Prediction of Life Expectancy of Lung Cancer Patients Post Thoracic Surgery using K-Nearest Neighbors and Bat Algorithm

Muhamad Nur Arifiansyah [1*], Anggyi Trisnawan Putra [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: mnurarifiansyah@students.unnes.ac.id

ARTICLE INFO

Article history
Received 26 October 2022
Revised 4 November 2022
Accepted 7 November 2022

Keywords
Data Mining
Thoracic Surgery
Bat Algorithm
KNN Algorithm

ABSTRACT

Lung cancer is one of the deadliest cancers, accounting for 11.6% of cancer diagnoses in the world. Death in lung cancer patients can occur in various ways and one of the treatments for lung cancer patients that can be done is thoracic surgery. Thoracic surgery is generally considered a medium risk procedure, but thoracic surgery has a high risk, one of the risks is that if the patient loses blood which will result in the death of the patient. In this study, the method used to implement predictive life expectancy in post-thoracic surgery patients is the bat algorithm for feature selection and the KNN algorithm for classifying data. The dataset used in this study was obtained from the UCI Machine Learning Repository, namely the thoracic surgery dataset which contains 470 data with 16 attributes. The results of the study in predicting the life expectancy of patients after thoracic surgery were carried out with 3 tests. The first test is testing the population with the best accuracy of 87.23%, the second test is convergent testing with the best accuracy of 87.23% and the third test is the comparison test of KNN which produces the best accuracy of 87.23%. The bat algorithm succeeded in increasing the accuracy of the KNN classification by 5.23% from 81.91%.

1 Introduction

Cancer is a major cause of death and a significant obstacle to increasing life expectancy in every country in the world. According to estimates by the World Health Organization (WHO) in 2019, cancer is the first or second cause of death before the age of 70 in many countries. One of the deadliest cancers is lung cancer which accounts for 11.6% of all cancer diagnoses in the world (GLOBOCAN, 2020). There are several options for treating patients with lung cancer, namely thoracic surgery, radiotherapy, or chemotherapy. One of the treatments for lung cancer patients that can be done is thoracic surgery (Duma et al., 2019). Thoracic surgery is one of the rapidly growing specialties in all fields of surgery, both technically and technologically for the treatment of chest diseases (Sihoe, 2022).

Early treatment can be done by reducing mortality after thoracic surgery, one of which is collecting data in the form of information about lung cancer patients after thoracic surgery. Many attributes in data cannot produce accurate information, then a feature selection or attribute selection is needed for produce accurate information. Feature selection is the process of selecting a minimal representative feature subset from the original feature set to meet the measurement criteria. One type of algorithm used for feature selection is the bat algorithm. The bat algorithm is a metaheuristic algorithm inspired by the echolocation habit of bats (Chakri et al., 2018).

Need a machine learning method that is used for classification after optimization with the bat algorithm. One of the machine learning methods that is often used is K-Nearest Neighbors

(KNN). KNN is a method that classifies unknown data by measuring the distance or similarity of a known data and then comparing it with a data set (Pawlovsky, 2018). The principle of the KNN method is to find the most similar data samples from the same class and have a high probability. In general, this method begins by finding the k closest neighbors of a query in the training data set and then predicts the query as the main class of the KNN (Zhang et al., 2018).

The disadvantage of KNN is that it must initialize the closest number of k parameter values and must know the attributes or features selected in distance-based learning to get the best results because sufficient computation requires calculations from each data test (Sugiarta et al., 2019).

The combination of the bat and KNN algorithms or the BA-KNN algorithm is a combination of the binary bat and KNN algorithms (Sugiarta et al., 2019). Combining this algorithm, combining the echolocation process of bats with feature selection for the thoracic surgery dataset, then a machine learning method, namely KNN, is used for classification which is used as the value of the fitness function.

2 The Proposed Method

2.1 Bat Algorithm

Bat algorithm or BA is an inspiration from the echolocation habit of bats which is applied to a metaheuristic algorithm (Chakri et al., 2018). Echolocation is the ability of bats to identify objects using ultrasonic sound in their surroundings. Bats use this echolocation ability to avoid objects and find food. BA imitates the echolocation ability to create new and improved metaheuristic algorithms (Sugiarta et al., 2019).

