SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

JOSCEX

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Optimization of support vector machine using information gain and adaboost to improve accuracy of chronic kidney disease diagnosis

Eka Listiana[1], Rini Muzayanah[2], Much Aziz Muslim[3], Endang Sugiharti[4]
[1, 2,3, 4]Departement of Computer Science, Universitas Negeri Semarang, Indonesia
[3] Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

Article Info

Article history:

Received September 25, 2023
Revised September 27, 2023
Accepted September 28, 2023

Keywords:

Data mining
Support vector machine
Adaptive boosting
Information gain
Chronic kidney disease

ABSTRACT

Today's database is growing very rapidly, especially in the field of health. The data if not processed properly then it will be a pile of data that is not useful, so the need for data mining process to process the data. One method of data mining used to predict a decision in any case is classification, where in the classification method there is a support vector machine algorithm that can be used to diagnose chronic kidney disease. The purpose of this study is to determine the level of accuracy of the application of information gain and AdaBoost on the support vector machine algorithm in diagnosing chronic kidney disease. The use of information gain is to select the attributes that are not relevant while AdaBoost is used as an ensemble method commonly known as the method of classifier combination. In this study the data used are chronic kidney disease (CKD) dataset obtained from UCI repository of machine learning. The result of experiment using MATLAB applying information gain and AdaBoost on vector machine support algorithm with k-fold cross validation default k = 10 shows an accuracy increase of 0.50% with the exposure of the result as follows, the support vector machine algorithm has accuracy of 99.25 %, if by applying AdaBoost on the support vector machine has an accuracy of 99.50%, whereas if applying AdaBoost and information gain on the support vector machine has an accuracy of 99.75%.

Corresponding Author:

Eka Listiana,
Departement of Computer Science,
Universitas Negeri Semarang, Indonesia.
Email: ekalisti@students.unnes.ac.id

https://doi.org/10.52465/joscex.v4i3.218

1. INTRODUCTION

The development of databases in the present is growing very rapidly, especially the existing data in the field of health. From a large collection of such data when not in use it becomes just a collection of data that is not useful. Therefore, from the pile of data that is not useful can be used as a source of data that is then processed so that more useful commonly referred to as the term data mining. Data mining is a process used to find data that has not been known by the user with a model so that can be understood and used as the basis for decision making [1]. Data mining is a stage to find patterns or information in a set of data by using certain techniques and algorithms. Data mining is often called Knowledge Discovery in Databases (KDD) [2]. In data mining, data is stored electronically and processed automatically by computer using certain techniques and calculations [3]. Various learning techniques have been developed, including supervised, unsupervised, and reinforcement learning [4]. One method or technique used to predict a decision is classification. Classification is the process of classifying test data on a class that has been determined based on the learning algorithm [5]. Classification is a data mining technique that can be used to predict group membership for data instances [6]. Also classification can also be used as a way to solve the scheduling problem, where in the scheduling is classified into several kinds of attributes used [7]. Classification technique is a systematic approach to construct a classification model of an input data set [8]. Classification is a process used to find models (or functions) by describing and differentiating data classes or concepts [9]. Popular algorithm for data classification is support vector machine (SVM) [10]. SVM is a new method used for data classification of both linear and nonlinear data, which creates discrete hyperplane in the descriptor space of the training data and is classified based on the hyperplane side residing [11]. SVM maps nonlinear input data to some of the higher dimension spaces where data can be separated linearly, thus providing a large classification or data regression [12]. Using the kernel trick to map the training sample from the input space to the high-dimensional feature space [13]. Although the basis for SVM has existed since the 1960s this method makes SVM a promising new method for classifying data, both linear and nonlinear data.

Ensemble method is a method used to improve classification algorithm accuracy by constructing several classifiers of training data then at the time of classification of this method using the voting / aggregating of the classifier [14]. Adaboost is a very popular boosting algorithm to improve classification accuracy [15]. In adaboost this training set that is used for each base classifier is selected based on the performance of the previous klasifier. In boosting, samples that are not correctly predicted by the classifier in the circuit will be selected more often than correctly sampled samples. Adaboost and its variants have been successfully applied in several fields because of their strong theoretical basis, accurate predictions and great simplicity [16].

