CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(2) (2024), Pages 320-328
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Optimizing Data Classification in Support Vector Machines Using Metaheuristic Algorithms

Qonita Ilmi Awalin[1], Ika Hesti Agustin[1,2,*], Alfian Futuhul Hadi[1], Dafik[1,2], R. Sunder[3]

[1]Department of Mathematics, University of Jember, Indonesia
[2]PUI-PT Combinatorics and Graph, CGANT-University of Jember, Indonesia
[3]School of Computer Science and Engineering, Galgotias University, India

Email: ikahesti.fmipa@unej.ac.id

ABSTRACT

This study aims to optimize the classification of Chronic Kidney Disease (CKD) diagnosis data by comparing the performance of Support Vector Machine (SVM) algorithms enhanced with two metaheuristic optimization methods, namely Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). The data used consist of secondary medical records of CKD patients, which are split into 80% for training and 20% for testing. An oversampling technique was applied to address the issue of imbalanced data. In this study, both models were optimized through hyperparameter γ tuning using the RBF kernel in SVM. The evaluation of the models' performance was conducted using accuracy and error rate metrics, calculated through a confusion matrix. The results show that the SVM-PSO model achieved an accuracy of 97.54%, slightly higher than the SVM-GA model with an accuracy of 97.37%. Additionally, the SVM-PSO model exhibited a lower classification error rate (2.46%) compared to SVM-GA (2.63%). The findings suggest that PSO is more effective in enhancing the classification performance of SVM. The contribution of this research is to provide empirical evidence that hyperparameter optimization using PSO results in a more accurate and convergent classification model compared to GA in the context of CKD diagnosis.

Keywords: Chronic Kidney Disease; classification; SVM-PSO; SVM-GA

INTRODUCTION

Access to instant products containing artificial sweeteners is widespread. The widespread availability of these products has led to increased consumption, which may have negative public health implications. One significant health issue linked to excessive consumption of these products is Chronic Kidney Disease (CKD), a condition where the kidneys progressively lose their ability to filter blood and perform their normal functions effectively. CKD is characterized by a gradual decline in kidney function over time, with diagnoses typically based on a combination of medical history, clinical symptoms, and laboratory results [1]. Early detection and accurate diagnosis are critical to preventing the progression of CKD to end-stage renal disease. Thus, the classification of patient diagnosis data is essential to understanding the factors that influence CKD.

Submitted: September 12, 2024 Reviewed: September 23, 2024 Accepted: October 23, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i2.29320

Classification is a widely used technique in Machine Learning (ML) to predict the class of observed objects based on identified statistical patterns [2]. As a subset of Artificial Intelligence (AI), ML involves systems that can perform tasks and learn from data similarly to humans, enabling the automation of complex data-driven decision-making processes [3]. One of the most popular ML algorithms for classification tasks is the Support Vector Machine (SVM), which is known for its ability to handle both linear and non-linear classification problems [4]. However, optimizing the performance of SVM typically requires tuning hyperparameters, which significantly influences the model's accuracy and robustness. Two commonly used metaheuristic algorithms for hyperparameter optimization are Particle Swarm Optimization (PSO) and Genetic Algorithm (GA).

The PSO algorithm, inspired by the social behavior of bird flocks, has strong exploration capabilities but often converges prematurely to local optima [5]. On the other hand, GA, inspired by natural selection and evolution, is adept at exploring the global solution space but can also suffer from premature convergence [6]. These two algorithms have been applied to various classification tasks, including medical diagnoses. For example, Saputra et al. [6] showed that PSO could effectively optimize SVM for heart disease classification, demonstrating a significant improvement in accuracy. Similarly, Awalullaili et al. [7] applied GA for hypertension classification and highlighted its potential for optimizing SVM performance. These studies emphasize the importance of choosing the right optimization technique based on the dataset and classification problem.

Recent research has focused on improving CKD diagnosis using advanced optimization techniques. Sawhney et al. [8] investigate AI models for CKD prediction and stress the critical role of hyperparameter optimization. Jerlin Rubini and Perumal [9] proposed a multi-kernel SVM combined with the Fruit Fly Optimization Algorithm (FFOA), achieving significant improvements in CKD classification accuracy, demonstrating the effectiveness of hybrid optimization methods. Aprilianto [10] applied binary PSO with feature selection for CKD diagnosis, showing that this combination could further enhance classification performance. In addition, Poonia et al. [11] and Rubini and Perumal [12] explored hybrid optimization methods such as Grey Wolf Optimization (GWO) integrated with SVM, improving model performance in CKD classification. Sahu [13] introduced a combination of GA and Principal Component Analysis (PCA) for dimensionality reduction, which further enhanced the accuracy of CKD classification. Additionally, Lambert and Perumal [14] and Balakrishnan [15] demonstrated the effectiveness of feature selection techniques like Teaching Learning-Based Optimization (TLBO) in enhancing CKD diagnosis, particularly by refining the feature space and improving model generalization.

