CLASSIFICATION OF KIDNEY DISEASE USING GENETIC MODIFIED KNN AND ARTIFICIAL BEE COLONY ALGORITHM

Ardina Ariani[1], Samsuryadi[2*]
[1]Graduate Program in Computer Science, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
[2]Department of Computer Science, Faculty of Computer Science, Universitas Sriwijaya, Indonesia

Abstract
The health care system is currently improving with the development of intelligent artificial systems in detecting diseases. Early detection of kidney disease is essential by recognizing symptoms to prevent more severe damages. This study introduces a classification system for kidney diseases using the Artificial Bee Colony (ABC) algorithm and genetically modified K-Nearest Neighbor (KNN). ABC algorithm is used as a feature selection to determine relevant symptoms used in influencing kidney disease and Genetic modified KNN used for classification. This research consists of 3 stages: pre-processing, feature selection, and classification. However, it focuses on the pre-processing stage of chronic kidney disease using 400 records with 24 attributes for the feature selection and classification. Kidney disease data is classified into two classes, namely chronic kidney disease and not chronic kidney disease. Furthermore, the performance of the proposed method is compared with other methods. The result showed that an accuracy of 98.27% was obtained by dividing the dataset into 280 training and 120 test data.

INTRODUCTION

Classification is one of the most commonly used data grouping mining techniques in the health sector. This technique is used to detect diseases based on symptoms. Various classification techniques are used to detect and classify diseases. Several diseases are cancer, parkinsonism, heart, liver, diabetes, hepatitis, thalassemia, and Alzheimer's [1].

The research carried out by [2, 3, 4] stated that the most widely used method in classifying various applications and diseases is the K-Nearest Neighbor (KNN). This method has been used to classify hepatitis with better performance than the Neural Network (NN) [5].

Fast training and simple algorithms are the hallmarks of the KNN. However, [6] stated that the number of features that influence the nearest neighbour's distance and the accuracy level is one of the weaknesses associated with the KNN. The condition leads to low accuracy on the multidimensional datasets. Therefore, it is necessary to determine the features that are best used in classification. Furthermore, to overcome the low accuracy in KNN, a new algorithm is needed, which modifies the KNN, enabling it to validate the training data and weigh the calculated votes. Modified KNN also has weaknesses in determining the parameter k. Therefore, the proven Genetic Algorithm needs optimization [7] to produce k automatically with the best results. Optimization of the Genetic Algorithm (GA) in the Modified KNN is called Genetic Modified KNN (GMKNN). Also, features are selected to obtain the attributes that affect classification techniques based on the best model obtained.

Kidney disease is a worldwide health problem. It leads to decreased glomerular filtration rate and cardiovascular disease, which causes death [8]. The situation is a killer disease because most infected people fail to obtain early diagnoses. The dataset derived from the UCI Machine Learning outlined 24 symptoms (features) used to determine and group people with and without kidney disease. Therefore, there are many features. Ancillary equipment components do not need to use as a determining

factor in classifying someone suffering from chronic kidney disease.

In the study conducted by [9], Particle Swarm Optimization (PSO) was applied to feature selection on form-based diagnosis in Cluster Microcalcification in Mammography. The Ant Colony Optimization (ACO) method was modified into an efficient multivariate feature selection method to select the optimal feature from the initial set. This method shows a higher efficiency in reducing the number of features by maintaining maximum classification accuracy.

Research from [10] combined the Artificial Bee Colony (ABC) with the ACO method to optimize feature selection. The research produced a suitable ABC performance method to reduce the feature subset's size with a better accuracy and speed level. Furthermore, this study examines the performance of the Genetic Modified KNN (GMKNN) classification method by applying the ABC algorithm for feature selection in classifying kidney diseases using the chronic dataset derived from the UCI Repository Machine Learning.

