Effect of Using GLCM and LBP+HOG Feature Extraction on SVM Method in Classification of Human Skin Disease Type

Friska Eka Khoirunisa[1], Novrido Charibaldi[2]

[1,2]Informatics Department, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia
[1]friska744@gmail.com
[2] novrido@upnyk.ac.id (*)

Abstract— Skin diseases have been ranked third out of ten diseases suffered by outpatients in many hospitals in Indonesia. The public often underestimates these diseases because they are considered not to cause death. In general, dermatologists diagnose skin diseases using the biopsy process, but the biopsy process is quite expensive and can cause injury to the skin. Each skin disease has different texture and shape characteristics, so classification can be used to distinguish the type of skin disease. This study compares LBP+HOG and GLCM feature extraction with the SVM classification method to determine the best feature extraction in skin disease classification. This research compares GLCM and LBP+HOG feature extraction using the SVM method. GLCM is used to measure the relationship between pixel intensities in the image, while LBP+HOG combines information about the texture and shape of the image. The test results show that the GLCM feature extraction method with SVM classification achieves an accuracy of 74%. In this test, the parameters C=100, and the features used are homogeneity, contrast, energy, correlation, ASM, and dissimilarity. Meanwhile, extraction with LBP+HOG resulted in an accuracy of 68%.

Keywords— Human Skin Disease; LBP; HOG; GLCM; SVM.

I. INTRODUCTION

Skin diseases have been ranked third out of ten diseases suffered by outpatients in many hospitals in Indonesia [1]. The public often underestimates this disease because it is considered not to cause death. Public awareness of skin diseases is still low, mainly due to a lack of understanding of them and how to properly treat them. Skin diseases that are not treated properly can cause disability with expensive treatment costs and can even be fatal [2].

Dermatologists generally diagnose skin diseases using a biopsy process, where skin tissue is analyzed in a laboratory. The biopsy process is expensive and can cause injury to the human skin. The first step in treating skin diseases is to identify the type of skin disease so that appropriate treatment can be provided. Each skin disease has different characteristics of texture and shape, so classification can be used to distinguish the type of skin disease. With today's technological advancements, skin disease detection can be done by utilizing digital image processing [1].

Many studies have been conducted to detect skin diseases using image processing. For example, research conducted by [3] Using Gray Level Co-occurrence Matrix (GLCM) and Backpropagation methods to identify acne types based on texture, using 120 images for training data and 18 images for test data. The research used 4 GLCM features: contrast, correlation, energy, and homogeneity. The study resulted in an accuracy of 56.67%. However, the shape of the acne pattern has similarities, so the system has difficulty identifying objects.

In addition, research was carried out by [4] Using Gray Level Co-occurrence Matrix (GLCM) as a texture extraction and K-Nearest Neighbour (KNN) as a classification method to detect melasma skin disease on the face. The study used 20 facial image data divided into 16 images for training and four for test data. The study succeeded in achieving an accuracy of 98%. However, when using image data, one must pay attention to stable lighting to be appropriately processed.

Then, research was conducted [5] on skin disease classification by comparing machine learning algorithms using colour and texture features. The study used 157 skin disease image data with three classes. The Support Vector Machine (SVM) classification method produces 80% accuracy, but processing time increases compared to using the LDA, Artificial Neural Networks (ANN), and Naïve Bayes methods.

As has been done in previous studies, the complexity of skin diseases and their variety of textures and shapes requires an approach to feature extraction and the right classification method. Therefore, comparing GLCM and LBP+HOG feature extraction using the SVM method in classifying human skin disease types by utilizing the texture and structure information from each feature extraction can provide deeper insight into which one is more effective in skin disease classification.

