Classification of Colposcopy Data Using GLCM-SVM on Cervical Cancer

Muhammad Thohir
Department of Arabic Language Education
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
muhammadthohir@uinsby.ac.id

Ahmad Zoebad Foeady
Department of Mathematics
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
deddyzoebad711@gmail.ac.id

Dian Candra Rini Novitasari
Department of Mathematics
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
diancrini@uinsby.ac.id

Ahmad Zaenal Arifin
Department of Mathematics
Universitas PGRI Ronggolawe
Tuban, Indonesia
az_arifin@unirow.ac.id

Bunga Yuwa Phiadelvira
Department of Islamic Education
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
yuwabunga@uinsby.ac.id

Ahmad Hanif Asyhar
Department of Mathematics
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
hanif@uinsby.ac.id

***Abstract*— cervical cancer is the second deadliest disease for women. To reduce the number of deaths caused by this disease, it is necessary that there is prevention by early detection of cancer. The method used to identify the presence of cervical cancer is to make visual observations that produce image data. However, a visual observation also has weaknesses, so it needs to be done computer-based observation to facilitate early detection. In this study, the computer-based observation method used is pre-processing, followed by a feature extraction process using the Gray Level Co-occurrence Matrix (GLCM) and Support Vector Machine (SVM) as a classification method. The best SVM classification results are using the polynomial kernel and GLCM feature extraction with an angle of 450. The accuracy rate obtained is 90%.**

***Keywords*— cervical cancer, colposcopy, GLCM, SVM**

I. INTRODUCTION

Cancer is a disease caused by the growth of abnormal and uncontrolled body cells that enter normal body tissue and then suppress normal tissue growth [1], [2]. These cells will continue to grow and can spread to other parts of the body that can cause death in humans. Cancer is one of the diseases that contribute to the highest mortality rate in the world, with a mortality rate of 7.6 million or around 13% every year [3]–[5].

According to the World Health Organization (WHO), Indonesia is a country that has the most cervical cancer sufferers in the world, with the number of cases every year not less than 150,000 cases [6]. In Indonesia, there are 5 types of cancers that are suffered by humans that cause death, and one of them is cervical cancer. Cervical cancer is a type of cancer that attacks the reproductive organs in women, namely the cervix (uterus) [7]. Basically, this cancer attacks older women, but in reality, this disease also attacks women at the age of 20 to 30 years. After breast cancer, cervical cancer is the second cancer that causes many deaths of women [8]. Cancer is influenced by several things such as gene factors, poor lifestyle and also Human Papilloma Virus (HPV).

HPV is a virus that is the main cause of cervical cancer. HPV has several types, but types 16 and 18 are the types that most influence the occurrence of cervical cancer. Transmission of this virus can occur through skin contact, which usually occurs during sexual intercourse [9]. To reduce the mortality rate caused by this disease, it is necessary to prevent and detect this cancer early.

Early detection can be in several ways, one way is to use colposcopy data [10]. From the data, the cervical identification process can be said to be normal or cancerous. The detection process is carried out by a doctor by making direct visual observations. Visual observation has several weaknesses that cause a lack of accuracy. Because of that, observation based on a computer is needed that is expected to provide results with a value that can be accounted for by the medical authorities in the early detection of cervical cancer [6]. Observation-based on a computer begins with pre-processing followed by a feature extraction process using the Gray Level Co-occurrence Matrix (GLCM) which has been successful in describing and improving image features [11]. Then the results obtained from the pre-processing process are needed input on the classification process using Support Vector Machine (SVM). The SVM method was chosen as the classification method because it was considered to have a high generalization ability [12].

In several previous research, many researchers used SVM as a method for classification such as An Intelligent System for Automated Breast Cancer Diagnosis and Prognosis Using SVM Based Classifier with an accuracy value is 97% [13], Automatic Acne Detection of Face Image Case Study on Javanese Skin with an accuracy value is 93.13% [14], MRI Brain Cancer Classification Using Hybrid Classifier (SVM-KNN) with an accuracy value is 96% [15], Automated Diagnosis System of Diabetic Retinopathy Using GLCM Method and SVM Classifier with an accuracy value is 82.35% [16].

