Research Article

Nurchaya Pradana Taufik Prakisyah*, Febri Liantoni, Puspanda Hatta, Yusfia Hafid Aristyagama, and Andika Setiawan

Utilization of K -nearest neighbor algorithm for classification of white blood cells in AML M4, M5, and M7

<https://doi.org/10.1515/eng-2021-0065>

received October 30, 2020; accepted March 29, 2021

Abstract: Acute myeloid leukemia (AML) M4, M5, and M7 are subtypes of leukemia derived from myeloid cell derivatives that influences the results of the identification of AMLs, which includes myeloblast, monoblast, and megakaryoblast. Furthermore, they are divided into more specific types, including myeloblasts, promyelocytes, monoblasts, promonocytes, monocytes, and megakaryoblasts, which must be clearly identified in order to further calculate the ratio value in the blood. Therefore, this research aims to classify these cell types using the K -nearest neighbor (KNN) algorithm. Three distance metrics are tested, namely, Euclidean, Chebychev, and Minkowski, and both the weighted and unweighted were tested. The features used as parameters are area, nucleus ratio, circularity, perimeter, mean, and standard deviation, and about 1,450 objects are used as training and testing data. In addition, to ensure that the classification is not overfitting, K -fold cross validation was conducted. The results show that the unweighted Minkowski distance acquired about 240 of 290 objects at $K = 19$, which is the best. Therefore, the unweighted Minkowski distance is selected for further analysis. The accuracy, recall, and precision values of KNN with unweighted Minkowski distance obtained from fivefold cross validation are 80.552, 44.145, and 42.592%, respectively.

Keywords: acute myeloid leukemia, classification, K -nearest neighbor, white blood cells

1 Introduction

Blood cancer or leukemia is a type of cancer discovered in the blood and bone marrow, which is caused by the rapid production of abnormal white blood cells, and it is one of the deadliest diseases in the world. Furthermore, its symptoms, which are sometimes difficult to detect, make it quite dangerous. Furthermore, it is caused by an excessive number of immature white blood cells in the human body, and the large number of immature blood cells actually inhibit the functioning of organs, which can lead to other diseases [1].

Leukemia is of two types, depending on the rate of growth of immature cells in the blood, acute and chronic leukemia. Both produce excess white blood cells that cannot properly function as antibodies. Acute leukemia can be recognized by the very fast multiplication of blast cells. It can lead to death in a matter of weeks or even days unless immediately and properly treated, while blast cells in chronic leukemia multiply slower than acute leukemia [2].

Acute myeloid leukemia (AML) is one of the main types, which is derived from the calculation of white blood cells from the offspring of myeloid cell. Its growth is so quick that the people with AML must get the proper and immediate treatment. Furthermore, AML is divided into eight groups of diseases based on the number of components of the white blood cells, namely, M0, M1, M2, M3, M4, M5, M6, and M7 [3]. AML subtypes are described in Table 1.

Some subtypes of AML such as M4, M5, and M7 are affected by the same type of precursor cells, and their precursor cells are myeloblast, monoblast, and megakaryoblast. They need to be precisely distinguished so that

* Corresponding author: Nurchaya Pradana Taufik Prakisyah, Universitas Sebelas Maret, Faculty of Teacher Training and Education, Jl Ir Sutami 36A, Surakarta, Jawa Tengah, Indonesia, e-mail: nurchaya.ptp@staff.uns.ac.id

Febri Liantoni, Puspanda Hatta, Yusfia Hafid Aristyagama: Informatics Education Department, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Jl Ir Sutami 36A, Surakarta, Jawa Tengah, Indonesia

Andika Setiawan: Informatics Engineering, Department of Production and Industrial Technology, Institut Teknologi Sumatera, Jl Terusan Ryacudu, Way Huwi, Jati Agung, Lampung Selatan, Indonesia

Table 1: Subtypes of acute myeloid leukemia

AML subtype	Name
M0	Undifferentiated acute myeloblastic leukemia
M1	Acute myeloblastic leukemia with minimal maturation
M2	Acute myeloblastic leukemia with maturation
M3	Acute promyelocytic leukemia
M4	Acute myelomonocytic leukemia
M5	Acute myelomonocytic leukemia with eosinophilia
M6	Acute erythroid leukemia
M7	Acute megakaryoblastic leukemia

every cell can be counted [4]. These types of cells can be grouped into more specific types, which are used as the main factor ratio for AML M4, M5, and M7. They are myeloblast, promyelocyte, monoblast, promonocyte, monocyte, and megakaryoblast. Sample images for each white blood cell types are shown in Table 2.

