LONTAR KOMPUTER VOL. 10, NO. 2 AUGUST 2019 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2019.v10.i02.p06 **e-ISSN 2541-5832**
Accredited B by RISTEKDIKTI Decree No. 51/E/KPT/2017

SVM Optimization Based on PSO and AdaBoost to Increasing Accuracy of CKD Diagnosis

Amanah Febrian Indriani[a1], Much Aziz Muslim[a2]

[a]Department of Computer Science, Universitas Negeri Semarang
Semarang, Indonesia
[1]amanahfebrian@students.unnes.ac.id
[2]a212muslim@yahoo.com

Abstract

Classification is data mining techniques which used for the purposes of diagnosis in the medical field as measured by the high accuracy produced. The accuracy of classification algorithm is influenced by the use of features and dimensions in dataset. In this study, Chronic Kidney Disease (CKD) dataset was used where the data is one of the high dimension datasets. Support Vector Machine (SVM) algorithm is used because its ability to handle high-dimensional data. In the dataset, it consists of 24 attributes and 1 class which if all are used results accuracy of classification will be diminished. Method for selecting features with Particle Swarm Optimization (PSO) is applied to reduce redundant features and produce optimal features. In addition, ensemble AdaBoost also applied in this research to increase performance of entirety classification algorithm. The results showed that the optimization of SVM algorithm by using PSO as a selection and ensemble feature of AdaBoost with an average of selected features of 18 features could increase the accuracy of 36.20% to 99.50% in the diagnosis of CKD compared to the SVM algorithm without optimization only resulting in accuracy 63.30%. This research can be used as a reference for further research in focusing on the preprocessing stage.

Keywords: *Data Mining, Support Vector Machine, Particle Swarm Optimization, AdaBoost, Chronic Kidney Disease*

1. Introduction

Currently, from various sources data can be collected and become very large. If this data is not utilized, it will only become a pile of useless data. Large and hidden databases can be extracted into useful knowledge with data mining techniques [1] [2]. Therefore, data mining can be considered as a tool to obtain knowledge from raw data, and the data that does not mean in the medical field. There are 3 stages of data mining, namely: data processing, data modeling, and processing of data posts. In data modeling, data mining tasks are divided into two, namely: predictive/classification algorithms and regression algorithms that are learned through a supervised learning process [3]. From a patient's medical record, data mining can be used to predict disease with classification [4].

In the medical field, there are two types of kidney failure, namely Chronic and Acute Kidney Disease which occurs when the kidneys cannot filter waste from the blood [5]. Patients with CKD are increasing as the population grows rapidly throughout the world, even within 10 years, the Global Burden of Disease notes that Chronic Kidney Disease (CKD) disease rises 9 ranks from the initial rank 27 to 18th place [6]. Medical examinations performed on patients produce very large data. However, in a very large volume of data, there are still some missing data, therefore good classification techniques are needed and produce high accuracy for detecting chronic kidney disease based on datasets [7].

There are several classification techniques in data mining include Neural Network (NN), Decision Tree (DT), Logistic Regression (LR), Naïve Bayesian (NB), and Support Vector Machine (SVM) [8]. SVM is a learning machine that utilizes the space of linear function hypotheses in high dimensional feature space, based on optimization theory obtained from statistic learning theory [9]. SVM has the concept of looking for hyperplane based on the best vector support and margins that function as the boundary of two classes and have been successfully applied to many classification cases with high accuracy [10].

To increase the accuracy of the classification algorithm, an ensemble technique used to combining several weak classifiers using the AdaBoost algorithm [11]. One of the most promising algorithms with convergence fast and easy to implement is Adaboost, because AdaBoost does not require knowledge from the weak learner and can be easily combined with other methods such as SVM [12].

Another way to increase accuracy is by selecting features at the preprocessing stage. Feature selection is a data preprocessing step that is used to delete some features in a data set so that the process runs faster, and data visualization is easier [13]. Feature selection methods usually implicate heuristic or random search strategies to avoid complexity [14]. PSO is a heuristic algorithm that has been proven to provide optimization of value [15]. In some cases, it has been proven that PSO is more competitive when compared to genetic algorithms to overcome the feature selection problem [16].

The goals of this study were to improve the accuracy of the SVM algorithm that had been optimized using the PSO algorithm as an Adaboost selection and ensemble feature in the diagnosis of CKD.

