Journal of Holistic Medical Technologies
Vol.1 No.1 (2024) 10 – 16
Available online at https://journal.innoscientia.org/index.php/jhmt

Classification of gastric tissue images based on texture characteristics using the Random Forest method

Hesti Windyasari[1*], Putri Zulfikah[1], Hanin Aisya Fakihati[1], Nabila Triwahyuni Handayani[1], Fitria Kholbi Azizah[1], and Wahyu Malda Sere[1]
[1]Department of Physics, Faculty of Science and Technology, Universitas Islam Negeri Walisongo Semarang, Indonesia
*Corresponding author's e-mail: hestiwindyaaa@gmail.com

ABSTRACT

Gastric cancer is a group of malignant diseases caused by many factors, including genetics, lifestyle, and environment. This study aims to create additional tools for distinguishing gastric cancer and normal in microphysical biopsy images from the Kaggle database; the dataset includes 98 gastric cancer and 95 normal. The method used in this research utilizes the coarse and delicate nature of the extracted image based on Histogram and Gray Level Co-occurrence Matrix (GLCM) texture features. Image classification uses the Random Forest method in WEKA software. The results showed that the highest accuracy was 94% in folds 15, 20, and 25, while the lowest accuracy was 93% in folds 5 and 10. This research can be an additional tool for differentiating microphysical biopsy images.

Keywords:
Cancer, Gastric cancer, Stomach, Biopsy.

Introduction

According to the Great Dictionary of the Indonesian language online, the disease is a condition that results in disturbances in living things, health disorders produced by bacteria, viruses, or fatal system or tissue abnormalities in organs or living things. Every individual has the potential to develop a disease, and many factors can trigger the disease, such as diet, sleep patterns, environment, psychological stress, and more. Therefore, humans must take care of their health by exercising, getting enough rest, eating healthy foods, and drinking at least eight glasses of water daily (Raharjo et al., 2016).

Gastrointestinal tumors are still one of the main problems in gastroenterology. In general, gastrointestinal cancer is a significant disease in the structure of developmental morbidity and mortality in oncology, which is a broad and complex field of medicine, suggesting that gastric cancer has long been the primary research model for many cancer problems. It is natural because, in the first decade of the last century, gastric cancer was often considered the most common tumor disease. Many existing clinical and oncological diagnostic postulates have been studied in detail in patients with gastric cancer. Currently, the picture of the incidence and death of cancer due to various types of cancer in the world has changed somewhat (Portnoy, 2006).

Symptoms of gastric cancer in the early stages are often found in adults aged 30 years and older, with 80% of cases occurring in those aged 40 years and above. Although some patients have no history of gastric disease, they may experience signs such as fullness in the upper abdomen, loss of appetite, diarrhea, anemia, fatigue, and changes in the stool. Laboratory tests are necessary to confirm the diagnosis of gastric disease (Raharjo et al., 2016).

Stomach cancer is a group of malignant diseases caused by multiple factors, including genetics, lifestyle, and environment. Hereditary Diffuse Gastric Cancer (HDGC) can be caused by genetic mutations on chromosome 16. In addition to genetic factors, an unhealthy diet, smoking, and alcohol consumption are also risk factors for developing stomach cancer. An unhealthy diet can lead to Helicobacter pylori bacteria in a person's stomach, which can then develop into a malignant bacterium. Epstein-Barr virus (EBV) infection is another risk factor for stomach cancer, aside from H. pylori bacterial infection. EBV infection in stomach cancer patients indicates their immune system, which can affect their prognosis (Chudri, 2020).

The stomach plays a central role in regulating the digestive process, a fact that is often underestimated. Furthermore, gastric acid secretion in recent decades has been viewed as a 'bystander' with a minor role but with potential harm to itself and the surrounding organs, such as the esophagus and duodenum. Consequently, pharmacological approaches have led to the development of more potent acid-inhibiting drugs. Due to the growing awareness of gastrointestinal dysfunctions, the role of the stomach has been re-evaluated as the origin of dyspepsia symptoms. Recently, attention has been focused on the stomach due to its control functions in food intake and its contribution to maintaining metabolic balance (Hunt et al., 2015).

In some cases, biologists use microscopic histopathology biopsy images, which are images of a patient's microscopic tissue structure. A sample image of stomach cancer is obtained through a biopsy. A biopsy is a process that involves taking a piece of tissue or a sample of cells from the body to be tested in a laboratory. A biopsy may be performed when a patient has certain signs and symptoms, or if a doctor has identified an area of concern. A biopsy can provide information about whether a patient has cancer or other diseases. Consequently, analyzing biopsy images is an important technique for cancer diagnosis (Cataldo et al., 2010).

