CNN Based Autoencoder Application in Breast Cancer Image Retrieval

Agus Eko Minarno
Informatics Department
Universitas Muhammadiyah Malang
Malang, Indonesia
aguseko@umm.ac.id

Kharisma Muzaki Ghufron
Informatics Department
Universitas Muhammadiyah Malang
Malang, Indonesia
kharisma.muzaki@gmail.com

Trifebi Shina Sabrila
Informatics Department
Universitas Muhammadiyah Malang
Malang, Indonesia
trifebiss@gmail.com

Lailatul Husniah
Informatics Department
Universitas Muhammadiyah Malang
Malang, Indonesia
husniah@umm.ac.id

Fauzi Dwi Setiawan Sumadi
Informatics Department
Universitas Muhammadiyah Malang
Malang, Indonesia
fauzisumadi@umm.ac.id

2021 International Seminar on Intelligent Technology and Its Applications (ISITIA) | 978-1-6654-2847-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISITIA52817.2021.9502205

***Abstract*—Content Based Medical Image Retrieval (CBMIR) is considered as a common technique to retrieve relevant images by comparing the features contained in the query image with the features contained in the image located in the database. Currently, the study related to CBMIR on breast cancer image however remains challenging due to inadequate research in such area. Previous study has a low performance and misinformation emphasizing the feature extraction process. Therefore, this study aims to utilize the CNN based Autoencoder method to minimize misinformation in the feature extraction process and to improve the performance result. The dataset used in this study is the BreakHis dataset. Overall, the results of image retrieval in breast cancer applying the CNN based Autoencoder method achieved higher performance compared to the method used in the previous study with an average precision of 0.9237 in the mainclass dataset category and 0.6825 in the subclass dataset category.**

***Index Terms*—Image retrieval, Breast cancer, Autoencoder, CNN**

I. Introduction

Breast cancer is acknowledged as a type of cancerous disease which provides a severe threat to women. In fact, breast cancer itself becomes one of the significant sources of death for women worldwide [1] with a total of 15% of deaths from breast cancer and a 24% increase in new cases as reported in 2018 [2]. This notion has attracted attention for early detection of breast cancer. A notable attempt which has been regarded to offer a supporting tool in diagnosing breast cancer is the CBIR system for histopathological images of breast cancer.

Content-Based Image Retrieval (CBIR) or commonly referred to as content-based image retrieval is defined as a technique retrieving the relevant image by comparing the features contained in the query image with the features contained in the database [3]. The features comprise several items such as colors, shapes, textures, borders or other features derived from the image. In addition, the performance of CBIR system is significantly dependent on the selected feature [4]. The growing application of CBIR has been based on a content-based medical image retrieval technique, acknowledged as Content Based Medical Image Retrieval (CBMIR). The mechanism of CBMIR works applies similar technique as in CBIR, focusing on a compilation of medical images with the main objective to assist the general practitioner to craft the informed decisions regarding the patient disease diagnosis [4].

Numerous researches have been devoted related to Content Based Medical Image Retrieval. However, research on CBMIR emphasizing the cases of breast cancer images is inadequate. Prior research unfortunately presents insufficient results; hence, further research is required regarding CBMIR targeting the cases of breast cancer images.

In previous research, Content Based Medical Image Retrieval on breast cancer images was conducted by Elaheh Mahraban Nejad et. al, applying BreakHis dataset by engaging the VGG-19 pre-trained deep CNN model for feature extraction and Euclidean Distance to calculate the similarity distance of the captured images. Of the 5 conducted experiments, the highest trial results yielded a MAP value of 0.80 [5]. Yun Gu, et al, similarly conducted a research on CBMIR on breast cancer images engaging similar dataset of BreakHis. The utilized method in this study is CNN-based Densely-Connected Multi-Magnification Hashing (DCMMH), in which the applied data has various magnification levels of: 40X, 100X, 200X, and 400X. This study further applies the two strategies such as low-magnification and high-magnification images. The test results indicate that the overall employed methods were eligible to produce an improved accuracy in the image retrieval task compared to the previous model, with an average MAP for each magnification level of 0.95. However, this study barely performs image retrieval on the main-class dataset [6]. Further, in 2019, Yun Gu, et al, reconducted another research with simi-