Despite the promising results from previous studies, the optimal method for improving CKD diagnosis accuracy through SVM hyperparameter optimization using PSO and GA remains unclear. Therefore, this study aims to compare the performance of these two algorithms in optimizing SVM hyperparameters for CKD classification. The ultimate goal is to provide more reliable diagnostic tools that improve the accuracy and efficiency of CKD detection.

METHODS

This study uses a quantitative approach to optimize the classification of Chronic Kidney Disease (CKD) diagnosis using Support Vector Machines (SVM) enhanced by metaheuristic algorithms. The study is based on secondary data from the Kaggle website,

specifically a CKD dataset created by Akshay Kumar Singh, published on February 2, 2020. The research focuses on patients diagnosed with CKD, with the target population comprising CKD patients, and the sample limited to CKD patients in India. The independent variables (X) include educational background, medical history, and laboratory results, while the dependent variable (Y) is the CKD diagnosis outcome.

The research follows several steps. First, the relevant problems are identified based on the literature, followed by an exploration of data processing techniques. The CKD dataset is preprocessed to remove irrelevant variables and handle missing values. The data is then divided into training (80%) and testing (20%) sets for model evaluation. Two metaheuristic optimization algorithms, Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), are applied to optimize the hyperparameters of the SVM model, specifically C and γ. The PSO algorithm is inspired by the social behavior of bird flocks and is initialized by setting ranges for C, determining the number of particles, setting maximum particle speed, and configuring psi 1 and psi 2 values to control individual and social influences. In SVM, γ controls how far the influence of individual data points spreads in the RBF kernel, with higher values focusing on nearby points (risking overfitting) and lower values spreading influence more broadly (risking underfitting). The fitness function evaluates model performance based on accuracy. The particle positions are updated iteratively, based on velocity, individual best positions, and the global best position, continuing until convergence or the maximum number of generations is reached.

Simultaneously, GA-SVM optimization is performed. GA is inspired by natural selection and involves setting the population size, mutation probabilities, and crossover probabilities. The population is initialized with random individuals representing different C and γ values. The fitness of each individual is evaluated using the same accuracy-based function. The selection, crossover, and mutation operators help evolve the population toward an optimal solution, with the process repeating until convergence or the maximum number of generations is reached. Both optimization techniques aim to reduce error rates and improve CKD diagnosis accuracy.

Data collection relies on secondary data from Kaggle, and data analysis uses Python libraries such as scikit-learn for SVM implementation and performance evaluation. Key metrics, including accuracy, precision, recall, and confusion matrices, are used to compare the performance of SVM optimized with PSO versus SVM optimized with GA. This approach aims to determine which optimization technique offers better performance in classifying CKD data.

The findings are expected to contribute to medical diagnostics by providing a more accurate and efficient method for CKD diagnosis through optimized machine learning models. The use of metaheuristic algorithms is anticipated to enhance SVM's robustness, offering better support for medical professionals in early CKD detection and classification.

RESULTS AND DISCUSSION

Data Classification Using SVM-PSO Algorithm

The classification process with the Google Collaboratory program uses several packages. The data preprocessing stage requires numpy which provides numerical computing operations for arrays and matrices. Storing and managing tabular data in the form of dataframes requires pandas which is capable of analyzing data structures. Package sklearn provides various machine learning tasks such as SVM as a classification algorithm, model selection, and evaluation in the form of confusion matrix and

accuracy_score. The classification algorithm is responsible for improving accuracy and reducing error. In the case of unbalanced data, the algorithm will focus on the majority class, namely the diagnosis of patients with CKD. This results in high accuracy because the majority class is given more attention, while the minority class is ignored.

The error rate using standard SVM is relatively high at 25.1%, so PSO was integrated to optimize the hyperparameters of SVM. Defining variables and functions is done first: Initially populate and function(x). The populate variable with a range of x_1 and x_2 is used to create an initial population as the starting point of the simulation to be run. The function generates random values that are Uniformly distributed (0, 1) between x_1 and x_2.

