Scientific Journal of Informatics
Vol. 10, No. 3, Aug 2023

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Performance Analysis for Classification of Malnourished Toddlers Using K-Nearest Neighbor

Syahrani Lonang[1*], Anton Yudhana[2], Muhammad Kunta Biddinika[3]
[1*,3]Program Studi Magister Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]Program Studi Teknik Elektro, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Abstract.
Purpose: Malnutrition in toddlers is a nutritional issue that Indonesia is still dealing with. Toddlers can suffer from decreasing cognitive and physical abilities, as well as being categorized as having a high risk of death. Early detection is crucial for preventing this and providing appropriate treatment if malnutrition is detected. Classification is a machine-learning technique widely used in disease detection. Because it is simple and easy to implement, K-Nearest Neighbor is the most used classification algorithm. Detecting malnutrition can be done automatically and more quickly by utilizing classification and machine learning algorithms. The aim of this study was to analyze performance to find out which model is best for detecting malnutrition by evaluating the performance of classification using KNN with the Euclidean distance function.
Methods: The dataset used in this study is the nutritional status of toddlers from Puskesmas Ubung. The classification method proposed in this research is the KNN algorithm with Euclidean distance. There are three scenarios for the classification model that will be used. Performance classification will compare each model in terms of accuracy, precision, recall, f1-score, and mean absolute error.
Results: The experimental results show that KNN k = 15 using the first model generates excellent classification when classifying malnourished toddlers using the Euclidean distance function. The model obtains 91% accuracy, 86.6% precision, 83.8% recall, 85.2% recall, and a mean absolute error of 0.09.
Novelty: In this experiment, we analyzed the performance of the KNN to classify malnourished children using a nutritional status dataset, which resulted in an excellent classification that could be used for early detection.

Keywords: K-nearest neighbor, Machine learning, Classification, Malnutrition

Received June 2023 / **Revised** July 2023 / **Accepted** August 2023

INTRODUCTION

Malnutrition is one of the nutritional challenges that the world is currently dealing with on an individual and community level [1]. Malnutrition is a severe nutritional condition where the nutritional status of toddlers is far below the standard [2]. Indonesia is one of the developing countries with nutritional challenges. According to 2017 Nutritional Status Monitoring data, the prevalence of malnourished toddlers in Indonesia is 17.8% [3]. Malnourished toddlers tend to be triggered by a lack of protein and energy intake consumed daily for a long period of time [4]. Malnutrition causes impairment of cognition, delayed physical growth, a higher risk of death, and disease exposure [5], [6]. Therefore, early detection of malnourished toddlers is needed to prevent this from happening.

In this regard, machine learning techniques have already been implemented in health sector, such as classification, clustering, and association for disease detection purposes [7]. One of the well-known machine learning algorithms for disease classification is K Nearest Neighbor (KNN). Prediction of the possibility of getting COVID-19 [8], heart disease prediction [9], and diagnosis of osteoporosis [10] are some examples of classification with KNN. Not only in the health sector, KNN can also be applied to image classification, face identification, text classification, spam email classification, and even leaf disease detection for chili plants [11]–[15].

[*]Corresponding author.
Email addresses: lonangsyahrani3@gmail.com (Lonang)
DOI: 10.15294/sji.v10i3.45196

The KNN is a classification technique that is frequently utilized for classifying data input into pre-defined classes [16]. The prediction of the KNN classification model is solely based on neighbor values, with no assumptions about the dataset. The 'K' in KNN stands for the number of nearest neighbor data values. The KNN algorithm decides on classifying the given dataset based on 'K' or the number of nearest neighbors [17]. The confusion matrix is one of several parameters that can be used to evaluate the performance of a classification algorithm [18],[19]. Once the confusion matrix is generated, the following metric values are generated: accuracy is the percentage of accurate prediction; precision is the ratio of positively predicted instances among the retrieved instances; recall is the ratio of positively predicted instances among all the instances; and the f1-score states equilibrium between precision and recall [20],[21]. The mean absolute error for each model is also taken into consideration for evaluation. MAE is a metric that indicates how well a forecast or prediction matches the actual outcome [22]. MAE characterizes the difference between the original and predictable values and is mined as the total alteration mean of the dataset [23] .