II. RESEARCH METHODOLOGY

The research method used in this research is quantitative. The type of research is implementation research, which consists of design and development. This research designs a skin classification system that is made using the Python language and implemented into the application by comparing the feature extraction of Grey Level Co-Occurrence Matrix (GLCM) and Histogram of Oriented Gradients (HOG) + Local Binary Pattern (LBP) using the Support Vector Machine (SVM) methods, and the research stages in Figure 1.
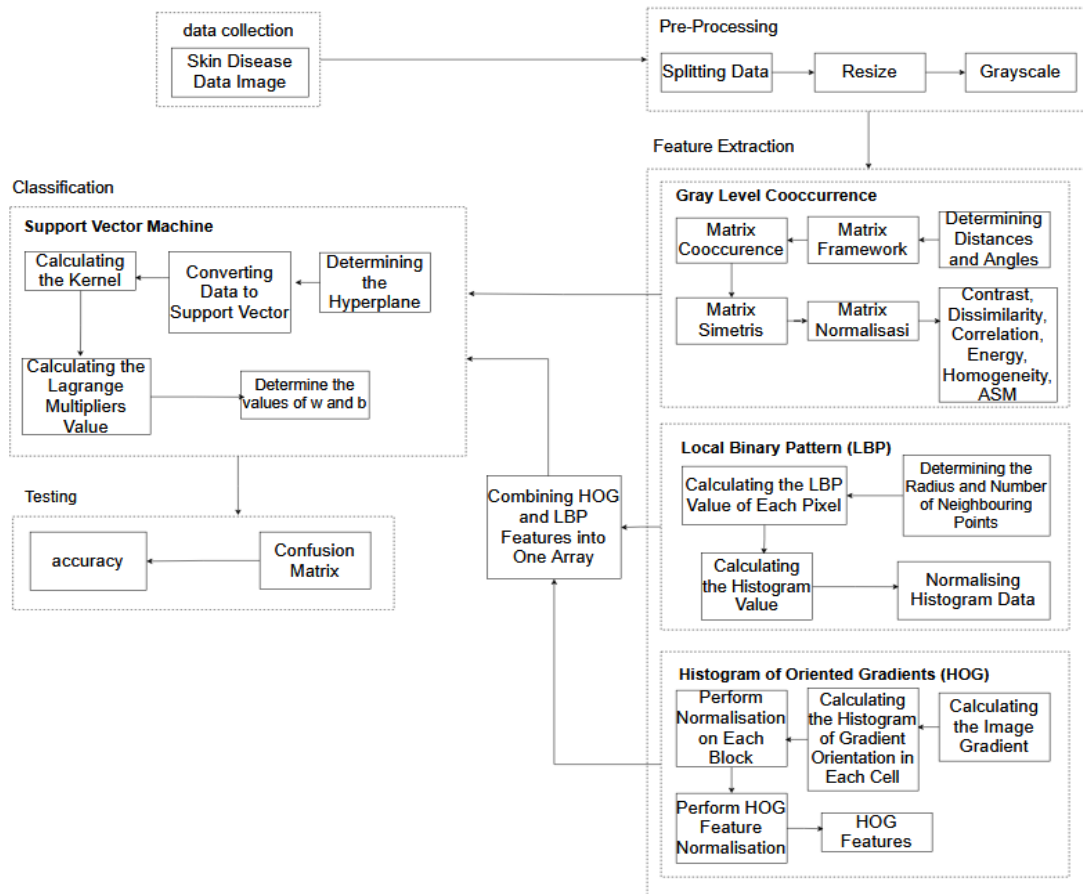
Figure 1. Research Methodology

A. Data Collection

Several skin diseases in humans include Eczema, Keratosis, and Melanocytic Nevi. The factors that cause the appearance of these skin diseases are viruses, fungi, and bacteria that grow in dirty environments, exposure to UV rays, extreme climate change, and allergies to certain things [6].

This research uses secondary data sourced from the Kaggle website. In this dataset, 1,050 skin disease image data are divided into three classes, namely 350 eczema image data, 350 keratosis image data, and 350 melanocytic nevi image data. Figure 2 shows the skin disease datasets used: Eczema, Keratosis, and Melanocytic nevi.

Figure 2. Skin Disease Dataset

B. Data Preprocessing

The preprocessing process is carried out to prepare raw data, aiming to improve the original image's quality. The data can be processed during the feature extraction and classification process. The preprocessing stage in this study performs data splitting, which aims to separate data according to its usefulness so that the entire dataset collected can meet the needs of each process that will be run. The data composition used is 20% for testing data and 80% for training data. Then, changing the entire image size to uniform the size of each data by changing the image size to 224x224 pixels and finally changing the RGB image in the input image to grayscale aims to store in 8bit format for each pixel sample, reduce the colour characteristics of existing data, and reduce computational complexity [7].