lar dataset as the previous one. This research is utilized as the author's main reference paper in this study In this research, Yun Gu, et al, conducted a retrieval process for the two categories of class datasets, including main-class and sub-class. The method used was Multi-Magnification Correlation Hashing (MMCH) Overall, the proposed method achieved satisfactory performance on the main-class dataset with the highest MAP value of 0.9416. However, in the sub-class category, the resulting performance was comparatively low, ranging from 0.61 to 0.66. This result is due to the misinformation of the image when hashing, leading to the less-relevant image query response [7].

This study proposes a Content Based Medical Image Retrieval (CBMIR) for histopathological images of breast cancer by employing the CNN-based Autoencoder method to amend the misinformation occurring in previous studies. The researchers propose the CBMIR method as this method demonstrates a feature extraction process and learning process to minimize the misinformation during feature extraction and to develop the resulting performance, especially in the subclass dataset category.

II. Method

A. Data collection

The utilized dataset in this study is BreakHis dataset retrieved from https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ containing a set of microscopic images of breast tumor tissue utilizing different magnification sizes (40X, 100X, 200X, and 400X) with a total data of 7909 images. This dataset itself is divided into 2 main classes including benign class (benign tumor with 2480 images) and malignant class (malignant tumor with 5429 images), each of which is further divided into 4 different subclasses. Benign class category includes: adenosis, fibroadenoma, phyllodes tumor, and tubular adenona. Malignant class category includes: ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. An example of the dataset image as employed in this study is illustrated in Fig.1 as benign tumor image and Fig. 2. as malignant tumor image.

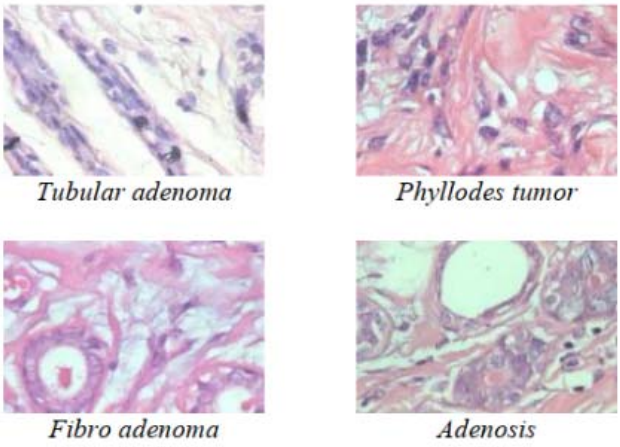

Fig. 1. Example of Benign Tumor Image

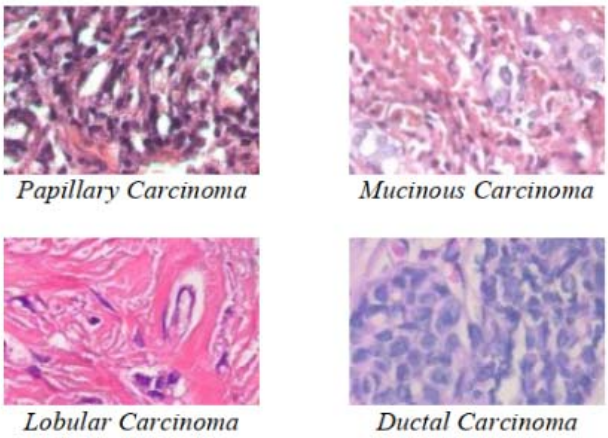

Fig. 2. Example of Malignant Tumor Image

B. Training

The training process in this study will initially begin with an image separation process of 80% for training, 10% for validation, and the remaining 10% for measuring the retrieval value (test). The original image in this study will be converted into a reconstructed image by utilizing the CNN-based Autoencoder method through the 2 stages of encoder and decoder.

