Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 157-167. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1702

Endoscopic Image Classification Using ConvNeXt for GERD and Polyp Identification

Muhammad Faqih, Okta Qomaruddin Aziz, Ajib Hanani

Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University Maulana Malik Ibrahim Malang, Malang, Indonesia

E-mail: 220605110069@student.uin-malang.ac.id

Abstract

Early and accurate detection of gastrointestinal abnormalities, such as gastroesophageal reflux disease (GERD) and intestinal polyps, is essential for preventing severe clinical complications. However, manual interpretation of endoscopic images is often constrained by inter-observer variability and time limitations. This study proposes a ConvNeXt-Tiny-based deep learning framework for multi-class classification of gastrointestinal endoscopic images. Experiments were conducted using the GastroEndoNet v3 dataset, which contains 4,006 images categorized into four classes: GERD, GERD Normal, Polyp, and Polyp Normal. A total of twelve experimental scenarios were designed to systematically evaluate the effects of dataset-provided augmentation, ImageNet-based normalization, and batch size on model performance. The optimal configuration, combining augmentation, normalization, and a batch size of 64, achieved a test accuracy of 99.75% and a macro-averaged F1-score of 0.9977, indicating stable convergence and strong generalization on unseen data. The results demonstrate that ConvNeXt-Tiny effectively captures disease-relevant visual patterns in endoscopic images while maintaining consistent performance across varying training conditions. Comparative evaluation with a transformer-based baseline further indicates that modern convolutional architectures remain competitive for gastrointestinal image classification tasks. The proposed framework establishes a reliable and lightweight baseline for automated gastrointestinal disease detection. Extensions to video-based endoscopy would require incorporating temporal information across consecutive frames, which is beyond the scope of the current image-based study.

Keywords: *ConvNeXt, Endoscopic Image Classification, Medical Imaging, CNN*

1. Introduction

Gastroesophageal reflux disease (GERD) and intestinal polyps are two common gastrointestinal conditions that may progress into severe complications if not detected early. GERD is a chronic disorder characterized by the reflux of stomach contents into the esophagus, resulting in symptoms such as epigastric pain (heartburn), nausea, acid regurgitation, and dysphagia. If left untreated, GERD can lead to esophagitis and Barrett's esophagus, which may develop into esophageal cancer [1]. Although its prevalence in Asia is lower than in Western countries, recent studies show a concerning increase; for instance, in Japan and Taiwan, the prevalence of GERD has reached approximately 13–15% [1]. Intestinal polyps, on the other hand, are abnormal tissue growths in the colon wall that often remain asymptomatic in their early stages. Approximately 35% of colorectal cancer cases in Indonesia originate from untreated polyps, and most occur among individuals under 40 years of age [2]. Early detection of both conditions is therefore critical, as prompt treatment is more effective in preventing disease progression, especially in healthcare environments with limited specialist availability.

Endoscopy serves as a minimally invasive medical procedure that allows direct visualization of the upper and lower gastrointestinal tract using a flexible camera, making it the primary diagnostic tool for disorders such as GERD and polyps. The resulting endoscopic images contain crucial visual cues regarding mucosal abnormalities. However, manual interpretation of these images heavily depends on clinician expertise and is prone to inter-observer variability and human error, which can result in delayed or inaccurate diagnosis [3]. Moreover, the visual complexity of endoscopic imagery such as subtle tissue color variations, uneven illumination, and

imaging artifacts adds further challenges to manual assessment [20]. These limitations highlight the need for automated, reliable, and efficient diagnostic tools that can assist physicians in clinical decision-making.

Recent advances in artificial intelligence (AI), particularly deep learning, have demonstrated promising capabilities in automating medical image analysis. Deep learning models based on convolutional neural networks (CNNs) can learn discriminative visual features directly from large datasets, enabling consistent and rapid recognition of pathological patterns. By leveraging thousands of endoscopic images, AI-based classification systems can perform diagnosis more efficiently and with higher consistency than manual evaluation, potentially improving both diagnostic accuracy and healthcare service quality. Traditional CNN architectures such as ResNet and EfficientNet have shown strong performance in image classification tasks, yet their computational demands often make them impractical for deployment in resource-constrained hospitals [4], [18].

To address this limitation, the present study employs ConvNeXt-Tiny, a modern convolutional architecture that integrates the efficiency of CNNs with design principles derived from Vision Transformers. ConvNeXt-Tiny achieves a balance between performance and computational efficiency, showing superior results on major computer vision benchmarks, including ImageNet classification, object detection, and semantic segmentation [5]. Benchmark results on ImageNet-1K and ImageNet-22K indicate that ConvNeXt variants consistently outperform ResNet models and rival or surpass Vision Transformers such as Swin-B, offering higher accuracy and faster inference throughput while maintaining architectural simplicity [5]. These characteristics make ConvNeXt particularly suitable for medical applications that demand high precision within limited computational capacity [6]. Furthermore, its modular and lightweight structure facilitates flexible fine-tuning on small to medium-sized medical datasets, enabling efficient model adaptation for specific diagnostic tasks [7], [8].

This research utilizes the GastroEndoNet: Comprehensive Endoscopy Image Dataset for GERD and Polyp Detection [9] to develop and evaluate an endoscopic image classification model based on the ConvNeXt-Tiny architecture. The model is trained to automatically detect GERD and intestinal polyps, with performance evaluated through standard medical image classification metrics, including the confusion matrix, accuracy, sensitivity, specificity, and F1-score [6]. The primary objective of this study is to construct an accurate and computationally efficient ConvNeXt-Tiny–based model capable of supporting automatic endoscopic image classification for GERD and polyp detection. The proposed system is expected to enhance diagnostic accuracy and efficiency, particularly in healthcare facilities with limited access to gastroenterology specialists.

