گـۆڤـاری زانـکۆی ڕاپــەڕیــن
Journal of University of Raparin.
مجلة جامعة رابرين

E-ISSN: 2522 – 7130 P-ISSN: 2410 – 1036

DOI: 10.26750/87w4cn57

Date Received: 12/01/2025
Date Accepted: 03/06/2025
Date Published: 29/08/2025

Optimization of Support Vector Regression for Improved Prediction Accuracy in Cross-Sectional Data Using Genetic Algorithms

Hindreen Abdullah Taher [1,4]- Aras Jalal Mhamad[2] - Ahsan Abdalkhaliq Taha[3]

hindreen.taher@univsul.edu.iq - aras.mhamad@univsul.edu.iq - ahsan.taha@univsul.edu.iq

[1] Department of Information Technology, College of Commerce, University of Sulaimani, Sulaymaniyah, Kurdistan Region, Iraq.

[2, 3] Department of Statistics and Informatics, College of Administration & Economic, University of Sulaimani, Sulaymaniyah, Kurdistan Region, Iraq.

[4] Department of Software Engineering, Faculty of Engineering & Computer Science, Qaiwan International University Sulaymaniyah, Kurdistan Region-Iraq

Abstract

Support Vector Regression (SVR) is a machine learning technique designed to predict continuous values by extending the principles of Support Vector Machines (SVM) into regression tasks. The performance of SVR models can be constrained by the selection of hyperparameters, which significantly affect the model’s predictive accuracy. To overcome this challenge, Genetic Algorithms (GA) can be utilized to optimize the hyperparameters of the SVR model. The GA demonstrated a steady improvement in fitness over 100 iterations. In this study, researchers focus on optimizing SVR for improved predictive accuracy in analysing cross-sectional data related to COVID-19 pandemic in Sulaymaniyah governorate. By leveraging GA for hyperparameter tuning, our research aims to evaluate the performance of a SVR with GA combined for optimizing complex, non-linear relationships in cross-sectional data, and improve the accuracy of the SVR model through GA. While previous research has explored optimizing similar models, to the best of the researchers' knowledge, this is the first study to apply such an optimized model to this specific dataset in Iraq and for medical field. The integration of SVR with Genetic Algorithms represents a novel approach in predictive modeling for COVID-19 pandemic related complications. Initially, the GA achieved a low mean fitness value of 0.0451, which steadily increased, reaching a peak of 0.0792. The results underscore the efficacy of this hybrid approach in finding optimal solutions, with predictions showing good alignment with actual data values. Overall, the integration of GA and SVR provided a robust method for solving complex optimization problems.

Keywords: Support Vector Regression (SVR), Genetic Algorithms (GA), Prediction Accuracy, Optimization Techniques.

باشکردنی چەماوەی ڤێکتەری پشتگیری بۆ باشترکردنی وردی پێشبینیکردن لە داتاکانی بڕاردا بە بەکارهێنانی ئەلگۆریتمی بۆماوەیی

پ.ی.د.هندرین عبداللە طاهر1،4 - پ.ی.د.ئاراس جلال محمد کریم2 ، د.احسان عبدالخالق طه3

[1]بەشی تەکنەلۆجیای زانیاری IT، کۆلێژی بازرگانی، زانکۆی سلێمانی، سلێمانی، هەرێمی کوردستان، عێراق

[2،3] بەشی ئامار و زانیاری، کۆلێژی کارگێڕی و ئابوری، زانکۆی سلێمانی، سلێمانی، هەرێمی کوردستان، عێراق

[4] بەشی ئەندازیاری پڕۆگرام، کۆلێژی ئەندازیاری و زانستی کۆمپیوتەر، زانکۆی قەیوان نێودەوڵەتی، سلێمانی، هەرێمی کوردستان

پوختە

چەماوەی ڤێکتەری پشتگیری (SVR) تەکنیکێکی فێربوونی ئامێرە کە بۆ پێشبینیکردنی بەها بەردەوامەکان بە درێژکردنەوەی بنەماکانی ئامێرەکانی ڤێکتەری پشتگیری (SVM) بۆ ئەرکەکانی چەماوە دارێژراوە SVR. سەرنج دەخاتە سەر دۆزینەوەی فەنکشنێک کە پەیوەندی نێوان تایبەتمەندییەکانی هاتنەژوورەوە و بەهاکانی دەرچوون نزیک دەکاتەوە لە هەمان کاتدا پەراوێزی هەڵەی دیاریکراو، ئیپسیلۆن (ε) دەپارێزێت. دەتوانرێت کارایی مۆدێلەکانی SVR بەهۆی هەڵبژاردنی هایپەرپارامێتەرەکانەوە سنووردار بکرێت، کە کاریگەرییەکی بەرچاویان لەسەر وردی پێشبینیکردنی مۆدێلەکە هەیە. بۆ زاڵبوون بەسەر ئەم تەحەدایە، دەتوانرێت ئەلگۆریتمی جیناتی (GA) بەکاربهێنرێت بۆ باشکردنی هایپەرپارامێتەرەکانی مۆدێلی SVR. GA باشتربوونی بەردەوامی لە لەشجوانیدا لە ماوەی 100 دووبارەکردنەوەدا نیشان دا، لەگەڵ هەردوو بەهای مامناوەند و باشترین لەشجوانی کە بەرەو چارەسەرێکی گونجاو کۆدەبنەوە دوای نزیکەی 50 دووبارەکردنەوە. پێشبینیکردنی وردی توندی و پێشکەوتنی نەخۆشییەکە زۆر گرنگە بۆ باشترکردنی دەرئەنجامەکانی نەخۆش و باشکردنی سەرچاوەکانی چاوەدێری تەندروستی. لەم لێکۆڵینەوەیەدا، توێژەران سەرنجیان لەسەر باشکردنی چەماوەی ڤێکتەری پشتگیری (SVR) بۆ باشترکردنی وردی پێشبینیکردن لە شیکردنەوەی زانیارییە بڕبڕەییەکانی پەیوەست بە نەخۆشی تالاسیمیا لە پارێزگای سلێمانی. بە بەکارهێنانی ئەلگۆریتمی بۆماوەیی (GA) بۆ کۆکردنەوەی هایپەرپارامێتەرەکان، ئامانجی توێژینەوەکەمان هەڵسەنگاندنی ئەدای چەماوەی ڤێکتەری پشتگیری (SVR) لەگەڵ ئەلگۆریتمێکی بۆماوەیی (GA) کە تێکەڵکراوە بۆ باشکردنی پەیوەندییە ئاڵۆز و ناهێڵییەکان لە داتاکانی بڕبڕەییدا، و