Maintaining health is very important in human life [17]. The classification algorithm can be utilized and assist the medical expert in diagnosing a disease, one of them being chronic kidney disease (CKD). CKD is a heterogeneous disorder that affects the structure and function of the kidney progressively 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 [18]. CKD is also a pathogenic physiology process with various variations that will lead to significantly decreased renal function (gland in the kidney) where ultimately kidney failure occurs [19]. This condition is called chronic because the damage happens slowly over a long period of time. It is a disease that affects people all over the world [20]. Chronic Kidney Disease (CKD) also known as Chronic Renal Failure (CRF) is one of the diseases that its timely diagnosis has a significant positive impact on the patient’s chances of recovery, otherwise, the disease progresses and gets out of control [21].

This study uses CKD dataset. The dataset used in this study was obtained from the UCI repository of machine learning datasets. Datasets are a collection of objects and properties or characteristics of an object itself (attribute). In the chronic kidney disease dataset, there may be many irrelevant atibuts in the data to be used so that it needs to be removed. There are also many mining algorithms that do not do their job well because of the large number of features and attributes [14]. Therefore, it is necessary to apply techniques that can evaluate and optimize attributes on the dataset. Optimizing the work of the classifier can be done by selecting relevant features using feature selection [22]. One of the techniques that can be used is feature selection. Feature selection is the process of selecting relevant features, or a subset of feature candidates. Evaluation criteria are used to obtain optimal feature parts [23]. Based on the above background description, this study uses information gain as feature selection and ensemble learning Adaptive boosting (Adaboost) as a technique to improve the classification level applied to classify SVM algorithm in diagnosing chronic kidney disease.

2. METHOD

Systematically there are stages of the methods undertaken in this study, starting from the stage of data processing, classification, and evaluation where will result in improved accuracy of the support vector machine algorithm. In Figure 1 is a classification flow diagram using adaboost and information gain on the support vector machine.

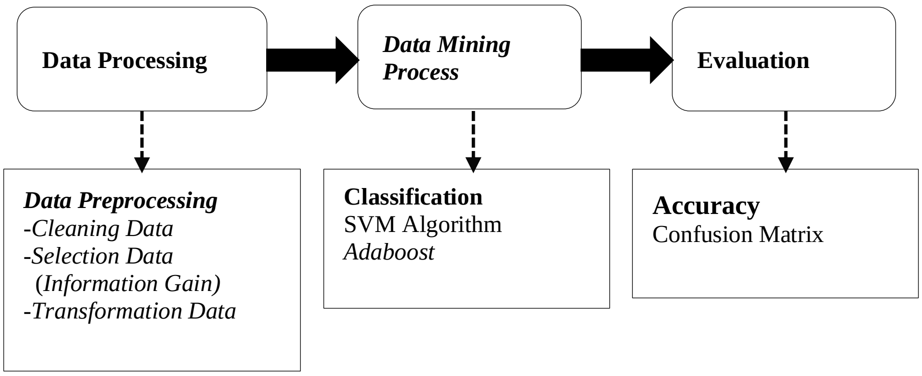

Figure 1. Flowchart of research methods

2.1 Data Processing

Stage of data processing in this research use some process such as data cleaning, data integration, data selection, and data transformation. The data used are chronic kidney disease dataset taken from UCI repository of machine learning. The number of record data in the dataset, ie 400 records consisting of 24 attributes and 1 class attribute. In the dataset used still have a record there is no value / missing value, so it needs to be applied data cleaning process. In the missing value handling can be filled with the average value. The average calculation to replace missing value data can be shown in Equation 1.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Description:

$\boldsymbol{x_i}$: Attribute Value

n : Number of attributes that have values

In the data selection stage, dimensional dimensions occur on the dataset to optimize attributes that will affect the accuracy of the algorithm in mining datasets. Dimensional dimensions of this atibut are performed using techniques in the future selection of information gain. Information gain is an expectation of the reduction of entropy generated from the dataset object's partition based on certain features [24]. Attribute deletion is done one by one from the attribute that has the smallest information gain value and then it will be mining [25]. The value of information gain can be seen in Table 1. The disposal and mining process will stop when the accuracy of each algorithm has decreased. The process of obtaining an entropy value for each attribute is obtained from Equation 2.