In another research, KNN was used to classify the nutritional status of toddlers in a similar study with the highest accuracy of 85.24% with K = 3, 5, 7, and 9 [24]. Another research used different classifiers to compare the best performance for predicting diabetes. It was found that KNN performed best from SVM classifier achieving an accuracy of 83.15% [25]. In another research, KNN was used to recognize human emotions based on electroencephalogram signals, with up to 200 records collected using a Neurosky Mindwave mobile device. With a value of k = 15, the highest accuracy was 93.33% [26].

The purpose of this research is to analyze the performance of each classification model that will be tested using the toddler nutrition examination dataset. The distance function used is the Euclidean distance with K values of 1, 3, 5, 7, 9, 11, 13, and 15. Each classification result is evaluated, and the resulting level of accuracy is analyzed. Then compared to find the best and considered the most effective model used to classify malnourished toddlers.

METHODS

The analysis was carried out on performance metrics, namely accuracy, precision, recall, and F1 score, using the k-nearest neighbor algorithm. The dataset will go through a preprocessing process to check for completeness, consistency, or the absence of noise or inconsistency in the data. Preprocessing is one of the stages of clearing problems that could cause problems with the data processing's results [27]. When data is obtained through an experiment, the next step is to model the data for the purpose of extracting useful information [28].

There are three scenarios for the classification model that will be used. In the first classification model, the dataset will be broken down into 80% training data and 20% testing data. For the second classification model, the dataset will be broken down into 70% training data and 30% testing data. For the third classification model, the dataset will be broken down into 60% training data and 40% testing data. The following dataset can be seen in Figure 1.

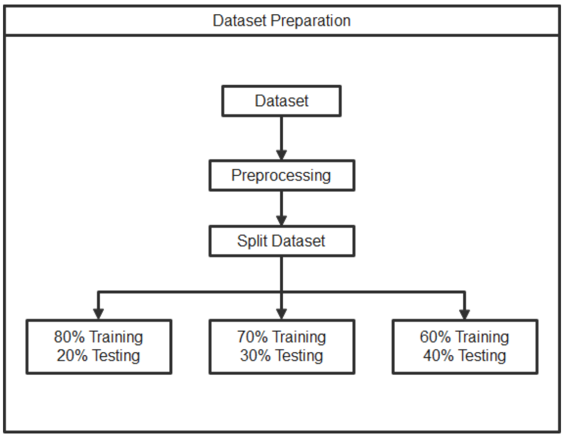


Figure 1. Dataset preparation

Each classification will use a value of k = 1, 3, 5, 7, 9, 11, 13, 15. There are twenty-four performance metrics results, namely eight from the first model, eight from the second model, and finally eight from the third model. The highest performance metric value will be chosen as the best method in this study using the KNN method. The effect of the k value and model on each classification will also be analyzed to see how big the effect is. All classification results are processed using the Python programming language and the Skicit-Learn library. The following development evaluation model to select the best model is seen in Figure 2.

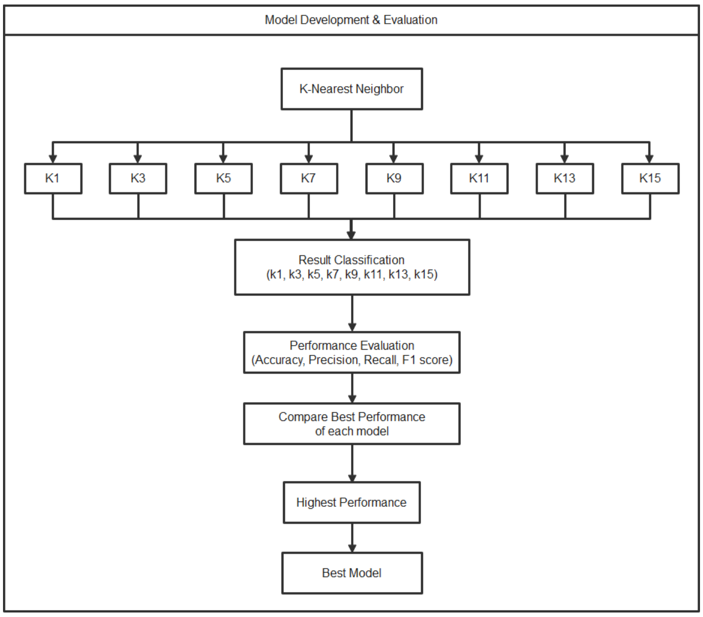


Figure 1. Model development and evaluation

Dataset

This study uses data on the nutritional status of toddlers from Puskesmas Ubung. The data collected is the data of 2021, which is hereinafter referred to as the research dataset. Each entity of the dataset, hereinafter referred to as data, represents a single nutritional status that has nine attributes and one binary class. Each attribute or criterion represents the aspects assessed to indicate the child's nutritional status, namely gender, age, weight of the toddler at the time of the checkup, height of the toddler, body weight according to height (bb/tb), z-score bb/tb, body height according to age (tb/u), z-score tb/u, and z-score body weight based on age (bb/u). Table 1 shows a snippet of the dataset.