2. Related Works

Recent years have witnessed remarkable progress in deep-learning-based endoscopic analysis. Jha et al. [10] introduced ColonSegNet for real-time polyp segmentation on the Kvasir-SEG dataset (Dice 0.82; 180 FPS), outperforming YOLOv4 and Faster R-CNN. Cao et al. [11] improved small-polyp detection using YOLOv3 with multi-level feature fusion, achieving 91.6% precision and 88.8% F1-score. For upper-GI diseases, Chan et al. [12] employed DenseNet121 with a Global Attention Block for GERD grading (accuracy 74.69%, F1 69.87%), emphasizing interpretability through attention maps. Wang et al. [13] applied the Vision Transformer (ViT) to colonoscopic datasets, obtaining 4–6% higher accuracy than CNN baselines.

Beyond endoscopy, modern convolutional backbones have also been explored in other medical imaging domains. Nergiz [15] tested four ConvNeXt variants on the MHIST dataset, where ConvNeXt-L achieved 88.9% accuracy and 91.21% F1-score, outperforming ResNet and DenseNet even in few-shot settings. Huan and Dun [14] proposed MSMP-Net, a multi-scale ConvNeXt backbone for monkeypox classification, reaching 87.03% accuracy with efficient end-to-end inference.

Masum et al. [16] introduced an explainable deep-ensemble transformer framework that combines the Vision Transformer (ViT) and Swin Transformer for multi-class gastrointestinal disease classification. Their study utilized the GastroEndoNet dataset, comprising 4,006 endoscopic images across four categories: GERD, GERD Normal, Polyp, and Polyp Normal. The ensemble model integrated the complementary strengths of both architectures, achieving balanced F1-scores across all classes and an overall accuracy of 87%. The approach demonstrated the ability of transformer-based models to capture global contextual relationships within endoscopic images, while incorporating explainable AI techniques such as Grad-CAM and Grad-CAM++ for clinical interpretability. This work represents one of the earliest efforts applying transformer ensembles to the

GastroEndoNet dataset and provides a relevant baseline for subsequent studies exploring alternative architectures such as modernized convolutional networks.

Collectively, these studies reveal a clear trajectory: from handcrafted features to CNNs, and more recently toward hybrid ConvNeXt- and Transformer-based frameworks. Nevertheless, most prior works focus on a single disease domain either GERD or polyps and rarely explore multi-class endoscopic classification under limited computational settings. Moreover, the application of ConvNeXt to upper-GI endoscopy remains under-investigated despite its proven robustness in other medical fields. Therefore, this research introduces a ConvNeXt-Tiny–based lightweight framework for endoscopic image classification encompassing both GERD and intestinal polyps. The approach aims to fill this gap by providing an accurate, efficient, and interpretable model tailored for real-world medical environments with constrained resources.

3. Method

3.1. Research Design

This study adopts an experimental research design aimed at developing and evaluating a deep learning–based classification model for the automatic identification of gastroesophageal reflux disease (GERD) and intestinal polyps from endoscopic images. The overall workflow of the research is illustrated in Figure 1, which consists of five main stages: data acquisition, preprocessing, model development, classification, and performance evaluation.

The process begins with the acquisition of the GastroEndoNet dataset, a publicly available endoscopic image collection comprising four diagnostic categories: GERD, GERD-Normal, Polyp, and Polyp-Normal. The dataset is divided into training, validation, and testing subsets to ensure unbiased model evaluation. Each image undergoes preprocessing operations, including resizing to 224×224 pixels, rescaling pixel values to the [0,1] range, and normalization using the ImageNet mean and standard deviation.

The preprocessed images are then fed into the ConvNeXt-Tiny architecture, which serves as the primary feature extractor. This modernized convolutional neural network incorporates depthwise convolution, Layer Normalization, and GELU activation functions to efficiently learn hierarchical visual representations from the endoscopic imagery. The extracted features are subsequently passed through a fully connected classification layer with a Softmax activation function to categorize each image into one of the four diagnostic labels.

Finally, the trained model is evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score, to quantitatively assess its diagnostic capability. This experimental design ensures a reproducible, end-to-end workflow that bridges raw medical imaging data and automated disease classification through a robust and computationally efficient ConvNeXt-based framework.

3.2. Dataset Description

This study utilizes the GastroEndoNet dataset, a publicly available endoscopic image collection introduced by Bitto et al. (2025) GastroEndoNet [9]. The dataset was developed to support computer-aided diagnosis (CAD) research for Gastroesophageal Reflux Disease (GERD) and gastric polyps. It contains a total of 24,036 high-quality endoscopic images, categorized into four classes: GERD, GERD-Normal, Polyp, and Polyp-Normal.

The dataset comprises 4,006 primary images collected from endoscopic procedures at Zainul Haque Sikder Women's Medical College & Hospital in Dhaka, Bangladesh, and was expanded through six augmentation techniques including rotation (60° and 90°), horizontal flipping, zooming, brightness adjustment, and shearing to increase the total number of images to 24,036. Each image was resized to 512 × 512 pixels in JPG format to ensure uniformity and facilitate model training.