Autoencoder will extract the output image to reconstruct the input image and to compare with the original input image [8]. Subsequently at this stage, the learning process will be conducted. After several iterations, this process will retrieve and save the optimal model with the lowest loss value against the established model, further applied in the feature extraction and retrieval stages. The data used in this training process comprises the data that has been broken down into training data and data validation. Additionally, the validation data will be employed to evaluate the previously established models. In simple terms, Autoencoder includes 3 main steps of: Encoder, Decoder, and Calculating Function and Optimization Errors [8]. Further details on the stages in the training process are illustrated in Fig.3. At the encoder stage, the input data will be converted into smaller dimensions or also entitled as compression, which in this study will be converted into a latent dimension with a size of 48 nodes by using Conv2D (2D Convolution Layer). Conv2D layer generates a tensor of outputs by convolving a convolution kernel with the layer input, in this study two modified default parameter were used. The first one is strides, stride is set to 2 for striding of the convolution along the height and width, another one is using "same" padding value which mean padding convolution with zero values is evenly distributed to the left/right or up/down of the input, resulting in output with the same height or

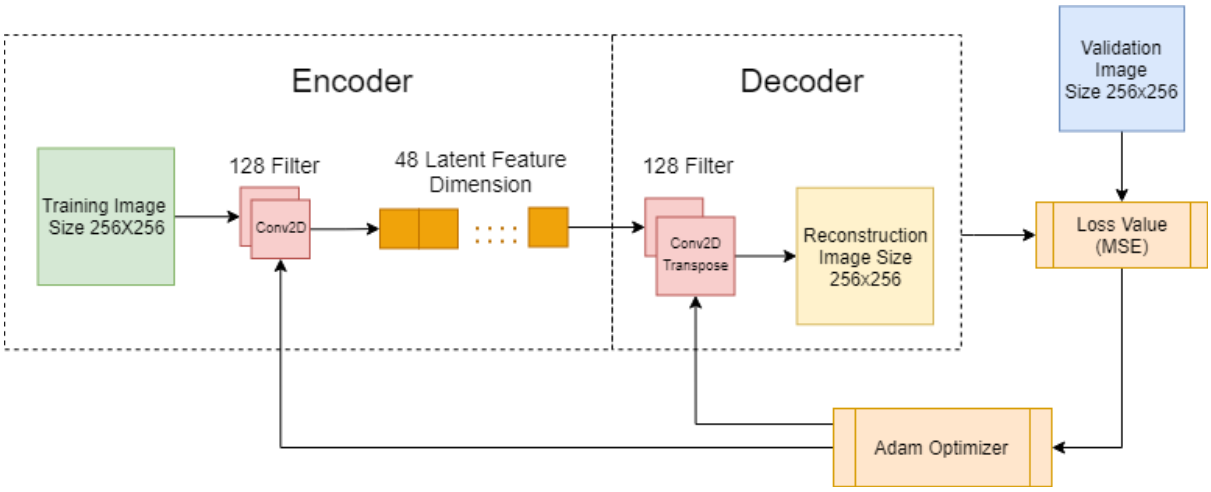

Fig. 3. Flow of Training Stages

width dimension as the input. The number of filters used is 128 with a size of 3x3 pixels. The activation function employed in convolution is LeakyRelu, which is widely applied in the field of artificial intelligence and machine learning, especially in deep learning, compared to the activation functions of its predecessors such as sigmoid and TanH, LeakyRelu is regarded simpler to overcome the problem of gradient loss; the equation in the LeakyRelu function is as follows [9]:

$$f(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \quad (1)$$

The parameter used on LeakyRelu is alpha as negative slope coefficient with 0.2 as the value. Consequently, upon performing the encoder function, the stage will proceed to the decoder stage to reconstruct the code or input data that has been converted into a latent dimension back into its original form, accredited as the reconstructed image.

The decoder has a same configuration of convolution method with encoder using 128 filters and 3x3 pixels as the size of filter, and using same activation function but differently in processing. The decoder using transpose data from the latent dimension to the reconstructed image. Upon the completion of the encoder and decoder stages, the stage is progressed to the error calculation and optimization phase. At this stage, the loss calculation for each function is conducted in each iteration to obtain the function with the lowest loss value. The loss function used in this study is the Mean Square Error (MSE). The equation used in the calculation of Mean Square Error (MSE) is as follows [10]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \quad (2)$$

Where:

f_i = Predicted value

y_i = Actual value

The weighted values and bias of the convolution are updated through optimization by applying the Adam's algorithm, by using a standard value of 1e-3 as the initiation of the learning rate and the learning rate scheduler using decay parameter with a rate of 2e-5 [11].