C. LBP+HOG Feature Extraction

Histogram of Oriented Gradient (HOG) is a method in image processing by taking the edge of the gradient intensity distribution of an image to detect an object [8]. Meanwhile, Local Binary Pattern (LBP) is a method to describe the local texture of an image [9]. By combining LBP and HOG feature extraction, information about the texture and shape of the image can be obtained more thoroughly. Figure 4 shows the feature extraction process using a combination of LBP and HOG. The LBP+HOG feature extraction process is in Figure 3.

The LBP calculation process involves thresholding the grey value of each pixel. This thresholding is done with the centre of the matrix as the centre. Pixels with values equal to or greater

than the centre point are given a value of 1, while those whose values are less than the centre point are given a value of 0, the LBP computation Equation (7) [9].
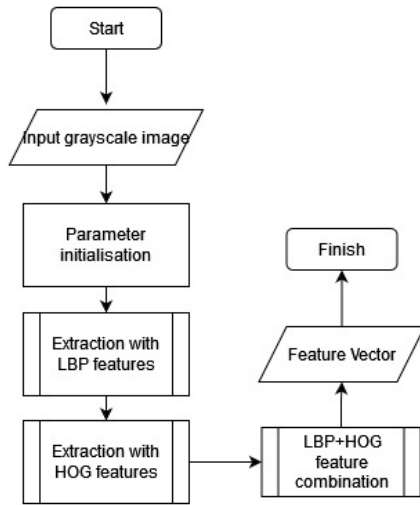

Figure 3. LBP+HOG Flowchart

$$LBP_{P,R}(X_C, Y_C) = \sum_{P=0}^{P-1} S(g_p - g_c) 2^p \qquad (7)$$

After obtaining the LBP value in each neighborhood for the NxM image size, the image's texture is represented by forming a histogram [10].

Next, the Histogram of Oriented Gradient (HOG) process calculates the histogram of the gradient orientation of each cell. Each cell pixel's histogram is based on the value generated in the gradient calculation, which is then normalized in each block. Filtering the horizontal and vertical kernels calculates an image's horizontal and vertical gradients. Then, the amount of image gradient is calculated using the Equation (8) [7].

$$G = \sqrt{g_x^2 + g_y^2} \qquad (8)$$

Then, the gradient is transformed into axis coordinates with angles 0^0 to 180^0. The calculation of the gradient orientation (θ) can be calculated with Equation (9), where g_x and g_y are the horizontal, and vertical gradients.

$$\theta| = tan^{-1} \frac{g_x}{g_y} \qquad (9)$$

The next step is to calculate the histogram of the gradient orientation of each cell. Each cell pixel has its histogram value based on the value generated in the gradient calculation, which is then normalized in each block. After that, the orientation bin divides the image into several smaller regions called cells. Each cell will form an orientation histogram. The orientation histogram divides the angles by a fixed number in the specified bins. By default, the histogram is distributed by nine bins with a multiple of 20^0 in each bin.

Then, normalize the block features. The block features are normalized so that the values in the histogram are more balanced and less affected by small changes in lighting conditions or contrast. This can help to obtain more stable features [11]. The results of the LBP and HOG histograms are combined into one feature vector, which is later used for classification.

D. GLCM Extraction

Feature extraction using GLCM with grey-comatrix and greycoprops() libraries. Retrieval of texture features using homogeneity, contrast, correlation, energy, dissimilarity, and ASM with a distance of 1,2,3 and angles $0^0, 45^0, 90^0$, and 145^0. The research will use GLCM feature extraction to get homogeneity, contrast, correlation, energy, dissimilarity, and ASM, calculating GLCM features using Equations (1) to (6).

$$Homogeneity = \sum_{i=0} \sum_j \frac{p(i,j)}{1+|(i-j)|} \qquad (1)$$

$$Contrast = \sum_{i,j=0}^{N-1} p_{i,j}(i-j)^2 \qquad (2)$$

$$Correlation = \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{(i-\mu_i')(j-\mu_j')(GLCM(I,J))}{(\sigma_i' \sigma_j')} \qquad (3)$$

