Buletin Ilmiah Sarjana Teknik Elektro
Vol. 5, No. 2, June 2023, pp. 201-209
ISSN: 2685-9572, DOI: 10.12928/biste.v5i2.7548

Implementation of Discretisation and Correlation-based Feature Selection to Optimize Support Vector Machine in Diagnosis of Chronic Kidney Disease

Dwika Ananda Agustina Pertiwi [1], Pipit Riski Setyorini [2], Much Aziz Muslim [3], Endang Sugiharti [4]
[1,2,3,4] Department of Computer Science, Universitas Negeri Semarang, Indonesia
[1,3]Postgraduate Student, Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Malaysia

ARTICLE INFORMATION

Article History:

Submitted 25 January 2023
Revised 22 May 2023
Accepted 27 May 2023

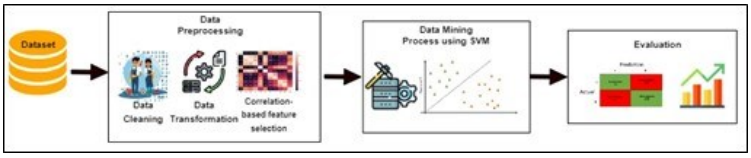
Keywords:

Support Vector Machine;
Discretization;
CFS;
Chronic Kidney Disease

Corresponding Author:

Dwika Ananda Agustina Pertiwi,
Universitas Negeri Semarang,
Sekaran, Gunungpati,
Semarang, Indonesia.
Email:
dwikapertiwi13@gmail.com

ABSTRACT



This study aims to improve the accuracy of the classification algorithm for diagnosing chronic kidney disease. There are several models of data mining. In classification, the Support Vector Machine (SVM) algorithm is widely used by researchers worldwide. The data used is a chronic kidney disease dataset taken from the UCI machine learning repository. This data consists of 25 attributes and 11 numeric data attributes, and 14 negative attributes. To call continuously, discrete data is used. Meanwhile, data is selected using Correlation-based Feature Selection (CFS) to reduce irrelevant and redundant data. The research results by applying discretization and feature selection based on correlation for classification in the SVM algorithm with 10-fold cross-validation show an increase in accuracy of 0.5%. The classification of the vector machine support algorithm in the diagnosis of chronic kidney disease produces an accuracy of 99.25%, and after applying discretization and correlation-based feature selection, produces an accuracy of 99.75%. Implementation of discretion and correlation-based feature selection to optimize support vector machine for diagnosis of chronic kidney disease has increased accuracy by 0.5%. The proposed method is feasible as a method of diagnosing chronic kidney disease.

Document Citation:

D. A. A. Pertiwi, P. R. Setyorini, M. A. Muslim, and E. Sugiharti, "Implementation of Discretisation and Correlation-based Feature Selection to Optimize Support Vector Machine in Diagnosis of Chronic Kidney Disease," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 5, no. 2, pp. 201-209, 2023, DOI: 10.12928/biste.v5i2.7548.

1. INTRODUCTION

The database is overgrowing because it is needed in every field, such as the health sector. A database is needed to store patient medical record data in the health sector. The amount of medical recorded data is increasing daily, so data accumulation exists. A collection of lots of data, if not used properly, then only becomes a collection of data that is not useful. A system is needed to process data so that important information can be obtained that can be used for the development of the field. Data processing is known as data mining. Data mining is the analysis of data that has been there before to solve a problem and is defined as finding patterns in data [1], [2]. Data mining methods study the science of extracting knowledge or finding patterns from data [3]. Data mining is the process of finding data that is not yet known by users with a model so that it can be understood and used as a basis for decision-making [4]. Some methods often mentioned in the data mining literature include clustering, classification, and others. C-Means is one of the clustering algorithms. C-Means applies fuzzy grouping, which means that each data can be a member of several clusters with various levels of membership in each cluster [5]. Besides clustering, there is another data mining model, namely classification. The classification algorithm in data mining is one of data analysis's fundamental functions. The final goal of classification is to build models that can correctly predict classes of different objects. Input for this method is training data (training data), class-owned (dependent variable), and a set of variables that describe the characteristics of objects that are different (independent variable) [6]. Classification is the process of grouping test data on classes determined based on the learning algorithm [7], [8]. In addition, classification can also be used to overcome scheduling problems, where scheduling is classified into several types of attributes that are used [9]. In the field of classification, the Support Vector Machine (SVM) algorithm is the algorithm that occupies the most number that is often used by researchers around the world [10], [11]. Support Vector Machine (SVM) is included in the Artifical Neural Network (ANN) class which in conducting the classification needs the training stages and testing stages [12]–[14].