C. Feature Extraction

Feature extraction comprises a process to retrieve features from an image, further analyzed for the next process. Afterwards, the obtained features from each image are identified to distinguish an image from another image. [12]. In this study, the feature extraction process is performed from all the images contained in the data train to achieve the output in the form of a feature frequently identified as a vector feature of each existing image. An outline of the feature extraction process is illustrated in Fig.4, performed by applying the best stored model in the previous training stage through the encoder stage in the Autoencoder model. The results of the feature extraction are further indexed into a .json (JavaScript Object Notation) file. Apart from the features, the image file name and label on the image are also saved in the .json file. The collection of indexed images in this stage will be utilized as a database to compare the query image at the retrieval stage.

D. Retrieval

Retrieval comprises a process to retrieve a similar image based on the input query image. The data that has been separated into test data will be conducted by feature extraction as in the previous stage. Upon successfully extracting the images, the calculation is performed between the similarities between the features in the image in the test data and the features of the image in the training previously indexed in the .json file. Calculations are conducted by performing the distance calculation methods such as Manhattan Distance, Euclidean Distance, Haversine, and others.

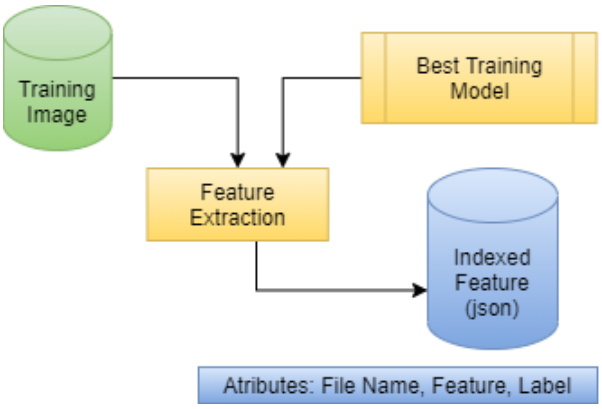

Fig. 4. Feature Extraction Process Flow

This study applies the Euclidean Distance to measure the similarity distance of two vectors.

Euclidean distance is one of the most widely applied distance calculation methods to measure the similarity distance of two vectors, whether two-dimensional, three-dimensional, or even more [13] [14]. The two feature vectors extracted from the feature are comparable each other by calculating the distance between them. The following equation for measuring distance using Euclidean Distance is most often used [14]:

$$d_E = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{3}$$

Where:

d_E = euclidean distance

n = number of vectors

a_i = vector image query (test data)

b_i = vector image in database (train data)

Upon calculating the distance by each image in the test data with all images in the train data, the results of the distance will be compared, in which smaller Euclidean value leads to more similar image. The image with the smallest distance value will be included in this retrieval process. In this study, the image retrieval process is performed based on the five most similar images or five images with the smallest Euclidean value.

E. Retrieval Evaluation

Evaluation serves as the stage to determine the performance of the system that has been built in research [15]. Precision and recall are engaged in this study to determine the success rate of the system. Retrieval process flow and retrieval evaluation is illustrated in Fig.5. The retrieval process flow first works by receiving input as an image from test data, then using saved encoder model in training to generate latent feature and forward to retrieval evaluation to calculating Euclidean distance based on indexed .json file while get on past step named feature extraction process flow, the final output is precision and recall value.

