Premier University, Chattogram

Project Report on

Endoscopic Image Analysis for GERD and Polyp Detection

Course Title: Machine Learning Laboratory

Course Code: CSE 458

Submitted To:

Md Tamim Hossain
Lecturer
Department of CSE, PUC

Submitted By:

Farzana Nasir Barsha
ID: 2104010202292
Shabah Tasfiya
ID: 2104010202294
Abdullah Al Rohan
ID: 2104010202288

Semester: 8th, Batch: 40th, Sec: D
Session: Spring 2025
Submission Date: 23rd November, 2025

Department of CSE
Premier University, Chattogram

Abstract—In this paper, we introduce *GastroEndoNet*, a large-scale gastrointestinal endoscopy image dataset for aiding automatic GERD and colorectal polyps detection and classification. The dataset is composed over a total of 24,036 images with the original samples (4,006) augmented using six transformation strategies—rotation, flipping, shifting, zooming, brightness and noise addition—to increase diversity as well as to limit overfitting and improve model generalization. The data are classified into four clinics informative classes: GERD, GERD Normal, Polyp and Polyp normal for holistic analysis of both upper and lower GI abnormalities.

We compare the performance of the four approached to provide state-of-the-art for lung lobe segmentation in order to benchmark the dataset and develop a reliable computer-aided diagnostic framework where we test six Convolutional neural network (CNN) architectures: EfficientNet-B0, Resnet50, MobileNetV2, InceptionV3, DenseNet121 and ConvNeXt-Tiny. Typical pre-processing, including image rescaling, normalization and focused augmentation was followed to overcome class imbalance and cope with the varying imaging conditions caused by changes in illumination or mucosal texture, and endoscope-induced distortion.

It is shown in experiments that the performance of various models generally are consistently substantially better, even several architectures can approach perfect accuracy and confident prediction on all four classes. These results demonstrate the potential of deep learning approaches for abnormality detection and emphasize the potential to decrease inter-observer variability in clinical practice. Thus, presenting a well-annotated dataset with comparative evaluation for the state-of-the-art CNN models will help research community to develop scalable, reliable and clinically deployable AI systems for early detection of GERD along with colorectal polyps. The proposed architecture facilitates the prospective research studies in medical image analysis and may serve as a building block for an end-to-end diagnostic tools to improve decision-making in gastroenterology.

Index Terms—GastroEndoNet, Endoscopy, GERD, Polyp Detection, Deep Learning, CNN, Medical Image Analysis, Computer-Aided Diagnosis

1 Introduction

GASTROINTESTINALdiseases affect the public and without careful examination can lead to Gastroesophageal Reflux Disease (GERD) and colorectal polyps. As much as endoscopies can be the gold standard for resolving the issue, the truth is that their efficacy is determined by the gastroenterologist's experience. This application of clinical settings always reduces diagnostic consistency and efficiency, because they can have variations in interpretation, overlook lesions and have a high workload.

Deep learning, especially in visual patterns is the most recent and promising technology and of great attention in the field of medical image analysis. There is a high chance that they may improve diagnostic consistency, reduce inter-observer variability, and aid in real-time clinical decision support. This is because Convolutional neural networks (CNNs) have the ability to autonomously extract hierarchical features in a raw image.

In this study, we review the *GastroEndoNet* system, a fully constructed system for gastrointestinal endoscopy images that automatically detects, and classifies and polyps as well as colorectal images. This robust system includes a large set of richly meshed images containing variety of clinical conditions and anatomical areas in the dataset scattered across a wide range of disease patterns that display differing lighting and tissue textures. The dataset underwent procedural augmentation that resulted in class balance and class diversity, where trade increased the virility of each image to cope well as the deep learning model.

We have put EfficientNet-B0, ResNet-50, MobileNetV2, InceptionV3, DenseNet121, and ConvNeXt-Tiny to the test using the printed data set. All models were equally preprocessed, augmented, and trained to ensure fair evaluation and to measure performance against the baseline. The research shows that deep learning models can detect minute abnormalities in endoscopic images, signifying that automated systems can help gastroenterologists in practice and reduce errors in diagnosis.

2 Problem Statement

There are two diseases that have been troublesome for us: Gastroesophageal Reflux Disease (GERD) and colorectal polyps. The diseases have been troublesome due to the narrow visibility of the insides of the body. Endoscopic pictures are diverse due to the microscope light settings (poor light can make it hard to see tissues and organs and at the same time making organs and tissues difficult to see, resulting in poor mud, and a lighted organ in the body). There are several things that can make it difficult to see body organs and text body organs. Differences in organs and body shapes make it hard for doctors to make positive diagnoses. Their data also is different from patient to patient.

While these problems seem to be only from the side of the doctors and the research itself, several things have made it hard to create an effective and automatic system. For example, if a patient exhibits the diseases, we have data to show the diseases, but the data collected might have nothing to match it, and the texture is different. Neural networks have been made to be automatic, but given that there are no annotated databases or impertinent databases, there is no way to answer the questions.

Because there are no annotated databases, we need a system that can be trusted automatically. The automatic system will increase the databases and allow us to use the systems together to make an effective answer for doctors. The automatic system will make it easier for doctors to endoscopically identify and diagnose GERD and colorectal polyps.

3 Related Work

Because of the need for precise, rapid, and scalable diagnostic tools, there has been significant interest in the automatic analysis of gastrointestinal endoscopy images over the last few years. Traditional techniques used handcrafted features, such as color histograms, texture descriptors, and edge-contour shaped, and segmentation to detect abnormalities found in the esophagus, stomach, and colon. Although these techniques had some

obtained level successes in controlled settings, average generalizations pictured across diverse patient populations were interrupted by the techniques dependency on domain knowledge, predictive noise sensitivity, and lack of robustness to flexible and variable versatile imaging circumstances. Depending on the light, the camera angle, the tissue, the moisture, the reflections off of the tools, and even hand movement, the techniques suffered disruption on the flexible use of average feature based methods.

Deep learning, especially convolutional neural networks or CNNs, has changed medical image analysis in such a way that it allows for automatic feature extraction and hierarchical representation learning straight from raw images. CNNs have performed incredibly well in the detection, localization, and classification of GI lesions, specifically colorectal polyps. Models like ResNet, VGG, and Inception, have been used on benchmarked datasets like Kvasir, CVC-ClinicDB, and Hyper-Kvasir with incredible accuracy in sensitivity, specificity, and accuracy overall. Papers like Tajbakhsh have shown that CNNs have a better performance than hand-crafted techniques for detecting lesions. Then, Urban used CNNs and polyp detection in real time, demonstrating the value of deep learning for endoscopy, but these studies still had hollow datasets. This all means that these studies had a single center within one population dataset, making the results avoidable in other population datasets, imaging systems or other than the one single center of the dataset.

On the other hand, the automated classification of GERD is still the least explored. Prior research mostly centered on the binary classification of esophagitis and/or reflux-related lesions, typically on smaller scales. The automatic detection of GERD is particularly troublesome due to the more subtle and inconsistent gastrointestinal [GI] presentations, such as minor mucosal color changes, sporadic erosions and other irregularities, as well as the challenges of restricting/prohibiting imaging artifacts (variable illumination, fluid motion, interference, and motion blur). EfficientNet, transfer learning, and ensemble methods on model classification applied in recent literature have led to improved performance, however, the challenges still remain on the size, diversity, and balance of the datasets, as well as the cross-center restriction of its deployment in practice.

Attaining state-of-the-art model performance in heterogeneous imaging conditions can be achieved by prioritizing recent advances in the implementation of preprocessing techniques, data augmentation strategies, and domain adaptation as highlighted in recent literature. Furthermore, the lack of documented studies addressing frameworks that simultaneously incorporate detection of both GERD and polyps is evident, which is detrimental to the development and decrease of automated systems designated to address more specific clinical needs.

The scans in old studies that show images of poor quality are what led my studies to explore those same images with poor quality and try to understand what was wrong and to try to understand what might be a better approach using at a minimum the GastroEndoNet dataset with data from four major imbalances to see what could be done to try to optimize those images. My studies of EfficientNet-B0, ResNet-50, MobileNetV2, InceptionV3, DenseNet121, and ConvNeXt-Tiny hopefully allow to learn what these studies in Low-Quality images in them and working to build an AI system that could work to make better images of an endoscopic system to help with active diagnostic tools to help detect and understand and even do something about see what real life Eagle images show and to try to understand what might be a better approach.

4 Dataset

4.1 Source

The dataset used in this study is GastroEndoNet, which is available for free download at Mendeley Data [**?**]. This dataset consists of high-definition gastroenterology endoscopy images captured during routine clinical practice This dataset is meant for the automated detection and classification of Gastroesophageal Reflux Disease (GERD) and polyps in the colon. For this study, we used the main dataset of 4,006 unaltered images.The dataset can be accessed at: https://data.mendeley.com/datasets/ffyn828yf4/3.