SVM also has difficulty in distinguishing between influential and non-influential attributes in the prediction process. By using feature selection can improve the performance of the classification process which means that feature selection produces fewer features with the same accuracy or even greater than the classification results in a pure dataset [15]–[17]. Feature selection is one of the most important techniques and is often used in data mining preprocessing, especially for knowledge discovery and discovery science. This technique reduces the number of features involved in determining a target class value, reduces irrelevant, redundant features and data that causes misunderstanding of the target class [17]. Feature selection using the method of correlation-based feature selection can be obtained attributes that influence thereby eliminating redundant attributes [18]–[21]. Correlation based feature selection is the most stable tribute selection method compared to information gain, ration gain, chi-squared, symmetrical uncertainty, and relief [19], [22], [23]. The essence of correlation-based feature selection is a heuristic technique for evaluating the value or price of a feature subset. The method used to calculate the level of correlation of each variable is symmetrical uncertainty.

To simplify the original data and make it more efficient, the discretization process is carried out. Discretization is the transformation of data from continuous attributes to categorical attributes. This discretization process becomes important because good results in this process will affect the performance results of data mining algorithms [24]. Some classification and clustering algorithms do not only deal with nominal attributes and cannot handle attributes measured on a numerical scale. In general datasets, numerical attributes must first be discretized into a few different small ranges [25].

The algorithm can be used for diagnosis of diseases such as chronic kidney disease (CKD) [26]–[29]. Chronic kidney disease is a heterogeneous disorder that progressively affects the structure and function of the kidneys and is difficult to recover, where the body is unable to maintain metabolism and fails to maintain fluid and electrolyte balance resulting in increased ureum [30]–[33]. Chronic kidney disease is caused by physiological processes of pathogens with various variations that will cause a significant decrease in renal function (glands in the kidneys) which will eventually lead to kidney failure [34]–[37]. This study uses a chronic kidney disease dataset obtained from the UCI repository of machine learning. Based on the background description above, this study uses correlation-based feature selection and discretization as a technique to improve the level of classification applied to the support vector machine algorithm in the diagnosis of chronic kidney disease. So, the contribution of this research is to improve the accuracy performance of the Support Vector Machine by implementing discretization and correlation-based feature selection for the diagnosis of chronic kidney disease.

2. METHODS

The stages of the method used in this study, starting from the data preprocessing, classification, and evaluation stages, will result in increased accuracy in the support vector machine algorithm. In Figure 1 is a classification flow diagram using correlation-based feature selection and discretization in support vector

machines. In Figure 1 there are stages of research carried out starting from the collection of chronic kidney disease datasets taken from the UCI repository of machine learning, then the datasets go through the initial process, namely preprocessing, this is done so that the data is ready to be processed into classification modeling, which then from the modeling process will be evaluated using a confusion matrix. In detail, the stages of the research are presented in the sub-section below.

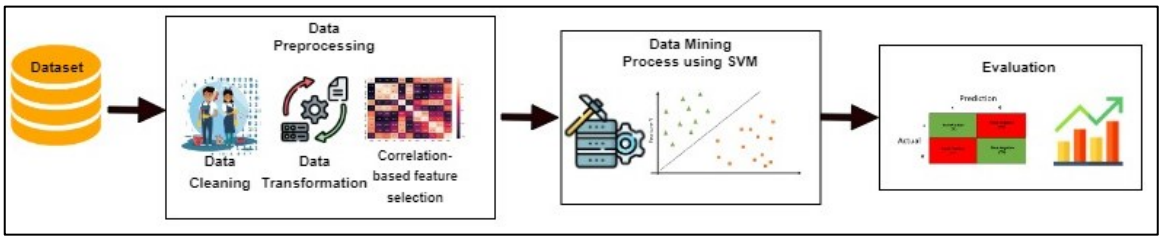


Figure 1. Flowchart of research methods

2.1. Data Preprocessing

Preprocessing data in this study includes data cleaning, data selection, and data transformation [38]. The data used is chronic kidney disease dataset taken from the UCI repository of machine learning. The number of data records in the dataset is 400 records consisting of 24 attributes and 1 class attribute. The data cleaning stage is the process of cleaning up incomplete, empty, or null data. Blank data can be filled using the average model which can replace the wrong value with the average value based on the value available on the feature, for the i-th feature. In this chronic kidney disease dataset, there is data that has a value "?" Or what is called a missing value. Calculation of the average to replace the missing value data can be shown in Equation (1), in which the variable x_i is the attribute value, and n present the number of attributes that have values.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

The discretization process is carried out at the data transformation stage using Entropy-based Discretization on numeric type attributes that are continuous. This method uses entropy as part of the continuous data hose separation process. So, numeric type attributes can be transformed into two categories represented in numeric form 0 and 1. To find the best separator value, split point value, information gain value of entropy between 2 samples and the formula must be calculated:

