2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE), 21–22 October 2022, Penang, Malaysia

Cervical Pre-cancerous Cell Image Classification System Using Histogram of Oriented Gradients and K-Nearest Neighbor Algorithms

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

[2]Maryza Intan Rahmawati
Department of Electrical Engineering
Faculty of Engineering,
Universitas Muhammadiyah
Yogyakarta
Yogyakarta, Indonesia

[3]Siti Noraini Sulaiman
Center for Electrical Engineering
Studies,
Universiti Teknologi MARA
Penang, Malaysia

2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE) | 978-1-6654-8339-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCSCE54767.2022.9935660

***Abstract*—Cervical cancer is a dangerous disease, with more than 99% of which contain Human Papillomavirus (HPV), threatening women worldwide. The Global Burden of Cancer Study (Globocan) has recorded 36,633 cervical cancer cases, ranking second in Indonesia. Analysis of Pap smear results manually as an early detection effort possesses many weaknesses. Therefore, as an early detection step in diagnosing cervical pre-cancerous cell images, an artificial intelligence system is highly required to assist medical personnel in providing fast and accurate diagnostic evaluations. This study utilized 972 training image data and 108 testing image data, with a cervical pre-cancerous cells image classification system using Histogram of Oriented Gradients (HOG) algorithms for feature extraction and KNN machine learning for the classification system. The gray level in the contrasting images between the texture of the nucleus, cytoplasm, and background had different pixel and bit depth intensity values. Hence, HOG obtained bin orientation for each pixel in the cell. The cosine KNN model demonstrated the best matrix performance, acquiring classification results of 0.8 for accuracy, 0.8 for precision, 0.889 for recall, 0.846 for specificity, and 0.771 for f-score. Moreover, the training data generated an accuracy of 69.3% and the fastest training time of 4.2359s.**

Keywords*—*cervical pre-cancerous cell images, training, testing, HOG, KNN

I. Introduction

Cervical cancer, more than 99% of which contains Human Papillomavirus (HPV), is a dangerous disease menacing women all over the world. This disease can originate from a history of pregnancy, sexual behavior, contraceptive use, smoking, nutrition, and genetics [1]. The Global Burden of Cancer Study (Globocan) from the World Health Organization (WHO) noted that the number of cancer cases in Indonesia in 2020 reached 396,914, with a total death of 234,511. It ranks second with 36,633 cases or 9.2% of the total cancer cases after breast cancer, which mostly occurs in Indonesian women [2].

Cervical cancer occurs in the female reproductive organ (cervix), the entrance to the uterus. It is located between the uterus and the female intercourse hole (vagina). The leading cause is infection with high-risk types of Human Papillomavirus (HPV), namely 16 and 18, which can be transmitted through secretions and sexual contact. One of the most significant factors for cervical cancer is having the first sexual intercourse too early. Mucous cells in the genitals are susceptible to changes, changing the normal state of other cells into an abnormal one or leading to malignancy [3].

The incidence of cervical cancer can be reduced by primary prevention efforts followed by early detection through a Pap smear or visual inspection using acetic acid (IVA). Unfortunately, the coverage of screening in Indonesia through Pap smear and IVA is still shallow (around 5%), even though this step effectively declines morbidity and mortality due to cervical cancer by 85% [4]. Analysis of Pap smear results manually requires many pathologists, and the result interpretation takes longer and tends to be objective, causing a lack of speed in handling cervical cancer. With advances in science and computers performing image processing of cervical pre-cancerous cells as an early detection step, it can help pathologists provide fast and accurate diagnostic evaluations. Thus, artificial intelligence algorithms can be implemented to classify cervical pre-cancerous cells based on digital images [5][6][7].

A study in 2019 [8] developed a texture analysis method and classification of the characteristic echo patterns of cervical nodules on ultrasound images. It performed feature extraction based on histograms, encompassing Gray Level Co-occurrence Matrix (GLCM) and gray level Run-length matrix (GLRLM), as well as classification using the multi-layer perceptron (MLP). The evaluation unveiled that using histogram features was the most appropriate extraction method, acquiring accuracy of 97.06%, sensitivity of 98.31%, specificity of 95.35%, PPV of 96.67%, NPV of 97.62%, and kappa of 0.9395.