2 Research method

The research consists of several steps. It started with data acquisition and ended with analyzing the results. The training and testing processes are carried out as the number of folds used in cross validation. These steps are inspired from Harjoko et al. previous research [1]. All the steps are illustrated in Figure 1.

2.1 Data acquisition

The AML M4, M5, and M7 extract features data were provided by Setiawan et al. [4]. These were basically blood smear images obtained from RSUP Sardjito Yogyakarta. The images were captured by using a 21-megapixel resolution digital camera attached to Olympus ocular lens with 1,000 times magnification. Furthermore, the features were extracted to obtain the numerical data. Six features are used as inputs for this research [1,4].

- Cell area: the number of pixels that form an area of white blood cell, including nucleus and cytoplasm.

- Perimeter: the outermost part of the cell object located right next to the background image.
- Roundness: a degree of curvature measurement of an object that forms a circle.
- Nucleus ratio: the value obtained from the ratio of nucleus area divided by the area of the cell body.
- Mean: In this case, it is the average distribution of gray intensity values of each pixel in a gray scale image.
- Standard deviation: measurement of the variation or dispersion of a set of value relative to its mean. It is also known as the square root of the variance.

In all, 1,450 cell objects were used as training and testing data. Each of the data rows was then labeled with white blood cell for validation purpose. In addition, six labels were given to the object, namely, myeloblast, promyelocyte, monoblast, promonocyte, monocyte, and megakaryoblast. The detailed number of objects in each AML preparation are shown in Table 3.

Ethical approval: The use of blood smear digital image dataset has been complied with all the relevant national regulations, institutional policies and in accordance the tenets of the Helsinki Declaration, and has been approved by the Medical and Health Research Ethic Comitee (MHREC) Faculty of Medicine Gadjah Mada University – Dr. Sardjito General Hospital.

2.2 Data training

2.2.1 K-nearest neighbor

The *K*-nearest neighbor or KNN is used as the proposed classifier. KNN is a classification algorithm based on the distance between the object and its neighbors. Its purpose is to classify the new object based on attributes. Furthermore, the *K* samples of the training data in KNN is a number of nearest neighbors that are included in the contribution of the voting process [5]. The number *K* depends on the case where it is applied. When the number *K* is large, the time and storage costs are higher. However, if it is small, the nearby meter will be extremely

Figure 1: Research steps.

Table 2: Sample images of AML cells

Source: Setiawan et al. (2018).

Table 3: Number of objects

AML subtype	Myeloblast	Promyelocyte	Monoblast	Promonocyte	Megakaryoblast	Monocyte
M4	195	11	83	0	0	0
M5	0	0	274	163	0	2
M7	0	0	0	0	722	0
Total	195	11	357	163	722	2

small due to the poor information [6]. It is important to find the best value of K , and, therefore, a trial-and-concept error needs to be conducted [7].

The accuracy of the KNN algorithm is greatly influenced by the absence or presence of irrelevant features. It is also influenced by the weight of the feature, which is not equivalent to its relevance for classification [8]. In the training phase, this algorithm stores features and class vectors of the training data. In the testing phase, the same features are calculated for the testing data. When new data are entered, its classification is unclear. The distance from the new one to all the learning data vectors is calculated, and the closest number of K is taken. The new points are predicted to be included in the most classifications of these points [9].

The training data are projected into a multidimensional space in which each dimension contains a set of feature data. This space is divided into several sections consisting of a collection of learning data. A point in this space is marked as a certain class if the class is the most common classification in the closest K of that point. KNN is modeled in Figure 2.

Figure 2 displays two classes, A and B. Therefore, two K s, $K = 3$ and $K = 6$. In addition, a test data is located right in the center of the circle. If the K used is 3, it is easy to see that the proximity of the data is more inclined toward class A. However, if the K used is 6, then the test data is simply recognized as class B because it has a greater closeness to class B. Neighbors are usually calculated based on distances, and the distance between two points

Figure 2: K-nearest neighbor model sample.

can be calculated using certain distance metrics such as Euclidean, Chebychev, Minkowski, etc. [10,11].

