International Journal of Informatics and Information System ISSN 2579-7069
Vol. 4, No. 1, March 2021, pp. 13-20

Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer

Henderi [1], Tri Wahyuningsih [2,*], Efana Rahwanto [3]

[1,2,3] Informatics Engineering Master Program, Raharja University, Indonesia
henderi@raharja.info; triwahyuningsih@raharja.info *; efana@raharja.info
* corresponding author

(Received February 4, 2021 Revised February 22, 2021 Accepted February 27, 2021, Available online March 1, 2021)

Abstract

The purpose of this study was to examine the results of the prediction of breast cancer, which have been classified based on two types of breast cancer, malignant and benign. The method used in this research is the k-NN algorithm with normalization of min-max and Z-score, the programming language used is the R language. The conclusion is that the highest k accuracy value is k = 5 and k = 21 with an accuracy rate of 98% in the normalization method using the min-max method. Whereas for the Z-score method the highest accuracy is at k = 5 and k = 15 with an accuracy rate of 97%. Thus the min-max normalization method in this study is considered better than the normalization method using the Z-score. The novelty of this research lies in the comparison between the two min-max normalizations and the Z-score normalization in the k-NN algorithm.

Keywords: K-nearest neighbors, Min-Max Normalization, Z-Score Normalization, Breast Cancer

1. Introduction

In some datasets, there are different ranges of values for each attribute. The difference in the range of values for each attribute causes the malfunction of the attribute which has a much smaller value, compared to other attributes. Therefore, it is necessary to transform data with normalization, to equalize the range of values for each attribute with a certain scale, in order to produce well-normalized data. Data transformation with normalization can be done in several ways, namely Min-Max normalization, Z-Score normalization, Decimal Scaling normalization, Sigmoidal normalization, and Softmax normalization [1]. In this study, two normalization methods were used, namely Min-Max and Z-Score. The algorithm used was K-Nearest Neighbors (kNN), while the data used in this study was a dataset of breast cancer types.

k-NN is an algorithm or method used to classify data [2] [3]. Classification is an important stage in data mining. Classification is grouping new data or objects into classes or labels based on certain attributes. KNN is one of the nonparametric machine learning algorithms (models). A nonparametric model is a model that does not assume anything about the distribution of instances in the dataset. Nonparametric models are usually more difficult to interpret, but one advantage is that the class decision lines generated by the model can be very flexible and nonlinear.

In this research, the dataset being analyzed is related to breast cancer. Based on data, breast cancer ranks second as a cause of cancer death in women after lung cancer [4]. Today, about 1 in 8 women (12%) will develop breast cancer in their lifetime. The American Cancer Society estimates that in 2017, about 252,710 women will be diagnosed with invasive breast cancer and about 40,610 will die from the disease. Only 5% of 10% of breast cancers occur in women with a clear genetic predisposition for this disease. Most breast cancers are "sporadic" which means there is no immediate family history of the disease. The risk for developing breast cancer increases as a woman ages.

Research conducted by Pandey and Jain (2017) [5] compared data normalization using the min-max and Z Score methods on the IRIS dataset with 100% accuracy at k = 1 using the min-max normalization method and 85.71% using the z score [5]. Research conducted by Chamidah et al (2012) [1] obtained optimal classification results in breast cancer cases with an accuracy of 96.86% for the min-max normalization method and 95.68% for the Z-Score method. In the research of Nasution et al (2019) [6] comparing Data Normalization for Wine Classification Using the

k-NN Algorithm on the wine dataset, the results obtained are 65.92% for the method using min-max normalization and 65.85% for Z-Score normalization.

In this study, we used the k-NN method classification by comparing the normalized min-max and Z Score to test the accuracy of breast cancer types.

The purpose of this study was to examine the results of the prediction of breast cancer, which have been classified based on two types of breast cancer, malignant and benign. The novelty proposed lies in the comparison between the two min-max normalizations and the Z-score normalization in the k-NN algorithm.