According to the International Agency for Research on Cancer (IARC), in 2000, 10.1 million new cases of cancer were diagnosed in all locations. There were 6.2 million cancer deaths and a total of 22 million cancer patients during the 5-year survival period. Compared to 1990 data, the cancer incidence in 2000 increased by 22%. The cancer profile varies greatly depending on the parameters studied. The incidence of stomach cancer is much higher worldwide. Nearly two-thirds of all stomach cancer incidences occur in less developed countries. However, the incidence of this disease is also high in some economically developed countries. It includes countries in Eastern Europe, East Asia, South and Central America, some countries of the former Soviet republics, and most African countries (Portnoy, 2006).

In previous research conducted by Wang et al. (2022), through experimental verification on SIER stomach cancer patient data, the Weighted Random Forest algorithm increased accuracy by 0.79% compared to the original Random Forest. Regarding AUC, the macro-average increased by 2.32%, and the micro-average increased by 0.51% on average. Among 10 public datasets, the accuracy of Weighted Random Forest performed best in 6 datasets, with an average accuracy increase of 1.44% and an average AUC increase of 1.2%. This study differs from previous research in data processing techniques. This study aims to differentiate between normal stomach images and stomach cancer images to serve as an additional tool for healthcare professionals in diagnosis.

Methods

Data processing procedures
The study used gastric tissue images from the kaggle.com database, which included 95 images of cancer and 98 images of normal images. Gastric tissue image extraction was carried out using Python in Google Colab, then classification using WEKA machine learning algorithm. The stages of the research can be seen in Figure 1. There are three main stages: preprocessing, texture feature extraction, and using the Random Forest method for classification.

Preprocessing
There are several processes, namely taking images from kaggle.com, then collecting images, and converting RGB input images to grayscale.

Texture Feature Extraction
The results of this extraction process yield significant data that can then be used to obtain important information. This information, in turn, can provide a comprehensive description and interpretation of an object (Purwaningsih et al., 2015). The texture feature extraction in this study includes Histogram and GLCM (Gray Level Co-Occurrence Matrix) (Rizal et al., 2019).

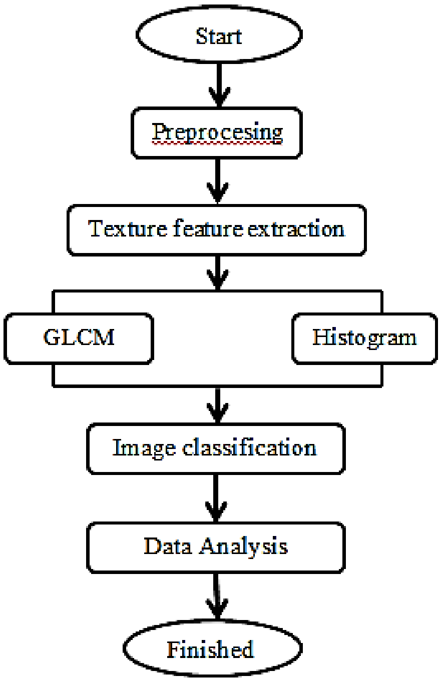

Figure 1. Flowchart research

1. Histogram

Histogram equalization will enhance the brightness and contrast of dark and low-contrast images and make features not visible in the initial image observable by distributing the gray levels in the image so that each gray level is equal (Istianah & Sumarti, 2020). One of the easiest methods to describe texture using statistical methods related to the image intensity histogram includes (Zaheeruddin et al., 2012):

[1] The mean is the average used to calculate the image histogram intensity.

$$\mu = \sum_{n=0}^{N} z_n p(z_n) \tag{1}$$

where z_n is related to the range of gray level intensities in the image, $p(z_n)$ is the image's pixel value with the same gray level as n, μ is the mean value, and N is the highest gray level.

[2] Variance $(\sigma)^2$ is a measure of contrast useful for describing the level of smoothness.

$$\sigma^2 = \sum_{i=0}^{I} (z_n - \mu)^2 p(z_n) \tag{2}$$

[3] Skewness (α_3) is a criterion for the degree of symmetry of the histogram, measuring whether the distribution of gray levels is symmetric or not.

$$\alpha_3 = \frac{1}{\alpha_3} \sum_{n=0}^{N} (z_n - \mu)^3 p(z_n) \tag{3}$$

[4] Entropy (S) is a measure of the randomness of variation and will have a value of zero for a uniform image.

$$S = -\sum_{n=0}^{N} p(z_n)^2 \log_2 p(z_n) \tag{4}$$

[5] Standard deviation (σ) is a measure of the average contrast.