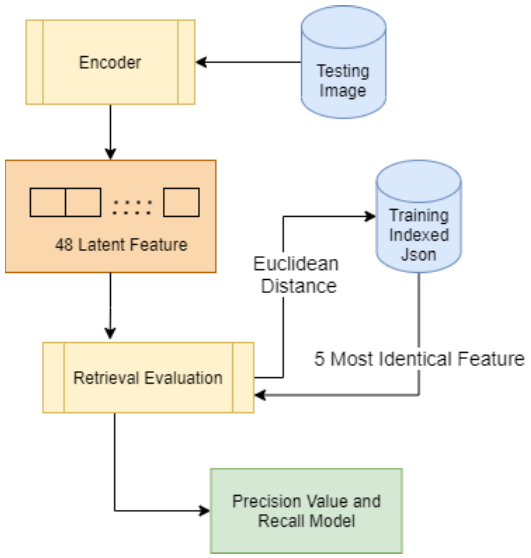

Fig. 5. Retrieval Process and Retrieval Evaluation Flow

The equations used to calculate precision and recall are defined as follows [16]:

$$Precision = \frac{R_v}{n} \tag{4}$$

$$Recall = \frac{R_v}{M} \tag{5}$$

Where:

R_v = Relevant image found

n = Number of images captured

M = Number of relevant images in the database

III. Results and Discussion

The BreakHis dataset employed in this study merely includes images with magnification sizes of 40X, 100X, 200X, and 400X for each class, both in the main class dataset category, and in the subclass dataset category. From the collected dataset, data will be divided both in the main class and subclass categories for each magnification into the data train, data validation, and test data with a ratio of 80:10:10. This separated data is further utilized in the testing process of this study. The training process employs 50 epochs for all magnification models and classroom scenarios. At each step, the 32 images were utilized to obtain 45 steps in each epoch. Image samples were obtained from the training stage when reconstructing the original image by applying the Autoencoder (seen in Fig.6), where the image is the result of a transpose using a 2-dimensional convolution of the latent feature hash value.

In Fig.7 and Fig.8, the training stage as performed in the testing process of the dataset in both categories of main class dataset and subclass dataset produces loss values gradually decreasing to epoch 50. Meanwhile, the main class dataset has the lowest error value obtained at 200x magnification. In the sub class, the lowest loss value is at 400x magnification. The

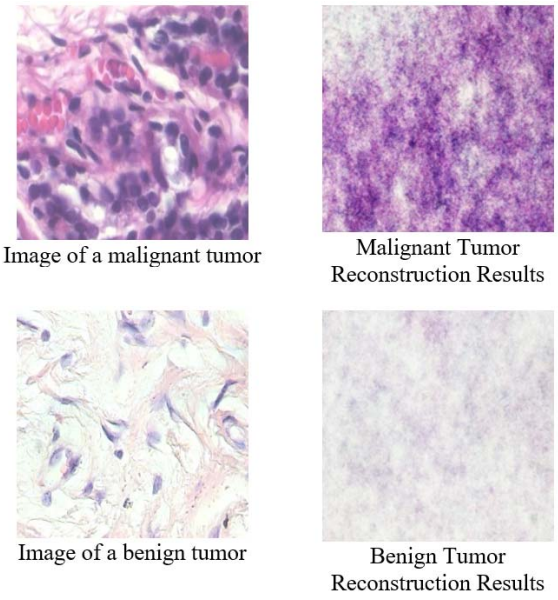

Fig. 6. Image Sample Reconstruction Results during Training

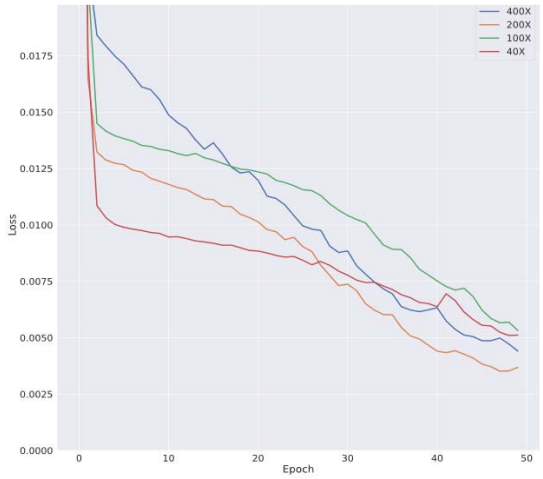

Fig. 7. Graph of Loss Value when Training Main Class Dataset

results of the loss value in the overall magnification value were obtained below 0.01 in which the retrieval process was possible.

Image retrievals which are relevant to the input query image are selected based on the similarity value generated through the Euclidean Distance calculation. The results of image retrieval by the established CBMIR system for each magnification measure of the queries are illustrated in Fig.9.