2. Research Method

This study uses the data mining classification method with the k-NN algorithm and uses R programming. Figure 1 is the steps taken in conducting research:

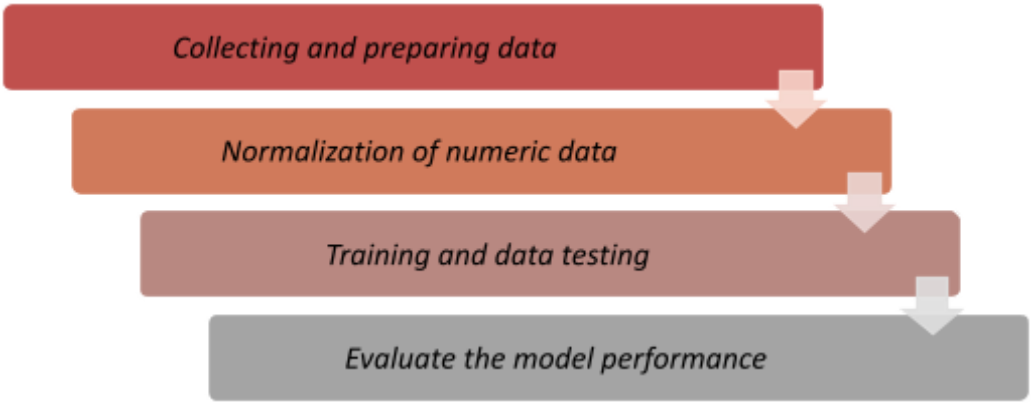

Fig. 1. Research Steps

2.1. Classification Algorithm k-NN

In general, data mining is a scientific discipline that studies methods for extracting knowledge or finding patterns from large data. Data mining is also an interactive and interactive process to get an interesting new pattern. This pattern will certainly be very useful. An interactive process means a process that still requires human interaction to be carried out. Meanwhile, the interactive process means a process that is not only done once, it needs an iterative process to get the important data in question. Models generated from the data mining process are usually perfect so that they can be generalized for future purposes. Data mining is an activity that includes collecting, using historical data to find regularities, patterns, and relationships in large data sets [7] [8]. Data mining performs extraction to obtain important information that is implicit and previously unknown, from data [9]. Other names for data mining are knowledge discovery in databases (KDD), big data, business intelligence, knowledge extraction, pattern analysis, and information harvesting. The purpose of data mining is to extract and identify data for certain information related to a large database or big data [10]. The main function of data mining according to estimation, forecasting, classification, clustering, and association [11].

Classification is the process of finding a model (or function) that describes and differentiates data classes or concepts that aim to be used to predict the class of objects whose class label is unknown [12]. The algorithms used in the classification are Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, Adaptive Credal C4.5), Naive Bayes (NB), K-Nearest Neighbor (k-NN), Linear Discriminant Analysis (LDA), Logistic Regression (LogR), and others.

The K-nearest neighbors or k-NN algorithm is an algorithm that functions to classify data based on learning data (train datasets), which are taken from k nearest neighbors [13]. Where k is the number of closest neighbors. K-nearest neighbors perform classification with learning data projections in multi-dimensional space. This space is divided into sections that represent the learning data criteria. Each learning data is represented as points c in many-dimensional space. The new classified data is then projected on a multi-dimensional space that contains c points of learning data. The classification process is carried out by finding the nearest c point from the new c (nearest neighbor). A common technique of finding the nearest neighbor is done using the Euclidean distance formula [14] which can be calculated using the following formula.

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \quad \text{..... Equation 1}$$

