Comparison of the Normalization Method of Data in Classifying Brain Tumors with the k-NN Algorithm

Rinci Kembang Hapsari[1(✉)], Abdullah Harits Salim[2], Budanis Dwi Meilani[3], Tutuk Indriyani[1], and Aery Rachman[4]

[1] Department of Informatics, Faculty of Electrical and Information Technology, Institute Teknologi Adhi Tama Surabaya, Surabaya, Indonesia
{rincikembang,tutuk}@itats.ac.id
[2] Department of Mathematics, Faculty of Science and Data Analytics, Institute Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3] Department of Information Systems, Faculty of Electrical and Information Technology, Institute Teknologi Adhi Tama Surabaya, Surabaya, Indonesia
budanis@itats.ac.id
[4] Department of Information Systems, Universitas Trunojoyo Madura, Madura, Indonesia

Abstract. One way to examine patients with brain tumors is the radiological examination, including Magnetic Resonance Image (MRI) with contrast. The classification process is needed to differentiate MRI images of people with brain tumors from those without brain tumors. The classification was based on MRI image feature extraction results with statistical features. Different **statistical** feature scale values for each dataset parameter can complicate the classification process. An unbalanced range of values can affect the quality of the classification results. For this reason, it is necessary to pre-process the data. The pre-processing method used is data transformation with normalization. Three normalization methods are used in data transformation: Min-Max normalization, z-score normalization, and T-Score Normalization. Data processed from each normalization method will be compared to see the results of the best classification accuracy using the K-NN algorithm. The k used in the comparison are 3, 5, 7, and 11. The normalized data from the dataset is divided into test data and training data with k-fold cross-validation. Based on the results of the classification test with the K-NN algorithm shows that the best accuracy lies in the Brain Tumor dataset, which has been normalized using the Min-Max normalization method with K = 3 of **85.92%. The average obtained is 79.68%.**

Keywords: Data Normalization · brain tumors · classification · k-NN

1 Introduction

Abnormal cell growth in the brain will disrupt the working system of the brain and affect nerve control in the human body. This abnormal cell growth is called a tumor. Cases of brain tumors in the world are increasing every year. Every year in Indonesia, 300

I. H. Agustin (Ed.): ICONNSMAL 2022, AISR 177, pp. 21–29, 2023.
https://doi.org/10.2991/978-94-6463-174-6_3

patients are diagnosed with brain tumors. Not only adults, but brain tumors also attack children at a relatively young age. The analysis of tumors in detecting the characteristics of cancer is a difficult task because of the variable nature of tumors and their similar properties to other brain regions. Cases of brain tumors have increased rapidly in the past. Rapid lifestyle changes and environmental conditions have had this impact [1].

The diagnosis of brain tumors is based on clinical and radiological information. Magnetic resonance imaging (MRI) is a mainstay for assessing patients with brain tumors [2]. MRI is a pioneer for imaging brain tumors in clinical practice providing structural, micro, functional, and metabolic information [3]. In addition, new advanced imaging techniques are continuously being developed to improve the identification, characterization, and assessment of brain tumor response [4]. Therefore, many artificial intelligence applications (AI) in brain tumor imaging are based on MRI. For more information on brain tumors, please refer to [5]. The evolution of brain tumor detection has resulted in various diagnostic tools and new technologies being developed to improve the performance of more accurate estimates. With the latest developments, automation in the detection of brain tumors requires an analysis of the diagnosis of brain tumors for an area to present accurate decisions [6].

In several research articles, brain tumor detection is carried out through the application of Machine Learning [7] and Deep Learning [8] algorithms. When this system is applied to MRI images, brain tumor prediction is carried out very quickly, and greater accuracy helps provide treatment to patients [8]. Computer-assisted diagnosis has proven useful in supporting medical practitioners [9]. This diagnosis can be made using different techniques, including machine learning (ML)[7].