Table 1. Snippet of the dataset

JK	Umur	Berat	Tinggi	Z-Score BB/TB	BB/TB	Z-Score TB/U	TB/U	Z-Score BB/U	Malnutrisi
1	44	6,7	81,0	-5,02	1	-4,71	1	-5,80	1
1	39	11,5	89,5	-0,97	3	-2,00	3	-1,83	0
1	38	12,0	90,4	-0,68	3	-1,63	3	-1,39	0
1	37	11,0	87,7	-1,06	3	-2,17	2	-2,01	1
1	50	12,0	85,0	0,58	3	-4,37	1	-2,40	1
0	32	12,8	89,3	0,08	3	-1,30	3	-0,62	0
0	54	13,7	100,3	-1,48	3	-1,52	3	-1,87	0
1	32	12,5	86,0	0,79	3	-1,82	3	-0,44	0
0	49	12,5	98,8	-2,33	2	-1,22	3	-2,22	1
0	53	13,6	101,2	-1,78	3	-1,14	3	-1,82	0
0	56	13,5	101,3	-1,89	3	-1,48	3	-2,10	1
0	39	14,3	95,0	0,21	3	-0,83	3	-0,34	0

K-Nearest Neighbor

KNN is an algorithm that is included in the supervised machine learning algorithm, which is easy to implement and can solve quite complex tasks [29]. There are two crucial factors that determine the performance of the KNN: the first is the distance function used, and the other is the selected k value [30], [31]. KNN, as a lazy earning model, requires that all training data instances be saved. Then, for each unseen case and each training instance, it computes a pairwise distance or similarity measure [32]. The disadvantage is that we have to compute distances for each new sample with each training sample in the dataset [33].

KNN is referred to as non-parametric because it does not assume any underlying data distribution [34],[35] .The working principle of KNN is to calculate the distance of a new sample data point from the point closest to it [36]. The KNN working principle is shown below in Figure 4. There are 2 different classes, namely class A in the shape of a red square and class B in the shape of a green triangle. The yellow circle point "?" is data that will be predicted by KNN to be included in class A or class B.

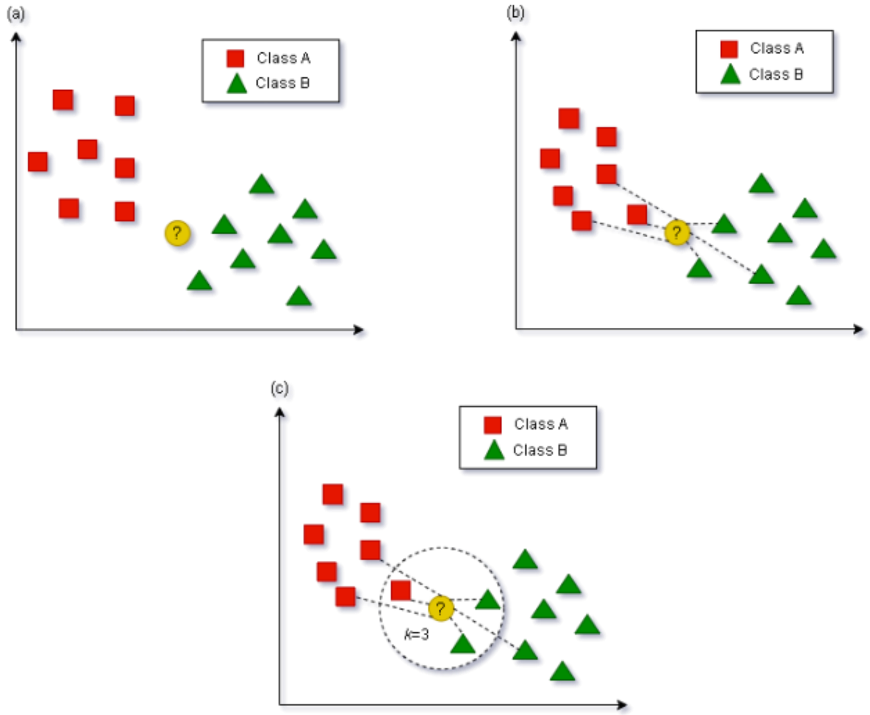


Figure 3. KNN principle. (a) inisial data, (b) calculate instance, (c) find neighbor and vote