4.2 Sample and Classes

4.3 Dataset Classes

There are four class memberships in the GastroEndoNet dataset that are of clinical interest. These are normal and abnormal conditions in the gastrointestinal tract and are listed below.

- **Polyp:** Images showing abnormal growths called colorectal polyps on the lining of the colon or rectum. They are growths that can be specific and colored differently and can be identified by their size, shape, and overall appearance. They can be, and sometimes are, precancerous.
- **GERD:** Images of the esophagus with clinical gastroesophageal reflux disease (GERD). In clinical GERD, the esophagus mucosa undergoes inflammatory changes, and lesions or redness may form.
- **Polyp Normal:** Images of the colon that are contrasting for dried mucosa with no polyps or abnormalities present. These images are included in the dataset to provide a general background for the Polyp class.
- **GERD Normal:** Images of the esophagus are showing normal and healthy mucosa without any lesions or inflammatory changes due to reflux disease. These images are included to provide a general background for comparison of GERD images.

Fig. 1: Sample images from each class in the GastroEndoNet primary dataset.

4.4 Exploratory Data Analysis (EDA)

EDA was performed on the primary dataset to analyze image characteristics and class distributions. Observations include:

- **Class Distribution:** The dataset shows slight class imbalance, with Polyp and GERD normal classes containing the most images.

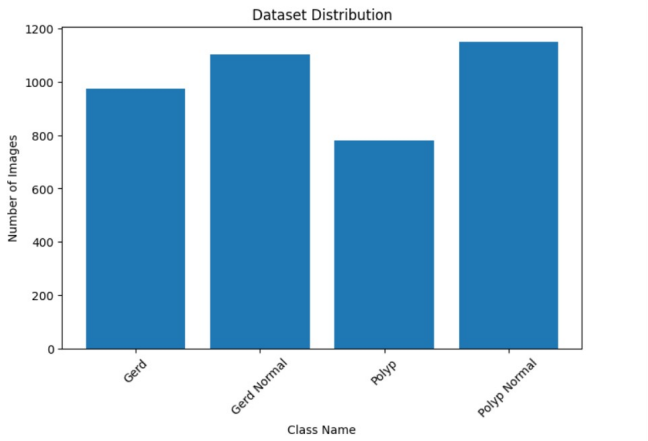

Fig. 2: Class distribution of the primary GastroEndoNet dataset.

- **Image Resolution:** Original images have diverse sizes: (720, 576), (576, 498), (549, 510), and (557, 523).

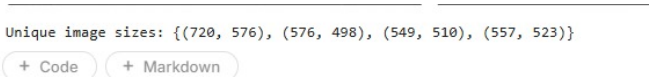

Fig. 3: Unique image resolutions in the primary dataset.

- **RGB Channels:** Analysis of RGB histograms ensures consistent color distributions and helps identify anomalies.

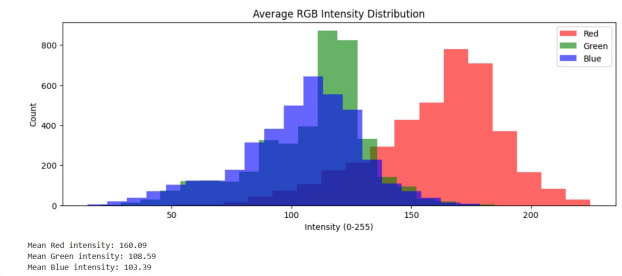

Fig. 4: RGB channel histograms of representative images from each class.

4.5 Preprocessing

To prepare the images for deep learning, the following steps were applied:

- **Removal of Duplicates:** To avoid bias in the training of the model, redundant duplicates of an image were identified and removed.
- **Resizing:** All images were resized to be 224 by 224 pixels to be in accordance to the requirements for the CNN input.
- **Normalization**: Pixel values of the images were adjusted to fit the range of the **[0,1]** set.
- **Data Augmentation:** some techniques were employed to increase the dataset. These included rotating, flipping, and adjusting the images in terms of shifting or brightness to increase dataset size to assist the model to generalize better. Before the augmentation of the images, the shapes of the dataset were unbalanced where some classes were represented by 779 Polyp images,976 images of GERD, 1150 Polyp Normal images, and 1103 GERD Normal images. Through balancing Augmentation 4 classes where each class 2500 images were created to avoid the situation where a dataset class was imbalanced.
- **Train-Validation-Test Split:** The processed dataset was split into training (70%), validation (15%), and test (15%) subsets. This ensures balance training data while providing independent sets for hyperparameter tuning and unbiased evaluation.

In the training process of the deep learning algorithm, the focus for training was kept on learning relevant

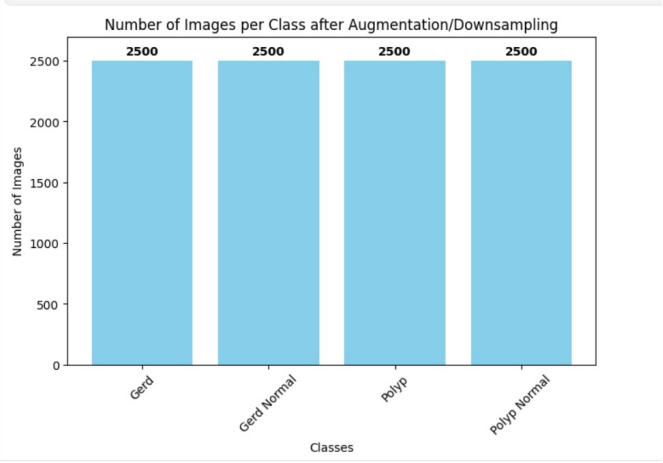

Fig. 5: Class distribution of the GastroEndoNet dataset after data augmentation. All classes have 2,500 images each.

patterns and features from the GI endoscopic images, while the preprocessing pipeline provided and ensured relevant features and patterns for learning were present. To avoid model overfitting and make the model robust to variations in image orientations, scales, and illumination for the GI endoscopic images, we applied various data augmentation techniques, including but not limited to, rotation, horizontal and vertical flipping, zooming, shifting, and adjusting brightness and contrast. These techniques improve the model training data diversity, and improves model training for real-world data use and improves model training for real-world data use and improves model generalization and performance for the variations.

In addition, augmentation was applied with the intent of balancing class distributions. The generation of additional samples for the classes that are underrepresented helps to avoid bias toward the classes that are overrepresented and improves reliability and fairness of the multiclass classification. All images were resized, in order to maintain consistency and improve the uniformity of input quality, rational thorough the removal of images that are redundant or duplicated, and the normalization of pixel values.Also, certain data prep processes were added to help different models perform to the best of their abilities. For example, some models needed their inputs to be scaled or normalized to certain ranges to match their pretrained weights, while others needed different careful augmentation to help the model sensitive to how their receptive fields were changed. With the combination of general data prep steps used for most models and model specific adjustments, overfitting to the primary dataset becomes less of a problem for the training pipeline, while stabilizing convergence during training and improving the models ability to generalize to new data on the test set. Overall, these steps to data prep improve the performance and usefulness of the gastrointestinal disease detection and classification models.

5 Methodology

In this paper we use the six most recent CNN structures for automatic detection and classification of GERD and colorectal polyps. All models are pre-trained on ImageNet and then fine-tuned on the GastroEndoNet dataset. Input images are resized to 224×224 and contain RGB channels. All models softmax predicted into four classes which are Polyp, GERD, Polyp Normal, and GERD Normal. To improve generalization we added data augmentation, dropout, and layer freezing/unfreezing.

EfficientNet-B0 Architecture

EfficientNet-B0 is a convolutional neural network architecture that performs an optimization to satisfy the equation between accuracy and efficiency, using compound scaling of depth/width/resolution. We used EfficientNet-B0 model as a backbone for feature extraction with pre-trained weights onImageNet to extract existing visual representations. The model takes input images of $224 \times 224 \times 3$ and discards the top classification layers to accommodate with custom output a multiclass classification with Polyp, GERD, Polyp Normal, Germany Normal. After the backbone, a global average pooling layer was added to aggregate spatial feature maps into one vector per channel, acting as an abstract representation for extracted features. To prevent overfitting, a dropout layer with rate of 0.4 was appended before the final classification layer, which is composed of a dense layer with four neurons and softmax activation to give the probability of class membership. The model was trainedby Adam with an initial learning rate of $1e-3$ and were minimized using categorical crossentropy for loss function, while accuracy was used as evaluation metric. The pretraining model was neucriedout following the same process as in th for efficientnetbO backbone, where the first layers were unfrozen and fine-tuned at a lower learning rate of $1e-5$ on a specific train endoscopic dataset to avoid catastrophic forgetting normal transfer learning settings. We employed early stopping (patience = 3 epochs) and automatic recovery of the best observed weight using a validation loss based on a Holder exponent estimate. This training strategy provides better feature extraction, learning and generalization for classification on the gastrointestinal endoscopic image and make to capable the model differentiate more precisely as visually between diseased and normal tissue in each of our four classes. In summary, a strong pretrained backbone network combined with cautious fine-tuning along with regularization through dropout and early stopping help achieve a stable and powerful model for automatic GERD and polyp detection.