In this study, the classification task is formulated as a single-label multi-class classification problem involving four mutually exclusive categories: GERD, GERD-Normal, Polyp, and Polyp-Normal. Each endoscopic image is assigned exactly one ground-truth label, and no image belongs to multiple classes simultaneously. Therefore, the task is not treated as multi-label classification, and GERD and polyp conditions are modeled as separate and independent diagnostic classes within a unified four-class framework.

Table 1. Category-wise data distribution.

Category	Original Images	Augmented Images	Total
GERD	974	4,870	5,844
GERD Normal	1,103	5,515	6,618
Polyp	779	3,895	4,674
Polyp Normal	1,150	5,750	6,900
Total	**4,006**	**20,030**	**24,036**

The class-wise distribution of both original and augmented images is presented in Table 1, while Table 2 describes each diagnostic category with representative endoscopic visuals.

3.3. System Architecture

The proposed endoscopic image classification system is designed as a modular, end-to-end framework that integrates data preprocessing, hierarchical feature extraction, classification, and evaluation stages. The overall architecture is illustrated in Figure 2, which shows the sequential flow from the GastroEndoNet dataset to the final classification output. The system consists of five main components: (1) input acquisition, (2) preprocessing, (3) feature extraction using the ConvNeXt-Tiny backbone, (4) classification through a fully connected Softmax layer, and (5) performance evaluation. Each stage is constructed to ensure diagnostic robustness, computational efficiency, and reproducibility.

Raw endoscopic images from the GastroEndoNet dataset are standardized to ensure consistent dimensions and illumination profiles. Each image is resized to 224 × 224 pixels, conforming to the ConvNeXt-Tiny input specification. A per-channel normalization is applied using the ImageNet mean and standard deviation values ($\mu = [0.485,0.456,0.406], \sigma = [0.229,0.224,0.225]$) to align pixel intensity distributions with the pretrained model domain. To enhance generalization, repository-provided augmented dataset using horizontal flipping, mild rotation, and brightness adjustment is employed to simulate real-world variability in endoscopic imaging conditions. These steps mitigate dataset bias and improve the model's sensitivity to subtle pathological cues such as mucosal erythema or small polyps.

In experimental scenarios where normalization is disabled, this ImageNet-based per-channel normalization step is omitted. No dataset-specific z-score normalization or contrast normalization is applied in this study. The normalization setting is controlled solely at the scenario level to evaluate its impact on training stability and generalization.

The ConvNeXt-Tiny backbone serves as the core feature extractor, modernizing the traditional ResNet structure through Transformer-inspired principles while retaining the convolutional inductive bias. It is composed of four hierarchical stages with progressively reduced spatial resolutions and increased channel depths, and strictly follows the default architectural configuration defined in [5]. The channel dimensions across the four stages are set to [96, 192, 384, 768], with the corresponding block distributions per stage adopted without structural modification. This configuration ensures compatibility with pretrained ImageNet weights and provides a standardized backbone setting, with no manual adjustment of stage width or depth hyperparameters performed in this work.

Each stage contains multiple ConvNeXt blocks, combining depthwise convolution for spatial filtering, Layer Normalization for feature scaling, GELU activation for smooth non-linearity, and 1×1 pointwise projections with residual connections for stable optimization. This hierarchical design captures both local textural patterns and high-level semantic cues crucial for differentiating between GERD-induced lesions and benign mucosal structures.

After feature extraction, Global Average Pooling (GAP) condenses the spatial feature maps into a compact descriptor vector. This vector passes through a fully connected layer to produce four logits representing the diagnostic categories. A Softmax activation transforms these logits into normalized probabilities, providing interpretable predictions consistent with clinical classification standards.

Predicted labels are compared with ground-truth annotations from the test subset using standard performance metrics: accuracy, precision, recall, and F1-score. These metrics collectively assess diagnostic reliability, class-wise sensitivity, and robustness to class imbalance. This architecture unifies the efficiency of convolutional processing with modern Transformer-like design cues, achieving a balance between precision and computational feasibility. The macro-level system flow is visualized in Figure 2, while the micro-level composition of the ConvNeXt block is elaborated mathematically in Section D.

3.4. ConvNeXt-Tiny Feature Extraction and Classification

The ConvNeXt-based classification framework serves as the core component of the proposed endoscopic image analysis system. It modernizes the traditional convolutional neural network (CNN) paradigm by incorporating several architectural refinements inspired by Vision Transformers (ViT), such as large-kernel convolutions, Layer Normalization, and the Gaussian Error Linear Unit (GELU) activation,

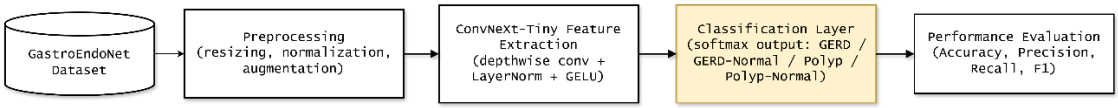


Figure 1. Overall research workflow.