2.2 Binary Bat Algorithm

Binary bat algorithm or BBA is a continuation of the bat algorithm. In the bat algorithm, the algorithm works well for continuous-valued problems, that is, each bat has a continuous-valued position in a certain search space. However, in discrete and combinatorial cases it is recommended to change the algorithm called binary bat (Gupta et al., 2019). In BBA, artificial bats can move around the search space by utilizing position and velocity vectors that are updated in a continuous state (Ma & Wang, 2018).

2.3 K-Nearest Neighbors Algorithm

K-Nearest Neighbor or KNN is one of the classification methods in data mining, where KNN can classify datasets based on training data that are classified or labelled. KNN is included in the supervised learning group, namely the results of the newly classified query based on most of the proximity to the categories in KNN (Dewi & Dwidasmara, 2020).

This algorithm is based on the shortest distance from the test data to the training data to determine it. Then the most data is taken to be used as a prediction from the testing data. Near or far neighbors are calculated using eucledian distance (Cahyanti et al., 2020).

2.4 BA-KNN Algorithm

BA-KNN is a combination of algorithms between BBA and KNN. BBA can improve the accuracy of KNN by finding and selecting better attributes. The result of feature selection from BBA can be used for KNN classification. This classification will result in better accuracy. In addition, due to the few features that are used, the classification required for computation time is faster (Sugiarta et al., 2019).

The basis of BA–KNN is to use the accuracy of KNN used for the fitness function of the BBA algorithm. Therefore, BA-KNN is a BBA algorithm but uses the accuracy of KNN classification with selected features for fitness function retrieval. BA-KNN can be considered as an optimization variation of the optimized BBA for feature selection purposes (Gupta et al., 2019).

3 Method

In this study, the first stage in the research method is to prepare the research object and normalize the data for the research object to be applied to the BA-KNN algorithm. Then do the classification using the BA-KNN algorithm. After being applied to the algorithm, the next step is to compare

which algorithm gives better accuracy results between the KNN model before using feature selection and after. Overall, the research method can be seen in the system flowchart in figure 1.