In the following year, another research [1] entitled "Segmentation of Cervical Cancer Image Using Markov Random Field and K-Means Algorithms" revealed that pre-processing using grayscale channels and low pass filters generated the highest results, with 75.76% for the nucleus and 66.43% for the cytoplasm. Moreover, the classification results of two classes using KNN obtained the highest accuracy of 89.29%.

The previous research signifies that using the histogram features is the most appropriate extraction method. Hence, the authors developed an image processing-based application system using a Histogram of Oriented Gradients (HOG) for

978-1-6654-8339-1/22/$31.00 ©2022 IEEE

feature extraction and K-Nearest Neighbor (KNN) in Matlab R2020a for classification [9].

This system extracts texture features from the ct-scan images, where the texture of the cervical smear images is divided into three parts (nucleus, cytoplasm, and background area). The characteristics of the cells on the Pap smear, used as a distinguishing feature of texture between normal and abnormal cells, vary in color, shape, size, and texture. Using the HOG algorithms in image processing of pre-processed cervical pre-cancerous cells, the contrast between the nucleus, cytoplasm, and background can be exhibited. The gray level in the contrasting images has different pixel and bit depth intensity values. Thus, HOG obtains the bin orientation for each pixel in the cell. The process is continued with the classification after the feature value representation is attained from the extraction to detect the presence of normal and abnormal cells. It is expected to facilitate the detection to acquire faster and more accurate results.

II. Methodology

A. Data Collection

Image data obtained from cervical cancer patients at the Universiti Sains Malaysia Hospital were verified and possessed a code of ethics. The image data were cervical pre-cancerous cell images, comprising 324 High-grade Squamous Intraepithelial Lesion (HSIL) images, 324 Low-grade Squamous Intraepithelial Lesion (LSIL) images, and 324 Normal images in the training data, and 108 images in the testing data.

B. System Design

System design and testing using MATLAB R2020a with the input of cervical pre-cancerous cell images underwent several stages, such as pre-processing, feature extraction, training, and testing, as illustrated in Figure 1. MATLAB version R2020a was applied as software for processing the cervical image system design. Table I displays the computer hardware specifications.

TABLE I. Hardware Specifications

Processor	Intel® Core™ i5-9400F
RAM Memory	16384MB
GPU	Nividia RTX 2060 6GB

C. Pre-Processing

Pre-processing refers to processing the original image data to prepare them before being used for classification. The image data were classified into three classes: HSIL, LSIL, and Normal, and divided into 972 training data (90% of total image data) and 108 testing data (10% of total image data).

Then, pre-processing was carried out [10] with image enhancement to improve the quality of digital images by sharpening, followed by cropping to determine which part was used and resized to 299×299 pixels for the entire image resolution of cervical pre-cancerous cells. Then, augmentation was performed using Flip Vertical, Flip Horizontal, and Flip Horizontal-Vertical, resulting in increased image data for machine learning to obtain optimal performance [11], [12], [13].

The last pre-processing was the conversion of the pre-processed images of the previous research data into grayscale images to facilitate the calculation of the gradient images on the HOG to obtain the characteristics of cervical pre-cancerous cells, as portrayed in Table II.