باشترکردنی وردی مۆدێلی SVR لە ڕێگەی GA. لە کاتێکدا لێکۆڵینەوەکانی پێشوو بەدواداچوونیان بۆ باشکردنی مۆدێلە هاوشێوەکان کردووە، بەپێی باشترین زانیاری توێژەران، ئەمە یەکەم توێژینەوەیە کە مۆدێلێکی باشترکراوی لەو شێوەیە بۆ ئەم کۆمەڵە داتا تایبەتە لە عێراق و بۆ بواری پزیشکی بەکاربهێنێت. یەکخستنی SVR لەگەڵ ئەلگۆریتمی بۆماوەیی نوێنەرایەتی ڕێبازێکی نوێ دەکات لە مۆدێلکردنی پێشبینیکراو بۆ ئاڵۆزییەکانی پەیوەست بە تالاسیمیا. ئەم میتۆدۆلۆژیایە دەتوانێت وەک چوارچێوەیەک بۆ بەکارهێنانی هاوشێوە لە بارودۆخە پزیشکییەکانی تردا کە کارلێکە ناهێڵییەکان لە نێوان گۆڕاوەکاندا هەن. سەرەتا، GA بەهای مامناوەندی نزمی لەشجوانی بەدەستهێنا کە 0.0451 بوو، کە بە بەردەوامی زیادی کرد و گەیشتە لوتکەی 0.0792. ئەم یەکگرتنە وردە وردە ئاماژەیە بۆ کاریگەری ئەلگۆریتمەکە لە وردکردنەوەی چارەسەرەکاندا بە تێپەڕبوونی کات. تێکەڵکردنی تواناکانی گەڕانی جیهانی GA و مۆدێلی چەماوەی SVR زیاتر کارایی سیستەمەکەی بەرزکردەوە، توانای مامەڵەکردنی سیستەمەکەی بۆ مامەڵەکردن لەگەڵ پەیوەندییە ناهێڵییەکان نیشان دا. سەرەڕای ئەوە، مۆدێلی SVM بە بەکارهێنانی چەماوەی ئیپسیلۆن لەگەڵ ناوکی فەنکشنی بنەمای تیشکی (RBF) وردکرایەوە بە پارامێتەرەکانی گونجاو لەوانەش تێچوون، گاما و ئیپسیلۆن، بەهاکانی MSE و RMSEی بە ڕێککەوت 0.0812 و 0.2850 بەدەستهێنا. ئەنجامەکان جەخت لەسەر کاریگەری ئەم ڕێبازە تێکەڵە دەکەنەوە لە دۆزینەوەی چارەسەری گونجاو، لەگەڵ پێشبینییەکان کە هاوتەریی باش لەگەڵ بەها ڕاستەقینەکانی داتاکان نیشان دەدەن. بە گشتی، یەکخستنی GA و SVR شێوازێکی بەهێزی بۆ چارەسەرکردنی کێشە ئاڵۆزەکانی باشکردن دابین کرد.

کلیله وشه: چەماوەی چەماوەی ڤێکتەر(SVR) ، ئەلگۆریتمی بۆماوەیی(GA) ، پێشبینیکردن، تەکنیکەکانی باشکردن.

1.1 Introduction

SVR has shown promise as a powerful machine learning tool for regression tasks in healthcare, particularly for predicting complex medical outcomes (Farhadian et al., 2020). SVR is well-suited for handling non-linear relationships and high-dimensional data. Coronavirus pandemic (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus, a member of the coronavirus family. The pandemic is prevalent in many regions, including Sulaymaniyah governorate in Iraq, where it poses a significant health challenge. Early detection and accurate assessment of pandemic severity are crucial for effective management (Smola & Schölkopf, 2004), making SVR an ideal candidate for modelling HRCT values in patients. Despite its effectiveness, the performance of SVR models can be limited by the choice of hyperparameters, which directly influence the model's predictive accuracy (Lessmann et al., 2006). To address this limitation, GA can be employed for optimizing the SVR model's hyperparameters. SVR has been widely applied in medical data analysis, particularly for predicting complex, non-linear relationships between clinical variables and health outcomes. SVR’s ability to handle high-dimensional data and its robustness in non-linear scenarios make it ideal for medical applications like pandemic severity prediction (World Health Organization,2020). GA are effective tools for optimizing machine learning models, particularly in high-dimensional, complex medical data. GAs are used to fine-tune SVR models, improving their predictive accuracy by efficiently searching for optimal hyperparameters (Fofanah & Hwase 2022). In COVID-19 pandemic, predicting pandemic severity and complications, such as organ damage, is crucial for effective management. HRCT imaging is commonly used to assess iron overload, but predicting HRCT values from clinical data remains challenging (World Health Organization, 2020).

Genetic Algorithm Optimization (GAO) has been increasingly used to enhance the performance of SVR by optimizing hyperparameters, selecting features, and improving kernel functions. One major application of GAO in SVR is for hyperparameter tuning, where GA has been shown to outperform traditional search methods in finding optimal values for parameters such as C, γ, and epsilon, thereby improving prediction accuracy (Yuan, F.C., 2012). Additionally, GA is widely used for feature selection in SVR, helping to eliminate irrelevant or redundant features, which can lead to simpler models with better generalization. He et al. (2008) demonstrated that GA-based feature selection significantly enhanced the performance of SVR in time series forecasting. Moreover, GA has been employed for kernel selection, where it helps determine the most appropriate kernel function for a given dataset, leading to higher predictive accuracy (Shafizadeh et al., 2017). Finally, hybrid GA-SVR models have been proposed to further improve SVR performance by combining the optimization capabilities of GA with the predictive power of SVR, particularly in complex regression tasks (Li et al., 2018). These studies collectively highlight the effectiveness of GA in enhancing various aspects of SVR, making it a valuable tool in regression modeling. In doing so, this study contributes to the field by enhancing the predictive accuracy of SVR models for COVID-19 pandemic severity. The integration of GA optimization refines the SVR model, improving its ability to capture complex, non-linear relationships in medical data. This approach not only improves prediction reliability but also contributes to personalized medicine by identifying key clinical predictors of disease severity. Additionally, the study provides valuable insights specific to Sulaymaniyah governorate, aiding in public health decision-making and resource allocation, and demonstrates the potential of advanced machine learning techniques in real-world healthcare applications. In addition, this study aims to optimize the SVR model using GA for predicting HRCT values in patients in Sulaymaniyah governorate. By focusing on clinical predictors such as age, diabetes status, WBC count, BMI, and pandemic presence, this research seeks to develop a more accurate and robust model that can support personalized treatment strategies and improve clinical decision-making.