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i log_2 p_i \quad (2)$$

Description:

S = space (data) of samples used,

n = the number or partition

pi = the proportion of Si to S (the number of positive resolution (+) and negative resolution (-) in the data for a given criterion)

Table 1. The value of the data gain attributes the CKD dataset

No.	Attribute Name	Information Gain
1	(age)Age	0,06884774
2	(bp) Blood Pressure	0,165895585
3	(sg) Specific Gravity	0,449386996
4	(al) Albumin	0,465192841
5	(su) Sugar	0,115043167
6	(rbc) Red Blood Cells	0,08633873

7	(pc) Pus Cell	0,14764293
8	(pcc) Pus Cells Clumps	0,078416627
9	(ba) Bacteria	0,036848979
10	(bgr) Blood Glucose	0,349264124
11	(bu) Blood Urea	0,333740277
12	(sc) Serum Creatinine	0,580280931
13	(sod) Sodium	0,188281608
14	(pot) Potassium	0,064832727
15	(hemo) Hemoglobin	0,692144437
16	(pcv) Packed Cell	0,653457091
17	(wc) White Bloodcell	0,077142271
18	(re) Red Blood Cell	0,41820288
19	(htn) Hypertension	0,337771462
20	(dm) Diabetes Mellitus	0,306352333
21	(cad) Coronary Artery Disease	0,061014911
22	(appet) Appetite	0,16127238
23	(pe) Pedal Edema	0,14764293
24	(ane) Anemia	0,11294013
25	(class) Class	0,954434003

The transformation stage uses Entropy-Based Discretization which is used to transform numeric data into recognizable patterns to obtain entropy values aimed at finding information gain values from numeric attributes. So for numeric attributes can be transformed into two categories represented in numerical form 0 and 1 [26].

2.2 Data Mining Process

This study applies information gain and adaboost to improve the accuracy of the support vector machine algorithm in diagnosing chronic kidney disease.

2.2.1 Classification of Support Vector Machine (SVM) Algorithm

SVM is an algorithm that works using a nonlinear mapping to convert the original training data to a higher dimension, in a new dimension, then will look for optimal linear hyperplane separators.

The first step of an SVM algorithm is to define the equation of a separating hyperplane written with Equation 3.

$$w.X + b = 0 \tag{3}$$

Where w is a vector weight, that is $w = \{w_1, w_2, \dots, w_n\}$; n n is the number of attributes and b is a scalar called bias. If it is based on attribute A1, A2 with training tuple layout of X = $(x_{1,}x_2)$, x1 and x2 is the value of the attribute A1 and A2, and if b is considered an additional weight w_0, then the equation of a separating hyperplane can be rewritten as in Equation 4.

$$w_0 + w_1x_1 + w_2x_2 = 0 \tag{4}$$

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

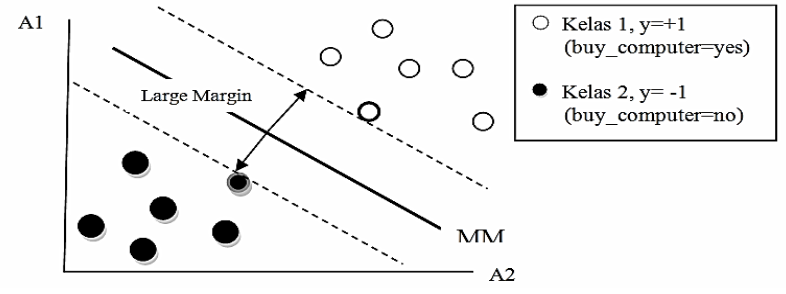

Figure 2. Separation of two data classes with maximum margins

In Figure 2, SVM finds the maximum separator hyperlane, the hyperlane that has the maximum distance between the nearest training tuples. Support vector is shown with bold constraints at the tuple point. Thus, any point located above the dividing hyperplane satisfies Eq. 5.

$$w_0 + w_1x_1 + w_2x_2 > 0 \quad (5)$$

Whereas, the point located below the separating hyperlane satisfies the formula as in Equation 6.