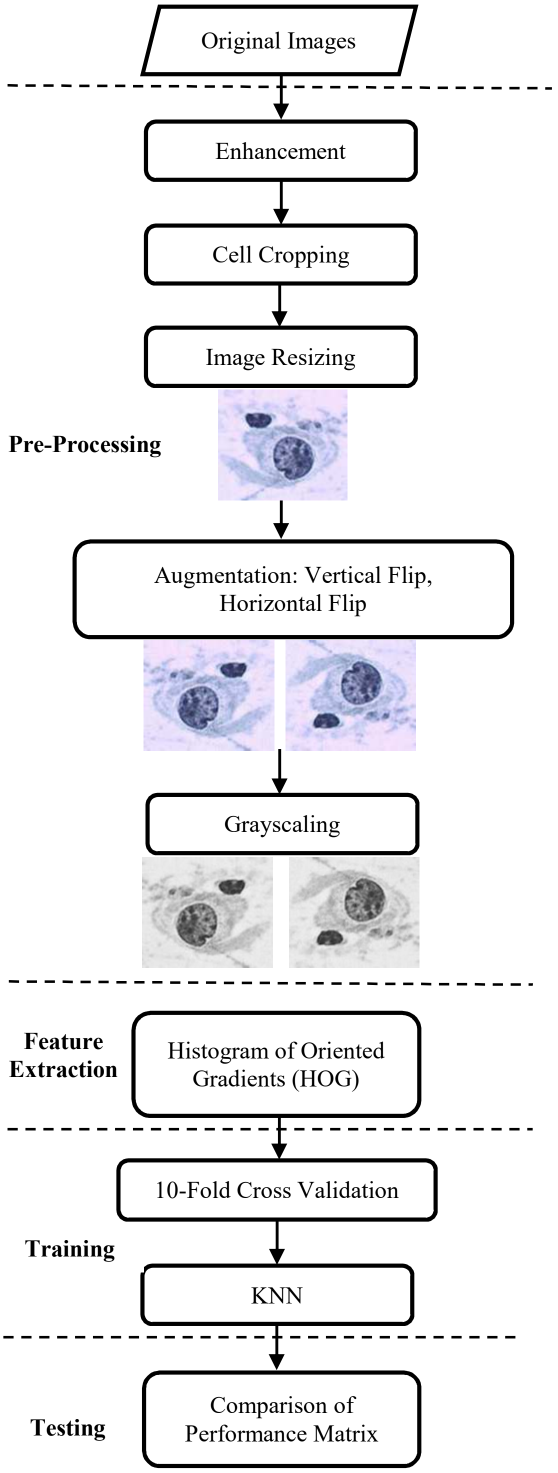
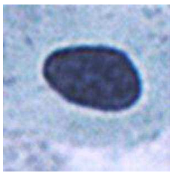
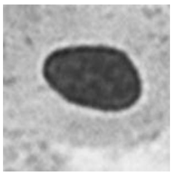
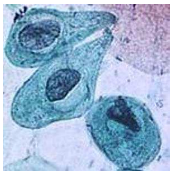
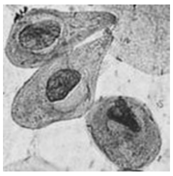
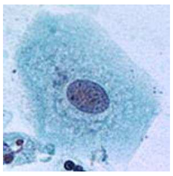
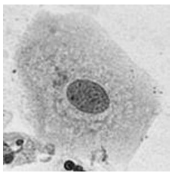


Fig. 1. Flowchart of the System Design

TABLE II. GRAYSCALE IMAGE RESULTS

Class	Pre-Processing Image	Grayscale Image
Class 1 **HSIL**		
Class 2 **LSIL**		
Class 3 **NORMAL**		

D. Feature Extraction

The feature extraction using HOG is a feature taking of a shape. The level of gray color in the contrasting images between the nucleus, cytoplasm, and background had different pixel and bit depth intensity values. Hence, HOG obtained bin orientation for each pixel in the cell. The extraction results were 576 feature values from each class to distinguish one object from another in the following stage, classification. Table III exhibits the quantitative data of the resulting feature values.

TABLE III. FEATURE EXTRACTION RESULTS

No.	Feature	Mean ± Stdv		
		Class 1 (HSIL)	Class 2 (LSIL)	Class 3 (Normal)
1.	Feature 1	0.134 ± 0.045	0.119 ± 0.045	0.131 ± 0.049
2.	Feature 2	0.116 ± 0.041	0.100 ± 0.041	0.108 ± 0.048
3.	Feature 3	0.118 ± 0.040	0.103 ± 0.041	0.109 ± 0.046
4.	Feature 4	0.114 ± 0.045	0.106 ± 0.046	0.106 ± 0.051
5.	Feature 5	0.161 ± 0.045	0.152 ± 0.051	0.161 ± 0.049
6.	Feature 6	0.128 ± 0.046	0.137 ± 0.054	0.133 ± 0.056
7.	Feature 7	0.139 ± 0.041	0.146 ± 0.046	0.151 ± 0.052
8.	Feature 8	0.134 ± 0.043	0.136 ± 0.049	0.143 ± 0.053
9.	Feature 9	0.142 ± 0.044	0.133 ± 0.048	0.144 ± 0.049
10	Feature 10	0.158 ± 0.040	0.154 ± 0.040	0.161 ± 0.042
…	…	… ± …	… ± …	… ± …
576	Feature 576	0.155 ± 0.057	0.155 ± 0.040	0.156 ± 0.040