2. Reseach Methods

The combination of several proposed algorithms aims to improve the accuracy of the diagnosis of Chronic Kidney Disease. Steps that will be carried out in this experiment include preprocessing, feature selection using PSO, and SVM classification with AdaBoost ensemble.

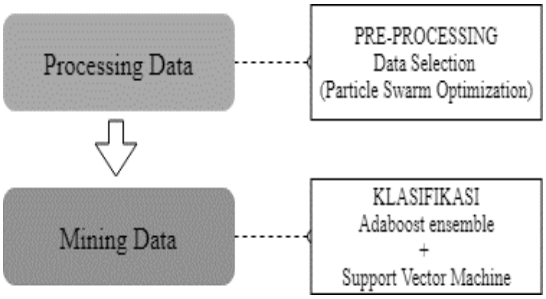

Figure 1. Research Method

The work step in this study began by inputting the CKD dataset. Then the data will be processed in the data preprocessing stage, was by cleaning data. Data cleaning is done by removing the missing value in the CKD dataset. Still, in the preprocessing stage, the data that has been filled in with the missing value will then be feature selection using the PSO algorithm. Based on the selected features, the classification process will be carried out with the SVM algorithm combined with AdaBoost. Then the classification model is tested using data testing and evaluated using a confusion matrix to produce accuracy values. The flowchart of the research method carried out in Figure 2.

