Comparison of Colorectal Cancer Classification between K-Nearest Neighbors (K-NN) and Neural Network

F Zhafarina\textsuperscript{1,*}, Z Rustam\textsuperscript{1}, Y Amalia\textsuperscript{1}, I Wirasati\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, University of Indonesia, Depok 16424, Indonesia

E-mail*: fildzah.zhafarina@sci.ui.ac.id

Abstract. Machine learning is one of the technologies used in medicine. Machine learning can help detect various kinds of problems in the medical field and enables a process to be faster and more efficient. Cancer is one of the most dangerous diseases in the world. Machine learning is widely used in bioinformatics and particularly in cancer diagnosis. One of the most popular methods is K-nearest neighbors (K-NN) and Neural Network. There are supervised learning methods. Using K-NN, the quality of the results depends largely on the distance and the value of the parameter “k” which represents the number of the nearest neighbors. This research is explains the classification of colorectal cancer by using K-NN with different k values and Neural Network Classification. Our work will be performed on the Colorectal Cancer dataset obtained by the Al-Islam Hospital, Bandung, Indonesia and it consists of benign cases 163 and malignant cases 47 samples. Thus, the final result indicates better performance for K-nearest neighbors’ accuracy is 0.786 in K-parameter equal to 7, 9, 11 has the same accuracy with 60% data training and Neural Network reached 0.904 with 90% of data training.

1. Introduction

Cancer is a disease that causes death in second place in the world [1]. Detecting these diseases when still at an early stage is associated with markedly improved survival prospects [2]. Cancer is a disease that can be detected by machine learning[3]. It can be seen from the development of data mining in medical science is increasing rapidly[4]. This increase can be seen from the high prediction results, which can reduce treatment costs, increase the chances of recovery of patients, and decisions to save lives [4]. Colorectal cancer is cancer with the third death rate. responsible for around 600,000 per year worldwide [5].

Various methods can be used to cure this disease, including systemic chemotherapy, radiotherapy, and resection surgery [6]. However, there is no efficient treatment for colorectal cancer patients, especially in patients with liver metastases (CRLM) [7]. Liver metastases are the main cause of death in this disease [8]. The treatments commonly used in CRLM are chemotherapy and liver resection. However, its toxicity and side effects limit the use of this chemotherapy [9].

There are some methods of previous researches for processing colorectal cancer datasets, such as Neural Network[10], Fuzzy C-Means [11], and Random Forest [12]. This research uses another method for classifying the data, which is K-Nearest Neighbor and Neural Network. Therefore, this method will help the medical staff to classify a medical problem easily.
2. Methods
This research using Jupyter Notebook as software for running the program of K-NN with different k-values and Neural Network Classification. The programming language used for this paper is Python 3.

2.1. Data
The Colorectal Cancer dataset was taken from Al-Islam Hospital, Bandung, Indonesia. This dataset consists of begin cases 163 and malignant cases 47. The colorectal cancer dataset consists of 6 variables, as shown in Table 1.

| Attributes     | Description                                           |
|----------------|-------------------------------------------------------|
| Age            | The number of age patients who are in check           |
| CA 19-9        | The number of cancer antigen units per milliliter of blood |
| Hemoglobin     | The number of hemoglobin gram per deciliter of blood  |
| Leukocyte      | The number of leukocyte cell per uL of blood          |
| Hematocrit     | Hematocrit or the volume percentage of red blood cells|
| Platelets      | The number of thrombosis cell per uL of blood         |

Table 1. Colorectal Cancer Dataset Variable

This research uses logistic regression and random forest for classification. This method are evaluated using 43-random state. Table sample of dataset shown in Table 2.

| Age | CEA  | Hemoglobin | Leukocyte | Hematocrit | Platelets | Diagnosis |
|-----|------|------------|-----------|------------|-----------|-----------|
| 54  | 0.94 | 12.1       | 6500      | 39.2       | 286000    | 0         |
| 48  | 0.82 | 12.3       | 11100     | 39.2       | 503000    | 0         |
| 83  | 1.14 | 12.7       | 6100      | 38.5       | 462000    | 0         |
| 56  | 12.38| 7.8        | 21000     | 23.4       | 223000    | 1         |
| 68  | 1.18 | 13.1       | 5800      | 38.7       | 195000    | 0         |
| 63  | 1.3  | 12.8       | 12300     | 40.7       | 329000    | 0         |
| 76  | 1.28 | 8.3        | 16400     | 20.8       | 165000    | 0         |
| 58  | 40.38| 14.7       | 17600     | 44.4       | 429000    | 1         |