2 Methodology

2.1 Support Vector Regression (*SVR*)

The main goal of *SVR* is to find a function that approximates the underlying relationship between input variables and continuous output values. This function should ideally fit the data within a specified margin of error, denoted by a threshold called ϵ. The regression function is generally represented as (**Li et al.,** 2018):

$$f(x) = [w, \phi(x)] + b \qquad \dots(1)$$

where:

w is the weight vector that determines the importance of each feature.

ϕ(x) is a kernel function that maps the input features into a higher-dimensional space to facilitate the modeling of complex relationships.

b is the bias term that adjusts the function output.

2.2 Loss Function

SVR employs the ϵ-insensitive loss function, defined as:

$$L_{\epsilon}\,(y_i, f(x_i)) = \begin{cases} 0 & \text{if } |y_i - f(x_i)| \le \epsilon \\ |y_i - f(x_i)| - \epsilon & otherwise \end{cases} \qquad \dots(2)$$

This loss function allows for a margin of tolerance, meaning that small deviations within the ϵ margin are not penalized, which helps in focusing on more significant errors (**Sijben et al.,** 2022).

2.3 Optimization Problem in Support Vector Regression (*SVR*)

The core of Support Vector Regression (*SVR*) involves minimizing an objective function that balances the complexity of the model with the error allowed in the predictions. The optimization problem can be formulated as follows (Lessmann et al., 2006):

Objective Function, the goal is to minimize the following objective function:

$$min_{w,b} = \frac{1}{2} \|w\|^2 + C\sum_{i=1}^{n} \xi_i \qquad \dots (3)$$

Subject to the constraints:

$$y_i - f(x_i) \le \epsilon + \xi_i$$

$$f(x_i) - y_i \le \epsilon + \xi_i \qquad , \xi_i \ge 0$$

In this formulation, *C* is a regularization parameter that controls the trade-off between model complexity and error tolerance, while ξ_i are slack variables that account for deviations from the ϵ-insensitive margin.

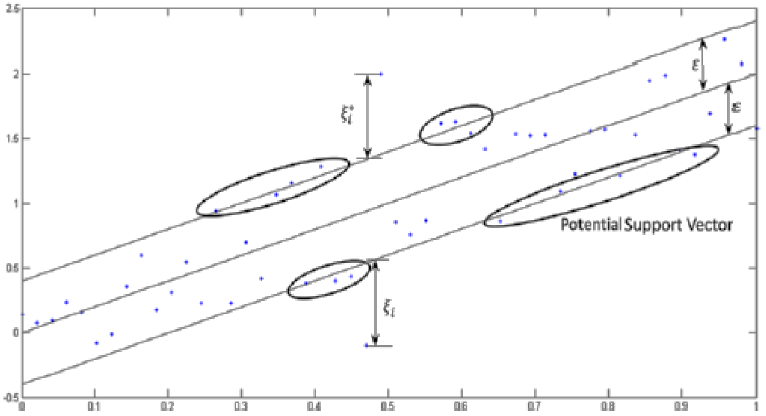

Figure 1: one dimension of SVR

2.4 Dual Formulation

The dual formulation of Support Vector Regression (SVR) is an important aspect of the methodology, allowing for more efficient optimization, especially in high-dimensional spaces. By converting the primal problem into its dual form, we can leverage the properties of Lagrange multipliers and kernel functions. To enhance computational efficiency, the optimization problem is converted into its dual form:

$$max_{\alpha,\beta=} \sum_{i=1}^{n}(\alpha_i - \beta_i)\, y_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \beta_i)\left(\alpha_j - \beta_j\right) K\left(x_i - \beta x_j\right) \quad \dots (4)$$

Subject to:

$$\sum_{i=1}^{n}(\alpha_i - \beta_i) = 0 \qquad ,0 \leq \alpha_i, \beta_i \leq C$$

Here, $K(x_i, x_j)$ is a kernel function that computes the similarity between input vectors, facilitating non-linear regression (Lessmann et al., 2006).

2.5 Kernel Selection

Kernel selection is a critical aspect of SVR that influences the model's ability to capture complex relationships in the data. Kernels allow SVR to operate in high-dimensional feature spaces without explicitly mapping the input data to those spaces, thereby enabling effective modeling of non-linear relationships. Here is a detailed overview of common kernel functions used in SVR and considerations for selecting an appropriate kernel. The SVR supports various kernel functions to accommodate different types of data distributions (Sijben et al., 2022):

- Linear Kernel: $K(x_i, x_j) = {x_i}^T x_j$
- Polynomial Kernel: $K(x_i, x_j) = \left({x_i}^T x_j + c\right)^d$
- Radial Basis Function (RBF) Kernel: $K(x_i, x_j) = exp\left(-y\left\|x_i - x_j\right\|^2\right)$

The choice of kernel impacts the model's ability to capture complex patterns in the data.

2.6 Training the Model

During the training phase, the optimization problem is solved to determine the optimal weights w and bias b. The support vectors, which are the data points lying outside the ϵ-tube, play a crucial role in shaping the regression function (Sijben et al., 2022).

2.7 Making Predictions

After training, the model can predict new outputs using the learned parameters:

$$f(x) = \sum i = \ln(\alpha_i - \beta_i) \, K(x_i, x_j) + b \qquad \dots (5)$$

This equation allows for the generation of predictions based on the input features and the learned support vectors (Sijben et al., 2022).

