediction dataset

because they are only available at a few Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG) stations. Additionally, soil fertility data are typically updated infrequently, every two to five years, and are not available at the necessary spatial resolution.

Rice yields in Madura exhibit significant annual variation, as shown in Figure 1. Analysis of this dataset shows that these agricultural variables have a strong nonlinear relationship with harvest yield. The complexity of this relationship requires a prediction method capable of handling nonlinear patterns and operating with a relatively small sample size. Therefore, this study utilises the SVR model with GA optimisation. However, before that, the study performed data preprocessing to improve data quality, thereby increasing the accuracy of subsequent predictions. This step is crucial to ensure input quality, given that the SVR model is sensitive to data scale and distribution. This ensures that the resulting predictive analysis achieves higher accuracy and greater stability. Furthermore, given the

characteristics of the dataset in this study, applying the SVR-GA model is more appropriate than other methods, such as PSO-SVR or Artificial Neural Networks, including multi-layer perceptrons. This is because the multi-layer perceptron model requires large datasets and long sequences to effectively learn temporal patterns, thereby producing low prediction stability. Moreover, the selection of the GA model as a parameter optimiser is also due to its greater stability in global optimisation. GA employs a crossover-mutation mechanism that more effectively explores the solution space, thereby avoiding local traps. In addition, given the fluctuating and nonlinear nature of agricultural data, GA can yield more consistent parameter-tuning performance [29]. In SVR, the RBF kernel, when properly optimised by a GA, can model the relationship effectively.

Table 2. Sample dataset harvesting of rice plant

No.	Year	Area (Ha)	Planted Area (Ha)	Harvested Area (Ha)	Productivity (Kw/Ha)	Production Result (Ton)
1	2019	1411	1719	1657	73	11148
2	2019	1458	1309	1201	55	11868
3	2019	1007	1570	1423	71	11988
4	2019	497	960	932	34	11534
5	2019	1885	1402	1294	61	11769
6	2019	1418	1251	1946	59	8122
7	2019	1406	1238	1192	57	11817
8	2019	1801	1731	1554	58	11451
9	2019	1691	1037	1920	53	7815
10	2019	1411	1520	3425	60	11275
...	....	...	....	....	...	....
341	2023	1213	1258	1843	46.65	5032
342	2023	1093	1485	1187	56.94	9752
343	2023	1562	1280	1096	61.13	7460
344	2023	887	1696	649	46.26	3006
345	2023	1790	1008	1676	49.65	2072
346	2023	1351	1827	1374	57.99	388
347	2023	1274	1240	11480	60.93	7558
348	2023	1352	1723	1345	54.23	7128
349	2023	1308	1228	1953	5.86	2319
350	2023	1217	1807	1620	56.22	7425
...	...	....	....	....	...	...
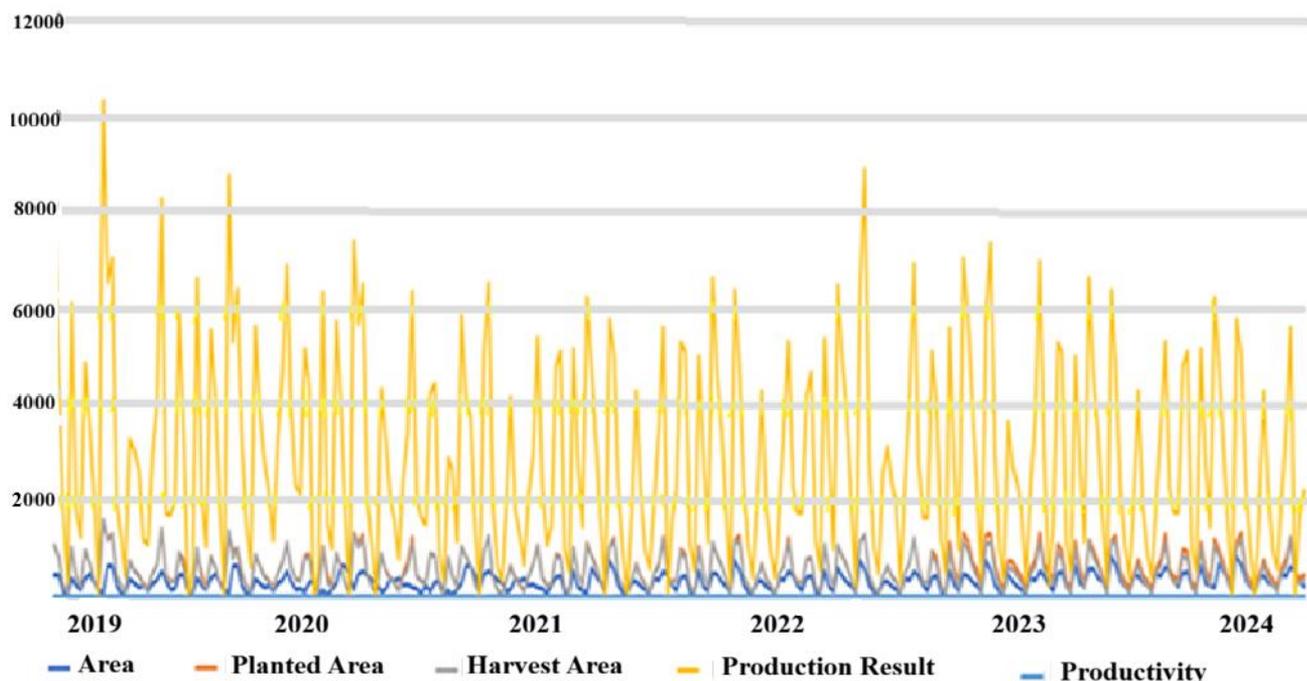
Figure 1. Data on Madura rice yields

2.3 Data preprocessing

Data preprocessing is a critical step in dataset preparation, enabling the production of high-quality data that improves the performance of predictive models. There are several stages in the data preprocessing process, including identifying null or empty values, replacing each column with the column mean (setting missing values to 0), removing outliers, and deleting temporary columns not used for prediction [30]. In this study, data preprocessing is performed not only as a technical step

but also as an analytical step to reduce noise, address missing values, and scale the data to suit the model's characteristics [31]. Data cleaning is the process of identifying missing data, invalid zeros, and duplicate entries. When empty or meaningless zero values are found, replacement is performed using the column average value (mean imputation). Analytically, this method was chosen because it preserves the underlying distribution of the data without reducing the sample size, thereby minimising information loss in relatively small datasets, such as rice harvest data with 420 observations. Data cleaning results will affect the information generated by data mining techniques, as the processed data will be reduced in quantity and complexity. Meanwhile, outlier data are identified as extreme values that fall far outside the normal range. The presence of outliers is known to cause significant bias in nonlinear regression models, including SVR. Therefore, outliers are removed or revised using the Interquartile Range method or Z-score analysis. Analytically, removing outliers improves model stability by reducing variance and avoiding the formation of hyperplanes that deviate from the majority data pattern [32]. The presence of outliers will disrupt the overall data analysis, as it can bias the conclusions drawn. The next stage is the data transformation process, in which techniques modify data values and types to conform to the analysis requirements. One data transformation technique is data normalisation. This study used the Min-Max normalisation technique. The selection of Min-Max in this research is based on the following analytical considerations:

(a) SVR sensitivity to the range of values, not to the form of distribution. SVR models, especially those with RBF kernels, are more sensitive to feature scaling than to distributional normality. Because RBF operates based on Euclidean distance, equalising the value range is a top priority, and Min-Max provides the most consistent range control (0–1).

(b) Parameter stability in GA optimisation. GA explores SVR parameter combinations within a solution space heavily influenced by the range of input values. Min-Max produces fixed, standardised value boundaries, thereby accelerating GA convergence and reducing the risk of overly broad parameter exploration when using the Z-score method, which has an unconstrained scale.

(c) Feature distributions are not extremely skewed. Initial descriptive analysis of the six features indicates that skewness remains moderate. Therefore, applying a log or Z-score transformation does not significantly improve distributional evenness, whereas Min-Max preserves comparative information across features without substantially altering the curve's shape.

(d) Consistency on a limited dataset size. The dataset of 420 data points is relatively small for a log or Z-score transformation, which is more stable when the sample size is large. Min-Max is safer to use on small datasets because it is less sensitive to fluctuations in the mean or standard deviation. Considering these factors, Min-Max normalisation was chosen as the most appropriate method for this case. The Min-Max normalisation technique can be found using Eq. (1) [33]:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where, x' is the normalized data, x_{min} is the minimum value of the data per column, x_{max} is the maximum value of the data

in each column, x is the actual data. This technique scales all data to a predetermined minimum-maximum range, typically with the minimum set to 0 and the maximum to 1. After the data normalisation process, the final stage of preprocessing is to denormalise the data, which is the opposite of normalisation, namely returning the data that has been normalised in the range of zero to one to its original scale. Data denormalisation helps improve query performance or ignore significant storage anomalies by using Eq. (2) [34]. SVR models, especially when optimised with GA, are very sensitive to input quality, so proper preprocessing has a direct impact on the final accuracy of the model.

$$x_i = x_{scaled}(x_{max} - x_{min}) + x_{min} \qquad (2)$$

The application of Min-Max yields a homogeneous range of values, maintains the stability of the SVR distance calculation, and enables the GA optimisation process to reach the optimal point more efficiently. This combination directly improves the accuracy and consistency of crop-yield predictions.