Based on the results of image retrieval using the CBMIR system that has been built, evaluation results or system performance are obtained in the form of precision and recall values. The results of the performance comparison between the systems established in the previous study and the proposed system for each dataset category are illustrated in Table 1 and Table 2. The results of the evaluation indicate that the proposed method has succeeded in providing satisfactory

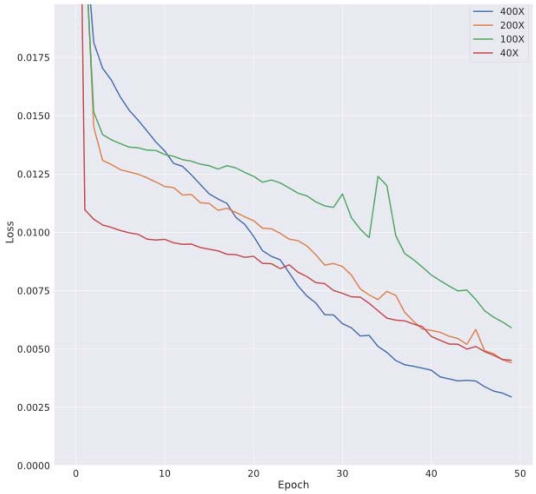

Fig. 8. Graph of the Value of Loss during the Sub Class Dataset Training

TABLE I
THE RESULTS OF SYSTEM PERFORMANCE COMPARISONS IN THE MAINCLASS DATASET CATEGORY

Magnification	Method	Precision	Recall
40X	MCCH [7]	0.9416	-
	Proposed Method	**0.9523**	0.7728
100X	MCCH [7]	**0.9202**	-
	Proposed Method	0.7781	0.4972
200X	MCCH [7]	0.9084	-
	Proposed Method	**0.9246**	0.7619
400X	MCCH [7]	0.8951	-
	Proposed Method	**0.9263**	0.6889

performance with an average precision value of 0.92 in the main class dataset category.

However, the results are less optimal compared to that in the subclass dataset category. However, when compared with the MCCH method in previous studies, the resulting performance in the subclass dataset category has already exceeded the performance generated by previous studies in this category.

IV. CONCLUSION

This study proposes an image retrieval system by employing the CNN-based Autoencoder method for

TABLE II
THE RESULTS OF SYSTEM PERFORMANCE COMPARISONS IN THE SUBCLASS DATASET CATEGORY

Magnification	Method	Precision	Recall
40X	MCCH [7]	0.6674	-
	Proposed Method	**0.7585**	0.4341
100X	MCCH [7]	0.6578	-
	Proposed Method	**0.6787**	0.3229
200X	MCCH [7]	0.6553	-
	Proposed Method	**0.6619**	0.3478
400X	MCCH [7]	0.6344	-
	Proposed Method	**0.7069**	0.3162

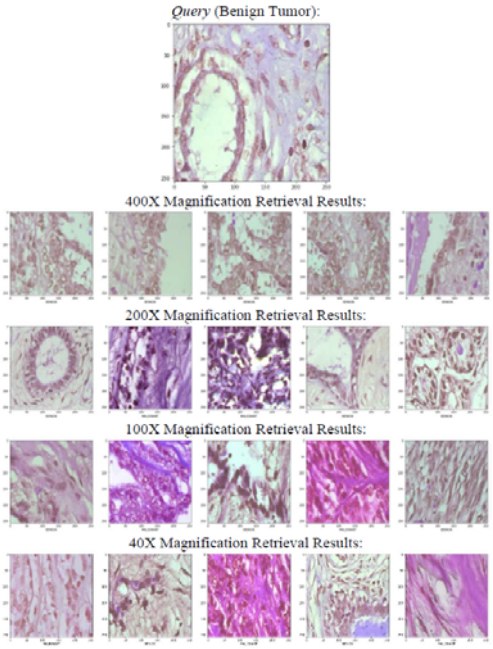

Fig. 9. Model of Retrieval Results in Benign Tumors

breast cancer images. In the proposed method, a learning process is conducted to minimize the misinformation during feature extraction and to improve the performance generated by previous research. Overall, the obtained results indicate that the proposed method has succeeded in achieving higher performance when compared to the MCCH method applied in previous studies, both in the main class dataset category and the subclass dataset category.

