ISSN: 2581-1274

Comparison of Convmixer Method and Resnet Method in Classification and Detection of Gastrointestinal Diseases Using Kvasir Dataset

Yuliana [a,1,*], Ivan Septa A P [a,2], Riska Veny F [a,3]

[a] *University Pamulang, Jl. Raya Puspiptek No. 46, Kel. Buaran, Kec. Serpong, South Tangerang City, Banten 15310, Indonesia*
[1] *dosen02557@unpam.ac.id*;* [2] *ivanseptaa@gmail.com;* [3] *riskaveny31@gmail.com*
*Corresponding author

ARTICLE INFO

Article history:
Published
July 16, 2025

Keywords:
ConvMixer
ResNet
Deep learning
Kvasir
Gastrointestinal
Evaluation metrics

ABSTRACT

This research discusses the comparison of ConvMixer and ResNet methods in the classification and detection of gastrointestinal diseases using the Kvasir dataset. Gastrointestinal diseases are often difficult to detect early due to the similarity of visual patterns in endoscopy images, requiring an efficient deep learning-based solution. The purpose of this research is to compare and evaluate the models used. The research used a quantitative approach with an experimental method. Endoscopy image data was processed through augmentation, normalization, and division of the dataset into train, validation, and test. ConvMixer and ResNet were trained with customized hyperparameters, and evaluated using accuracy, precision, recall, and F1-score metrics.The results showed that ResNet excelled with 86% accuracy, slightly higher than ConvMixer which recorded 84% accuracy. ResNet's residual structure overcomes the problem of vanishing gradients, while ConvMixer offers better training speed. Both models showed high performance, although the challenge of similar patterns between classes was still an obstacle. As a result, ResNet provides better results in detecting gastrointestinal diseases, but ConvMixer is also a promising alternative. Further development with more diverse datasets is needed to improve model performance.

I. Introduction

Gastrointestinal (GI) disorders are disorders or diseases of the food/digestive tract. Gastrointestinal diseases include disorders of the esophagus (eshopagus), stomach (gaster), small intestine (intestinum), colon (colon), liver (liver), bile duct (biliary tract) and pancreas [1].

In the context of healthcare, rapid and accurate diagnosis is essential to determine the right treatment, especially for diseases that are progressive and high-risk. However, in practice, many cases are not detected early due to limited time, human resources, and full reliance on manual interpretation from medical personnel[2].

One of the much-needed alternatives in disease detection accuracy is deep learning, which is a promising approach in the development of clinical decision support systems, especially in medical image processing and interpretation[3]. In recent years, with the continuous development of medical technology, new gastrointestinal endoscopy technology has been widely used in clinical practice. Gastrointestinal endoscopy has gradually evolved from a traditional tool to a diagnostic and tarapeuk tool [4] Endoscopy not only helps in the diagnosis of diseases, but also helps in the cure of certain disorders[5]. The Kvasir dataset is a widely used data source in the field of medical image analysis, especially for gastrointestinal disease detection. This dataset was introduced as part of the MediaEval

medical multimedia challenge and consists of thousands of images collected through endoscopic procedures[6].

Objective In this study, ConvMixer and ResNet models are proposed to detect and classify gastrointestinal diseases from endoscopy images. The proposed ConvMixer and ResNet models present a depp learning model that is tested for comparison to find the maximum accuracy. A thorough analysis was performed on the Kvasir dataset, which contains high-resolution endoscopic images of various gastrointestinal diseases and normal tissues, to test the effectiveness of the proposed ConvMixer and ResNet models [7]. However, the similarity of the visual patterns of the samples within each class of this dataset is an issue that can directly impact the accuracy of these models. This research addresses this issue by using data augmentation techniques that serve to make the data vary in order to improve the performance of the model for better test results.

II. Method

The research method used is a literature review or literature study, which contains theories that are relevant to the research problems that have been made. In this section, an assessment of the concepts and theories used is carried out based on the available or collected literature, especially from articles published in various scientific journals.

This research also uses a quantitative approach with an experimental method that compares two deep learning models, ResNet and ConvMixer, to measure the effectiveness and accuracy of the models [8].