$$w_0 + w_1x_1 + w_2x_2 < 0 \quad (6)$$

Looking at the two conditions above, we get two hyperplane equations, as in Equations 7 and 8.

$$H_1: w_0 + w_1x_1 + w_2x_2 \geq 0 \quad (7)$$
$$H_2: w_0 + w_1x_1 + w_2x_2 \leq 0 \quad (8)$$

Based on the lagrangian formulation, Maximum Margin Hyperplane (MMH) can be rewritten as a decision boundary written with Equation 9.

$$d(X^T) = \sum_{i=1}^{l} y_i a_i \; X_i X^T + b_0 \quad (9)$$

$\boldsymbol{y_i}$ is the class label of the support vector $\boldsymbol{X_i}$, $\boldsymbol{X^T}$ is a test tupel. $\boldsymbol{a_i}$ and $\boldsymbol{b_o}$ is a numerical parameter determined automatically by the optimization of SVM algorithm and $\boldsymbol{l}$ is the number of vector support.

2.2.2 Classification of SVM Algorithm Based Adaboost Ensemble and Information Gain

The ensemble method, known as the classifier combination method, is a technique that can be used to improve the accuracy of the classification in data mining. In adaboost this training set that is used for each base classifier is selected based on the performance of the previous klasifier. Application of Adaboost (Adaptive Boosting) technique on vector machine support algorithm first divide data into 2 parts, that is training data for model formation process and data testing to measure ability of classification model that formed. Data sharing is done by using k-fold cross validation with default value k = 10.

Then suppose $\mathcal{X}$ denoted as instance and $\mathcal{Y}$ as a set of class labels. Supposed $\mathcal{Y} = \{-1, +1\}$. Then given base or weak learning algorithm and a *training set* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The *Adaboost* algorithm works like the following: first each training example (x_i, y_i) $(i \in \{1, \ldots . m\})$ given the same weight. Denote the weight distribution in the learning round t-th as D_t. From the training set and D_t the Adaboost algorithm generates a weak or base learner $h_t : \mathcal{X} \rightarrow \mathcal{Y}$ by calling its basic learning algorithm. Then the training example is used to test of h_t, and the weights from the wrong classification sample will increase. Thus, a weight distribution has been updated D_{t+1} obtained. From training set and D_{t+1} *Adaboost* generate another weak learner by calling the basic learning algorithm again. The process is repeated for T rounds, and the final model is obtained by weighted majority voting from a T weak learner pool, where the weight of the learner is determined during the training or training process [27]. After the process is complete then select the attributes that are not relevant based on the value of the gain that has been sorted.

2.3 Evaluation Step

Measurement of classification performance can be done by means of confusion matrix [28]. Confusion matrix is a tool used to evaluate the classification model to estimate the correct or incorrect object and its accuracy, precision, and recall can be calculated [29]. A matrix of prediction that will be compared to the original class of input or in other words contains actual and predicted value information on the classification [30]. Tool or system used is using software matlab R2014a. The workflow stage is illustrated in Figure 3.

.