Table 2. Sample of Colorectal Cancer Dataset from Al-Islam Hospital, Bandung, Indonesia

2.2. K-Nearest Neighbors (K-NN)
The closest K-Nearest Neighbor (K-NN) is a well-known non-parametric classifier that is widely used in pattern recognition [13]. K-Nearest Neighbor is a method for classifying a test object data by calculating the value of the adjacency value based on the distance function and the pattern of similarity to certain objects [14]. K-Nearest Neighbor is used to determining class membership with the highest environmental value. In this method, there are several formulas for calculating distances, one of which is:

Euclidean Distance
The Euclidean Distance is an equation to calculate the value of the distance between \(X_i\) and \(Y_i\) [15]. This value can be calculated using Equation (5):

\[
d_{m,n} = \sqrt{\sum_{i=1}^{n}(X_{mi} - Y_{ni})^2} \quad (1)
\]
2.3. Neural Network

Artificial Neural Networks are machine learning which is used widely in many application areas. This is due to their higher classification performance \[16\]. This algorithm uses a universal function approach that can model linear and non-linear data \[17\]. A type of artificial neural network, a simple three-layer feedforward network with well-established generalization properties is called a multilayer perceptron \[18\]. The classification process uses an Artificial Neural Network, which is entering data into the classifier through the input layer to produce output in the hidden layer, such as:

\[ H_j = \varphi_H(\sum_{i=1}^{d} w_{j,i}x'_i) \]  

\( d \) = the number of inputs neurons  
\( w_{j,i} \) = input and hidden layer’s weights  
\( H_j \) = the hidden layer’s activation function  
\( \varphi_H \) = activation function at the hidden layer

The classification at the output layer as follows:

\[ y_k = \varphi_O(\sum_{j=1}^{M} w_{k,j}H_j) \]  

\( y_k \) = the output layer’s activation function  
\( M \) = neurons at the hidden layer  
\( w_{k,j} \) = the hidden and output layer’s weights  
\( O \) = the output layer’s activation function

The hidden layer and the output layer use the tan-sigmoid and softmax activation functions. The tan-sigmoid function:

\[ \varphi_H = \frac{2}{1+e^{-2}} - 1 \]  

\( \varphi_H \) = tan-sigmoid’s activation function

The softmax function:

\[ \varphi_O = \frac{e}{\sum_{i=1}^{e}} \]  

\( \varphi_O \) = softmax’s activation function

Backpropagation learning algorithms are used to train neural networks. This is done by updating the weights backward from the output layer through the hidden layer to the input layer by spreading the difference between the desired output and the predicted output \[19\]. Can be seen in the equation:

\[ E = \frac{1}{2} \sum_{n=1}^{N}(y_n - z_n)^2 \]  

\( E \) = error of output
\( y_n \) = the predicted value
\( z_n \) = the actual value
\( E \) = the mean squared error.
\( N \) = the number of value output

Errors are minimized by adjusting the weight vector for each iteration, taking into account several hidden neurons and different training algorithms that are optimized simultaneously with feature weights to maximize classification accuracy on colorectal cancer data.

2.4. Confusion Matrix

One of the methods used to calculate accuracy in the concept of data mining or decision support systems is the Confusion Matrix [20]. Accuracy is the ratio of the true predictions in the whole data [21].