$$Energy = \sum_i \sum_j p(i,j)^2 \qquad (4)$$

$$Dissimilarity = \sum_{i,j=0}^{levels-1} P_{i,j}|i-j| \qquad (5)$$

$$ASM = \sum_{i,j=0}^{levels-1} p_{i,j}^2 \qquad (6)$$

E. Support Vector Machine (SVM)

The SVM algorithm is a binary classification model that produces an optimal separation between two data classes: a straight line in two dimensions. For high-dimensional data, an ideal decision plan is required as a reference for segmentation. The principle of SVM emphasizes solving classification problems. The maximum distance between the nearest sample point and the decision surface should be achieved, and the distance between the two classes of sample points should be enlarged to effectively separate the samples [12]. In SVM, only a selected amount of data is used in classification, and SVM only stores a small part of the training data for prediction. This is an advantage of SVM because not all training data is involved during training [13]. SVM algorithms are constantly improving and can now solve binary classification and classification with more than two classes. Two approaches to solving the SVM multiclass problem exist building and combining several binary classifiers or directly considering all data in one optimal formation. SVM can classify multiple classes by utilizing one against one and one against the rest [14].

SVM is a learning machine that focuses on the Structural Risk Minimisation (SRM) principle, which is used to find the best hyperplane to separate two classes in the input space. A hyperplane is a linear line that separates two classes. In the case of multiclass classification, the hyperplane can take the form of a higher dimensional separating surface. The goal of the

hyperplane is to maximize the margin, which is the distance between the hyperplane and the closest data points of the two classes. The closest point is called a support vector machine (SVM) using Equation (10) [15].

$$f(x) = x.w + b \tag{10}$$

The margin must be maximized to obtain good learning results[16]. Data that is included in class +1 is data that fulfills Equation (11).

$$x_i.w + b \geq +1 \ for \ y_i = +1 \tag{11}$$

The data that is included in class -1 is data that fulfills Equation (12). The *W* variable is the field normal, and b is the field position relative to the coordinate centre.

$$x_i.w + b \leq +1 \ for \ y_i = -1 \tag{12}$$

Linear data in SVM cannot be found in a solution. The kernel trick must assist the Equation in this formula. Kernel trick is a method used to map low-dimensional data and transform it into a higher-dimensional space. The goal is to facilitate data classification by finding a hyperplane that can separate the dataset linearly well [16].

F. Multiclass Support Vector Machine

Multiclass SVM is a modification of SVM to handle situations with more than two classes. SVM was initially designed for binary classification problems. However, the SVM algorithm is continuously improving. Two approaches to solving the multiclass SVM problem exist building and combining multiple binary classifiers and directly considering all the data in one optimization formulation. SVM can classify multiple classes by utilizing one-against-one and one-against-rest approaches [14]

G. Confusion Matrix

Confusion matrix is a method used to measure the performance of classification methods where the content of the confusion matrix is information that compares the classification results by the system with what should be. The confusion matrix has four terms for the representation of classification results when measuring performance, namely Total True Positive (TTP), Total True Negative (TTN), Total False Positive (TFP), and Total False Negative (TFN). The confusion matrix table used in this study is in Table I.

TABLE I
CONFUSION MATRIX

Classes		Predicted Class		
		A	B	C
Actual Class	A	TN	FP	TN
	B	FN	TP	FN
	C	TN	FP	TN

The true positive value is the number of positive samples that are correctly classified by the system as positive, false negative is the number of positive samples that are misclassified by the system as negative, true negative is the number of negative samples that are correctly classified by the system as negative, false positive is the number of negative samples that are miss-classified by the system as positive. Based on the TTP, TTN, TFP, and TFN values, the accuracy, precision, recall, and f1-score values can be obtained [17].

III. RESULT AND DISCUSSION

This section describes the implementation of human skin disease classification based on the methodology listed in Figure 1. This research uses secondary data from the Kaggle site, which has the title "Skin Diseases Dataset" created by Rikimartua and contains skin disease data in JPG format. The datasets used are 1,050 skin disease images divided into three classes: Eczema 350 image data. Keratosis consists of 350 image data and Melanocytic Nevi 350 image data. Then, the data is preprocessed by dividing it into 20% training data taken randomly, and 80% is used as testing data. After division, the image is converted into grey and resized to 224x224 pixels.

Feature extraction using Gray Level Co-occurrence Matrix (GLCM) is performed to extract texture features including contrast, dissimilarity, correlation, energy, homogeneity, and ASM (Angular Second Moment) using a distance of 1,2,3 pixels and angles 0^0, 45^0, 90^0, and 135^0. The GLCM feature extraction process starts with building the co-occurrence matrix of the initial image. Next, its transpose matrix is calculated and normalized by dividing each element by the total number of elements. The following GLCM feature values of each skin disease class with a distance of 1. The GLCM calculation results in this study can be seen in Table II and Table IV.