Based on previous research, the SVM method has the accuracy and generalization value higher than other classification methods. Therefore this research uses the GLCM method and SVM classification as Computer-Aided Diagnosis (CAD) to improve early detection of cervical cancer.

II. LITERATURE REVIEW

A. Cervical Cancer

Cervical cancer is a disease that attacks the reproductive organs in women, namely in the cervix. This disease can occur when cells in the cervix grow abnormally and are out of control [17]. Figure 1 can be seen in the condition of the cervix affected by cervical cancer.

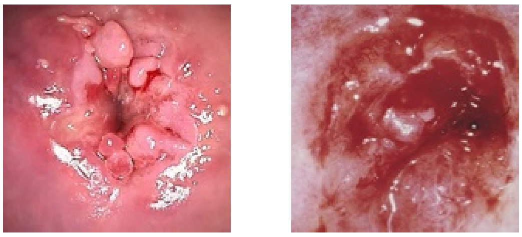


Fig. 1 Cervix affected by cancer

Cervical cancer is the second disease that is considered as the main cause of death in women after breast cancer. About 14,000 people in the United States have cervical cancer every year, and about 5,000 people are dying. Around 85% of deaths from cervical cancer occur in developing countries, and the mortality rate is 18 times more than in developed countries [18].

B. Colposcopy

Colposcopy is one of the treatment processes to determine the condition of cervical cancer. The colposcopy process begins with the administration of acetic acid to the cervix and then enlarges to see abnormalities in the cervix [19].

C. Median Filter

A median filter is one of the techniques used in pre-processing to eliminate "salt and pepper" noise in image data. The probability that each pixel has in the image is $\frac{p}{2}(0 < p <)$ 1, which is dominated by black and white points [20]. The step in the median filter is to sort the grey value in the neighbor value and then take the median value to replace the noise value in the image [21]. The median filter equation can be seen in Equation 1.

$$g(x,y) = median\{f(x-i, y-j), i,j \in W \quad (1)$$

Where $f(x,y)$ is real image data, $g(x,y)$ is the result of the median filter and W is the 2D mask in the form of an order matrix $n \times n$.

D. Histogram Equalization

Histogram equalization is also a technique used in pre-processing to manage brightness or contrast levels and improve the quality of image histograms to obtain image results that will be easier to analyze later [22]. The histogram equalization equation can be seen in Equation 2.

$$X' = T(x) = \sum_{i=0}^{x} n_i \frac{MaxIntensity}{N} \quad (2)$$

Where *X'* is a new intensity of image value. *n* is pixel data, and *N* is the amount of data.

E. Gray Level Co-Occurance Matrix (GLCM)

GLCM is an important part of the classification process that uses statistical texture calculations in the second-order to calculate the probability of a close relationship between two pixels at a certain distance. The GLCM feature is calculated through 4 different directions, namely 0^0, 45^0,90^0 and 135^0 [23]. To extract features, GLCM involves parameters as well as several properties that can be seen in equations [23].

1) Contrast

$$contrast = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(i-j)^2 M(i,j) \quad (3)$$

2) Correlation

$$Correlation = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} ij\, M(i,j) \quad (4)$$

3) Energy

$$Energy = \sqrt{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} M^2(i,j)} \quad (5)$$

4) Homogenitas

$$Homogeneity = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \frac{M(i,j)}{1+|i-j|} \quad (6)$$

In equation (3) - (6), M (i, j) is a probability with the value is start from 0 to 1, that is the value of the elements in the cohesion matrix, N is many pixels and i, j is the number of rows and columns.

