Comparison between Support Vector Machine and K-Nearest Neighbor Algorithms for Leukemia Images Classification using Shape Features

1st Yessi Jusman
Department of Electrical Engineering
Faculty of Engineering, Universitas Muhamadiyah Yogyakarta
Yogyakarta, Indonesia
*Corresponding Email:
yjusman@umy.ac.id

2nd Aisyah Nur Hasanah
Department of Electrical Engineering
Faculty of Engineering, Universitas Muhamadiyah Yogyakarta
Yogyakarta, Indonesia

3rd Kunnu Purwanto
Department of Electrical Engineering
Faculty of Engineering, Universitas Muhamadiyah Yogyakarta
Yogyakarta, Indonesia

4th Siti Nurul Aqmariah Mohd Kanafiah
School of Mechatronics Engineering, Universiti Malaysia
Perlis, Malaysia

5th Slamet Riyadi
Department of Information Technology
Faculty of Engineering, Universitas Muhamadiyah Yogyakarta
Yogyakarta, Indonesia

6th Rosline Hassan[a], 7th Zeehaida Mohamed[b]
[a]D*epartment of Haematology*
[b]D*epartment of Microbiology and Parasitology, Universiti Sains Malaysia*
Kelantan, Malaysia

2021 25th International Computer Science and Engineering Conference (ICSEC) | 978-1-6654-1197-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSEC53205.2021.9684585

***Abstract*—Leukemia occurs when the body produces abnormal white blood cells in amounts exceeding the normal limit, making them misfunctioning. It is highly influential on the human immune system. Currently, medical personnel require a long time to recognize leukemia, and it is difficult to distinguish between acute leukemia cells and normal cells. Hence, this study aims to build a system program using white blood cell images with image processing using feature extraction with the Hu moments invariant and the Support Machine Machine (SVM) and K-Nearest Neighbor (K-NN) classification methods. The samples used were 800 blood images divided into two classes, acute and normal, with each class consisting of 400 sample images. Based on the test results from comparing the average value of accuracy and training time in both methods, the highest accuracy value was in the SVM method, with an accuracy of 87.97% and the K-NN method of 83.96%. The fastest training time was in the K-NN method of 2.43 seconds and the SVM method of 3.73 seconds.**

Keywords—Hu Moments, K-Nearest Neighbor, Leukemia, Acute Leukemia, Normal Leukemia, Support Vector Machine

I. INTRODUCTION

Leukemia is a cancer of the blood cells originating from the bone marrow. The proliferation of white blood cells usually characterizes it by manifesting abnormal cells in the peripheral blood (blast cells) in excess and causing the suppression of normal blood cells, resulting in impaired function.

Identification and classification of leukemia cancer are crucial because treatment varies according to the subtype of leukemia. The conventional approach to classify cancer based on morphological characteristics has been found to be inadequate due to the underlying complexity and ambiguity in cancer classification. Thus, it takes highly skilled resources to detect differences among tumour cells. This procedure is immensely time-consuming and expensive. In other words, such a handling procedure is an inappropriate solution. Cells can appear morphologically the same but react in stark contrast to drugs and cytotoxic treatments.

Several studies have attempted to develop a computer-assisted system with digital image processing methods and different classification methods to help deal with this leukemia problem. Starting from research using hybrid hierarchical classifiers and Fuzzy C Means (FCM) based on morphological contour segmentation to the application of the watershed algorithm and Gray Level Co-occurrence Matrix (GLCM) in leukemia cells images are discussed [1], [2], [3], [4], [5], [6], [7], [8], [9], [10].

Machine learning system is mostly be applied to classification purpose of biomedical images. The system can be aided the expert to obstacle the time consuming procedure in the manual diagnosis. SVM and K-NN methods are mostly implemented in many researches related to classification system for diagnosis purposes [11], [12]. Several machine learning for leukemia have been presented and discussed [13], [14], [15], [16], [17].

Rawat et al. designed leukemia computer aided system based on texture and shape features with 89.8% of accuracy [13]. Another researcher segmented cell nucleus by FCM algorithm and extracted features (i.e. geometric and statistical features) obtained from the nucleus [14]. Two Bare-bones Particle Swarm Optimization (BBPSO) algorithms proposed to identify the most significant discriminative characteristics of healthy and blast cells could achieved 94.94% of accuracy [15].