1. Sort the data subset from the smallest to the largest.
2. Calculate the average value per 2 contiguous data used for split points. Each average value is a point value that might be a split point (split_point) to choose the best point, the data will be broken down according to the requested point. The formula for doing split points in Equation (2).

$$Split_{point} = \frac{a_i + a_{i+1}}{2} \tag{2}$$

3. Calculate the information value of the two samples (Sa). Then T (split point) which has the smallest information value is taken as the node boundary. The formula used to find the entropy and value of information is shown by Equations (3) and (4).

$$E\ (S,T) = \frac{|S_1|}{|S|}\ Ent\ (S_1) + \frac{|S_2|}{|S|}\ Ent\ (S_2) \tag{3}$$

$$Information\ Gain = Gain\ (A, S_a) - E(S,T) \tag{4}$$

At the data selection stage, the process of minimizing the amount of data used for the data mining process is carried out. At this stage, data selection is performed to reduce irrelevant and redundant data. Irrelevant attributes are attributes that contain information that is not useful for performing data mining tasks directly, while redundant attributes are attributes that duplicate much, or all of the information contained in one or more other attributes [39]. Dimension reduction on this attribute is done by using correlation based Feature Selection (CFS) technique [40]–[42]. Deleting a variable is done on attributes that have a Symmetrical Uncertainty (SU) value that is less than the minimum value.

The first step is to calculate the entropy value of each variable. Entropy of variable X is defined in Equation (5), in which $H(x)$ is an entropy of variable X, $P\ (x_i)$ present the probability of variable X.

$$H(x) = -\sum_i P\,(x_i)\,\left(P\,(x_i)\right) \tag{5}$$

Next perform the entropy calculation of a variable by referring to other variables defined in Equation (6). in which $H(X \mid Y)$ is an entropy variable X to variable Y, and $P(y_j)$ present the probability of variable Y, then $P\,(x_i \mid y_j)$ denotes the probability of variable X to variable Y.

$$H(X \mid Y) = -\sum_j P(y_j) \sum_i P\,(x_i \mid y_j)\, log_2\, \left(P\,(x_i \mid y_j)\right) \tag{6}$$

From the entropy, the Information Gain (IG) value is calculated in Equation (7), in which $IG(X \mid Y)$ denotes the information gain of variable X to variable Y.

$$IG(X \mid Y) = H(X) - H(X|Y) \tag{7}$$

To measure the correlation between features, symmetrical uncertainty is used. Symmetrical uncertainty values range from 0 to 1. Symmetrical uncertainty is formulated in Equation (8).

$$SU\,(X,Y) = 2\,\left[\frac{IG\,(X|Y)}{H(X)+H(Y)}\right] \tag{8}$$

2.2. Data Mining Process

Data mining is the stage to find patterns or information in a set of data using certain techniques and algorithms. This study applies discretization and correlation-based feature selection to improve the accuracy of the support vector machine algorithm in the diagnosis of chronic kidney disease.

The initial step of an SVM algorithm is to define the equation of a separating hyperplane written in Equation (9), in which *w* is an a vector weight is w={w1,w2,…,wn}, *n* present the number of attributes, and *b* is an a scalar called bias.

$$w.X + b = 0 \tag{9}$$

If based on attributes A1, A2 with the example of training tuples X = (x_ (1,) x_2), x1 and x2 are values of attributes A1 and A2, and if b is considered an additional weight w_0, then the equation of a separating hyperplane can be written repeat as in Equation (10).

$$w_0 + w_1x_1 + w_2x_2 \tag{10}$$

After the equation can be defined, the values x_1 and x_2 can be entered into the equation to find the weights w_1, w_2, and w_0 or b. A graph of the separation of two data classes with maximum margins can be seen in Figure 2.

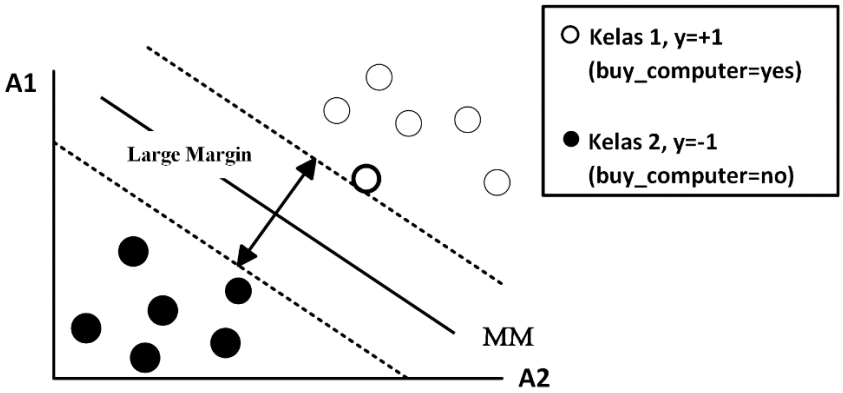


Figure 2. Separation of two data classes with Maximum Margins

In the Figure 2, SVM finds the maximum separating hyperlane, the hyperlane which has the maximum distance between the closest training tuples. Support vectors are indicated by thick limits at the tuple point. Thus, each point located above the separating hyperplane fulfills Equation (11).