METHOD

Several research stages are needed to obtain the best results from the classification of chronic kidney diseases. Figure 1 shows various stages carried out in this study.
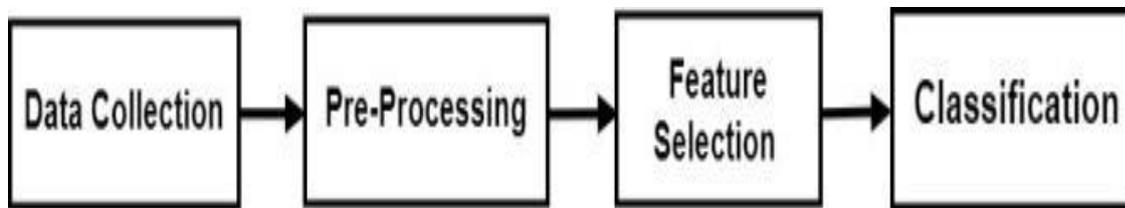

Figure 1: Research Stages

Data Collection

The data used in this study were obtained from the Center for Chronic Kidney Disease at the University of California warehouse (UCI Repository Machine Learning Benchmark). This dataset consisted of 200 patients with 24 symptoms divided into chronic and non-chronic groups, as shown in Table 1.

Table 1. Distribution of Chronic Kidney Disease Datasets

Class	
CKD (Chronic Kidney Disease)	Not CKD (Chronic Kidney Disease)
250	150

Many factors greatly influence the success of the performance of the Machine Learning method. The representation and quality of data are one of essential attributes. The condition happened because irrelevant data and noise make it difficult for knowledge discovery in the training process [11]. Therefore, it is important to pre-process the data by cleaning and normalizing it from noise to obtain the right information.

Pre-processing

At this stage, the data pre-processing was conducted by implementing the mean, median, and mode calculation techniques. Meanwhile, the same scale of data, normalization process, was performed using the min-max scalar technique.

Feature selection is an effective data pre-processing technique that was conducted by taking a portion of the selected features from the original dataset. The purpose of feature selection is to eliminate noisy, irrelevant, and excessive features. The application of this selection enables a reduction in computational costs and increased accuracy in the data analysis process [11]. The reduced dataset was normalized with the min-max scalar technique is shown in (1) [12].

$$S' = \frac{s - \min(S_k)}{\max(S_k) - \min(S_k)} \tag{1}$$

Where:
s = value from the input data
$\min(S_k)$ = the minimum value of data
$\max(S_k)$ = the maximum value of data

Feature Selection

The next step is a feature selection using the Artificial Bee Colony (ABC) algorithm. It comprises a simple concept that is easy to implement with little control parameters. The ABC algorithm's three important components are food sources and waiting for bees in the hive [13].

In the ABC algorithm, the food source position represents the solution that allows for the optimization problem and the nectar of the number of food sources according to the related solution's quality. The number of observer bees is equal to the solutions in the population.

Each search cycle consists of the following three steps:

1. Sending employed bees to their food source and evaluating the amount of nectar,
2. Selecting food source areas by observer bees and evaluating the amount of nectar from food sources,
3. Determining surveillance bees and randomly sending them to possible new food sources. At the initialization stage, a set of food sources is randomly chosen by bees, with the amount of nectar determined.

The main steps of the ABC algorithm are as follows [14]:

1. Initializing Population Repeat
2. Place bees that are employed in food sources
3. Place the observer bee on a food source
4. depending on the amount of nectar
5. Send a lookout to the search area to determine
6. a new food source
7. Memorize the best food sources found
8. Repeat this process until the requirements are met.

Classification

In this last stage, the classification of kidney disease used the Modified K-Nearest Neighbor (MKNN) method combined with the GA method. The kidney disease data used results from feature selection using the ABC algorithm in the previous stage.

The KNN method has two parameters: the distance between the test data and training data and the nearest neighbour's distance (parameter k) [15]. The classification process of the KNN method begins by calculating the distance between the test data and training data. Then, sorting based on the smallest distance, determining the closest k parameter, after that, selecting the shortest distance according to the number of k parameters. Finally, and voting on the largest class so that the class can be predicted.