The classification process results in table 1 get an accuracy of 97.54% and γ of 0.00367. The classification error rate of the SVM-PSO model is 2.46%, indicating a high performance. A total of 289 patients were predicted positive and actually had CKD. 306 patients were predicted to be negative and actually did not have CKD. It can be seen in Table 1 that the model still predicts the wrong class, such as patients who do not have the disease are predicted as positive and vice versa.

Table 1. Confusion matrix results on SVM implementation

	Diagnosis Yes	**Diagnosis No**
Predict Yes	289	8
Predict No	7	306

Data Classification Using SVM-GA Algorithm

The next optimal search algorithm in SVM is the genetic algorithm, which is inspired by the process of biological evolution to gradually find better solutions in each generation. Classification using the SVM-GA algorithm is also carried out through several stages, namely the data import process as well as data preprocessing. The application of SVM-GA can be two processes, namely GA negative which looks for minimum values and GA positive which looks for maximum values. This research produces the best accuracy by using GA negative which gets a smaller value of γ.

The results of ten generations of observations show the highest fitness value of each individual in all generations. In applying the SVM-GA algorithm to classify CKD patient data, an accuracy result of 97.37% was obtained, which means that the SVM-GA algorithm is able to predict the class with an error rate of 2.63%, and Table 4.6 is presented for the confusion matrix. A total of 308 patients were predicted negative and did not have the disease, while 286 were predicted positive and actually had the disease. Similar to the previous two classification algorithms, SVM-GA also still predicted the wrong class. When comparing the accuracy value with SVM, the SVM-GA algorithm is far superior. Meanwhile, there is a slight difference between the accuracy value of SVM-GA and the SVM-PSO algorithm, which is 0.17%.

Table 2. Confusion matrix results on SVM-GA implementation

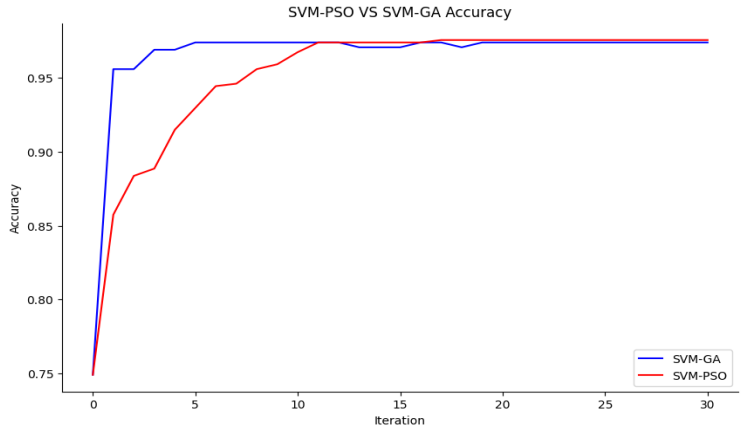
	Diagnosis Yes	**Diagnosis No**
Predict Yes	286	11
Predict No	5	308

Comparison of SVM-PSO and SVM-GA Classification Results

In this study, three algorithms have been applied for the classification of Chronic Kidney Disease patient data. Initially, the value of C was left at the SVM default, while the first optimization was performed on the value that produced good results. The C value was not further optimized and remained the default value in each classification algorithm. The quality of the classification model is measured through the accuracy results, the value of, and the confusion matrix. After integrating PSO and GA, the accuracy improved significantly compared to the standard SVM without the use of metaheuristic algorithms.

Figure 1. The dynamics of accuracy change at each iteration of the normal view

Based on Table 3, it can be seen that the SVM-PSO and SVM-GA algorithms only have small differences in accuracy. However, the most superior algorithm in classifying CKD patient



diagnosis data is SVM-PSO, with an accuracy of 97.54% and a γ value of 0.00367. The difference in value is 0.001 compared to the value in SVM-GA which achieves an impact on accuracy. Value optimization with PSO and GA allows for finer class separation, so models with smaller values are able to distinguish data classes more accurately. This is reflected in the lower scores and higher accuracy of SVM-PSO compared to SVM-GA.