\( TP \) (True Positive): Number of samples having colorectal cancer diagnosed correctly.
\( FP \) (False Positive): Sum of healthy people that were incorrectly identified to have colorectal cancer.
\( TN \) (True Negative): Number of healthy individuals correctly spotted.
\( FN \) (False Negative): The number of samples with colorectal cancer that were incorrectly classified as healthy.

| Actual Value | Recognize Value | Positive | Negative |
|--------------|-----------------|----------|----------|
| Positive     | TP              | FN       |
| Negative     | FP              | TN       |

From Table 3 it can build the formula for accuracy that can be seen in the equation below:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7)
\]

3. Experiments Results

This research using Jupyter Notebook as software for running the program of K-NN and Neural Network in processing colorectal cancer classification problem. In this test, the number of data training is equal to 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% which will be used on the results of the dataset. The results of accuracy is given by K-NN and Neural Network Classifier method are shown in table 4 and 5:

| Train size | Accuracy of K-NN |
|------------|------------------|
|            | k = 1 | k = 5 | k = 7 | k = 9 | k = 11 |
| 10%        | 0.619 | 0.704 | 0.725 | 0.783 | 0.783 |
| 20%        | 0.661 | 0.690 | 0.714 | 0.714 | 0.780 |
| 30%        | 0.653 | 0.707 | 0.728 | 0.741 | 0.776 |
| 40%        | 0.683 | 0.738 | 0.770 | 0.770 | 0.770 |
| 50%        | 0.695 | 0.743 | 0.752 | 0.771 | 0.771 |
| 60%        | 0.679 | 0.738 | **0.786** | **0.786** | **0.786** |
| 70%        | 0.683 | 0.762 | 0.778 | 0.778 | 0.778 |
| 80%        | 0.619 | 0.738 | 0.738 | 0.738 | 0.738 |
| 90%        | 0.476 | 0.571 | 0.571 | 0.571 | 0.571 |
Based on Table 4, it is shown that the number of data training is affecting by the values of accuracy. In this research, the highest accuracy value was recorded when the data training is 60% with 0.786 while the lowest accuracy value was recorded when the data training is 90% with $k = 1$ the accuracy is 0.476%.

### Table 5. The results of colorectal cancer classification using neural network

| Train size | Accuracy of Neural Network |
|------------|---------------------------|
| 10%        | 0.746                     |
| 20%        | 0.797                     |
| 30%        | 0.802                     |
| 40%        | 0.785                     |
| 50%        | 0.704                     |
| 60%        | 0.809                     |
| 70%        | 0.841                     |
| 80%        | 0.857                     |
| 90%        | **0.904**                 |

Based on Table 5, it is shown that the number of data training is affected by the values of accuracy. In this research, the highest accuracy value was recorded when the data training is 90% with 0.904 while the lowest accuracy value was recorded when the data training is 50% with 0.704.

### 4. Conclusion

Colorectal cancer detection quickly is very important. It is useful for handling cancer quickly before being infected to all organs of the body. However, this is difficult because colorectal cancer has no specific symptoms. The K-NN and Neural Network methods can help detect colorectal cancer based on blood tests and age. Thus, the final result indicates better performance for K-nearest neighbors accuracy is 0.786 in K-parameter equal to 7, 9, 11 has the same accuracy with 60% data training and Neural Network reached 0.904 with 90% of data training.

### Acknowledgement

This research is fully supported financially by The Indonesian Ministry of Research, Technology, with a KEMENRISTEK/BRIM PDUPT 2021 research grant scheme.

### References

[1] Stewart B W and Kleihues P Eds 2003 *World cancer report*

[2] Chari S T 2007 August *Detecting Early Pancreatic Cancer: Problems and Prospects* (Seminars in oncology) *Vol 34* No 4 pp 284-294 WB Saunders

[3] Hornbrook M C, Goshen R, Choman E, O’Keeffe-Rosetti M, Kinar Y, Liles E G and Rust K C 2017 *Early Colorectal Cancer Detected By Machine Learning Model Using Gender, Age, And Complete Blood Count Data* Digestive diseases and sciences *62*(10) pp 2719-27

[4] Wu X, Zhu X, Wu G Q and Ding W 2013 *Data Mining with Big Data* IEEE transactions on knowledge and data engineering *26*(1) pp 97-107

[5] Dora L, Agrawal S, Panda R and Abraham A 2017 *Optimal Breast Cancer Classification Using Gauss–Newton Representation Based Algorithm* Expert Systems with Applications *85* pp 134-145