The success of a Machine Learning (ML) method depends on data quality [10]. Therefore, the pre-processing data phase is crucial for improving ML performance [11]. Data normalization is one of the processes carried out in the pre-processing data phase. In normalization, the values are scaled back so that it can make processing easier [12]. In addition, data normalization does not lead to a large increase in memory workload and processing power requirements.

2 Research Method

In this paper we use data mining to classify using the K-Nearest Neighbor (k-NN) algorithm. The input data used is an MRI. Figure 1 shows the steps carried out in this research.

The feature extraction process is one of the important processes aiming to measure each pixel's quantitative feature size. The feature extraction results represent the characteristics of an object that can distinguish object classes properly using the 3D-GLCM method [13] in this study using six statistical characteristics, namely Max Probability, Entropy, Energy, Correlation, Contrast, and homogeneity.

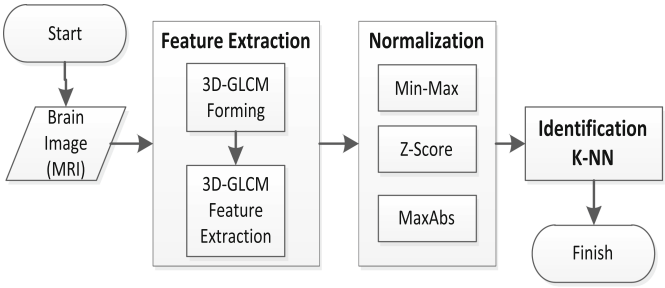

Fig. 1. Research Steps

Because the values of the six statistical features produced have high intervals, normalization is carried out. This study will compare the results of the classification accuracy of k-NN with normalization, Min-Max Normalization, Z-Score Normalization, and MaxAbs normalization methods.

A. *Research Dataset*

The dataset used in this research is the MRI Tumor dataset, which was taken from www.kaggle.com. Brain MRI dataset consisting of 2 groups, brain tumors consisting of 155 images, and no brain tumors consisting of 98 images.

B. *Normalization*

The normalization method used in this study is by transforming data into the range 0 and 1. The methods used are:

- **Min-max Normalization**. Min-Max normalization is a normalization method that carries out a linear transformation of the original data to produce a balance of comparative values between the data before and after processing [14, 15, 16]. The equation for calculating Min-Max Normalization is as follows:

$$X_{new} = \frac{X - M\,\mathrm{in}(X)}{\max(X) - \mathrm{Min}(X)} \tag{1}$$

with,

X_{new} = the new X value is the result of the original normalization.
X = value to be normalized.
$Max(X)$ = the maximum value of an attribute in the dataset.
$Min(X)$ = the minimum value of an attribute in the dataset.

- **Z-Score Normalization.** Z-score Normalization is a normalization method based on the mean and standard deviation of the data. This method is very useful if the actual minimum and maximum values of the data are unknown [17, 18]. The equation for calculating the Z-Score Normalization is as follows:

$$X_{new} = \frac{X - \mu}{\sigma} \tag{2}$$

with,

X_{new} = the new X value is the result of the original normalization
X = value to be normalized
μ = population mean
σ = standard deviation value

- **AbsMax Normalization**. MaxAbs normalization is a data normalization method that divides all values by the absolute value of the maximum value. This changes the maximum value to 1. This method does not change the data sparsity because this method does not center the data [18]. The equation for calculating MaxAbs Normalization is as follows:

$$X_{new} = \frac{X}{|Max(X)|} \tag{3}$$

with,
X_{new} = the new X value is the result of the original normalization
X = value to be normalized
$Max(X)$ = the maximum value of an attribute in the dataset

C. *k-Nearest Neighbor (k-NN)*

The K-Nearest Neighbor (K-NN) algorithm is one of the most popular NN-based methods. The K-nearest neighbor or k-NN algorithm is an algorithm that functions to classify data based on learning data (train dataset) taken from the k-closest neighbors [19]. The value of k represents the number of nearest neighbors involved in determining class label predictions in the test data.