2.2.2 Euclidean distance

The Euclidean distance is the most common distance metric used for KNN. It is a straight line distance between two data points (x, y) , where $x_i, y_i \in R$ in N -dimensional vector plane [12]. The equation for Euclidean distance is represented as mentioned in equation (1).

$$d_{\text{euclidean}}(x_i, y_i) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

The distance between two points is simply calculated by finding the root of squared difference in x and y . This formula is similar to the Pythagorean theorem formula, and, therefore, it is also known as the Pythagorean.

2.2.3 Chebychev distance

The Chebychev distance or chessboard distance or $P \infty$ metric is a distance metric defined on a vector space where the distance between points (x, y) where $x_i, y_i \in R$ is the maximum absolute distance in one-dimension of two N -dimensional points [11]. The Chebychev distance is shown in equation (2).

$$d_{\text{chebychev}}(x_i, y_i) := \max_i (|x_i - y_i|) \quad (2)$$

2.2.4 Minkowski distance

The Minkowski distance is a metric in a normed vector space, which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. It is used as the dissimilarity measurement between two vectors where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ where $x_i, y_i \in R$ in the N -dimensional vector space [13]. The Minkowski distance is represented in equation (3).

$$d_{\text{minkowski}}(x_i, y_i) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \quad (3)$$

The Minkowski distance is a generalized distance metric. The formula above can be manipulated by substituting “ p ” to calculate the distance between two data points in different ways. For the case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the case of $p = \infty$, the Minkowski distance gives the Chebychev distance [14].

2.2.5 Weighted KNN

Weighted KNN is a modified version of nearest neighbor algorithm. One of the many issues that affect the performance of the KNN algorithm is the choice of the hyperparameter K , which is sometimes less representative. In order to overcome this issue, a weight variable is added to the distance calculation formula [15]. The weight is calculated by the reciprocal of the distance, as shown in equation (4).

$$w(x_i, y_i) = \frac{1}{d(x_i, y_i)} \quad (4)$$

Where $d(x_i, y_i)$ is the distance metric function. Therefore, the formula of weighted KNN is shown as in equation (5).

$$\hat{d}(x_i, y_i) = \frac{\sum_{i=1}^n w(x_i, y_i) * d(x_i, y_i)}{\sum_{i=1}^n w(x_i, y_i)} \quad (5)$$

In this research, either unweighted or weighted models were applied to the three distance metrics proposed above so that a total of six algorithms will be compared.

2.2.6 K-fold cross validation

The validation is vital in classification to ensure the model is clean, correct, and reliable. K -fold cross validation is used as a validation method. Furthermore, K -fold is one of the most common cross-validation methods by folding data to a number of K and repeating (iterating) the training and testing process as much as K [16]. Similarly, for every single iteration, onefold is used as the test set and the rest is used as the train set. The role as the test data takes turns in accordance with the order of the K index [1]. Figure 3 is the example for fivefold cross validation.

Figure 3 shows a set of data, which is divided into five segments or folds. In the first iteration, the first segment is used as the test data, and the number of test data set is $1/5 * n$, where n is the total number of data set. Similarly, the other four segments are used as train set and then the second fold of data set is used as the test set, and the rest is used as the train set including the very first fold. This iteration is conducted five times as $K = 5$.

2.3 Data testing and validation

The data testing and validation are carried out in three stages. The first one is dividing the data into two parts,

Figure 3: K-fold cross validation.

Figure 4: Line graphs of 50 K -nearest neighbor with Euclidean, Chebychev, and Minkowski distance.

one is for training data and the rest is for testing data. Similarly, to be proportional, out of 1,450 feature data set, 1,160 objects are considered to be the training data and 290 objects as the testing data, and the selection of the training and testing classes was randomly chosen.

The next stage is testing three distance metrics to find out the best based on the most number of correctly predicted objects and the minimum K . Furthermore, each metric was tested weighted and unweighted in an increasing value of K neighborhood. This increased gradually starting from 0 and ending at 50. The metric with the highest predicted value and the lowest K value will be considered as the best. Line graphs of changes in the number of correctly predicted objects from these metrics are shown in Figure 4.

Figure 4 shows six lines representing the sum of correctly predicted object for every distance metric. Y-axis represents the number of cells and X-axis represents the number of K neighborhood. Each metric has at least one K neighborhood value that is able to obtain the highest number of cells.

The results show that the unweighted Euclidean distance successfully identified 229 objects at $K = 20$ while the weighted Euclidean distance only obtained 220 objects at $K = 11$. Both unweighted Chebychev distance at $K = 10$ and weighted Chebychev distance at $K = 27$ correctly identified 235 objects. The unweighted Minkowski distance obtained 240 objects at $K = 19$ while the weighted Minkowski distance obtained the same result at $K = 27$. Therefore, unweighted Minkowski distance continued to be analyzed in the third stage.