2.2. Collecting and preparing data

This study uses the Wisconsin Breast Cancer Diagnostic [20] dataset from the UCI Machine at http://archive.ics.uci.edu/ml. The breast cancer data included 569 cancer biopsy samples, with 32 variables including ID, diagnosis and 30 other variables were laboratory measurements. The id variable is the medical record number of a patient with breast cancer. Diagnosis variable is the determination of the health condition currently experienced by breast cancer sufferers, the diagnosis is divided into two, namely "M" and "B". For "M" denotes Malignant (malignant) and "B" denotes Benign (benign). The other 30 variables are laboratory measurements consisting of mean, se (standard error) and wort. The three variables each consist of 10 different characteristics, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. Figure 2 is the Wisconsin Breast Cancer Diagnostic dataset.

```
> str(wbcd)
'data.frame':   569 obs. of  32 variables:
 $ id                : int  87139402 8910251 905520 868871 9012568 906539 92529
1 87880 862989 89827 ...
 $ diagnosis         : chr  "B" "B" "B" "B" ...
 $ radius_mean       : num  12.3 10.6 11 11.3 15.2 ...
 $ texture_mean      : num  12.4 18.9 16.8 13.4 13.2 ...
 $ perimeter_mean    : num  78.8 69.3 70.9 73 97.7 ...
 $ area_mean         : num  464 346 373 385 712 ...
 $ smoothness_mean   : num  0.1028 0.0969 0.1077 0.1164 0.0796 ...
 $ compactness_mean  : num  0.0698 0.1147 0.078 0.1136 0.0693 ...
 $ concavity_mean    : num  0.0399 0.0639 0.0305 0.0464 0.0339 ...
 $ points_mean       : num  0.037 0.0264 0.0248 0.048 0.0266 ...
 $ symmetry_mean     : num  0.196 0.192 0.171 0.177 0.172 ...
 $ dimension_mean    : num  0.0595 0.0649 0.0634 0.0607 0.0554 ...
 $ radius_se         : num  0.236 0.451 0.197 0.338 0.178 ...
 $ texture_se        : num  0.666 1.197 1.387 1.343 0.412 ...
 $ perimeter_se      : num  1.67 3.43 1.34 1.85 1.34 ...
 $ area_se           : num  17.4 27.1 13.5 26.3 17.7 ...
 $ smoothness_se     : num  0.00805 0.00747 0.00516 0.01127 0.00501 ...
 $ compactness_se    : num  0.0118 0.03581 0.00936 0.03498 0.01485 ...
 $ concavity_se      : num  0.0168 0.0335 0.0106 0.0219 0.0155 ...
 $ points_se         : num  0.01241 0.01365 0.00748 0.01965 0.00915 ...
 $ symmetry_se       : num  0.0192 0.035 0.0172 0.0158 0.0165 ...
 $ dimension_se      : num  0.00225 0.00332 0.0022 0.00344 0.00177 ...
 $ radius_worst      : num  13.5 11.9 12.4 11.9 16.2 ...
 $ texture_worst     : num  15.6 22.9 26.4 15.8 15.7 ...
 $ perimeter_worst   : num  87 78.3 79.9 76.5 104.5 ...
 $ area_worst        : num  549 425 471 434 819 ...
 $ smoothness_worst  : num  0.139 0.121 0.137 0.137 0.113 ...
 $ compactness_worst: num  0.127 0.252 0.148 0.182 0.174 ...
 $ concavity_worst   : num  0.1242 0.1916 0.1067 0.0867 0.1362 ...
 $ points_worst      : num  0.0939 0.0793 0.0743 0.0861 0.0818 ...
 $ symmetry_worst    : num  0.283 0.294 0.3 0.21 0.249 ...
 $ dimension_worst   : num  0.0677 0.0759 0.0788 0.0678 0.0677 ...
```

Fig. 2. Wisconsin Breast Cancer Diagnostic Dataset

The patient id variable is a unique number for each patient in the data and does not provide meaningful information, so in this study the id variable was excluded from the model. Diagnosis variables are used to predict, this variable indicates whether the cancer is malignant or benign cancer. The following table 1 is the frequency of diagnosis of malignant and benign cancer.

Table. 1. Diagnosis frequency

Diagnose	Total	Percentage
Malignant	212	37.3
Benign	357	62.7

For the other 30 variables, numerical values have different measurements from each of the 10 characteristic values. The smoothness_mean level starts from 0.05263 to 0.16340, the radius_mean ranges from 6,981 to 28,110 and the area_mean ranges from 143.5 to 2501, this will have an impact on the calculation of area_mean which has a value much greater than the calculation of smoothness distance. This impact creates a classification problem, so the data needs to be normalized to change the feature scale.