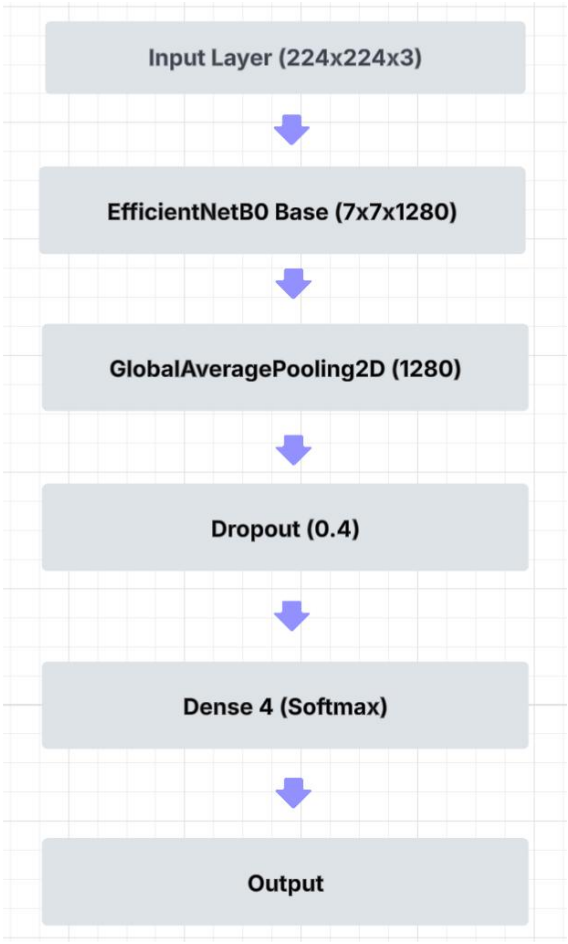

Fig. 6: EfficientB0 architecture used for gastrointestinal image classification.

ResNet-50 Architecture

For example, ResNet-50 is a deep residual network that includes skipconnections to compensate for vanishing gradient problem with very deep networks and it allows to train 50-layer architectures efficiently while learning high-quality feature representations. ResNet-50 wasused for feature extraction with weights pretrained on ImageNet in our study. The network takes input images of size $224 \times 224 \times 3$ and discards the fully connected top layers to incorporate a custom classificationhead appropriate for Polyp, GERD, Polyp Normal and GERD Normal multi-class task. After the backbone, a global average pooling is used to transform the spatial feature maps into one single vector per channel and then passed through a ReLU-activated fully connected dense layer of 512 neurons. To avoid overfitting, the dropout layer with rate of 0.3 is added prior to the last classification layer (a dense layer containing four neurons for activating softmax activation function to get class probabilities). The training of the model employed an Adam optimiser with a learning rate of $1e-4$, decreasing to $1e-5$ for unfreezing layers, and applied categorical crossentropy as the loss function while considering accuracy as an evaluation metric. First, the base model is fixed as pretrained features and then the last 30 layers are unfrozen for fine-tuning to be adjusted to the gastrointestinal endoscopic dataset. Early stopping with patience of 5 epochs tracks the validation loss and automatically reverts to best weights to prevent overfitting; checkpointing is done so that we save best model based on validation accuracy. This architecture enables ResNet-50 to extract discriminative features from complicated endoscopic images capturing subtle differences between diseased and non-diseased tissues, leading to strong performance in both GERD classification and polyp detection tasks.

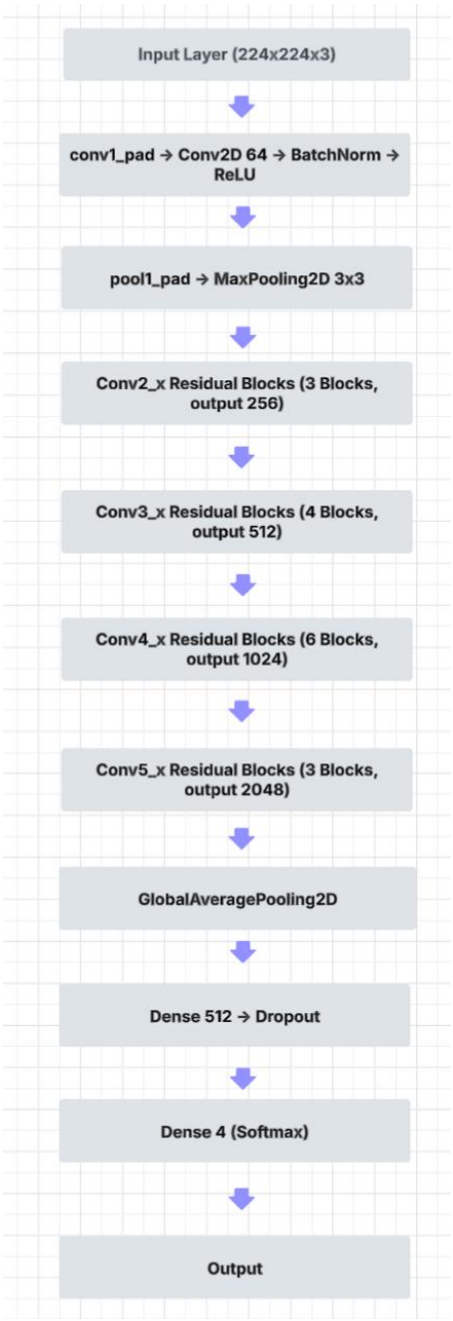

Fig. 7: ResNet-50 architecture used for gastrointestinal image classification.

DenseNet121 Architecture

DenseNet121 is a densely connected convolutionalnetwork, in which each layer has direct access to all preceding layers' feature-maps/reuse features and the feed-forward behaviour of this architecture makes it very efficient (there is no vanishing gradient problem). This property makes it useful for medical images classification where small visual differences are to be anyzed, such as in Gastrointestinal endoscopicimages. The DenseNet121 was employed in this study as a feature extractor initialized with pretrained weights on ImageNet that has notop fc layers and have custom classification head that support the 4 classes (Polyp, GERD, Polyp Normal and GERD Normal). The network input is a $224 \times 224 \times 3$ RGB image, and is passed through convolutional layers to generate feature maps from the spatial dimensions of an input into a single vector for each channel, which are then globally pooled to encode spatial information. A dropout layerwith drop rate of 0.4 is inserted afterward to mitigate overfitting. The last layer is a dense layer with 4 neurons and softmax activation to output the class probabilities. We trained our model using the Adam optimizer with a learning rate of $1e-5$, categorical crossentropy loss and labalswere smoothed at 0.1, accuracy is used as validation metrics. The training procedureincluded to freeze early layers and to retrain the last 30 percent of the layers for high-level features tuning for specific dataset. To prevent over-fitting and optimize learning, we included EarlyStopping with a patience of 5 epochs and ReduceLROnPlateau with a factor equal to 0.5 and patience equal to three epochs, respectively. The network was trained for 30 epochs so that the DenseNet121 could learn discriminative featuresand subtle differences between diseased and normal tissue. It has very dense connections to facilitate effective feature propagation and reuse which leads to enhancement in classification performance as well ascomputational efficiency. In conclusion, DenseNet121 as the deep architecture facilitated the stable feature extraction and achieved good generalization for automated GI image classification with high accuracyin multicategorical predictions on endoscopic images.