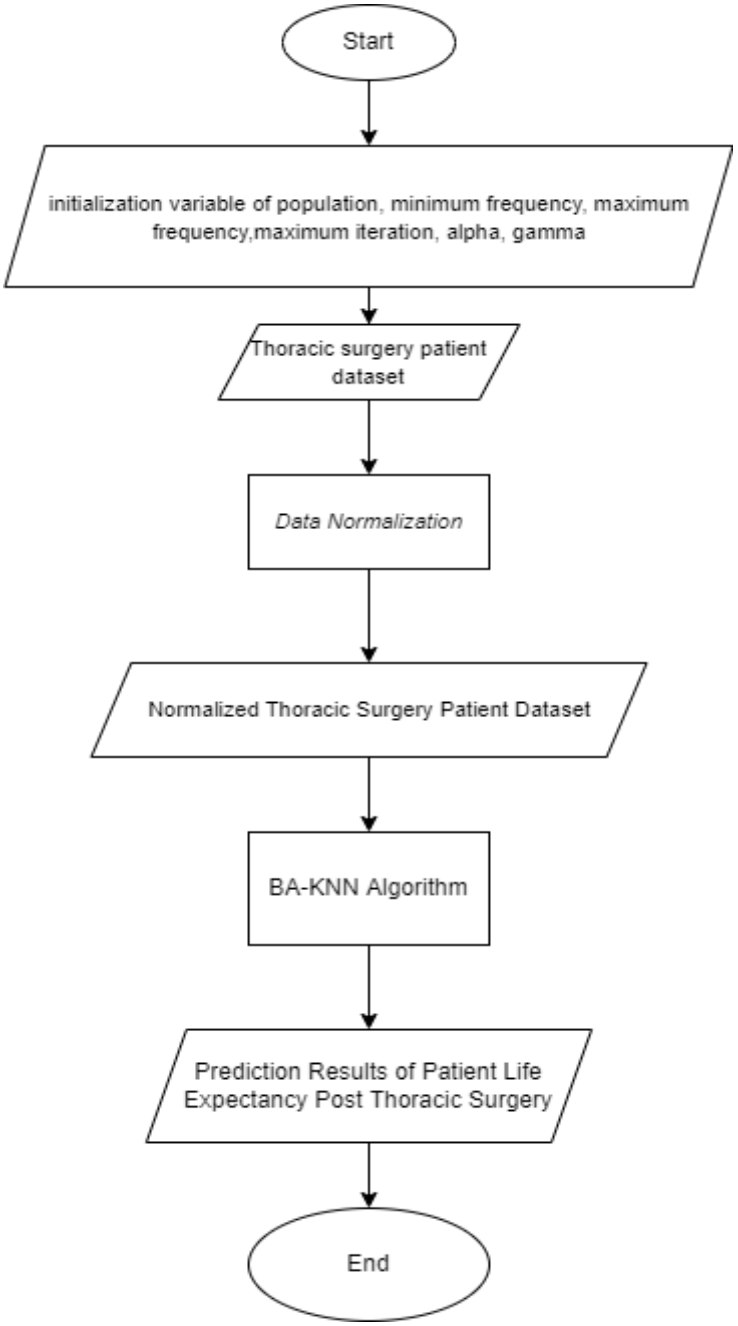

Figure 1. Flowchart of proposed method

4 Results and Discussion

This section is divided into two parts, results and discussion. The results are a description of the data and findings obtained using the methods and procedures described in the data collection method. The discussion is an explanation of the results that answer research questions more comprehensively.

4.1 Results

The application of the BA-KNN algorithm to predict the life expectancy of lung cancer patients after thoracic surgery has five stages of research. The five stages include the data collection stage, the data processing stage, the model evaluation stage, the classification stage, and the system

implementation stage. A more complete explanation regarding the results of the research stages will be described as follows.

*4.1.1 **Data Collection***

In this study, the data used to be processed in the study is secondary data obtained from the Machine Learning Repository University Irvine California (UCI). Wroclaw Thoracic Surgery Center is secondary data used in this study which contains a collection of data, the contents of the dataset are patient data from 2009 to 2014 with lung cancer patients who underwent thoracic surgery.

This dataset has 16 attributes about each patient that represent the condition before and after the patient underwent thoracic surgery. The data types in the dataset are binary, numeric, and nominal. This dataset has two classes, namely, surviving and dying within one year (die) with a total sample of 400 samples for the survival class and 70 for the die class. Table 1 shows the 16 attributes used along with their descriptions and data types.

Table 1. Thoracic surgery dataset

No.	Attribute	Description	Data Type
1.	DGN	Diagnosis of specific combinations of ICD-10 codes for primary and secondary tumors and more than one tumor, if any	Nominal
2.	PRE4	Amount of air that can be forcibly exhaled from the lungs after taking as deep a breath as possible (FVC)	Numeric
3.	PRE5	The amount of air that has been exhaled at the end of the first second of FVC (FEV1)	Numeric
4.	PRE6	A measure of the general ability of cancer patients in daily activities (Zubrod Scale)	Nominal
5.	PRE7	Pain before surgery	Binary
6.	PRE8	Haemoptysis before surgery	Binary
7.	PRE9	Dyspnoea before surgery	Binary
8.	PRE10	Cough before surgery	Binary
9.	PRE11	Weak condition before surgery	Binary
10.	PRE14	Tumor size	Nominal
11.	PRE17	Diabetes	Binary
12.	PRE19	Myocardial Infarction (MI) up to 6 months	Binary

13.	PRE25	Diseases that attack the arteries/blood flow (PAD)	Binary
14.	PRE30	Smoke	Binary
15.	PRE32	Asthma	Binary
16.	AGE	Age at operation	Numeric
17.	RISK	The patient's ability to survive after 1 year	Binary

*4.1.2 **Normalization Data***

Data normalization is done to balance data values by mapping data into certain ranges. The data normalization process is carried out using the Min-max Normalization calculation. The following steps are used for data normalization.

4.1.2.1 Finding the minimum and maximum values for each attribute

In the process of normalizing the data required minimum and maximum values. The minimum and maximum values of the attributes are shown in table 2.

Table 2. Minimum and maximum value of dataset

No	Attribute	Minimum	Maximum
1.	DGN	1	8
2.	PRE4	1.44	6.3
3.	PRE5	0.96	86.3
4.	PRE6	0	2
5.	PRE7	0	1
6.	PRE8	0	1
7.	PRE9	0	1
8.	PRE10	0	1
9.	PRE11	0	1
10.	PRE14	1	4
11.	PRE17	1	1
12.	PRE19	1	1
13.	PRE25	1	1
14.	PRE30	1	1
15.	PRE32	1	1
16.	AGE	21	32
17.	RISK	0	1

4.1.2.2 Calculate the value of each attribute

The next step is the process of calculating the value itself using normalization data equation. Following are the results of normalization based on the 4 selected attributes shown in table 3.