In KNN, there are various distance functions that can be used, such as Euclidean distance, Minkowski distance, or others. The distance function that is most often used is the Euclidean distance because of its simplicity. Therefore, this research uses Euclidean distance. The function is discussed as follows:

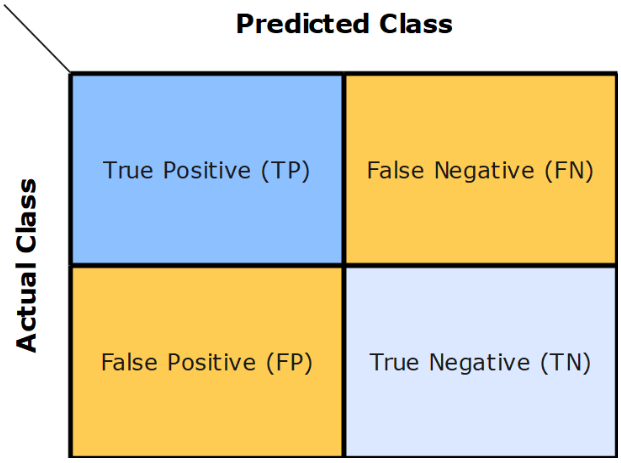
$$Euclidean\ d_{(a,b)} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (1)$$

Where:

$d_{(a,b)}$	= distance
a	= training data
b	= testing data
i	= number of attributes
n	= dimension data

Performance Evaluation

The classification will be evaluated using the performance metrics that have been considered. This metric is obtained with the help of a confusion matrix. The confusion matrix can be used to analyze more detailed algorithms in machine learning [37]. Confusion matrix is shown in Figure 5.



	Predicted Class	
Actual Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Figure 4. Confusion matrix for binary classification.

Confusion matrix represents the sum of the actual and predicted values [19]. The confusion matrix contains four numbers that are used to determine the measurement metrics of the classifier. These four numbers are: True Positive (TP) is total number of toddlers who been properly classified as malnourished. True Negative (TN) is the total number of correctly classified toddlers who are normal. False Positive (FP) is the total number of misclassified toddlers with malnourished but normal. False Negative (FN) is total number of toddlers misclassified as normal but malnourished [38].

Accuracy, precision, recall, and F1 score are performance metrics for an algorithm that are calculated based on the TP, FP, TN, FN mentioned above.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1\,Score = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)} \tag{5}$$

The classifier's accuracy is measured by the proportion of correctly classified instances to all other instances in the dataset that are represented [39]. The precision of an algorithm is the percentage of positive predictions that are correct. The recall metric is the percentage of positive labeled instances predicted as positive; sensitivity is another name for recall. F1 score or F measure is precision between recall harmonic mean [40].

RESULTS AND DISCUSSIONS

Three classification models will be used to evaluate each model's classification performance. Each classification model generates results including accuracy, precision, recall, and f1-score, each with a different k value. The classification of malnourished toddlers in this research uses the Python programming language and the Skicit-Learn library. The KNeighborClassifier function in the library is used to classify. The first step is to take the Sklearn library and import the required functions such as KNeighbor Classifier, train_test_split, confusion matrix, and classification metrics.

The second step is to load data using pandas as pd to get the dataset file (.xlsx), declare the dataset in the df variable, and check if the dataset has a missing value using function variable.isnull(). The third step is declaring the features and targets. The fourth step is to split the dataset using the train_test_split() function with three parameters: feature, target, and test size. The fifth step of generating the model is creating a KNN classifier, training the model using training sets, and predicting the response for the test dataset.

The final step is to print performance metrics using the confusion_matrix() function by comparing actual test set values and predicted values and to generate reports using metrics. Determining the best model is done using three scenarios on the dataset. The best model is generated with 80% training data and 20% testing data using k = 15. The confusion matrix of each of the best k from each existing model can be seen in Tables 2, 3, and 4.