2.4 Analysis Support Vector Regression model

Prediction is the process of estimating future conditions or values from relevant historical data [35]. With advances in computing technology, modern prediction methods increasingly rely on machine learning approaches, including SVR, LSTM, Random Forest, and hybrid methods such as PSO-SVR and SVR-GA. These methods can model nonlinear relationships commonly observed in agricultural data, which are strongly influenced by weather, soil, seasonality, and cropping patterns. The core concept of SVR is structural risk minimisation, which bounds the generalisation error to prevent overfitting to the training data. SVR maps inputs to a high-dimensional space via kernel functions to better capture nonlinear patterns. SVR seeks to find a linear or nonlinear function that best captures the data pattern while being minimally affected by noise. To do this, SVR defines a tolerance margin (ε) around the prediction line, where minor errors are still considered acceptable. Only data points outside this margin affect the model's shape; such points are called support vectors. This approach helps SVR maintain a balance between accuracy and generalizability, making it suitable for agricultural data that tends to be volatile and nonlinear. SVR models can use various kernel types, such as linear, RBF, and polynomial kernels. Essential parameters in SVR are: C (complexity constant), which controls the model's tolerance for error, ε (epsilon-insensitive loss function), which determines the deviation tolerance margin, and γ, which determines the range of influence of points in the RBF kernel. These parameters significantly impact model performance. Tests were conducted by varying C, ε, and γ to assess how model performance changed with these parameters. These kernel functions are used to select the optimal C parameter for the model. The C parameter is a hyperparameter that controls the extent to which the model tolerates errors in the training data [36]. Testing is performed by varying the C parameter in each kernel to assess how model performance changes with C. In addition, SVR is used to find a function *f(x)* which has deviation ε, the largest of the actual values *y*, for all training data. An SVR with an ε-sensitive loss function produces a regression model, *y = f(x)*, used to predict outputs for the dataset points. The value of the prediction function *f(xᵢ)* is determined by a subset of the training data, referred to as

support vectors. In general, the SVR uses Eq. (3) [37]:

$$f(x) = w'a + b \qquad (3)$$

where, the weighting vector is symbolised by w'. The mapping function x is a and b for preference. In nonlinear regression, a kernel function is applied to map the data to a higher-dimensional space. By adding a vector dimension to the λ variables as the preference value can be reformulated into an Eq. (4). Next, to train the data, use a sequential learning process by forming a Hessian matrix using Eq. (5). Then calculate the error value (E) shown in Eq. (6). And calculate $\delta\alpha_i^*$ and $\delta\alpha_i$ by using Eq. (7). In addition, the value calculation is carried out Lagrange Multipliers by using Eq. (8). Repeating the sequential learning process in Eq. (6) until it reaches the maximum number of iterations or has a stopping condition with Eq. (9). If the data meets the requirements $(\alpha_i^* - \alpha_i)$ is not equal to 0. It can be referred to as a support vector.

$$f(x) = \sum_{j=1}^{i} (\alpha_j^* - \alpha)(K(x_i, x_j) + \lambda^2) \qquad (4)$$

$$R_{ij} = K(x_i, x_j) + \lambda^2 \; untuk \; i, j = 1, 2, \ldots, l \qquad (5)$$

$$E_i = y_i - \sum_{j=1}^{i}(\alpha_j^* - \alpha) \, R_{ij} \qquad (6)$$

$$\delta\alpha_i^* = \min\{\max[\gamma(E_i - \varepsilon), -\alpha_i^*], C - \alpha_i^*\} \qquad (7)$$

$$\delta\alpha_i = \min\{\max[\gamma(E_i - \varepsilon), -\alpha_i], C - \alpha_i\} \\ \alpha_i^* = \alpha_i^* + \delta\alpha_i^* \qquad (8)$$

$$\alpha_i = \alpha_i + \delta\alpha_i \\ \max(|\delta\alpha_i^*|) < \varepsilon \; and \; \max(|\delta\alpha_i|) < \varepsilon \qquad (9)$$

where, α_j^* and α is a symbol of Lagrange multipliers, $K(x_i, x_j)$ for the kernel to be used, and λ represents data deviation, R_{ij} is the Hessian matrix of the ith row of the jth column, y_i is the actual value, γ is the value of Learning Rate, ε is the loss value, and C is the complexity value, α_i^* and α_i are Lagrange Multipliers.

In SVR, several kernels can be used in testing, including [38]:

(a) A linear kernel is used for datasets that are linear with a straight hyperplane. The distance between two vectors is the sum of the products of each pair of input values. The linear kernel is calculated using Eq. (10):

$$K(x_i, x_j) = \sum x_i x_j \qquad (10)$$

(b) The polynomial kernel has a curved or nonlinear hyperplane shape so that it can separate data with more complex shapes. The polynomial kernel is calculated using Eq. (11):

$$K(x_i, x_j) = 1 + \sum (x_i x_j)^m \qquad (11)$$

Shaped kernel similar to a linear hyperplane. This indicates that the degree needs to be determined manually during the training process.
(c) Kernel RBF is a kernel function that can map input space

into a space with infinite dimensions. RBF can be calculated using Eq. (12):

$$K(x_i, x_j) = e\left(-\gamma \sum (x_i x_j)^2\right) \qquad (12)$$

A high γ value in the RBF kernel can cause overfitting. Although SVR is capable of modelling nonlinearity, its performance is very sensitive to the choice of C, γ, and ε. Several studies [19, 20] show that the RMSE of SVR decreases greatly when parameters are optimised. SVR without optimisation often gets stuck in non-ideal local configurations. Prediction accuracy is highly dependent on the optimal parameter combination. Therefore, a global optimisation algorithm, such as the GA or PSO, is needed to select the best fit. In terms of accuracy, this model also exhibits greater consistency in searching the SVR parameter space, making the SVR-GA model combination more effective for nonlinear agricultural patterns. This also occurs in LSTM models that require large datasets; the risk of overfitting is very high in small datasets [39]. Thus, the selection of SVR-GA in this study is analytical, not merely descriptive, because it is supported by evidence of performance and the characteristics of the agricultural dataset used.

2.5 Genetic Algorithm optimization

GA uses a natural analogy based on natural selection [40]. GA consists of several stages: chromosome representation, crossover, mutation, and selection. GA begins by generating alternative solutions (a population). Each solution is represented as an individual or chromosome. Fitness values are used to evaluate prediction results; the fitness can be computed as the minimum or maximum value. The fitness value generated by the SVR in this study was computed after the MAPE was calculated in the GA optimisation. Fitness values are inversely proportional to MAPE values, where the lower the better, while a large fitness value indicates a low MAPE result. The use of SVR with GA optimisation is to minimise MAPE by taking the maximum fitness value as the best result. This is achieved using Eq. (13) not only because GA can explore a broad search space, but also because of several identified relative advantages over other optimisation methods, such as PSO. PSO is relatively fast but more prone to local optimums, especially with nonlinear objective functions such as SVR hyperparameter optimisation. GA is used as an optimisation method to determine the best parameters in SVR, namely C and γ. GA works through the general stages of chromosome representation, selection, crossover, and mutation [10, 41]. Each individual in the population represents a set of SVR parameters, and its fitness is computed from the model's MAPE. Because fitness is inversely proportional to MAPE, optimisation efforts focus on identifying the individual with the highest fitness value. The crossover and mutation processes were described in Eqs. (14) and (15), respectively, process to ensure extensive exploration of the solution space, increasing the chance of finding optimal parameters.

$$fitness = \frac{1}{1 + MAPE} \qquad (13)$$

Crossover is a technique of crossing over, or the process of mating between two parent chromosomes to form a new

chromosome (offspring). Crossover produces offspring by combining the values of the two parents. This combination is selected randomly from the population. For example, P_1 and P_2 are parents that have previously been chosen for crossover, so that offspring (C_1 and C_2) can be generated using Eq. (14), where the variable α is randomly selected with a range of $[0,1]$.

$$C_1 = P_1 + \alpha(P_2 - P_1)$$
$$C_2 = P_2 + \alpha(P_1 - P_2)$$
$$(14)$$

Mutation is a process that creates new individuals by modifying one or more genes within the same individual. The mutation process aims to change genes lost during selection and replace them with genes not present in the initial population. The mutation used in this study is random: it selects one parent from the population at random, as shown in Eq. (15):

$$C_i = x_i + r(max_i - min_i) \quad (15)$$

The final stage is selection. The goal of selection is to determine which individuals will be selected for reproduction and how many offspring each individual will produce. The selection method used in this study is elitist selection. This process ranks all individuals (parents and offspring) by fitness from highest to lowest and passes the highest-fitness individuals into the next generation.