K means clustering, marker-controlled watershed and HSV color-based segmentation algorithm and SVM are used for leukemia cell classification [16]. Monte Carlo cross-validation nested by 10-fold cross-validation was used to rank clinical variables on the randomly split training sets and a forward feature selection algorithm was employed to find the shortest list of most discriminatory variables. The classification of leukemia cells using random forest model achieved best accuracy of 82.9% [17].

The shape of the imagery features has the most crucial role in classifying the images effectively and efficiently. Hu moments are normally extracted from the silhouette or outline of an object in an image. By describing the silhouette or outline of an object, we are able to extract a shape feature vector (i.e. a list of numbers) to represent the shape of the object. Some studies have used Hu moment invariant algorithms in recognizing the shape [18], [19], [20], [21], [22], [23]. Based on the literature review, the implementations

978-1-6654-1197-4/20/$31.00 ©2021 IEEE

of hu moment invariant algorithm for feature extraction are limited. Thus, this research will take the gap in this research topic.

This study discusses the classification of normal white blood cells and acute leukemia cells using the Hu moment invariants algorithm for feature extraction with the linear type Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) classification methods by comparing the results of the two methods. The best value of the results can differentiate between acute and normal leukemia cells. This research is expected to help the medical realm in handling white blood cell cancer quickly and accurately. Thus, it inhibits white blood cell cancer from spreading to other tissues and eases medical personnel to diagnose the type of leukemia patients suffer, thereby reducing the mortality rate due to delays in handling and diagnosing the type of leukemia.

II. METHODOLOGY

A. System Design

System design is one of the essential stages in applying a research concept. Thus both programs and program results can run as intended. In this study, the system design was carried out using algorithms, and training was performed on research data. This study required software and hardware as performance support devices to assist in the system design process. The hardware used is demonstrated in Table 1, while the software was MATLAB R2019b using an application on the MATLAB system in the form of a classification learner.

TABLE I. HARDWARE SPECIFICATIONS

Specification	Description
Processor	Intel(R) Core(TM) i5-4200U @ 2,30 GHz
System Type	64-bit Operating System
RAM	8.00 GB

This study utilized 100 normal images and 100 acute images (Figure 1) obtained from patients at the Universiti Sains Malaysia hospital. The preprocessing step is augmentation process for the 100 normal and 100 acute images. Thus, total of 800 images of leukemia cells consisting of 400 acute and 400 normal images are used as the image data. The images were divided into ten datasets to facilitate the training and testing stages. The images were divided into 10-fold using the K-fold cross-validation method; each fold encompassed training image data and testing image data with a ratio of 90:10 comprising 720 training images and 80 testing images.

The system design process consisted of two main stages: the training and testing. The first stage was training on the data training, comprising the feature extraction process with the hu moment invariants, and the training process using the linear type SVM and the fine type K-NN. In the SVM and K-NN training process, a database was obtained in the forms of accuracy, confusion matrix, and training time used as a classification process in the testing stage and comparing the two methods. The stages of the training process are displayed in Figure 2.

