

Federated Transformer-Enhanced MobileNet V2 for Decentralized Polyp Detection in Colonoscopy Images

Kumar Janardan Patra

Schools of Computer Sciences, Odisha University of Technology and Research, Bhubaneswar, Odisha, 751003, India
E-mail: janardanpatra1997@gmail.com
ORCID iD: <https://orcid.org/0009-0005-0208-6244>

Jibitesh Mishra

Schools of Computer Sciences, Odisha University of Technology and Research, Bhubaneswar, Odisha, 751003, India
E-mail: jmishra@outr.ac.in
ORCID iD: <https://orcid.org/0000-0003-2761-470X>

Soumya Ranjan Nayak*

School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar-751024, Odisha, India
E-mail: nayak.soumya17@gmail.com
ORCID iD: <https://orcid.org/0000-0002-4155-884X>
*Corresponding author

Received: July 23, 2025; Revised: December 18, 2025; Accepted: January 21, 2026; Published: August 8, 2026

Abstract: Early polyp detection is vital in avoiding colorectal cancer, a top contributor to cancer-related deaths globally. Automated polyp detection has been greatly improved by deep learning, but extensive deployment tends to be hampered by patient data privacy fears. For this purpose, we suggest a federated learning (FL) paradigm that enables decentralized model training without sharing raw patient data while having high diagnostic accuracy. We first performed an extensive evaluation using DL models on a polyp dataset that we gathered. MobileNetV2 was the best performing model as per important metrics like accuracy, precision, recall, and F1 score. In order to further expand its representational power, we incorporated a transformer module into MobileNetV2 so that the model can better capture long-range dependencies and context information. Our new Transformer-Enhanced MobileNetV2 model was then implemented on several simulated clients in a federated learning scenario. This configuration enabled training over decentralized clinical data without violating patient privacy. We utilized standard FL algorithms for model averaging and evaluated the system in accuracy, precision, recall, F1 score, and convergence time. The devised approach performed extremely well with an accuracy of 98.37% and an F1 score of 0.971 while converging effectively in 27-33 rounds. These findings imply that a promising path for safe and effective medical image analysis is to combine transformer designs with lightweight models in a federated condition.

Index Terms: Polyp Detection Federated Learning, MobileNetV2, Transformer, Privacy-Preserving AI, Medical Image Analysis, Deep Learning, Colonoscopy, Lightweight CNN, Decentralized Training.

1. Introduction

As one of the primary causes of cancer-related deaths, colorectal cancer continues to be a major global health concern. A high percentage of these instances are caused by polyps abnormal tissue growths on the colon or rectum that, if they are discovered and excised before the lesions become malignant, can be prevented from causing cancer. Colonoscopy is presently the standard for the detection of polyps, but it is time-consuming and operator-dependent, which raises the risk of false-negative diagnoses, particularly for small or flat lesions. In response to the growing demand for reliable and economical screening, deep learning-based AI methods have become essential in the progression of computer-aided detection in medical imaging applications.

CNNs with deep architectures have been extensively successful in segmenting and classifying polyps from colonoscopy images. [2] Nevertheless, conventional training of high-achieving CNN models entails large amounts of annotated medical data obtained from varied patient populations. In reality, this is a major challenge because there are stringent rules around data privacy, [3] like HIPAA and GDPR, which ban centralizing sensitive patient data among institutions. In addition, heterogeneity among imaging protocols and patient populations at different medical centers also compounds the challenge in gathering an integrated dataset. Federated learning (FL) is a recent paradigm that has been identified as a promising solution to these challenges. In contrast to conventional centralized methods [4], FL allows for model training collaboratively on multiple decentralized data sources without exposing raw data to a central server. Parameters or gradients, rather than raw data, are sent to a central server for aggregation only. This method not only alleviates privacy threats but also facilitates greater inclusion of hospitals and clinics, culminating in stronger and more generalizable models. With its promise, FL has attracted greater attention in numerous healthcare domains, such as diabetic retinopathy classification, COVID-19 diagnosis from chest X-rays, and brain tumor segmentation [5].