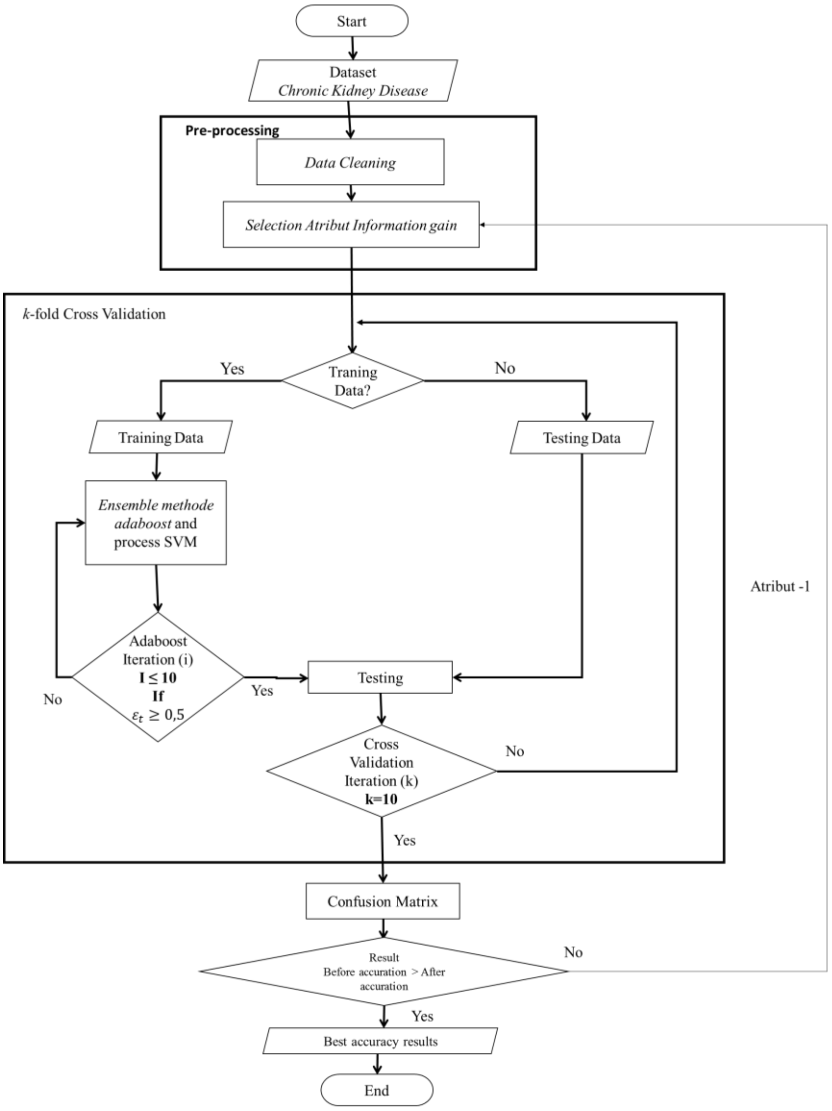

Figure 3. Flowchart support vector machine by applying the adaboost ensemble technique and information gain

3. RESULTS AND DISCUSSIONS

The testing process in this research using Matlab R2014a software. By utilizing Matlab, data analysis, algorithm development, and modeling and application can also be made visual appearance of a program so that it can facilitate user [31]. Matlab is a high performance (high performance) language for computational engineering problems [32]. Data used for model test is chronic kidney disease dataset taken from UCI repository of machine learning. Specifications used for this research, namely Intel processor (R) Core (TM) i3-3227U CPU @ 1.90GHz (4 CPUs), ~ 1.9GHz; 2 GB memory; and Windows 8.1 operating system. In this section the system is used to analyze the performance evaluation of data mining algorithms used. The result of comparison of accuracy of vector machine support algorithm to diagnose chronic kidney disease using k-fold cross validation can be seen in Table 2.

Table 2**.** Accuracy comparison of support vector machine algorithm

Algorithm	*Without Discretization*
SVM Classification	99,25%
SVM Classification and *Adaboost*	99,50%
SVM Classification using *Adaboost* + IG	99,75%

The improved graph of accuracy in the support vector machine algorithm can be seen in Figure 4.

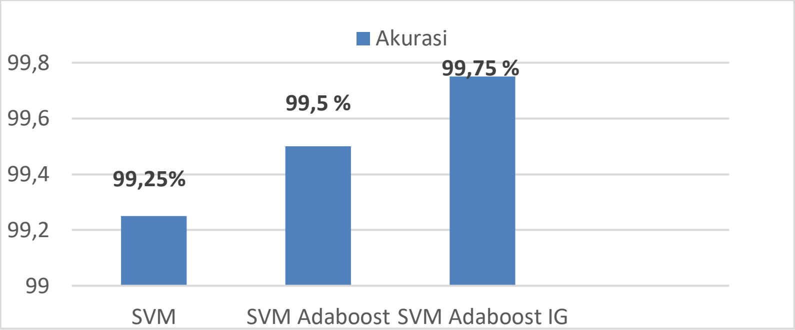

Figure 4. Graph of increasing SVM accuracy, adaboost SVM, and SVM adaboost IG

Based on the results of system implementation it can be seen that the accuracy of classification algorithm support vector machine before applying information gain and adaboost generate 99.25%. The result of accuracy of classification algorithm of support vector machine and adaboost produce 99,50%. Whereas after applied adaboost and information gain in classification algorithm of support vector machine have increased 0,50% so yield accuracy equal to 99,75%.