The final stage is conducting a cross validation with the number of fold 5, which was performed on the unweighted Minkowski distance. Each fold contains 290 data that alternate in each iteration. This method is conducted to prevent the results from overfitting.

3 Results and discussion

The experimental results show that some data can be identified properly. Every data that has been tested, either correctly or incorrectly predicted, was counted and labeled as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Furthermore, a TP is a result where the objects correctly predict the positive class. Similarly, a true negative is a result where the model correctly predicts the negative class. A false positive is a result where the model incorrectly predicts the positive class, and a false negative is a result where the model incorrectly predicts the negative class [17].

A confusion matrix is created to calculate the detailed accuracy, recall, and precision value from the best metric distance, the unweighted Minkowski distance. Accuracy is the ability of a classification model to match the actual value of the quantity measured [18]. The equation for calculating the accuracy is represented in equation (6).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \quad (6)$$

Recall is the ability of a model to find all the relevant cases within a dataset [17]. The equation for calculating recall is represented in equation (7).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \quad (7)$$

Precision is the ability of a classification model to identify only the relevant class label [17]. The equation for calculating precision is shown in equation (8).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \quad (8)$$

The prediction results from 5-fold cross validation KNN with the unweighted Minkowski distance are shown in Table 4, and some mispredicted data are shown in Table 4. Misclassifications occurred because the features possessed by several cells were very similar to each other in such a way that they have very close degrees of neighborhood. Furthermore, these data were aggregated by

Table 4: Confusion matrix of KNN with unweighted Minkowski distance

		Actual values					
		Myeloblast	Promyelocyte	Monoblast	Promonocyte	Megakaryoblast	Monocyte
Predicted values	Myeloblast	160	1	54	22	3	0
	Promyelocyte	0	0	0	0	0	0
	Monoblast	30	10	285	134	2	2
	Promonocyte	2	0	13	6	0	0
	Megakaryoblast	3	0	5	1	717	0
	Monocyte	0	0	0	0	0	0

Table 5: Recall and precision values of KNN with unweighted Minkowski distance

Blood cells	Recall (%)	Precision (%)
Myeloblast	82.0513	66.667
Promyelocyte	0	0
Monoblast	79.8319	61.5551
Promonocyte	3.681	28.5714
Megakaryoblast	99.3075	98.7603
Monocyte	0	0
Average	44.145	42.592

category, i.e., TPs, TNs, FPs, and FNs. These results were written in a confusion matrix table.

The confusion matrix is subsequently used as a basis to calculate the value of the real accuracy, recall, and precision. Similarly, each class has the same accuracy in such a way that the total accuracy for KNN with unweighted Minkowski distance is 80.552%. The average recall and precision values obtained are 44.145% and 42.592%, respectively. Table 5 shows the detailed recall and precision values for each blood cells.

By inspecting the results in Table 5, it was concluded that the KNN with unweighted Minkowski distance metric provides optimal results for accuracy only. The value of recall and precision are pretty average. This is because some object classes were not identified.

4 Conclusion

Comparing the classification of white blood cells gives interesting results. Most of the 1,450 blood cell objects were correctly identified, and the error occurred because of the wide variety of variations in white blood cells. Even some cell types have not been identified, because they

have similar characteristics, which make the classification process more difficult. The unweighted Minkowski distance's accuracy is the highest among others, with a value of 80.552%. However, the moderate recall and precision values make it less suitable for practical purpose. The given suggestion for the future research is to consider increasing the amount of data acquired from other sources. It will increase the variety of objects which will consequently allow a better generalization to the implementations of the unweighted Minkowski distance.

Acknowledgment: Authors applied the FLAE approach for the sequence of authors.

Funding information: Authors state no funding involved.

Authors contribution: Nurcahya Pradana Taufik Prakisy conducted the software coding and prepared the manuscript with contributions from all co-authors. Febri Liantoni made the conclusions. Puspanda Hatta created the model code. Yusfia Hafid Aristyagama carried out the data testing and validation. Andika Setiawan was responsible for data gathering and labeling.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The data that support the findings of this study are available from Medical and Health Research Ethic Comitee (MHREC), Faculty of Medicine Gadjah Mada University – Dr. Sardjito General Hostpital but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Medical and Health Research Ethic Comitee (MHREC), Faculty of Medicine Gadjah Mada University – Dr. Sardjito General Hostpital.