III. Results and Discussion

A. Analysis of system design

The system that will be proposed by the author is a development of the current system that only uses a consultation system and direct analysis of images obtained through endoscopy procedures into a system in the form of a web-based application to detect and classify gastrointestinal diseases using two deep learning models whose training model data sources come from the Kvasir dataset, the author hopes that the system built can help doctors and patients get accurate and efficient diagnostic results [9]. Based on the author's analysis that has been carried out, the author proposes a system design that has the following procedures:

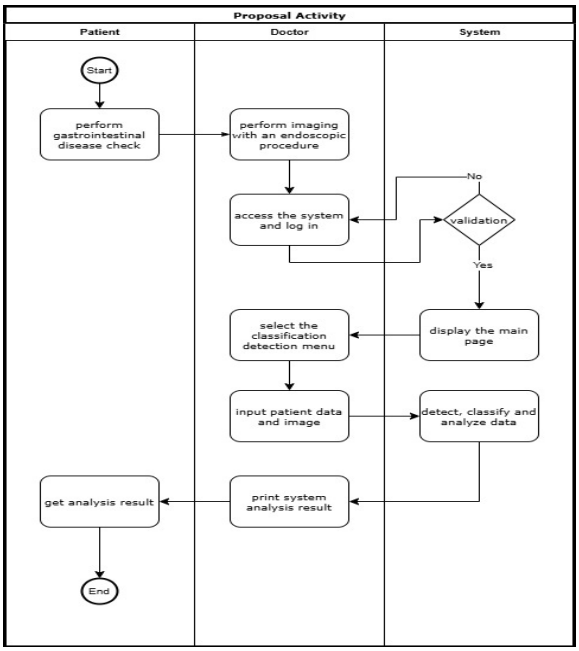

Fig. 1. Analysis of system design

In the proposed System Analysis above, the author proposes that the patient conducts a gastrointestinal disease examination and then the doctor takes pictures with an endoscopy procedure where the images are used for system analysis to help doctors diagnose diseases based on the patient's endoscopic images.

B. Data collection

Data collection in this study was from images taken using an endoscope. With respect to direct use in the multimedia research area, the main application area of Kvasir is the automatic detection, classification, and localization of endoscopic pathological findings in images taken in the gastrointestinal tract. As such, the provided dataset can be used in several scenarios where the goal is to develop and evaluate algorithmic analysis of images. By using the same dataset, researchers can easily compare experimental approaches and results, and the results can be easily reproduced [7].

The training dataset is separated into 3 groups consisting of train, validation and test datasets, each group has 9 classes in it. The distribution of datasets that the author uses is as follows:

Table 1. Dataset distribution

Class	Data *Training*	Data *Validation*	Data *Testing*	amount
Dyed-lifted-polyps	350	100	50	500
Dyed-resection-margins	350	100	50	500
Esophagitis	350	100	50	500
Normal-cecum	350	100	50	500
Normal-pylorus	350	100	50	500
Normal-z-line	350	100	50	500
Polyps	350	100	50	500
Ulcerative-colitis	350	100	50	500
Unknown	350	100	50	500
	Total Data			4500

In the process of dividing the model training dataset, the dataset consisting of 4500 endoscopy images is divided into three main groups: train, validation and test. The training group trains the model to learn the patterns and characteristics of the data, while the validation group evaluates the performance of the model during the training process, helps to tune the hyperparameters, and prevents interference and The test group is used to evaluate the final performance of the model after the training and tuning process is completed. This data is never used during the training or validation process, so it can provide an objective picture of the model's capabilities against data that has never been seen before.

The data in each class is divided proportionally according to the ratio, so that each class has an equal amount of data in all three groups. For example, out of a total of 500 data in the “Dyed-lifted-polyps” class, 350 is allocated to train, while 100 is allocated to validation and 50 to testing before data preprocessing. So that the model can be trained and evaluated fairly, this division ensures consistent class representation within each group.