2.3. Normalization of numeric data

Min-Max normalization is a normalization method by performing linear transformations of the original data so as to produce a balance of value comparisons between data before and after the process [15] [16]. This method can use the following formula [17]:

$$X_{new} = \frac{X - min(X)}{max(X) - min(X)} \quad \text{..... Equitation 2}$$

X_{new} = The new value from the normalized results

X = Old value

Max (X) = Maximum value in the dataset

Min (X) = Minimum value in the dataset

Z-score normalization is a method of normalization based on the mean (mean value) and standard deviation (standard deviation) of the data [18] [19]. This method is very useful if the actual minimum and maximum values of the data are not known. The formula used is as follows:

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - Mean(X)}{StdDev(X)} \quad \text{..... Equitation 3}$$

X_{new} = The new value from the normalized results

X = Old value

μ = Population mean

σ = Standard deviation value

2.1. Training and testing data

From the data, 569 biopsies were classified as benign or malignant. In this study, it will be tested how well the model has been given after normalization. The data will be divided into two, namely training data and testing data. Training data is used to build the k-NN model while testing data is used to estimate the accuracy of the model. This study used 469 training data while testing data used 100 data to stimulate new cancer patients. Table 4. Is the percentage comparison of the frequency of breast cancer diagnosis:

Table. 2. Percentage comparison of the frequency of breast cancer diagnosis

Diagnose	Original dataset	Training dataset	Testing dataset
Malignant	37.3	36.9	39
Benign	62.7	63.1	61

Table 4 shows that the training data used for malignant is 36.9% and for benign is 63.1%, while the comparison of testing data for malignant is 39% and benign is 61%.

3. Result and Discussion

In the data mining classification algorithm, there is an evaluation to determine the level of accuracy of the classification algorithm. The classification algorithm is divided into 2 data, namely training data and testing data. Training data is used to create a pattern in forming a classification model. Meanwhile, data testing is used to measure the accuracy of the classification algorithm whether it succeeds in classifying correctly. Evaluation uses a Confusion matrix to provide decisions obtained in training and testing. Confusion matrix, provides an assessment of classification performance based on true or false objects. To get better accuracy results, an experiment was carried

out. From the experiments conducted in this study is to calculate the overall average value. The confusion matrix provides an assessment of the classification performance based on true or false objects. Figure 3 is the result of the value of the min-max normalization method with k = 21

	wbcd_test_pred		
wbcd_test_labels	Benign	Malignant	Row Total
Benign	61	0	61
	1.000	0.000	0.610
	0.968	0.000	
	0.610	0.000	
Malignant	2	37	39
	0.051	0.949	0.390
	0.032	1.000	
	0.020	0.370	
Column Total	63	37	100
	0.630	0.370	

Fig. 3. Min-max normalization method with k = 21

Figure 3 shows the values are divided into four categories, namely true negative, true positive, false negative and false positive. The Benign column shows that 61 is true negative, this value is the case where the cancer is benign and the k-NN algorithm identifies it correctly, while the False positive category shows the number 0.The malignant column shows that a true positive result of 37 predicts a truly positive one. malignant. Whereas 2 states false negative, this indicates that the k-NN approach does not agree with the actual column, so in this case the predicted value will be benign, even though the cancer is actually malignant. This mistake is very dangerous because it can cause the patient to believe that the patient is free of malignant cancer, even though in fact the patient is exposed to malignant cancer so that if not treated properly, the cancer will continue to spread. The following figure 4 compares the accuracy with different k in the min-max normalization method.

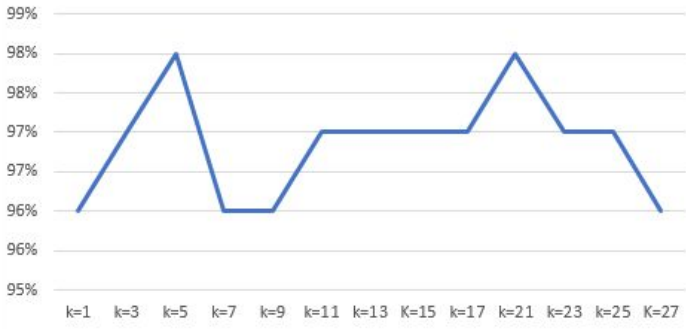

Fig. 4. Comparison of accuracy in the min-max normalization method

Based on Figure 4 above, it can be seen by the min-max normalization method the highest accuracy value at k = 5 and k = 21 with an accuracy value of 98% and the lowest result at k = 1, k = 7, k = 9 and k = 27 with values 96% accuracy.