2.6 Evaluation method

In testing, cross-validation is used to assess the accuracy of a model trained on a specific dataset. Cross-validation is a method for selecting a model's parameters by evaluating error on the test data. K-fold cross-validation is a method for estimating a system's average performance by randomly partitioning the data into k subsets and iteratively assessing the model on each subgroup [42, 43]. K-fold cross-validation is performed by dividing the data into n folds. In this study, K-fold cross-validation will be conducted with K values ranging from 2 to 5. Table 3 shows 5-fold cross-validation, with the test data highlighted in yellow and the training data in white. In this paper, K-Fold cross-validation with K = 5 is used because many experimental results indicate that 5-fold is the best choice for obtaining accurate estimates. Accuracy is evaluated by monitoring the evolution of MAPE, RMSE, MAE, and R². The testing technique will focus on the results of the modelled data mining process. The accuracy of the prediction process will be evaluated using the MAPE, as defined in Eq. (16) [44]. The prediction value evaluation results are shown in Table 4.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - y_i'}{y_i} \times 100\% \quad (16)$$

Table 3. K-fold cross-validation for rice yield prediction evaluation

Experiment	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Experiment I	Test	Train	Train	Train	Train
Experiment II	Train	Test	Train	Train	Train
Experiment III	Train	Train	Test	Train	Train
Experiment IV	Train	Train	Train	Test	Train
Experiment V	Train	Train	Train	Train	Test

Table 4. Mean Absolute Percentage Error (MAPE) evaluation results in a prediction

MAPE (%)	Evaluation
MAPE ≤ 10%	High prediction accuracy
10% < MAPE ≤ 20%	Good prediction accuracy
20% < MAPE ≤ 50%	Medium prediction accuracy
MAPE > 50%	Prediction accuracy is inaccurate

RMSE quantifies the difference between the actual and predicted values. This metric gives greater weight to significant errors (outliers) because of the squaring operation, making it sensitive to outliers. The purpose of RMSE is: The smaller the RMSE value (closer to 0), the more accurate the predictive model is. The RMSE equation is shown in Eq. (17) [45].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (17)$$

MAE is the average absolute error between the actual and predicted values. Unlike RMSE, MAE gives equal weight to all errors and is less sensitive to outliers. The purpose of MAE is that the smaller the MAE value (closer to 0), the better the model's performance. The MAE equation is shown in Eq. (18) [46].

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (18)$$

Meanwhile, R² measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. R² provides an understanding of how good the model is overall, not the magnitude of the average error. And the purpose of R² is: The higher the R² value (closer to 1), the better the model is at explaining the data. The R² equation is shown in Eq. (19) [47].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (19)$$

3. MAIN RESULTS

This section explains the prediction workflow for the development of the SVR model and the integration of the SVR and GA models for the Madura rice harvest forecasting system. The parameters C, ε, γ, P_m, and P_c influence the SVR-GA model's forecasting performance, producing the lowest MAPE and results that are close to the actual values.

3.1 System flow and Support Vector Regression model

This section discusses the flow of the Madura rice harvest forecasting system using the SVR method, as shown in Figure 2. This study not only performed preprocessing, including handling missing values, outliers, and normalisation, but also conducted an initial evaluation of the relevance and contribution of each feature in the dataset. Correlation and multicollinearity analyses were performed to assess potential redundancy among six features, such as land area, planted area, and harvested area, which have moderate correlation but are not identical. All three represent distinct phases in the

cultivation process and therefore cannot be removed. Productivity is an indicator of efficiency, not a physical measure of land, and thus has a unique contribution. Year and location provide the temporal and spatial dimensions necessary to capture variation in production across time and space. Therefore, no redundant or functionally duplicated features were found, as each variable contains distinct structural information that contributes to yield. Although SVR is not theoretically a model with inherent importance relative to Random Forest, this study still conducted a feature-importance analysis, and the results showed that productivity, harvested area, and planted area were the three most influential features. Year and location contributed less but still improved model stability by capturing temporal and spatial variation not captured by agronomic features. No feature removal improved accuracy; removing one feature increased MAPE by 0.8%–3%. Therefore, the justification for selecting six features in SVR-GA modelling remains intact, as each captures a distinct aspect of the rice production system. The nonlinear relationship between features and the target is quite strong, so feature reduction could potentially eliminate important information. The SVR model ensures that relevant features remain optimised, while GA adjusts model parameters to address variations in scale and contribution between features. No features increase noise or decrease accuracy in feature sensitivity testing. By ensuring that all features are relevant and not redundant, the SVR-GA modelling process can run stably and produce a regression function that maps complex relationships in the data. Ensuring that features are used optimally provides a strong methodological basis that the model is not only accurate but also informationally efficient.

The preprocessing process involves several stages. The first stage is identifying missing values and outliers. Missing values are problems in which data are incomplete or missing [48]. Therefore, missing values often become a bottleneck because they are crucial data, which can lead to inefficient analysis and decreased accuracy. In addition to missing values, the rice harvest dataset contains outliers that can bias the model and reduce prediction accuracy. Although the SVR model is relatively resistant to outliers compared with other methods, detecting and handling outliers and missing values is crucial for improving model performance. After checking and handling missing values, data normalisation is performed. The purpose of data normalisation is to apply a linear transformation to the original data, thereby balancing the comparison values before and after the process. This can help the SVR achieve faster and more stable convergence during training and ensure that all features have a comparable influence on the model. The results of data normalisation are presented in Table 5. The results of data normalisation indicate that the data falls within a narrower range, improving data organisation and eliminating data anomalies, as shown in the overall data graph in Figure 3. Based on the normalised data, the data appear more structured, have a consistent scale, facilitate analysis, and improve the quality of the results. This is the main objective of this study: to conduct data normalisation, which is performed at the beginning of the process, before entering the forecasting optimisation, to avoid scale differences between attributes that can introduce errors, outliers, and incomplete data, which can affect model performance.

In the data division process, both training and test data are used: the training data are used to train the system, and the test data are used to evaluate the trained system's performance. Training data are used to train an SVR model to learn patterns and relationships between inputs and outputs. After the modelling has completed learning, the resulting model will be used to make predictions on new data. Test data is used to assess the performance of a model designed to handle unseen data, or to evaluate the model's ability to overcome the problems it faces. Data division and initialisation of the SVR model parameters used to test several parameter-pair variations on the best hyperplane are shown in Table 6. Based on the test data, this study conducted a trial by applying three kernels (linear, RBF, and polynomial) across a range of parameter values for each SVR parameter, and by splitting the data for training and testing.