E. Classification

After that, the extracted feature values were applied as input to the KNN models to perform classification, consisting of training data and model testing to determine the class of cervical pre-cancerous cells (HSIL, LSIL, and Normal) being diagnosed from the patients. The training was carried out using the feature values extracted from the training data in excel form as input values for classification. Training using training data of 972 cervical pre-cancerous cell images consisting of 324 HSIL images, 324 LSIL images, and 324 Normal images using the K-Fold Cross Validation method, with a value of K = 10, indicates that 90% of the data for training and 10% data for validation were determined randomly by the pre-training model. The training was performed ten times to produce ten different training data for each model. The classification results on training data were Receiver Operating Characteristics (ROC) data in three KNN models (Medium KNN, Fine KNN, Cosine KNN), accuracy values, and training time.

The training resulted in three models, which were then tested for the outcomes using the feature values of the testing data extraction results totaling 36 HSIL images, 36 LSIL images, and 36 Normal images. The classification results were displayed in a confusion matrix comparing the classification results by the system (model) and the actual classification results and the performance matrix in accuracy, precision, sensitivity (recall), specificity, and f-score.

III. RESULTS AND DISCUSSION

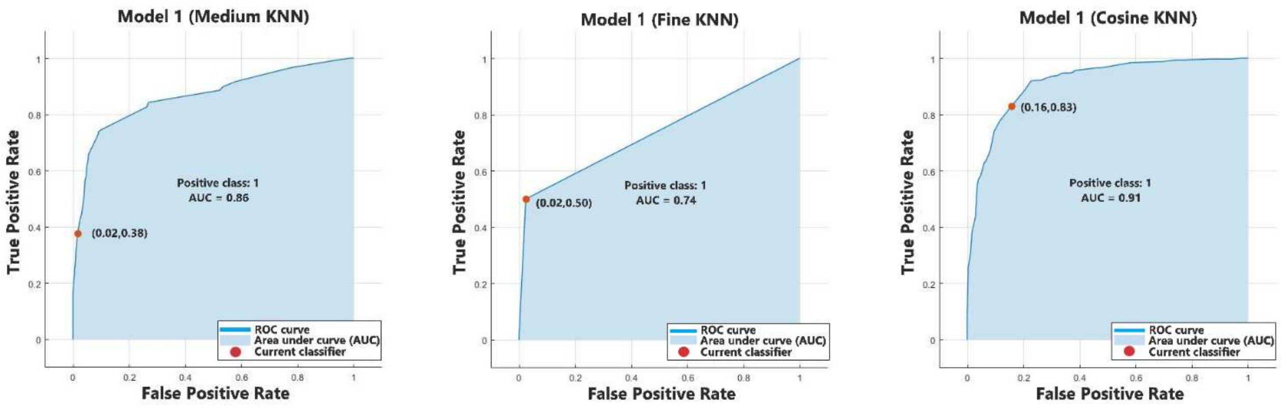
A. Training Results

Training using training data with the K-Fold Cross Validation method, with a value of K = 10, produced ten different training data on three KNN models. The results revealed the accuracy and training time for the Medium KNN, Fine KNN, and Cosine KNN models, as displayed in Table IV. The training data in Table IV describe that the Medium KNN model acquired the highest accuracy of 57%, the fastest training time of 3.8674s, and an error rate of 43%. The Fine KNN model attained the highest accuracy of 58.3%, the fastest training time of 3.8028s, and an error rate of 41.7%. Meanwhile, the Cosine KNN model obtained the highest accuracy of 69.3%, the fastest training time of 4.2359s, and an error rate of 30.7%.