The working principle of k-Nearest Neighbor is to find the shortest distance between the data to be evaluated and the k neighbors in the training data. From the k closest neighbors who were selected, a class vote was then carried out from the k nearest neighbors. The class with the highest number of neighboring votes is given as the predicted class label in the test data.

K-nearest Neighbor Algorithm [20]:

1. Determine Parameter K (Number of nearest neighbors).
2. Calculates the Euclidean distance (query instance) of each object against the given test data. Euclidean equation, as follows:

$$d_{ij} = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - c_{kj}\right)^2} \tag{4}$$

With,
d_{ij} =the distance between the training data point x and the testing data point to be classified
x_{ij} = training data
c_{kj} = testing data
j = data dimensions (number of attributes used)

3. Then sort these distances into groups based on the smallest Euclidean distance.
4. Gather category k (Nearest Neighbor Classification).

5. By using the most majority Nearest Neighbor category, it is used as the predicted result.

D. *Performance measurement*

Classification performance measurement is done by calculating the specificity, accuracy, and sensitivity. Accuracy measures how close the system results are to the actual value. Specificity is the precision in identifying the background area.

The sensitivity indicates how well the classification method identifies. Based on the confusion matrix, classification performance measurement can be calculated using the sensitivity, specificity, and accuracy expressed in Eqs. (5), (6), (7) [21] [22]

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} * 100\% \tag{5}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{6}$$

$$Specificity = \frac{TN}{(FP + TN)} \tag{7}$$

3 Result and Analysis

After feature extraction and normalization, identification is carried out. In this process, there are two training sub-processes and a testing sub-process. In solving the limited amount of internal data, the training and testing process uses cross-validation [23]. The classification model is formed based on the data patterns resulting from data training. Meanwhile, the algorithm's accuracy and the success rate of correctly classifying it are measured based on the results of testing the data.

The evaluation uses the confusion matrix to provide decisions obtained in training and testing. The confusion matrix provides an assessment of classification performance based on true or false objects. Based on the Confusion matrix, the performance measurement

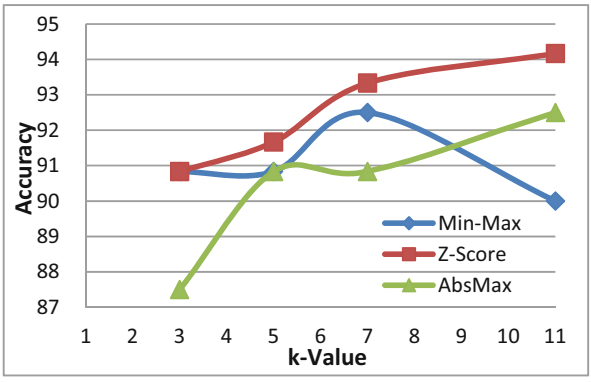

Fig. 2. Test results for the accuracy of the 3 Normalization methods

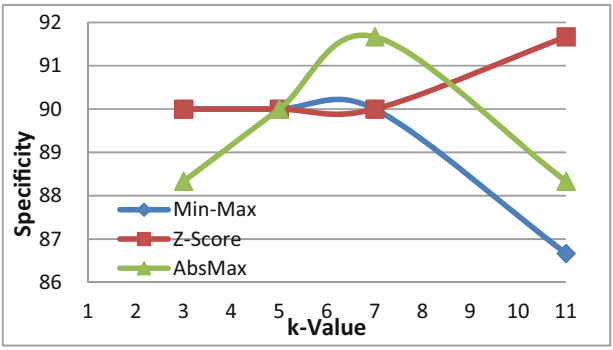

Fig. 3. Test results for the specificity values for the 3 Normalization methods

of the k-NN algorithm is measured by calculating its accuracy, specificity, and sensitivity values. Testing was carried out with 6-ford cross-validation of the entire dataset.

Feature extraction values from the dataset were transformed using the normalization method. The normalization methods compared are Min-Max normalization, Z-Score normalization and AbsMax normalization.