This algorithm's performance is also great, depending on the quality of the data used [16]. The KNN method requires development to improve its performance, namely the Modified KNN. In KNN, the classification process is only done by calculating the distance and voting based on the k parameter. While in the MKNN method, there is a calculation of the value of validity and weighted voting, which can increase the resulting level of accuracy compared to the traditional KNN [17].

The Modified K-Nearest Neighbor (MKNN) method is an initial KNN algorithm. At MKNN, there is an additional step of classification, with the class labeling determined based on the parameter k at the training's validated data points.

Furthermore, the KNN is weighted on the test data.

MKNN pseudo-code algorithm, according to [6]:

```
Output_label := MKNN (train_set, test_sample)
Begin
    For i := 1 to train_size
        Validity(i) := Compute Validity of i-th
        sample;
    End For;
    Output_label := Weighted_KNN(Validity,
    test_sample);
    Return Output_label;
End
```

Every training data needs to go through a validation process. After all the training data has been validated, this process provides more information on points [18].

$$Validity(\text{x}) = \frac{1}{H}\sum_{i=1}^{H} S(lbl(x), lbl(N_i(x))) \qquad (2)$$

Where:
 H   = Number of closest points
 $lbl(x)$ = Class x
 $N_i(x)$ = Label the closest point class x
 S   = The function used to calculate the similarity between the x point and the i-th nearest neighbor data

$$S(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \qquad (3)$$

Where:
 a   = Class A on training data
 b   = Other classes besides A on the training data

In addition to using the closest neighbor k, weighted voting was conducted in each of the major voting samples. Furthermore, the validity of each training data is multiplied by the weight based on the Euclidean distance.

$$W(i) = Validity(i) \left(\frac{1}{Ed + 0,5}\right) \qquad (4)$$

Where:
 W(i)   = Weight voting calculation
 Validity(i) = Validity value
 Ed = Euclidean distance
Euclidean distance (Ed) equation as follows:

$$dist(x_l, x_u) = \sqrt{\sum_{i=1}^{n} (x_{li} - x_{ui})^2} \qquad (5)$$

Where:
 $dist(x_l, x_u)$ = Euclidean distance
 n = Number of data
 x_l = Training data ($x_l = x_{l1}, x_{l2}, ..., x_{ln}$)
 x_u = Testing data ($x_u = x_{u1}, x_{u2}, ..., x_{un}$)
 i = $1, 2, ..., n$

A genetic algorithm is a proven and effective technique used in data mining and pattern recognition. Besides being useful for optimization, it can also be directly used as a classifier. The genetic algorithm method is very popular because it has been successfully applied to many optimization problems. One of its advantages is the ability to complete searches with enormous dimensions of space. In this study, the GA method was combined with the MKNN method to classify kidney disease where GA played a role in determining the optimal k parameter. The GMKNN-ABC method in the classification of kidney disease is compared with other methods to determine the proposed method's performance. The performance of GMKNN-ABC is compared to the MKNN method without using ABC as feature selection and GA optimization, and the MKNN method uses ABC as feature selection. The method's performance applied in this kidney disease classification study was measured by a confusion matrix, namely accuracy, precision, recall, and f1-score. These values were measured compared to the three methods, namely MKNN, MKNN-ABC, and GMKNN-ABC.

Experimental Setup

The dataset used to carry out this research was obtained from the Chronic Kidney Disease center at the University of California repository (UCI Repository Machine Learning Benchmark). The dataset can be downloaded by open access at
http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

The kidney disease dataset consists of 400 records with 24 features and two classes, namely 250 CKD classes and 150 not CKD classes. The dataset is divided into 280 training data and 120 testing data.

In this study, the classification of kidney disease was tested in several methods, namely MKNN, MKNN-ABC, and GMKNN-ABC. It aims to compare the performance of each method in classifying kidney disease.