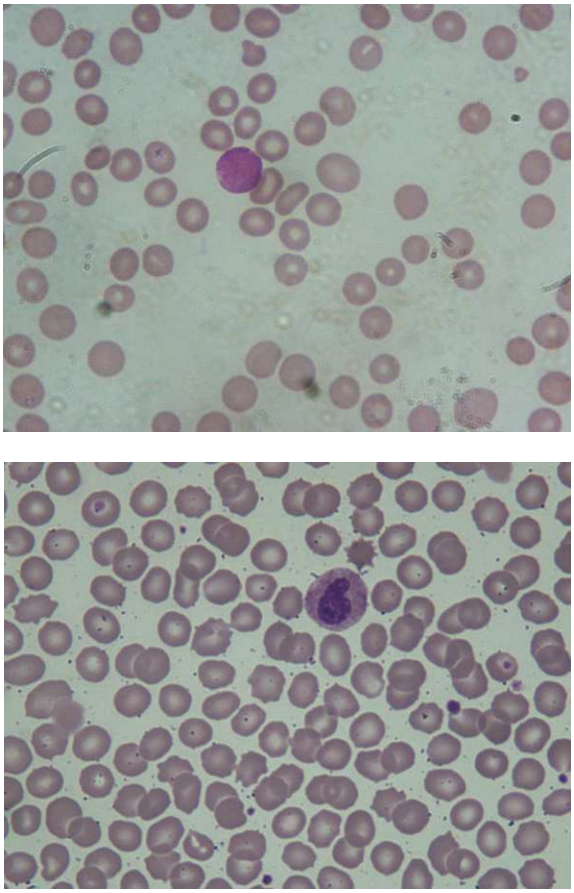

Fig. 1. Example of Leukemia Images

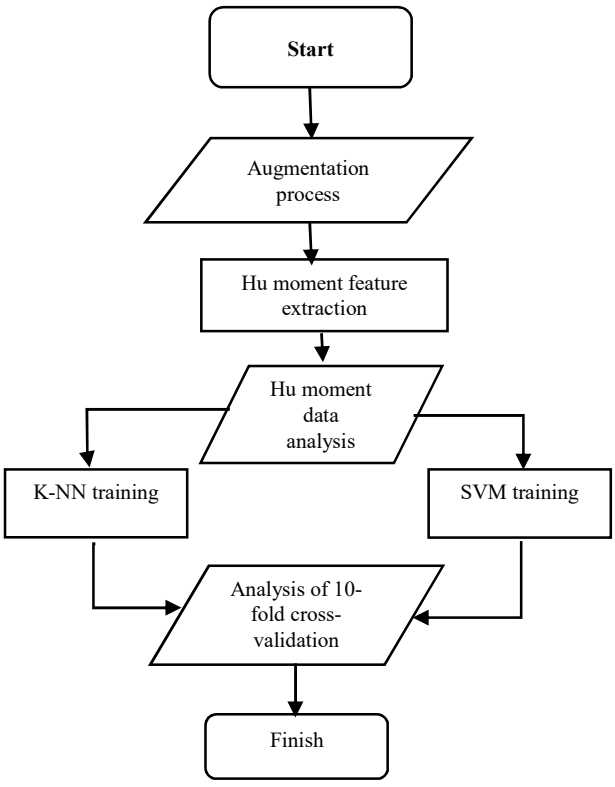

Fig. 2. Design of the training process system

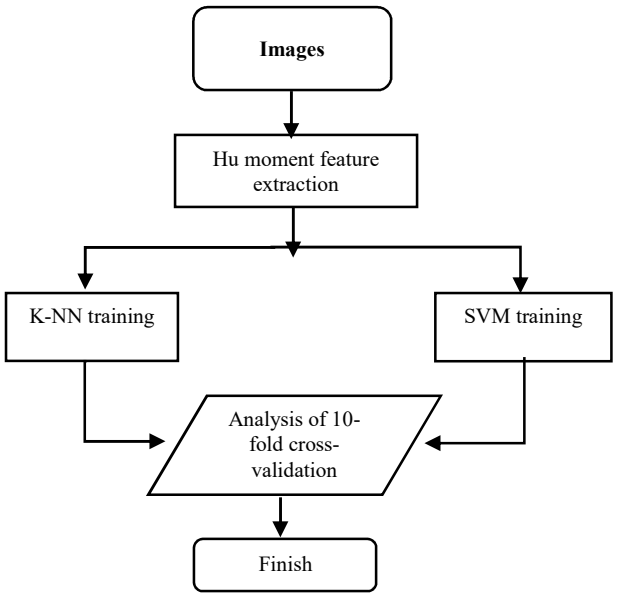

Fig. 3. Design of the testing process system

At the testing stage on the testing data, several processes were carried out to classify the images. The testing processes comprised a feature extraction process using the Hu moments invariant algorithm. The hu moments are normally extracted from the outline of the cells in an image. It is used in this research due to the different of the cells shape between normal and abnormal cells based on the shape feature vectors.

The results from the linear type SVM training process and the previous fine type K-NN were employed in this classification testing process. Figure 3 illustrates the system design flow of the testing process. The detail information of the system is presented.

1. Feature extraction with Hu moment invariant

Feature extraction aims to obtain significant data (i.e. shape features vectors) on the images. Hence, cells in the images can be distinguished from one another. The shape features of the cells have the most crucial role in identifying the shape effectively and efficiently to distinguish the images of acute and normal leukemia cells.

2. Hu moment invariants data analysis

The data from the feature extraction of Hu moment invariant algorithm were in the form of seven feature moments (Hu's seven-moment invariants) in phi values totaling seven values consisting of phi 1, phi 2, phi 3, phi 4, phi 5, phi 6, and phi 7. The seven phi values were analyzed by calculating the average value of each 360 acute and 360 normal images and the standard deviation value. Image features at the training stage from the feature extraction results were labeled on each image data following its class.