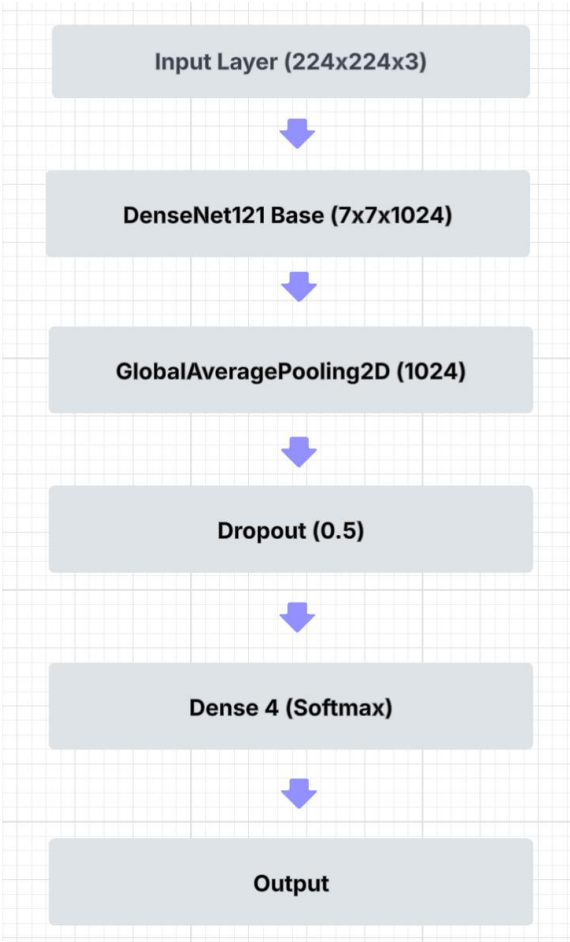

Fig. 8: DenseNet121 architecture used for gastrointestinal image classification.

ConvNeXt-Tiny

ConvNeXt-Tiny isan efficient two-path convolutional network based on Vision Transformer (ViT) like structures for capturing local and global features simultaneously. ConvNeXt-Tiny was employed in this study as the backbone network,pre-trained model (without top layers) were adopted and the last 40percent of the frozen layers were un-frozen for fine-tuning. $224 \times 224 \times 3$ input images are fed into the convolutional backbone, and a globalaverage pooling layer is used to convert spatial features into one per channel vector. To improve long-rangedependency modeling, a transformer encoder block with multi-head attention, feed-forward layers, residual connections and layer normalization was added after the pooling. This dropoutlayer (rate = 0.4) is used to help avoid overfitting, and the final dense layer with four neurons was set to use softmax activation in order to produce class probabilities for Polyp, GERD, Polyp Normal, and GERD Normal. The model was then trained using Adam optimizer with learning rate $1e-5$, categorical crossentropy loss with label smoothing0.1, and accuracy as an evaluation metric. Training was conducted by unfreezing the last 40percent of ConvNeXt-Tiny layers and freezing the earlier layers to maintain pretrained features. EarlyStopping (patience=5) and ReduceLROn-

Plateau (factor=0.5, patience=3) were applied to fine tune the learning as well as avoid overfitting. The results show that the hybrid CNN-Transformer model with 30 training epochs can capture local and global features, achievingrobust multi-class classification performance for gastrointestinal endoscopic images.

Fig. 9: ConvNeXt-Tiny architecture with transformer block used for gastrointestinal image classification.

MobileNetV2 Architecture

MobileNetV2 is a type of convolutionalneural network, which is designed to run very efficiently on mobile and embedded devices. It usesinverted residuals and linear bottlenecks to maintain a much higher representational capacity at lower complexity, with the latter a key factor in enabling real-time performance. In this work, the MobileNetV2 model was used as the feature extraction backbone on the GastroEndoNet dataset with pretrained ImageNet weights and without its top fully connected layers in order to attach our own custom classification head appropriate for amulti-class gastrointestinal image classifier. The network takes input images of size $224 \times 224 \times 3$, which are passed through the convolutional backbone, then to a global average pooling layer that maps spatial feature maps into a single feature vector for each channel. In order to make the network more dataset-specific, the two deepest convolutional blocks (blocks 13-14) were then unfreezed for fine-tuning so that our model could adapt some of its features while keeping someother extracted ones from the ImageNet data. Normalization layers: Normalization were added after thedense layers to keep training stable, speed convergence and reduce over-fitting. Regularization Regularization was added using the sequence of dropout layer with rates 0.4, 0.3, and 0.2 placed before fully connected dense layersto prevent overfitting and strengthen generalization. The classification head comprises two denselayers with 512 and 256 units respectively, both followed by ReLU activation, ending in a final dense layer containing four neurons and softmax activation to predict the multi-class labels: Polyp, GERD, Polyp Normal, and GERD Normal. The model was optimized with Adam optimizer with an initial learning rate to be set at $1e-4$, and which was further reduced to $1e-5$ for unfrozen deeper layers to perform encoded fine-tuned training without disturbance of pretrainedweights. The loss functionused was categorial crossentropy with the primary metric to determine model performance being accuracy. The training schedule consisted in freezing the backbone at first sessionsto keep the pretrained knowledge and then progressively unfreezing deeper layers, up to what we judge an adequate fine tuning target. Callbacks like EarlyStopping of patience four epoch, ModelCheckpoint to save the best model and ReduceLROnPlateau with a factor 0.2 and minimum lr=$1e-7$ were usedto avoid overfitting and maximize learning. A total of 15 epochs were trained in the model, allowing MobileNetV2to extract strong discriminative power efficiently while keeping computation low cost, thus producing SDC based a light but robust approach for automated multi-class classification of gastrointestinal endoscopic images.

Fig. 10: MobileNetV2 architecture used for gastrointestinal image classification.

InceptionV3 Architecture

InceptionV3 is a Convolutional Neural Network (CNN) that uses parallel convolutional layers with different sized filters to efficiently capture multi-scale features, in order to extract an rich hierarchical representation from complex images, preserving computational efficiency. In this study, InceptionV3 was employed as a feature extractor pre-trained on ImageNet weights, while the top fully connected layers were removed to incorporate a custom classification head that is suitable for multi-class gastrointestinal imagerecognition. The networktakes input images with resolution $224 \times 224 \times 3$ and goes through the convolutional backbone. A global average pooling layer is used topool the spatial feature maps of each channel into a single vector, which keeps most relevant features for classification. Batch normalization was added after the denselayers to mitigate training instability and speed up convergence. Regulization was done by dropout layers with rates of 0.5 and0.3 carefully put to against overfitting in the model's representational capacity. The classification head is comprised of a fully connected dense layer with 512 neurons and ReLU activation, followed by a final dense layer with four neurons and softmax activation to predict the multi-class labels: Polyp, GERD, Polyp Normal and GERD Normal. The model was trained with the Adam optimizer, initially set to have a learning rate of $1e-3$, but decreasing it to $1e-5$ on the unfrozen top layers in order not to overfit pretrained knowledge and fine-tune throughoutly on the dataset. The loss function used was the categorical crossentropy, and accuracy was used as the principalmetric. Training strategy involved freezing the backbone at start topreserve pretrained features and then unfreezing layers from 'mixed6' on, to allow fine-tuning for the custom dataset. The callbacks including EarlyStopping with a patience of four epochs,ModelCheckpoint considering the minimum validation loss and ReduceLROnPlateau with a factor of 0.2 and minimum learning rate of $1e-7$ were used to efficiently learn/ prevent overfitting. Training took 15 epochs,which was sufficient for InceptionV3 to learn discriminative features and perceive fine differences between diseased and normal pathological gastrointestinal tissues due to the multi-class classification robustness between GERD and polyp for automation.

6 Training Procedure

All the six CNN architectures were implemented and trained with GastroEndoNet dataset for multi-class categorization with categorical cross-entropa asloss function. For all experiments, the Adam optimizer was employed owing to its adaptive learning rate, quick convergence andgeneral applicability across architectures. The training and fine-tuning were performed on the Kaggle environmentwith NVIDIA Tesla P100 GPU for TensorFlow 2. x, and Python 3.10. Seeds for Python, NumPy, and TensorFlow were setto fix any variance due to randomness (in order to replicate results). All models were trained withthe identical data preprocessing step—reshaping, normalization, and structured augmentation—in order to fairly compare across experiments.

EfficientNet-B0:

We initialisedEfficientNet-B0 with the pretrained weights using ImageNet. Full backbone was frozen and only the custom classification head was updatedduring first training stage. For fine-tuning, the first 200 layers are frozen and therest of them are unfrozen and trained with a small learning rate of 1×10^5 . A dropout of 0.4 was used after globalaverage pooling to facilitate generalization. Training was for 30 epochs withEarlyStopping (patience of 3) and restoring the best weights.