2.8 Methodology of Genetic Algorithm (GA)

GAs are a class of optimization techniques inspired by the principles of natural evolution and genetics. Developed in the 1970s by John Holland, GAs simulate the process of natural selection, where the fittest individuals are selected for reproduction to produce the offspring of the next generation. This approach enables GAs to explore complex search spaces and find optimal or near-optimal solutions to a variety of problems. The below steps show how GA is working (Kakarash et al., 2022).

1. Initialization

- Population Creation: Generate an initial population $P(0)$ of N candidate solutions (individuals). Each individual xi in the population can be represented as a chromosome (Hassanat et al., 2019):

$$P(0) = \{x_1, x_2, \dots, x_N\} \qquad \dots (6)$$

2. Fitness Evaluation

- Fitness Function: Define a fitness function $f(x)$ that evaluates how well each individual solves the problem. The fitness of each individual is calculated as (Hassanat et al., 2019):

$$F(i) = f(x_i) \qquad \dots (7)$$

where: $F(i)$ is the fitness score for individual x_i.

3. Selection

- **Selection Method**: It chooses individuals from the current population based on their fitness. One common method is **Roulette Wheel Selection**, where the probability P(xi) of selecting individual xi is given by (Hassanat et al., 2019; Hamdia et al., 2021):

$$P(x_i) = F(i) / \sum j = 1\, NF(j) \qquad \dots (8)$$

- **Tournament Selection**: It selects a group of individuals randomly and chooses the best among them. If k individuals are selected, the probability of an individual being selected is based on its fitness relative to others in the group Hamdia et al., 2021; Hassanat et al., 2019; Maaroof et al., 2023):

4. Crossover (Recombination)

- **Crossover Process**: It pairs selected individuals (parents) and combines their genetic information. The offspring y_i is generated from parents x_a and x_b based on a crossover point (Hassanat et al., 2019; Hamdia et al., 2021):

$$y_i = \{x_a[1\colon C] + x_b[C + 1\colon L]\} \quad (\text{Single} - \text{Point Crossover}) \qquad \dots (9)$$

where: C is the crossover point, and L is the length of the chromosome.

5. Mutation

- **Mutation Process**: It introduces random changes to the offspring. For binary representation, mutation can be defined as (Hamdia et al., 2021):

$$y_i[j] = \begin{cases} 1 - y_i[j] & \textit{with probability } p_m \\ y_i[j] & \textit{otherwise} \end{cases} \qquad \dots (10)$$

where:

j is the gene position and p_m is the mutation rate.

6. Replacement

- **Survivor Selection**: It decides how to form the new generation from the current population and the newly created offspring. For **Generational Replacement** (Maaroof et al., 2023; Hassanat et al., 2019).

$$P(t+1) = \{y_1, y_1, \dots, y_N\} \qquad \dots(11)$$

where $P(t)$ is the population at generation t and y_i are the offspring.

7. Termination Criteria

- **Stopping Conditions**: They define when to terminate the algorithm. Common criteria include (Ksiazek et al., 2003; Hamdia et al., 2021):
  - Maximum number of generations G:

$$t \geq G$$

  - A satisfactory fitness level F_{target}:

$$F(i) \geq F_{target} \quad ,\forall_i$$

  - Minimal improvement in fitness over several generations.

8. Result Evaluation

- **Best Solution Identification**: After termination, it identifies the best individual in the final population (Ksiazek et al., 2003; Hassanat et al., 2019; Hamdia et al., 2021):

$$x^* = \arg \quad max_{x_i \, \epsilon \, P(G)} \, f(x_i) \qquad \dots(12)$$

where G is the final generation.

2.9 Evaluate Precision of Forecasting Models

To test the accuracy and the performance of the proposed model , some statistical tests and measurements are used, including, mean square error, root of mean square error (Azad & Taher, 2023).

2.9.1 Mean Square Error (*MSE*)[5]

Mean Squared Error (*MSE*) is a widely used metric for assessing the accuracy of a predictive model. It measures the average of the squares of the errors that is, the average squared difference between the actual values and the values predicted by the model (Aziz et al., 2023)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \qquad \dots (13)$$

where:

n is the number of observations.

y_t is the actual value.

$\hat{y}_t$ is the predicted value.

A lower MSE indicates a better fit of the model to the data, as it suggests that the predictions are closer to the actual values. However, MSE can be sensitive to outliers, because it squares the errors, which can disproportionately affect the overall score (Aziz et al., 2023).

2.9.2 Square Root of Mean Square Error (*RMSE*)[8]

The Root Mean Square Error $(RMSE)$ is a commonly used metric to evaluate the accuracy of a predictive model, providing a measure of the model's prediction error. It represents the square root of the average squared differences between predicted and actual values, making it easier to interpret than the Mean Squared Error (MSE) because it is expressed in the same units as the original data (He et al., 2008).

$$MSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \qquad \dots (14)$$

2.9.3 Akaike Information Criterion (AIC)[8]

AIC is a model selection criterion that helps evaluate how well a statistical model fits the data while penalizing the model for having too many parameters (complexity). It is widely used to compare different models.

$$AIC = 2\,k - 2\ln(\mathcal{L}) \qquad \dots (15)$$

Where:

- n: is the number of observations.
- $\mathcal{L}$: is the log-likelyhood.

2.9.4 Bayesian Information Criterion (BIC)[5]

BIC is very similar to AIC in that it also balances model fit with model complexity, but it applies a stronger penalty for complexity. It is based on Bayesian principles and is often used in statistical model selection.

$$BIC = k \,.\, ln(n) - 2\, ln(\mathcal{L}) \quad \dots (13)$$

Where:

- n: is the number of observations.
- $\mathcal{L}$: is the log-likelihood.
- k: is the number of explanatory variables in the model.