Table 3. Classification algorithm performance comparison results

Performance	SVM	SVM-PSO	SVM-GA
Accuracy	74,91%	97,54%	97,37%
Value γ	0,00667	0,00367	0,00467
TP	230	289	286
FP	67	8	11
FN	86	7	5
TN	227	306	308

Based on the accuracy values generated from both algorithms, SVM-PSO (Support Vector Machine with Particle Swarm Optimization) is superior to SVM-GA (Support Vector Machine with Genetic Algorithm) in terms of accuracy. The X-axis in Figure 1 represents the number of iterations or training cycles, while the Y-axis shows the accuracy of the models. Both models demonstrate a sharp increase in accuracy during the initial iterations, indicating that the optimization processes for hyperparameter tuning are effective early on in improving the model's performance.

The SVM-PSO model reaches a high level of accuracy more quickly than SVM-GA, as evident from the steeper rise in its accuracy curve. This suggests that PSO is more efficient in finding optimal or near-optimal hyperparameters in the early stages. As the iterations progress, both models exhibit convergence, meaning the accuracy stabilizes and does not

fluctuate significantly. This indicates that the models have identified hyperparameters close to their optimal values, and further iterations do not lead to substantial improvements. Overall, while both optimization algorithms show similar performance after convergence, SVM-PSO achieves a slightly higher and faster accuracy improvement compared to SVM-GA.

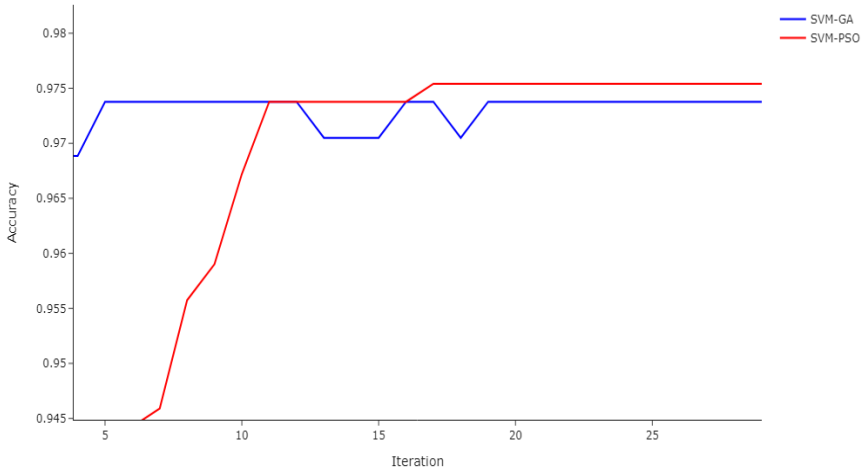


Figure 2. The dynamics of accuracy change at each iteration of the zoomed-in view

The comparison of the two algorithms' performance revealed that SVM-PSO achieved both higher accuracy and faster convergence, reaching optimal performance at the 17th iteration, while SVM-GA converged slightly later, at the 19th iteration. Although the difference in accuracy between the two models is small, it is essential to explain why PSO was more effective in optimizing the SVM hyperparameter γ. The PSO algorithm benefits from its ability to balance exploration and exploitation, as particle updates are influenced by both individual best positions and the global best. This consistent update mechanism allows PSO to maintain a steady improvement in accuracy over iterations, as seen in Figure 2, where PSO's accuracy curve steadily increases without significant fluctuations.

In contrast, GA relies on random crossover and mutation operations, which introduce more variability in each generation. This variability can cause temporary drops in accuracy (as shown by the fluctuating GA performance in Figure 2) before converging to a stable solution. These fluctuations slow down the overall convergence of GA, explaining why it required two additional iterations to achieve a near-optimal solution.

The slight difference in the convergence rate also reflects how PSO is generally more efficient in fine-tuning hyperparameters like γ, especially in cases where the solution space is complex but well-structured, such as in the classification of CKD patient data. The faster convergence of SVM-PSO can be attributed to its deterministic particle updates, which prevent it from being affected by the same level of randomness as GA, making PSO better suited for this type of optimization problem.

Despite these differences, both PSO and GA proved to be effective in enhancing the accuracy of the SVM model. The results confirm that hyperparameter optimization is crucial for improving the performance of machine learning models, particularly in medical data classification tasks like CKD diagnosis. However, SVM-PSO outperforms SVM-GA in terms of convergence speed and stability, making it the more efficient choice for this specific application.