TABLE II
GLCM FEATURE VALUE OF ECZEMA CLASS

	contrast	Dissimilarity	Correlation	Energy	Homogeneity	ASM
0^0	24,86	3,35	0,99	0,03	0,31	0,00
45^0	39,44	4,34	0,98	0,02	0,23	0,00
90^0	15,36	2,76	0,99	0,03	0,33	0,00
135^0	33,96	4,09	0,99	0,03	0,24	0,00

TABLE III
GLCM FEATURE VALUE OF KERATOSIS CLASS

	contrast	Dissimilarity	Correlation	Energy	Homogeneity	ASM
0^0	20,61	2,19	1,00	0,15	0,56	0,02
45^0	40,09	3,05	1,00	0,15	0,50	0,02
90^0	20,53	2,22	1,00	0,16	0,56	0,03
135^0	39,29	3,04	1,00	0,15	0,50	0,02

TABLE IV
GLCM FEATURE VALUE OF MELANOCYTIC NAVI CLASS

	contrast	Dissimilarity	Correlation	Energy	Homogeneity	ASM
0^0	93,57	2,25	0,99	0,06	0,36	0,00

	contrast	Dissimilarity	Correlation	Energy	Homogeneity	ASM
45^0	11,30	2,49	0,99	0,05	0,34	0,00
90^0	4,84	1,62	1,00	0,06	0,45	0,00
135^0	11,94	2,55	0,99	0,05	0,33	0,00

In Table 2 to Table 4, the results of GLCM feature extraction from skin disease images of Eczema, Keratosis, and melanocytic nevi classes are displayed using six GLCM features and a distance of 1,2,3 pixels. The number of features produced is 72 features. After obtaining values from GLCM feature extraction, these values are used in the SVM method modelling.

In the extraction process using LBP and HOG, the extracted features are combined into one feature vector, which is then added to the feature list. The results of these features are used to train the image classification model with the SVM method. The LBP feature extraction process uses a value of 3 to determine the distance between neighbouring points and the centre and a value of 24 to determine the number of neighbours to be used in the LBP calculation. After that, the LBP histogram is calculated and normalized to ensure the number of pixels equals 1.

Figure 4 shows histograms of LBP feature extraction from skin disease images with Eczema. In the LBP histogram, the horizontal line represents the bins or range of LBP values, while the vertical line represents the number of occurrences of LBP values in an image. The bins on the LBP histogram refer to the bin values used to represent the texture in the image, while the frequency indicates the number of occurrences of each bin value in the image.
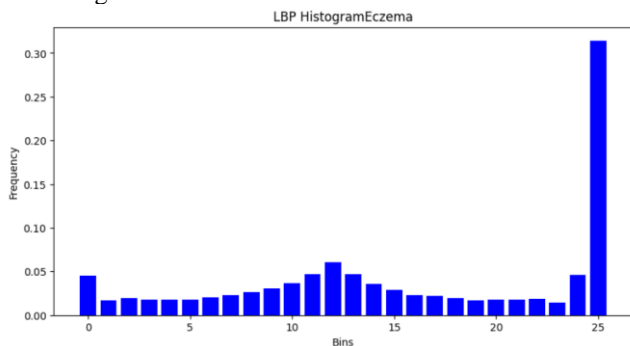

Figure 4. Histogram of LBP in Eczema Class

The HOG process generates feature vectors that describe the distribution of gradient directions in the image. Parameters such as orientation, pixels per cell, and cells per block are used to set the HOG feature extraction configuration. In Figure 5, there is a HOG histogram result, a feature description containing information about the distribution of orientation gradients in the image. The white lines in the HOG histogram represent the number of occurrences of orientation gradients in an image block or cell. Each cell in the histogram measures the orientation gradient distribution represented by several bins that represent the direction and magnitude of the gradient. The white line in the histogram indicates the dominant direction in the block.