The dataset will be transformed again using a different normalization method. The next normalization method used is the z-score normalization. The formula used in this method can be seen in equation 2. Normalization of the Z-score is done by processing the mean and standard deviation of the attribute values. Figure 5 is the result of the value of the z-score normalization method with k = 21.

wbcd_test_labels	wbcd_test_pred Benign	Malignant	Row Total
Benign	61	0	61
	1.000	0.000	0.610
	0.924	0.000	
	0.610	0.000	
Malignant	5	34	39
	0.128	0.872	0.390
	0.076	1.000	
	0.050	0.340	
Column Total	66	34	100
	0.660	0.340	

Fig. 5. Z-Score normalization method with k = 21

Figure 5 informs that the use of the Z-Score method in the Benign column shows the same information as the min-max method, namely 61 is true negative and false positive is 0. However, the malignant value shows that 34 true positive results are predicted to be true positive and malignant numbers. 5 indicates false negatives. Meanwhile, the comparison of accuracy with different k in the Z-Score normalization method is shown in Figure 6.

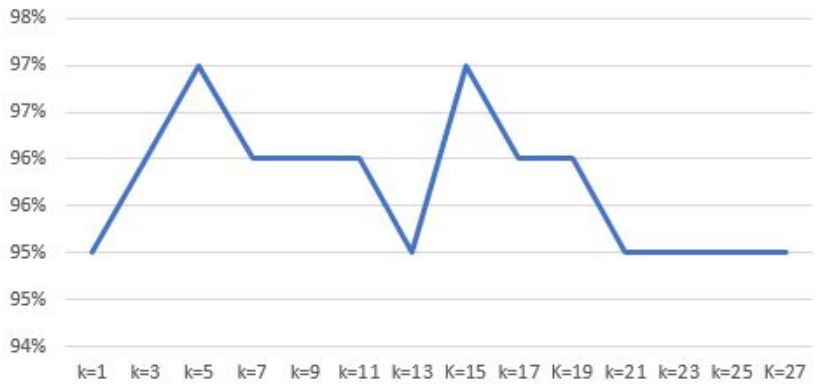

Fig. 6. Comparison of accuracy in the Z-Score normalization method

Figure 6 also shows that testing the z-score normalization method gets the highest accuracy at k = 5 and k = 15 with an accuracy value of 97% and the lowest result at k = 1, k = 13, k = 21, k = 23, k = 25 and k = 27 with an accuracy value of 95%. The test results in this study inform that the z-score normalization method has a stable accuracy between 95% to 97%. The accuracy value of the z-score method found in this study is higher than the results of research conducted by Pandey and Jain (2017) [5] on the IRIS data set, and Nasution et al (2019) [6] regarding the wine data set.

4. Conclusion

Breast cancer is classified into two, namely benign and malignant. Benign breast cancer in this model was 62.7% and malignant breast cancer was 37.3%. The training data set used for malignant breast cancer was 36.9% and for benign breast cancer was 63.1%. As for the testing dataset, 39% for malignant breast cancer and 61% for benign breast cancer. Based on the test results using the min-max normalization method, it was found that 61% of benign breast cancer was predicted to be truly benign (true negative) while benign breast cancer that was predicted to be malignant (False positive) did not exist, while malignant breast cancer was predicted to be benign. (false negative) by 2% and malignant breast cancer that was predicted to be malignant (true positive) by 37%. At the min-max normalization, the highest k accuracy value is k = 5 and k = 21 with an accuracy rate of 98%. The test results using the z-score normalization method showed that 61% of benign breast cancer was predicted to be truly benign (true negative), while benign breast cancer that was predicted to be malignant (False positive) did not exist, while malignant breast cancer was predicted to be benign (false negative) by 5% and malignant breast cancer that was really predicted to be malignant (true positive) by 34%. For the Z-score method, the highest accuracy is at k = 5 and k = 15 with an

accuracy rate of 97%. Thus the min-max method in this study is considered better than the normalization method using the Z-score and this study strengthens previous research conducted by Pandey and Jain (2017) [5].