$$w_0 + w_1x_1 + w_2x_2 > 0 \tag{11}$$

Meanwhile, the point located below the hyperplane separator fulfills the formula as in Equation (12).

$$w_0 + w_1 x_1 + w_2 x_2 < 0 \quad (12)$$

Looking at the two conditions above, two hyperplane equations are obtained, as in Equations (13) and (14).

$$H_1 = w_0 + w_1 x_1 + w_2 x_2 \geq 0 \quad (13)$$

$$H_2 = w_0 + w_1 x_1 + w_2 x_2 \leq 0 \quad (14)$$

The formulation of the SVM model uses mathematical tricks namely Lagrangian Formulation. Based on the Lagrangian formulation, the Maximum Maximum Hyperplane Margin (MMH) can be rewritten as a decision boundary in Equation (15), in which y_i denotes the class label of the support vector Xi, the X^T is a test tupel, and a_i and b_0 describe a numerical parameter determined automatically by the optimization of SVM algorithm, 1 is the number of vector support.

$$d\,(X^T) = \sum_i^1 y_i a_i\, X_i\, X^T + b_0 \quad (15)$$

3. RESULT AND DISCUSSION

The tool used to test research in applying discretization and correlation-based feature selection on the support vector machine algorithm in the diagnosis of chronic kidney disease in the form of a system design to obtain accuracy results based on the model using Matlab R2014a software. The workflow stages of the SVM algorithm classification by discretization and correlation-based feature selection in the diagnosis of chronic kidney disease are illustrated in Figure 3.

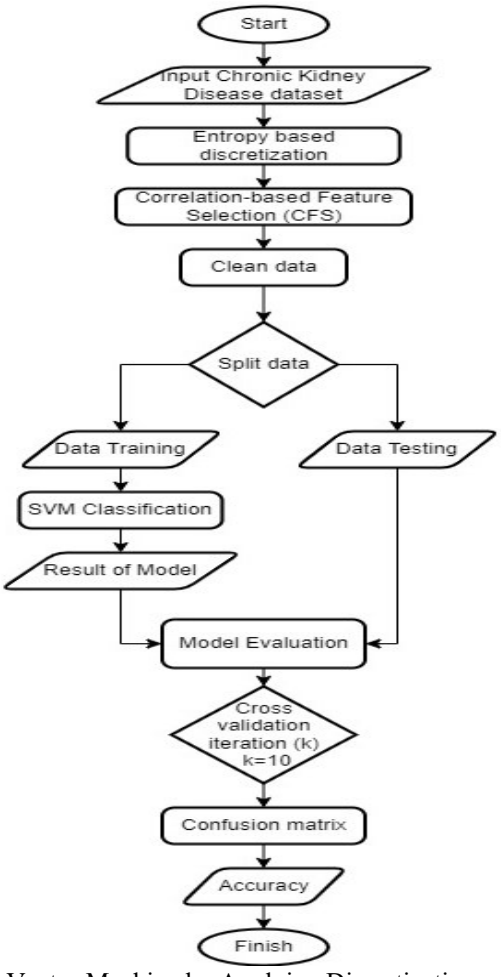


Figure 3. Flowchart Support Vector Machine by Applying Discretization and Correlation based Feature Selection

In Figure 3 is an illustration of the design classification model in this study, where SVM is optimized by discretization and correlation-based feature selection implemented in the diagnosis of chronic kidney disease. So, from the proposed modeling, the research obtained a comparison of the performance results from the step by step optimization carried out, which is SVM without optimization, and result of SVM after optimization.

In this section, the experimental results are analyzed to evaluate the performance of the data mining algorithm used. The system created applies discretization and correlation-based feature selection to the support vector machine algorithm for the diagnosis of chronic kidney disease. The data used is a chronic kidney disease dataset that has passed the data cleaning process and then the data transformation stage is performed. Nominal type data is converted into numeric type. In the continuous type of attribute, the discretization process is carried out using Entropy - based Discretization. Chronic kidney disease data that has been carried out data transformation process, then performed data selection using correlation-based feature selection. From the data that has been transformed, counted the number of cases, the number of cases for the ckd class, the number of cases for the notckd class of each attribute.