3. SVM and K-NN training data

At the training stage, the training data from excel data were converted into matrix data to facilitate calculations. Then, they were entered into the application on the MATLAB system in the form of a classification learner using the K-fold cross-validation with the linear type SVM and the fine type K-NN, resulting in the value of training accuracy, confusion matrix analysis results, and training time.

4. Data analysis

Based on the training data results using the data preparation method of 10-fold cross-validation, the best weight from each training process, both with the SVM and K-NN models, were taken to be used as the weight of the system being built. Testing was carried out to determine whether the system built could work and function properly. Data from classification results were in the form of the system's success in distinguishing acute and normal image cells in leukemia regarding the accuracy, specificity, and sensitivity.

5. System testing using testing data

Figure 2 exhibits the system testing process. In the testing data, the identification process was run by feature extraction with the Hu moments to obtain the 7 phi value, with the aim of the system being able to classify the types of leukemia by studying the training data and then classifying the testing data using the SVM and K-NN methods.

6. Classification of SVM and K-NN

The classification stage was the decision stage in determining the image class. This study employed two classes, normal and acute. To classify images after determining the class, the training data were inputted with the testing image data. The classification process utilized the results of the SVM and K-NN training and the value of the feature matrix from the feature extraction results.

B. Analysis

The success of the classification system using the SVM and K-NN methods was when the system could identify the type of leukemia cells appropriately. The results of the classification were based on the confusion matrix. Based on the confusion matrix obtained with 10-fold cross-validation, the performance of the system was assessed based on the values of accuracy, sensitivity, specificity, and training time of the classification system.

III. RESULTS AND DISCUSSIONS

Table 2 demonstrates the accuracy, sensitivity, specificity results ang training time of the linear type SVM.

TABLE II. THE RESULTS OF THE LINEAR TYPE SVM

Datasets	Accuracy	Sensitivity	Specificity	Time (sec)
Dataset 1	93.7	100	87.5	7.49
Dataset 2	93.7	97.5	90	5.34
Dataset 3	85	82.5	87.5	0.86
Dataset 4	95	100	90	0.89
Dataset 5	83.7	77.5	90	0.83
Dataset 6	93.7	100	87.5	6.69
Dataset 7	93.7	100	87.5	5.2
Dataset 8	75	70	80	0.86
Dataset 9	93.7	97.5	90	5.69
Dataset 10	72.5	65	80	3.41
Average	87.97	89	87	3.73

Tables 2 and 3 display the values of accuracy, sensitivity, specificity, and training time using the fine type K-NN and the linear type SVM. The results of accuracy, sensitivity, specificity, and training time using the SVM and K-NN were applied as percentages to discover the best method between the two methods.

TABLE III. THE RESULTS OF THE THE FINE TYPE K-NN

Datasets	Accuracy	Sensitivity	Specificity	Time (sec)
Dataset 1	90	90	90	5.18
Dataset 2	96.2	95	97.5	1.94
Dataset 3	78.7	75	82.5	1.15
Dataset 4	91.2	92.5	90	1.08
Dataset 5	78.7	72.5	85	1.04
Dataset 6	88.7	90	87.5	2.05
Dataset 7	86.2	82.5	90	1.17
Dataset 8	71.2	67.5	75	0.9
Dataset 9	88.2	87.5	90	1.69
Dataset 10	70	62.5	77.5	8.07
Average	83.96	81.5	86.5	2.47

The comparison of the average results from calculating the accuracy, sensitivity, and specificity values of the two methods for leukemia classification is presented in Table 4 and graphical form in Figure 4. Meanwhile, the comparison of the highest results from calculating the accuracy, sensitivity, and specificity values of the two methods for leukemia classification is listed in Table 6 and graphical results as presented in Figure 5.

TABLE IV. THE COMPARISON OF THE AVERAGES OF THE K-NN AND SVM METHODS (%)

No.	Classification Method	Accuracy	Sensitivity	Specificity
1.	K-NN	83.96	81.5	86.5
2.	SVM	87.97	89	87

The classification results (averages values) using linear type SVM and fine type K-NN were compared to determine which one had the best value between the two. From the comparison data, the average value of accuracy, the sensitivity, and specificity values were in the linear type SVM with an accuracy value of 87.97% compared to the fine type K-NN with an accuracy value of 83.96%. The sensitivity value of SVM of 89% is greater than the K-NN of 81.5%. The specificity value of SVM of 87% is higher than the K-NN of 86.5%.