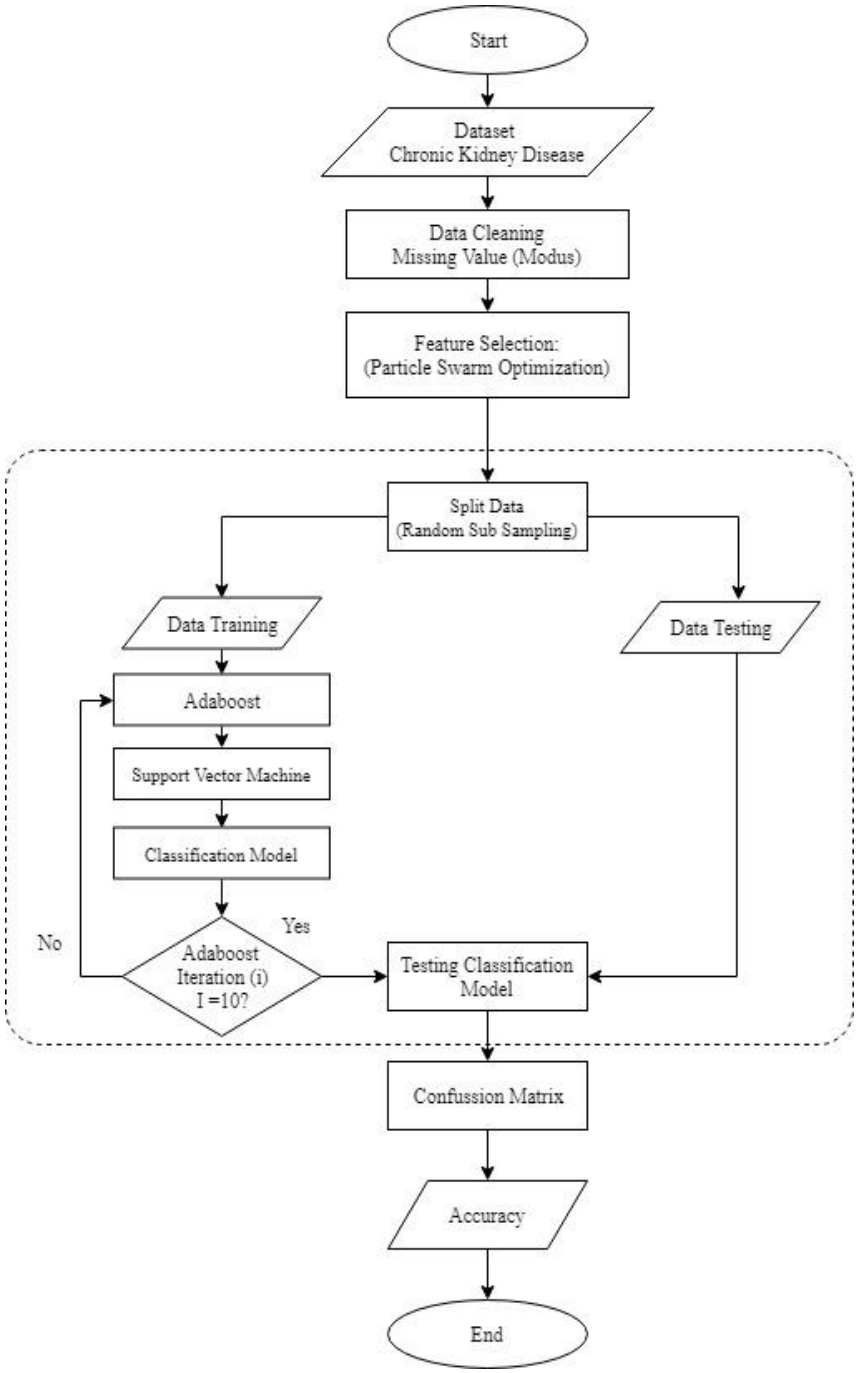

Figure 2. Flowchart Support Vector Machine Algorithm with PSO and Adaboost

2.1. Preprocessing

The dataset used in this study is a CKD dataset that was collected and uploaded by the Apollo hospital, India in 2015 at the UCI Machine Learning Repository. The collected data amounts to 400 instances and 25 attributes consist of 11 numeric attributes and 14 nominal attributes. The attribute description of the CSD dataset can be seen in Table 1.

Tabel 1. Description of Chronic Kidney Disease Dataset

Features	Type
Age (Age)	Numeric
Blood pressure (Bp)	Numeric
Appetite (appet)	Nominal
Specific gravity (Sg)	Nominal
Albumin (Al)	Nominal
Sugar (Su)	Nominal
Red blood cells (Rbc)	Nominal
Pus cell (Pc)	Nominal
Pus cell clumps (Pcc)	Nominal
Bacteria (Ba)	Nominal
Hypertension (Htn)	Nominal
Haemoglobin (Hemo)	Numeric
Serum creatinine (Sc)	Numeric
Blood glucoses (Bgr)	Numeric
Blood urea (Bu)	Numeric
Coronary artery disease (Cad)	Nominal
Sodium (Sod)	Numeric
Potassium (Pot)	Numeric
Pedal edema (Pe)	Nominal
Packed cell volume (Pcv)	Numeric
White blood cell (Wbcc)	Numeric
Red blood cell count (Rc)	Numeric
Diabetes mellitus (Dm)	Nominal
Anemia (Ane)	Nominal
Class (class)	Nominal

In the preprocessing stage, nominal type attributes will be transformed into numeric. There are a number of 14 nominal attributes that will be transformed into numeric type attributes. Then, of the 400 instances in the CKD dataset, 250 of them are labeled with the *ckd* class while the other 150 are labeled the *notckd* class. In the CKD dataset, there are more than 50% of the missing value, so it is necessary to handle missing values to produce higher accuracy. Filling in the missing value is done by the mode method, namely by replacing the empty value with the most frequency of each attribute.

2.2. PSO for Feature Selection

The PSO algorithm in the selection of features tries to get the best composition of features in a problem space. PSO has the ability to get the optimal subset by finding the best position around the local position and global position [17].

Although PSO was initially introduced to optimize real number problems, now PSO can also display discrete or qualitative differences between variables, it is called Binary Particle Swarm Optimization (BPSO). In BPSO, each particle will be represented in binary variables 0 or 1. Then, velocity is transformed into a change in probability, that is, the probability of a binary variable takes a value of 1. However, the velocity must be limited to the range [0,1] [18].

There are stages in the BSPO algorithm as feature selection by initializing random positions and velocities of particles. Then, the fitness value of each particle in the population will be evaluated. After that, unite if the fitness value of particle i is less than *pBest* value, then *pBest* from particle i to particle position i, but if *pBes*t is updated and fitness value is less than current *gBest* value, set *gBest* to *pBest* at this time from particle i. Then, update the speed and position of the

particle. If the best fitness value or iteration is fulfilled if it has not returned to the fitness calculation stage then stop iteration [19].

2.3. Adaboost Ensemble

Ensemble learning usually consists of several basic learning algorithms that are usually generated from training data. Ensemble methods are widely used because they can improve the basic learning algorithm and make highly accurate predictions [20]. The Ensemble method can be used to improve overall accuracy by studying and combining a series of individual classifier models [21].