C. Data Pre-processing

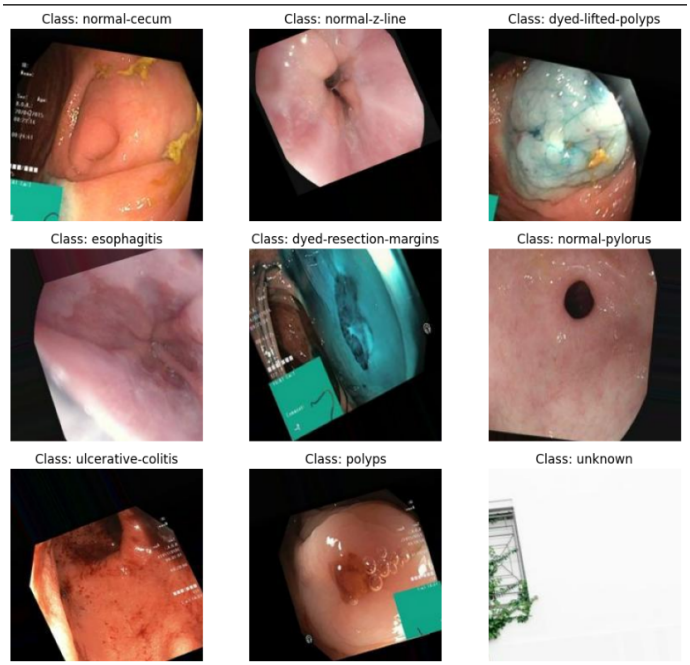

Fig. 2. Data pre-processing

The figure above is the result of the data pre-processing process carried out to increase the variety of data aimed at improving the performance and learning of the model in the training process.

D. Model Arsitektur

1. ConvMixer

The ConvMixer model was chosen because it has high learning capability and good accuracy, especially when used on large datasets. The ConvMixer method offers various architectures, which are designed based on experimental results from previous research. Below is a picture of the ConvMixer model architecture.

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 224, 224, 3)	0	-
conv2d (Conv2D)	(None, 24, 24, 256)	62,464	input_layer[0][0]
depthwise_conv2d (DepthwiseConv2D)	(None, 24, 24, 256)	12,800	conv2d[0][0]
batch_normalization (BatchNormalization)	(None, 24, 24, 256)	1,024	depthwise_conv2d[0][0]
add (Add)	(None, 24, 24, 256)	0	conv2d[0][0], batch_normalization[0…
conv2d_1 (Conv2D)	(None, 24, 24, 256)	65,792	add[0][0]
batch_normalization_1 (BatchNormalization)	(None, 24, 24, 256)	1,024	conv2d_1[0][0]
depthwise_conv2d_1 (DepthwiseConv2D)	(None, 24, 24, 256)	12,800	batch_normalization_1…
batch_normalization_2 (BatchNormalization)	(None, 24, 24, 256)	1,024	depthwise_conv2d_1[0]…
add_1 (Add)	(None, 24, 24, 256)	0	batch_normalization_1… batch_normalization_2…
conv2d_2 (Conv2D)	(None, 24, 24, 256)	65,792	add_1[0][0]
batch_normalization_3 (BatchNormalization)	(None, 24, 24, 256)	1,024	conv2d_2[0][0]
depthwise_conv2d_2 (DepthwiseConv2D)	(None, 24, 24, 256)	12,800	batch_normalization_3…
batch_normalization_4 (BatchNormalization)	(None, 24, 24, 256)	1,024	depthwise_conv2d_2[0]…

Fig. 3. Model ConvMixer

2. ResNet

The ResNet model was chosen for its superior ability to overcome the vanishing gradient problem, thus allowing for very deep network training. With its residual architecture, ResNet is able to maintain high accuracy and show consistent performance on various types of datasets, including large datasets. The following is the architecture of ResNet.

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 224, 224, 3)	0	-
conv2d (Conv2D)	(None, 112, 112, 64)	9,472	input_layer[0][0]
batch_normalization (BatchNormalization)	(None, 112, 112, 64)	256	conv2d[0][0]
activation (Activation)	(None, 112, 112, 64)	0	batch_normalization[0…
max_pooling2d (MaxPooling2D)	(None, 56, 56, 64)	0	activation[0][0]
conv2d_1 (Conv2D)	(None, 56, 56, 64)	36,928	max_pooling2d[0][0]
batch_normalization_1 (BatchNormalization)	(None, 56, 56, 64)	256	conv2d_1[0][0]
activation_1 (Activation)	(None, 56, 56, 64)	0	batch_normalization_1…
conv2d_2 (Conv2D)	(None, 56, 56, 64)	36,928	activation_1[0][0]
batch_normalization_2 (BatchNormalization)	(None, 56, 56, 64)	256	conv2d_2[0][0]
add (Add)	(None, 56, 56, 64)	0	batch_normalization_2… max_pooling2d[0][0]
activation_2 (Activation)	(None, 56, 56, 64)	0	add[0][0]
conv2d_3 (Conv2D)	(None, 56, 56, 64)	36,928	activation_2[0][0]
batch_normalization_3 (BatchNormalization)	(None, 56, 56, 64)	256	conv2d_3[0][0]