The advantages of this study is to apply information gain and adaboost can improve accuracy in diagnosing chronic kidney disease so it can be used by further researchers as a reference in conducting data mining classification research in diagnosing chronic kidney disease, on the other hand this research has a deficiency that is not the implementation pre-processing complete data such as data integration.

4. CONCLUSION

Optimization of support vector machine algorithm using adaboost and information gain in diagnosing chronic kidney disease with three experiments has an accuracy increase of 0.50%. The dataset used is chronic kidney disease taken from the UCI repository of machine learning. The first experiment, using SVM algorithm, yielded 99.25% accuracy. The second experiment by adding adaboost yields an accuracy of 99.50%. While in the third experiment feature selection information gain and adaboost on SVM classification has an accuracy of 99.75%.

REFERENCES

[1] E. Sugiharti, S. Firmansyah, and F. R. Devi, “Predictive evaluation of performance of computer science students of unnes using data mining based on naÏve bayes classifier (NBC) algorithm,” *J. Theor. Appl. Inf. Technol.*, vol. 95, pp. 902–911, Feb. 2017.

[2] M. Muslim, I. Kurniawati, and E. Sugiharti, “Expert system diagnosis chronic kidney disease based on mamdani fuzzy inference system,” vol. 78, pp. 70–75, Aug. 2015.

[3] 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, May 2017, doi: 10.15294/sji.v4i1.9503.

[4] A. Nurdina and A. B. I. Puspita, “Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis,” *J. Inf. Syst. Explor. Res.*, vol. 1, no. 2, Jul. 2023, doi: 10.52465/joiser.v1i2.167.

[5] K. P. Wijaya and M. A. Muslim, “Peningkatan Akurasi pada Algoritma Support Vector Machine dengan Penerapan Information Gain untuk Mendiagnosa Chronic Kidney Disease,” in *Seminar Nasional Ilmu Komputer*, Semarang: Universitas Negeri Semarang, 2016, pp. 22–27.

[6] I. A. Ashari, M. A. Muslim, and A. Alamsyah, “Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing,” *Sci. J. Informatics*, vol. 3, no. 2, pp. 149–158, Nov. 2016, doi: 10.15294/sji.v3i2.7911.

[7] A. Nurzahputra and M. A. Muslim, “Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit,” *Seminar Nasional Teknologi dan Informatika 2017*. Indonesia, 2017.

[8] E. Sugiharti and M. Muslim, “On-line clustering of lecturers performance of computer science department of semarang state university using K-MeansAlgorithm,” vol. 83, pp. 64–71, Jan. 2016.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[10] D. Aprilianto, “SVM Optimization with Correlation Feature Selection Based Binary Particle Swarm Optimization for Diagnosis of Chronic Kidney Disease,” *J. Soft Comput. Explor.*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/joscex.v1i1.1.

[11] V. S and D. S, “Data Mining Classification Algorithms for Kidney Disease Prediction,” *Int. J. Cybern. Informatics*, vol. 4, no. 4, pp. 13–25, Aug. 2015, doi: 10.5121/ijci.2015.4402.

[12] X. Li, L. Wang, and E. Sung, “AdaBoost with SVM-based component classifiers,” *Eng. Appl. Artif. Intell.*, vol. 21, no. 5, pp. 785–795, Aug. 2008, doi: 10.1016/j.engappai.2007.07.001.

[13] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. in Pearson International Edition. Pearson Addison Wesley, 2006. [Online]. Available: https://books.google.co.id/books?id=_XdrQgAACAAJ

[14] K. Kumar and R. Chezian, “Support vector machine and K- nearest neighbor based analysis for the prediction of hypothyroid,” *Int. J. Pharma Bio Sci.*, vol. 5, pp. B447–B453, Jan. 2014.

[15] H. A. Prihanditya and A. Alamsyah, “The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease,” *J. Soft Comput. Explor.*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/joscex.v1i1.8.

