JURNAL INFOTEL
Informatics - Telecommunication - Electronics

Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

Breast cancer recurrence prediction system using k-nearest neighbor, naïve-bayes, and support vector machine algorithm

I Ketut Agung Enriko[1]*, Melinda Melinda[2], Agnesia Candra Sulyani[3], I Gusti Bagus Astawa[3]
[1] Institut Teknologi Telkom Purwokerto
[2] Universitas Syiah Kuala
[3] PT Telkom Indonesia
[1]Jl. D.I Panjaitan No.128, Purwokerto 53147, Indonesia
[2]Jl. Teuku Nyak Arief No. 441, Banda Aceh 23111, Indonesia
[3]Jl. Gatot Subroto Kav 52, Jakarta Selatan 12710, Indonesia
*Corresponding email: agungenr@gmail.com

Received 24 July 2021, Revised 28 September 2021, Accepted 22 October 2021

Abstract —Breast cancer is a serious disease and one of the most fatal diseases in the world. Statistics show that breast cancer is the second common cancer worldwide with around two million new cases per year. Some research has been done related to breast cancer, and with the advancements of technology, breast cancer can be detected earlier by using artificial intelligence or machine learning. There are popular machine learning algorithms that can be used to predict the existence or recurrence of breast disease, for example, k-nearest neighbor (kNN), naïve bayes, and support vector machine (SVM). This study aims to check the prediction of breast cancer recurrence using those three algorithms using the dataset available at the University of California, Irvine (UCI). The result shows that the kNN algorithm gives the best result in terms of accuracy to predict breast cancer recurrence.

Keywords – k-Nearest Neighbor, machine learning, breast cancer prediction.

I. INTRODUCTION

Breast cancer is known as a fatal disease, especially for women. World Health Organization (WHO) reports that globally, there are about two million breast cancer new cases and causing more than 600,000 deaths in 2018 [1]. The treatment of breast cancer takes years and effort, from the diagnosis, surgery (if needed), and therapy (radiation, chemotherapy, and medicines) [2].

Meanwhile, in developing countries like Indonesia, the number of doctors and general practitioners is still low. Statistics show that the ratio of doctor and population is 0.4 doctor per 1000 population [3], the second-worst in South East Asia. This condition urges healthcare researchers to innovate in terms of improving people's healthcare, with the use of technology.

There are a lot of studies related to healthcare technology, for example [4,5] discussed the implementation of heart disease prediction system. They can detect the type of heart disease of a patient using the k-nearest neighbor (kNN) machine learning algorithm. Similar to this, another research offered a heart disease prediction system using a Hybrid Random Forest with a Linear Model (HRFLM) [6]. Meanwhile, [7] utilized data mining algorithms like kNN and Bayesian to predict diabetes disease of some patients. For breast cancer disease, research [8] studied about prediction system of benign or malignant breast cancer using data mining techniques like naïve bayes, RBF Network, and J48 Decision Tree, while [9] still studied the same topic but using five different machine learning algorithms: C4.5, support vector machine (SVM), naïve bayes, and kNN.

This study discusses how breast cancer can be detected in terms of its recurrence using three popular machine learning algorithms: kNN, naïve bayes, and SVM. These algorithms are selected since they are robust and recognized as top algorithms frequently used in machine learning research [10,11,12,13]. The dataset

used for this study is from the University of California, Irvine (UCI) which available on the Internet [14]. The dataset consists of 286 records and will be analyzed using those three algorithms to be checked from the accuracy and speed aspects.

This paper is organized as follows: Section 1 for the introduction, Section 2 is about research method, Section 3 is the result, Section 4 is for discussion, and Section 5 provides the conclusion.

II. RESEARCH METHODS

A. Dataset

The dataset used in this research is taken from UCI Irvine (286 records) with attributes listed in Table 1.

Table 1. UCI irvine breast cancer dataset attributes

Num	Attribute	Data Type	Complete/ Incomplete
1	Age	Range (10)	Complete
2	Menopause	Options (ge40, let40, premeno)	Complete
3	Tumor size	Range (5)	Complete
4	Inv-Nodes	Range (3)	Complete
5	Code-Caps	Binary (Yes/No)	Complete
6	Deg-Malig	Options (1,2,3)	Complete
7	Breast	Binary (Left/Right)	Complete
8	Breast-Quad	Options	Complete
9	Irradiat	Binary (Yes/No)	Complete
10	Class	Binary (Recurrence/No)	Complete

B. Data Preprocessing

For analysis purposes, the data contents should be reformatted. For example, the data type "range" should be converted to a median number. For the "Age" attribute which originally is grouped by 10 years range (for example 0-9, 10-19, 20-29, and so on), the median value is taken so it can be calculated by Weka (converted to 5, 15, 25, and so on). Table 2 shows the attribute data after being converted.

