Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

An Improved K-NN Algorithm and Bagging for Liver Disease Classification

Anindya Khrisna Wardhani[1], Lakhmudien[2], Astrid Novita Putri[3], Salim Fathi Salim Ashour[4]

[1.2,] Rekam Medis dan Informasi Kesehatan, [3]Teknik Informatika, [4]Computer Science
[1.2,]Politeknik Rukun Abdi Luhur, [3] Universitas Semarang, [4]Elmergib University
E-mail : anindya.khrisna@poltekun.ac.id[1], lakhmudien@poltekun.ac.id[2], astrid@usm.ac.id[3], salemashour992@gmail.com[4]

ARTICLE INFO

History of the article:
Received January 12, 2022
Revised Mei 21, 2022
Accepted June 17, 2019

Keywords:
K-NN, Bagging,
Data Mining,
Clasification,
Liver

Correspondece:
E-mail:
Anindya.khrisna@poltekun.ac.id

ABSTRACT

The function of the liver is to detoxify toxins in the human body and control cholesterol and fat in the human body. If the liver is damaged, health will be disturbed, even death. A lot of data on the liver disease can be used to predict liver disease. This study aims to improve the accuracy of liver disease classification using K-NN and bagging methods. The experimental results in this study are the bagging method can improve the performance accuracy of the K-NN prediction model even though it is based on the T-test even though there is only a slight change in accuracy. In this study, the accuracy value using the K-NN method was 78.56%. For the highest accuracy value of 99.83% using the K-NN model which is integrated with bagging. From the results of experiments carried out in this study, the K-NN model with bagging can certainly improve performance on the prediction model of liver disease classification. So that the predictions made can be more accurate and can be used to predict liver disease.

INTRODUCTION

The liver is a vital organ for humans. This organ is located in the right abdominal cavity, just below the diaphragm. There are several functions of the liver, including as an antidote and penetration of toxins, regulate hormones, regulating the composition of blood containing fat, sugar, protein, and other substances. The liver also works to make bile, a substance that helps digest fats. Liver disease is a disorder of any liver function. The liver is responsible for critical functions in the body, where these functions can cause significant damage to the body. The liver is the only organ in the body that can easily replace damaged cells, but if these cells are lost, the liver cannot meet the body's needs. Liver disease is often referred to as a silent killer because it may not cause symptoms. The problem that usually occurs is the difficulty recognize the liver disease early, even when it has spread. Even though knowing the symptoms of liver disease from an early age is very much needed, so that patients can take proper treatment(Elly Pusporani, 2019).

In the health and medical industry, predicting disease is critical and requires effective decisions in analysing and predicting the accuracy of a patient's disease. In diagnosing the presence or absence of liver disease, references from the results of liver function tests carried out in the laboratory can be used. These tests include serum transaminases, alkaline phosphatase, total bilirubin, conjugated bilirubin, total protein, albumin, and the ratio of albumin and globulin. The test results show that the test results are significant as a feature of whether or not liver function disorders are present (Handayani et al., 2019). One method that

can use for classification is to use data mining (Wardhani et al., 2018). Data mining is a series of processes to explore added value from a data set in the form of knowledge that has not been known manually(Bimantoro & Wardhani, 2020)

Several methods can be used for classification cases, including logistic regression, nave Bayes, k-nearest neighbour (K-NN), and support vector machine (SVM) (Wardhani, 2017b). This study uses the K-Nearest Neighbor (K-NN) algorithm. The K-NN algorithm is a classification algorithm based on the proximity of data to other data(Abdul Mukid & Rusgiyono, 2017). Q-dimensional can calculate the q-dimensional data and the distance from the data to other data in the K-NN algorithm. This distance value is used as the value of the proximity/similarity between the test data and the training data(Wardhani, 2017a). The value of K in K-NN means the closest K-data from the test data. The advantage of K-NN has a simple principle works based on the shortest distance between the test sample and the training sample. The K-NN algorithm proved outstanding, with high accuracy and low statistical error (Mandong & Munir, 2018).