Table 5. Normalisation of Madura rice harvest data

| No.	Area	Planted Area	Harvested Area	Productivity	Production Result
...	...	...	...	...	...
11	0.5655	0.6509	0.6513	1	0.6825
12	0.5723	0.4957	0.4936	0.7534	0.5248
13	0.0736	0.1262	0.1225	0.9726	0.1351
14	0.0138	0.0225	0.0267	0.4657	0.0366
15	0	0.6305	0.6280	0.8356	0.7506
16	0.2775	0.2344	0.2203	0.8082	0.3517
17	0.1313	0.1048	0.1077	0.7808	0.1051
18	0.4773	0.5873	0.5805	0.7945	0.9545
19	0.6059	0.4138	0.4113	0.7260	0.7852
20	0.3698	0.1873	0.1868	0.8219	0.5194
...	...	...	...	...	...
401	0.3924	0.3636	0.3422	0.6390	0.9329
402	0.2306	0.2495	0.2358	0.7800	0.6892
403	0.1538	0.1719	0.1658	0.8375	0.8632
404	0.0563	0.0055	0.0086	0.6337	0.9340
405	0.4757	0.2831	0.2672	0.6802	0.6827
406	0.4123	0.4003	0.3763	0.7944	0.7400
407	0.6902	0.7488	0.7042	0.8347	0.8809
408	0.5570	0.3292	0.3102	0.7429	0.9306
409	0.2616	0.2329	0.2208	0.0803	0.8173
410	0.1040	0.1414	0.2208	0.7701	0.8510
...	...	...	...	...	...
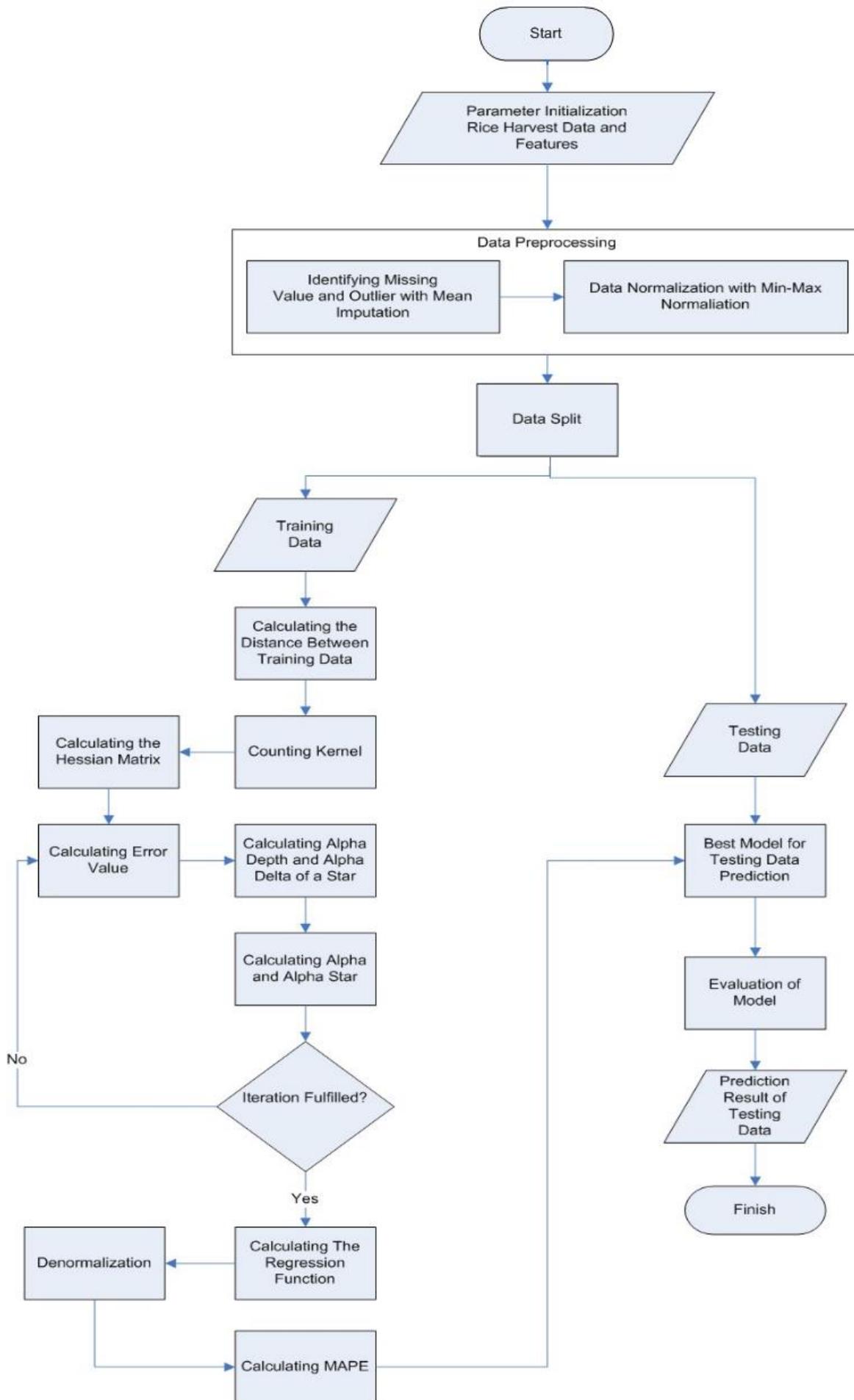
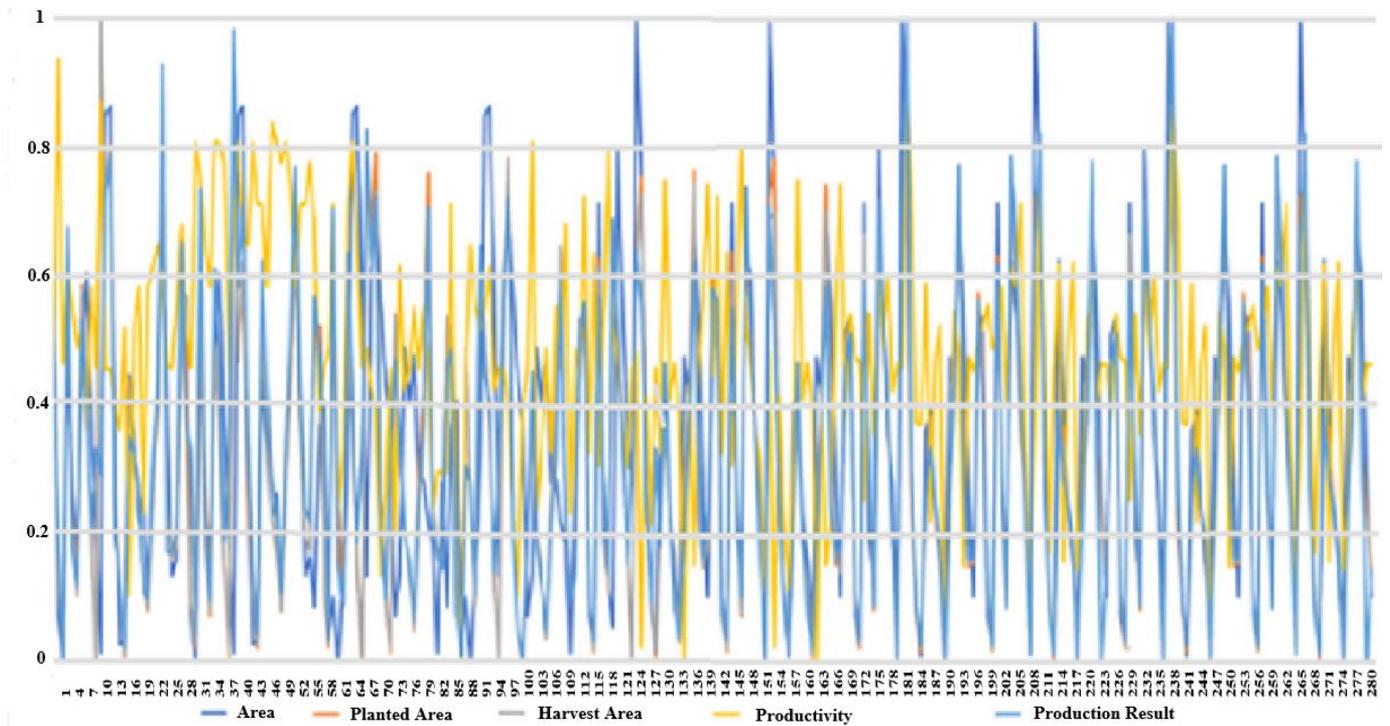**Figure 2.** Support Vector Regression (SVR) model process flow diagram

Figure 3. Min-Max normalisation with Madura rice yields data

Table 6. Data training parameters in the Support Vector Regression (SVR) model

Kernels	Parameters			Range Data Split (%)	
	Cost (C)	Epsilon (ε)	Gamma (γ)	Training Data	Testing Data
Linier	[0.1-1000]	[0.0001-10]	-	50	50
				60	40
				70	30
				80	20
				90	10
Polynomial	[0.1-1i00]	[0.0001-10]	[0.0001-10]	50	50
				60	40
				70	30
				80	20
				90	10
Radial basis function (RBF)	[0.1-1000]	[0.0001-10]	[0.0001-10]	50	50
				60	40
				70	30
				80	20
				90	10

Table 7. Distance calculation results between training data

| No.	21	22	23	24	25	…	248	249	250	251	252	…
…	…	…	…	…	…	…	…	…	…	…	…	…
21	0	0.3803	0.8010	1.3896	0.1540	…	0.8887	0.3375	0.3037	0.9005	0.8662	…
22	0.3803	0	0.3803	0.7711	0.0799	…	0.1397	0.0518	0.1483	0.1156	0.3343	…
23	0.8010	0.7847	0	0.2459	0.7850	…	0.7317	0.4404	1.3709	0.9577	0.2489	…
24	1.3896	0.7711	0.2459	0	0.9933	…	0.4254	0.4872	1.5309	0.6273	0.1008	…
25	0.1540	0.0799	0.7850	0.9933	0	…	0.3958	0.0994	0.0878	0.3743	0.5278	…
..	..	..	…	..	..	…	….	…	…	…	…	…
248	0.8887	0.1397	0.7317	0.4254	0.3958	…	0	0.1479	0.5484	0.0203	0.1490	…
249	0.3375	0.0518	0.4404	0.4872	0.0994	…	0.1479	0	0.3014	0.1866	0.1704	…
250	0.3037	0.1483	1.3709	1.5309	0.0878	…	0.5484	0.3014	0	0.4434	0.9012	…
251	0.9005	0.1156	0.9577	0.6273	0.3743	…	0.0203	0.1866	0.4434	0	0.2752	…
252	0.8662	0.3343	0.2489	0.1008	0.5278	…	0.1490	0.1704	0.9012	0.2752	0	…
…	…	…	…	…	…	…	…	…	…	…	…	…
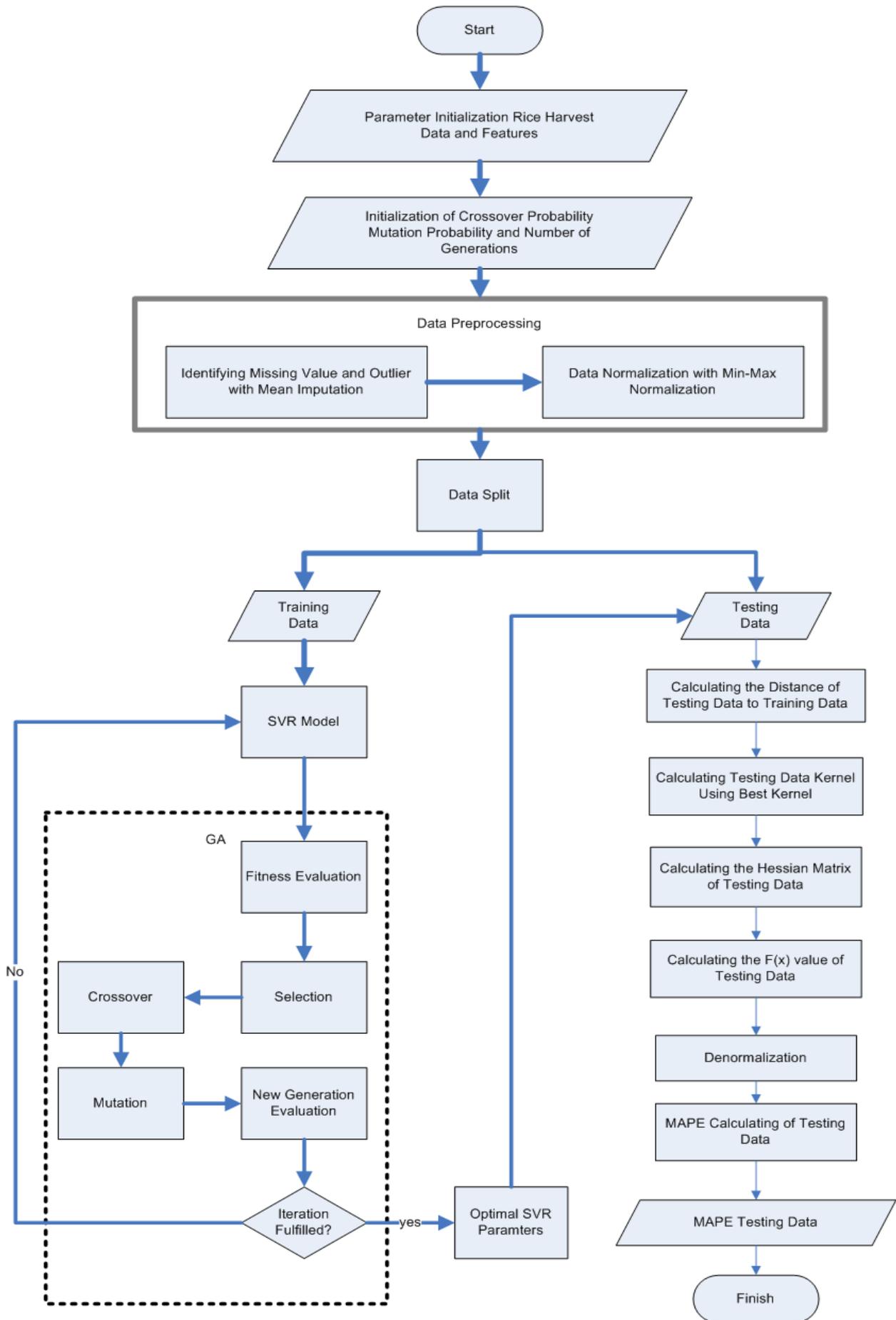
Figure 4. SVR and Genetic Algorithm model hybrid process flow diagram

