has an average of 7377.0530 pixels. Furthermore, myeloblast's maximum area is only 14675 pixels, which makes it the cell with the highest roundness and nucleus ratio compared to the other three. Table 5 shows that its roundness average is of 0.8629, and its nucleus ratio average is 0.8314.

Promyelocyte and monoblast cell types are closely related. They both have an average value of four geometrical features that are not much different. However, both have differences in the value of color features, such as mean and standard deviation. The cytoplasmic color of monoblasts is purer blue while in the cytoplasm the promyelocyte cells appear pinker because there are visible granules.

Support cells have a relatively diverse range of feature values, and that is because there are more than only just one form of cells in the support cells type. Examples include lymphocytes, myelocytes, plasma and segment cells. They were deliberately grouped into a new type for easy classification.

Before entering the training phase, the raw data had to be normalized first. This method needed to be done because the extracted feature data still had a wide range of values. The range and data type of each feature can be described as follows:

- The cell area and perimeter have a range of original integer values.
- The roundness and nucleus ratio has a range of real number values between 0 and 1.
- Mean and standard deviation in the form of real numbers with range limitations between 0 and 255.

After the normalization, all features range were changed from varying scale from between 0 to 1. This scale would simplify the calculation process in classification.

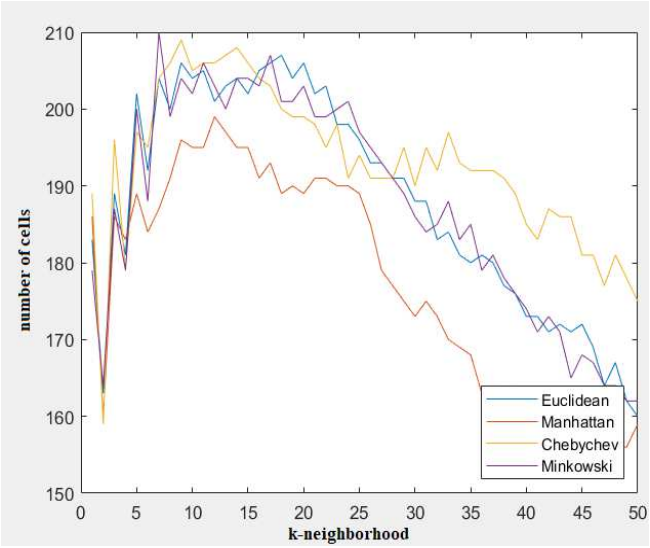

Fig. 4 Line graph of 50 K-nearest neighbor.

In the first stage, selecting the best distance metrics was carried out by dividing training and testing data into 434 and 300 through random data sharing. Four distance metrics were tested to find out the best one based on the maximum number of correctly predicted objects and the minimum K number. Furthermore, each metrics were tested in increasing value of K-neighborhood. It increased gradually starting from 0 and ending at 50. The result of 50 times k-NN iteration is shown in Figure 4.

X-axes is the number of k-neighborhood, and y-axes are the number of correctly predicted objects. The number of correct objects is perpendicular regardless of the size of k-neighborhood. Furthermore, the blue, red, yellow, and purple lines represent K-NN testing with Euclidean, Manhattan, Chebyshev and Minkowski distances, respectively.

Figure 4 shows that the Euclidean distance successfully identified 207 out of 300 objects at K=18. The second metric, Manhattan distance, only correctly identified 197 out of 300 objects at K=13. Chebyshev distance got 209 out of 300 correct objects at K=9. Meanwhile, Minkowski distance could acquire 210 out of 300 objects at K=7. Thus, Minkowski was chosen for the next steps as the best distance metric for classifying AML cells.

It was then later validated by using k-fold cross-validation in stage two. Also, the number of k-folds used in this study was 5. Therefore, each fold was 1/5 of the total data. Thus, it can be said that every fold could hold 147 data.