In a previous study, based on SVM and K-NN accuracy scores on experimental test results, SVM was better in predicting liver disease performance than the K-NN algorithm (Assegie, 2021). Based on the value of accuracy and precision, the SVM method gives the best results but is based on recall, and the K-NN method gives the best results. Although SVM gives the highest accuracy and precision values, there is a large discrepancy between the resulting precision and recall values compared to the difference in accuracy and recall values from the K-Nearest Neighbor method(Sari, 2020). The low value of K-NN accuracy can be caused by several conditions, such as unstable class conditions. Class problems with unbalanced data can hurt K-NN performance which can cause the algorithm to overfit data and poor accuracy (Fakhruzi, 2018). An unbalanced data class is a condition where the distribution of data classes is not balanced. The number of data classes (instances) is less or more than the number of other data classes. The data class group with the minor data is called the minority group (minority), while the data group with the most data is called the majority group (majority)(Sashidhar & Kutz, 2021). Data with unbalanced classes usually have problems classifying machine learning because the amount of data per class is not distributed evenly/normally. This condition is usually found in credit, health and other data (Giovanini et al., 2018).

Researchers have taken many approaches to overcome the problem of unbalanced classes. One effective way to deal with this problem is the ensemble method(Athanasopoulos et al., 2018). In the ensemble method, boosting and bagging are two popular ways that have been proven to increase the accuracy of prediction models or learning algorithms. However, the bagging method performs better for class imbalance problems than other methods (Assegie, 2021)

In addition to the unbalanced class problem, in the K-NN Algorithm, there is one parameter, namely k. Based on previous research, the selection of the k value is important because it will affect the performance of the K-NN algorithm (Angreni et al., 2019). If the value of k is too small, then the classification results will be more affected by noise. On the other hand, if the value of k is too high, it will reduce the effect of noise on the classification but make the boundaries between each classification more blurred. A good value of k can be selected by parameter optimization, for example, by using cross-validation.

Based on the description of the problem, the researcher will optimize the K-NN algorithm using bagging and cross-validation when selecting the K value on the K-NN method to increase accuracy so that the predictions made can be more accurate and can be used to predict liver disease.

RESEARCH METHODS

This study uses a computer with specifications for the Windows 10 operating system, an Intel Core i5 processor, and 4 GB of memory. Analysis of data processing in this study using a rapid miner. The research took the dataset from the UCI machine learning repository, a dataset of liver disease patients in India. This data set contains 416 liver patient records and 167 non-liver patient records. Data collected the data set from northeast Andhra Pradesh, India. The class label is divided into two groups (liver patients or not). This data set contains 441 male patient records and 142 female patient records. The attributes used were patient age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, natural aminotransferase, aspartate aminotransferase, total protein, albumin (ALB), albumin ratio and globulin ratio.

This research consists of several stages. First, the dataset will be preprocessed before being processed to produce an accurate classification, and then use the ensemble bagging method to deal with class imbalances. As determined by the K-NN algorithm, the bagging method will perform as many iterations (factors). Validation in this study uses cross-validation. The purpose of this experiment is to produce the performance of the classification model using K-NN. K-NN can measure performance based on accuracy.

The K-NN method is one of the methods used in grouping data. The working principle of K-NN is to find the closest distance between the data to be evaluated and its k closest neighbours in the training data (Sari, 2020). This algorithm only performs feature vector storage and classification of learning data in the learning phase. In the same vector, classification is calculated for the learning data. The same vector classification is calculated for the test data (whose classification is unknown). Finally, the distance from this new vector to all learning data vectors is calculated, and the closest k numbers are taken. In its application K-NN has the following steps in its application:

1. Determine the parameter k (number of close neighbours). There is no exact formula for determining the value of k. However, one tip that can be considered is if the class number is even then the k value should be odd, otherwise if the class number is odd then the k value should be even. In this case we use k = 3.
2. After determine the parameter k, calculate the square of the object's Euclidean distance to the given training data.

$$d_{ij} = \sqrt{\sum_{i=1}^{p}(x_{ik} - x_{jk})^2} \qquad (1)$$

with,

X_{ik} = Data Training

X_{jk} = Data Testing

ij = Variabel Data

d = Distance

p = Dimention Data

3. Sort the calculation results of step number 2 in ascending order.
4. Collect category Y (Classification of nearest neighbours based on the value of k).
5. The object category can be predicted by using the majority nearest neighbour category.