Table 2. Confusion matrix k = 15 model 1

True label	Predicted label	
	0	**1**
0	130	8
1	10	52

Table 3. Confusion matrix k = 5 model 2

True label	Predicted label	
	0	**1**
0	194	13
1	17	76

Table 4. Confusion matrix k = 1 model 3

True label	Predicted label	
	0	**1**
0	265	1
1	23	94

Table 2 shows the confusion matrix of K = 15, which produces the best performance in the first model. With a True Negative value of 130, a true positive value of 52, a false positive value of 8, and a false negative value of 10, Similarly, Table 3 and Table 4 are the best K confusion matrices in the second and third models, respectively. TN = 194, TP = 76, FP = 13, FN = 17 on the second model; TN = 265, TP = 94, FP = 18, FN = 23.

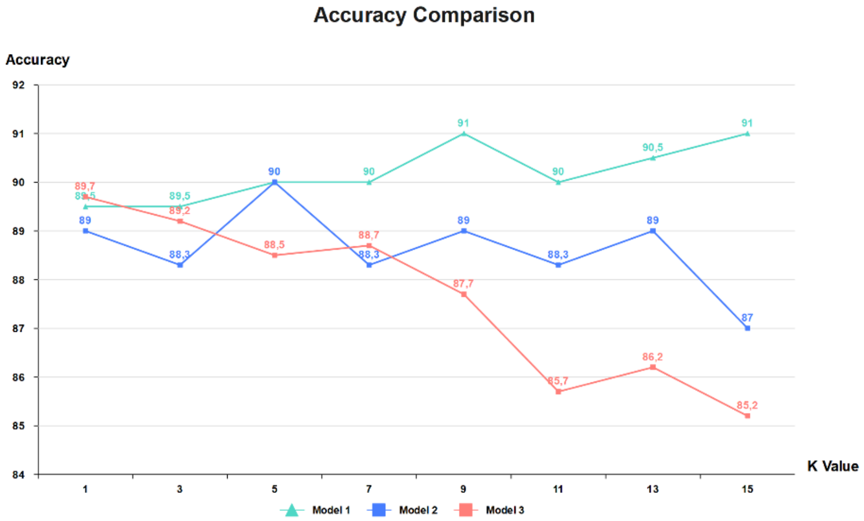


Figure 5. Accuracy comparison

The first classification model indicates that the value k = 15 produces the best performance. 91% is the highest accuracy achieved using this model. When the accuracy value is compared to another value, k=15 remains superior. Excellent classification is obtained not only on the value k = 5, but also on the values k = 9. The increase in the value of k has a direct correlation with the accuracy of the model. With an accuracy of 89.5%, the k = 3 value generates the model's lowest results. 80% of the training data set is obtained by this model. This increase impacts not only accuracy but also precision, recall, and f1-score values.

Figure 5 shows that KNN with k = 5 is superior to the other k values in this model in terms of accuracy. The second model uses more data testing and less data training than the first model. This model's best accuracy was 91.3%, while its lowest was 84.3%. The performance metrics for the third model, which has less data training than the prior two models. With an accuracy of 89.7%, k = 1 is the most optimal value of k. When compared to the prior model, the k value in this model is significantly lower. This implies that models with large training data must choose the appropriate k value to achieve high accuracy. In contrast to the third model, which used less training data, the k value required to get the best accuracy requires a low k value and every time the value of k increases, the accuracy of the third model decreases.