Table 8. Regression function results on training data

Data	F(x)	F(x) Denormalization
1	0.01744	3956.095
2	0.02775	4305.651
3	0.01999	3750.366
4	0.01802	3609.066
5	0.04255	5364.433
6	0.03192	4604.135
7	0.02196	3891.124
8	0.02997	4464.530
9	0.02399	4036.835
10	0.01873	3659.996
…	…	…
369	0.02874	3464.550
370	0.05530	6364.453
371	0.03358	4704.235
372	0.08406	8197.761
373	0.03873	4804.213
374	0.04706	5564.389
375	0.02074	3791.587
376	0.06131	7364.360
377	0.03227	4704.457
378	0.07680	7197.008

Calculating distances between training data points is the first step in the SVR model. In this study, we tried to use a division of 90% training data = 378 data and 10% testing data = 42 data, C = 100, ε = 0.001, and γ = 0.1, and the kernel used is RBF as a test parameter to produce the distance between training data as seen in Table 7. In the kernel calculation, the σ (sigma) parameter is added. Using a very small σ value makes the forecasts smoother or more similar. If the σ value is too large, the forecasting results will be coarser and less accurate. For example, using σ of 0.5. From this distance calculation, the difference between the training data is calculated and then squared. The resulting output is used to calculate the training Hessian matrix. The purpose of computing the Hessian matrix is to map data using a kernel to quantify the closeness relationships between data points. The number of iterations for repetition in the sequential learning calculation process is set to a maximum, and the algorithm searches for error values across all training data until the optimal α^* and α values are obtained. If they are met, then the iteration is stopped. After completing the sequential learning process, the most appropriate regression function (hyperplane) is sought that can predict continuous values with minimal error.

The regression function in this study was trained to minimise the deviation between the predicted and actual target values across the training data, and to maximise the margin around the regression line, thereby predicting the value of the continuous target variable from the independent variables. If the value is 0, a perfect regression equation is obtained. Data points within the margin are called margin epsilon. Data points that lie within distance ε of the margin are called support vectors. The regression function results are presented in Table 8. To calculate the value of the regression function, the values of α^* and α, as well as the Hessian matrix, are required. Subsequently, a matrix of i rows is formed in the alpha-star and alpha matrices, and then denormalised in the subsequent process. Then, MAPE and fitness metrics are computed on the training data to assess how well the SVR model fits.

$$MAPE = \frac{1}{378}\left(\left(\frac{51148.49-3956.095}{51148.49}\right)\times 100\% + \cdots + \left(\frac{39480.84-3659.996}{39480.84}\right)\times 100\%\right) = 3.02\%$$

3.2 System flow and Support Vector Regression–Genetic Algorithm model

The process of predicting rice yields by optimising SVR parameters using the GA model is shown in Figure 4. The first step of this process is to initialise the GA parameters, namely, forming a chromosome with three genes (parameters C, ε, and γ) obtained from the previous SVR training calculation. After the chromosome is formed, the next stage is to train an SVR using these parameters to obtain fitness results over 1000 iterations. The number of iterations greatly affects the size of the fitness value. This is shown in the study [21]: the more iterations performed, the higher the fitness value, yielding more optimal optimisation results. Furthermore, other parameters, such as population size (pop-size), P_m, and P_c, were initialised to achieve higher fitness. For example, in this experiment, a population size of 20 was used, P_m was 0.2, and P_c was 0.8. Sun and Ren [49] reported that using a Pc value greater than the Pm value can yield a higher fitness value. Therefore, this study has resulted in the formation of an initial population, as shown in Table 9, which generated five initial individuals and populations and initialised the probability of crossover and n mutations. Each chromosome consists of three genes generated from the SVR process, namely C, ε, and γ, on the training data and weighted by evaluating the fitness function to determine performance according to predetermined parameters.

Table 9. Initial population using a pop-size five individuals

Individual	C	ε	γ	MAPE (%)	Fitness
1	0.1	1	0.0001	0.8012	0.5551
2	1	0.1	0.001	0.7435	0.5735
3	10	0.01	0.01	0.0613	0.9422
4	**100**	**0.001**	**0.1**	**0.0302**	**0.0007**
5	1000	0.0001	10	0.6451	0.6078

Note: C: cost, γ: gamma, and ε: epsilon.

Table 10. Crossover results after the first generation crossing

Individual	γ	C	ε	Fitness
1	5.00005	500.005	0.05000	0.769230769
2	0.00550	0.550	0.00450	0.408163265
3	2.50500	5.500	0.02900	0.980969198
4	2.50005	5.005	0.09359	0.636942675
5	5.00050	500.050	0.05050	0.766283525

Table 11. The first-generation selection probability

No	Selection Probability	Cumulative (%)
P1	0.01307	13.072
P2	0.19245	19.24
P3	0.19909	19.90
P4	0.26648	26.64
P5	0.211223	21.12
Total	1	100

The next stage is the crossover process, in which this study uses a one-point crossover to exchange gene values at a single parameter gene point. For example, in the first generation that forms a new individual, parents are used in individuals 1 and 5; the results are shown in Table 10. This process combines the genes of both parents to produce offspring. Each individual becomes a parent twice, all at random. Therefore, before conducting a crossover in this study, it is necessary to determine the parents through a selection process. Selection in

this study uses the Roulette Wheel method, which computes the probability as the fitness value divided by the total fitness of each chromosome. These probability values are cumulative, summing to 100%, and are used to select individuals who will become parents, as shown in Table 11. In Table 11, the probability values are sorted by each individual's fitness. The next step is the mutation process; however, random values were previously generated for each gene, as in Table 12. In Table 12, several random values did not undergo mutation because their mutation probabilities were lower than the random values drawn from the population. For instance, the random value C = 0.0708 is smaller than P_m = 0.2.

After carrying out the crossover process, the final step is the mutation process. The mutation in this study was carried out to prevent the results from being trapped in local optimal conditions. The results of the mutation process can be seen in Table 13. Through several GA processes that have been carried out, the best parameters obtained from the training data are shown in Table 14. Based on Table 14, to measure the performance of the testing data, the optimal parameters obtained were the values of C, ε, and γ with the highest fitness values in the previous GA process. The new population was

obtained by selecting the best parameters, which resulted in the highest fitness values, from both the initial population process and the mutation process for each individual. Specifically, γ = 2.505, C = 495.52, and ε = 0.029 yielded a fitness value of 0.98096, with an MAPE of 1.94%.

Table 12. The first-generation random mutation value

γ	C	ε
0.763259775	**0.070888228**	0.519890166
0.225610351	0.834148239	0.878851746
0.790416984	**0.150934410**	0.498964396
0.302140710	0.429398869	**0.015626096**
0.475433879	0.622341917	0.847123166

Table 13. The first-generation mutation results

Individual	γ	C	ε	Fitness
1	5.00005	**928.65**	0.05000	0.76923076
2	0.00550	0.550	0.00450	0.40816326
3	2.50500	**495.52**	0.02900	0.98096919
4	**4.96000**	5.005	0.09359	**0.63694267**
5	5.00050	500.050	0.05050	0.76628352

Table 14. List of the best individuals in the first generation

Individual	γ	C	ε	Fitness	MAPE (%)
1	**2.50500**	**495.52**	**0.02900**	**0.980969190**	**1.94**
2	2.50005	5.005	0.02501	0.665201592	5.03
3	2.505	5.5	0.029	0.649834282	5.38
4	5	10	0.05	0.569230769	7.56
5	5.00005	500.005	0.05001	0.569230769	7.56

4. DISCUSSION

This section presents several test analyses of the performance of the SVR and GA-optimised SVR models in predicting rice yields. The analysis includes an evaluation of accuracy, such as MAPE, RMSE, MAE, and R², the effect of

kernel variations, and an interpretation of computational time. Furthermore, this section reviews the model's generalizability, based on reviewer feedback, to assess whether the developed model can be applied to other regions with different characteristics.