References

- 1 Harjoko A, Ratnaningsih T, Suryani E, Wiharto W, Palgunadi S, Prakisyta NPT. Classification of acute myeloid leukemia subtypes M1, M2 and M3 using active contour without edge segmentation and momentum backpropagation artificial neural network. MATEC Web of Conferences; 2018.
- 2 Suryani E, Wiharto W, Palgunadi S, Prakisyta NPT. Classification of acute myelogenous leukemia (AML M2 and AML M3) using momentum back propagation from watershed distance transform segmented images. J Phys Conf Ser. Institute of Physics Publishing. 2017;801:012044.
- 3 Suryani E, Wiharto W, Palgunadi S, Putra YR. Cells identification of acute myeloid leukemia AML M0 and AML M1 using *K*-nearest neighbour based on morphological images. 2017 International Conference on Data and Software Engineering (ICoDSE); 2018 Jan. p. 1–6.
- 4 Setiawan A, Harjoko A, Ratnaningsih T, Suryani E, Wiharto W, Palgunadi S. Classification of cell types in acute myeloid leukemia (AML) of M4, M5 and M7 subtypes with support vector machine classifier. 2018 International Conference on Information and Communications Technology (ICOIAC); 2018. p. 45–9.
- 5 Li G, Zhang J. Music personalized recommendation system based on improved KNN algorithm. 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC); 2018. p. 777–81.
- 6 Vaishnav MP, Suganya Devi K, Srinivasan P, Arutperumjothi G. Detection and classification of groundnut leaf diseases using KNN classifier. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN); 2019. p. 1–5.
- 7 Zhang N, Karimoune W, Thompson L, Dang H. A between-class overlapping coherence-based algorithm in KNN classification. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017 Jan. p. 572–7.
- 8 Wisnu H, Afif M, Ruldevyani Y. Sentiment analysis on customer satisfaction of digital payment in Indonesia: a comparative study using KNN and Naïve Bayes sentiment analysis on customer satisfaction of digital payment in Indonesia: a comparative study using KNN and Naïve Bayes. J Phys Conf Ser. Institute of Physics Publishing. 2020;1444:012034.
- 9 Harijanto B, Amalia EL, Mentari M. Recognition of the character on the map captured by the camera using *K*-nearest neighbor. IOP Conference Series: Materials Science and Engineering. Vol. 732; 2020. p. 012043.
- 10 Williams J, Li Y. Comparative study of distance functions for nearest neighbors. Advanced techniques in computing sciences and software engineering. Dordrecht: Springer; 2010 Jan. p. 79–84.
- 11 Zaman FHK, Yassin IM, Shafie AA. Ensembles of large margin nearest neighbour with grouped lateral patch arrangement for face classification. 2016 IEEE 4th International Symposium on Robotics and Intelligent Sensors (IRIS); 2017 Dec. p. 6–12.
- 12 Singh A, Pandey B. An Euclidean distance based KNN computational method for assessing degree of liver damage. 2016 International Conference on Inventive Computation Technologies (ICICT); 2016.
- 13 Klimo M, Škvarek O, Tarabek P, Šuch O, Hrabovský J. Nearest neighbor classification in Minkowski quasi-metric space. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA); 2018. p. 227–32.
- 14 Khaldi B, Harrou F, Cherif F, Sun Y. Improving robots swarm aggregation performance through the Minkowski distance function. 2020 6th International Conference on Mechatronics and Robotics Engineering (ICMRE); 2020. p. 87–91.
- 15 Fan GF, Guo YH, Zheng JM, Hong WC. Application of the weighted *k*-nearest neighbor algorithm for short-term load forecasting. Energies. 2019;12(5):916.
- 16 Yadav S, Shukla S. Analysis of *k*-fold cross-validation over hold-out validation on colossal datasets for quality classification. 2016 IEEE 6th International conference on advanced computing (IACC); 2016. p. 78–83.
- 17 Rajpurohit S, Patil S, Choudhary N, Gavasane S, Kosamkar P. Identification of acute lymphoblastic leukemia in microscopic blood image using image processing and machine learning algorithms. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2018. p. 2359–63.
- 18 Uddhavraokatkar B, Barkade V. Prediction of isolated erythrocytes formation from compound erythrocytes in blood cells images. 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA); 2018. p. 5.