RESULTS AND DISCUSSION

Classification of kidney disease is conducted by using the ABC method for feature selection and Genetic Modified KNN. GMKNN-ABC method performance testing results are seen from the resulting confusion matrix, namely accuracy, precision, recall, and f1-score. Testing is conducted by limiting the parameters k = 1 to

50. Out of the 24 features in the chronic kidney disease dataset, only 1, 2, 5, 7, 11, 13, 14, 19, and 21 were used. These include blood pressure, specific gravity, red blood cells, pus cell clumps, serum creatinine, potassium, haemoglobin, diabetes mellitus, and appetite. The selected features are used in the classification of kidney disease. The results of the confusion matrix based on the parameter k are shown in Table 2.

The confusion matrix results show that the values of accuracy, precision, recall, and f1-score provided high enough results, with an average above 90%. In the form of a graph, the confusion matrix test results are shown in Figure 2.

The confusion matrix graph shows that the large value of parameter k has no significant effect on the level of accuracy, precision, recall, and f1-score with the production of high results above 90%. To test the GMKNN-ABC method's performance, the MKNN, MKNN-ABC, and GMKNN-ABC are compared, as shown in Figure 3. Figure 3 shows a curve of the comparison of accuracy levels.

Table 2. Confusion Matrix Testing the GMKNN-ABC Method

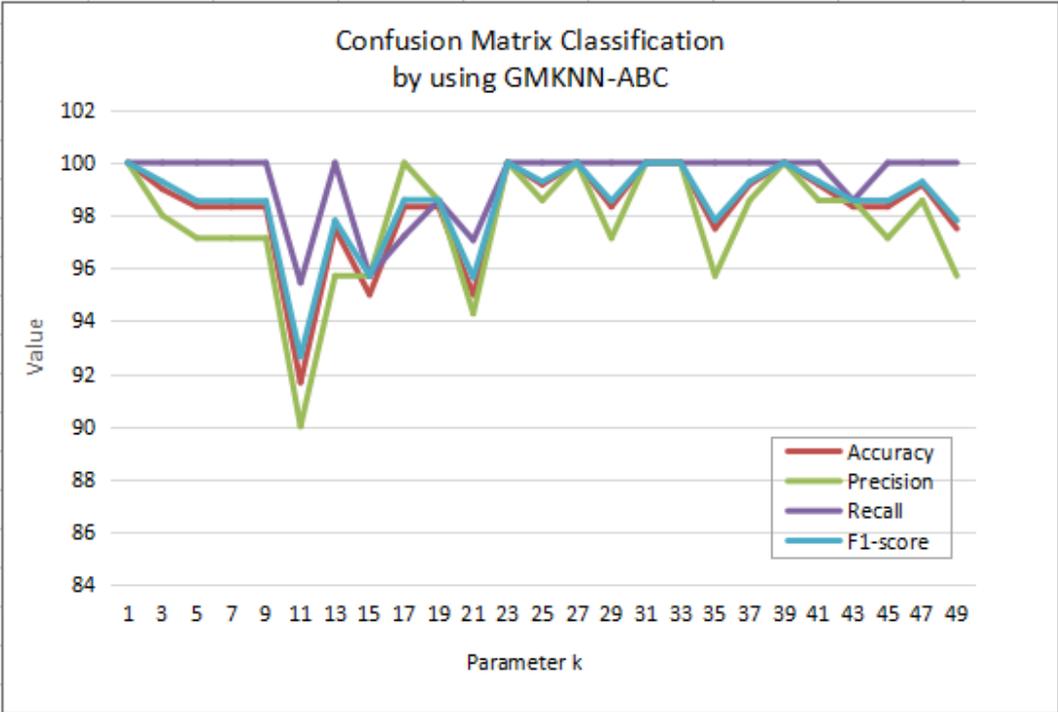| Parameter k	Confusion Matrix (%)			
	Accuracy	Precision	Recall	F1-score
1	100.00	100.00	100.00	100.00
3	99.17	98.57	100.00	99.28
5	98.33	97.14	100.00	98.55
7	98.33	97.14	100.00	98.55
9	98.33	97.14	100.00	98.55
11	91.67	90.00	95.45	92.65
13	97.50	95.71	100.00	97.81
15	95.00	95.71	95.71	95.71
17	98.33	100.00	97.22	98.59
19	98.33	98.57	98.57	98.57
21	95.00	94.28	97.06	95.65
23	100.00	100.00	100.00	100.00
25	99.17	98.57	100.00	99.28
27	100.00	100.00	100.00	100.00
29	98.33	97.14	100.00	98.55
31	100.00	100.00	100.00	100.00
33	100.00	100.00	100.00	100.00
35	97.50	95.71	100.00	97.81
37	99.17	98.57	100.00	99.28
39	100.00	100.00	100.00	100.00
41	99.17	98.57	100.00	99.28
43	98.33	98.57	98.57	98.57
45	98.33	97.14	100.00	98.55
47	99.17	98.57	100.00	99.28
49	97.50	95.71	100.00	97.81