The working principle of K-NN is to find the closest distance between the data to be evaluated and the closest K-NN in the training data. Equation 1 shows the calculation formula to find the closest distance using the Euclidean formula (Sutanto, 2009).

After performing classification calculations, a bagging method is given to determine the best accuracy. Bagging stands for bootstrap aggregating, it uses a sub-dataset (bootstrap) to generate a training set L (learning), L trains the learning base using an unstable learning procedure, and then, during testing, takes the average[30]. Bagging is good for classification and regression. In the case of regression, to be more robust, one can take the average when combining predictions. Bagging is a learning algorithm that is stable at small changes in the training set causing large differences in the resulting learners, namely the learning algorithm on data that has high variance (noise). Bagging is able to improve accuracy significantly greater than individual models, and is stronger against noise and overfitting effects than the original training data[31]

Bagging can improve performance by combining (ensemble) algorithms such as Decision Tree (DT), Neural Network (NN), K-NN and Support Vector Machine (SVM)[31]. Datasets with high noise cause errors in generalization of classification, so it takes the right algorithm to be combined (ensemble) with the neural network so that prediction accuracy can increase.

The Bootstrap technique is a computer-based method for estimating the standard error of an empirical distribution. This technique estimates the significance level of the messy statistic when the shape of the distribution is unknown. This method takes advantage of being completely automated and requires no theoretical calculations or assumptions on the original distribution. The population distribution method is based on the bootstrap principle which can be used by plug-in principles to estimate the parameters of the sample. Bagging Algorithm [31] :

Loop for b = 1, 2, . . ., B

1. Generate a boostrap sample $\{(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_n^*, Y_n^*)\}$ by randomly replacing the training data $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ matches the C_b classifier turned on on the bootstrap-compliant sample.
2. Final classifier output:

$$(x) = B{-}1 \sum_{b=1}^{B} Cb\ (x)$$

More details can be seen in Figure 1 below:

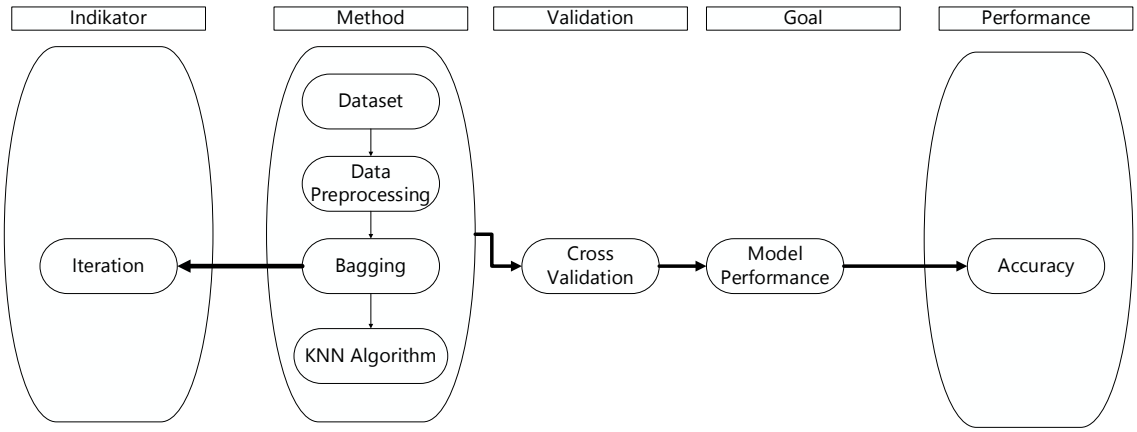

Figure 1 Research Flow

The purpose of this study is to determine or improve the accuracy of K-NN. The researcher uses the K-NN algorithm which is integrated with bagging so that it is expected to increase the accuracy of data

classification. Researchers use the following method to reduce errors in determining the classification. More details can be seen in Figure 2 below:

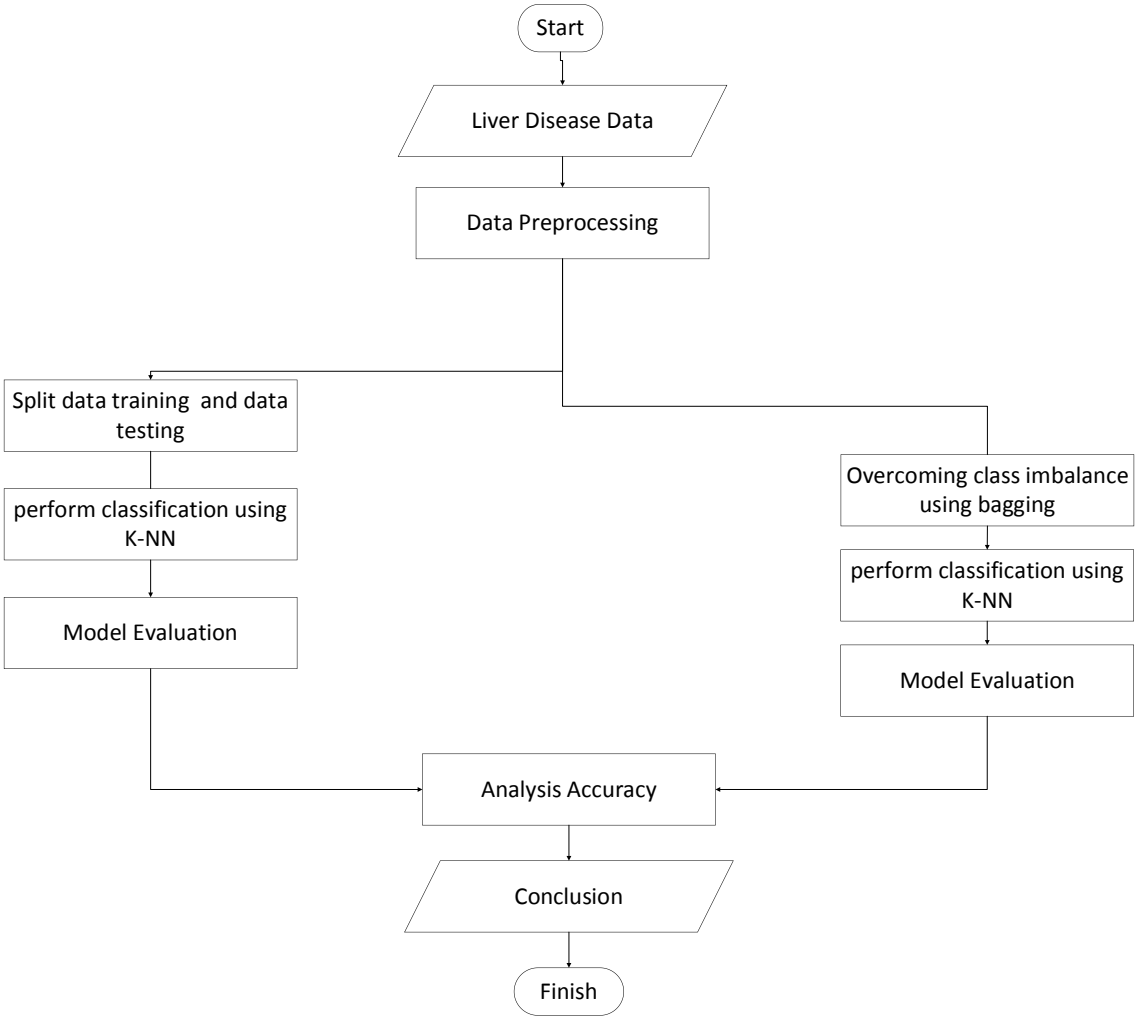

Figure 2 System Flowchart

RESULTS AND DISCUSSION

At the initial stage, data preprocessing will be carried out using a sample liver disease dataset from northeast Andhra Pradesh, India, as shown in Table 1. The variables contained in the dataset consist of age, gender, total bilirubin (TB), Alkaline Phosphatase (Alkphos), Alamin Aminotransferase (Sgpt), Aspartate Aminotransferase (Sgot), Total Protein (TP), Albumin (ALB), Albumin Ratio and Globulin Ratio (A/G), Selector Field.