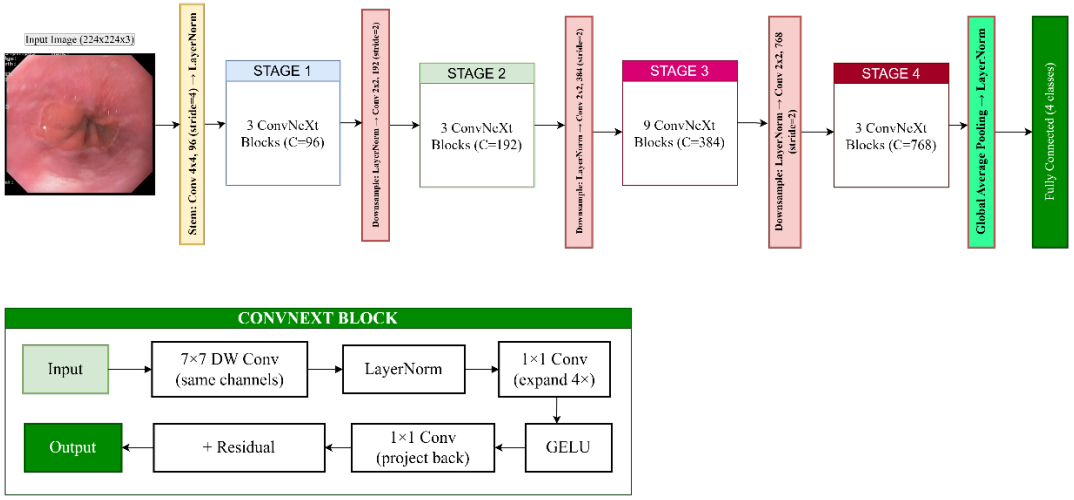
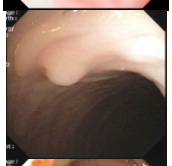
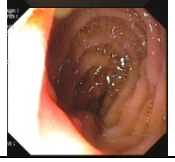


Figure 2. Proposed system architecture for ConvNeXt-based endoscopic image classification adapted from [5].

Table 2. Description of the GastroEndoNet dataset with representative samples.

Class name	Description	Visualization
GERD	This class includes endoscopic images showing reflux-related inflammation in the lower esophagus. Typical visual cues comprise erythematous mucosa, mucosal breaks, irregular Z-line structure, and other erosive manifestations associated with acid exposure. These findings represent the pathological spectrum of gastroesophageal reflux disease.	
GERD-Normal	Images in this class display the normal esophageal lining with a smooth, glistening surface and well-defined vascular patterns. No erosions or inflammatory changes are visible, making these images suitable as negative controls for reflux classification tasks.	
Polyp	The Polyp category contains images of abnormal protrusions on the gastric or colonic mucosa. Polyps vary in appearance—ranging from small sessile bumps to larger pedunculated lesions—and often exhibit distinct color or texture contrasts that signal neoplastic potential.	
Polyp-Normal	This class represents healthy gastric mucosa without visible masses or protrusions. The surface appears uniform and free of structural irregularities, providing reference examples for distinguishing normal anatomy from polypoid growths.	

while maintaining the computational efficiency and inductive bias of convolutional operations. The detailed structure of the ConvNeXt block previously depicted in the lower section of Figure 2 forms the core of the proposed system architecture.

At the input stage, each endoscopic image $I \in R^{H\times W\times C}$ is transformed through a stem convolutional layer into an initial feature representation. This process effectively divides the image into small spatial patches and encodes their local texture and color information. The first major operation in each ConvNeXt block is the depthwise convolution, which processes each channel independently to capture spatial dependencies while preserving channel-specific information, as defined in equation (1).

$$Y_{dw}(x,y,c) = \sum_{i=-k}^{k}\sum_{j=-k}^{k} W_{dw}(i,j,c)\cdot X(x+i,y+j,c) \quad (1)$$

where W_{dw} represents the depthwise kernel of size $(2k+1) \times (2k+1)$, and (X) denotes the input feature map. Unlike standard convolution, this formulation greatly reduces the number of trainable parameters by removing inter-channel mixing at this stage. In medical imaging contexts such as endoscopy, depthwise convolution helps isolate subtle local features such as fine mucosal texture, vascular patterns, or small erosions without oversmoothing them across feature channels.

Following this, a pointwise convolution (1×1) is applied to mix channel-wise information and enhance feature expressiveness. The operation can be formulated in equation (2).

$$Y_{pw} = W_{pw} * Y_{dw} + b \tag{2}$$

where W_{pw} represents the 1×1 convolution weight matrix and (b) is the bias term.

Together, these two operations form a depthwise separable convolution, a design that balances accuracy and efficiency, reducing redundancy in inter-channel processing.

To stabilize feature distribution during training, the network replaces the conventional Batch Normalization with Layer Normalization (LN), which computes normalization independently for each sample and is thus less sensitive to batch size variation. The transformation is expressed as equation (3).

$$\mathrm{LN}(x_i) = \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \cdot \gamma + \beta \tag{3}$$

where μ_i and σ_i denote the mean and standard deviation of features in sample i, and γ and β are learnable affine parameters. LN improves numerical stability and allows consistent convergence even with small mini-batches, a common limitation in GPU-constrained medical imaging environments.

The non-linear activation used in ConvNeXt is the Gaussian Error Linear Unit (GELU), which introduces a smooth probabilistic gating behavior and outperforms ReLU in gradient propagation and model expressivity. GELU is defined in equation (4).

$$\mathrm{GELU}(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{4}$$

Unlike ReLU, which discards negative activations, GELU softly weights them according to their probability under a Gaussian distribution. This property improves gradient continuity and yields smoother feature manifolds, which is crucial when differentiating between visually similar classes such as mild GERD and normal mucosa.