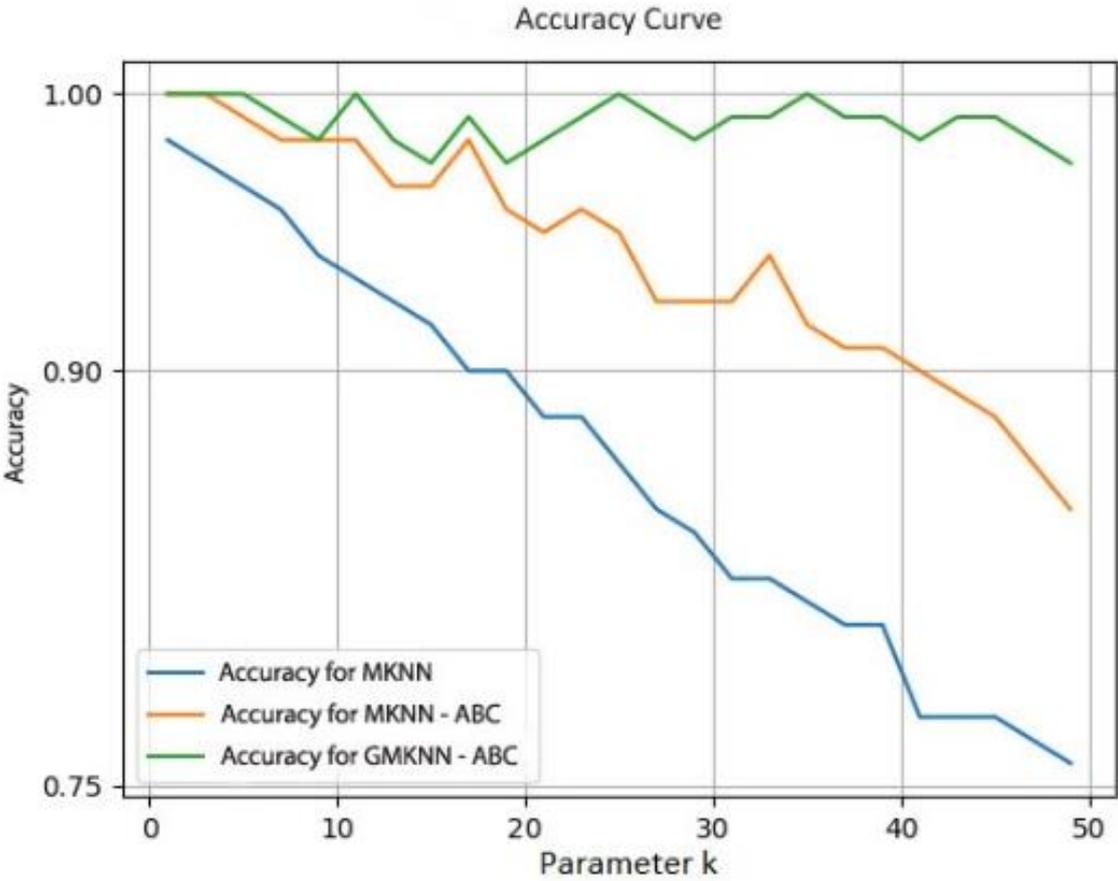Figure 2. GMKNN-ABC Confusion Matrix Test Results