Mathematical Aspects of the SVM-PSO and SVM-GA Algorithms

The Support Vector Machine (SVM) is a supervised learning model used for classification and regression tasks. The core of SVM's mathematical formulation revolves around finding the optimal hyperplane that maximizes the margin between two classes of data points. This margin is defined by support vectors, which are the critical data points closest to the decision boundary. Mathematically, the optimization problem in SVM can be written as follows.

$$min_{w,b}\frac{1}{2}\|w\|^2 \text{ subject to } y_i(w \cdot x_i + b) \geq 1, \forall i$$

where w is the weight vector, b is the bias term, x_i represents the input features, and y_i are the class labels. The hyperparameters C (penalty for misclassification) and γ (the width of the Gaussian kernel in non-linear classification) play significant roles in controlling the trade-off between maximizing the margin and minimizing classification errors. Optimizing these hyperparameters is critical for improving the SVM's performance, and this is where metaheuristic algorithms like Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) are applied.

Particle Swarm Optimization (PSO) in SVM

PSO is a nature-inspired optimization algorithm that mimics the social behavior of bird flocks. Each particle in the swarm represents a candidate solution, and the fitness of each particle is evaluated based on the objective function, which in this case is the accuracy of the SVM model. The position of each particle, p_i, is updated based on its velocity, which is influenced by both the particle's own historical best position and the global best position of the entire swarm. The mathematical equations for updating velocity v_i and position p_i in PSO are given by follows.

$$v_i^{t+1} = wv_i^t + c_1r_1\left(p_{best,i} - p_i^t\right) + c_2r_2(g_{best} - p_i^t)$$
$$p_i^{t+1} = p_i^t + v_i^{t+1}$$

Where w is the inertia weight, c_1 and c_2 are cognitive and social coefficients, r_1 and r_2 are random numbers uniformly distributed between 0 and 1, $p_{best,i}$ is the particle's best-known position, and g_{best} is the global best-known position.

Genetic Algorithm (GA) in SVM

The Genetic Algorithm (GA) is another evolutionary optimization technique inspired by the process of natural selection. In GA, a population of candidate solutions (individuals) is evolved through the operations of selection, crossover, and mutation. The fitness of each individual is evaluated based on the accuracy of the SVM model, and better-performing individuals are more likely to be selected for reproduction. The key mathematical steps in GA are given by follows.

1. Selection: Individuals are selected based on their fitness value, usually using methods like roulette wheel selection or tournament selection.
2. Crossover: Two selected individuals exchange portions of their genetic information (in this case, the values of C and γ) to produce offspring.
3. Mutation: Random alterations are applied to the offspring's genes (hyperparameter values) to maintain diversity in the population and avoid premature convergence.

The iterative process of selection, crossover, and mutation continues until the fitness function stabilizes or the maximum number of generations is reached. In this study, the SVM-GA model achieved an accuracy of 97.37% and an error rate of 2.63%, with a slightly higher γ value of 0.00467.

Convergence and Optimization Efficiency

The convergence behavior of both algorithms was compared in this study. Figure 1 from the results shows that the SVM-PSO algorithm converged faster, reaching its optimal accuracy at the 17th iteration, while SVM-GA required 19 iterations to stabilize. This difference in convergence can be attributed to the nature of the algorithms. PSO tends to provide smooth and continuous updates to the hyperparameters, whereas GA introduces more randomness through mutation and crossover, leading to more fluctuating behavior in fitness improvement. Mathematically, the faster convergence of PSO can be explained by its reliance on both individual and global best positions, allowing it to exploit known good solutions while still exploring the solution space. In contrast, GA's dependence on crossover and mutation introduces more variance in each generation, which, while maintaining diversity, can slow down convergence towards the global optimum.

Comparison of Performance Metrics

The confusion matrix results of both models reveal that SVM-PSO and SVM-GA performed similarly in terms of classification accuracy, with SVM-PSO slightly outperforming SVM-GA (97.54% vs. 97.37%). However, the mathematical implications of these small differences in accuracy can be significant in medical diagnosis applications. The improved separation of classes achieved by PSO suggests finer optimization of the decision boundary in the feature space, as reflected in the lower false positive and false negative rates. Moreover, the difference in the optimized γ values between SVM-PSO and SVM-GA (0.00367 vs. 0.00467) indicates that PSO was able to achieve a finer balance between overfitting and underfitting the training data. This aligns with PSO's stronger exploitation capabilities, which can find hyperparameter values that yield better generalization performance.