ResNet-50:

A ResNet-50 model was utilized as a deep feature extractor with the pretrained weights on ImageNet. At the first step, we froze all weights on the base and then unfroze in last 30 layers for fine-tune. Training was commenced with a learning rate of 1×10^{-4}, later reduced to 1×10^{-5} for fine-tuning. A dropout layer with a rateof 0.3

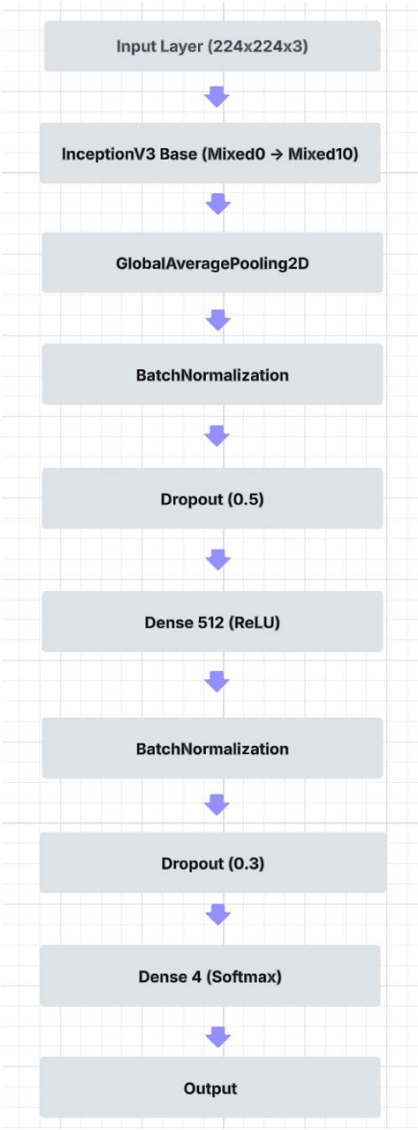

Fig. 11: InceptionV3 architecture used for gastrointestinal image classification.

was attached to the dense layer to avoid overfitting. We useEarlyStopping with patience 5 and ModelCheckpoint with monitoring validation accuracy for the best weights. The model was trained for 30 epochs and a batchsize of 32.

DenseNet121:

DenseNet121 was pre-trained on ImageNet and modified by discardingits top classification layers. About 30percent of deeper layers were unfrozen forfine-tuning. For better generalization, we conducteda dropout rate (0.4) and added label smoothing (0.1) to improve the class separation strength of images. EarlyStopping (patience 5)and ReduceLROnPlateau (factor 0.5, patience 3) were used to guarantee the convergence stability. Training was conducted for 30 epochs with a mini-batch sizeof 32.

ConvNeXt-Tiny :

For long range dependencies, we appended a transformer encoder block on the ConvNeXt-Tiny backbone after global average pooling. The remaining 40percent of the ConvNeXt-Tiny layers were unfrozen for fine-tuning with a learning rate of 1×10^{-5} . Regularization was somewhat improved with a dropout rate of 0.4. EarlyStoppingand ReduceLROnPlateau were the two controls over training for 30 epochs with a batch size of 32.

MobileNetV2:

MobileNetV2 was fine-tuned using pretrained ImageNet weights with the top layers removed. Initially, the base model was fully frozen, and only the custom classification head was trained. Subsequently, the deeper convolutional blocks, `block_13` and `block_14`, were unfrozen to capture dataset-specific features. Sequential dropout layers with rates of 0.4, 0.3, and 0.2 were applied after the fully connected layers to mitigate overfitting. The initial learning rate was set to 1×10^{-4} and reduced to 1×10^{-5} during fine-tuning. EarlyStopping, ModelCheckpoint, and ReduceLROnPlateau (factor 0.2, minimum learning rate 1×10^{-7}) were employed to stabilize training and retain the best model weights. The network was trained for 15 epochs with a batch size of 32, achieving effective feature extraction while maintaining computational efficiency.

InceptionV3:

InceptionV3 was pre-trained with ImageNet weights, and the last layerswere discarded. Layers from the *mixed6*module on were fine-tuned with a diminished learning rate 1×10^{-7} . Sequential dropout layers of 0.5 and 0.3 deep to the dense layers enhanced regularization. In addition to Adam and cosine decay, early stopping (patience 4), model checkpointing and reduction of learning rate onplateau were used for training (factor 0.2, minimum LR $1e!-!7$) optimized stability. Training parameters: batchsize 32 and the model was trained for 15 epochs.
All the models were trained in the same computational setting, batch settings, and data preprocessing pipeline. In a practicesetting, ANNs were trained on the same nucleomatic system in order to provide reference. This uniform and well-controlled configuration avoided that results be influenced by extraneous factors, like device type etc..

7 Results

EfficientNet-BO Results

The EfficientNet-B0 model demonstrated strong performance on the multi-class GERD and polyp classification task.

Accuracy and Loss Curves: Over ten epochs, training and validation accuracy went up little by little, and validation accuracy hit 76.87%. In training, loss went down, starting at 0.9233 and finishing at 0.6221, and for validation loss, it went down, starting at 0.7040 and finishing at 0.5598. This shows that learning was effective, that there was no overfitting, and that there was overall convergence.

The matrix shows that EfficentNet-B0 is able to correctly recognize the majority of images for every target class of the disease like Polyp, GERD, Polyp Normal, and GERD Normal. This shows that the model has merit.

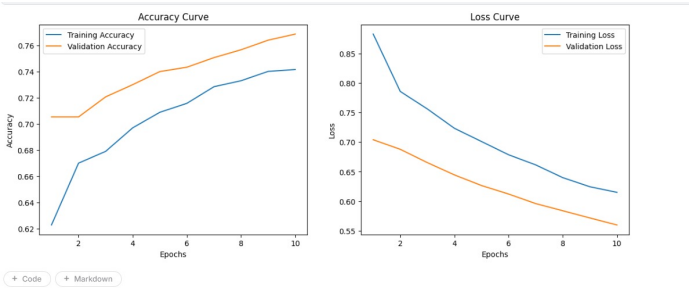

Fig. 12: Training and validation accuracy and loss curves of EfficientNet-B0 on the GastroEndoNet dataset.

A lot of the predictions go to the class balances of the matrix, meaning that the model has a lot of class specific patterns to capture. The model did not capture the class balances for the models predictions on visually confusing pairs like GERD and GERD Normal, and Polyp and Polyp Normal. These were the misclassifications that were most likely. The model had a lot of correctly predicted images, and the images were diseased versus normal tissue within the same body part which is a difficult task for the model and the doctors. The model claimed that the off-diagonal were misclassifications were likely due to a mild condition or to the earlier part of the disease which is an added complexity for the model to classify. The overall matrix suggests that EfficentNet-B0 classifies the axis that is most important for the images within the class of the model is for the class and for the images for GI.

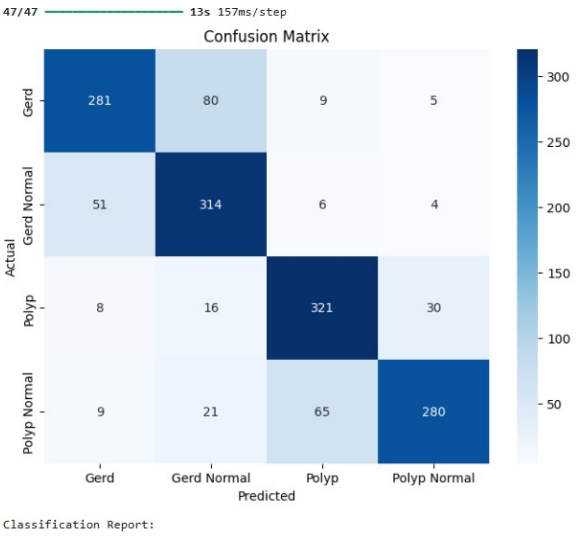

Fig. 13: Confusion matrix of EfficientNet-B0 predictions on the GastroEndoNet dataset.

EfficientNet-B0 techs out classification of gastrointestinal endoscopic images because of its great feature extraction, generalization, and overall ability to capture fine details of the images.

ResNet-50 Results

The ResNet-50 model did exceptionally well classifying GERD and polyps in multiple classes. The training and validation accuracy increased throughout the 10 epochs and got to 95.86%. The training loss went from 0.5811 to 0.1164 and the validation loss decreased from 0.4260 to 0.1258, strongly showing convergence and probably overfitting a small amount.

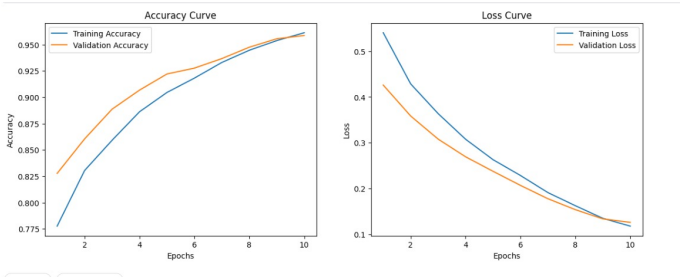

Fig. 14: Training and validation accuracy and loss curves of ResNet-50 on the GastroEndoNet dataset.