Despite this progress, limited research has explored the integration of FL with optimized, lightweight, and attention-augmented models specifically tailored for polyp detection [6]. Existing federated learning studies in medical imaging often rely on conventional CNN architectures, which, while powerful, are computationally intensive and not well-suited for deployment on resource-constrained client devices like endoscopy equipment or mobile diagnostic tools. The combined architecture, inspired by the principles of MobileViT, [7] was then deployed across a simulated federated environment comprising multiple clients. Each client represented a local medical site and retained its own subset of the polyp dataset. During each communication round, clients trained the hybrid model locally using their private data and sent the updated model parameters to a central server, where [8] federated averaging (FedAvg) was performed. This iterative process allowed us to collaboratively train a strong global model without compromising patient privacy.

Our experiment is evaluated using publicly available colonoscopy image datasets within a realistic federated learning simulation environment. Model performance is assessed using standard metrics, including accuracy, precision, sensitivity, F1-score, and convergence speed. Experimental results demonstrate that the Transformer-enhanced MobileNetV2 consistently outperforms its non-Transformer counterpart and achieves competitive performance compared with centralized training baselines. Moreover, the model exhibits stable and efficient convergence under non-IID data distributions, reflecting practical variations in imaging conditions, hardware, and patient populations commonly encountered in real-world federated settings. To the best of our knowledge, this work represents one of the earliest efforts to integrate a lightweight proposed architecture within a scalable federated learning framework for colorectal polyp detection, supported by a comprehensive comparison with recent federated medical imaging approaches. It is important to note that the privacy benefit of the proposed framework arises from the decentralized nature of federated learning, which avoids raw data sharing. Formal differential privacy mechanisms are not incorporated in the current study. In summary, this work makes the following key contributions:

1. We present a detailed comparison of five widely-used CNN models for polyp classification and identify MobileNetV2 as the most effective and lightweight backbone for further optimization.
2. We enhance MobileNetV2 with a transformer module to capture richer contextual information, thereby improving detection accuracy and robustness.
3. We design a privacy-preserving federated learning system that enables decentralized training of the hybrid model across multiple clients.
4. We validate our approach on realistic polyp datasets and show that it offers superior performance in terms of classification metrics and communication efficiency while preserving data confidentiality.

The remaining portions of this work are organized as follows: Section 2 examines relevant research in federated learning and polyp detection. The suggested technique, including the model architecture, transformer integration, and federated training process, is explained in Section 3. The experimental design, datasets, and evaluation measures are described in Section 4. The findings are shown in Section 5, together with a comparative analysis. Section 6 wraps up the study and addresses potential avenues for future research.

2. Literature Review

Peter O.O.E., et al. [9] presents a novel approach for augmenting colonoscopy datasets using fine-tuned text-to-image generative models like Stable Diffusion and Dream Booth. It achieves high-quality synthetic image generation with a 2.36 Inception Score and significantly boosts the performance of classification models, with ViT achieving 93% accuracy. For segmentation, the use of SAM-generated masks and FPN yields strong results (IoU 0.64, Dice 0.77). Lima A.C.M., et al. [11] introduces an effective two-stage polyp detection method using transformers, combining saliency and depth map extraction with image channel fusion for enhanced accuracy. Evaluated on four public colonoscopy datasets, the method achieved impressive Average Precision scores 91% on CVC-ClinicDB, 92% on Kvasir-SEG, and 84% on CVC-ColonDB. By leveraging transformers and visual cues, the approach significantly improves polyp localization, addressing key limitations in traditional techniques. Ribeiro E. et al. [10] investigate deep

learning for automated colonic polyp classification using CNNs trained from scratch and pretrained architectures. "KoloPol," a real-time computer-aided polyp detection software (APDS), is evaluated by MD Peter K. et al. [12] in standard colonoscopy conditions. The system showed encouraging viability when used in 55 procedures, detecting 75.3% of all polyps and obtaining a polyp detection ratio (PDR) of 50.9%.