Each ConvNeXt block further employs an inverted bottleneck structure, consisting of two successive 1×1 convolutions that expand and then project feature dimensions. The transformation can be described in equation (5).

$$Z = W_2 \cdot \mathrm{GELU}\big(W_1 \cdot \mathrm{LN}(Y_{dw})\big) + b \tag{5}$$

where W_1 and W_2 correspond to the expansion and projection matrices, respectively, respectively, as defined in equation (6). This design enhances the model's ability to represent complex feature interactions without substantially increasing the number of parameters. The resulting tensor (Z) is then added back to the input via a residual connection to promote stable gradient flow and facilitate optimization of deeper layers.

$$Y_{out} = Y_{in} + Z \tag{6}$$

Residual learning allows ConvNeXt to maintain performance stability even under deeper hierarchies, effectively mitigating vanishing-gradient problems while preserving low-level spatial cues from earlier layers.

Once features have passed through all four ConvNeXt stages, they are summarized using Global Average Pooling (GAP), which computes the mean activation across spatial dimensions for each channel, as formulated in equation (7).

$$f_c = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} F(x, y, c) \tag{7}$$

This operation condenses the information into a compact descriptor vector $f = [f_1, f_2, \ldots, f_C]$, which is subsequently passed into a fully connected layer producing four logits corresponding to the diagnostic categories GERD, GERD-Normal, Polyp, and Polyp-Normal. These logits are converted into probabilities through a Softmax activation as presented in equation (8).

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{4} e^{z_j}} \tag{8}$$

The model is trained end-to-end using the categorical cross-entropy loss, optimized via Adam to ensure stable and efficient convergence. A learning-rate scheduler and early stopping criterion are employed to promote generalization and prevent overfitting. By integrating these elements, the ConvNeXt-based framework effectively balances local spatial precision and global contextual reasoning two characteristics

essential for accurate endoscopic diagnosis. Its combination of depthwise separable convolution, LayerNorm, GELU, and residual mapping yields a lightweight yet powerful classifier capable of identifying pathological tissue patterns with both efficiency and interpretability.

3.5. Evaluation Metrics

The performance of the proposed model is evaluated using standard classification metrics, including the confusion matrix, accuracy, precision, recall, and F1-score, which are interpreted in the context of medical image diagnosis. The confusion matrix provides a comprehensive view of the model's predictions by comparing true and predicted diagnostic labels, enabling the identification of specific misclassification patterns among endoscopic image categories. Accuracy measures the overall proportion of correctly classified endoscopic images across all diagnostic classes. Precision reflects how reliably the model's predicted disease labels, such as GERD or polyp cases, correspond to true pathological findings. Recall (sensitivity) assesses the model's ability to correctly detect images that truly contain disease-related abnormalities, which is particularly important in medical applications to minimize missed pathological cases. Finally, the F1-score, defined as the harmonic mean of precision and recall, provides a balanced evaluation of the model's diagnostic consistency, especially under class imbalance conditions. Collectively, these metrics reflect the reliability and clinical relevance of the proposed model as a computer-aided diagnostic support tool.

4. Results

4.1. Experimental Configuration

All experiments were conducted using Google Colab equipped with an NVIDIA Tesla T4 GPU (16 GB). The implementation was carried out in the PyTorch framework using the Adam optimizer with an initial learning rate of 0.001 and the Cross-Entropy Loss function, which is suitable for multi-class classification tasks. Each training run was executed for a maximum of 20 epochs with an early-stopping mechanism (patience = 10) based on validation loss to prevent overfitting and unnecessary computation. To ensure reproducibility, a fixed random seed of 42 was applied across all experiments.

The experiments utilized the GastroEndoNet dataset (version 3) [9], which consists of endoscopic images categorized into four classes: GERD, GERD Normal, Polyp, and Polyp Normal. The dataset provides both original and with augmented image versions. To prevent potential data leakage, data splitting was performed exclusively on original images. Specifically, the 4,006 original images were divided into 70% for training (2,804 images), 20% for validation (801 images), and 10% for testing (401 images). All validation and test samples consisted solely of original images. The 24,036 augmented images were included only in the training set when augmentation was used, and never appeared in validation or testing subsets.

All images were resized to 224 × 224 pixels to match the ConvNeXt-Tiny input specification. When enabled, per-channel ImageNet normalization was applied using the standard mean and standard deviation values to align input distributions with the pretrained model domain.

4.2. Experimental Scenarios and Quantitative Results

To systematically evaluate the effects of augmented data, normalization, and batch size on model performance, twelve experimental scenarios (A–L) were designed. The scenarios followed a factorial structure combining two binary factors—augmentation (with/without) and normalization (with/without)—with three batch sizes (16, 32, and 64). This design enables an explicit analysis of both individual and interaction effects of preprocessing and optimization parameters on learning stability and generalization.

All scenarios were trained under identical conditions except for the specified configuration differences. Table 3 summarizes the quantitative results on the test set using macro-averaged accuracy, precision, recall, and F1-score.

The results reveal several consistent and interpretable patterns. Scenarios A–C, which applied both augmented dataset and normalization, achieved the highest and most stable performance across all batch sizes, with F1-scores exceeding 0.99. Scenario C (batch size 64) yielded the best overall result, achieving an accuracy of 99.75% and a macro-averaged F1-score of 0.9977. Increasing the batch size from 16 to 64 consistently improved training stability and generalization under this configuration.

In contrast, disabling normalization while retaining augmented dataset (Scenarios D–F) resulted in highly unstable behavior. Scenarios D and E (batch sizes 16 and 32) exhibited severe performance degradation, with F1-scores

Table 3. Combined summary of experimental configurations and performance metrics (test set, macro-averaged).