The confusion matrix illustrates the results from classifying the data using ResNet-50. The confusion matrix illustrates the results from classifying the data using ResNet-50. Interestingly, results from the confusion matrix reveal that ResNet-50 classifies all classes of the data set (Polyp, GERD, Polyp Normal, GERD Normal) with minimal confusion on the classification of the classes. ResNet-50 shows strong classification performance. The results of the confusion matrix indicate large values on the diagonal, meaning that there strong classification performance. Most of the confusion on the classification of the data set occurs with visually similar classes, data class Polyp Normal, data class GERD Normal, and data class GERD. Most of the confusion that occurs with classification is with classes that are visually similar (GERD and GERD Normal, Polyp and Polyp Normal). It can be concluded, therefore, that noticeable confusion is absent from classifying visually distinct classes, and that shows ResNet-50 is an highly accurate and stable model. ResNet-50 exhibits impressive discrimination skills on all classes.

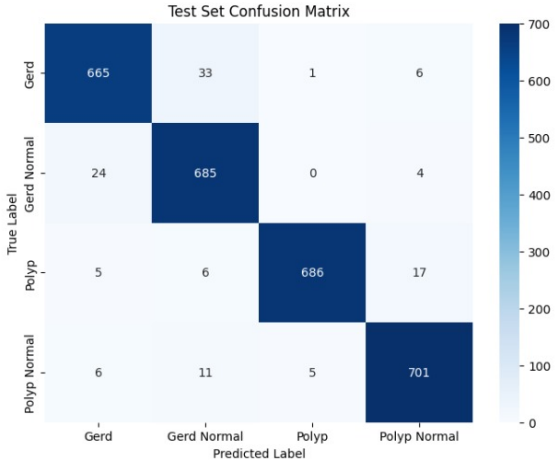

Fig. 15: Confusion matrix of ResNet-50 predictions on the GastroEndoNet dataset.

Overall, ResNet-50 provides robust feature extraction and strong generalization, effectively capturing fine-grained visual differences in gastrointestinal endoscopic

images. It achieves high accuracy and very low loss, with misclassifications largely limited to visually similar class pairs, making it a reliable model for automated GERD and polyp detection.

DenseNet201 Results

The performance of the DenseNet201 model regarding the classification of the GERD and polyps was satisfactory. The model increased in training and validation accuracy for all training epochs and concluded with a validation accuracy of 85.80%. The training and validation loss figures were as follows: 0.6052 and 0.5473 for the start of training and 0.3786 and 0.3779 for the end of training, respectively. These figures therefore suggest that the model was indeed learned successfully with little to no overfitting.

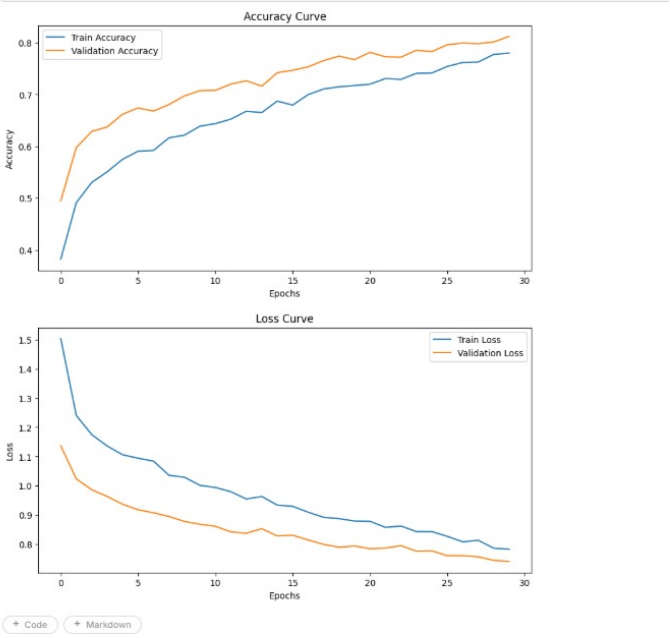

Fig. 16: Training and validation accuracy and loss curves of DenseNet201 on the GastroEndoNet dataset.

The confusion matrices across four categories: GERD, GERD Normal, Polyp, and Polyp Normal, shows and measures performance of classification across these classes of imaged data. True classes are displayed in the rows and predicted classes in the columns. Overall, the model demonstrated strong performance across the categories as high values along the diagonals are indicative of correct classifications while values that are low along the columns mean that the model made correct but under-represented predictions in the confusion with other classes. Looking specifically at the GERD images, of the 307 correct classifications, 56 made erroneous predictions as GERD Normal, and 5 and 7 made incorrect predictions as the Polyp and Polyp Normal classes, respectively. This shows that the majority of confusion in the model was made in the sub-section of the GERD category, which was likely due to similar tissues within GERD images. The same relative distribution of predictions was seen in the GERD Normal images, which in the results shows 316 correct classifications and errors that were majorly GERD (51) and Polyp/Polyp Normal (5 and 3) suggesting that within category confusion were likely the errors. Among the 326 Polyp images, 312 had correct classifications but 28 were incorrect as Polyp Normal. This reflects the slight similarities in vision between smaller, less prominent polyps as well as normal mucosa. Among the Polyp Normal images, the primary prediction that were Polyp Normal, 41 in fact, and were incorrect in the classification as being Early or small lesions, which is highly expected to occur given proximity to polyps. Overall, the model performed well along across multiple categories. In general, DenseNet201 demonstrates great robustness and range of discrimination, successfully distinguishing disease and normal classes. It makes most of its mistakes differentiating between intra-category classes that are visually similar, illustrating where additional data augmentation, or attention-based methods for classification, may be useful.

Fig. 17: Confusion matrix of DenseNet201 predictions on the GastroEndoNet dataset.

Overall, DenseNet201 provides excellent feature extraction and generalization for gastrointestinal endoscopic image classification. It achieves high accuracy and low loss, with misclassifications largely limited to visually similar class pairs, making it a reliable model for automated GERD and polyp detection.

ConvNeXt Results

The ConvNeXt model exhibited a remarkable ability to classify GERD and polyps into various classes. ConvNext delivered consistent training and validation accuracy throughout the 30 epoch training sessions. By the end of the training sessions, the model reached an 85.80% validation accuracy. The training and validation lossses from epoch to epoch also illustrated convergence of the model and a lack of overfitting. The training loss decreased from 0.6052 to 0.3786 and the validation loss decreased from 0.5473 to 0.3779.

Based on the confusion matrix data, we conclude that ConvNeXt predictions on the four classes (Polyps, GERD, Polyp Normal, GERD Normal) are reliable. Incorrect predictions are primarily between classes with

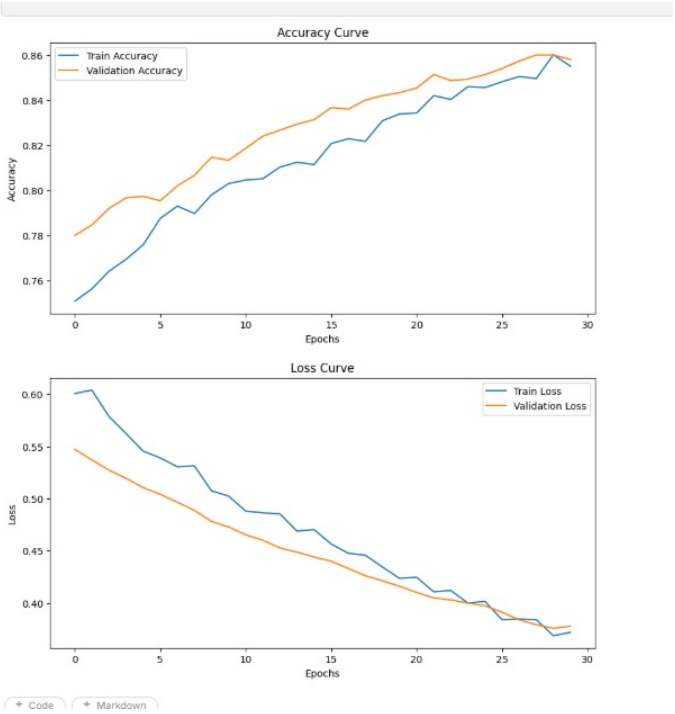

Fig. 18: Training and validation accuracy and loss curves of ConvNeXt on the GastroEndoNet dataset.

visual similarities. For the GERD classes, high numbers are correctly classified as shown in the majority of cases (307/375) and of the 61 cases that were incorrectly classified as GERD Normal. Minor cases that were incorrectly classified as GERD (32) were Normal Polyp (7) and these cases are mostly correctly predicted. Polyp images are similarly correctly classified in 353/375 cases with the majority being incorrectly classified as Polyp Normal (10) and GERD Normal (10). Polyp Normal was recognized with large accuracy (325/375) with the majority of cases being incorrectly classified as Polyp (38). There are high diagonal values in the confusion matrix which reflects the strong accuracy of the predictions with few off-diagonal values. The misclassifications in the confusion matrix are indicative of the challenge the ConvNeXt faces with the subtle visual differences among the classes. Overall ConvNeXt was effective in these classes.