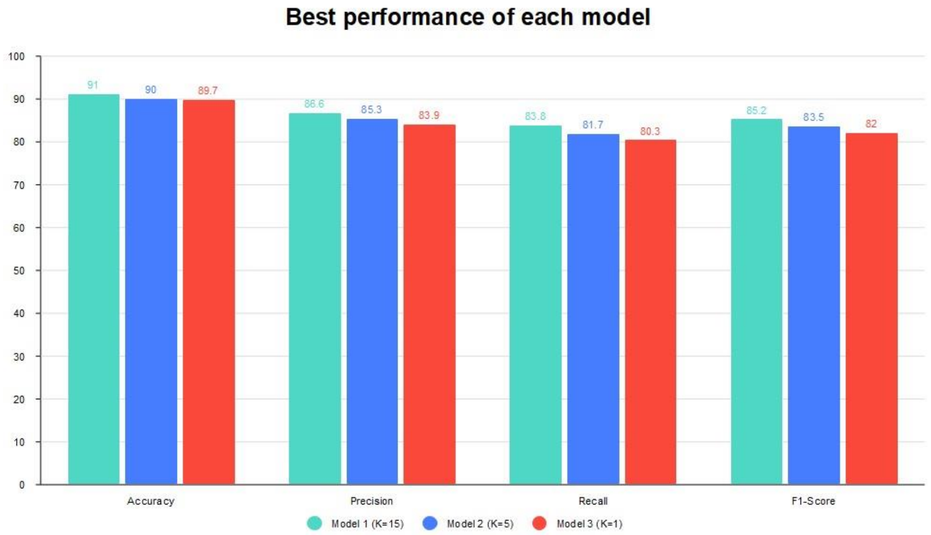


Figure 6. Best Performance of each model

Based on Figure 6, the three models tested were two that obtained excellent classification, namely the first model with an accuracy of 91%, precision of 86.6%, recall of 83.8%, and f1-score of 85.2%, and the second model with an accuracy of 90%. K=1 in the third model gets the best classification performance among the other k values. But when compared with the first-best model and second-best model, the resulting performance does not have a very significant difference—only 1.3% compared to the first model and 0.3% when compared to the second model. A comparison of the error rate for each model can also be seen in Figure 6.

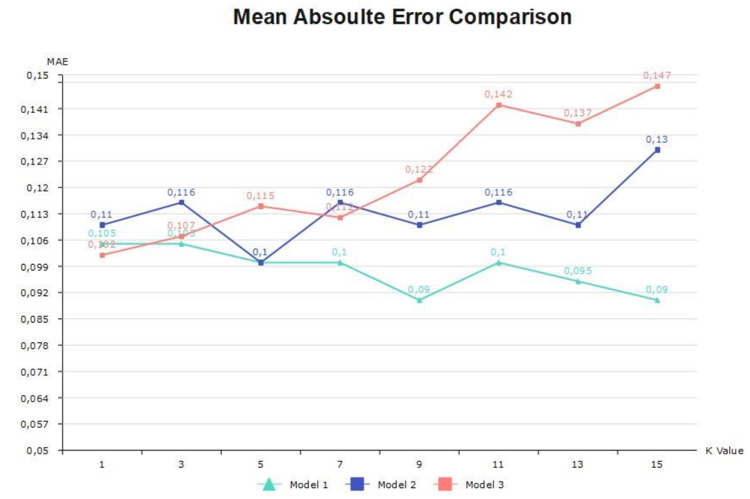


Figure 7. MAE Comparison

Based on Figure 7, When comparing the error rates of each model, the first model is still the best because it generates the lowest mean absolute error (0.09) when k = 9 and k = 15. At k = 15, the first model in this study produces the highest accuracy. In the second model, k = 5 not only generates the highest accuracy but also the lowest mean absolute error (0.10). The third model, where k = 1, generates the lowest error rate compared to the others. As value of k Increases in the model, the mean absolute error rate also rises. The best model generated in this study is the first model, with a value of k = 15.

Based on the findings of this research, which has an approach in classification methods and research objects in [24] using k-fold as the basis of the model and obtained the highest accuracy of 85.24% with a total of 210 data used. Another study [25] with the same classification method to predict diabetes obtained an accuracy of 83.15%. This indicates that this study has greater accuracy than some of the other research.

CONCLUSION

This paper uses one of the most popular classification algorithms, KNN. The distance function used is the Euclidean distance with K values of 1, 3, 5, 7, 9, 11, 13, and 15, which are implemented in the toddler nutrition examination dataset. The purpose of this research is to analyze the performance of each classification model and find the best model among the other models tested. Each model is compared while considering the different performance evaluation metrics such as accuracy, precision, recall, f1-score, and mean absolute error. The first model, with a k value of 15, has the best model performance. The highest accuracy obtained is 91%, with the smallest mean absolute error of (0.09).

The performance of each model, in terms of accuracy, precision, recall, etc., is affected by the splitting of the amount of training data and testing data as well as the value of k used. This study has an assortment of limitations, including the use of only one distance function, a limited number of k values, and three classification scenario models. This study provides recommendations for further research to add a distance function that is used to see the effectiveness of KNN in classifying. The dataset also still has a problem of imbalance in one of the classes; it is hoped that further research will solve this problem by using the Under-sampling method or the Over-sampling method. Based on these various explanations, the KNN method is effective in this case because it obtains a high level of accuracy to classify malnourished toddlers.

REFERENCES

[1] C. Lowe *et al.*, “The double burden of malnutrition and dietary patterns in rural Central Java, Indonesia,” *Lancet Reg. Heal. - West. Pacific*, vol. 14, 2021, doi: 10.1016/j.lanwpc.2021.100205.

[2] R. N. Rachmawati and N. H. Pusponegoro, “Spatial Bayes Analysis on Cases of Malnutrition in East Nusa Tenggara, Indonesia,” *Procedia Comput. Sci.*, vol. 179, pp. 337–343, 2021, doi: 10.1016/j.procs.2021.01.014.