F. Support Vector Machine (SVM)

In 1992 for the first time, the Support Vector Machine (SVM) was introduced by Vapnik as one of the pattern recognition methods. SVM has the concept to determine the optimal hyperplane in the form of a linear line to separate two classes that are divided into classes +1 and -1 [16]. The optimal hyperplane can be searched by measuring the closest distance between the hyperplane with the closest pattern of each class. The purpose of the SVM method itself is to get optimal vector support. The SVM equation can be seen in Equation 7.

$$f(x) = w(x) + b \quad (7)$$

Where *w* is the weight, *x* is the data as the input variable and *b* is the bias. Maximizing the margin value is a method used to get the optimal hyperplane value. The optimal hyperplane image can be seen in Figure 2.

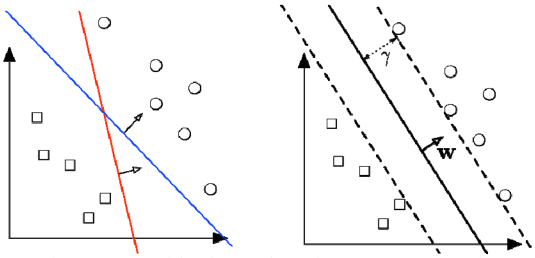


Fig. 2. SVM with a hyperplane that separates two classes [24]

G. Confusion Matrix

Confusion Matrix is a method used to measure the accuracy of a classifier that can produce accuracy, specificity and sensitivity values. To get accuracy in a classifier, we need True Negative (TN), False Positive (FP), False

Negative (FN), True Positive (TP) values on the confusion matrix. The confusion matrix can be seen in table 1.

Table 1. *Confusion Matrix*

Actual Data	Classification Result	
	+	-
+	*True Positive (TP)*	*False Negative (FN)*
-	*False Positive (FP)*	*True Negative (TN)*

After TN, FP, FN, and TP values are obtained, the accuracy, sensitivity, and specification values can be calculated using equations (8) - (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

$$Sensitivity = \frac{TN}{TN+FP} \tag{10}$$

III. RESEARCH METHODS

A. *Research Type*

The type of research on early detection of cervical cancer using Support Vector Machine is an image data and quantitative data because in the classification process using mathematical calculations.

B. *Data Collection and Analysis*

The data used in this study is colposcopy data taken from [25]. The amount of data used in this study collected 500 data which was divided into 400 training data and 100 testing data.

C. *Research Steps*

The steps taken in this research are:

1) *Pre-Processing*

Pre-processing is a process carried out to improve data and identify shapes and sizes to facilitate image processing. Many ways can in pre-processing such as changing the image into grayscale, reducing noise and also adjusting the contrast and brightness levels picture.

2) Feature Extraction

Feature extraction is a process that has the purpose of extracting information that is characteristic of each class that makes the classification process easier.

3) Classification

After getting the results of the training process that produces weight and bias values, the next step is the testing process that uses data from the output of the training process to classify data into five classes namely normal, cervical cancer stage1, stage 2, stage3, and stage 4.

For facilitate understanding, the research steps can be seen in Figure 3.

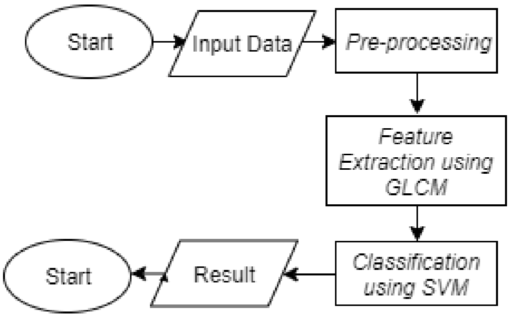


Fig. 3. Research flowchart

IV. RESULT AND DISCUSSION

This research consists of several processes, namely pre-processing, feature extraction using GLCM and classification using SVM. Data samples used for the classification process are data samples of normal cervical conditions, cervical cancer stage 1, stage2, stage3, and stage 4 so in this research, a classification was carried out using five classes. The more classes that are used, the more helpful the medical authorities will be in more detailed identification. The image obtained is still raw data that has not been processed and still has a lot of noise that must be processed using preprocessing, which will then be extracted features to get the test parameters. This parameter will be input to the classification process and affect the success of classification. Examples of data samples used can be seen in Figure 4.