Adaptive boosting (AdaBoost) is one of several variants in the boosting algorithm. AdaBoost is an Ensemble of learning that is often used in boosting algorithms [22]. Adaboost and its variants have been successfully applied in several fields because of its strong theoretical basis and great simplicity. The steps of the Adaboost algorithm are [23]:

a. Input: A collection of training samples with labels *{(x_i,y_i),…,(x_N, y_N)}*, a basic learning algorithm, the number of T turns.
b. Initialize: Weight of a training sample w = 1 / N, for i = 1, ..., N.
c. Do for t = 1, ..., T.
   1) Use the basic learning algorithm to train a classification component, ht, on the training weight sample
   2) Calculate the training error at h_t: $\varepsilon_t = \sum_{i=1}^{N} w_i^t , y_i \neq h_t(x_i)$
   3) Set the weight for component classifier $h_t = \alpha_t = \frac{1}{2} \ln (\frac{1-\varepsilon_t}{\varepsilon_t})$
   4) Update the training sample weight $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(x_i)\}}{c_t}$, i = 1, ..., N C_t is a normalization constant.
d. Output $f(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$ to make predictions using the last model.

Therefore, the core of the iterative AdaBoost process is iteratively AdaBoost by in circles, updating the sample to find the best weak classifier distribution at the moment, and then calculate the error rate of each weak classifier, and finally build a weak classifier into a strong classifier several times [24].

2.4. Support Vector Machine Classification

Classification is the process of classifying a collection of objects, data or ideas into groups, where each member has one of the same characteristics. In classification, classes cannot be contested before examining data so that it is often called supervised learning [25].
SVM is used for linear and nonlinear data classifications. SVM with nonlinear mapping functions to convert the original training data into higher dimensions, this is done when the data is not linearly separated. Data from 2 classes separated by hyperplane found by SVM uses margin and support vector [26].

SVM uses kernel tricks to connect training sample input space to high dimensional feature space and identify optimal separator hyperplane. The RBF (Radial Basis Function) kernel with gamma parameters is used. To control the complexity of the model and training errors, regulatory parameter C is used. Choosing the right gamma and C values, solving the problem of overfitting. A low parameter C value makes a smooth decision, while a high C goals to classify all training samples correctly. The function of the SVM decision for binary classification problems is defined as follows [27].

$$f(x) = [w, \varphi(x)] + b \quad (1)$$

Mapping sample x from input space to high dimensional feature space is represented by φ (x). Dot product in the feature space is displayed as [..., ...]. The ideal value w and b are achieved by doing the following optimization.

Minimize : $g(w,\varepsilon) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\varepsilon_i$ (2)

Subject to : $y_i([w,\varphi(x_i)] + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0$ (3)

Where ε_i is a variable slack.

$$k(x_i, x_j) = [\varphi(x_i), \varphi(x_j)] \quad (4)$$

The kernel function k (xi, xj) is used to map input vectors non-linearly to the appropriate feature space using the RBF function.

3. Result and Discussion

In this study, the proposed algorithm was tested using the Python programming language by utilizing a little-known library and the Pyswarms library. This experiment was carried out 5 times with the provisions of the PSO parameters shown in Table 2 as follows.

Table 2. PSO Parameter setting

Parameter	Values
Swarm size	30
Cognitive parameters (c_1)	2
Social parameters (c_2)	2
Inertia weight	1
Number of iteration	100

The parameters of SVM in this study are arranged as follows. Parameters C = 0.1 and gamma parameters = 1 / number of features. While for AdaBoost in this study, 10 iterations will be conducted. Data collection is done randomly with ratio 3:7 for each data testing: training data. The comparison is taken because it can produce high accuracy.

The application of the PSO algorithm as a feature selection in this study after performing 5 times the execution produces a feature that is not always the same for each execution. Because in each execution there are several differences in the features selected, it produces different accuracy. After feature selection, classification process was carried out using the SVM algorithm. In this research, ensemble Adaboost was applied to improve the accuracy of the SVM classification algorithm. The results of the feature selection process with PSO and the accuracy of each SVM execution with the application of PSO as feature selection without ensemble adaboost and with the application of ensemble adaboost can be seen in Table 3 as follows.