CONCLUSION

Based on the research results, both SVM-PSO and SVM-GA are effective for classifying Chronic Kidney Disease (CKD) diagnosis data, with hyperparameters optimized iteratively. However, SVM-PSO, achieving an accuracy of 97.54%, outperforms SVM-GA with an accuracy of 97.37%. The superiority of PSO lies in its ability to balance exploration and exploitation more efficiently, leading to smoother convergence and more consistent accuracy improvements. This allows PSO to better navigate the solution space and avoid premature convergence, making it particularly effective for optimizing complex, high-dimensional datasets like CKD. Consequently, SVM-PSO is better suited for CKD classification due to its robust and efficient optimization process.

REFERENCES

[1] Lameire, N. H., Levin, A., Kellum, J. A., Cheung, M., Jadoul, M., Winkelmayer, W. C., ... & Srisawat, N. (2021). Harmonizing acute and chronic kidney disease definition and classification: report of a Kidney Disease: Improving Global Outcomes (KDIGO) Consensus Conference. *Kidney international*, *100*(3), 516-526. https://doi.org/10.28996/2618-9801-2023-1-11-25

[2] Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare* (Vol. 10, No. 3, p. 541). MDPI. https://doi.org/10.3390/healthcare10030541

[3] Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. *Clinical pharmacology & therapeutics*, *107*(4), 871-885. https://doi.org/10.1002/cpt.1796

[4] Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, *24*, 100631. https://doi.org/10.1016/j.imu.2021.100631

[5] Ma, F., Sun, T., Liu, L., & Jing, H. (2020). Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Generation Computer Systems*, *111*, 17-26. https://doi.org/10.1016/j.future.2020.04.036

[6] Saputra, D., Dharmawan, W. S., & Irmayani, W. (2022). Performance Comparison of the SVM and SVM-PSO Algorithms for Heart Disease Prediction. *International Journal of Advances in Data and Information Systems*, *3*(2), 74–86. https://doi.org/10.25008/ijadis.v3i2.1243

[7] Awalullaili, F. O., Ispriyanti, D., & Widiharih, T. (2023). Klasifikasi Penyakit Hipertensi Menggunakan Metode SVM Grid Search dan SVM Genetic Algorithm (GA). *Jurnal Gaussian*, *11*(4), 488–498. https://doi.org/10.14710/j.gauss.11.4.488-498

[8] Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. *Decision Analytics Journal*, *6*, 100169. https://doi.org/10.1016/j.dajour.2023.100169

[9] Jerlin Rubini, L., & Perumal, E. (2020). Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm. *International Journal of Imaging Systems and Technology*, *30*(3), 660-673. https://doi.org/10.1002/ima.22406

[10] Aprilianto, D. (2020). SVM optimization with correlation feature selection based binary particle swarm optimization for diagnosis of chronic kidney disease. *Journal of soft computing exploration*, *1*(1), 24-31. https://doi.org/10.52465/joscex.v1i1.1

[11] Poonia, R. C., Gupta, M. K., Abunadi, I., Albraikan, A. A., Al-Wesabi, F. N., & Hamza, M. A. (2022, February). Intelligent diagnostic prediction and classification models for detection of kidney disease. In *Healthcare* (Vol. 10, No. 2, p. 371). MDPI. https://doi.org/10.3390/healthcare10020371

[12] Rubini, L. J., & Perumal, E. (2020). Hybrid kernel support vector machine classifier and grey wolf optimization algorithm based intelligent classification algorithm for chronic kidney disease. *Journal of Medical Imaging and Health Informatics*, *10*(10), 2297-2307. https://doi.org/10.1166/jmihi.2020.3177

[13] Sahu, S. K. (2021). Novel ensemble model with genetic algorithm and principal components analysis for classification of chronic kidney disease. *International Journal of Applied Evolutionary Computation (IJAEC)*, *12*(4), 1-17. https://doi.org/10.4018/ijaec.2021100101

[14] Lambert, J. R., & Perumal, E. (2021). Optimal feature selection methods for chronic kidney disease classification using intelligent optimization algorithms. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, *14*(9), 2886-2898. https://doi.org/10.2174/2666255813999200818131835

[15] Balakrishnan, S. (2020). Feature selection using improved teaching learning based algorithm on chronic kidney disease dataset. *Procedia Computer Science*, *171*, 1660-1669. https://doi.org/10.1016/j.procs.2020.04.178