TABLE IV. TRAINING RESULTS

Dataset	Medium KNN		Fine KNN		Cosine KNN	
	Acc (%)	Time (s)	Acc (%)	Time (s)	Acc (%)	Time (s)
Run 1	54.3	3.8860	57.2	3.8323	69.3	4.3091
Run 2	55.9	4.3933	56.8	4.7311	69.1	4.2844
Run 3	56.8	4.3590	57.3	3.8028	67.7	4.2881
Run 4	56.2	3.9055	57.2	3.8455	68.0	4.2784
Run 5	55.9	4.3638	57.1	4.3392	66.8	4.2359
Run 6	56.0	4.3629	58.0	4.3468	69.2	4.6788
Run 7	55.7	4.3887	58.3	4.3679	69.0	4.3038
Run 8	56.6	4.3863	57.9	3.8306	68.8	4.2953
Run 9	56.5	3.8674	57.5	4.3123	68.0	4.2922
Run 10	57.0	4.4137	57.0	3.8481	68.4	4.3110
Mean ± Stdv	**56.09 ± 0.760**	**4.232 ± 0.239**	**57.43 ± 0.485**	**4.125 ± 0.331**	**68.43 ± 0.804**	**4.327 ± 0.125**

In addition to accuracy and training time, the results of training data uncovered a ROC curve for each model with TP class = 1 (HSIL), illustrating the relationship between TPR (y-axis) and FPR (x-axis). The Medium KNN model, with the best ROC curve, depicted the lowest left point or the current classifier of 0.02.0.38, resulting in an AUC of 0.86, signifying a good classification. The Fine KNN model, with the best ROC curve, demonstrated the lowest left point or the current classifier of 0.03,0.48, yielding an AUC of 0.73, implying a fair classification. Furthermore, the Cosine KNN model, with the best ROC curve, exhibited the lowest left point or the current classifier of 0.16.0.83, producing an AUC of 0.91, close to 1. Hence, the accuracy indicates excellent classification. Figure 2 displays the best ROC curves for the three KNN models.



(a) (b) (c)

Fig. 2. Receiver Operating Characteristic (ROC) Results of Training Data, (a) Medium KNN; (b) Fine KNN; (c) Cosine KNN

B. Testing Results

After completing the training, three KNN models were tested to classify the testing data. The classification results were employed to determine the performance of three KNN models on cervical pre-cancerous cell image classification, with a comparison of the performance matrix in accuracy, precision, sensitivity (recall), specificity, and f-score, exhibited in Table V and represented in the bar chart in Figure 3.

Table V demonstrates the most incredible performance matrix during classification using the Cosine KNN model, with an accuracy of 0.8, precision of 0.8, recall of 0.889, specificity of 0.846, and f-score of 0.771.

TABLE V. PERFORMANCE MATRIX RESULTS

Model	Performance	HSIL	LSIL	NORMAL
Medium KNN	*accuracy*	0.642	0.559	0.553
	precision	1.000	0.200	0.200
	recall	0.194	0.350	0.333
	specificity	1.000	0.616	0.616
	f-score	0.326	0.255	0.250
Fine KNN	*accuracy*	0.625	0.505	0.575
	precision	1.000	0.158	0.188
	recall	0.167	0.261	0.353
	specificity	1.000	0.579	0.629
	f-score	0.286	0.197	0.245
Cosine KNN	*accuracy*	0.800	0.752	0.792
	precision	0.681	0.780	0.800
	recall	0.889	0.667	0.727
	specificity	0.746	0.830	0.846
	f-score	0.771	0.719	0.762

Then, the classification system compared the classification results by the system (model) and the actual classification results in the confusion matrix. The Cosine KNN model, as the best model, correctly classified images following their classes: 32 images into class 1 (HSIL), 20 images into class 2 (LSIL), and 24 images into class 3 (Normal). However, the system misclassified four HSIL images as LSIL, eight LSIL images into HSIL, eight images into Normal, seven Normal images as HSIL, and five images into LSIL. Figure 4 exhibits the most appropriate confusion matrix graph between the three KNN models.