Table 3. Thoracic surgery dataset after normalization

No	Diagnosis	Pain	…	Cough	Risk_1Y
1.	0.142857	0.0	…	1.0	0.0
2.	0.285714	0.0	…	0.0	0.0
3.	0.285714	0.0	…	1.0	0.0
4.	0.285714	0.0	…	0.0	0.0

5.	0.285714	0.0	…	1.0	1.0
6.	0.142857	0.0	…	1.0	0.0
7.	0.285714	0.0	…	0.0	0.0
8.	0.285714	0.0	…	0.0	0.0
9.	0.285714	0.0	…	1.0	0.0
10.	0.285714	0.0	…	1.0	0.0
…	…	…	…	…	…
470	0.285714	0.0	…	0.0	0.0

*4.1.3 **Splitting Data***

The distribution of data in the research dataset aims to divide the data into training data and testing data. This split data stage uses k fold cross validation, with divide the data 0.2 or 20% will be used as testing data and 0.8 or 80% is used as training data with a k-fold value of 5.

*4.1.4 **Data Mining***

At the data mining stage, there are two mining processes. First, the classification process using the KNN Algorithm on the Thoracic Surgery Prediction Dataset. Second, the classification process using the bat algorithm for feature selection, and KNN algorithm for classification. In this study, the BA-KNN algorithm carried out 3 tests, namely testing the number of populations, testing convergence and testing KNN comparisons.

4.1.4.1 Population test

The population size test is a test that aims to determine the best value of a population variable on the BA-KNN for the thoracic surgery dataset. The population variable in this test is the number of bats in the BA-KNN algorithm. Where each population will find the best solution, the more population values, the more bats can find a solution to the problem. In this test, it is done by changing the value of the population parameter for the number of bats in the BA-KNN algorithm. In this test, it is done by changing the parameter value of the population with values of 2, 4, 6, 8, and 10. The results of the accuracy and selected features with changes in the population size parameter shown in table 4.

Table 4. Result of population test

Number of Population	Accuracy	Atrribute Selected
2	87.23 %	*['Diagnosis', 'Pain', 'Cough', 'Weakness', 'diabetes', 'MI_6months']*
4	86.17 %	*['Forced_Capacity', 'Weakness', 'Size_of_tumor', 'MI_6months', 'Age']*
6	87.23 %	*['Diagnosis', 'Forced_Expiration', 'Zubrod_scale', 'Cough', 'Weakness', 'Size_of_tumor', 'Asthmatic']*
8	87.23 %	*['Diagnosis', 'Forced_Capacity', 'Zubrod_scale', 'Pain',*

10	87.23 %	*'Dyspnoea', 'MI_6months', 'PAD']* *['Zubrod_scale', 'Pain', 'Cough', 'Weakness', 'Size_of_tumor', 'diabetes', 'MI_6months']*

Based on table 4, it is shown that the test results with several parameter values of the population and the results of the BA-KNN test by testing the number of populations get the highest accuracy of 87.23% for the total population of 2, 6, 8, and 10 by selecting from sixteen features to seven features. Selected with one of the highest accuracy values. Seven features selected from one of the highest accuracy scores were 'Zubrod_scale', 'Pain', 'Cough', 'Weakness', 'Size_of_tumor', 'diabetes', 'MI_6months'.

4.1.4.2 Convergent Test

Convergent testing is a test to determine the convergence of BA on BA-KNN, convergent conditions are where BA has found a solution and the value of the solution does not change from several iterations (Sugiarta et al., 2019). Changes in parameter values from the maximum iterations carried out in this test. The maximum parameter values for iterations of the BA–KNN algorithm tested are 2, 4, 6, 8, and 10. The results of the accuracy and selected features with changes in the maximum iteration parameter shown in table 5.

Table 5. Result of convergent test

Maximum Iteration	Accuracy	Attribute Selected
2	86.17 %	*['Diagnosis', 'Forced_Expiration', ' Haemoptysis', 'Dyspnoea', 'Weakness', 'MI_6months', 'Age']*
4	87.23 %	*['Diagnosis', 'Cough', 'Weakness', 'MI_6months', 'Smoker']*
6	87.23 %	*['Diagnosis', 'Forced_Expiration', 'Zubrod_scale', 'Dyspnoea', 'Cough', 'Weakness', 'MI_6months', 'Asthmatic']*
8	87.23 %	*['Forced_Capacity', 'Size_of_tumor', 'diabetes', 'Asthmatic', 'Age']*
10	87.23 %	*['Forced_Capacity', 'Forced_Expiration', 'Pain', 'Cough', 'Size_of_tumor', 'diabetes', 'PAD']*