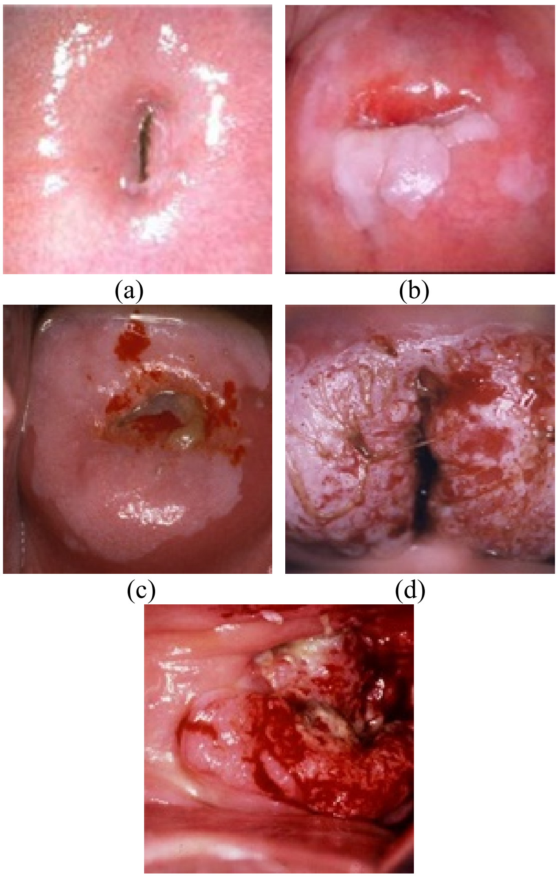
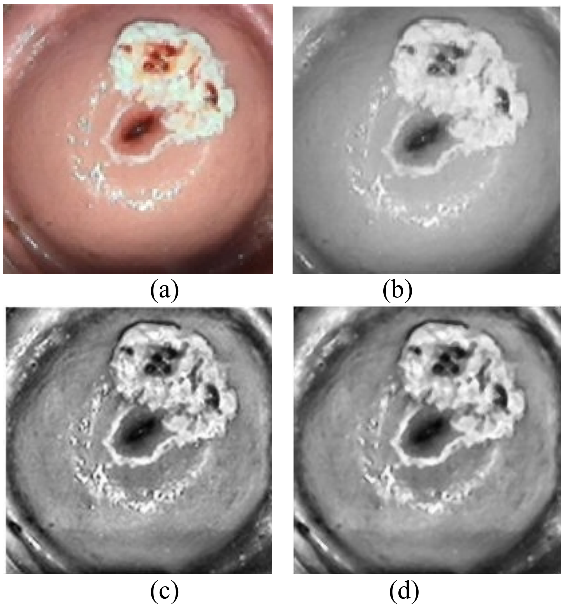

(a) (b) (c) (d) (e)

Fig 4. (a) cervix normal, (b) stage 1 cervical cancer, (c) stage 2 cervical cancer, (d) stage 3 cervical cancer, (e) stage 4 cervical cancer

The steps taken in this research are:

A. *Pre-processing*

Pre-processing is the first step carried out to get a new and good image value to be used at a later stage. Pre-processing begins with changing the initial image data into grayscale because the method to be used in feature extraction is GLCM, then the image must be changed first to grayscale, the next step is to adjust the brightness and contrast levels in the image using histogram equalization. This aims to equalize the level of brightness and contrast in all existing image data. The next step taken is to reduce noise or unwanted objects using a median filter, and the size of the matrix used as a 2D mask value in this study is 3 x 3. The use of a matrix size 3 x 3 as a 2D mask is because if the size of the matrix taken is too large, then a lot of textures on the features will be wasted. The results of pre-processing can be seen in Figure 5.a as a grayscale result, 5. b as a result of histogram equalization and 5.c as a result of the median filter.