Furthermore, entropy, information gain and symmetrical uncertainty are calculated for each attribute. Data selection uses correlation-based feature selection based on the symmetrical uncertainty value of each attribute. The results of the calculation of correlation-based feature selection consisting of entropy values, information gain and symmetrical uncertainty for each attribute are shown in Table 1.

Table 1. The calculation result of correlation-based feature selection

Atribut	Entropy	Information Gain	Symmetrical Uncertainty
Age (*age)*	0.862174529	0.068847739	0.075798102
Bp (*blood pressure*)	0.741482739	0.165895584	0.195641189
Sg (*specific gravity*)	2.057428544	0.594194675	0.394569583
Al (*Albumin*)	1.929573917	0.529213900	0.366998922
Su (Sugar)	1.308742586	0.217086838	0.191842597
Rbc (*red blood cells*)	0.522131740	0.086338730	0.116945325
Pc (*pus cell*)	0.701471459	0.147642930	0.178322897
Pcc (*pus cell clumps*)	0.492328842	0.078416626	0.108402876
Ba (*Bacteria*)	0.296923323	0.036848979	0.058894415
Bgr (*blood glucoses*)	0.948733167	0.337771461	0.354957217
Bu (*blood urea*)	0.938454382	0.938454382	0.340029618
Sc (*serum creatinine*)	0.999549110	0.574500957	0.588030626
Sod (*Sodium*)	0.946755449	0.334540562	0.351927644
Pot (*Potassium*)	0.436469817	0.064832727	0.093223882
Hemo (*Haemoglobin*)	0.989587521	0.688619681	0.708448618
Pcv (*packed cell volume*)	0.994149171	0.649194450	0.666324598
Wc (*white blood cell*)	0.991254007	0.110054556	0.113126622
Re (*red blood cell count*)	0.992774453	0.633138670	0.650303944
Htn (*Hypertension*)	0.948733167	0.337771461	0.354957217
Dm (*diabetes mellitus*)	0.927190965	0.306352333	0.325625284
Cad (*coronary artery disease*)	0.419556496	0.061014910	0.088814167
Appet (*Appetite*)	0.731816079	0.161272379	0.191279314
Pe (*pedal edema*)	0.701471459	0.147642930	0.178322897
Ane (*Anemia*)	0.609840304	0.112940130	0.144399392
Age (*age)*	0.862174529	0.068847739	0.075798102
Bp (*blood pressure*)	0.741482739	0.165895584	0.195641189
Sg (*specific gravity*)	2.057428544	0.594194675	0.394569583
Al (*Albumin*)	1.929573917	0.529213900	0.366998922
Su (Sugar)	1.308742586	0.217086838	0.191842597
Rbc (*red blood cells*)	0.522131740	0.086338730	0.116945325

The selection process is carried out by entering the minimum value of symmetrical uncertainty. In this study, researchers gave a minimum value of symmetrical uncertainty is 0.2. Attributes used for the classification process support vector machine algorithm are attributes that have a symmetrical uncertainty more than equal to (≥) a predetermined minimum value of 0.2. While the attributes that are not used or the attributes that are deleted are those that have a symmetrical uncertainty less than (<) the minimum value that has been determined. There are 13 attributes used and 11 attributes removed. The attributes used and removed for the diagnosis of chronic kidney disease by applying discretization and correlation-based feature selection to the support vector machine algorithm are shown in Table 2.

Data that has passed the stage of discretization and correlation based feature selection is used for the classification process of the support vector machine algorithm. The evaluation technique given is k-fold cross validation with a default value of k = 10. The accuracy results obtained are 99.75%. The results of the accuracy comparison of the support vector machine classification algorithm for the diagnosis of chronic kidney disease by testing using k-fold cross validation can be seen in Table 3.

Table 2. Attributes that are used and deleted in the classification of the SVM algorithm with discretization and correlation based feature selection

Used Attributes	Removed Attributes
Sg (*specific gravity*)	Age
Al (*Albumin*)	Bp (*blood pressure*)
Su (Sugar)	Bgr (*blood glucoses*)
Rbc (*red blood cells*)	Bu (*blood urea*)
Pc (*pus cell*)	Sc (*serum creatinine*)
Pcc (*pus cell clumps*)	Sod (*Sodium*)
Ba (*Bacteria*)	Hemo (*Haemoglobin*)
Pot (*Potassium*)	Pcv (*packed cell volume*)
Wc (*white blood cell*)	Rc (*red blood cell count*)
Cad (*coronary artery disease*)	Htn (*Hypertension*)
Appet (*Appetite*)	Dm (*diabetes mellitus*)
Pe (*pedal edema*)	
Ane (*Anemia*)	