Figure 3. The Accuracy Curve of MKNN, MKNN-ABC and GMKNN-ABC

The size of the parameter k greatly influences the level of accuracy of the MKNN test. Therefore, the greater the parameter k, the more the level of accuracy is obtained. However, the lowest accuracy is achieved when the parameter interval k ranges from 40 to 50. Meanwhile, the MKNN-ABC test gives a higher accuracy level than the MKNN, decreasing its accuracy level for the increasing k parameter. In GMKNN-ABC, the level of accuracy produced is more stable and more elevated than MKNN and MKNN-ABC. The size of the parameter k does not influence the high level of accuracy produced. Table 3 shows the comparison of the average Confusion Matrix results from the MKNN, MKNN-ABC and GMKNN-

ABC. Table 4 display a comparison with other related works.

A comparison of the average performance of the MKNN, MKNN-ABC, and GMKNN-ABC methods in a confusion matrix is displayed in the form of a bar graph in Figure 4.

Table 3. Comparison of Average Confusion Matrix

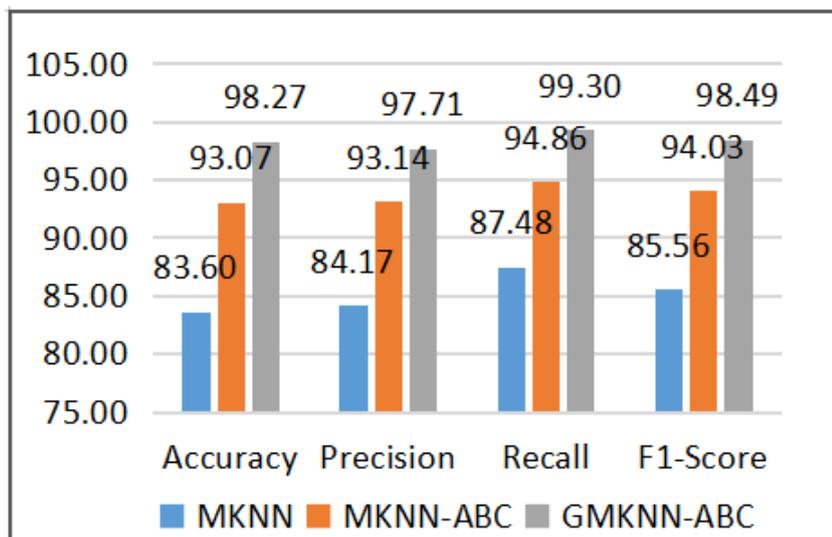| CONFUSION MATRIX	MKNN	MKNN-ABC	GMKNN-ABC
ACCURACY	83.60	93.07	98.27
PRECISION	84.17	93.14	97.71
RECALL	87.48	94.86	99.30
F1-SCORE	85.56	94.03	98.49

Figure 4. Comparison of Average Confusion Matrix Graphs

Table 4. Comparing with other related works

No.	Author	Year	Method	Accuracy (%)
1.	Kunwar et al.	2016	ANN [19]	72.73
2.	Gharibdousti et al.	2017	NB [20]	96.70
3.	Subas, Alickovic, and Kevric	2017	ANN [21]	98.00
4.	Fadilla, Adikara, and Perdana	2018	ELM [22]	96.70
5.	Abdelaziz et al.	2019	Hybrid LR and NN [23]	97.80
6.	Elhoseny, Shankar, and Uthayakumar	2019	D-ACO [24]	95.00
7.	Rady and Anwar	2019	Probabilistic Neural Network [25]	96.70
8.	Hamedan et al.	2020	Fuzzy Expert System [26]	92.13
9.	**Proposed method**	**2020**	**MKNN***	**83.60**
			MKNN-ABC*	**93.07**
			GMKNN-ABC*	**98.27**