(a) (b)

(c) (d)

Fig 5. (a) Image of cervical cancer (b) Grayscale image (c) Extraction of histogram equalization (d) Extraction of a median filter

B. *Feature Extraction*

The feature extraction process aims to recognize the characteristics of each image that will be used in the classification process. Each picture must have its own characteristics, then the extraction of feature features on each image will be recognized. In this research, the feature extraction that is used is GLCM by using data from the results of previous preprocessing. The results of feature extraction in the sample data can be seen in Table 2.

In Table 2, which contains the results of feature extraction cannot be identified directly because each parameter has almost the same value. Based on these problems, we need machine learning to help in pattern recognition. The machine learning used is SVM. The SVM method will help share data and identify cancer automatically.

Table 2. The result of feature extraction using GLCM

Feature Extraction				Class
Contrast	Correlation	Energy	Homogeneity	
0.137749	0.944056	0.196599	0.934016	Normal
0.216131	0.899401	0.230481	0.922629	Normal
0.216131	0.899401	0.230481	0.922629	Normal
0.065428	0.965698	0.270725	0.968411	Stage 1
0.103204	0.940013	0.254888	0.951572	Stage 1
0.094897	0.971223	0.190179	0.955693	Stage 1
0.179844	0.924411	0.18834	0.924589	Stage 2
0.13886	0.958423	0.194506	0.933311	Stage 2
0.13886	0.958423	0.194506	0.933311	Stage 2
0.152446	0.966701	0.165549	0.933085	Stage 3
0.162269	0.925382	0.197353	0.923683	Stage 3
0.162269	0.925382	0.197353	0.923683	Stage 3
0.114669	0.88449	0.536621	0.953897	Stage 4
0.252721	0.893642	0.210037	0.910637	Stage 4
0.252721	0.893642	0.210037	0.910637	Stage 4

C. *Classification using SVM*

Feature data from the extraction process using the GLCM method consisting of contrast, correlation, energy, and homogeneity will be used as input data in the classification process using SVM. The results of the classification process will be validated using a confusion matrix to get the percentage values of accuracy (Ac), sensitivity (Sn), and Specificity (Sp). In this research, the classification process uses four directions in GLCM using four kernels in SVM. The results of the classification can be seen in table 3.

Table 3. Classification results

Kernels	**Degrees**			
		Ac	**Sn**	**Sp**
Linier	0^0	57.00 %	62.40 %	57.00 %
	45^0	64.00 %	64.79 %	64.00 %
	90^0	64.00 %	65.48 %	64.00 %
	135^0	61.00 %	61.86 %	61.00 %
Polynomial	0^0	82.00 %	82.26 %	82.00 %
	45^0	90.00 %	89.94 %	90.00 %
	90^0	87.00 %	87.23 %	87.00 %
	135^0	88.00 %	88.57 %	88.00 %
Gaussian	0^0	83.00 %	84.31 %	83.00 %
	45^0	84.00 %	84.16 %	84.00 %
	90^0	83.00 %	84.77 %	83.00 %
	135^0	82.00 %	84.60 %	82.00 %

From the testing data result in table 2, it can be seen that the testing using three kernels in SVM are linear, polynomial, and Gaussian. The accuracy results generated in linear kernels mostly is 61.5%. This means that the linear kernel is not good at identifying cervical cancer. The poor accuracy results are caused by the result data from GLCM that cannot be separated properly with a hyperplane. The best results are obtained in testing using the polynomial kernel. This is because the GLCM data is centralized, so when the polynomial kernel function is applied, it can be separated properly. The best results on the polynomial kernel are obtained from generating data on GLCM with a degree of 45°. This determines that the data is more easily distinguished when observing the neighbouring co-occurrence matrix towards the diagonally on the right. This accuracy can be a reference for medical authorities in detecting cervical cancer if they use the GLCM and SVM methods. Each testing will produce a confusion matrix.

Confusion matrix results in GLCM data of 45 degrees and using polynomial kernels in SVM can be seen in Table 4.