Table 3. The results of the accuracy comparison of the support vector machine classification algorithm for the diagnosis of chronic kidney disease by testing using k-fold cross validation

Algorithm	Accuracy
SVM	99.25%
SVM + correlation based feature selection	99.25%
SVM + discretization + correlation based feature selection	99.75%

In this study, researchers applied discretization and correlation based feature selection to the support vector machine algorithm in the diagnosis of chronic kidney disease. The data used is a chronic kidney disease dataset taken from the UCI repository of machine learning. The pre-processing stage is carried out namely the data cleaning stage, the data transformation stage and the data selection stage. At the cleaning stage, the handling of missing values is done with the average model. Data that has been shared are carried out by a discretization process using Entropy-based Discretization. Then the data selection stage is performed using correlation based feature selection. At this stage, the symmetrical uncertainty value is used for attribute selection. After the data selection stage, the attributes that will be used and the attributes that are deleted are obtained. The selected attributes are used in the classification process using the support vector machine algorithm.

From the results of the application implementation, it can be seen that for the results of the accuracy of the support vector machine classification algorithm in the diagnosis of chronic kidney disease before applying discretization and correlation based feature selection produces 99.25%. While accuracy by applying correlation based feature selection is 99.25%. Meanwhile, after applying discretization and correlation based feature selection it has increased by 0.50% so that it produces an accuracy of 99.75%.

So, we compare with the other research in the case of Chronic Kidney Disease Diagnosis, research by [43] applies discretization and correlation-based feature selection in the C4.5 algorithm obtains an accuracy of 97.5%. Furthermore, research [44] was applied SVM in Chronic Kidney Disease Diagnosis obtaining an accuracy of 97.75%. So, from comparison performance gain from this study with previous research, shows research has a contribution to improve accuracy performance in Chronic Kidney Disease Diagnosis.

4. CONCLUSIONS

From the results of experiments by applying discretization and correlation based feature selection on the support vector machine algorithm in the diagnosis of chronic kidney disease has increased accuracy by 0.50% from 99.25% to 99.75%. The dataset used is chronic kidney disease obtained from the UCI repository of machine learning. The first experiment using the SVM algorithm produced an accuracy of 99.25%. The next experiment added correlation based feature selection before the SVM algorithm classification process resulted in an accuracy of 99.25%. The third experiment added discretization and correlation based feature selection resulting in an accuracy of 99.75%. The application of correlation based feature selection in the chronic kidney disease dataset can minimize the amount of data and reduce irrelevant data used for the data mining process. While the application of discretization in this study can change the attributes of continuous type into discrete and can simplify the original data and make it more efficient. The application of discretization and correlation based feature selection can improve the accuracy of the support vector machine algorithm. Despite the promising contributions presented in this paper, the use of discretization and correlation-based feature selection to improve the prediction performance of CKD remains open for further research.

REFERENCES

[1] M. S. Kukasvadiya and N. H. Divecha, "Analysis of data using data mining tool orange," *Int. J. Eng. Dev. Res.*, vol. 5, no. 2, pp. 1836–1840, 2017, https://www.ijedr.org/viewfull.php?&p_id=IJEDR1702288.

[2] A. Kumar, P. Kumar, A. Srivastava, V. D. Ambeth Kumar, K. Vengatesan, and A. Singhal, "Comparative analysis of data mining techniques to predict heart disease for diabetic patients," in *International Conference on Advances in Computing and Data Sciences*, pp. 507–518, 2020, https://doi.org/10.1007/978-981-15-6634-9_46.

[3] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," in *Intelligent natural language processing: Trends and Applications*, pp. 373–397, 2018, https://doi.org/10.1007/978-3-319-67056-0_18.

[4] S. E. Bibri and J. Krogstie, "The big data deluge for transforming the knowledge of smart sustainable cities: A data mining framework for urban analytics," in *Proceedings of the 3rd International Conference on Smart City Applications*, pp. 1–10, 2018, https://doi.org/10.1145/3286606.3286788.

[5] R.-J. Kuo, T. C. Lin, F. E. Zulvia, and C. Y. Tsai, "A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis," *Appl. Soft Comput.*, vol. 67, pp. 299–308, 2018, https://doi.org/10.1016/j.asoc.2018.02.039.

[6] S. Feng, H. Zhou, and H. Dong, "Using deep neural network with small dataset to predict material defects," *Mater. Des.*, vol. 162, pp. 300–310, 2019, https://doi.org/10.1016/j.matdes.2018.11.060.