Gross S. et al. [13] demonstrates the effectiveness of a computer-based algorithm for detect colorectal polyps using narrow-banding image (NBI). Evaluated on 434 polyps from 214 patients, the algorithm achieved high diagnostic performance (accuracy of 93.1%, 95.0% of sensitivity and specificity of 90.3%), comparable to expert endoscopists and significantly outperforming nonexperts. The results remained consistent even for polyps 5 mm or smaller. Li H. et al. [14] presents NSPNet, a robust DL model for dynamic polyp detection in endoscopic videos. By combining wavelet-based noise suppression, multivariate regression refinement, and a Combined High–Low Frequency Feature Aggregation (CHLFFA) framework, NSPNet enhances feature extraction under challenging visual conditions. Ali S. et al. [15] gives a detailed insight into the Endoscopy Artefact Detection (EAD) challenge, highlighting artefact detection as having a pivotal role in improving quantitative endoscopy. Using a large, varied dataset, 23 algorithms were assessed for artefact classification, localization, and segmentation. Krenzer A. et al. [16] introduce two computerized systems for colorectal polyp classification according to the Paris (form) and NICE (texture/surface) criteria. With the use of a two-stage transformer-based methodology, the Paris classification reaches state-of-the-art accuracy of 89.35%. For NICE classification which suffers from limited data, a few-shot Deep Metric Learning algorithm provides a promising accuracy of 81.13%. The method incorporates heatmap-based explanation and large-scale ablation studies. Younas F. et al. [17] proposes a weighted ensemble of well-optimized deep learning models for multiclass colorectal polyp classification, successfully tackling challenges induced by hard-to-detect types such as serrated adenomas. With weak learners combined and hyper parameter fine-tuned, the model significantly outperforms all current methods. Tested on the UCI and PICCOLO datasets, it produces high accuracy (96.3%, 81.2%) and stable performance in precision, Cohen's Kappa, F1-score, and recall. Hsu C.M. et al. [18] introduce a DL approach for identifying and categorizing colorectal polyps from grayscale medical images, presenting a light-weight yet robust alternative to RGB and narrow-band imaging. Benchmarked on CVC-Clinic and clinical information, the model had a 95.1% detection accuracy using grayscale images which was superior to that of RGB images. The model had high recall (95.2%) but performed poorly with small polyps. Korbar B. et al. [19] describe a deep learning approach to automatic classification of five typical types of colorectal polyps in whole-slide histopathology images. It was trained on 2,074 labeled crops and tested on 239 test slides, with the model—employing residual network architecture being able to reach 93% accuracy in demonstrating its strength and consistency. Eriyanti N. A. et al. [20] offers a digital image processing system employing preprocessing, segmentation, feature extraction, and SVM classification. The workflow includes scaling, bilateral filtering, and watershed segmentation, followed by measurement of polyp characteristics. Tested on 20 images, the system achieved strong performance with 93.54% accuracy, 94.73% sensitivity, 85.71% precision, and 93.02% specificity. Lo C. M. et al. [21] proposes a quantitative polyp classification system using textural and deep convolutional neural network (DCNN) features to support rapid malignancy assessment during colonoscopy. The study employed a 1991 imaging dataset that included adenomatous and hyperplastic polyps. Among the models that were examined were DenseNet-201, ResNet-101, Inception-V3, and AlexNet. With an accuracy of 96.4%, AlexNet, which was trained from scratch, outperformed both handcrafted features and transfer learning. Gao H. et al. [22] introduces FedADT, a novel federated learning algorithm that improves convergence and robustness by incorporating adaptive step sizes and gradient differences in client updates. A moving average decay is applied to reduce noise in the derivative term. Theoretical analysis shows a convergence rate of for non-convex objectives. Experiments on MNIST and Fashion MNIST demonstrate improved performance, with ROC curves highlighting classification accuracy.

Recent advances in colorectal polyp analysis have showcased the power of deep learning, federated learning, and hybrid systems in enhancing detection, classification, and segmentation accuracy. Multiple studies demonstrate the effectiveness of CNNs, transformers, ensemble learning, and image augmentation techniques in clinical and histopathological applications. Federated learning methods preserve privacy while maintaining high performance, and few-shot learning approaches address data scarcity. These developments highlight AI's growing role in improving diagnostic accuracy, clinical decision-making and real-time support during colonoscopy, contributing significantly to colorectal cancer prevention and patient care. Despite significant progress, challenges remain in ensuring model generalizability, robustness across diverse datasets, and interpretability of results. Many studies focus on specific datasets or conditions, which may not reflect real-world variability in image quality, polyp types, or endoscopic equipment. Furthermore, although several architectures have achieved high accuracy, no single method consistently performs best across all tasks (detection, classification, segmentation). Data imbalance, especially with rare polyp types like serrated adenomas, also limits the performance of current models. Additionally, there is limited integration of these AI models into real-time clinical work flows, which hinders their practical utility.