Table 1 Sample Data
(Source: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)#)

Age	gender	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G	Selector field
65	Female	0,7	0,1	187	16	18	6,8	3,3	0,9	Liver
62	Male	10,9	5,5	699	64	100	7,5	3,2	0,74	Liver
62	Male	7,3	4,1	490	60	68	7	3,3	0,89	Liver
58	Male	1	0,4	182	14	20	6,8	3,4	1	Liver
72	Male	3,9	2	195	27	59	7,3	2,4	0,4	Liver

46	Male	1,8	0,7	208	19	14	7,6	4,4	1,3	Liver
26	Female	0,9	0,2	154	16	12	7	3,5	1	Liver
29	Female	0,9	0,3	202	14	11	6,7	3,6	1,1	Liver
55	Male	0,7	0,2	290	53	58	6,8	3,4	1	Liver

Data preprocessing is a set of techniques applied to the database to remove noise, missing values, and inconsistent data. Data preprocessing is divided into several steps: data cleaning, transformation, and reduction. After the data is preprocessed, the data will be split using cross-validation to separate the testing and training data. After being separated using cross-validation, the next step is calculating the classification using the K-NN algorithm. After being processed using the K-NN algorithm, the bagging method performs 10 models and performs a classification based on the results of the best classification accuracy values. Based on the results of the K-NN modelling with the bagging method with a value of k = 3 in the dataset, it calculates the model performance by creating a confusion matrix, as shown in Table 2.

Tabel 2 Confusion Matrix Calculation Result of K-NN Classifier

	True Liver	**True NonLiver**
Pred. Liver	374	83
Pred. NonLiver	42	84

Based on the confusion matrix table data, it can be concluded that:

a. True Positive (TP) = 374 data from one liver class that data can predict correctly in the liver class.

b. True negative (TN) = 83 data from one NonLiver class that data can predict correctly in the NonLiver class

c. False positive (FP) = 42 data from conditions in which the Liver class has a wrong prediction in the NonLiver class, while

d. False negative (FN) = 84 data from conditions in the NonLiver class, which data is predicted to be wrong in the liver class.

e. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{374+84}{374+84+42+83} = 78{,}56\ \%$

From the results of research using the K-NN algorithm, an accuracy of 78.56% is obtained. In the next stage, the researcher will compare using the K-NN algorithm, which is integrated using bagging to overcome unbalanced classes. Like the previous stages, the data will be preprocessed before processing. Based on the results of the K-NN modelling with the bagging method with a value of k =3 in the dataset, it calculates the model performance by creating a confusion matrix, as shown in Table 3.

Table 3 Confusion Matrix Calculation Result of K-NN Classifier + Bagging

	True Liver	**True NonLiver**
Pred. Liver	416	1
Pred. NonLiver	0	166

Based on the confusion matrix table data, it can be concluded that:

a. True Positive (TP) = 416 data from one liver class that data can predict correctly in the liver class.

b. true negative (TN) = 166 data from one NonLiver class that can be predicted correctly in the NonLiver class

c. false positive (FP) = 0 data from the condition where the Liver class's prediction is wrong in the NonLiver class,

d. false negative (FN) = 1 data from conditions in the NonLiver class, which data predicted to be wrong in the liver class.*Accuracy* = $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{416+166}{416+166+0+1} = 99{,}83\ \%$

Based on the results of the accuracy value, accuracy can see that the K-NN method using bagging has a superior accuracy value of 99.83% compared to the K-NN method without bagging, which only reaches 78.56% accuracy. Furthermore, this high accuracy shows that the accuracy value of K-NN using bagging is also superior to the SVM method in the previous study (4), which only achieved an accuracy of 75.21%.