Table 2. Dataset after reformatted

Num	Attribute	Data Type	Complete/ Incomplete
1	Age	Median of range	Complete
2	Menopause	Category (1,2,3)	Complete
3	Tumor Size	Median of range	Complete
4	Inv-Nodes	Median of range	Complete
5	Code-Caps	Binary (0/1)	Complete
6	Deg-Malig	Category (1,2,3)	Complete
7	Breast	Binary (0/1)	Complete
8	Breast-Quad	Category (1,2,3,4,5)	Complete
9	Irradiat	Binary (0/1)	Complete
10	Class	Binary (0/1)	Complete

C. Data Processing

The data are now ready to be processed with the Weka tool. Weka is a popular and free data mining tool created by The University of Waikato [15]. The Weka settings for dataset evaluation are shown in Table 3.

Table 3. Weka settings used for data analysis

Method	Options
kNN	• 10-fold cross validation. • k value from 1 to 9 (trial & error). • distanceFunction: Euclidean • distanceWeighting: No. • nearestNeighbourSearchAlgorithm: LinearNNSearch
naive bayes	10-fold cross validation
SVM	• 10-fold cross validation • Calibrator: Logistic • filterType: normalize training data • kernel: PolyKernel

The overall flowchart of the research method is depicted in Fig.1.

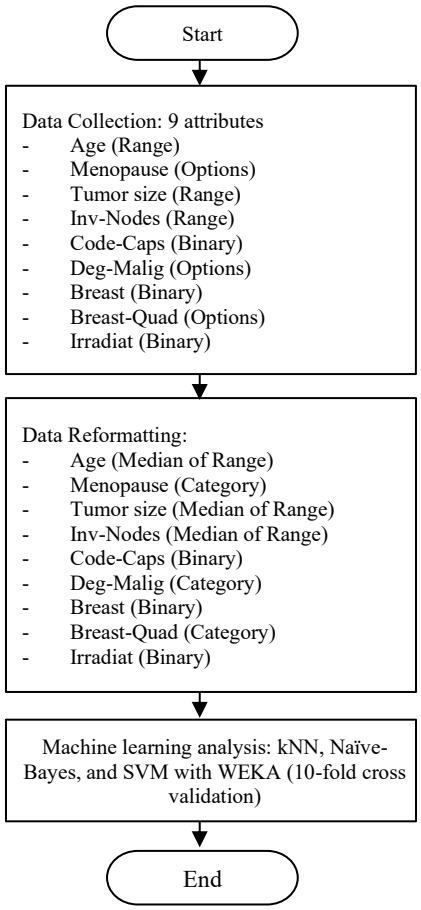

Fig.1. Flowchart of the research method

III. RESULTS

A. kNN Analysis

The first simulation with the dataset is by using the kNN algorithm. Since the k value for kNN cannot be determined, trial and error simulations should be done to choose the optimum value for k to yield the best accuracy. In the Weka tool, we can change the k value by changing the configuration script as shown in Fig. 1 below. The k parameter is written in bold font (-K 1), where k=1. The k value can be changed with other odd numbers (1,3,5,7,9, and so on) to determine the correct class via voting.

By filling the k number from 1 to 9 (odd values only), the result is shown in Table 4. From the table, we can see that the best accuracy is 77.98% when the *k* value is 7. Besides accuracy, the other important parameter is speed. Weka tool informs that the time taken to build the model is 0 second (or less than 1 second).