Table 15. Evaluation results with the Support Vector Regression model

| K-Fold Validation	Kernel	MAPE (%)	RMSE	MAE	R²	Time (s)	Optimal Parameters		
							C	γ	ε
	Linier	14.94	221593,7	231264.1	0.799	60	10	-	0.001
5	RBF	13.92	210546.3	229048.8	0.893	38	100	0.1	0.01
	Polynomial	21.11	297533.1	287157.2	0.538	60	1000	10	0.1
	Linier	16.31	250943.7	245236.6	0.765	60	10	-	0.001
4	RBF	15.73	241632.8	238324.7	0.743	30	100	0.1	0.01
	Polynomial	21.48	292521.9	284415.8	0.521	35	1000	10	0.1
	Linier	16.84	253814.3	241593.4	0.767	56	10	-	0.001
3	RBF	16.55	254995.2	242682.1	0.761	29	100	0.1	0.01
	Polynomial	20.86	285786.1	273771.2	0.623	56	1000	10	0.1
	Linier	15.09	246677.5	234860.4	0.725	20	10	-	0.001
2	RBF	14.79	227168.4	215959.3	0.799	17	100	0.1	0.01
	Polynomial	18.26	270259.6	266045.5	0.692	28	1000	10	0.1
	Linier	15.08	248346.3	237134.4	0.748	24	10	-	0.001
1	RBF	14.93	229520.2	219213.1	0.791	12	100	0.1	0.01
	Polynomial	17.97	261431.4	258322.5	0.702	26	1000	10	0.1

4.1 Performance Support Vector Regression model for prediction analysis without Genetic Algorithm optimization

In this test scenario, the SVR model is tested using three types of kernels, such as linear, RBF, and polynomial, to

determine the performance characteristics of each kernel without parameter optimisation. The MAPE, RMSE, MAE, and R² results on the test data with K-Fold values ranging from 1 to 5 are shown in Table 15. Based on these results, the Linear and RBF kernels yield relatively low MAPE values and differ little. This indicates that the two kernels capture the

relationship between the input and output variables quite well. The best MAPE was achieved with the RBF kernel (MAPE = 13.92%, RMSE = 210546.3, MAE = 229048.8, R^2 = 0.893). An R^2 value close to 0.9 indicates that the RBF kernel explains most of the variation in the data. The performance of the linear kernel is competitive, but remains slightly below that of the RBF kernel. This condition is consistent with the characteristics of SVR, where the RBF kernel is typically more flexible in modelling non-linear relationships and can therefore yield lower errors than the linear kernel, which models only linear relationships. In contrast, the polynomial kernel performs worst, with a substantially higher MAPE than the other two kernels. This pattern indicates that the complexity of the polynomial kernel does not align with the characteristics of rice yield data, which does not exhibit a high degree of nonlinearity. Increasing the percentage of test data increases MAPE for all kernels, particularly the polynomial kernel. This indicates that the model's generalisation ability decreases as the training data size decreases, particularly for high-complexity models.

Regarding computation time, the results in Figure 5 show that the RBF kernel requires the shortest time in the scenario without GA. However, in this research, computation time serves only as a supporting metric. The primary focus remains on prediction accuracy, and the results confirm that the RBF kernel is the best choice overall.
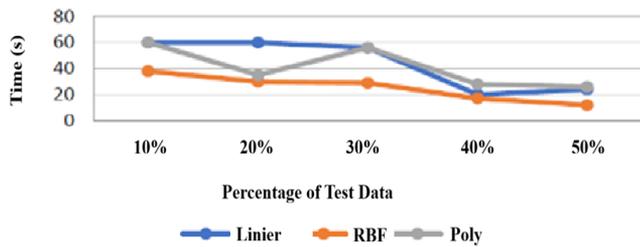
Figure 5. Comparison of computation time for rice yield testing data percentage with the Support Vector Regression model

4.2 Performance Support Vector Regression model for prediction analysis using Genetic Algorithm optimization

In the test scenario with the SVR model and parameter optimisation using GA for predicting Madura rice yields, MAPE results were compared across several kernels (linear, RBF, and polynomial) and test data distributions using K-Fold cross-validation with folds up to 5. The results of this scenario are shown in Table 16. Table 16 indicates that the RBF kernel has displayed excellent model performance, as evidenced by its MAPE graph line being below 10%, specifically 5.35%, RMSE = 171384.7, MAE = 106339.4 and R^2 = 0.968, this shows that the accuracy of a resulting value is shown, the smaller the MAPE value, the smaller the RMSE and MAE, while the R-Squared getting bigger and closer to 1. The Linear kernel also experienced performance improvements after optimisation, but was still unable to surpass the RBF kernel. This confirms that although GA can find optimal parameters, the kernel function's shape remains a major factor in the model's ability to capture data patterns.

On the other hand, the polynomial kernel produces the highest MAPE, indicating that the polynomial kernel model performs the worst among the RBF and linear kernels in the prediction model. Furthermore, the higher the percentage of test data used, the higher the model's MAPE, indicating that increasing the test data rate reduces model performance. This emphasises that speed metrics cannot substitute for prediction accuracy. Given the high MAPE, the polynomial kernel is not recommended in this case. Meanwhile, the computational time for each test kernel and the percentage of test data are shown in Figure 6. As shown in Figure 6, the polynomial kernel requires less computational time than the RBF and linear kernels. The computation-time graph for the polynomial kernel consistently attains its lowest value for each percentage of the test data used. Although the polynomial kernel requires less computation time, the resulting MAPE is the worst among the RBF and linear kernels.

Therefore, kernel selection must not only consider computational speed but also stability and prediction accuracy. Overall, the scenarios with GA reinforce the conclusion that the RBF kernel is the most optimal choice, the linear kernel is good but not superior, and the polynomial kernel provides the lowest performance despite having the fastest computation time. Furthermore, increasing the percentage of test data still produces the same pattern: an increase in MAPE for all kernels. This confirms that a smaller amount of training data negatively impacts the model's ability to recognise patterns and generalise.

Table 16. Evaluation results with the Support Vector Regression–Genetic Algorithm model

K-Fold Validation	Kernel	MAPE (%)	RMSE	MAE	R^2	Time (seconds)	Optimal Parameters		
							C	γ	ε
5	Linier	5.53	233405.3	646880.5	0.913	40	250	-	0.001
	RBF	5.35	171384.7	106339.4	0.968	19	500	2.5	0.02
	Polynomial	24.47	310546.3	229048.8	0.893	21	1000	10	0.4
4	Linier	6.42	264710.6	225343.8	0.900	24	250	-	0.001
	RBF	7.34	250558.8	181100.9	0.928	22	500	2.5	0.02
	Polynomial	31.35	233752.5	171051.7	0.920	15	1000	10	0.4
3	Linier	7.81	304253.4	227620.8	0.897	17	250	-	0.001
	RBF	8.25	311538.8	225343.8	0.909	17	500	2.5	0.02
	Polynomial	27.30	329316.7	240043.4	0.784	12	1000	10	0.4
2	Linier	7.57	251891.8	182225.4	0.945	12	250	-	0.001
	RBF	7.99	306262.4	229731.8	0.899	13	500	2.5	0.02
	Polynomial	24.10	351856.5	857120.9	0.721	9	1000	10	0.4
1	Linier	8.36	394449.7	124338.2	0.786	7	250	-	0.001
	RBF	8.50	310546.3	229048.8	0.893	10	500	2.5	0.02
	Polynomial	22.56	351110.3	105307.7	0.754	7	1000	10	0.4
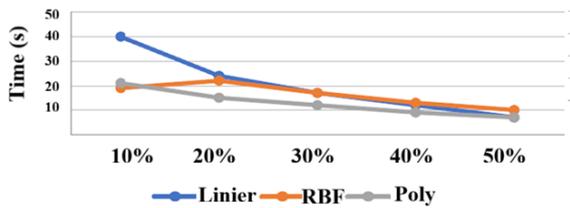
Figure 6. Comparison of computation time for rice yield testing data percentage with the SVR-GA model

4.3 Potential generalization of the model to other regions

Generalisation is essential for assessing the extent to which the SVR and SVR-GA models can be applied to regions outside Madura. The results show that the SVR-GA model, particularly with the RBF kernel, can learn the relationship between climate/environmental variables and crop yields with high accuracy. However, model generalisation to other regions is highly dependent on the following factors:

i. Data availability and similarity

The model was trained on rice-yield data from Madura, which has distinct characteristics, such as soil types and cropping patterns. If other regions have similar characteristics, the model has the potential to perform well. Conversely, areas with significantly different agro-climatic conditions require retraining or parameter adjustments.

ii. Input feature diversity

The more comprehensive and relevant the features used (e.g., rainfall intensity, temperature, soil moisture content), the greater the model's potential for generalisation. If the available features differ in other regions, model performance may decline.

iii. Additional testing on cross-regional data

To ensure generalisation, the model should ideally be tested on data from other regions. This study did not include cross-regional evaluation. Therefore, generalisations remain potential rather than conclusive.

Therefore, although the model performed very well on the Madura data, its use in other regions requires additional validation, such as retraining with local data or fine-tuning parameters.