Based on Table 3, it is known that the accuracy of each kernel is calculated in 4 directions and the results of the classification that has the greatest accuracy are the SVM classification using a polynomial kernel with a GLCM of 45^0. Confusion matrix values that have the greatest accuracy are presented in table 3 with rows as actual data and columns as classification data results.

In the first column, it can be seen that as many as 16 data included in the normal group are correctly predicted. Then 2 data should be included in the stage 1 group and 1 data that should be included in the stage 3 group but predicted to be included in the normal group. In the second column as much as 18 predicted data correctly entered in the stage 2 group, then 1 data should be included in the normal group and 1 data that should be included in the stage 2 group predicted in the stage 1 group.

Table 4. Confusion Matrix with the best SVM classification results

Actual data	**Prediction results**				
	Normal	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
Normal	16	1	3	0	0
Stage 1	2	18	0	0	0
Stage 2	0	1	17	2	0
Stage 3	1	0	0	19	0
Stage 4	0	0	0	0	20

In the third column, as many as 17 predicted correct data entered in the stage 2 group, then 3 data should be included in the normal group predicted entered in the stage 2 group. The results in the third column are quite good, although misdiagnosed but not endangering the patient. In the fourth column, 19 predicted true data entered in stage 3 and 2 data that should be included in stage 2 were predicted to enter stage 3 category. In the fourth table, 20 predicted data correctly entered in stage 4. In the first column, there was 1 data stage 3 is detected normally, and also two stages 1 data is detected normally. This is more dangerous because it results in patients not knowing the actual condition that they have cervical cancer.

Some research says that there are other methods that can be used for feature extraction processes, namely the GLRLM method. In this research, the result of feature extraction using GLRLM has an accuracy rate of 99.5% and GLCM, which has an accuracy rate of 97.75%. It means that GLRM has a higher accuracy value than using GLCM [19]. So, it is hoped that the next research on cervical cancer detection can use GLRLM as feature extraction.

V. Conclusion

The results of research on early identification of cervical cancer using GLCM and SVM methods are parameters of accuracy, sensitivity, and specifications. This result is expected to be applied to an application using the same methods to facilitate the medical side to recognize the normal cervix, although to determine the stage of cancer. This research produces the best accuracy of 90% obtained through GLCM of 45o and SVM using the polynomial kernel. Based on the best accuracy results, it will be able to cooperate with the medical parties so that it can be applied to help early identification of cancer efficiently without doubting the results

REFERENCES

[1] X. Meng, J. Zhong, S. Liu, M. Murray, and A. M. Gonzalez-Angulo, “A new hypothesis for the cancer mechanism,” *Cancer Metastasis Rev.*, vol. 31, no. 1–2, pp. 247–268, 2012.

[2] A. Sudhakar, “History of Cancer, Ancient and Modern Treatment Methods,” *J. Cancer Sci. Ther.*, vol. 01, no. 02, pp. i–iv, 2009.

[3] N. N. Abdullah, W. Al-Kubaisy, and M. M. Mokhtar, “Health Behaviour Regarding Cervical Cancer Screening Among Urban Women in Malaysia,” *Procedia - Soc. Behav. Sci.*, vol. 85, pp. 110–117, 2013.

[4] M. J. Thun, J. O. DeLancey, M. M. Center, A. Jemal, and E. M. Ward, “The global burden of cancer: Priorities for prevention,” *Carcinogenesis*, vol. 31, no. 1, pp. 100–110, 2009.

[5] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics. CA Cancer J Clin. 2011;,” *Ca Cancer J Clin*, vol. 61, no. 1, pp. 69–90, 2011.

[6] D. A. Dharmawan, “Deteksi Kanker Serviks Otomatis Berbasis Jaringan Saraf Tiruan LVQ dan DCT,” vol. 03, no. 04, pp. 3–6.

[7] J. Giftson Senapathy, P. Umadevi, and P. S. Kannika, “The present scenario of cervical cancer control and HPV epidemiology in India: An outline,” *Asian Pacific J. Cancer Prev.*, vol. 12, no. 5, pp. 1107–1115, 2011.