Tests on the three normalization methods have been carried out using the k-NN classification method, with k = 3, 5, 7, and 11. It was found that the best accuracy value using the Z-Score normalization method is shown in Fig. 2. The average accuracy value for Min-Max normalization is 91.04, the normalized Z-Score is 92.50, and the normalized AbsMax is 90.42. Moreover, of the three normalization methods, the smallest average accuracy is AbsMax normalization, and the largest is normalized Z-Score.

Figure 3 shows the test results from measuring the specificity values of the three normalization methods. The average specificity value of the normalized Min-Max method was 89.17, the normalized Z-Score method was 90.42, and the normalized AbsMax method was 89.57. Moreover, the smallest average specificity value of the three normalization methods is the Min-Max normalization, and the Z-Score normalization is the largest.

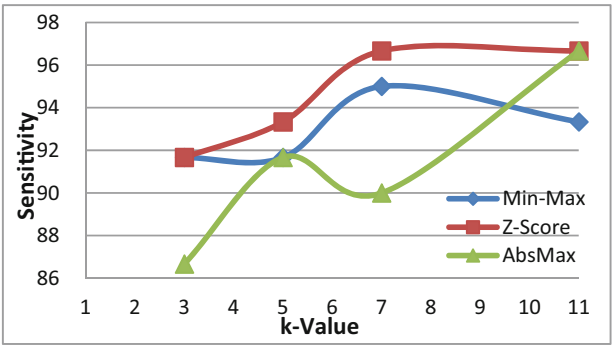

Fig. 4. Sensitivity values for the 3 Normalization methods

Figure 4 shows the test results of measuring the sensitivity values of the three normalization methods. The average sensitivity value of the normalized Min-Max method was 92.92, the normalized Z-Score method was 94.58, and the normalized AbsMax method was 91.25. And of the three normalization methods, the lowest average sensitivity is AbsMax normalization, and the largest is normalized Z-Score.

4 Conclusion

This paper presents the use of three normalization techniques in predicting brain tumors using k-NN. From the experiments conducted, it is suggested that k-NN can produce better accuracy, specificity and specifications using Z-Score Normalization compared to the other two techniques (MinMax and AbsMax). And the k-NN method will produce good performance when k = 11.

Acknowledgment. We thank the research team and Institut Teknologi Adhi Tama Surabaya (ITATS) for supporting us in holding the ICCGANT 2022 in Jember.

References

1. A. Vienne-Jumeau, C. Tafani, and D. Ricard, Environmental risk factors of primary brain tumors: A review, vol. 175, no. 10. 2019. https://doi.org/10.1016/j.neurol.2019.08.004.
2. J. E. Villanueva-Meyer, M. C. Mabray, and S. Cha, "Current clinical brain tumor imaging," Clin. Neurosurg., vol. 81, no. 3, pp. 397–415, 2017, https://doi.org/10.1093/neuros/nyx103.
3. J. T. Grist et al., "Hyperpolarized C MRI : A novel approach for probing cerebral metabolism in health and neurological disease," J. Cereb. Blood Flow Metab., 2020, https://doi.org/10.1177/0271678X20909045.
4. W. B. Overcast, K. M. Davis, C. Y. Ho, G. D. Hutchins, and M. A. Green, "Advanced imaging techniques for neuro-oncologic tumor diagnosis , with an emphasis on PET-MRI imaging of malignant brain tumors," vol. 8, 2021, https://doi.org/10.1093/neuonc/nov088.
5. M. C. Mabray and S. Cha, "Current Clinical Brain Tumor Imaging," Neuro Surg., vol. 0, no. 0, pp. 1–19, 2017, https://doi.org/10.1093/neuros/nyx103.
6. S. Hamdani, N. Dar, and R. Reshi, "Histopathological spectrum of brain tumors: A 4-year retrospective study from a single tertiary care facility," Int. J. Med. Sci. Public Heal., vol. 8, no. 0, p. 1, 2019, https://doi.org/10.5455/ijmsph.2019.0616504062019.
7. H. Peni and A. Tjahyaningtijas, "Evolution in diagnosis and detection of brain tumor – review," 2021, https://doi.org/10.1088/1742-6596/2115/1/012039.
8. S. Grampurohit, V. Shalavadi, V. R. Dhotargavi, M. Kudari, and S. Jolad, "Brain Tumor Detection Using Deep Learning Models," Proc. - 2020 IEEE India Counc. Int. Subsections Conf. INDISCON 2020, pp. 129–134, 2020, https://doi.org/10.1109/INDISCON50162.2020.00037.