Table 3. The Result of Feature Selection PSO

Execution	Feature Set	Total Feature	Accuracy	
			PSO + SVM	PSO + SVM + *Adaboost*
1	0 1 1 1 1 1 1 1 1 0 0 1 0 1 1 0 0 1 1 1 1 1 1 1	18	97,25%	100%
2	0 0 1 1 1 1 1 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 1 1 1	18	98,75%	99,16%
3	0 1 1 1 1 1 1 1 1 1 0 0 1 0 1 1 0 0 1 1 1 1 1 1 1	18	97,25%	100%
4	0 0 1 1 1 1 1 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 1 1 1	18	98,75%	99,16%
5	0 0 1 1 1 1 1 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 1 1 1	18	98,75%	99,16%
Average			**98,15%**	**99,50%**

Information:
Selected features : 1
Unselected feature : 0

From the experimental data, it can be seen that the application of PSO as a feature selection can produce a high level of accuracy, besides that adaboost ensemble application can also increase the accuracy of the SVM + PSO classification, which is increased by 1.35% compared to before added Adaboost. So that it can be seen the accuracy comparison of the SVM classification method without optimization and SVM after being optimized using PSO as a feature selection and Adaboost ensemble. Comparison of the results of accuracy can be seen in Table 4 as follows.

Table 4. Accuracy Result of Feature Selection PSO

Algorithm	Accuracy
SVM	63,30%
PSO + SVM	98,15%
PSO + SVM + Adaboost	99,50%

From this study, the classification with the SVM algorithm obtained an accuracy of 63.30% while the classification with the SVM algorithm optimized by PSO and ensemble Adaboost algorithms produced an average accuracy of 99.50%. By applying the algorithm PSO and ensemble Adaboost can increase accuracy by 36.20%. Significant accuracy increases due to the application of the PSO algorithm to select the optimal feature set in the classification algorithm SVM performs an optimal solution based on the swarm intelligence concept where each particle in the search area represents a classification process. In addition, the determination of the parameter values used in the application of the PSO algorithm also affects the selection of optimal features so that it can provide high accuracy results. Besides that, ensemble Adaboost in this combination can also improve the performance of the SVM algorithm classification for the CKD dataset or which has the same characteristics. From this study, it is known that by applying the PSO and ensemble Adaboost algorithms on the SVM algorithm it can improve the accuracy of the diagnosis of CKD so that it can be used by researchers as a reference in conducting research into the diagnosis of CKD.

4. Conclusion

Based on this research, the application of PSO and ensemble AdaBoost algorithms to optimize the SVM classification algorithm in the CKD dataset taken from the UCI Machine Learning Repository. PSO algorithm is used to get the best combination of features for the classification process, while AdaBoost is used as an ensemble method to improve SVM accuracy results as weak classifiers to become strong classifiers. The results of this study, obtained the accuracy of the application of the SVM algorithm without optimization of 63.30% while after being optimized using the PSO + AdaBoost feature selection the average accuracy increased by 36.20% to 99.50% with the selected feature average numbering 18 features. Thus, it can be concluded that the application of the PSO and ensemble AdaBoost algorithms can get optimal features and can improve the accuracy of the SVM algorithm. In future works, the spilted reduction feature can be applied to many types of the dataset with the same characteristics as the CKD dataset. It also compares with other feature algorithms to determine the impact of the model in increasing the accuracy of classifiers.

References

[1] M. H. Elhebir and A. Abraham, "A Novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification," *International Journal of Computer Information Systems and Industrial Management Applications,* vol. 7, pp. 189-195, 2015.

[2] G. A. Afzali and S. Mohammadi, "Privacy Preserving Big Data Mining: Association Rule Hiding," *Journal of Information System and Telecomunication,* vol. 4, no. 2, pp. 70-77,

2016.

[3] H. Hamidi and A. Daraei, "Analysis and Evaluation of Techniques for Myocardial Infraction Based on Genetic Algorithm and Weight by SVM," *Journal of Information System and Telecommunication,* vol. 4, no. 2, pp. 85-91, 2016.

[4] M. A. Muslim, E. Sugiharti, B. Prasetiyo and S. Alimah, "Penerapan Dizcrretization dan Teknik Bagging UNtuk Meningkatkan Akurasi Klasifikasi Berbasis Enseble pada Algoritma C4.5 dalam Mendiagnosa Diabetes," *LONTAR KOMPUTER: Jurnal Ilmiah Teknologi Informasi,* vol. 8, no. 2, pp. 135-143, 2017.

[5] L. J. Rubini and P. Eswaran, "Generating Comparative Analysis of Early Stage Prediction of Chronic Kidney Disease," *International Journal of Modern Engineering Research (IJMER),* vol. 5, no. 7, pp. 49-55, 2015.