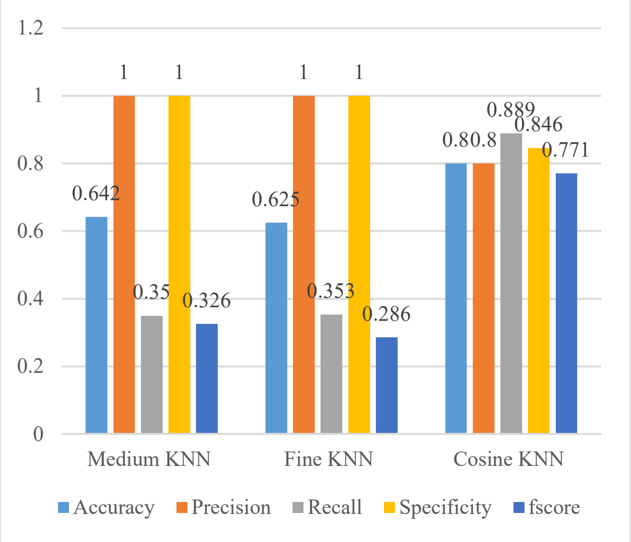


Fig. 3. The Best Testing Performance of the KNN Models

The cervical pre-cancerous cell image classification system using HOG algorithms with three KNN models revealed that the Cosine KNN model performed the best, proven by the highest training accuracy of 69.3%, the fastest training time of 4.2359s, and the lowest error rate of 30.7%. The ROC curve depicted an AUC of 0.91, indicating an excellent classification. The comparative analysis of the performance matrix during classification unveiled that the Cosine KNN model yielded the most significant performance matrix of 0.8 for accuracy, 0.8 for precision, 0.889 for recall, 0.846 for specificity, and 0.771 for f-score.

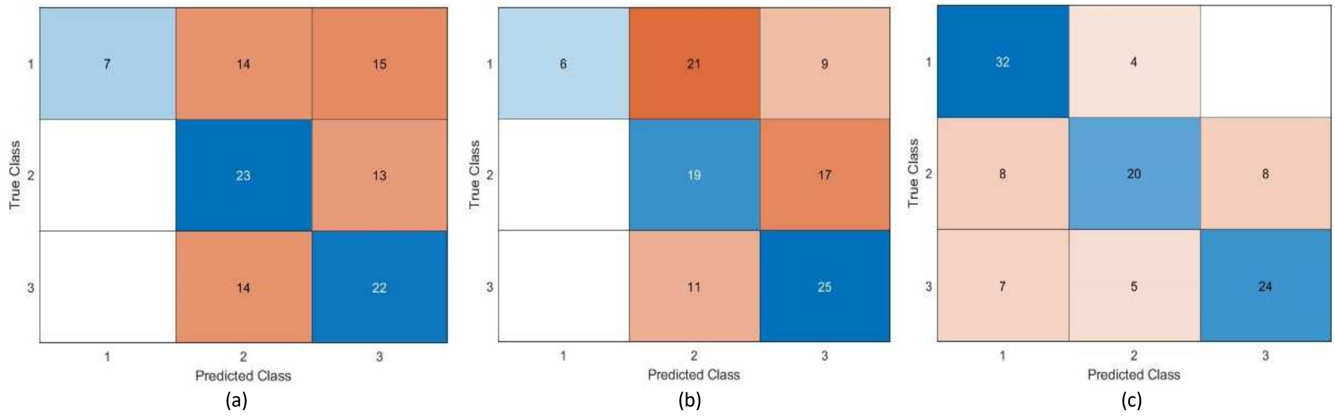


Fig. 4. Confusion Matrix (CM) Results of Training Data, (a) Medium KNN; (b) Fine KNN; (c) Cosine KNN

IV. Conclusions

In conclusion, the cervical pre-cancerous cell image classification system with HOG and machine learning algorithms with three KNN models: Medium KNN, Fine KNN, and Cosine KNN, ran well. The cosine KNN model depicted the most outstanding performance, generating classification results of 0.8 for accuracy, 0.8 for precision, 0.889 for recall, 0.846 for specificity, and 0.771 for f-score. With training data yielding an accuracy of 69.3% and the fastest training time of 4.2359s, this model fell into the category of a good system because the obtained accuracy results reached and exceeded the lower threshold of tolerable accuracy, equal to 60%. In the medical field, the error rate identifies the system's limitations. Hence, this system should be developed again to achieve a lower error rate.