[8] P. Elayaraja and M. Suganthi, “Automatic approach for cervical cancer detection and segmentation using neural network classifier,” *Asian Pacific J. Cancer Prev.*, vol. 19, no. 12, pp. 3571–3580, 2018.

[9] and G. Jennifer, Lisa, Brooke, Jessica, Silvia, Rachel, “Human Papillomavirus Type Distribution In Invasive Cervical Cancer and High-Grade Cervical Lesions,” *Int. J. Cancer*, pp. 621–632, 2007.

[10] H. I. Hyacinth, O. A. Adekeye, J. N. Ibeh, and T. Osoba, “Cervical Cancer and Pap Smear Awareness and Utilization of Pap Smear Test among Federal Civil Servants in North Central Nigeria,” *PLoS One*, vol. 7, no. 10, pp. 1–8, 2012.

[11] F. Roberti de Siqueira, W. Robson Schwartz, and H. Pedrini, "Multi-scale grey level co-occurrence matrices for texture description," *Neurocomputing*, vol. 120, pp. 336–345, 2013.

[12] M. Behzad, K. Asghari, M. Eazi, and M. Palhang, "Generalization performance of support vector machines and neural networks in runoff modelling," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7624–7629, 2009.

[13] I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, “An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers,” *Appl. Intell.*, vol. 30, no. 1, pp. 24–36, 2009.

[14] Yanuangga and L. Zaman, “Deteksi Jerawat Otomatis Pada Citra Wajah Studi Kasus Pada Kulit Penduduk Jawa,” *Semin. Nas. “Inovasi dalam Desain dan Teknol.*, pp. 400–408, 2015.

[15] K. Machhale, H. B. Nandpuru, V. Kapur, and L. Kosta, “MRI brain cancer classification using hybrid classifier (SVM-KNN),” *2015 Int. Conf. Ind. Instrum. Control. ICIC 2015*, no. Icic, pp. 60–65, 2015.

[16] A. Z. Foeady, D. C. R. Novitasari, A. H. Asyhar, and M. Firmansjah, “Automated Diagnosis System of Diabetic Retinopathy

Using GLCM Method and SVM Classifier," *2018 5th Int. Conf. Electr. Eng. Comput. Sci. Informatics*, pp. 154–160, 2019.

[17] G. A. Mishra, S. A. Pimple, and S. S. Shastri, "An overview of prevention and early detection of cervical cancers," *Indian J. Med. Paediatr. Oncol.*, vol. 32, no. 3, pp. 125–132, 2011.

[18] W. Small *et al.*, "Cervical cancer: A global health crisis," *Cancer*, vol. 123, no. 13, pp. 2404–2412, 2017.

[19] K. Bhattachan *et al.*, "Evaluation of Abnormal Cervix with Visual Inspection under Acetic Acid and Colposcopy," *J. Nepal Health Res. Counc.*, vol. 17, no. 1, pp. 76–79, 2019.

[20] N. Gopinath, "Graph Based Image Segmentation Method for Identification of Cancer in Prostate MRI Image," *J. Comput. Appl.*, vol. IV, no. 4, pp. 104–108, 2011.

[21] Y. Zhu and C. Huang, "An Improved Median Filtering Algorithm for Image Noise Reduction," *Phys. Procedia*, vol. 25, pp. 609–616, 2012.

[22] R. Dahiya, "Histogram Equalization Based Image Enhancement Techniques For Brightness Preservation And Contrast Enhancement," *Int. J. Adv. Res. Educ. Technol.*, vol. 2, no. 2, 2015.

[23] D. Gadkari, "Image Quality Analysis Using GLCM," University of Pune, 2004.

[24] J. Nayak, B. Naik, and H. S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges," *Int. J. Database Theory Appl.*, vol. 8, no. 1, pp. 169–186, 2015.

[25] Guneva Foundation, "Geneva Foundation for Medical Education and Research," 2018. [Online]. Available: www.gfmer.ch. [Accessed: 20-Sep-2019].