$$\sigma = \sqrt{\sigma}^{4} \tag{5}$$

[6] Kurtosis (α^4) is a metric that indicates whether the data distribution is flat or skewed.

$$\alpha^4 = \frac{1}{\alpha^4}\sum_{n=0}^{N} (F_n - \mu)^4 \, p(F_n) - 3 \tag{6}$$

2. Grayscale Emergence Matrix (GLCM)

Grayscale Emergence Matrix (GLCM) is used to analyze textures or extract features. GLCM is a matrix representation that describes how often pairs of two pixels with a certain level of grayness appear within a specified distance and direction in an image (Prasetyo, 2011). The GLCM calculates the values of Homogeneity, Contrast, Correlation, and Energy (Zhou et al., 2017):

[1] Homogeneity assesses how close the pixel intensities in the co-occurrence matrix are to the mean intensity value of the entire image.

$$\sum_{i,j=0}^{g-1} \frac{\mathrm{p}_{\mathrm{i,j}}}{1 + |i - j|} \tag{7}$$

where g−1 represents the number of gray levels, and $\mathrm{p}_{\mathrm{i,j}}$ is the normalized GLCM matrix.

[2] Contrast measures the extent to which pixel intensities differ in the co-occurrence matrix.

$$\sum_{i,j=0}^{I-1} |i - j|^2 \mathrm{p}_{\mathrm{i,j}} \tag{8}$$

where I−1 is the number of gray levels.

[3] Correlation assesses the extent of the linear relationship between pixel intensities in the co-occurrence matrix.

$$\sum_{i,j=0}^{g-1} \frac{(i - \mu_i)(j - \mu_j)\ \mathrm{p}_{\mathrm{i,j}}}{\sigma_{i\times}\sigma_j} \tag{9}$$

where μ_i and μ_j are the mean values of the matrix $\mathrm{p}_{\mathrm{i,j}}$, and σ_i and σ_j are the standard deviations of the matrix $\mathrm{p}_{\mathrm{i,j}}$.

[4] Energy represents texture homogeneity by measuring how evenly pixel intensities distribute energy within the matrix.

$$\sum_{i,j=0}^{I-1} {\mathrm{p}_{\mathrm{i,j}}}^2 \tag{10}$$

Random Forest

This research began with biopsy and data collection, the Random Forest method is a machine learning ensemble method used to improve the accuracy of the classification method by combining the random forest method. Random Forest is an evolution of the decision tree method that uses multiple decision trees, each trained on individual samples, and each feature set is split into trees selected from a randomly chosen subset of features. Random Forest offers several advantages, including improved accuracy when data is missing, resilience to outliers, and practical data storage. It features a process for feature selection that can use the best features to enhance classification model performance and can operate effectively on large datasets with complex parameters (Givari et al., 2022).

Data Analysis Techniques

The accuracy of diagnostic values in classification can be illustrated by the evaluation index values obtained from data classification results using the WEKA machine learning algorithm. In this study analysis, accuracy, sensitivity, and specificity will be explained and measured as follows (Saifullah & Suryotomo, 2021):

a) Accuracy

It is the ratio of the overall correct prediction to the total data.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

b) Precision

The number of positive correct predictions is divided by the total positive predictions. It shows how many of the classes predicted to be positive are actually positive. Measure how accurate the model is in detecting disease.

$$Presisi = \frac{TP}{TP + FP} \quad (12)$$

c) Recall

The number of positive correct predictions is divided by the total positive actual instances. It shows how many of the positive actual instances are successfully predicted as positive. Measure how well the model identifies all actual cases of the disease.

$$Recall = \frac{TN}{TP + FN} \quad (13)$$

where True Positive (TP) refers to cases correctly identified as malignant tumors, False Positive (FP) refers to cases incorrectly identified as malignant tumors, True Negative (TN) refers to cases correctly identified as benign tumors, and False Negative (FN) refers to cases incorrectly identified as benign tumors.

Research and Discussions

Figure 2(a)shows a picture of a normal gastric diagnosis that is biopsied using a microscope. A normal stomach tends to be cleaner, and there is no rather dark color, as in the picture of stomach cancer. Normal gastric bypass is a condition in which the cells in the stomach wall grow normally and uncontrollably. The stomach consists of three regions: the cardia, fundus, and pylorus. In Figure 1(b), some cells grow abnormally in the stomach lining, which becomes practical. The extraction features used are feature texture using histograms and the Gray Level Co-occurrence Matrix (GLCM).