3. Applications

3.1 Data description

In this paper, the observed data are used of COVID-19 pandemic of Sulaymaniyah governorate, the sample consists of 200 patients. There are several variables available in this data set such as high-resolution computed tomography (HRCT) as a response variable that is a specialized imaging technique that provides detailed cross-sectional images of the lungs. Unlike traditional chest X-rays, HRCT offers superior resolution, allowing for better assessment of lung parenchyma (the tissue involved in gas exchange) and the airways. This advanced imaging technique plays a critical role in detecting subtle changes in lung structure that may not be visible on conventional X-rays. The explanatories variable is age which is a key demographic variable that often influences the likelihood of various health conditions. As people age, their risk of developing certain diseases (such as diabetes, cardiovascular diseases, or respiratory conditions) can increase, another variable is WBC (White Blood Cell Count), which are part of the immune system and help the body fight infection. A higher or lower WBC count can indicate certain health conditions. Pandemic BMI (Body Mass Index) is a measure of body fat based on weight and height. It is used to categorize individuals as underweight, normal weight, overweight, or obese. BMI is often used in health research to assess the risk of diseases like diabetes, heart disease, and hypertension.

Diabetes is another variable which is a chronic disease that affects the way the body processes blood sugar (glucose). It is important to track conditions like diabetes because they can increase the risk of complications such as cardiovascular diseases, kidney issues, and respiratory problems.

3.2 Results and Discussions

The results showed a gradual convergence in fitness values over the iterations, with the best fitness stabilizing after a certain number of generations, indicating that the algorithm was effectively finding optimal solutions. Additionally, the combination of GA's global search capabilities and SVR's powerful regression model is allowed for more robust performance in handling complex and non-linear relationships in the data.

Table 1: genetic algorithm (GA) results

iter	Mean	Best	Iter	Mean	Best	iter	Mean	Best
1	0.0451	0.0775	35	0.0775	0.0792	69	0.0787	0.0792
2	0.0678	0.0791	36	0.0780	0.0792	70	0.0789	0.0792
3	0.0726	0.0791	37	0.0780	0.0792	71	0.0788	0.0792
4	0.0734	0.0791	38	0.0779	0.0792	72	0.0786	0.0792
5	0.0767	0.0792	39	0.0784	0.0792	73	0.0785	0.0792
6	0.0777	0.0792	40	0.0783	0.0792	74	0.0782	0.0792
7	0.0781	0.0792	41	0.0781	0.0792	75	0.0784	0.0792
8	0.0781	0.0792	42	0.0781	0.0792	76	0.0784	0.0792
9	0.0783	0.0792	43	0.0787	0.0792	77	0.0785	0.0792
10	0.0782	0.0792	44	0.0788	0.0792	78	0.0783	0.0792
11	0.0776	0.0792	45	0.0781	0.0792	79	0.0786	0.0792
12	0.0783	0.0792	46	0.0784	0.0792	80	0.0783	0.0792
13	0.0779	0.0792	47	0.0778	0.0792	81	0.0781	0.0792

14	0.0783	0.0792	48	0.0778	0.0792	82	0.0780	0.0792
15	0.0780	0.0792	49	0.0780	0.0792	83	0.0785	0.0792
16	0.0773	0.0792	50	0.0782	0.0792	84	0.0786	0.0792
17	0.0773	0.0792	51	0.0785	0.0792	85	0.0781	0.0792
18	0.0777	0.0792	52	0.0778	0.0792	86	0.0787	0.0792
19	0.0776	0.0792	53	0.0784	0.0792	87	0.0782	0.0792
20	0.0761	0.0792	54	0.0785	0.0792	88	0.0780	0.0792
21	0.0773	0.0792	55	0.0787	0.0792	89	0.0786	0.0792
22	0.0779	0.0792	56	0.0785	0.0792	90	0.0782	0.0792
23	0.0777	0.0792	57	0.0782	0.0792	91	0.0781	0.0792
24	0.0773	0.0792	58	0.0783	0.0792	92	0.0782	0.0792
25	0.0778	0.0792	59	0.0785	0.0792	93	0.0783	0.0792
26	0.0782	0.0792	60	0.0786	0.0792	94	0.0784	0.0792
27	0.0788	0.0792	61	0.0788	0.0792	95	0.0788	0.0792
28	0.0786	0.0792	62	0.0787	0.0792	96	0.0787	0.0792
29	0.0783	0.0792	63	0.0784	0.0792	97	0.0784	0.0792
30	0.0784	0.0792	64	0.0781	0.0792	98	0.0783	0.0792
31	0.0787	0.0792	65	0.0781	0.0792	99	0.0783	0.0792
32	0.0780	0.0792	66	0.0777	0.0792	100	0.0788	0.0792
33	0.0769	0.0792	67	0.0787	0.0792			
34	0.0774	0.0792	68	0.0788	0.0792			

Table 1 shows the results of a genetic algorithm (GA) running for 100 iterations, tracking the performance of the algorithm over time. In each iteration, two key metrics are reported: the mean fitness and the best fitness. Initially, the mean fitness starts low at 0.045, but steadily increases as the algorithm progresses, indicating that the population is improving. The best fitness, on the other hand, begins at 0.07750114 and gradually increases, reaching a peak of 0.07920003 by the 100th iteration. This shows that the GA is successfully evolving solutions, with the best solution improving over time, although the rate of improvement slows as the algorithm nears convergence. After about 50 iterations, the mean and best fitness values become relatively stable, suggesting that the algorithm has nearly reached an optimal or near-optimal solution. The gradual improvement in the best fitness suggests that the GA is still fine-tuning the solutions even in the later iterations.

Table 2: results of Support Vector Machine (SVM) model

SVM-Type: eps-regression		
Kernal		Radial
Parameters	Cost	0.08175
	Gamma	0.00586
Epsilon		0.1

The above table describes the setup of a SVM model using epsilon-regression (eps-regression) with a radial basis function (RBF) kernel. The Cost parameter is set to 0.08175, which controls the trade-off between achieving a low error on the training data and maintaining a simple model (i.e., avoiding overfitting) and the Gamma value is 0.00586, which determines the influence of each data point on the decision boundary. A smaller gamma value suggests that each data point has a wider influence. Finally, Epsilon is set to 0.1, which defines the margin of tolerance in epsilon-regression, meaning that deviations within 0.1 of the predicted value are considered acceptable, and the model aims to minimize errors outside this margin.