Fig. 4. Model ResNet

E. Model training

In this training process, a series of experiments are conducted to identify the model that is effective in classifying the classes contained in the dataset. After undergoing several iterations, a combination of hyperparameters that produces optimal accuracy is finally obtained. The models used in this training are ConvMixer and ResNet. The following table summarizes the hyperparameters applied to the 2 models in this training:

Table 2. Hyperparameter

No	*Hyperparameter*	*Value*
1.	Optimizer	Adam
2.	Batch size	64
3.	Learning Rate	0.001
4.	Epoch	40
5.	Input shape	224x224

The hyperparameters used to train the ConvMixer and ResNet models on the gastrointestinal disease classification task include the Adam optimizer, which is able to adaptively adjust the learning rate to support faster convergence on datasets with complex feature variations. The batch size of 64 ensures training efficiency without sacrificing stability, while the learning rate of 0.001 is chosen as the optimal initial value for effective parameter updates. With 40 epochs, both models have enough iterations to learn patterns without the risk of overfitting. ConvMixer, with an input shape of 224x224, utilizes patch embedding to capture local patterns and channel mixer to understand global interactions between features, making it efficient in feature extraction through large kernel convolutions. Meanwhile, ResNet with the same residual architecture and input shape overcomes the vanishing gradient problem and captures hierarchical patterns in images, which is important for recognizing visual characteristics such as mucosal texture or tissue changes in gastrointestinal diseases.

1. ConvMixer

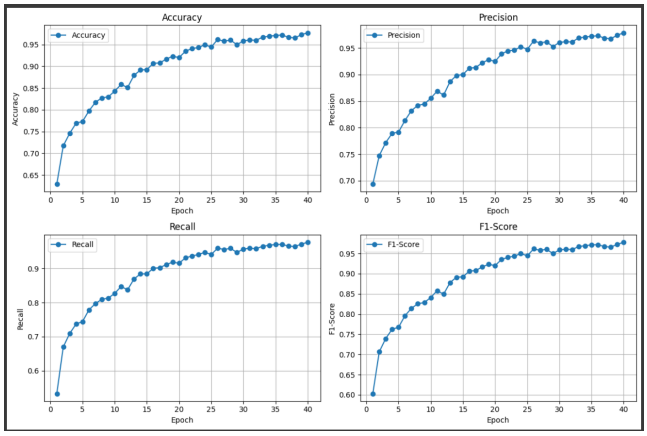

Fig. 5. ConvMixer training chart

The graph in the figure above shows a consistent improvement in all model evaluation metrics of accuracy, precision, recall, and F1-score over 40 training epochs. Accuracy increased from around 0.65 to over 0.95, while precision and recall each reached values close to 0.96, indicating the model was increasingly accurate in recognizing and classifying positive data. F1-score, as a balance between precision and recall, also experienced a similar trend, reaching more than 0.95. These results indicate that the model learned well and achieved optimal performance at the end of training.