Figure 4 shows that the GMKNN-ABC method provides the best results compared to MKNN and MKNN-ABC. Meanwhile, feature selection is very influential on the level of accuracy, precision, recall, and f1-score produced. The GMKNN-ABC method provides the highest average yield above 95% compared to the MKNN and MKNN-ABC.

CONCLUSION

This study showed that the ABC method was implemented as a GMKNN feature selection that influences kidney diseases classification. The comparison results of the average confusion matrix testing on the MKNN, MKNN-ABC, and GMKNN-ABC methods show that the GMKNN-ABC provides the best average results level accuracy, precision, recall, and f1-score compared to others. In addition, the GMKNN-ABC method produces an average accuracy of 98.27%, which is the highest compared to MKNN, MKNN-ABC and other methods. The method proposed in

this study, especially the GMKNN-ABC method, has never been carried out by previous studies. The increase in accuracy resulting from the GMKNN-ABC method in the classification of kidney disease contributes to the study.

REFERENCES

[1] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems with Applications,* vol. 41, no. 5, pp. 2239-2249, April 2014, DOI: 10.1016/j.eswa.2013. 09.022

[2] P. Shunmugapriya and S. Kanmani, "A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid)," *Swarm and Evolutionary Computation*, vol. 36, pp. 27–36, October 2017, DOI: 10.1016/j.swevo. 2017.04.002

[3] A. Noor et al., "Melastoma Malabathricum L. Extracts-Based Indicator for Monitoring Shrimp Freshness Integrated with Classification Technology using Nearest Neighbours Algorithm," *SINERGI,* vol. 25, no. 1, pp. 69-74, February 2021, DOI: 10.22441/sinergi.2021..1.009

[4] D. L. Z. Astuti, S. Samsuryadi and D. P. Rini, Real-Time Classification of Facial Expressions using a Principal Component Analysis and Convolutional Neural Network, *SINERGI,* vol 23, no. 3, pp. 239,244, October 2019, DOI: 10.22441/sinergi.2019.3.00

[5] A. Behori, "Komparasi Model Neural Network dan K-NN untuk Mengklasifikasi Kondisi Penderita Penyakit Hepatitis," *AiTech: Jurnal Manajemen Informatika dan Komputer,* vol. 3, no. 2, pp. 82–88, 2017

[6] H. Parvin, H. Alizadeh, and B. Minaei-bidgoli, "MKNN: Modified K-Nearest Neighbor," *Proceeding of the World Congress on Engineering and Computer Science,* San Francisco, USA, October 2008, 4 pages

[7] R. Verma, R. Handa and V. Ouri, "Feature Selection Using Genetic Algorithm for Cancer Prediction System,: In: Hura G., Singh A., Siong Hoe L. (eds) *Advances in Communication and Computational Technology. Lecture Notes in Electrical Engineering*, vol 668, Springer, Singapore, DOI: 10.1007/978-981-15-5341-7_91

[8] A. S. Levey *et al.,* "The definition, classification, and prognosis of chronic kidney disease: A KDIGO Controversies Conference report," *Kidney Int.*, vol. 80, no. 1, pp. 17–28, 2011, DOI: 10.1038/ki.2010.483

[9] I. Zyout, I. Abdel-qader, and C. Jacobs, "Embedded Feature Selection using PSO-kNN: Shape-Based Diagnosis of Microcalcification Clusters in Mammography," *Journal if Ubiquitoer System and Pervasive,* no. 3, vol. 1, 2011, pp. 7–11, 2011, DOI: 10.5383/JUSPN.03.01.002