Scenario	Aug.	Norm.	Batch	Accuracy	Precision	Recall	F1
A	Yes	Yes	16	0.9950	0.9943	0.9956	0.9949
B	Yes	Yes	32	0.9950	0.9936	0.9955	0.9945
C	Yes	Yes	64	0.9975	0.9978	0.9975	0.9977
D	Yes	No	16	0.2494	0.0623	0.2500	0.0998
E	Yes	No	32	0.2494	0.0623	0.2500	0.0998
F	Yes	No	64	0.9900	0.9905	0.9901	0.9902
G	No	Yes	16	0.7382	0.7383	0.7310	0.7300
H	No	Yes	32	0.7157	0.7347	0.7060	0.7052
I	No	Yes	64	0.7481	0.7430	0.7440	0.7425
J	No	No	16	0.3167	0.2927	0.2940	0.1900
K	No	No	32	0.3990	0.4228	0.3840	0.3051
L	No	No	64	0.6334	0.6786	0.6130	0.6124

below 0.10, indicating disrupted feature scaling and ineffective convergence. Notably, a larger batch size (Scenario F) partially mitigated this instability, recovering performance to an F1-score of 0.9902, suggesting that batch size can compensate for suboptimal preprocessing to a limited extent.

For configurations without augmentation but with normalization enabled (Scenarios G–I), the model demonstrated moderate and consistent generalization, achieving F1-scores in the range of 0.70–0.75. This indicates that normalization alone ensures training stability, but the absence of data diversity limits feature robustness. Finally, scenarios without both augmentation and normalization (Scenarios J–L) showed poor performance overall, although larger batch sizes again contributed to partial recovery.

Aggregated analysis further highlights normalization as the most critical factor for stable convergence, while augmentation significantly enhances generalization when combined with proper normalization. Batch size primarily affects optimization stability and amplifies the benefits of well-configured preprocessing.

4.2. Model Performance Analysis

To further examine the best-performing configuration, Scenario C (augmentation and normalization enabled, batch size = 64) was analyzed in detail. This configuration achieved the highest test-set accuracy of 99.75 % and a macro-averaged F1-score of 0.9977, representing the most stable and well-generalized configuration among all evaluated scenarios

The confusion matrix for the best-performing configuration, Scenario C, is presented in Figure 3. The model demonstrates balanced classification performance across all four categories, namely GERD, GERD Normal, Polyp, and Polyp Normal, with consistently high true-positive counts for each class. Minor misclassifications primarily occur between visually similar class pairs, particularly GERD versus GERD Normal, which reflects subtle variations in mucosal texture and illumination conditions in endoscopic imagery. Overall, the confusion matrix indicates robust class-wise performance without evidence of class imbalance or systematic bias.

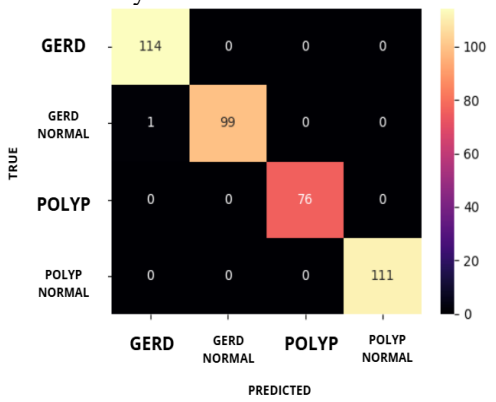


Figure 3. Confusion matrix of Scenario C showing per-class distribution of predictions on the test set.

Figure 4 illustrates the training and validation loss curves for Scenario C. Both curves exhibit a smooth and consistent decrease over 20 epochs, with no significant divergence between training and validation loss. The validation loss begins to plateau after approximately the eighth epoch, confirming the effectiveness of the early-stopping strategy in preventing overfitting and indicating stable convergence. This behavior highlights the contribution of ImageNet-based normalization in maintaining gradient stability and the role of dataset augmentation in improving generalization.

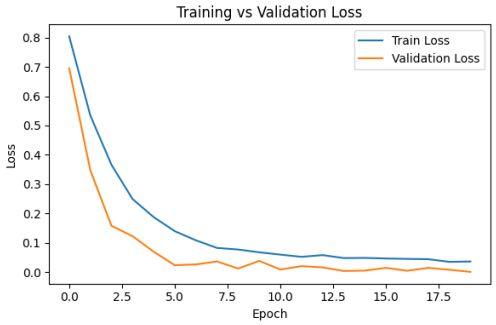


Figure 4. Training and validation loss curves for the best-performing configuration, Scenario C.

5. Discussion

5.1. Variable Influence Analysis

This section analyzes the influence of three key experimental variables, namely data augmentation, normalization, and batch size, on the performance of the ConvNeXt-Tiny model. To isolate the contribution of each variable, the mean and standard deviation of the macro-averaged F1-scores were computed by grouping experimental scenarios that share a constant variable while allowing the remaining factors to vary. The aggregated results are summarized in Table 4.

In terms of data preparation, the inclusion of dataset augmentation generally improved model performance, although its effectiveness was strongly dependent on complementary preprocessing strategies. Scenarios employing augmentation (A–F) achieved a mean F1-score of 69.6% ± 46.2%, whereas configurations without augmentation (G–L) attained a lower mean F1-score of 54.8% ± 24.0%. This performance gap indicates that augmentation contributes to enhanced generalization by exposing the model to a wider range of visual variations commonly observed in endoscopic imagery, such as changes in texture, illumination, and anatomical appearance [17].