[3] Kemenkes RI, “Buletin Jendela Data dan Informasi Kesehatan: Situasi Balita Pendek di Indonesia,” *Kementeri. Kesehat. RI*, p. 20, 2018.

[4] E. R. Arumi, Sumarno Adi Subrata, and Anisa Rahmawati, “Implementation of Naïve bayes Method for Predictor Prevalence Level for Malnutrition Toddlers in Magelang City,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 201–207, 2023, doi: 10.29207/resti.v7i2.4438.

[5] W. Hanandita and G. Tampubolon, “The double burden of malnutrition in Indonesia: Social determinants and geographical variations,” *SSM - Popul. Heal.*, vol. 1, pp. 16–25, 2015, doi: 10.1016/j.ssmph.2015.10.002.

[6] K. L. Perdue *et al.*, “Using functional near-infrared spectroscopy to assess social information processing in poor urban Bangladeshi infants and toddlers,” *Dev. Sci.*, vol. 22, no. 5, 2019, doi: 10.1111/desc.12839.

[7] V. Prasad, T. S. Rao, and M. S. P. Babu, “Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms,” *Soft Comput.*, vol. 20, no. 3, pp. 1179–1189, 2016, doi: 10.1007/s00500-014-1581-5.

[8] P. Theerthagiri, I. Jeena Jacob, A. Usha Ruby, and V. Yendapalli, “Prediction of covid-19 possibilities using knearest neighbour classification algorithm,” *Int. J. Curr. Res. Rev.*, vol. 13, no. 6 special Issue, p. S-156-S-164, 2021, doi: 10.31782/IJCRR.2021.SP173.

[9] R. Saxena, A. Johri, V. Deep, and P. Sharma, “Heart diseases prediction system using CHC-TSS evolutionary, KNN, and decision tree classification algorithm,” *Adv. Intell. Syst. Comput.*, vol. 813, pp. 809–819, 2019, doi: 10.1007/978-981-13-1498-8_71.

[10] K. A. Patil, K. V. M. Prashanth, and A. Ramalingaiah, “Texture feature extraction of Lumbar spine trabecular bone radiograph image using Laplacian of Gaussian filter with KNN classification to diagnose osteoporosis,” *J. Phys. Conf. Ser.*, vol. 2070, no. 1, 2021, doi: 10.1088/1742-6596/2070/1/012137.

[11] Z. Fan, J. K. Xie, Z. Y. Wang, P. C. Liu, S. J. Qu, and L. Huo, "Image Classification Method Based on Improved KNN Algorithm," *J. Phys. Conf. Ser.*, vol. 1930, no. 1, 2021, doi: 10.1088/1742-6596/1930/1/012009.

[12] N. K. A. Wirdiani, P. Hridayami, N. P. A. Widiari, K. D. Rismawan, P. B. Candradinata, and I. P. D. Jayantha, "Face Identification Based on K-Nearest Neighbor," *Sci. J. Informatics*, vol. 6, no. 2, pp. 150–159, Nov. 2019, doi: 10.15294/sji.v6i2.19503.

[13] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, and W. J. Huo, "The Lao text classification method based on KNN," *Procedia Comput. Sci.*, vol. 166, pp. 523–528, 2020, doi: 10.1016/j.procs.2020.02.053.

[14] E. Laksono, A. Basuki, and F. Bachtiar, "Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 377–383, 2020, doi: 10.29207/resti.v4i2.1845.

[15] A. Patil and K. Lad, "Chili Plant Leaf Disease Detection Using SVM and KNN Classification," *Adv. Intell. Syst. Comput.*, vol. 1187, pp. 223–231, 2021, doi: 10.1007/978-981-15-6014-9_26.

[16] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, and S. Abdulwahed, "Diabetes classification based on KNN," *IIUM Eng. J.*, vol. 21, no. 1, pp. 175–181, 2020, doi: 10.31436/iiumej.v21i1.1206.

[17] O. Altay and M. Ulas, "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children," *6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding*, vol. 2018-Janua, pp. 1–4, 2018, doi: 10.1109/ISDFS.2018.8355354.

[18] S. Reddi and G. V. Eswar, "Fake news in social media recognition using Modified Long Short-Term Memory network," *Secur. IoT Soc. Networks*, pp. 205–227, 2020, doi: 10.1016/B978-0-12-821599-9.00009-1.