[10] H. Ghimatgar, K. Kazemi, M. Sadegh, and A. Aarabi, "An improved feature selection algorithm based on graph clustering and ant colony optimization," *Knowledge-Based Systems*, vol. 159, pp. 270–285, November 2018, DOI: 10.1016/j.knosys.2018.06.025

[11] M. Schiezaro and H. Pedrini, "Data feature selection based on Artificial Bee Colony algorithm," *EURASIP Journal on Image and Video Processing,* no. 47, 2013, DOI: 10.1186/1687-5281-2013-47

[12] F. Wafiyah, N. Hidayat, and R. S. Perdana, "Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Demam," *Jurnal Pengembamgam Teknologi Informasi dan Ilmu Komputer,* vol. 1, no. 10, pp. 1210–1219, October 2017

[13] M. Afizi and M. Shukran, "Artificial Bee Colony based Data Mining Algorithms for Classification Tasks," *Modern Applied Science,* vol. 5, no. 4, pp. 217–231, 2011, DOI: 10.5539/mas/v5n4p217

[14] D. Karaboga and B. Akay, "A comparative study of Artificial Bee Colony algorithm," *Applied Mathematics and Computation*, vol. 214, no. 1, pp. 108–132, 2009, DOI: 10.1016/ j.amc.2009/03.090

[15] Y. Zhang, G. Cao, B. Wang, and X. Li, "A novel ensemble method for k -nearest neighbor," Pattern Recognit., vol. 85, pp. 13–25, 2019

[16] Z. Liu, Q. Pan, and J. Dezert, "A new belief-based K-nearest neighbor classification method," *Pattern Recognition,* vol. 46, no. 3, pp. 834–844, 2013, DOI: 10.1016/j.patcog. 2012.10.001

[17] M. R. Ravi and S. Adinugroho, "Implementasi Algoritme Modified K-Nearest Neighbor (MKNN) Untuk Mengidentifikasi Penyakit Gigi dan Mulut," *JPTIIK: Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 3, pp. 2596–2602, 2019

[18] H. Parvin, H. Alizadeh, and B. Minati, "A Modification on K-Nearest Neighbor Classifier," *Global Journal of Computer*

Science Technology, vol. 10, no. 14, pp. 37–41, November 2010

[19] V. Kunwar, K. Chandel, A. S. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, India, 2016, pp. 300-305, DOI: 10.1109/CONFLUENCE.2016.7508132

[20] M. S. Gharibdousti, K. Azimi, S. Hathikal, and D. H. Won, "Prediction of Chronic Kidney Disease using data mining techniques," *Proceedings of the 2017 Industrial and Systems Engineering Conferenc*e, May 2017, pp. 2135–2140

[21] A. Subas, E. Alickovic, and J. Kevric, "Diagnosis of Chronic Kidney Disease by Using Random Forest," *in Part of the IFMBE Proceedings book series,* 2017, vol. 62, pp. 589–594, DOI: 10.1007/978-981-10-4166-2

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

[23] A. Abdelaziz, A. S. Salama, A. M. Riad, and A. N. Mahmoud, "A Machine Learning Model for Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities," In: Hassanien A., Elhoseny M., Ahmed S., Singh A. (eds) Security in Smart Cities: *Models, Applications, and Challenges. Lecture Notes in Intelligent Transportation and Infrastructure*, Springer, Cham, DOI: 10.1007/978-3-030-01560-2_5

[24] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease," *Scientific Report*, vol. 9, no. 9583, 2019, 10.1038/s41598-019-46074-2

[25] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, no. 100178, March 2019, DOI: 10.1016/j.imu.2019.100178

[26] F. Hamedan, A. Orooji, H. Sanadgol, and A. Sheikhtaheri, "Clinical decision support system to predict chronic kidney disease: A fuzzy expert system approach," *International Journal of Medical Informatics*, vol. 138, no. 104134, June 2020, DOI: 10.1016/j.ijmedinf.2020.104134