As a result, kNN proves to have good accuracy in handling enough data and parameters and it performs very well with high speed, which is also shown in previous similar research [16].

```
weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearc
h -A \"weka.core.EuclideanDistance -R
first-last\""
```

Fig. 1. Weka configuration for changing *k* value (-K 1) where *k*=1

Table 4. kNN analysis results

k value	Accuracy
1	65.70%
3	70.04%
5	74.01%
7	**77.98%**
9	75.45%
11	75.09%

```
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -
W 1 -K
"weka.classifiers.functions.supportVector.P
olyKernel -E 1.0 -C 250007" -calibrator
"weka.classifiers.functions.Logistic -R
1.0E-8 -M -1 -num-decimal-places 4"
```

Fig. 2. Weka configuration for SVM analysis

B. Naïve Bayes Analysis

Using naïve bayes with Weka tool to analyze the data is quite straightforward. We just need to set the classifier setting to "naïve bayes" option. The result of this simulation is 73.65% of accuracy and the time taken to build the model is 0 second (or less than 1 second).

Naïve bayes proves a relatively good performance in analyzing simple data and binary class (recurrence or not) in this study, and it also performs at the fast speed.

C. SVM Analysis

In Weka tool, the SVM analysis is done by setting the configuration like depicted in Fig. 2.

The result from using the SVM method shows that the prediction accuracy is 71.12% and the time taken to build the model is 0.03 second, yet it is still less than 1 second.

In this study, SVM shows a fair performance in terms of accuracy, while the speed is quite impressive.

IV. DISCUSSION

The Weka simulations result that the accuracy of kNN, naïve bayes, and SVM algorithm are 77.98%, 73.65%, and 71.12% respectively. From the speed aspect, all three algorithms perform well with less than 1 second is needed to build the model. The overall result is shown in Table 5.

The result from these three algorithms is quite fair but still can be improved if the dataset has more than 286 records from UCI Irvine used in this research. Especially if we want to use advanced algorithms like deep learning, which need thousands of records to significantly improve the accuracy.

Table 5. kNN, naïve bayes, and SVM analysis result

Algorithm	Accuracy	Speed
kNN	77.98%	<1 second
naïve bayes	73.65%	<1 second
SVM	71.12%	<1 second

V. CONCLUSION

Nowadays, research about disease prediction is getting more attention, with the help of machine learning algorithms. kNN, naïve bayes, and SVM are some of frequently used algorithms. In this study, those three algorithms are used to predict the recurrence of breast cancer disease. A dataset from UCI Irvine which has 286 records is used, consists of 9 attributes and 1 binary class (recurrence or not). The result shows that kNN has the best performance with 77.98% accuracy and less than 1 second speed, followed by naïve bayes (73.65% accuracy and less than 1 second speed) and SVM (71.12% accuracy and less than 1 second speed).

For future works, the deep learning method can be used to do the analysis but possibly the amount of data needed should be added to achieve good accuracy.

ACKNOWLEDGMENT

This work is supported by a grant from PT Telkom Indonesia and The Indonesia Telecommunication and Digital Research Institute (ITDRI) task force.

REFERENCES

[1] The Global Cancer Observatory, "Cancer fact sheet," *World Heal. Organ.*, 2019.

[2] A. G. Waks and E. P. Winer, "Breast Cancer Treatment: A Review," *JAMA - Journal of the American Medical Association*. 2019.

[3] W. Bank, "Physicians (Per 1,000 People)," *World Bank Report*, 2020. [Online]. Available: https://data.worldbank.org/indicator/SH.MED.PHYS.ZS?most_recent_value_desc=true. [Accessed: 15-Feb-2021].

[4] Enriko, I. K. A., Suryanegara, M., & Gunawan, D, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 12, pp. 59–65, 2016.

[5] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease diagnosis system with k-nearest neighbors method using real clinical medical records," in *ACM International Conference Proceeding Series*, 2018.

[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, 2019.

[7] D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," in *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017*, 2018.

[8] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithms Comput. Technol.*, 2018.

[9] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," in *Procedia Computer Science*, 2016.

[10] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, 2008.

[11] L. Peterson, "K-nearest neighbor," *Scholarpedia*, 2009.

[12] D. Berrar, "Bayes' theorem and naive bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 2018.

[13] R. Gandhi, "Support Vector Machine - Introduction to Machine Learning Algorithms," *Towards Data Science Tutorial*, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed: 24-Feb-2021].

[14] U. of C. Irvine, "Breast Cancer Data Set," *UCI Dataset Repository*, 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer. [Accessed: 04-Jan-2021].

[15] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, 2004.

[16] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Comparative Study of Heart Disease Diagnosis Using Top Ten Data Mining Classification Algorithms," *J. Telecommun. Electron. Comput. Eng.*