Beyond dataset augmentation, input normalization emerged as the most influential factor affecting both convergence stability and overall performance. Configurations with ImageNet-based normalization enabled (A–C, G–I) achieved a mean F1-score of 86.1% ± 14.8%, significantly outperforming non-normalized configurations (D–F, J–L), which recorded a mean F1-score of only 38.3% ± 35.4%. The pronounced performance gap of nearly 48 percentage points underscores the critical role of normalization in aligning input image distributions with pretrained model parameters.

Without normalization, the model exhibited unstable gradient behavior and inconsistent convergence, particularly in small and medium batch settings, as evidenced by the near-collapse performance in scenarios D, E, J, and K. Conversely, normalized configurations demonstrated consistently high performance with substantially lower variance, indicating improved robustness across different batch sizes and augmentation conditions. These observations align with established findings that normalization preserves feature scale consistency and stabilizes early-layer activations in pretrained convolutional architectures, thereby facilitating reliable optimization [18].

In addition to normalization, batch size also played a notable role in training stability and performance consistency. Across all configurations, larger batch sizes generally yielded higher and more stable F1-scores. Scenarios with a batch size of 64 (C, F, I, L) achieved a mean F1-score of 83.6% ± 19.0%, compared to 52.6% ± 40.1% and 50.4% ± 43.0% for batch sizes of 32 and 16, respectively. This trend suggests that larger batches provide smoother gradient estimates, reducing optimization noise and mitigating the adverse effects of suboptimal preprocessing choices [19].

Notably, in the absence of normalization, increasing the batch size partially compensated for training instability. Scenario F, despite lacking normalization, achieved high performance when trained with a batch size of 64, whereas its smaller-batch counterparts collapsed. This behavior indicates that batch size can act as a secondary stabilizing factor, although it cannot fully replace the necessity of proper input normalization.

Ultimately, the results indicate that normalization is the primary determinant of stable convergence, while data augmentation enhances generalization only when combined with appropriate preprocessing. Batch size further modulates these effects by influencing optimization stability. The strongest performance was achieved when all three factors were jointly optimized, confirming that robust endoscopic image classification relies on the synergistic interaction between data preprocessing and training configuration rather than any single parameter in isolation.

5.2. Baseline Comparison

To contextualize the performance of the proposed approach, a comparative evaluation was conducted against a transformer-based baseline, Swin-Tiny, using the same data-splitting protocol and test set. Table 5 summarizes the quantitative comparison between the two models under their respective best-performing configurations.

The proposed ConvNeXt-Tiny framework achieved a macro-averaged F1-score of 0.9977, marginally outperforming the Swin-Tiny baseline, which obtained an F1-score of 0.9953. Although the performance gap is relatively small, the results demonstrate that modernized convolutional architectures can achieve performance comparable to transformer-based models for endoscopic image classification tasks.

Table 4. Summary of Mean ± Std F1-Scores by experimental variable

Variable	Condition	Mean ± Std F1 (%)	Main Interpretation
Augmentation	With augmentation (A–F)	**69.6 ± 46.2**	Improves generalization, but performance highly dependent on normalization; variance inflated by unnormalized cases.
	Without augmentation (G–L)	**54.8 ± 24.0**	Lower and less consistent performance; reduced robustness to data variation.
Normalization	Active (A–C, G–I)	**86.1 ± 14.8**	Critical for training stability and convergence; consistently high F1 across batch sizes.
	Inactive (D–F, J–L)	**38.3 ± 35.4**	Causes severe instability and performance collapse in small/medium batches.
Batch Size	16 (A, D, G, J)	**50.4 ± 43.0**	Small batches highly unstable; sensitive to preprocessing choices.
	32 (B, E, H, K)	**52.6 ± 40.1**	No significant improvement over batch 16; still unstable without normalization.
	64 (C, F, I, L)	**83.6 ± 19.0**	Large batches significantly improve stability and F1, especially with proper normalization.

Table 5. Performance comparison between the proposed ConvNeXt-Tiny model and a transformer-based baseline (Swin-Tiny) evaluated under identical data-splitting protocols.

Model	Architecture Type	Accuracy	Precision	Recall	F1-score
ConvNeXt-Tiny (Best Scenario)	Modernized CNN	0.9975	0.9978	0.9975	0.9977
Swin-Tiny (Baseline)	Vision Transformer	0.995	0.9953	0.9952	0.9953

In addition to competitive accuracy, ConvNeXt-Tiny offers architectural simplicity and reduced computational overhead compared to transformer-based designs. This characteristic may be advantageous in clinical or resource-constrained environments where efficiency and deployment feasibility are critical considerations. Overall, the baseline comparison confirms that the proposed model provides a strong balance between classification performance and architectural efficiency.

5.3. Synthesis and Implications

The overall findings demonstrate that the ConvNeXt-Tiny model provides robust and competitive performance for gastrointestinal endoscopic image classification when trained under appropriate preprocessing and optimization strategies. Using the GastroEndoNet v3 dataset [9], which consists of 4,006 original endoscopic images across four categories, namely GERD, GERD Normal, Polyp, and Polyp Normal, the best-performing configuration (Scenario C) achieved a test accuracy of 99.75% and a macro-averaged F1-score of 0.9977. These results confirm that modernized convolutional architectures can achieve highly reliable classification performance on challenging gastrointestinal imaging tasks.