[19] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," *Data Democr. Nexus Artif. Intell. Softw. Dev. Knowl. Eng.*, pp. 83–106, 2020, doi: 10.1016/B978-0-12-818366-3.00005-8.

[20] O. J. Awujoola, F. N. Ogwueleka, P. O. Odion, A. E. Awujoola, and O. R. Adelegan, "Genomic data science systems of Prediction and prevention of pneumonia from chest X-ray images using a two-channel dual-stream convolutional neural network," *Data Sci. Genomics*, pp. 217–228, 2023, doi: 10.1016/b978-0-323-98352-5.00013-6.

[21] D. Valero-carreras, J. Alcaraz, and M. Landete, "Computers and Operations Research Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, no. April 2022, p. 106131, 2023, doi: 10.1016/j.cor.2022.106131.

[22] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126.

[23] A. S. Rajawat, O. Mohammed, R. N. Shaw, and A. Ghosh, "Renewable energy system for industrial internet of things model using fusion-AI," in *Applications of AI and IOT in Renewable Energy*, Elsevier, 2022, pp. 107–128. doi: 10.1016/B978-0-323-91699-8.00006-1.

[24] S. Sendari, T. Widyaningtyas, and N. A. Maulidia, "Classification of Toddler Nutrition Status with Anthropometry using the K-Nearest Neighbor Method," in *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Oct. 2019, pp. 1–5. doi: 10.1109/ICEEIE47180.2019.8981408.

[25] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *Int. J. Informatics Vis.*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/joiv.3.2-2.229.

[26] A. Yudhana, A. Muslim, D. E. Wati, I. Puspitasari, A. Azhari, and M. M. Mardhia, "Human emotion recognition based on EEG signal using fast fourier transform and K-Nearest neighbor," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 6, pp. 1082–1088, 2020, doi: 10.25046/aj0506131.

[27] Junadhi, Agustin, M. Rifqi, and M. K. Anam, "Sentiment Analysis of Online Lectures using K-Nearest Neighbors based on Feature Selection," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 3, pp. 216–225, 2022, doi: 10.23887/janapati.v11i3.51531.

[28] Asundi Anand, "MATLAB for Photomechanics A Primer," p. 199, 2002.

[29] S. Ray, "A Quick Review of Machine Learning Algorithms," in *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.

[30] H. Abbad Ur Rehman, C. Y. Lin, and Z. Mushtaq, "Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease," *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 44, no. 1, pp. 77–87, 2021, doi: 10.1080/02533839.2020.1831967.

[31] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets," *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 43, no. 1, pp. 80–92, 2020, doi: 10.1080/02533839.2019.1676658.

[32] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, "kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data," *Knowledge-Based Syst.*, vol. 117, pp. 3–15, Feb. 2017, doi: 10.1016/j.knosys.2016.06.012.

[33] B. V. V. S. Prasad, S. Gupta, N. Borah, R. Dineshkumar, H. K. Lautre, and B. Mouleswararao, "Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach," *Prev. Med. (Baltim).*, vol. 174, 2023, doi: 10.1016/j.ypmed.2023.107619.

[34] A. Ali *et al.*, "Machine learning approach for the classification of corn seed using hybrid features," *Int. J. Food Prop.*, vol. 23, no. 1, pp. 1097–1111, 2020, doi: 10.1080/10942912.2020.1778724.

[35] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, "A Supervised Machine Learning Algorithms: Applications, Challenges, and Recommendations," pp. 1–10.

[36] D. M. Atallah, M. Badawy, A. El-Sayed, and M. A. Ghoneim, "Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier," *Multimed. Tools Appl.*, vol. 78, no. 14, pp. 20383–20407, 2019, doi: 10.1007/s11042-019-7370-5.

[37] B. Sahu, "Hybrid Approach for Breast Cancer Classification and Diagnosis," *EAI Endrosed Trans. Scalable Inf. Syst.*, doi: 10.4108/eai.19-12- 2018.156086.

[38] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," *Mach. Learn. Internet Med. Things Healthc.*, pp. 89–111, 2021, doi: 10.1016/B978-0-12-821229-5.00003-3.

[39] R. S. Moorthy and P. Pabitha, "Optimal Detection of Phising Attack using SCA based K-NN," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1716–1725, 2020, doi: 10.1016/j.procs.2020.04.184.

[40] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," *Knowledge-Based Syst.*, vol. 205, 2020, doi: 10.1016/j.knosys.2020.106270.