References

[1] N. Chamidah, . W., and U. Salamah, "Pengaruh Normalisasi Data pada Jaringan Syaraf Tiruan Backpropagasi Gradient Descent Adaptive Gain (BPGDAG) untuk Klasifikasi," *J. Teknol. Inf. ITSmart*, vol. 1, no. 1, p. 28, 2016, doi: 10.20961/its.v1i1.582.

[2] Z. Pan, Y. Wang, and Y. Pan, "A new locally adaptive k-nearest neighbor algorithm based on discrimination class," *Knowledge-Based Syst.*, vol. 204, p. 106185, 2020, doi: 10.1016/j.knosys.2020.106185.

[3] J. Jiang, Y. Chen, X. Meng, L. Wang, and K. Li, "A novel density peaks clustering algorithm based on k nearest neighbors for improving assignment process," *Phys. A Stat. Mech. its Appl.*, vol. 523, no. 20180101044, pp. 702–713, 2019, doi: 10.1016/j.physa.2019.03.012.

[4] E.Technical and P. Series, Guidelines for management of breast cancer. World Health Organization, 2006.

[5] A. Pandey and A. Jain, "Comparative Analysis of KNN Algorithm using Various Normalization Techniques," Int. J. Comput. Netw. Inf. Secur., vol. 9, no. 11, pp. 36–42, 2017, doi: 10.5815/ijcnis.2017.11.04.

[6] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," Comput. Eng. Sci. Syst. J., vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[7] Santoso, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis (1 ed.). Yogyakarta: Graha Ilmu.

[8] D. J. Hand, *Principles of data mining*, vol. 30, no. 7. 2007.

[9] A. Lausch, A. Schmidt, and L. Tischendorf, "Data mining and linked open data - New perspectives for data analysis in environmental research," *Ecol. Modell.*, vol. 295, pp. 5–17, 2015, doi: 10.1016/j.ecolmodel.2014.09.018.

[10] R. S. J. Baker, "Encyclopedia of Data Warehousing and Mining," *Encycl. Data Warehouse. Min.*, 2011, doi: 10.4018/978-1-59140-557-3.

[11] M. Maggioni, "What is ... Data Mining," *Bull. Am. Math. Soc.*, vol. 59, no. 4, pp. 532–534, 2012, [Online]. Available: http://www.ams.org/notices/201204/rtx120400532p.pdf

[12] Q. Zhang, S., Zhang, C., Yang, "Dara Preparation for Data Mining," *Appl. Artif. Intel.*, vol. 17(5–6), pp. 375–381, 2003, doi: 10.1080/08839510390219264.

[13] M. J. Zaki and W. Meira, "Fundamental concepts and Algorithms," 2014.

[14] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data Mining with Big Data Xindong," *Ieeexplore.Ieee.Org*, pp. 1–26, 2014, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6547630/.

[15] A. S. M. Al-rawahnaa, A. Yahya, and B. Al, "Data mining for Education Sector , a proposed concept," *J. Appl. Data Sci.*, vol. 1, no. 1, pp. 1–10, 2020.

[16] C. Saranya and G. Manikandan, "A study on normalization techniques for privacy preserving data mining," *Int. J. Eng. Technol.*, vol. 5, no. 3, pp. 2701–2704, 2013.

[17] 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, doi: 10.1109/melcon.2006.1653147.

[18] S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," *Iarjset*, pp. 20–22, 2015, doi: 10.17148/iarjset.2015.2305.

[19] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, 2005, doi: 10.1016/j.patcog.2005.01.012.

[20] UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Center for Machine Learning and Intelligent Systems.