REFERENCES

[1] A. Al Nahid, M. A. Mehrabi, and Y. Kong, "Histopathological breast cancer image classification by deep neural network techniques guided by local clustering," Biomed Res. Int., vol. 2018, 2018, doi: 10.1155/2018/2362108.

[2] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, and M. M. Fidler-Benaoudia, "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study," Lancet Glob. Heal., vol. 8, no. 8, pp. e1027–e1037, 2020, doi: 10.1016/S2214-109X(20)30215-1.

[3] J. Rejito, A. S. Abdullahi, Akmal, D. Setiana, and B. N. Ruchjana, "Image indexing using color histogram and k-means clustering for optimization CBIR in image database," J. Phys. Conf. Ser., vol. 893, no. 1, 2017, doi: 10.1088/1742-6596/893/1/012055.

[4] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, "Medical image retrieval using deep convolutional neural network," Neurocomputing, vol. 266, pp. 8–20, 2017, doi: 10.1016/j.neucom.2017.05.025.

[5] E. Mahraban Nejad, L. S. Affendey, R. B. Latip, I. Bin Ishak, and R. Banaeeyan, "Transferred Semantic Scores for Scalable Retrieval of Histopathological Breast Cancer Images," Int. J. Multimed. Inf. Retr., vol. 7, no. 4, pp. 241–249, 2018, doi: 10.1007/s13735-018-0157-z.

[6] Y. Gu and J. Yang, "Densely-Connected Multi-Magnification Hashing for Histopathological Image Retrieval," IEEE J. Biomed. Heal. Informatics, vol. 23, no. 4, pp. 1683–1691, 2019, doi: 10.1109/JBHI.2018.2882647.

[7] Y. Gu and J. Yang, "Multi-level magnification correlation hashing for scalable histopathological image retrieval," Neurocomputing, vol. 351, pp. 134–145, 2019, doi: 10.1016/j.neucom.2019.03.050.

[8] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network," IEEE Trans. Big Data, vol. 7790, no. c, pp. 1–1, 2017, doi: 10.1109/tbdata.2017.2717439.

[9] Y. Liu, X. Wang, L. Wang, and D. Liu, "A modified leaky ReLU scheme (MLRS) for topology optimization with multiple materials," Appl. Math. Comput., vol. 352, pp. 188–204, 2019, doi: 10.1016/j.amc.2019.01.038.

[10] D. Li, L. Deng, B. Bhooshan Gupta, H. Wang, and C. Choi, "A novel CNN based security guaranteed image watermarking generation scenario for smart city applications," Inf. Sci. (Ny)., vol. 479, pp. 432–447, 2019, doi: 10.1016/j.ins.2018.02.060.

[11] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

[12] E. A. Abdel Maksoud, S. Barakat, and M. Elmogy, Medical Images Analysis Based on Multilabel Classification. Elsevier Inc., 2019.

[13] S. P. Patel and S. H. Upadhyay, "Euclidean distance based feature ranking and subset selection for bearing fault diagnosis," Expert Syst. Appl., vol. 154, 2020, doi: 10.1016/j.eswa.2020.113400.

[14] H. Prasetyo and B. A. Putra Akardihas, "Batik image retrieval using convolutional neural network," Telkomnika (Telecommunication Comput. Electron. Control., vol. 17, no. 6, pp. 3010–3018, 2019, doi: 10.12928/TELKOMNIKA.v17i6.12701.

[15] M. K. Alsmadi, "Content-Based Image Retrieval Using Color, Shape and Texture Descriptors and Features," Arab. J. Sci. Eng., vol. 45, no. 4, pp. 3317–3330, 2020, doi: 10.1007/s13369-020-04384-y.

[16] T. Zhang, S. Song, S. Li, L. Ma, S. Pan, and L. Han, "Research on gas concentration prediction models based on lstm multidimensional time series," Energies, vol. 12, no. 1, 2019, doi: 10.3390/en12010161.