In the first iteration of 5-fold cross-validation, Fold number 1, which had 147 data, was used as the test data set. The rest four folds which contained 587 total data, were used as the train data set. This experiment was repeated five times, according to the proposed architecture. The test data partition position shifted in each iteration, in such a way that in the second iteration, the position of the test data set would be in the second fold and so on.

The experimental results show that some data can be appropriately identified. Every data that has been tested, whether correctly or incorrectly predicted, were counted. Table 6 shows the prediction results from 5-fold cross-validation.

TABLE VI
5-FOLD CROSS-VALIDATION

Fold	Correctly predicted	Incorrectly predicted	Subtotal
1	119	28	147
2	124	23	147
3	127	20	147
4	127	20	147
5	119	27	146
Total			**734**

There are some mispredicted data seen in Table 6. Misclassifications occurred because the features possessed by some cells were very similar such that they had very close degrees of neighborliness. Furthermore, these data are aggregated by category, i.e. true positive and negative, false positive and negative.

A true positive is an outcome where the objects correctly predict the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. Furthermore, a false positive is an outcome where the model incorrectly predicts the positive class, and a false negative is an outcome where the model incorrectly predicts the negative class [5]. Table 7 shows the confusion matrix from k-NN.

TABLE VII
CONFUSION MATRIX

		Actual values *Myeloblast*	*Promyelocyte*	*Monoblast*	*Others*
Predicted values	***Myeloblast***	347	19	5	20
Predicted values	***Promyelocyte***	13	96	15	8
Predicted values	***Monoblast***	0	3	2	0
Predicted values	***Others***	17	11	7	171

Confusion matrix was subsequently used as a basis in calculating the value of accuracy, recall and precision. Each class has the same accuracy value that totaled 83.9237% from this experiment. Table 8 shows detailed recall values for each class. The average recall value obtained from the table is 64.822%.

TABLE VIII
DETAILED RECALL OF EACH CLASS

Type	Recall
Myeloblast	92.0424403 %
Promyelocyte	74.4186047 %
Monoblast	6.8965517 %
Others	85.9296482 %
Average	**64.822 %**

Table 9 shows detailed precision values for each class. The average precision value obtained from the table was 77.788%.

TABLE IX
DETAILED PRECISION OF EACH CLASS

Type	Precision
Myeloblast	88.7468031 %
Promyelocyte	72.7272727 %
Monoblast	66.6666667 %
Others	83.0097087 %
Average	**77.788 %**

IV. CONCLUSION

Out of the four metrics offered in this study, Minkowski distance was chosen as the best metric capable of identifying white blood cell types forming leukemia M1, M2 and M3. This result is proved by the acquisition of the highest number of correctly predicted objects and the lowest k-neighborhood value obtained compared to the other three metrics in the test step with 210 out of 300 objects at k-neighborhood = 7. KNN with Minkowski distance was further analyzed with cross-validation to obtain accuracy, recall and precision. Although it can predict all object classes well, proved by 83.923% accuracy, it is less able to determine all the relevant class in the data set. Furthermore, the recall and precision obtained was 64.882% and 77.788%. The error occurred due to the variations in white blood cell that were too diverse. Some of which even had an only small portion of true positive results. They had a similar characteristic which makes the classification process more difficult. The given suggestions for the next research are the use of deep learning or genetic algorithm to classify blood cell types. Also, the amount of data needs to be increased so that the research validity can be better.

ACKNOWLEDGMENT

We would like to thank dr. Sardjito Hospital Yogyakarta and Harjoko for the permission to use the Acute Myeloid Leukemia: M1 M2 and M3 extracted features data.

REFERENCES

[1] A. Setiawan, A. Harjoko, T. Ratnaningsih, E. Suryani, Wiharto, and S. Palgunadi, "Classification of cell types in Acute Myeloid Leukemia (AML) of M4, M5 and M7 subtypes with support vector machine classifier," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, no. Cml, pp. 45–49, 2018.

[2] P. Sachin and R. Y. Kumar, "Detection and Classification of Blood Cancer from Microscopic Cell Images Using SVM KNN and NN Classifier," *Int. J. Adv. Res.*, vol. 3, no. 6, pp. 315–324, 2017.