4.4 Comparison of actual data and predicted data

Based on tests conducted with the SVR model itself and with SVR and GA optimisation, the lowest MAPE was obtained for the SVR-GA model during K-Fold Validation with K = 5 and the RBF kernel. The graph comparing actual test results with system predictions for the Madura rice harvest is shown in Figure 7, which visualises the closeness between predicted and actual data. This is demonstrated by predictions that closely approximate the actual value, as shown in a two-line graph. The yellow line represents the predicted data in tons, while the green line represents the actual data from the production column in tons. The x-axis represents the data index for each sub-district, and the y-axis represents the production value, indicating a close relationship with a slight difference. This indicates that the test results with the lowest MAPE are 5.35%, whereas the training results have a smaller MAPE of 1.94%. It can be concluded that the SVR and GA models have successfully predicted rice harvests.
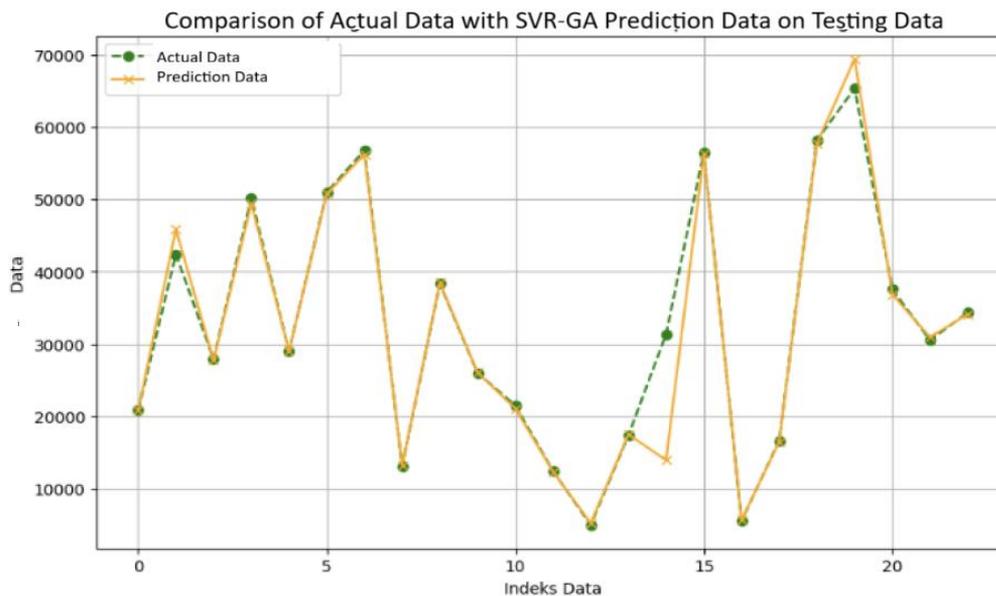
Figure 7. Comparison chart of actual data and predicted data

5. CONCLUSIONS

This study aims to evaluate the performance of SVR models optimised with a GA for predicting rice yields in Madura. Based on experimental results and analysis using various kernels (linear, RBF, and polynomial), several key conclusions can be drawn:

(a) The RBF kernel proved to be the model with the best accuracy, both in the SVR scenario without optimisation and the SVR optimised with GA. In the scenario without GA, the RBF kernel produced a MAPE of 13.92% and an R^2 of 0.893. After GA optimisation, accuracy increased significantly, with MAPE of 5.35% and R^2 of 0.968. This indicates that RBF is capable of capturing nonlinear patterns very well. The test required 38 seconds of computation time

with the RBF kernel, C = 100, γ = 0.1, and ε = 0.01.

(b) The application of GA consistently improved the prediction accuracy of the SVR, especially with the RBF kernel. GA successfully found the optimal parameter combination (C, γ, and ε), resulting in a significantly smaller error rate, with a smaller MAPE value of 1.94% on the training data compared to the test data results of 5.35% at C = 495.52, γ = 0.029, and ε = 2.505, P_m = 0.2 and P_c = 0.8, and a testing computation time of 19 seconds. This improvement confirms that parameter optimisation plays a significant role in improving the performance of the prediction model.

(c) The linear kernel demonstrated stable performance but did not outperform the RBF kernel, both before and after GA optimisation. This kernel can still be used as an alternative for linear data patterns, despite its lower accuracy.

(d) The polynomial kernel provided the lowest performance across all scenarios, even after the GA optimisation process. This indicates that the complexity of the polynomial kernel does not align with the characteristics of rice harvest data and exhibits low generalisation ability.

(e) The percentage of test data significantly impacts model performance. Increasing the proportion of test data (i.e., reducing the training data) increases MAPE across all kernels, indicating that smaller training data reduces the model's ability to recognise patterns and generalise.

(f) Computation time is not a primary factor in model selection, but is still recorded as a supporting parameter. Although some kernels have shorter computation times, accuracy remains a top priority. The model with the fastest computation time does not always provide the best accuracy.

(g) The model's generalisation potential has shown that the SVR model with the RBF kernel optimised by GA has excellent predictive ability on Madura data. However, model generalisation to other regions still depends on similar agroclimatic characteristics, the availability of relevant input features, and data patterns. Therefore, applying this model to other regions requires additional validation, such as retraining with local data or parameter adjustments.

Based on the research results on the application of SVR and GA optimisation for predicting rice harvests in Madura, several recommendations for further research can be made. In addition to GA, other optimisation techniques, such as PSO, differential evolution, grid search, random search, and Bayesian optimisation, can be used to optimise SVR parameters. Comparisons between optimisation methods will provide a more comprehensive picture of the most efficient and practical techniques for crop yield prediction. Furthermore, future research could incorporate additional features such as rainfall intensity, solar radiation, soil fertility, or irrigation patterns. These variables have the potential to improve prediction accuracy and strengthen the representation of agricultural patterns.

REFERENCES

[1] Arifin, T., Amri, S.N., Rahmania, R., Yulius, Ramdhan, M., Chandra, H., Adrianto, L., Bengen, D.G., Kurniawan, F., Kurnia, R. (2023). Forecasting land-use changes due to coastal city development on the peri-urban area in Makassar City, Indonesia. The Egyptian Journal of Remote Sensing and Space Science, 26(1): 197-206. https://doi.org/10.1016/j.ejrs.2023.02.002

[2] Iswahyudi. (2023). Agricultural activities and the Madura Salt Industry in the late 19th century to the 1930s. Britain International of Humanities and Social Sciences (BIoHS) Journal, 5(3): 179-192. https://doi.org/10.33258/biohs.v5i3.979

[3] Chen, K.Y., Wang, C.H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management, 28(1): 215-226. https://doi.org/10.1016/j.tourman.2005.12.018

[4] Hasanah, R.N., Indratama, D., Suyono, H., Shidiq, M., Abdel-Akher, M. (2020). Performance of genetic algorithm-support vector machine (GA-SVM) and autoregressive integrated moving average (ARIMA) in electric load forecasting. Journal FORTEI-JEERI, 1(1): 60-69. https://doi.org/10.46962/forteijeeri.v1i1.8

[5] Leksakul, K., Sopadang, A. (2012). Off-season supply forecasting in Thai longan supply chain with artificial neural network. Chiang Mai University Journal of Natural Sciences (Special Issue on Agricultural & Natural Resources), 11(1): 117-122.

[6] Suranart, K., Kiattisin, S., Leelasantitham, A. (2014). Analysis of comparisons for forecasting gold price using neural network, radial basis function network and support vector regression. In The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE), Chiang Rai, Thailand, pp. 1-5. https://doi.org/10.1109/JICTEE.2014.6804078

[7] Guo, Y., Han, S., Shen, C., Li, Y., Yin, X., Bai, Y. (2018). An adaptive SVR for high-frequency stock price forecasting. IEEE Access, 6: 11397-11404. https://doi.org/10.1109/ACCESS.2018.2806180

[8] Purnama, D.I. (2021). Peramalan Harga Emas Saat Pandemi COVID-19 Menggunakan model hybrid autoregressive integrated moving average-support vector regression. Jambura Journal of Mathematics, 3(1): 52-65. https://doi.org/10.34312/jjom.v3i1.8430

[9] Umiyati, A., Dasari, D., Agustina, F. (2021). Forecasting coal price Index using PSOSVR and IPSOSVR methods. Jurnal EurekaMatika, 9(1): 69-94. https://doi.org/10.17509/jem.v9i2.40064

[10] Borrero, J.D., Mariscal, J. (2023). Elevating univariate time series forecasting: Innovative SVR-empowered nonlinear autoregressive neural networks. Algorithms, 16(9): 423. https://doi.org/10.3390/a16090423

[11] Qiao, G.C., Yang, M.X., Zeng, X.L. (2022). Monthly-scale runoff forecast model based on PSO-SVR. Journal of Physics: Conference Series, 2189: 012016. https://doi.org/10.1088/1742-6596/2189/1/012016

[12] Ji, Y., Xu, K., Zeng, P., Zhang, W. (2021). GA-SVR algorithm for improving forest above ground biomass estimation using SAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 6585-6595. https://doi.org/10.1109/JSTARS.2021.3089151

[13] Zhan, A., Du, F., Chen, Z., Yin, G., Wang, M., Zhang, Y. (2022). A traffic flow forecasting method based on the GA-SVR. Journal of High Speed Networks, 28(2): 97-106. https://doi.org/10.3233/JHS-220682

[14] Chen, J., Chen, H., Huo, Y., Gao, W. (2017). Application of SVR models in stock index forecast based on different parameter search methods. Open Journal of Statistics, 7(2): 194-202. https://doi.org/10.4236/ojs.2017.72015