[7] M. R. Hidayah, I. Akhlis, and E. Sugiharti, "Recognition number of the vehicle plate using Otsu method and K-nearest neighbour classification," *Sci. J. Informatics*, vol. 4, no. 1, pp. 66–75, 2017, https://doi.org/10.15294/sji.v4i1.9503.

[8] S. T. Ikram and A. K. Cherukuri, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *J. King Saud Univ. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017, https://doi.org/10.1016/j.jksuci.2015.12.004.

[9] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, pp. 49–70, 2023, https://doi.org/10.52465/joiser.v1i1.104.

[10] H. A. Winarno, A. I. Poernama, I. Soesanti, and H. A. Nugroho, "Evaluation on EMG Electrode Reduction in Recognizing the Pattern of Hand Gesture by Using SVM Method," *J. Phys. Conf. Ser.*, vol. 1577, no. 1, 2020, https://doi.org/10.1088/1742-6596/1577/1/012044.

[11] A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV", *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 4, no. 1, pp. 12–21, May 2022, https://doi.org/10.12928/biste.v4i1.6079.

[12] Triwiyanto, O. Wahyunggoro, H. A. Nugroho, and Herianto, "Upper Limb Elbow Joint Angle Estimation Based on Electromyography Using Artificial Neural Network," in *2018 12th South East Asian Technical University Consortium (SEATUC)*, pp. 1–6, 2018, https://doi.org/10.1109/SEATUC.2018.8788877.

[13] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, pp. 101–121, 2020, https://doi.org/10.1016/B978-0-12-815739-8.00006-7.

[14] R. Rosita, D. A. A. Pertiwi, and O. G. Khoirunnisa, "Prediction of Hospital Intesive Patients Using Neural Network Algorithm," *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 8–11, 2022, https://doi.org/10.52465/joscex.v3i1.61.

[15] K. Jha and S. Saha, "Incorporation of multimodal multiobjective optimization in designing a filter based feature selection technique," *Appl. Soft Comput.*, vol. 98, p. 106823, 2021, https://doi.org/10.1016/j.asoc.2020.106823.

[16] C. Jie, L. Jiawei, W. Shulin, and Y. Sheng, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018, https://doi.org/10.1016/j.neucom.2017.11.077.

[17] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognit.*, vol. 64, pp. 141–158, 2017, https://doi.org/10.1016/j.patcog.2016.11.003.

[18] N. Gopika and A. M. Kowshalaya. M. E, "Correlation based feature selection algorithm for machine learning," in *2018 3rd international conference on communication and electronics systems (ICCES)*, pp. 692–695, 2018, https://doi.org/10.1109/CESYS.2018.8723980.

[19] Z. Chuanlei, Z. Shanwen, Y. Jucheng, S. Yancui, and C. Jia, "Apple leaf disease identification using genetic algorithm and correlation based feature selection method," *Int. J. Agric. Biol. Eng.*, vol. 10, no. 2, pp. 74–83, 2017, http://www.ijabe.org/index.php/ijabe/article/view/2166.

[20] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, 2018, https://doi.org/10.1016/j.asoc.2017.11.006.

[21] K. Yan, L. Ma, Y. Dai, W. Shen, Z. Ji, and D. Xie, "Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis," *Int. J. Refrig.*, vol. 86, pp. 401–409, 2018, https://doi.org/10.1016/j.ijrefrig.2017.11.003.

[22] A. K. Shrivas, S. K. Sahu, and H. S. Hota, "Classification of chronic kidney disease with proposed union based feature selection technique," in *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, pp. 26–27, 2018, https://doi.org/10.2139/ssrn.3168581.

[23] I. M. Nasir *et al.*, "Pearson correlation-based feature selection for document classification using balanced training," *Sensors*, vol. 20, no. 23, p. 6793, 2020, https://doi.org/10.3390/s20236793.

[24] F. Nojavan, S. S. Qian, and C. A. Stow, "Comparative analysis of discretization methods in Bayesian networks," *Environ. Model. Softw.*, vol. 87, pp. 64–71, 2017, https://doi.org/10.1016/j.envsoft.2016.10.007.

[25] S. S. Pal and S. Kar, "Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory," *Math. Comput. Simul.*, vol. 162, pp. 18–30, 2019, https://doi.org/10.1016/j.matcom.2019.01.001.