2. ResNet

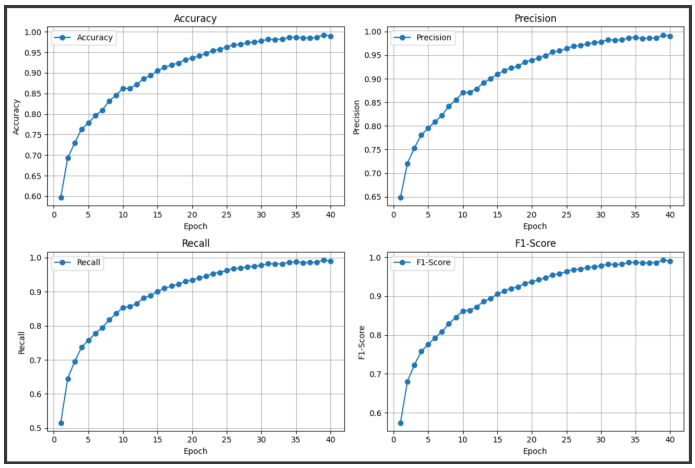

Fig. 6. ResNet training graph

The graph shows the improvement of the model's performance based on the four-evaluation metrics of accuracy, precision, recall, and F1-score, over 40 training epochs. Accuracy increased from around 0.7 to close to 1.0, indicating a very high ability of the model to predict correctly. Precision and recall also showed similar trends, each reaching values close to 1.0, indicating the accuracy and ability of the model to consistently recognize positive data. F1-score also increased significantly to near perfect values. This graph illustrates that the model achieved optimal performance with a good convergence rate at the end of training.

F. Model Evaluasi

In this study, ConvMixer and ResNet models were tested in classifying and detecting gastrointestinal diseases. Test groups were used to evaluate the ConvMixer and ResNet models. Accuracy, precision, recall, and F1 score are the metrics used to evaluate the performance of the models [10]. The results are presented in the form of a confusion matrix as below:

a. *ConvMixer*

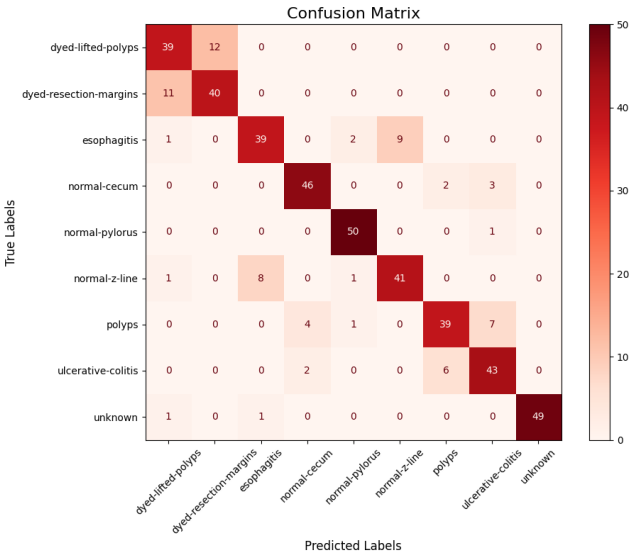

Fig. 7. Confusion matrix ConvMixer

Below are the steps to calculate the value generated through the confusion matrix.

1. Accuracy

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

$$\text{Accuracy} = \frac{39+40+39+46+50+41+39+43+49}{459} = \frac{386}{459}$$

$$\text{Accuracy} = \mathbf{0{,}84\ (84\%)}$$

2. Precision, recall and f1-score

$$\text{Precision} = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad F1\text{-}Score = 2x\frac{Recall\ x\ Presisi}{Recall + Presisi}$$

The following is the calculation of precision, recall and f1-score for each class:

- *Dyed-lifted-polyps*

  $\text{Precision} = \frac{39}{39+14} = 74\%$ $\qquad Recall = \frac{39}{39+12} = 76\%$

  $F1\text{-}Score = 2x\frac{76\ x\ 74}{76+74} = 75\%$

- *Dyed-resection-margins*

  $\text{Precision} = \frac{40}{40+11} = 78\%$ $\qquad Recall = \frac{40}{40+12} = 77\%$

  $F1\text{-}Score = 2x\frac{77x\ 78}{77+78} = 77\%$

- *Esophagitis*

  $\text{Precision} = \frac{39}{39+10} = 0{,}80\%$ $\qquad Recall = \frac{39}{39+11} = 78\%$

  $F1\text{-}Score = 2x\frac{78\ x\ 80}{78+80} = 79\%$

- *Normal-cecum*

  $\text{Precision} = \frac{46}{46+6} = 88\%$ $\qquad Recall = \frac{46}{46+5} = 90\%$