[16] A. Lestari and A. Alamsyah, “Increasing Accuracy of C4.5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease,” *J. Soft Comput. Explor.*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/joscex.v1i1.6.

[17] N. Reska and K. Tsabita, “Comparison of KNN, naive bayes, and decision tree methods in predicting the accuracy of classification of immunotherapy dataset,” *J. Student Res. Explor.*, vol. 1, no. 2, pp. 104–121, Jul. 2023, doi: 10.52465/josre.v1i2.170.

[18] A. S. Levey and J. Coresh, “Chronic kidney disease,” *Lancet*, vol. 379, no. 9811, pp. 165–180, Jan. 2012, doi: 10.1016/S0140-6736(11)60178-5.

[19] S. Beniwal and J. Arora, “Classification and feature selection techniques in data mining,” *Int. J. Eng. Res. Technol.*, vol. 1, Aug. 2012.

[20] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, “Chronic kidney disease prediction based on machine learning algorithms,” *J. Pathol. Inform.*, vol. 14, p. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.

[21] J. Qezelbash-Chamak, S. Badamchizadeh, K. Eshghi, and Y. Asadi, “A survey of machine learning in kidney disease diagnosis,” *Mach. Learn. with Appl.*, vol. 10, p. 100418, Dec. 2022, doi: 10.1016/j.mlwa.2022.100418.

[22] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, “Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis,” *Sci. J. Informatics*, vol. 6, no. 1, pp. 138–149, 2019.

[23] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection*. Wiley, 2008. doi: 10.1002/9780470377888.

[24] M. Mirqotussa’adah, M. A. Muslim, E. Sugiharti, B. Prasetiyo, and S. Alimah, “Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes,” *Lontar Komput. J. Ilm. Teknol. Inf.*, p. 135, Aug. 2017, doi: 10.24843/LKJITI.2017.v08.i02.p07.

[25] E. Listiana and M. A. Muslim, “Penerapan Adaboost untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi pada Diagnosa Chronic Kidney Disease,” *Seminar Nasional Teknologi dan Informatika 2017*. Indonesia, 2017.

[26] E. Prasetyo, *Data mining : Mengolah data menjadi informasi menggunakan matlab*. Yogyakarta, 2014.

[27] B. Wu and R. Nevatia, “Cluster Boosted Tree Classifier for Multi-View, Multi-Pose Object Detection,” in *2007 IEEE 11th International Conference on Computer Vision*, IEEE, 2007, pp. 1–8. doi: 10.1109/ICCV.2007.4409006.

[28] A. Aziz, R. Saptono, and K. P. Suryajaya, “Implementasi Vector Space Model dalam Pembangkitan Frequently Asked Questions Otomatis dan Solusi yang Relevan untuk Keluhan Pelanggan,” *Sci. J. Informatics*, vol. 2, no. 2, p. 111, Feb. 2016, doi: 10.15294/sji.v2i2.5076.

[29] F. Gorunescu, *Data Mining: Concepts, models and techniques*. 2011.

[30] N. Anam and P. Hendikawati, “Aplikasi Matrix Labolatory untuk Perhitungan Sistem Antrian dengan Server Tunggal dan Majemuk,” *Sci. J. Informatics*, vol. 1, no. 1, pp. 65–78, May 2015, doi: 10.15294/sji.v1i1.3642.

[31] A. Y. Nugroho, A. Suyitno, and R. Arifudin, “PERBANDINGAN ALGORITMA BRANCH AND BOUND DAN ALGORITMA GENETIKA UNTUK MENGATASI TRAVELLING SALESMAN PROBLEM (TSP) (Studi Kasus PT. JNE Semarang),” *Unnes J. Math.*, vol. 5, no. 2, pp. 135–143, 2017, doi: https://doi.org/10.15294/ujm.v5i2.13123.

[32] A. E. Pramadhani and T. Stiadi, “Penerapan Data mining Untuk Klasifikasi Prediksi Penyakit ISPA (Infeksi Saluran Pernafasan Akut) Dengan Algoritma Decision Tree (ID3),” *J. Sarj. Tek. Inform.*, vol. 2, no. 1, pp. 831–839, 2014.