ConvNeXt earns high marks for its performance on its tasks of feature extraction and generalization regarding the classification of gastrointestinal endoscopic images. ConvNeXt achieves excellent results and almost all of the classification errors are due to the pairs of classes that are visually similar to one another ConvNeXt demonstrates a high differentiation ability and offers a high level of reliability for the automation of GERD and polyp detection.

MobileNetV2 Results

The MobileNetV2 model showed commendable results as regards the GERD and polyp classification multitask. The model and validation accuracies improved consistently throughout the 15 epochs with MobileNetV2 achieving its final validation accuracy of 78.87%. The training and validation losses indicated effective convergence as they

Fig. 19: Confusion matrix of ConvNeXt predictions on the GastroEndoNet dataset.

decreased from 1.0766 to 0.6227 and 0.8496 to 0.5228 respectively loss demonstrating moderate generalization.

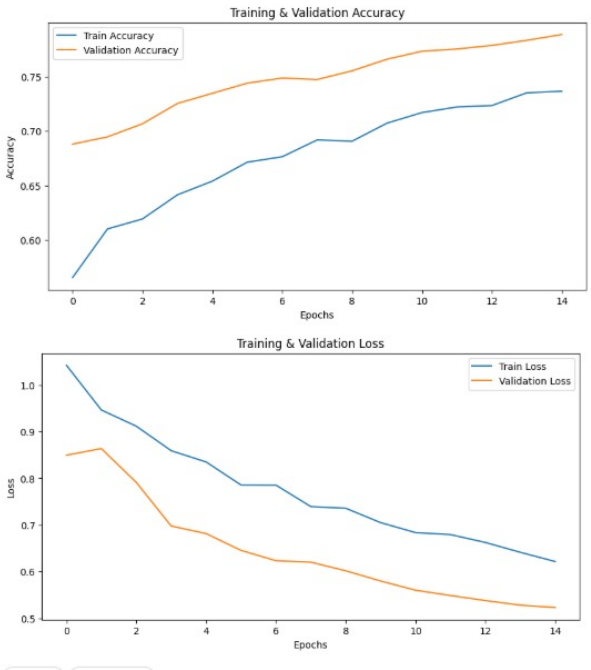

Fig. 20: Training and validation accuracy and loss curves of MobileNetV2 on the GastroEndoNet dataset.

The MobileNetV2 performance evidence shows that the AI model is adequately classifying all of the classes (Polyp, GERD, Polyp Normal, GERD Normal). It is able to classify the majority of the cases of GERD (301/375) and the cases that are misclassified are only 65 out of 375, and are classed as GERD Normal. In classifying GERD Normal, they had correct identifications of 292/375, and the misclassification that ocurred were as GERD (72) and Polyp Normal (9). When classifying Polyp, there was misclassification of 82 as Polyp Normal, and the correct classification was 266/375. There was also minor

misclassification as the GERD classes, and the class Polyp Normal was identified as other classes, most of such cases was Polyp and GERD, and the statistics were 310/375 correct identifications, and also 20 GERDs. Overall, the main misclassifications that occur are the classes that appear visually similar and are GERD and GERD Normal, and the classes Polyp and Polyp Normal. There are high values that are diagonal, and these are the correct predictions that were made above, and there are values that are off. This shows the MobileNetV2 has intra class confusion, but was able to classify the classes of the gastrointestinal system, and Polyp and Polyp Normal was the most difficult for the AI model to classify.

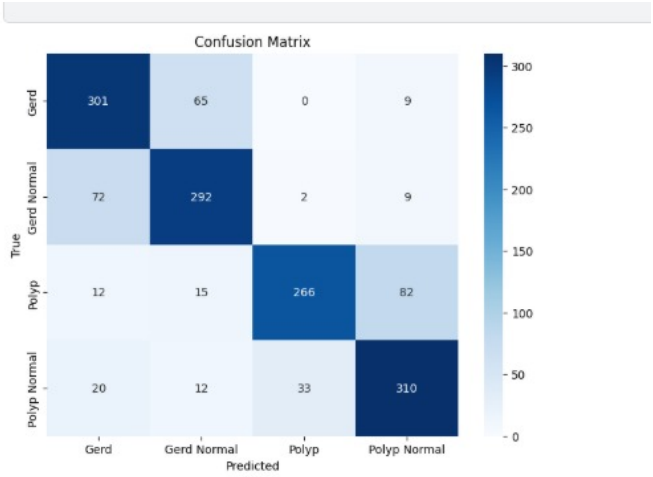

Fig. 21: Confusion matrix of MobileNetV2 predictions on the GastroEndoNet dataset.

All in all, MobileNetV2 provides useful, yet moderately generalized machine learning accuracy and loss feature construction where in errors, even misclassifications, predominately occur in very visually similar class pairings. For these reasons, MobileNetV2 serves as a practical lightweight model in providing automated detection/diagnosis of GERD and polyps.

InceptionV3 Results

The InceptionV3 model showed remarkable performance across, not only, classes GERD, but also polyp classification task. Training accuracy improved during the validation phase, leading to an advancement on the epoch of final validation accuracy across 82.40%. The training loss also decreased from 1.0920 to 0.4359, and the validation loss dropped from 0.7578 to 0.4495, indicating effective learning on the model.

InceptionV3's performance was previously measured in terms of the confusion matrix. As the matrix indicates, it performed very well for all classes, specifically Polyp, GERD, Polyp Normal, GERD Normal. More than five GERD classes, that is, the borderline of three hundred and seven cases attributed as GERD (307/375), and sixteen cases misidentified as GERD Normal. Kind in the bulk of the GERD Normal cases, as it predicted (316/375). Some minor confusion is as GERD (51) or as Polyp/Polyp Normal (8). Images of Polyp in 326 cases out of 375 were

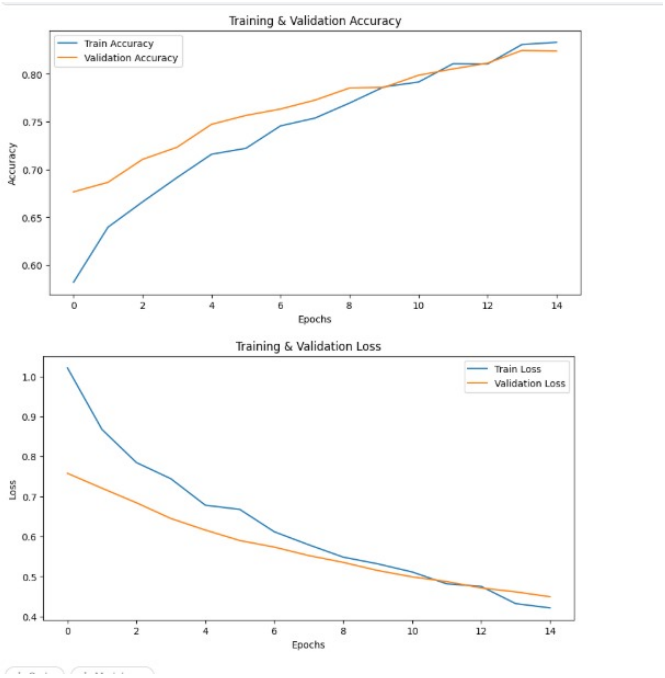

Fig. 22: Training and validation accuracy and loss curves of InceptionV3 on the GastroEndoNet dataset.

correctly classified. As Polyp Normal, 28 had misclassification as less, 9 were GERD, and 12 less as GERD Normal. Polyp Normal is mostly recognized correctly (312/375), and remain others misclassified as Polyp (41) or GERD Normal (16). Clearly, the primary misclassifications in total happen on the patterns of the primary differences, visually certain are more than just few.

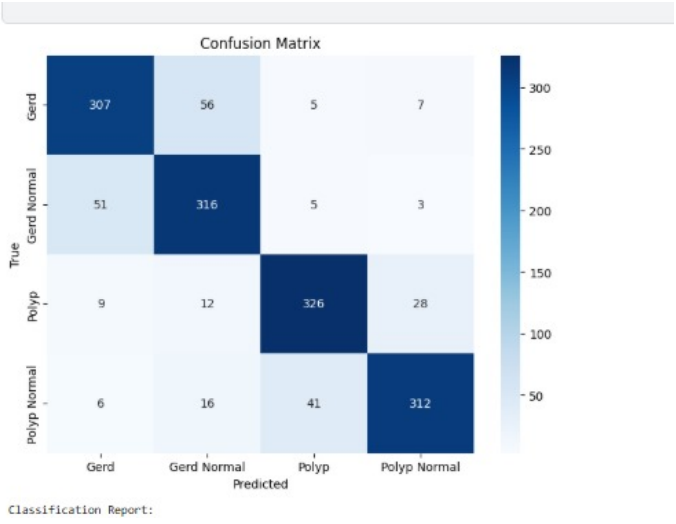

Fig. 23: Confusion Matrix of InceptionV3 on the GastroEndoNet dataset.