[3] E. Suryani, Wiharto, S. Palgunadi, and N. P. T. Prakisya, "Classification of Acute Myelogenous Leukemia (AML M2 and AML M3) using Momentum Back Propagation from Watershed Distance Transform Segmented Images," in *Journal of Physics: Conference Series*, 2017, vol. 801, no. 1.

[4] A. Harjoko, T. Ratnaningsih, E. Suryani, Wiharto, S. Palgunadi, and N. P. T. Prakisya, "Classification of acute myeloid leukemia subtypes M1, M2 and M3 using active contour without edge segmentation and momentum backpropagation artificial neural network," in *MATEC Web of Conferences*, 2018, vol. 154.

[5] S. Rajpurohit, S. Patil, N. Choudhary, S. Gavasane, and P. Kosamkar, "Identification of Acute Lymphoblastic Leukemia in Microscopic Blood Image Using Image Processing and Machine Learning Algorithms," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, no. Cll, pp. 2359–2363, 2018.

[6] M. H. Waseem *et al.*, "On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction," *IEEE Access*, vol. 7, pp. 141072–141082, 2019.

[7] S. S. Devi, A. Roy, M. Sharma, and R. H. Laskar, "kNN Classification Based Erythrocyte Separation in Microscopic Images of Thin Blood Smear," *Proc. - Int. Conf. Comput. Intell. Networks*, vol. 2016-Janua, pp. 69–72, 2016.

[8] M. P. Vaishnnave, K. Suganya Devi, P. Srinivasan, and G. Arutperumjothi, "Detection and classification of groundnut leaf diseases using KNN classifier," *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–5, 2019.

[9] N. Zhang, W. Karimoune, L. Thompson, and H. Dang, "A between-class overlapping coherence-based algorithm in KNN classification," *2017 IEEE Int. Conf. Syst. Man, Cybern. SMC 2017*, vol. 2017-Janua, pp. 572–577, 2017.

[10] B. Harijanto, E. L. Amalia, and M. Mentari, "Recognition of the character on the map captured by the camera using k-nearest neighbor," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 732, p. 012043, 2020.

[11] H. Wisnu, M. Afif, and Y. Ruldevyani, "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," 2020.

[12] N. Krithika and A. Grace Selvarani, "An individual grape leaf disease identification using leaf skeletons and KNN classification," *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*, vol. 2018-Janua, pp. 1–5, 2018.

[13] A. Singh and B. Pandey, "An Euclidean Distance based KNN Computational Method for Assessing Degree of Liver Damage," *Int. Conf. Inven. Comput. Technol.*, 2016.

[14] J. Williams and Y. Li, "Comparative Study of Distance Functions for Nearest Neighbors," *Adv. Tech. Comput. Sci. Softw. Eng.*, no. January, 2010.

[15] M. Klimo, O. Škvarek, P. Tarabek, O. Šuch, and J. Hrabovsky, "Nearest neighbor classification in minkowski quasi-metric space," *DISA 2018 - IEEE World Symp. Digit. Intell. Syst. Mach. Proc.*, pp. 227–232, 2018.

[16] B. Khaldi, F. Harrou, F. Cherif, and Y. Sun, "Improving robots swarm aggregation performance through the Minkowski distance function," *2020 6th Int. Conf. Mechatronics Robot. Eng. ICMRE 2020*, pp. 87–91, 2020.

[17] F. H. K. Zaman, I. M. Yassin, and A. A. Shafie, "Ensembles of large margin nearest neighbour with grouped lateral patch arrangement for face classification," *IRIS 2016 - 2016 IEEE 4th Int. Symp. Robot. Intell. Sensors Empower. Robot. with Smart Sensors*, no. December, pp. 6–12, 2017.

[18] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," *Proc. - 6th Int. Adv. Comput. Conf. IACC 2016*, no. Cv, pp. 78–83, 2016.