  $F1\text{-}Score = 2x\frac{90\ x\ 88}{90+88} = 89\%$

- *Normal-pylorus*

  $\text{Precision} = \frac{50}{50+2} = 96\%$ $\qquad Recall = \frac{50}{50+3} = 94\%$

  $F1\text{-}Score = 2x\frac{94\ x\ 96}{94+96} = 95\%$

- *Normal-z-line*

  $\text{Precision} = \frac{41}{41+10} = 80\%$ $\qquad Recall = \frac{41}{41+9} = 82\%$

$$F1\text{-}Score = 2x\frac{82\ x\ 80}{82\ +80} = 81\%$$

- *Polyps*

  Precision $= \frac{39}{39+11} = 78\%$ $Recall = \frac{39}{39+9} = 81\%$

  $F1\text{-}Score = 2x\frac{81\ x\ 78}{81\ +78} = 79\%$
- *Ulcerative-colitis*

  Precision $= \frac{43}{43+8} = 84\%$ $Recall = \frac{43}{43+12} = 78\%$

  $F1\text{-}Score = 2x\frac{78\ x\ 77}{78\ +77} = 77\%$
- *Unknown*

  Precision $= \frac{40}{40+2} = 95\%$ $Recall = \frac{40}{40+0} = 100\%$

  $F1\text{-}Score = 2x\frac{100\ x\ 95}{100\ +95} = 97\%$

b. ResNet

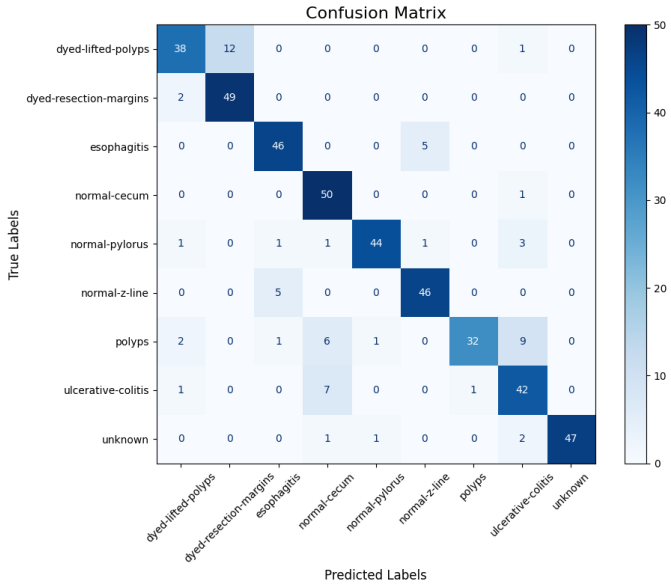

Fig. 8. Confusion matrix ResNet

Below are the steps to calculate the value generated through the confusion matrix.

1. Accuracy

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

   Accuracy $= \frac{38+49+46+50+44+46+32+42+47}{459} = \frac{394}{459}$

   Accuracy $=$ **0,86 (86%)**
2. Precision, *recall* dan *f1-score*

   Precision $= \frac{TP}{TP+FP}$ $Recall = \frac{TP}{TP+FN}$ $F1\text{-}Score = 2x\frac{Recall\ x\ Presisi}{Recall\ +Presisi}$

The following is the calculation of precision, recall and f1-score for each class:

- *Dyed-lifted-polyps*

  Precision $= \frac{38}{38+6} = 86\%$ $Recall = \frac{38}{38+13} = 75\%$

  $F1\text{-}Score = 2x\frac{75\ x\ 86}{75\ +86} = 80\%$
- *Dyed-resection-margins*