In the classification process, both methods required time in reading each dataset inputted. Table 5 depicts the average value of the training time of both methods. The average training time of both SVM and KNN are presented in Table 5.

TABLE V. AVERAGE TRAINING TIME FOR SVM AND K-NN METHODS

No	Classification Method	Training Time
1.	K-Nearest Neighbor	2.43 seconds
2.	Support Vector Machine	3.73 seconds

TABLE VI. THE COMPARISON OF THE BEST PERFORMANCES OF SVM AND K-NN METHODS (%)

No	Classification Method	Accuracy	Sensitivity	Specificity
1.	K-NN	96.2	95	97.5
2.	SVM	95	100	90

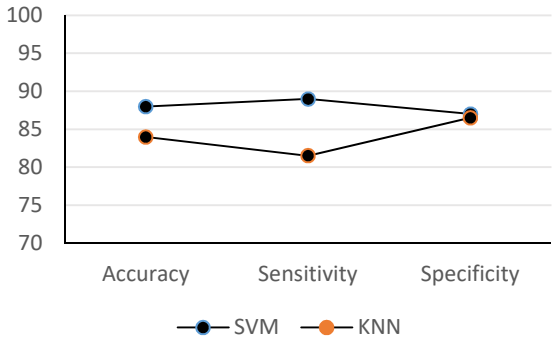

Fig. 4. Comparison graph of the average SVM and K-NN performances

The classification results (the highest values) using linear type SVM and fine type K-NN were compared to determine which one had the best value between the two. From the comparison data, the highest value of accuracy, the sensitivity, and specificity values were presented in Table 6 and Figure 5. The results of the fine type K-NN with an accuracy value of 96.2% is higher than be compared to the linear type SVM with an accuracy value of 95%. The sensitivity value of SVM of 100% is greater than the K-NN of 95%. The specificity value of K-NN of 97.5% is higher than the SVM of 90%. Figure 5 presents the comparison of the highest performances for both SVM and K-NN.

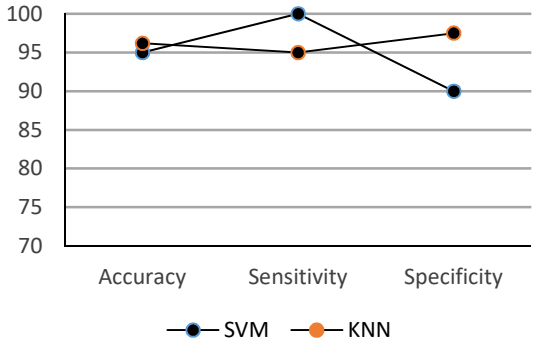

Fig. 5. Comparison graph of the highest SVM and K-NN performances

IV. CONCLUSIONS

Based on the classification system results for leukemia using feature extraction with Hu moments and two classification methods, the Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) methods, average feature extraction of 7 phi value was obtained in normal images, having a higher value than the average acute images. Classification with the SVM had an average accuracy value of 87.97% greater than the K-NN, only 83.96%. The sensitivity value of the SVM obtained a higher value of 89% than the K-NN method of only 81.5%, and the results of the specificity value of the SVM gained better results than the K-NN of only 86.5%. Regarding the required training time during the classification process, the K-NN had faster results (2.43 seconds) than SVM (3.28 seconds). In term of the highest performances, the results of the fine type K-NN with an accuracy value of 96.2% is higher than be compared to the linear type SVM with an accuracy value of 95%. The sensitivity value of SVM of 100% is greater than the K-NN of 95%. The specificity value of K-NN of 97.5% is higher than the SVM of 90%. Based on the results, both SVM and KNN classification achieved good performance to classify the leukemia images. the comparison of the two methods, particularly on the average accuracy value in the classification process on leukemia images with feature extraction of Hu moment invariant, linear type SVM was better than fine type K-NN.

ACKNOWLEDGMENT

This research is supported by Universitas Muhammadiyah Yogyakarta and a research project grant from the Ministry of Research and Technology of the Republic of Indonesia.