[26] N. Thein, K. Hamamoto, H. A. Nugroho, and T. B. Adji, "A comparison of three preprocessing techniques for kidney stone segmentation in CT scan images," in *2018 11th Biomedical Engineering International Conference (BMEiCON)*, pp. 1–5, 2018, https://doi.org/10.1109/BMEiCON.2018.8609996.
[27] P. Romagnani *et al.*, "Chronic kidney disease," *Nat. Rev. Dis. Prim.*, vol. 3, no. 1, pp. 1–24, 2017, https://doi.org/10.1038/nrdp.2017.88.
[28] Centers for Disease Control and Prevention, *Chronic kidney disease in the United States, 2019*. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, 2019, https://fluoridealert.org/studytracker/38332/.
[29] T. K. Chen, D. H. Knicely, and M. E. Grams, "Chronic kidney disease diagnosis and management: a review," *Jama*, vol. 322, no. 13, pp. 1294–1304, 2019, https://doi.org/10.1001/jama.2019.14745.
[30] A. C. Webster, E. V Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *Lancet*, vol. 389, no. 10075, pp. 1238–1252, 2017, https://doi.org/10.1016/S0140-6736(16)32064-5.
[31] W. Zheng *et al.*, "Improving crop yields, nitrogen use efficiencies, and profits by using mixtures of coated controlled-released and uncoated urea in a wheat-maize system," *F. Crop. Res.*, vol. 205, pp. 106–115, 2017, https://doi.org/10.1016/j.fcr.2017.02.009.
[32] J. L. Segar *et al.*, "Fluid management, electrolytes imbalance and renal management in neonates with neonatal encephalopathy treated with hypothermia," in *Seminars in Fetal and Neonatal Medicine*, vol. 26, no. 4, p. 101261, 2021, https://doi.org/10.1016/j.siny.2021.101261.
[33] S. Javaid, H. Awais, M. Usman, and U. Mukhtar, "Biochemical Changes in Chronic Kidney Disease (CKD) Patients and its Association with Hypertension and Diabetes Mellitus," *Asian J. Allied Heal. Sci.*, vol. 6, no. 2, 2021, https://jucmd.pk/journals/AJAHS/article/view/1415.
[34] K. L. Watts, P. Ghosh, S. Stein, and R. Ghavamian, "Value of nephrometry score constituents on perioperative outcomes and split renal function in patients undergoing minimally invasive partial nephrectomy," *Urology*, vol. 99, pp. 112–117, 2017, https://doi.org/10.1016/j.urology.2016.01.046.
[35] M. Liu *et al.*, "Personal exposure to fine particulate matter and renal function in children: a panel study," *Environ. Pollut.*, vol. 266, p. 115129, 2020, https://doi.org/10.1016/j.envpol.2020.115129.
[36] J. P. Kooman *et al.*, "Inflammation and premature aging in advanced chronic kidney disease," *Am. J. Physiol. Physiol.*, vol. 313, no. 4, pp. F938–F950, 2017, https://doi.org/10.1152/ajprenal.00256.2017.
[37] W. F. Clark *et al.*, "Effect of coaching to increase water intake on kidney function decline in adults with chronic kidney disease: the CKD WIT randomized clinical trial," *Jama*, vol. 319, no. 18, pp. 1870–1879, 2018, https://doi.org/10.1001/jama.2018.4930.
[38] F. Ridzuan and W. M. N. W. Zainon, "A review on data cleansing methods for big data," *Procedia Comput. Sci.*, vol. 161, pp. 731–738, 2019, https://doi.org/10.1016/j.procs.2019.11.177.
[39] C. B. Rjeily, G. Badr, A. Hajjarm El Hassani, and E. Andres, "Medical data mining for heart diseases and the future of sequential mining in medical field," in *Machine Learning Paradigms*, pp. 71–99, 2019, https://doi.org/10.1007/978-3-319-94030-4_4.
[40] I. S. Thaseen, J. Saira Banu, K. Lavanya, M. Rukunuddin Ghalib, and K. Abhishek, "An integrated intrusion detection system using correlation-based attribute selection and artificial neural network," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 2, 2021, https://doi.org/10.1002/ett.4014.
[41] D. Chutia, D. K. Bhattacharyya, J. Sarma, and P. N. L. Raju, "An effective ensemble classification framework using random forests and a correlation based feature selection technique," *Trans. GIS*, vol. 21, no. 6, pp. 1165–1178, 2017, https://doi.org/10.1111/tgis.12268.
[42] F. Hamedan, A. Orooji, H. Sanadgol, and A. Sheikhtaheri, "Clinical decision support system to predict chronic kidney disease: A fuzzy expert system approach," *Int. J. Med. Inform.*, vol. 138, p. 104134, 2020, https://doi.org/10.1016/j.ijmedinf.2020.104134.
[43] N. Cahyani and M. A. Muslim, "Increasing Accuracy of C4. 5 Algorithm by applying discretization and correlation-based feature selection for chronic kidney disease diagnosis," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 12, no. 1, pp. 25-32, 2020, https://jtec.utem.edu.my/jtec/article/view/4922.
[44] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101–111, 2019, https://doi.org/10.1016/j.compbiomed.2019.04.017.