Based on table 5, the results of the BA-KNN convergent test are shown with the first iteration 2 achieving 86.17% accuracy, the second iteration 4 achieving 87.23% accuracy, the third iteration 6 achieving 87.23% accuracy, the fourth iteration 8 achieving 87.23% accuracy, and the last iteration of 10 achieved an accuracy of 87.23%. In the convergent test, the highest accuracy achieved during the 4th, 6th, 8th, and 10th iterations with an accuracy of 87, 23% by selecting from sixteen features to seven selected features. Seven features selected from one of the highest accuracy scores were 'Forced_Capacity', 'Forced_Expiration', 'Pain', 'Cough', 'Size_of_tumor', 'diabetes', 'PAD'.

4.1.4.3 KNN Comparative Testing

The KNN comparison test is a test by changing the value of the k parameter on KNN and BA - KNN which aims to determine how influential BA is on BA - KNN by comparing the execution time and accuracy values of BA-KNN and KNN without the feature selection of the bat algorithm. This test makes changes to the value of the parameter k with values of 3, 5, 7, 9, and 11. The results of the accuracy and selected features with changes in the population size parameter shown in table 6.

Table 6. Result of KNN comparative test

Number of *k*	BA-KNN Accuracy	KNN Accuracy	KNN Execution Time (in seconds)	BA-KNN Execution Time (in seconds)
k = 3	87.23 %	81.91 %	0.37330	0.01209
k = 5	87.23 %	81.91 %	0.37512	0.03166
k = 7	86.17 %	81.91 %	0.36585	0.03328
k = 9	86.17 %	81.91 %	0.40116	0.04464
k = 11	86.17 %	81.91 %	0.41100	0.04595

Based on table 6, the results of the KNN comparison test with several combinations of parameters, the first k value with a k value of 3 has a BA-KNN accuracy of 87.23%, KNN accuracy of 81.91%, KNN execution time owned 0.37330 seconds and time BA-KNN's execution is 0.01209 seconds. The second test using the parameter value of k, namely 5, has an accuracy of 87.23% for BA-KNN, an accuracy of 81.91% for KNN, the execution time of KNN is 0.37512 seconds and the execution time of BA-KNN is 0.03166 seconds. The third test carried out with a k value of 7 obtained BA-KNN accuracy of 86.17%, KNN accuracy of 81.91%, KNN execution time of 0.36585 seconds and BA-KNN execution time of 0.03328 seconds. The fourth test uses a k value of 9, the BA-KNN accuracy is 86.17%, the KNN accuracy is 81.91%, the KNN execution time is 0.40116 seconds and the BA-KNN execution time is 0.04464 seconds. The last test with a k value of 11 obtained BA-KNN accuracy of 86.17%, KNN accuracy of 81.91%, execution time of 0.4100 seconds and BA-KNN's execution time of 0.04595 seconds.

4.2 Discussion

This study applies the bat algorithm as a feature selection to improve accuracy in predicting patient life expectancy after thoracic surgery. The classification algorithm used is the KNN algorithm. The dataset used in this study is a dataset obtained from the UCI Machine Learning Repository, namely the thoracic surgery dataset. In this study, a comparison was made between the KNN algorithm, the BA-KNN algorithm and the results of previous studies. The higher the accuracy of the model, the better the model used.

Table 7. Comparison of accuracy with previous research

Research	Methods	Results
(Setyadi et al., 2020)	Genetic Algorithm and Nave Bayes	85,31 %
(Prasetio & Susanti, 2019)	Boosted K-Nearest Neighbor Algorithm	85,11 %

(Roshan & Rohini, 2017)	Nave Bayes, Decision Stump, J48, random forest, J48 and random forest, J48 and Naïve bayes	The highest accuracy uses the J48 + Naïve Bayes algorithm with an accuracy of 88.73%
Proposed method	BA-KNN Algorithm	87, 23 %

The comparison results show that the accuracy obtained from the BA-KNN algorithm model is the second highest accuracy after the J48 + Naïve Bayes algorithm with an increase in accuracy of 1.92% from the research of Setyadi et al (2020), 2.12% from the research of Prasetio & Susanti (2019) and the comparison of the results of the classification system made for KNN is 5.23%. The J48 and Naïve Bayes algorithms have the highest accuracy assisted by attribute selection with the WEKA application using a ranker algorithm to sort attributes from 1 to 16 attributes while the BA-KNN algorithm uses the bat algorithm as a feature selection algorithm, but the drawback is that the feature selection is chosen randomly.