REFERENCES

[1] Jabar, F.H., et al., Image Segmentation using A Hybrid Clustering Technique and Mean Shift for Automated Detection Acute Leukaemia Blood Cells Images. Journal of Theoretical & Applied Information Technology, 2015. 76(1).

[2] Putzu, L., G. Caocci, and C. Di Ruberto, Leucocyte classification for leukaemia detection using image processing techniques. Artificial Intelligence in Medicine, 2014. 62(3): p. 179-191.

[3] Neoh, S.C., et al., An intelligent decision support system for leukaemia diagnosis using microscopic blood images. Scientific reports, 2015. 5: p. 14938.

[4] Viswanathan, P., Fuzzy C Means Detection of Leukemia Based on Morphological Contour Segmentation. Procedia Computer Science, 2015. 58: p. 84-90.

[5] Mishra, S., et al. Microscopic image classification using DCT for the detection of acute lymphoblastic leukemia (ALL). in Proceedings of International Conference on Computer Vision and Image Processing. 2017. Springer.

[6] Rawat, J., et al., Classification of acute lymphoblastic leukaemia using hybrid hierarchical classifiers. Multimedia Tools and Applications, 2017. 76(18): p. 19057-19085.

[7] Wang, Q., et al., Spectral-spatial feature-based neural network method for acute lymphoblastic leukemia cell identification via microscopic hyperspectral imaging technology. Biomedical optics express, 2017. 8(6): p. 3017-3028.

[8] Abdeldaim, A.M., et al., Computer-Aided Acute Lymphoblastic Leukemia Diagnosis System Based on Image Analysis, in Advances in Soft Computing and Machine Learning in Image Processing, A.E. Hassanien and D.A. Oliva, Editors. 2018, Springer International Publishing: Cham. p. 131-147.

[9] Negm, A.S., O.A. Hassan, and A.H. Kandil, A decision support system for Acute Leukaemia classification based on digital microscopic images. Alexandria Engineering Journal, 2018. 57(4): p. 2319-2332.

[10] Jusman, Y., et al. Application of Watershed Algorithm and Gray Level Co-Occurrence Matrix in Leukemia Cells Images. in 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT). 2020.

[11] Chandra, M.A. and S.S. Bedi, Survey on SVM and their application in image classification. International Journal of Information Technology, 2018.

[12] Deng, Z., et al., Efficient kNN classification algorithm for big data. Neurocomputing, 2016. 195: p. 143-148.

[13] Rawat, J., et al., Computer aided diagnostic system for detection of leukemia using microscopic images. Procedia Computer Science, 2015. 70: p. 748-756.

[14] MoradiAmin, M., et al., Computer aided detection and classification of acute lymphoblastic leukemia cell subtypes based on microscopic image analysis. Microscopy research and technique, 2016. 79(10): p. 908-916.

[15] Srisukkham, W., et al., Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization. Applied Soft Computing, 2017. 56: p. 405-419.

[16] Jagadev, P. and H. Virani. Detection of leukemia and its types using image processing and machine learning. in 2017 International Conference on Trends in Electronics and Informatics (ICEI). 2017. IEEE.

[17] Pan, L., et al., Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Scientific reports, 2017. 7(1): p. 1-9.

[18] Liu, Y., Y. Yin, and S. Zhang. Hand Gesture Recognition Based on HU Moments in Interaction of Virtual Reality. in 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics. 2012.

[19] Zhang, Y., et al., Pathological brain detection based on wavelet entropy and Hu moment invariants. Bio-Medical Materials and Engineering, 2015. 26: p. S1283-S1290.

[20] Sari, B.P. and Y. Jusman. Classification System for Cervical Cell Images based on Hu Moment Invariants Methods and Support Vector Machine. in 2021 International Conference on Intelligent Technologies (CONIT). 2021.

[21] Zhang, Y., et al., Pathological brain detection in MRI scanning via Hu moment invariants and machine learning. Journal of Experimental & Theoretical Artificial Intelligence, 2017. 29(2): p. 299-312.

[22] Zhang, Y.-D., et al., Alcoholism detection by medical robots based on Hu moment invariants and predator–prey adaptive-inertia chaotic particle swarm optimization. Computers & Electrical Engineering, 2017. 63: p. 126-138.

[23] Salleh, M.A.M., S.N.A.M. Kanafiah, and Y. Jusman. Features Extraction to Differentiate of Spinal Curvature Types using Hue Moment Algorithm. in Journal of Physics: Conference Series. 2020. IOP Publishing