[6] Kim E S, Putnam J B, Komaki R, Walsh G L, Ro J Y, Shin H J, ... and Khuri F R 2004 *Phase II Study Of A Multidisciplinary Approach With Induction Chemotherapy, Followed By Surgical Resection, Radiation Therapy, And Consolidation Chemotherapy For Unresectable Malignant Thymomas* Lung cancer *44*(3) pp 369-379

[7] Adam R and Kitano Y 2019 *Multidisciplinary Approach Of Liver Metastases From Colorectal Cancer* Annals of gastroenterological surgery *3*(1) pp 50-56
[8] Helling T S and Martin M 2014 *Cause Of Death From Liver Metastases In Colorectal Cancer* Annals of surgical oncology 21(2) pp 501-506

[9] Shangkuan W C, Lin H C, Chang Y T, Jian C E, Fan H C, Chen K H, ... and Chu C M 2017 *Risk Analysis Of Colorectal Cancer Incidence By Gene Expression Analysis* PeerJ 5 e3003

[10] Bottaci L, Drew P J, Hartley J E, Hadfield M B, Farouk R, Lee P W, ... and Monson J R 1997 *Artificial Neural Networks Applied To Outcome Prediction For Colorectal Cancer Patients In Separate Institutions* The Lancet 350(9076) pp 469-472

[11] Wang X Y, Garibaldi J and Ozen T 2003 *Application Of The Fuzzy C-Means Clustering Method On The Analysis Of Non Pre-Processed FTIR Data For Cancer Diagnosis* Int. Conf. on Australian and New Zealand Intelligent Information Systems (ANZIIS) pp 233-238

[12] Hajizadeh F, Masjedi A, Asl S H, Kiani F K, Peydaveisi M, Ghalamfarsa G, ... and Sevbitov A 2020 *Adenosine And Adenosine Receptors In Colorectal Cancer* International Immunopharmacology 87 106853

[13] Vajda S, and Santosh K C 2016 *A Fast K-Nearest Neighbor Classifier Using Unsupervised Clustering* Int. Conf on recent trends in image processing and pattern recognition (Springer, Singapore) pp 185-193

[14] Yu Q, Gong P, Clinton N, Biging G, Kelly M, and Schirokauer D 2006 *Object-Based Detailed Vegetation Classification With Airborne High Spatial Resolution Remote Sensing Imagery* Photogrammetric Engineering & Remote Sensing 72(7) pp 799-811

[15] Hu L Y, Huang M W, Ke S W and Tsai C F 2016 *The Distance Function Effect On K-Nearest Neighbor Classification For Medical Datasets* SpringerPlus 5(1) pp 1-9

[16] Dreiseitl S and Ohno-Machado L 2002 *Logistic Regression And Artificial Neural Network Classification Models: A Methodology Review* Journal of biomedical informatics 35(5-6) pp 352-359

[17] Baffi G, Martin E B and Morris A J 1999 *Non-linear projection to latent structures revisited (the neural network PLS algorithm)* Computers & Chemical Engineering 23(9) pp 1293-307

[18] Smits J R M, Melssen W J, Buydens L M C and Kateman G 1994 *Using artificial neural networks for solving chemical problems: Part I. Multi-layer feed-forward networks* Chemometrics and Intelligent Laboratory Systems 22(2) pp 165-189

[19] Dalwinder S, Birmohan S and Manpreet K 2020 *Simultaneous Feature Weighting And Parameter Determination of Neural Networks Using Ant Lion Optimization For The Classification Of Breast Cance* Biocybernetics and Biomedical Engineering 40(1) pp 337-351

[20] Darmawan S, Carolita I and Ananta E 2020 *Identification of Oil Palm Plantation on Multiscatter and Resolution of SAR Data Using Variety of Classifications Algorithm (Case Study: Asahan District, North Sumatera Province)* IOP Conf. Series: Earth and Environmental Science Vol 500 No 1 p 012075 IOP Publishing

[21] Rustam Z, Angie N, Pandelaki J and Yunus R E 2020 *Acute Sinusitis Classification Using Support And Fuzzy Support Vector Machines* In Journal of Physics: Conference Series Vol 1490 No 1 p 012029 IOP Publishing