CONCLUSIONS AND RECOMMENDATIONS

Based on the discussion discussed in the previous section, the conclusion can conclude that the K-NN method using the optimized bagging method using a value of k=2 can increase the accuracy of the K-NN method. With the increase in accuracy, it is hoped that high accuracy can use to implement liver disease classification.

REFERENCES

Abdul Mukid, M., & Rusgiyono, A. (2017). Analisis Credit Scoring Menggunakan Metode Bagging K-Nearest Neighbor. *Jurnal Gaussian*, *6*(1), 161–170. http://ejournal-s1.undip.ac.id/index.php/gaussian

Angreni, I. A., Adisasmita, S. A., Ramli, M. I., & Hamid, S. (2019). Pengaruh Nilai K Pada Metode K-Nearest Neighbor (Knn) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan. *Rekayasa Sipil*, *7*(2), 63. https://doi.org/10.22441/jrs.2018.v07.i2.01

Assegie, T. A. (2021). Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model. *Indonesian Journal of Electronics, Electromedical, and Medical Informatics (IJEEEMI)*, *3*(1), 9–14. http://ijeeemi.poltekkesdepkes-sby.ac.id/index.php/ijeeemi

Athanasopoulos, G., Song, H., & Sun, J. A. (2018). Bagging in Tourism Demand Modeling and Forecasting. *Journal of Travel Research*, *57*(1), 52–68. https://doi.org/10.1177/0047287516682871

Bimantoro, T., & Wardhani, A. K. (2020). Implementasi Algoritma Partitioning Around Medoids Dalam Pengelompokan Restoran. *Indonesian Journal of Technology, Informatics and Science (IJTIS)*, *2*(1), 33–36. https://doi.org/10.24176/ijtis.v2i1.5651

Elly Pusporani, S. Q. I. (2019). Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning. *INFERENSI, Vol. 2(1)*.

Fakhruzi, I. (2018). An artificial neural network with bagging to address imbalance datasets on clinical prediction. *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, *2018-Januari*(1), 895–898. https://doi.org/10.1109/ICOIACT.2018.8350824

Giovanini, L. H. F., Manffra, E. F., & Nievola, J. C. (2018). Evolutionary ensemble approach for behavioral credit scoring. *Springer Nature 2018, June*, 350–357. https://doi.org/10.1007/978-3-319-93713-7

Handayani, P., Nurlelah, E., Raharjo, M., Madya Ramdani, P., Nusa Mandiri Jakarta Jln Damai No, S., Jati, W., Satwa, M., Minggu, P., & Selatan, J. (2019). *Prediksi Penyakit Liver Dengan Menggunakan Metode Decision Tree Dan Neural Network* (Vol. 4, Issue 1).

Mandong, A.-M., & Munir, U. (2018). *Smartphone Based Activity Recognition using K-Nearest Neighbor Algorithm*. https://www.researchgate.net/publication/328738427

Sari, R. (2020). Analisis Sentimen Pada Review Objek Wisata Dunia Fantasi menggunakan Algoritma K-Nearest Neighbor (K-NN). *Jurnal Sains Dan Manajemen*, *8*(1). www.tripadvisor.com.

Sashidhar, D., & Kutz, J. N. (2021). *Bagging, optimized dynamic mode decomposition (BOP-DMD) for robust, stable forecasting with spatial and temporal uncertainty-quantification*. http://arxiv.org/abs/2107.10878

Wardhani, A. K. (2017a). *Penerapan Algoritma Partitioning Around Medoids Untuk Menentukan Kelompok Penyakit Pasien (Studi Kasus : Puskesmas Kajen Pekalongan)* (Vol. 6, Issue 1).

Wardhani, A. K., Widodo, C. E., & Suseno, J. E. (2018). *Information System for Culinary Product Selection Using Clustering K-Means and Weighted Product Method*.

Wardhani, A. Khrisna. (2017b). Penerapan Algoritma Partitioning Around Medoids Untuk Menentukan Kelompok Penyakit Pasien (Studi Kasus : Puskesmas Kajen Pekalongan). *Jurnal Kajian Ilmu Dan Teknologi (KILAT)*, *6*(1), 6–10.