After extracting the LBP and HOG features, they are combined into one feature vector and added to the feature list. In addition, the image class label has been added to the label list. Then, the LBP and HOG features are combined. The result of this combination is stored in one variable, which is the combined feature of LBP and HOG. Next, the combined feature vector is used for classification with SVM.
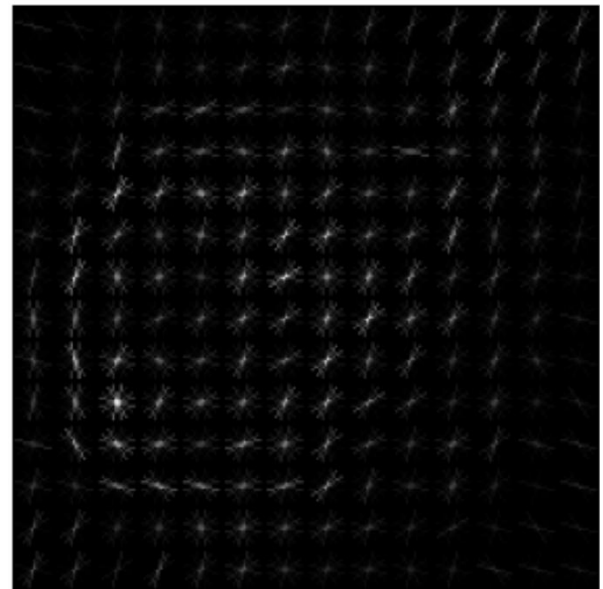

Figure 5.HOG Histogram

Then, training the model using SVM begins by initializing the required parameters. This research uses the One Versus Rest (OVR) multiclass SVM method with a C value of 100 and RBF kernel. Furthermore, testing uses a multiclass confusion matrix because this study has three labels. Testing was carried out using 210 test data. The following are the confusion matrix results for GLCM feature extraction and the combination of LBP and HOG. In Table V to Table VI, the confusion matrix results on GLCM and LBP+HOG feature extraction in this study.

TABLE V
CONFUSION MATRIX FOR GLCM FEATURE EXTRACTION

		Predicted		
		Eczema	Keratosis	Melanocytic Nevi
Actual	Eczema	51	6	6
	Keratosis	16	57	8
	Melanocytic Nevi	11	5	50

TABLE VI
CONFUSION MATRIX FOR LBP+HOG FEATURE EXTRACTION

		Predicted		
		Eczema	Keratosis	Melanocytic Nevi
Actual	Eczema	51	15	10
	Keratosis	9	54	9

	Predicted		
	Eczema	Keratosis	Melanocytic Nevi
Melanocytic Nevi	10	15	37

The confusion matrix results for GLCM feature extraction show that for the Eczema class, 51 data are correctly predicted (True Positive), 27 data are incorrectly predicted as Eczema (False Positive), and 12 Eczema data are incorrectly predicted as other classes (False Negative). The Keratosis class contained 57 TP, 11 FP, and 24 FN. The Melanocytic Nevi class had 50 TP, 14 FP, and 16 FN. The confusion matrix results for LBP+HOG feature extraction show 51 TP, 19 FP, and 25 FN for the Eczema class. The Keratosis class has 54 TP, 30 FP, and 18 FN. The Melanocytic Nevi class has 37 TP, 19 FP, and 25 FN.

These results show that the model with GLCM extraction tends to detect the Melanocytic Nevi class better than LBP+HOG, as it has a higher number of TPs and lower FPs and FNs. In contrast, for the Eczema and Keratosis classes, the model with LBP+HOG produced more errors (FP and FN) compared to GLCM.

IV. CONCLUSION

This research shows that the GLCM feature extraction method is more effective in classifying skin disease images than LBP+HOG. This study involved 1,050 skin disease data, including Eczema, Keratosis, and Melanocytic Nevi images obtained from the 'Skin Diseases Dataset' on Kaggle.com using a dermoscopy. The research process includes data sharing, preprocessing, feature extraction, and SVM classification. The results showed that GLCM with SVM achieved the highest accuracy of 74%, while LBP+HOG with SVM only achieved an accuracy of 68%. In the GLCM method, the extracted features include Homogeneity, Contrast, Energy, Correlation, ASM, and Dissimilarity. Using Pearson correlation in GLCM feature selection also helps improve accuracy. Therefore, it can be concluded that GLCM is more effective in feature extraction for skin disease image classification compared to LBP+HOG.

ACKNOWLEDGMENTS

REFERENCES

[1] R. Adawiyah and D. I. Mulyana, "Optimasi Deteksi Penyakit Kulit Menggunakan Metode Support Vector Machine (SVM) dan Gray Level Co-occurrence Matrix (GLCM)," *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, vol. 14, no. 1, 2022.