  Precision $= \frac{49}{49+2} = 96\%$ $Recall = \frac{49}{49+12} = 80\%$

$F1\text{-}Score = 2x \frac{80\ x\ 96}{80 + 96} = 87\%$

- *Esophagitis*

  Precision $= \frac{46}{46+7} = 86\%$ $Recall = \frac{46}{46+5} = 91\%$

  $F1\text{-}Score = 2x \frac{91\ x\ 86}{91 + 86} = 88\%$

- *Normal-cecum*

  Precision $= \frac{50}{50+15} = 77\%$ $Recall = \frac{50}{50+1} = 98\%$

  $F1\text{-}Score = 2x \frac{98\ x\ 77}{98 + 77} = 86\%$

- *Normal-pylorus*

  Precision $= \frac{44}{44+5} = 90\%$ $Recall = \frac{44}{44+4} = 92\%$

  $F1\text{-}Score = 2x \frac{92\ x\ 90}{92 + 90} = 91\%$

- *Normal-z-line*

  Precision $= \frac{46}{46+5} = 91\%$ $Recall = \frac{46}{46+6} = 88\%$

  $F1\text{-}Score = 2x \frac{88\ x\ 91}{88 + 91} = 89\%$

- *Polyps*

  Precision $= \frac{32}{32+11} = 74\%$ $Recall = \frac{32}{32+9} = 78\%$

  $F1\text{-}Score = 2x \frac{78\ x\ 74}{78 + 74} = 76\%$

- *Ulcerative-colitis*

  Precision $= \frac{42}{42+11} = 79\%$ $Recall = \frac{42}{42+14} = 75\%$

  $F1\text{-}Score = 2x \frac{75\ x\ 79}{75 + 79} = 77\%$

- *Unknown*

  Precision $= \frac{47}{47+4} = 92\%$ $Recall = \frac{47}{47+0} = 100\%$

  $F1\text{-}Score = 2x \frac{100\ x\ 92}{100 + 92} = 96\%$

IV. Conclusion

1. The implementation and evaluation results of the ConvMixer and ResNet deep learning models show that both models perform well in detecting and classifying gastrointestinal diseases using endoscopy images from the Kvasir dataset. ResNet achieved 86% accuracy, slightly higher than ConvMixer which recorded 84% accuracy.
2. The similarity of visual patterns in the images in the Kvasir dataset poses a challenge for the ConvMixer and ResNet models in detecting gastrointestinal diseases. Similar visual structures across classes can make it difficult for the models to distinguish the unique features of each disease class, potentially decreasing accuracy.

References

[1] N. I. Firdausi, "Implementasi Deep Learning Dalam Mendeteksi Penyakit Menggunakan Convolutional Neural Network Dan Tensorflow," *Kaos GL Derg.*, vol. 8, no. 75, pp. 147–154, 2020.

[2] Y. Pantha, "Enhancing Diagnostic Accuracy in Medicine Using Artificial Intelligence: a Deep Learning Approach," *J. Enhanc. Heat Transf.*, no. November, 2024.

[3] H. Laçi, K. Sevrani, and S. Iqbal, "Deep learning approaches for classification tasks in medical X-ray, MRI, and ultrasound images: a scoping review," *BMC Med. Imaging*, vol. 25, no. 1, 2025.

[4] C. Li, L. Li, and J. Shi, "Gastrointestinal endoscopy in early diagnosis and treatment of gastrointestinal tumors," *Pakistan J. Med. Sci.*, vol. 36, no. 2, pp. 203–207, 2020.

[5] F. F. Youssef, L. L. Branch, M. Kowalczyk, and T. J. Savides, "Endoscopic Approaches for Managing

Small Intestinal Disease," *Annu. Rev. Med.*, vol. 76, pp. 155–165, 2025.

[6] A. Ali, A. Iqbal, S. Khan, N. Ahmad, and S. Shah, "A two-phase transfer learning framework for gastrointestinal diseases classification," *PeerJ Comput. Sci.*, vol. 10, pp. 1–31, 2024.

[7] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," *Proc. 8th ACM Multimed. Syst. Conf. MMSys 2017*, pp. 164–169, 2017.

[8] Z. Muthiah, O. Triananda, and M. Iqbal, "Deep Learning Based Classification of TB and Normal Chest X-rays Using a Custom CNN with Minimal Epoch Training," vol. 10, no. 2, pp. 272–279, 2025.

[9] S. A. El-Ghany, M. A. Mahmood, and A. A. Abd El-Aziz, "An Accurate Deep Learning-Based Computer-Aided Diagnosis System for Gastrointestinal Disease Detection Using Wireless Capsule Endoscopy Image Analysis," *Appl. Sci.*, vol. 14, no. 22, 2024.

[10] A. A. Demirbaş, H. Üzen, and H. Fırat, "Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset," *Heal. Inf. Sci. Syst.*, vol. 12, no. 1, 2024.