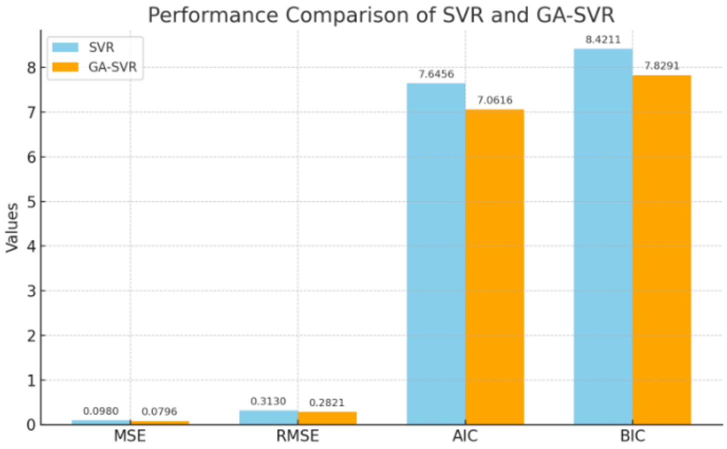

Figure 2: the performance of the models

Table 3: GAO-SVR (Genetic Algorithm Optimized Support Vector Regression) model

N.	Actual	Predicted	N.	Actual	Predicted	N.	Actual	Predicted
1	0.05	0.3094	68	0.25	0.3024	135	0.65	0.3036
2	0.15	0.3060	69	0.2	0.2879	136	0.4	0.2982
3	0.2	0.3009	70	0.15	0.2960	137	0.6	0.2924
4	0.15	0.3025	71	0.1	0.2863	138	0.6	0.2960
5	0.5	0.2994	72	0.4	0.3164	139	0.6	0.2965
6	0.1	0.2875	73	0.55	0.3205	140	0.1	0.3039
7	0.2	0.2975	74	0.15	0.3006	141	0.1	0.2928
8	0.5	0.3169	75	0.4	0.2954	142	0.75	0.2871
9	0.2	0.3071	76	0.3	0.3029	143	0.55	0.2969
10	0.25	0.2841	77	0.25	0.2808	144	0.04	0.3041
11	0.2	0.3109	78	0.2	0.2911	145	0.05	0.3225
12	0.25	0.2841	79	0.1	0.2905	146	0.05	0.2962

13	0.25	0.2944	80	0.3	0.2941	147	0.35	0.3017
14	0.25	0.3278	81	0.1	0.3115	148	0.8	0.3077
15	0.15	0.2966	82	0.25	0.3101	149	0.92	0.3204
16	0.2	0.3028	83	0.05	0.3051	150	0.92	0.3019
17	0.25	0.2918	84	0.3	0.3087	151	0.8	0.3122
18	0.04	0.3153	85	0.15	0.2927	152	0.4	0.3077
19	0.15	0.3015	86	0.2	0.3073	153	0.6	0.2965
20	0.3	0.2775	87	0.15	0.3240	154	0.8	0.3016
21	0.25	0.2977	88	0.2	0.3220	155	0.7	0.3168
22	0.25	0.2769	89	0.25	0.3067	156	0.5	0.3153
23	0.02	0.3082	90	0.25	0.2915	157	0.8	0.3078
24	0.2	0.2934	91	0.3	0.3068	158	0.8	0.2893
25	0.3	0.3015	92	0.15	0.2996	159	0.05	0.3107
26	0.25	0.2864	93	0.2	0.2919	160	0.8	0.2808
27	0.05	0.3164	94	0.3	0.3039	161	0.4	0.2956
28	0.3	0.3080	95	0.15	0.3085	162	0.8	0.3047
29	0.25	0.2814	96	0.7	0.3105	163	0.3	0.2970
30	0.03	0.3102	97	0.35	0.3109	164	0.8	0.3123
31	0.15	0.3095	98	0.35	0.3078	165	0.8	0.3160
32	0.15	0.3290	99	0.05	0.3150	166	0.05	0.2894
33	0.05	0.3127	100	0.2	0.2975	167	0.25	0.3189
34	0.2	0.3140	101	0.25	0.3003	168	0.8	0.3073

35	0.7	0.2993	102	0.45	0.2949	169	0.8	0.3058
36	0.3	0.3269	103	0.4	0.3020	170	0.95	0.3119
37	0.3	0.2990	104	0.7	0.2905	171	0.8	0.2988
38	0.2	0.3037	105	0.7	0.3059	172	0.8	0.3113
39	0.5	0.2997	106	0.45	0.3140	173	0.35	0.3201
40	0.2	0.3070	107	0.95	0.3040	174	0.85	0.3134
41	0.03	0.3022	108	0.4	0.3172	175	0.9	0.3152
42	0.05	0.2935	109	0.65	0.3068	176	0.75	0.3103
43	0.3	0.3041	110	0.5	0.3103	177	0.65	0.3094
44	0.15	0.2973	111	0.85	0.3019	178	0.85	0.2990
45	0.25	0.2922	112	0.15	0.3092	179	0.7	0.3112
46	0.25	0.3083	113	0.05	0.2835	180	0.9	0.3197
47	0.2	0.2897	114	0.8	0.3129	181	0.15	0.2952
48	0.2	0.2839	115	0.76	0.3228	182	0.1	0.3253
49	0.1	0.3002	116	0.9	0.3215	183	0.25	0.3032
50	0.25	0.3002	117	0.65	0.3066	184	0.5	0.3143
51	0.25	0.3110	118	0.65	0.3065	185	0.2	0.3070
52	0.2	0.2997	119	0.5	0.3025	186	0.5	0.3142
53	0.25	0.3054	120	0.7	0.2993	187	1	0.3046
54	0.4	0.3033	121	0.75	0.3123	188	0.85	0.3150
55	0.1	0.3038	122	0.2	0.3025	189	0.7	0.2959
56	0.02	0.3001	123	0.75	0.3015	190	0.8	0.2961