[2] R. Yohannes and M. E. Al Rivan, "Klasifikasi Jenis Kanker Kulit Menggunakan CNN-SVM," *Jurnal Algoritme*, vol. 2, pp. 133–144, 2022.

[3] Y. Fauzia Achmad, A. Yulfitri, and M. B. Ulum, "Identifikasi Jenis Jerawat Berdasarkan Tekstur Menggunakan GLCM dan Backpropagation," *Jurnal Sains Manajemen Informatika dan Komputer*, vol. 20, no. 2, 2021.

[4] W. I. Praseptiyana, A. W. Widodo, and M. A. Rahman, "Pemanfaatan Ciri Gray Level Co-occurrence Matrix (GLCM) Untuk Deteksi Melasma Pada Citra Wajah," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 11, 2019.

[5] P. R. Hegde, M. M. Shenoy, and B. H. Shekar, "Comparison of Machine Learning Algorithms for Skin Disease Classification Using Color and Texture Features," *IEEE Xplore*, 2018.

[6] M. Ezar Al Rivan and Orlando, "Klasifikasi Jenis Kanker Kulit Manusia Menggunakan Convolutional Neural Network," *MDP STUDENT CONFERENCE (MSC)*, vol. 2, 2023.

[7] G. Neela, K. Babu, and V. J. Peter, "Skin Cancer Detection Using Support Vector Machine with Histogram of Oriented Gradients Features," Ictact Journal on Soft Computing, vol. 11, no. 02, 2021.

[8] M. A. Rahayu, I. R. Magdalena, and R. Rahmania, "Identifikasi Penyakit Kulit Menggunakan Histogram of Oriented Gradients Dan Jaringan Syaraf Tiruan Backpropagation," *e-Proceeding of Engineering*, vol. 6, no. 2, 2019.

[9] H. S. Kusumastuti Aji Putri, C. Atika Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto, "Classification of Skin diseases types using naïve bayes classifier based on local binary pattern features" *2020 International Seminar on Application for Technology of Information and Communication (iSemantic),* 2020.

[10] N. Akmalia, P. Sihombing, and Suherman, "Skin Diseases Classification Using Local Binary Pattern and Convolutional Neural Network," *Telecommunication and Computer Engineering (ELTICOM)*, 2019.

[11] R. Pujiyanti Hidayat, I. Bambang Hidayat, drg Yuti Malinda, and M. Kes, "Identificaton of Lip Print Pattern on Human Identity Using Histogram of Oriented Gradients And Support Vector Machine For Biometric Forensic Application," *e-Proceeding of Engineering*, vol. 5, no. 3, 2018.

[12] D. D. Nur Cahyo *et al.*, "Sentiment Analysis for IMDb Movie Review Using Support Vector Machine (SVM) Method," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 8, no. 2, 2023.

[13] T. F. iraz Bintang Nusantara and R. Dwi Atmaja, "Klasifikasi Jenis Kulit Wajah Pria Berdasarkan Tekstur Menggunakan Metode Gray Level Co-Occurrence Matrix (GLCM) dan Support Vector Machine (SVM),", vol. 5, no. 2, 2018.

[14] N. L. Chusna, M. I. Shalahudin, U. Riyanto, and A. D. Alexander, "Klasifikasi Citra Jenis Tanaman Jamur Layak Konsumsi Menggunakan Algoritma Multiclass Support Vector Machine," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, 2022.

[15] L. M. Wisudawati, "Klasifikasi Tumor Jinak Dan Tumor Ganas Pada Citra Mammogram Menggunakan Gray Level Co-Occurrence Matrix (Glcm) Dan Support Vector Machine (SVM)," *Jurnal Ilmiah Informatika Komputer*, vol. 26, no. 2, 2021.

[16] D. P. N. Lyrawati, "Deteksi Ujaran Kebencian Pada Twitter Menjelang Pilpres 2019 Dengan Machine Learning," *Jurnal Ilmiah Matematika*, vol. 7, no. 3, 2019.

[17] R. Agustina, R. Magdalena, and N. K. C. Pratiwi, "Klasifikasi Kanker Kulit menggunakan Metode Convolutional Neural Network dengan Arsitektur VGG-16," *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 10, no. 2, 2022.