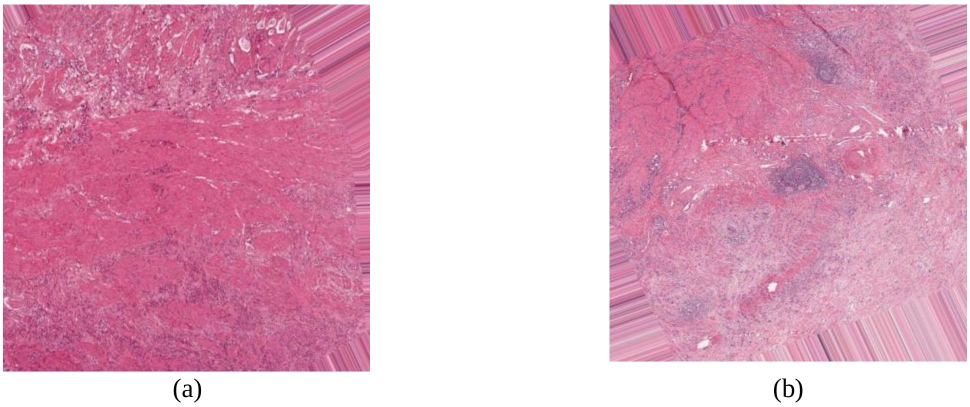
(a) (b)

Figure 2. (a) Normal stomach and (b) Stomach Cancer

Table 1 shows average results in the normal stomach and average stomach with a cancer diagnosis; there are 10 data, namely mean, standard deviation, variance, entropy, skewness, kurtosis, energy, contrast, correlation, and homogeneity. From these data, a difference with a large average was obtained, namely mean, variant, standard deviation, entropy, and contrast. Meanwhile, the average difference is slight: skewness, curtosis, energy, correlation, and homogeneity. The difference between this study and the previous research by Tong et al. (2016) lies in identifying stomach cancer using the Random Forest method with different data processing techniques, leading to differing results in the attributes. Their study obtained mean and standard deviation values.

Table 1. Average results in normal and cancer stomach

No	Atribut	Average in Normal	Average in Cancer
1	Mean	160.8503±16.8408	171.0086±14.8385
2	Standar Deviasi	22.6115±5.6032	28.3109±5.9976
3	Varian	542.6919±306.1116	837.6741±372.2126
4	Entropi	13.3536±0.0053	13.3497±0.0056
5	*Skewness*	0.2124±0.6732	0.1037±0.4681
6	Kurtosis	2.1618±1.6372	0.3134±1.0272
7	Energi	0.0244±0.0069	0.0304±0.0356
8	Contrast	220.3714±173.5746	295.7936±184.8686
9	Correlation	0.8035±0.0728	0.7997±0.1379
10	Homogenitas	0.1519±0.0424	0.1356±0.0663

The results in Table 2 of the analysis using random Forest also use a descriptive survey method with a cross-sectional approach, with diagnostic tests—the results from the table above help to calculate the accuracy value in gastric tissue sufferers. The results of the experiment from the table show that the highest accuracy results in folds 15, 20, and 25 are 94% while the lowest accuracy is in folds 5 and 10 with an accuracy of 93%, this accuracy value can help in making it easier to classify where it is a reference for the correct prediction ratio or one of the samples, The highest precision is almost the same as the result at folds 10, 15, 20, 25 by 92% while the lowest precision is almost close to 91% at fold 5, for the highest recall result is 97% at folds 15 and 20. The lowest recall, which is 95% in fold 5, the Random Forest method can provide gastric tissue image classification with a high degree of accuracy. This classification is obtained using the Random Forest method. Compared to the previous research by Xu et al. (2022), which predicted stomach cancer patient survival using the Random Forest method and achieved the highest accuracy of 85.54%, the differences are attributed to the amount of data used, image processing methods, and data processing techniques.

Table 2. Random Forest

Folds	TP (Data)	FP (Data)	FN (Data)	TN (Data)	Accuracy (%)	Precision (%)	Recall (%)
Training	98	0	0	95	100	100	100
5	93	9	5	86	93	91	95
10	93	8	5	87	93	92	95
15	95	8	3	87	94	92	97
20	95	8	3	87	94	92	97
25	94	8	4	87	94	92	96

The model's ability to recognize texture features related to specific pathological conditions can assist physicians and diagnosticians in early diagnosis and planning appropriate interventions. This study provides insights into the differences between normal stomach and stomach cancer based on texture features and can serve as an additional diagnostic tool for healthcare professionals. However, the limitations of this study include the lack of detailed values regarding tumor size, cancer type, and subsequent actions for healthcare professionals in treatment.