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] R. S. D. Wijaya, Adiwijaya, Andriyan B Suksmono, and Tati LR Mengko, "Segmentasi Citra Kanker Serviks Menggunakan Markov Random Field dan Algoritma K-Means," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 1, pp. 139–147, 2021, doi: 10.29207/resti.v5i1.2816.

[2] The Global Cancer Observatory, "Cancer Incident in Indonesia," Int. Agency Res. Cancer, vol. 858, pp. 1–2, 2020, [Online]. Available: https://gco.iarc.fr/.

[3] T. Conceição, C. Braga, L. Rosado, and M. J. M. Vasconcelos, "A review of computational methods for cervical cells segmentation and abnormality classification," Int. J. Mol. Sci., vol. 20, no. 20, 2019, doi: 10.3390/ijms20205114.

[4] F. Fuadah, S. Rejeki, H. Triana, and ..., "Deteksi Dini Kanker Serviks Melalui Pemeriksaan IVA Test Pada Wanita Usia Subur Di Desa Babakan Kecamatan Ciparay Kab Bandung," … Kpd. Masy. …, pp. 4–5, 2020, [Online]. Available: http://journal.unjani.ac.id/index.php/unex/article/view/30.

[5] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," Futur. Gener. Comput. Syst., vol. 102, pp. 643–649, 2020, doi: 10.1016/j.future.2019.09.015.

[6] R. Gupta, A. Sarwar, and V. Sharma, "Screening of Cervical Cancer by Artificial Intelligence based Analysis of Digitized Papanicolaou-Smear Images," Int. J. Contemp. Med. Res., vol. 4, no. 5, pp. 2454–7379, 2017, [Online]. Available: www.ijcmr.com.

[7] P. Wang, L. Wang, Y. Li, Q. Song, S. Lv, and X. Hu, "Automatic cell nuclei segmentation and classification of cervical Pap smear images," Biomed. Signal Process. Control, vol. 48, pp. 93–103, 2019, doi: 10.1016/j.bspc.2018.09.008.

[8] M. Rahmawaty and Y. Triyani, "Analisis Tekstur dan Klasifikasi Karakteristik Echo Pattern Pada Citra Ultrasonografi Leher Rahim," Pros. Semin. Nas. Kesehat., pp. 223–228, 2019.

[9] T. Xu et al., "Multi-feature based benchmark for cervical dysplasia classification evaluation," Pattern Recognit., vol. 63, pp. 468–475, 2017, doi: 10.1016/j.patcog.2016.09.027.

[10] Y. Jusman, S. Riyadi, A. Faisal, S. N. A. M. Kanafiah, Z. Mohamed, and R. Hassan, "Classification System for Leukemia Cell Images based on Hu Moment Invariants and Support Vector Machines," Proc. - 2021 11th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2021, pp. 137–141, 2021

[11] Y. Jusman, B.P. Sari, & S. Riyadi, "Cervical Precancerous Classification System based on Texture Features and Support Vector Machine," Paper presented at the 2021 1st International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), pp. 29-33, 2021

[12] B. P. Sari, & Y. Jusman, "Classification System for Cervical Cell Images based on Hu Moment Invariants Methods and Support Vector Machine," Paper presented at the 2021 International Conference on Intelligent Technologies (CONIT), pp. 1-5, 2021

[13] W. Tyassari, Y. Jusman, S. Riyadi, & S. N. Sulaiman, "Classification of Cervical Precancerous Cell of ThinPrep Images Based on Deep Learning Model AlexNet and InceptionV3," Paper presented at the 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), pp. 276-281, 2022