[6] I. Fadilla, P. P. Adikara and R. S. Perdana, "Klasifikasi Penyakit Chronic Kidney Disease (CKD) Dengan Menggunakan Metode Extreme Learning Machine (ELM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN 2548:964X,* vol. 2, no. 10, pp. 3397-3405, 2018.

[7] W. Abedalkhader and N. Abdulrahman, "Missing Data Classification of Chronic Kidney Disease," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 7, no. 5, pp. 55-61, 2017.

[8] A. Widodo and S. Handoyo, "The Classification Performance using Logistic regression and Support Vector Machine (SVM)," *Journal of Theoritical and Applied Information Technology,* vol. 95, no. 19, pp. 5184-5193, 2017.

[9] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin and Y. Effendi, "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction," *LONTAR KOMPUTER: Jurnal Ilmiah Teknologi Informasi,* vol. 9, no. 3, pp. 192-201, 2018.

[10] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi),* vol. 1, no. 1, pp. 19-25, 2017.

[11] M. Mohammadpour, M. Ghorbanian and S. Mozaffari, "AdaBoost Performance Improvement Using PSO Algorithm," in *2016 Eight International Conference on Information and Knowledge Technology (IKT)*, Iran, 2016.

[12] R. Wang, "AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review," *Physics Procedia*, vol. 25, pp.800-807, 2012.

[13] D. Panday, R. C. de Amorim and P. Lane, "Feature weighting as a tool for unsupervised feature selection," *Information Processing Letters,* vol. 129, pp. 44-52, 2018.

[14] M. H. Aghdam and S. Heidari, "Feature Seletion Using Particel Swarm Optimization in Text Categorization," *Journal of Artificial Intelligence and Soft Computing Research,* vol. 5, no. 4, pp. 231-238, 2015.

[15] F. Mar'i and A. A. Supianto, "Clustering Credit Card Holder Berdasarkan Pembayaran Tagihan Menggunakan Improved K-Means dengan Particle Swarm Optimization," *Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. 5, no. 6, pp. 737-744, 2018.

[16] M. A. Muslim, A. Nurzahputra and B. Prasetiyo, "Improving Accuracy of C4.5 Algorithm Using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction," in *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2018.

[17] F. Ardjani, K. Sadouni and M. Benyettou, "Optimization of SVM MultiClass by Particle Swarm (PSO-SVM)," in *International Workshop on Database Technology and Applications*, China, 2010.

[18] L. Y. Chuang, C. H. Ke and C. H. Yang, "A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification," in *Internation MiltiConference of Engineers and Computer Scientist 2008*, Hong Kong, 2008.

[19] S. Gunasundari, S. Janakiraman and S. Meenambal, "Multiswarm Heterogeneous Binary PSO using Win-Win approach for improved Feature Selection in Liver and Kidney disease Diagnosis," *Computerized Medical Imaging and Graphics,* vol. 70, pp. 135-154, 2018.

[20] Z. H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall: CRC, 2012.

[21] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Waltham, MA: Morgan Kaufman Publisher (Elsivier), 2012.

[22] A. Nurzahputra and M. A. Muslim, "Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan AdaBoost untuk Meminimalkan Resiko Kredit," in *Prosiding SNATIF*, Kudus, 2017, pp. 243-247.

[23] E. Listiana and M. A. Muslim, "Penerapan Adaboost Untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi Pada Diagnosis Chronic Kidney Disease," in *Prosiding SNATIF*, Kudus, 2017, pp. 875-881.

[24] Y. Wang and X. Li, "Improvement of RBF Neural Network by AdaBoost Algorithm Combined with PSO," *Telkomnika,* vol. 14, no. 3A, pp. 56, 2016.

[25] S. Vijayarani and S. Dhayanand, "Data Mining Classification Algorithm for Kidney Disease Prediction," *International Journal on Cybernetics & Informatics (IJCI),* vol. 4, no. 4, pp. 13-25, 2015.

[26] A. Subasi, "Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders," *Computers in biology and medicine,* vol. 43, no. 5, pp. 576-586, 2013.

[27] U. Bhosle and J. Deshmukh, "Mammogram classification using AdaBoost with RBFSVM and Hybrid KNN–RBFSVM as base estimator by adaptively adjusting γ and C value," *International Journal Information and Technology,* pp. 1-8, 2018.