Conclusion

The classification of gastric tissue images based on texture characteristics in gastric and normal cancer amounted to 193, with 98 cancer images and 95 normal images. From the table results obtained using the random forest method, the highest accuracy value is 94% at folds 15, 20, 25, the lowest accuracy is 93% at folds 5 and 10. The highest precision value is 92% in folds 10, 15, 20, and 25; the lowest is 91%

in fold 5. The highest recall value is 97% in folds 15 and 20, while the lowest recall is 95% in folds 5 and 10.

Acknowledgments

We would like to thank Universitas Islam Negeri Walisongo Semarang for providing the necessary facilities and resources for this work.

Conflicts of interest

The authors affirm that they have no conflicts of interest.

References

Cataldo, S. Di, Ficarra, E., Acquaviva, A., & Macii, E. (2010). Automated segmentation of tissue images for computerized IHC analysis. *Comput Methods Programs Biomed*, *100*(1), 1–15.

Chudri, J. (2020). Kanker lambung: kenali penyebab sampai pencegahannya. *Jurnal Biomedika Dan Kesehatan*, *3*(3), 144–152. https://doi.org/10.18051/jbiomedkes.2020.v3.144-152

Givari, M. R., Sulaeman, M. R., & Umaidah, Y. (2022). Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit. *Nuansa Informatika*, *16*(1), 141–149. https://doi.org/10.25134/nuansa.v16i1.5406

Hunt, R. H., Camilleri, M., Crowe, S. E., El-Omar, E. M., Fox, J. G., Kuipers, E. J., Malfertheiner, P., McColl, K. E. L., Pritchard, D. M., Rugge, M., Sonnenberg, A., Sugano, K., & Tack, J. (2015). The stomach in health and disease. *Gut*, *64*(10), 1650–1668. https://doi.org/10.1136/gutjnl-2014-307595

Istianah, L., & Sumarti, H. (2020). Classification of Pneumonia in Thoracic X-Ray images based on texture characteristics using the MLP (Multi-Layer Perceptron) method. *Journal of Natural Sciences and Mathematics Research*, *6*(2), 78–84.

Portnoy, L. M. (2006). *Radiologic Diagnosis of Gastric Cancer*. Springer Berlin, Heidelberg.

Prasetyo, A. B. (2011). *Formulasi Anti Nyamuk Spray Menggunakan Bahan Aktif Minyak Nilam*. Institut Pertanian Bogor.

Purwaningsih, N., Soesanti, I., & Nugroho, H. A. (2015). Ekstraksi Ciri Tekstur Citra Kulit Sapi Berbasis Co-Occurrence Matrix. *Seminar Nasional Teknologi Informasi Dan Multimedia*, 6–8.

Raharjo, J. S. D., Damiyana, D., & Hidayatullah, M. (2016). Sistem Pakar Diagnosa Penyakit Lambung dengan Metode Forward Chaining Berbasis Android. *Jurnal Sisfotek Global*, *6*(2).

Rizal, R. A., Gulo, S., Della, O., Sihombing, C., Bernandustahi, A., Napitupulu, M., Gultom, A. Y., & Siagian, T. J. (2019). Analisis Gray Level Co-Occurrence Matrix (Glcm) Dalam Mengenali Citra Ekspresi Wajah. *Jurnal Mantik Augustus: Manajemen, Teknologi Informatiak Dan Komunikasi*, *3*(2), 31–38.

Saifullah, S., & Suryotomo, A. P. (2021). Identification of chicken egg fertility using SVM classifier based on first-order statistical feature extraction. *ILKOM Jurnal Ilmiah*, *13*(3), 285–293. https://doi.org/10.33096/ilkom.v13i3.937.285-293

Tong, W., Ye, F., He, L., Cui, L., Cui, M., Hu, Y., Li, W., Jiang, J., Zhang, D. Y., & Suo, J. (2016). Serum biomarker panels for diagnosis of gastric cancer. *OncoTargets and Therapy*, *9*, 2455–2463. https://doi.org/10.2147/OTT.S86139

Xu, C., Wang, J., Zheng, T., Cao, Y., & Ye, F. (2022). Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Archives of Medical Science*, *18*(5), 1208–1220. https://doi.org/10.5114/aoms/135594

Zaheeruddin, Jaffery, Z. A., & Singh, L. (2012). Detection and shape feature extraction of breast tumor in mammograms. *Lecture Notes in Engineering and Computer Science*, *2198*, 719–724.

Zhou, J., Yan Guo, R., Sun, M., Di, T. T., Wang, S., Zhai, J., & Zhao, Z. (2017). The Effects of GLCM parameters on LAI estimation using texture values from Quickbird Satellite Imagery. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-07951-w