9. A. Chen, L. Zhu, H. Zang, Z. Ding, and S. Zhan, "Computer-aided diagnosis and decision-making system for medical data analysis: A case study on prostate MR images," J. Manag. Sci. Eng., vol. 4, no. 4, pp. 266–278, 2019, https://doi.org/10.1016/j.jmse.2020.01.002.
10. S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," Int. J. …, vol. 1, no. 2, pp. 1–7, 2006, https://doi.org/10.1080/02331931003692557.
11. Z. Mustaffa and Y. Yusof, "A comparison of normalization techniques in predicting dengue outbreak," Int. Conf. Bus. Econ. Res., vol. 1, pp. 345–349, 2011, [Online]. Available: http://www.ipedr.com/vol1/74-G10007.pdf
12. S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," Expert Syst. Appl., vol. 106, pp. 252–262, 2018, https://doi.org/10.1016/j.eswa.2018.04.008.
13. R. K. Hapsari, M. Miswanto, R. Rulaningtyas, and H. Suprajitno, "Identification of Diabetes Mellitus and High Cholesterol Based on Iris Image," J. Hunan Univ. (Natural Sci., vol. 48, no. 10, pp. 151–160, 2021.
14. S. Ribaric and I. Fratric, "Experimental evaluation of matching-score normalization techniques on different multimodal biometric systems," Proc. Mediterr. Electrotech. Conf. - MELECON, vol. 2006, pp. 498–501, 2006, https://doi.org/10.1109/melcon.2006.1653147.
15. S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," Iarjset, no. March, pp. 20–22, 2015, https://doi.org/10.17148/iarjset.2015.2305.
16. A. S. M. Al-rawahnaa, A. Yahya, and B. Al, "Data mining for Education Sector , a proposed concept," JournalofAppliedDataSciss, vol. 1, no. 1, pp. 1–10, 2020.
17. L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, "Data Mining: A Preprocessing Engine," J. Comput. Sci., vol. 2, no. 9, pp. 735–739, 2006, https://doi.org/10.3844/jcssp.2006.735.739.
18. I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," pp. 1–18, 2022.
19. M. J. Zaki, Data Mining and Analysis : Fundamental Concepts and Algorithms. Cambridge University Press, 2013. [Online]. Available: https://books.google.co.id/books?id=PX-7zQEACAAJ
20. A. Made, S. Indra, I. Bagus, and G. Dwidasmara, "Implementation Of The K-Nearest Neighbor (KNN) Algorithm For Classification Of Obesity Levels," vol. 9, no. 2, pp. 277–284, 2020.
21. T. Indriyani, I. Utoyo, and R. Rulaningtyas, "Comparison of image edge detection methods on potholes road images," J. Phys. Conf. Ser., vol. 1613, no. 1, 2020, https://doi.org/10.1088/1742-6596/1613/1/012067.
22. T. Indriyani, M. I. Utoyo, and R. Rulaningtyas, "A New Watershed Algorithm for Pothole Image Segmentation," Stud. Informatics Control, vol. 30, no. 3, pp. 131–139, 2021, https://doi.org/10.24846/v30i3y202112.
23. R. K. Hapsari, M. Miswanto, R. Rulaningtyas, H. Suprajitno, and G. H. Seng, "Modified Gray-Level Haralick Texture Features for Early Detection of Diabetes Mellitus and High Cholesterol with Iris Image," Int. J. Biomed. Imaging, vol. 2022, 2022.