The advantage of this study is that by applying the BA-KNN algorithm with the bat algorithm as a feature selection algorithm, it can increase accuracy in predicting life expectancy of patients after thoracic surgery so that it can be used by further research as a reference in conducting research. However, this research still has drawbacks, namely the application of the bat algorithm for feature selection does not necessarily get the optimal solution, because the results obtained can vary even though using the same parameters. However, the application of the bat algorithm has succeeded in being the best solution for optimizing the KNN model in terms of feature selection and execution time used in this study.

5 Conclusion

Based on the results of research and discussion related to the optimization of KNN with the bat algorithm as feature selection to increase accuracy in predicting life expectancy of patients after thoracic surgery using the thoracic surgery dataset obtained from the UCI Machine Learning Repository. Three tests of the BA-KNN algorithm used in predicting the life expectancy of patients after thoracic surgery using a thoracic surgery dataset that has been carried out, namely population testing, convergent testing and KNN comparison testing, the best accuracy results are 87.23% which has an execution time of 0.01209. seconds with an increase in accuracy of 5.32% from the accuracy of KNN without optimization of the bat algorithm, which is 81.91% which has an execution time of 0.37330 seconds.

6. References

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/https://doi.org/10.33096/ijodas.v1i2.13

Chakri, A., Ragueb, H., & Yang, X. S. (2018). Bat algorithm and directional bat algorithm with case studies. In *Studies in Computational Intelligence* (Vol. 744, pp. 189–216). Springer Verlag. https://doi.org/10.1007/978-3-319-67669-2_9

Dewi, A. M. S. I., & Dwidasmara, I. B. G. (2020). Implementation Of The K-Nearest Neighbor (KNN) Algorithm For Classification Of Obesity Levels. *Jurnal Elektronik Ilmu Komputer Udayana*, *9*(2). https://doi.org/10.24843/JLK.2020.v09.i02.p15

Duma, N., Santana-Davila, R., & Molina, J. R. (2019). non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment. *Mayo Clinic Proceedings*, *94*(8), 1623–1640. https://doi.org/10.1016/j.mayocp.2019.01.013

GLOBOCAN. (2020). *The global cancer observatory : All cancer [internet]*. https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf

Gupta, D., Arora, J., Agrawal, U., Khanna, A., & de Albuquerque, V. H. C. (2019). Optimized binary bat algorithm for classification of white blood cells. *Measurement: Journal of the International Measurement Confederation*, *143*, 180–190. https://doi.org/10.1016/j.measurement.2019.01.002

Ma, X. X., & Wang, J. S. (2018). Optimized parameter settings of binary bat algorithm for solving function optimization problems. *Journal of Electrical and Computer Engineering*, *2018*. https://doi.org/10.1155/2018/3847951

Pawlovsky, A. P. (2018). An ensemble based on distances for a knn method for heart disease diagnosis. *IEEE*. https://doi.org/10.23919/ELINFOCOM.2018.8330570

Prasetio, R. T., & Susanti, S. (2019). Prediksi Harapan hidup pasien kanker paru pasca operasi bedah toraks menggunakan boosted k-nearest neighbor. *Jurnal ResponsIF*, *1*(1), 64–69.

Roshan, & Rohini. (2017). Prediction of post-surgical survival of lung cancer patients after thoracic surgery using data mining techniques. *International Journal of Advanced Research*, *5*(4), 596–600. https://doi.org/10.21474/IJAR01/3852

Setyadi, Y. A., Asror, I., & Wibowo, Y. F. A. (2020). Prediksi harapan hidup pasca operasi toraks pada pasien penderita kanker paru-paru menggunakan metode genetic algorithm untuk feature selection dan naïve bayes classifier. *E-Proceeding of Engineering*, *7*(20), 8349–8360.

Sihoe, A. D. L. (2022). Thoracic surgery worldwide. *Journal of Thoracic Disease*, *14*(1), 216–217. https://doi.org/10.21037/jtd-2022-01

Sugiarta, K. A., Cholissodin, I., & Santoso, E. (2019). Optimasi k-nearest neighbor menggunakan bat algorithm untuk klasifikasi penyakit ginjal kronis. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(10), 10301–10308. http://j-ptiik.ub.ac.id

Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, *109*, 44–54. https://doi.org/10.1016/j.patrec.2017.09.036