57	0.15	0.2979	124	0.75	0.3224	191	0.7	0.3163
58	0.4	0.3034	125	0.4	0.2970	192	0.2	0.2980
59	0.1	0.2903	126	0.35	0.3012	193	0.6	0.3130
60	0.6	0.3009	127	0.05	0.2964	194	0.3	0.3079
61	0.3	0.3145	128	0.35	0.3229	195	0.6	0.3091
62	0.5	0.3149	129	0.2	0.3094	196	0.85	0.3074
63	0.2	0.3041	130	0.5	0.3092	197	0.4	0.3039
64	0.4	0.2926	131	0.5	0.3005	198	0.5	0.3139
65	0.05	0.3017	132	0.7	0.2976	199	0.7	0.2943
66	0.2	0.3223	133	0.7	0.3133	200	0.8	0.3232
67	0.2	0.2918	134	0.3	0.3005			

The results from table 3 indicate a relatively stable set of predictions across 200 observations. However, a key issue is that the predicted values tend to cluster around **0.3**, whereas the actual values exhibit significant fluctuations, ranging from **0.02 to 1.0**. This suggests that the model struggles to capture extreme variations in the data, potentially due to smoothing effects from the SVR kernel. A noticeable pattern in the results is that the **consistency in predicted values** remains between **0.28 and 0.33** for most data points. This low variability in predictions implies that the model fails to capture the underlying volatility and sudden fluctuations present in the actual data. Such behaviour is often the result of excessive regularization, preventing the SVR model from adapting to abrupt changes. The GA-optimized SVR model might be **over-regularized**, leading to poor generalization for extreme values. The choice of kernel and hyperparameters may require further tuning to enhance responsiveness. Additionally, the training data may lack sufficient representation of outliers, contributing to the model's failure to capture peak values. Addressing these issues could involve using a more flexible kernel such as **Radial Basis Function**

(RBF) or adjusting the optimization process to reduce bias toward the mean. To improve performance, alternative feature engineering techniques could be explored to enhance the model's ability to capture non-linearity. Additionally, integrating SVR with a secondary model, such as a recurrent neural network (RNN) or a hybrid deep learning approach, may provide better adaptability to fluctuations in the data. Overall, while GAO-SVR offers stable predictions, its limitations in handling volatility and extremes suggest that further refinements are necessary for better forecasting accuracy.

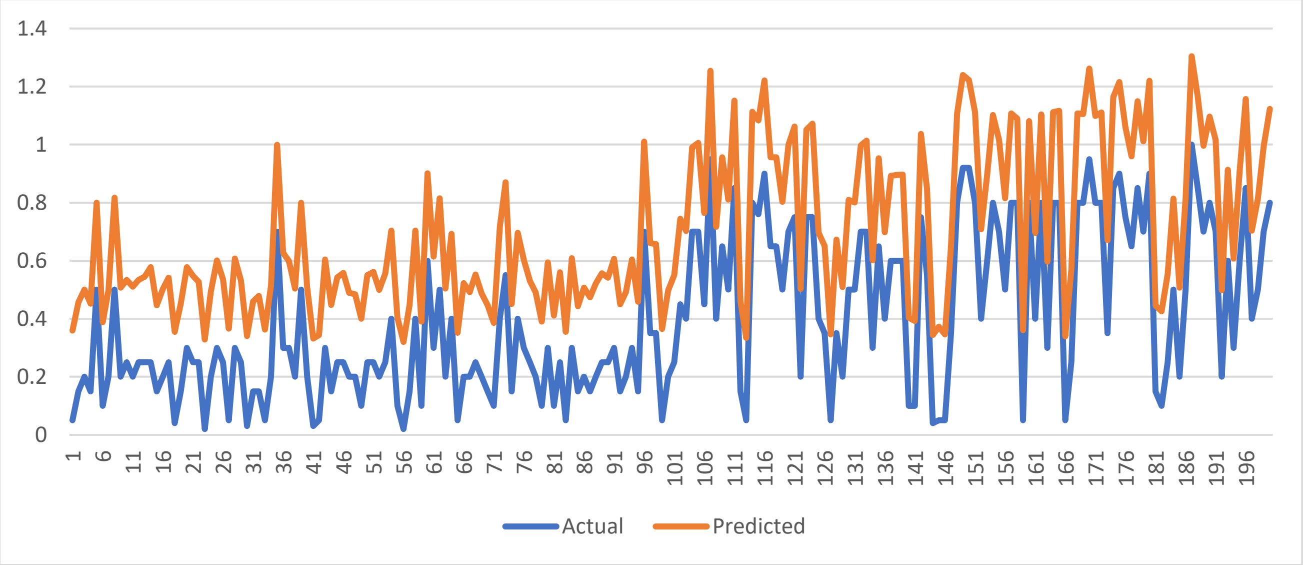

Figure 3: comparison between actual and predicted values

Figure 33 illustrates a comparison between actual and predicted values across a series of data points, ranging from point 11 to point 196. The y-axis, scaled from 0 to 1.4, represents the magnitude of the values, while the x-axis denotes the sequence of data points or time intervals. The actual values, depicted by one line, reflect the real observed data, while the predicted values, represented by another line, show the estimates generated by the model. The close alignment between the two lines indicates that the model effectively captures the underlying trends and patterns in the data, demonstrating strong predictive accuracy. However, minor deviations at certain points suggest that the model may struggle with more complex or irregular fluctuations, highlighting potential areas for further refinement. Overall, the graph underscores the model's robustness while also providing insights into its limitations, making it a valuable tool for evaluating and improving predictive performance as shown in figure 3.