[15] Shamsudina, H., Yusofa, U.K., Haijiea, Y., Isa, I.S. (2023). An optimizied support vector machine with genetic algorithm for imbalanced data classification. Jurnal Teknologi, 85(4): 67-74. https://doi.org/10.11113/jurnalteknologi.v85.19695

[16] Tsoku, J.T., Metsileng, D., Botlhoko, T. (2024). A hybrid of Box-Jenkins ARIMA model and neural networks for forecasting South African crude oil prices. International Journal of Financial Studies, 12(4): 118. https://doi.org/10.3390/ijfs12040118

[17] Permata, R.P., Muhaimin, A., Hidayati, S. (2024). Rainfall forecasting with an intermittent approach using hybrid exponential smooting neural network. BAREKENG: Journal of Mathematics and Its Applications, 18(1): 457-466. https://doi.org/10.30598/barekengvol18iss1pp0457-0466

[18] Anggraeni, W., Mahananto, F., Sari, A.Q., Zaini, Z., Andri, K.B., Sumaryanto. (2019). Forecasting the price of Indonesia's rice using hybrid artificial neural network and autoregressive integrated average (Hybrid NNs-ARIMAX) with exogenous variables. Procedia Computer Science, 161: 677-686. https://doi.org/10.1016/j.procs.2019.11.171

[19] Wati, P., Adriyansyah, A., Sulistiana, I. (2024). Comparison of rainfall prediction results in South Bangka Regency using support vector regression and SARIMA. CoreID Journal, 2(3): 86-92. https://doi.org/10.60005/coreid.v2i3.87

[20] Mahendra, F., Mutoi Siregar, A., Ahmad Baihaqi, K., Priyatna, B., Setyani, L. (2023). Implementation of linear regression algorithm and support vector regression in building prediction models fish catches of fishermen In Ciparagejaya Village. Edutran Computer Science and Information Technology, 1(1): 42-50. https://doi.org/10.59805/ecsit.v1i1.15

[21] Anshori, M.Y., Herlambang, T., Abu Yaziz, M.F. (2025). Optimizing occupancy of hospitality sector using Support Vector Regression and Genetic Algorithm. Journal of Revenue and Pricing Management, 1-7. https://doi.org/10.1057/s41272-025-00539-4

[22] Xia, Z., Mao, K., Wei, S., Wang, X., Fang, Y., Yang, S. (2017). Application of genetic algorithm-support vector regression model to predict damping of cantilever beam with particle damper. Journal of Low Frequency Noise, Vibration and Active Control, 36(2): 138-147. https://doi.org/10.1177/0263092317711987

[23] Parviz, L., Ghorbanpour, M. (2024). Assimilation of PSO and SVR into an improved ARIMA model for monthly precipitation forecasting. Scientific Reports, 14(1): 12107. https://doi.org/10.1038/s41598-024-63046-3

[24] Xiong, C., Zhang, Y., Wang, W. (2025). SD-LSTM: A dynamic time series model for predicting the coupling coordination of smart agro-rural development in China. Agriculture, 15(14): 1491.

https://doi.org/10.3390/agriculture15141491

[25] Pranolo, A., Setyaputri, F.U., Paramarta, A.K.A.I., Triono, A.P.P., et al. (2024). Enhanced multivariate time series analysis using LSTM: A comparative study of min-max and Z-score normalization techniques. ILKOM Jurnal Ilmiah, 16(2): 210-220. https://doi.org/10.33096/ilkom.v16i2.2333.210-220

[26] Sun, F., Meng, X., Zhang, Y., Wang, Y., Jiang, H., Liu, P. (2023). Agricultural product price forecasting methods: A review. Agriculture, 13(9): 1671. https://doi.org/10.3390/agriculture13091671

[27] Navarro, M.M., Navarro, B.B., Camino, J.L. (2023). Optimizing short-term forecasting of rice stock-commercial during the COVID-19 pandemic using GA-based holt-winters method. In Proceedings of the 6th European Conference on Industrial Engineering and Operations Management Lisbon, Portugal, pp. 1013-1022. https://doi.org/10.46254/EU6.20230295

[28] Ragunath, R., Rathipriya, R. (2025). Forecasting agriculture commodity price trend using novel competitive ensemble regression model. Information Technology and Computer Science, 17(3): 97-105. https://doi.org/10.5815/ijitcs.2025.03.07

[29] Prapcoyo, H., As'ad, M. (2022). The forecasting of monthly inflation in Yogyakarta city uses an exponential smoothing-state space model. International Journal of Economics, Business and Accounting Research (IJEBAR), 6(2): 1144-1152. https://doi.org/10.29040/ijebar.v6i2.4853

[30] Jadhav, A., Pramod, D., Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138

[31] Sharifnia, A.M., Kpormegbey, D.E., Thapa, D.K., Cleary, M. (2025). A primer of data cleaning in quantitative research: Handling missing values and outliers. Journal of Advanced Nursing. https://doi.org/10.1111/jan.16908

[32] Ben-Gal, I. (2009). Outlier Detection. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09823-4_7

[33] Pranolo, A., Setyaputri, F.U., Paramarta, A.K.A.I., Triono, A.P.P., et al. (2024). Enhanced multivariate time series analysis using LSTM: A comparative study of min-max and Z-score normalization techniques. ILKOM Jurnal Ilmiah, 16(2): 210-220. https://doi.org/10.33096/ilkom.v16i2.2333.210-220

[34] Alifiah, N., Kurniasari, D., Amanto, A., Warsono, W. (2023). Prediction of COVID-19 using the Artificial Neural Network (ANN) with K-fold cross-validation. Journal of Information Systems Engineering and Business Intelligence, 9(1): 16-27. https://doi.org/10.20473/jisebi.9.1.16-27

[35] Tsai, P.H., Berleant, D., Segall, R.S., Aboudja, H., Batthula, V.J.R., Duggirala, S., Howell, M. (2023). Quantitative technology forecasting: A review of trend extrapolation methods. International Journal of Innovation and Technology Management, 20(4): 2330002. https://doi.org/10.1142/S0219877023300021

[36] Rafsanjani, Z.A., Nurtiyasari, D., Syahputra, A. (2021). The dynamics of stock price change motion effected by Covid-19 pandemic and the stock price prediction using

multi-layered neural network. International Journal of Computing Science and Applied Mathematics, 7(1): 8. https://doi.org/10.12962/j24775401.v7i1.7023

[37] Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Mathematical and Computer Modelling, 58(3-4): 458-465. https://doi.org/10.1016/j.mcm.2011.11.021

[38] Yulianto, F., Mahmudy, W.F., Soebroto, A.A. (2020). Comparison of regression, support vector regression (SVR), and SVR-particle swarm optimization (PSO) for rainfall forecasting. Journal of Information Technology and Computer Science, 5(3): 235-247. https://doi.org/10.25126/jitecs.20205374

[39] Adhany, P.C., Wulandari, C., Intan, B., Santoso, B. (2025). Prediksi Padi Menggunakan Algoritma long short term memory. Journal of Informatics Management and Information Technology, 5(2): 120-127.

[40] Yuan, F.C. (2012). Parameters optimization using genetic algorithms in support vector regression for sales volume forecasting. Applied Mathematics, 3(10): 1480-1486. https://doi.org/10.4236/am.2012.330207

[41] Wang, Q., Han, Y., Zhao, L., Li, W. (2023). Water abundance evaluation of aquifer using GA-SVR-BP: A case study in the Hongliulin Coal Mine, China. Water, 15(18): 3204. https://doi.org/10.3390/w15183204

[42] Teodorescu, V., Obreja Brașoveanu, L. (2025). Assessing the validity of k-fold cross-validation for model selection: Evidence from bankruptcy prediction using random forest and XGBoost. Computation, 13(5): 127. https://doi.org/10.3390/computation13050127

[43] Aprihartha, M.A., Idham, I. (2024). Optimization of classification algorithms performance with k-fold cross

validation. Eigen Mathematics Journal, 7(2): 61-66. https://doi.org/10.29303/emj.v7i2.212

[44] Airlangga, G. (2024). A comparative analysis of machine learning models for predicting student performance: Evaluating the impact of stacking and traditional methods. Brilliance: Research of Artificial Intelligence, 4(2): 491-499. https://doi.org/10.47709/brilliance.v4i2.4669

[45] Ningsih, I.R., Faqih, A., Rinaldi, A.R. (2025). House price prediction analysis using a comparison of machine learning algorithms in the Jabodetabek Area. Journal of Artificial Intelligence and Engineering Applications, 4(2): 687-694. https://doi.org/10.59934/jaiea.v4i2.733

[46] Zhao, J., Guo, H., Han, M., Tang, H., Li, X. (2019). Gaussian process regression for prediction of sulfate content in lakes of China. Journal of Engineering and Technological Sciences, 51(2): 198-215. https://doi.org/10.5614/j.eng.technol.sci.2019.51.2.4

[47] Steurer, M., Hill, R.J., Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38(2): 99-129. https://doi.org/10.1080/09599916.2020.1858937

[48] Köse, T., Özgür, S., Coşgun, E., Keskinoğlu, A., Keskinoğlu, P. (2020). Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. BioMed Research International. https://doi.org/10.1155/2020/1895076

[49] Sun, W., Ren, C. (2021). Optimized Extreme Learning Machine (ELM) based on Genetic Algorithm (GA) to predict carbon price under the influence of multiple factors. Preprint. https://doi.org/10.21203/rs.3.rs-702953/v1