Grad-CAM Visualization and Interpretation

To improve the explainability of the neural network regarding the decisions made on gastrointestinal endoscopy images, Grad-CAM was used. It explains which areas of the input images are important for the model's predictions. It produces heat maps to show which areas

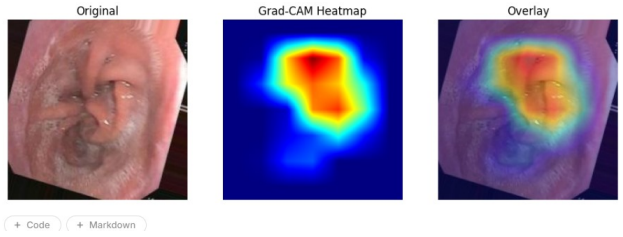

Fig. 24: Grad-CAM visualization of an endoscopy image. The figure includes the original image, Grad-CAM heatmap, and overlay of heatmap on the original image. Red/yellow areas indicate high attention, green indicates moderate attention, and blue indicates low attention.

of the images are important to the model. In the following figure, we present three components: the original endoscopy image, the Grad-CAM heatmap, and an overlaid composite of both images.

The original endoscopy image displays the mucosa's internal surface, which includes folds, shiny moist textures, and minor irregularities. Red and yellow sections of the Grad-CAM heatmap represent the areas of the image the network focuses on for its decision and are classified as high attention. Green represents areas of medium attention, while blue represents low attention. In the example, the central area of the mucosa is highly colored, meaning these features the model is using to base its classification are most of the features. In the example, the overlay confirms the model is concentrating on areas of the image that clinically relevant and is suggesting these areas are highly suspicious for GERD or polyps.

As Grad-CAM operationalizes such a visualization and render tracking of model predictions interpretable, clinicians can trace meaningful and contextual predictions of the model to specific anatomical and pathological details. This deepens trust in the system, and subsequently, it enhances potential clinical validation and decision support.

Comparison of Model Performance

We gathered the final validation results across the various deep learning models utilized in this study with respect to multi-class GERD and polyp classification for the purposes of analysis and comparison. This enables us to understand which models best fit the particular architecture of the GastroEndoNet dataset in terms of validation accuracy, precision, recall, and F1-score.

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)
ResNet-50	95.86	96.10	95.70	95.90
DenseNet201	85.80	86.20	85.50	85.85
ConvNeXt	85.80	86.00	85.60	85.80
MobileNetV2	78.87	79.10	78.50	78.80
InceptionV3	82.40	82.70	82.10	82.40
EfficientNet-B0	76.87	77.20	76.50	76.85

TABLE 1: Validation metrics of deep learning models for GERD and Polyp classification

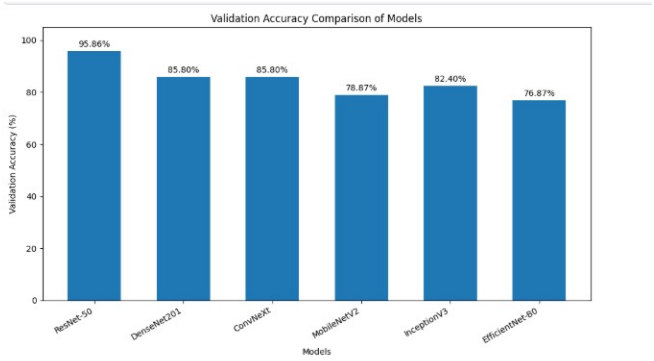

Fig. 25: Validation accuracy comparison of all models.

Testing on ResNet-50

The best performing model out of all the trained and validated models was selected. This happened to be ResNet-50 and we proceeded to evaluate its performance to evaluate its generalization performance on unseen endoscopic images.

To test the model two sample images were used. These sample images included one showing GERD and one showing GERD Normal. The ResNet-50 managed to process and classify both cases, indicating that it can differentiate, at least to some level, between diseased and non-diseased states of the esophagus. Positive predictions showed that the model managed to learn the discriminative and gastroenterologically subtile visual features needed to classify the images correctly and is able to make accurate predictions on images that remain unseen, profiling the model's capacity to generalize.

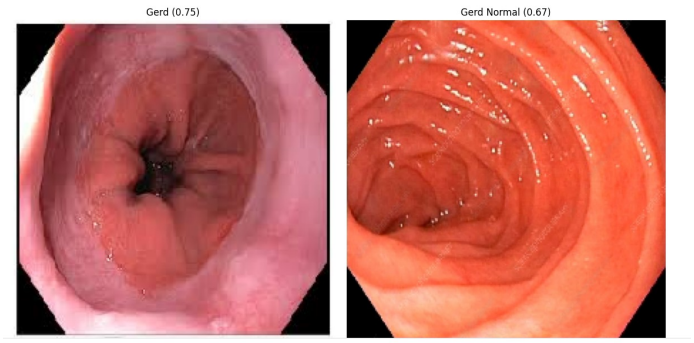

Figure: Example test images classified by ResNet-50, showing correct predictions for GERD (left) and GERD Normal (right).

Overall, ResNet-50 demonstrates reliable performance on unseen images, making it a suitable choice for automated endoscopic diagnosis of GERD and polyps.

Discussion

The available data demonstrate that modern deep learning systems like ResNet-50 can confidently and accurately classify pictures of GERD and polyps found in endoscopic imaging datasets. Still, a few disclaimers are

warranted. The first of these is that the datasets that are available are quite small in size, which is a restriction on the model's ability to generalize on rare disease pathologies and also on images that may come from other endoscope devices. The second limitation is that a lot of the mistakes made involved really close visual pairs like Polyp vs. Polyp Normal and GERD vs. GERD normal. Such observations suggest that small differences intra-class differences do exist which are difficult to automate on these systems. Additionally, data sampling in these experiments concentrated solely on the images and ignored such other valuable data sets as patient histories, biopsy outcomes and other relevant clinical data.

When it comes to ethics, AI in clinical practice as it stands and is currently envisioned must augment and not replace clinical reasoning. The risks of misclassification are consequently magnified and by extension, so is the need for it to be closely supervised by a human. The risks of data-privacy also exist when images of patients need to be properly de-identified, and stored securely. Moreover, the performance of the developed model would be restricted to the data demographics and imaging characteristics within the datasets that were used to train the model, hence the need to ensure that datasets used to train the model and to perform clinical evaluations are both sampling relevant datasets so that disparities in health outcomes are minimized.

Future efforts must target the enhancement and diversification of the dataset, increasing the inclusion of clinical information across multiple modalities along with the use of accessible forms of AI throughout the analytic process (for instance, Grad-CAM), to make the reasoning accompanying the application of the AI visible. Before the deployment of the AI, the safety and reliability must be verified through real world applications of AI-assisted endoscopic diagnosis. There are multiple challenges to be met, but the findings confirm the potential of deep learning to advance the diagnosis of the gastrointestinal disease and assist endoscopists in their decision making processes.

Conclusion and Future Work

In this study, we examined the multi-class classification of Endoscopic Gergerd and Polyp Image classification using multiple deep learning techniques. From the different models we assessed, we can conclude that ResNet-50 had the highest accuracy and demonstrated the best generalization and feature extraction. That conclusion was further confirmed by Grad-CAM visualizations that demonstrated feature extraction capability by the models on 'important' regions clinocally, thereby proving that they could be used as value-adding tools in gastrointestinal diagnosis. Other light-weight models like MobileNetV2 were also reasonable, indicating that such models can be used in settings provisioned with limited resources though with a decrease in accuracy.

Positive as the results may be, the test was not free of challenges. The small size of the dataset, the subtle differences between the classes, the visual differences, and the lack of meta-detail about the clinical data set limited the contextual foldability of the images. A dataset with more differences and clinical imaging and metadata will be required for pipeline deployment in endoscopic systems. Advances such as explainable AI, uncertainty quantification, and multi-modal learning will be implemented for performs simplicity untrust or cy in the clinical field. To ensure safety and reliability, thorough real-world validation is required, as well as equitable performance in diverse population.

This study shows that deep learning-based AI can successfully automate the classification of GI anomalies. This study is the work that will allow the development of AI-assisted tools for diagnostic support to endoscopists and to improve their ability to detect GI disorders in their early stages.

.

References

1) He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR*, 770–778.
2) Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *CVPR*, 4700–4708.
3) Tan, M., Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ICML*, 6105–6114.
4) Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*, 4510–4520.
5) Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *CVPR*, 2818–2826.
6) Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *ICCV*, 618–626.
7) GastroEndoNet Dataset. Available at: https://www.kaggle.com/datasets/username/gastroendonet