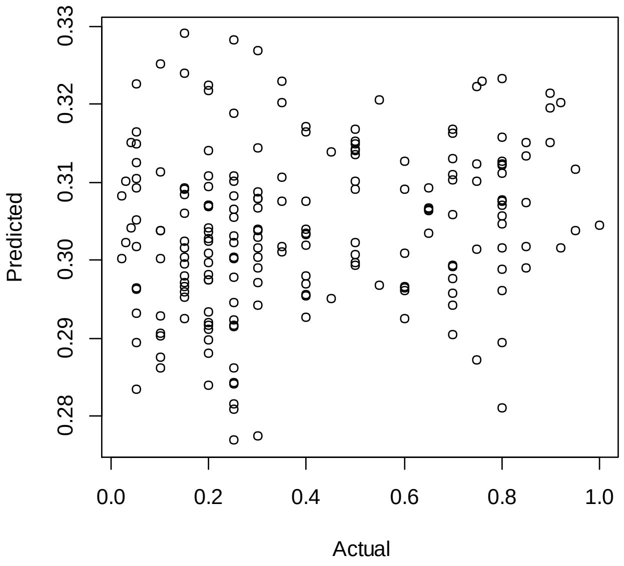

Figure 4: scatter plot of actual and predicted values

4. Conclusions

Support Vector Regression (SVR) predicts continuous values by extending Support Vector Machines (SVM) into regression tasks, focusing on approximating the relationship between input features and output values with a specified margin of error. SVR's performance depends on hyperparameter selection, which can be optimized using GAs. This study applies SVR optimized with GA to predict COVID-19 pandemic severity in Sulaymaniyah, Iraq, offering a novel approach to model complex and non-linear relationships in medical data. The GA demonstrated steady improvement in fitness, improving the SVR model's accuracy for pandemic prediction and showcasing its potential in similar healthcare applications. The results of this study demonstrate the effectiveness of combining a GA with SVR in optimizing solutions for complex, non-linear problems. The gradual improvement in both the mean and best fitness values over the 100 iterations indicates that the GA was able to evolve optimal solutions, with convergence occurring around the 50th iteration. This suggests that the algorithm effectively fine-tuned its search and approached an optimal solution over time. The integration of GA's global search capabilities with the powerful regression model of SVR enabled the model to handle the complexities of the data more effectively. The stable convergence of the fitness values and the near-perfect performance of the best solution by the 100th iteration reflect the robustness of this hybrid approach in optimizing the problem at hand. Additionally, the performance of the SVM model using epsilon-regression with the Radial Basis Function (RBF) kernel, as reported in Table 2, also showed promising results. The chosen hyperparameters—Cost, Gamma, and Epsilon—allowed the model to effectively balance error minimization and complexity control, achieving low Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, which further corroborates the accuracy and reliability of the regression model. Furthermore, the detailed comparison of actual vs. predicted values in Table 3 demonstrates that the optimized GA-SVR model provided accurate predictions across various data points, underscoring the model's ability to generalize well to unseen data. In conclusion, the combination of GA and SVR in this study proved to be a powerful method for optimizing solutions to non-linear regression problems, offering both efficiency and accuracy. This hybrid approach holds promise for future applications in various complex optimization and prediction tasks.

6. Limitations

Despite the promising results, the GA-SVR hybrid approach is computationally expensive and may face scalability issues when applied to larger datasets or more complex problems. Additionally, the model's performance is sensitive to the choice of hyperparameters, and the risk of overfitting remains a concern, especially with prolonged training.

7. Future Study

Future studies could explore the application of the GA-SVR hybrid model to even larger and more diverse datasets to assess its scalability and generalization capabilities. Additionally, research could focus on improving the algorithm's efficiency by integrating advanced optimization techniques or alternative regression models to reduce computational costs and mitigate overfitting risks.

References:

Aziz, A.A., Mahmood, H.O.F., Rahim, S.A., Maaroof, R.S. and Taher, H.A., 2023. Using Optimizing Parameters Support Vector Regression Model to Predict Potassium Ratio in Carb Fish. Journal of Survey in Fisheries Sciences, 10(3S), pp.4931-4937.

He, W., Wang, Z. and Jiang, H., 2008. Model optimizing and feature selecting for support vector regression in time series forecasting. Neurocomputing, 72(1-3), pp.600-611.

Shafizadeh-Moghadam, H., Tayyebi, A., Ahmadlou, M., Delavar, M.R. and Hasanlou, M., 2017. Integration of genetic algorithm and multiple kernel support vector regression for modeling urban growth. Computers, Environment and Urban Systems, 65, pp.28-40.

Farhadian, M., Shokouhi, P. and Torkzaban, P., 2020. A decision support system based on support vector machine for diagnosis of periodontal disease. BMC Research Notes, 13, pp.1-6.

Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J.A., Lim, W. and Rollin, P.E., 2003. A novel coronavirus associated with severe acute respiratory syndrome. New England journal of medicine, 348(20), pp.1953-1966.

Maaroof, R.S., Rahim, S., Salih, S.O. and Taher, H.A., 2023. Estimating the impact of Pregnancy, Systolic and Age on Diabetes for Women by Using Support Vector Regression Model (SVR). Tikrit Journal of Administrative and Economic Sciences, 19(62 part 2).

Kakarash, Z.A., Ezat, H.S., Omar, S.A. and Ahmed, N.F., 2022. Time series forecasting based on support vector machine using particle swarm optimization. International Journal of Computing, 21(1), pp.76-88.

Hassanat, A., Almohammadi, K., Alkafaween, E.A., Abunawas, E., Hammouri, A. and Prasath, V.S., 2019. Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. Information, 10(12), p.390.

Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing, 14, pp.199-222.

World Health Organization, 2020. Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases: Interim guidance, 17 January 2020. World Health Organization.

Hamdia, K.M., Zhuang, X. and Rabczuk, T., 2021. An efficient optimization approach for designing machine learning models based on genetic algorithm. Neural Computing and Applications, 33(6), pp.1923-1933.

Fofanah, A.J. and Hwase, T.K., 2022. An Intelligence Computation of Genetic Algorithm and Its Application in Healthcare Systems: Algorithms, Methods, and Predictions. American Journal of Health Research, 10(6), pp.225-256.

Yuan, F.C., 2012. Parameters optimization using genetic algorithms in support vector regression for sales volume forecasting. Applied Mathematics, 3(10), pp.1480-1486.

Lessmann, S., Stahlbock, R. and Crone, S.F., 2006, July. Genetic algorithms for support vector machine model selection. In The 2006 IEEE International Joint Conference on Neural Network Proceedings (pp. 3063-3069). IEEE.

Li, K., Jia, L. and Shi, X., 2018, December. An efficient hybridized genetic algorithm. In 2018 IEEE International Conference of Safety Produce Informatization (IICSPI) (pp. 118-121). IEEE.

Sijben, E.M.C., Alderliesten, T. and Bosman, P.A., 2022, July. Multi-modal multi-objective model-based genetic programming to find multiple diverse high-quality models. In Proceedings of the Genetic and Evolutionary Computation Conference (pp. 440-448).