The performance gains observed in this study are not attributable to architectural complexity alone, but rather to the synergistic interaction between data augmentation, ImageNet-based normalization, and an appropriately large batch size. Together, these factors stabilized gradient flow, reduced training variance, and enhanced generalization to unseen test images containing diverse mucosal textures, illumination conditions, and anatomical variations. This finding reinforces the importance of preprocessing and training configuration in maximizing the effectiveness of pretrained deep learning models in medical imaging applications.

In comparison with transformer-based approaches evaluated on the same dataset, the proposed ConvNeXt-Tiny framework exhibited performance that is comparable to, and slightly higher than, the Swin-Tiny baseline considered in this study. While transformer ensembles reported in prior work, such as Masum et al. [16], emphasize global contextual modeling through self-attention mechanisms, the results here indicate that carefully optimized convolutional networks can achieve similar levels of classification accuracy without relying on complex attention-based architectures. This observation is consistent with recent studies suggesting that architectural choice alone does not guarantee superior performance in medical image classification, particularly when dataset size and visual variability are limiting factors [12], [13].

From a practical perspective, the consistent performance of ConvNeXt-Tiny across multiple experimental conditions highlights its suitability as a lightweight and reliable backbone for computer-aided diagnosis systems in gastrointestinal endoscopy. Rather than positioning the model as a replacement for transformer-based methods, this study demonstrates that modern convolutional designs remain highly competitive when combined with appropriate regularization and training strategies.

6. Conclusion

This study presented a ConvNeXt-Tiny-based framework for automated multi-class classification of gastrointestinal endoscopic images using the GastroEndoNet v3 dataset. Twelve experimental scenarios were evaluated to analyze the effects of data augmentation, normalization, and batch size on convergence and generalization. The best-performing configuration combined dataset-provided augmentation, ImageNet-based normalization, and a batch size of 64, achieving a test accuracy of 99.75% and a macro-averaged F1-score of 0.9977.

The results demonstrate that ConvNeXt-Tiny effectively captures fine-grained mucosal patterns across four diagnostic categories (GERD, GERD Normal, Polyp, and Polyp Normal). Normalization proved critical for training stability, augmentation enhanced generalization when paired with appropriate preprocessing, and batch size influenced optimization robustness particularly without normalization. Compared to a transformer-based baseline under the same protocol, ConvNeXt-Tiny achieved comparable performance, confirming that modernized convolutional architectures remain competitive for endoscopic image classification. These findings underscore the importance of well-designed preprocessing and training strategies over architectural complexity in medical imaging applications.

References

[1] S. Rijal, K. M. Tayibu, I. M. Musa, P. Hapsari, and P. Natsir, "Karakteristik penderita gastroesophageal reflux disease," *Fakumi Medical Journal*, vol. 4, no. 5, pp. 402–411, 2024.

[2] Y. Lafau, Z. Azmi, and A. Calam, "Sistem pakar mendiagnosis polip usus pada manusia menggunakan metode certainty factor," *Jurnal Sistem Informasi TGD*, vol. 3, no. 5, pp. 602–609, 2024.

[3] E. Sivari *et al.*, "A new approach for gastrointestinal tract findings detection and classification: Deep learning-based hybrid stacking ensemble models," *Diagnostics*, vol. 13, no. 4, p. 720, 2023, doi: 10.3390/diagnostics13040720.

[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[5] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 11966–11976.

[6] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] H. C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[8] D. I. Emegano, M. T. Mustapha, I. Ozsahin, D. U. Ozsahin, and B. Uzun, "Advancing prostate cancer diagnostics: A ConvNeXt approach to multi-class classification in underrepresented populations," *Bioengineering*, vol. 12, no. 4, p. 369, 2025.

[9] A. K. Bitto *et al.*, "GastroEndoNet: Comprehensive endoscopy image dataset for GERD and polyp detection," *Data in Brief*, vol. 60, p. 111572, 2025, doi: 10.1016/j.dib.2025.111572.

[10] D. Jha *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[11] C. Cao *et al.*, "Gastric polyp detection in gastroscopic images using deep neural network," *PLOS ONE*, vol. 16, no. 4, e0250632, 2021.

[12] I. N. Chan *et al.*, "Multi-class gastroesophageal reflux disease classification system using deep learning techniques," in *Proc. 10th Int. Conf. Biomed. Bioinform. Eng.*, Kyoto, Japan, 2023.

[13] R. Wang *et al.*, "Transformer-based model for colorectal polyp segmentation in colonoscopy images," *Artificial Intelligence in Gastrointestinal Endoscopy*, vol. 3, no. 1, pp. 12–24, 2024.

[14] E Huan and H. Dun, "MSMP-Net: A multi-scale neural network for end-to-end monkeypox virus skin lesion classification," Applied Sciences, vol. 14, no. 20, p. 9390, 2024. doi: 10.3390/app14209390.

[15] M. Nergiz, "Classification of precancerous colorectal lesions via ConvNeXt on histopathological images," *Balkan Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 129–137, 2023.

[16] A. K. M. A. Masum *et al.*, "An explainable AI based deep ensemble transformer framework for gastrointestinal disease prediction from endoscopic images," *EAI Endorsed Transactions on AI and Robotics*, vol. 4, 2025, doi: 10.4108/airo.9795.

[17] T. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 60, pp. 1–48, 2019.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[19] P. Goyal *et al.*, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[20] S. Ali *et al.*, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Scientific Reports*, vol. 10, no. 1, p. 2748, 2020, doi: 10.1038/s41598-020-59413-