3. Proposed Framework

The proposed framework, as given in Fig. 1, represents a federated learning approach tailored for polyp detection using three major components: Client-End Learning, Server-End Learning, and Client-Server Communication. At the

client side, image harmonization, augmentation, and advanced techniques like Visual Transformers and ensemble methods are applied for local feature learning. The server handles model aggregation and ensures IID label distribution with Non-IID evaluations. Communication ensures privacy by avoiding raw data sharing, uses lightweight model updates, and emphasizes fast convergence for efficient learning. The conceptual design of the proposed framework is illustrated in Fig. 1, while its complete implementation is detailed in Section 4 (Methodology). The following subsections describe the client-end learning, server-end aggregation, and client-server communication mechanisms in a structured manner.

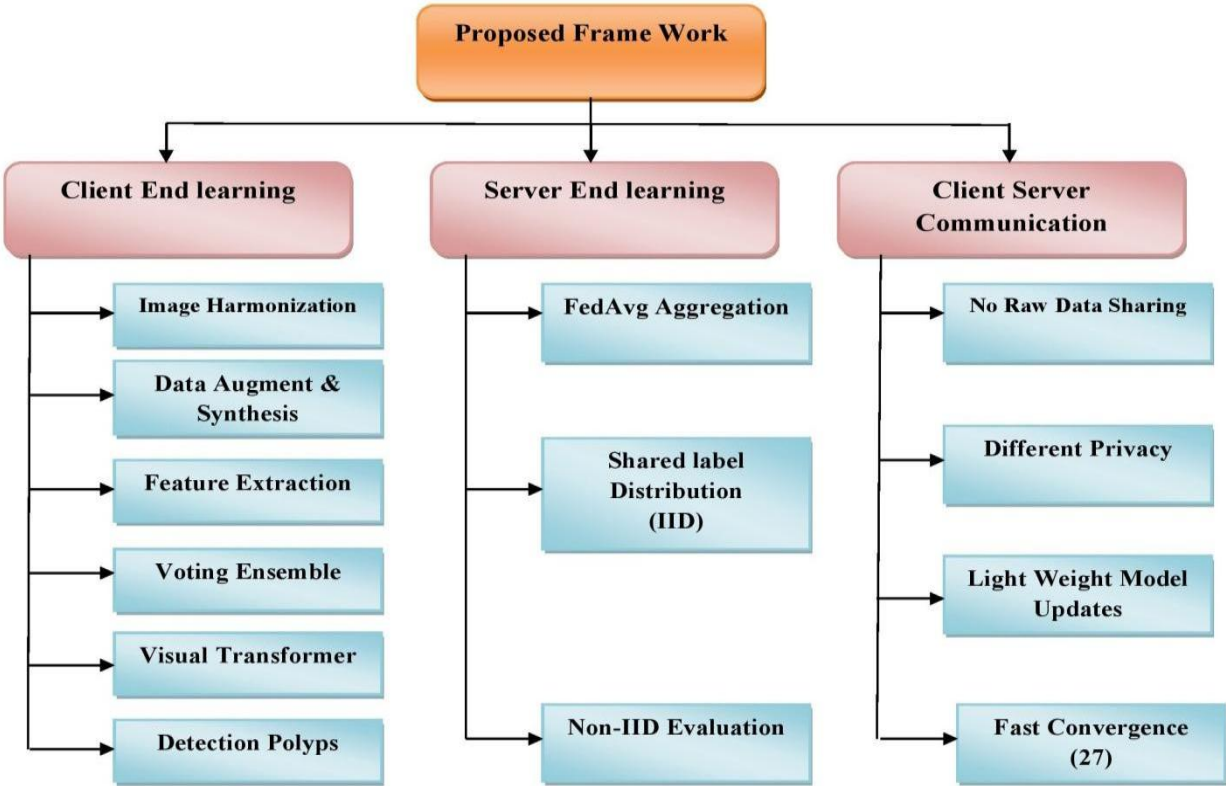


Fig. 1. Flow work of Proposed frame work Approach.

4. Methodology

4.1. Data Collection

The Kvasir dataset, which includes a variety of gastrointestinal (GI) disorders and anatomical locations, was used in this study [1, 9, 10, 23, 24]. It comprises eight categories: dyed-lifted-polyps, ulcerative colitis, polyps, normal pylorus, normal cecum, dyed-resection-margins AND normal z-line. Each category is represented by 500 images, providing a balanced distribution across all classes. These images, as illustrated in the figure, exhibit clear visual differences that are valuable for training classification models. In order to help with the automated identification and classification of gastrointestinal anomalies in endoscopic images, the dataset promotes the development of AI systems. Fig. 2 shows sample pictures.

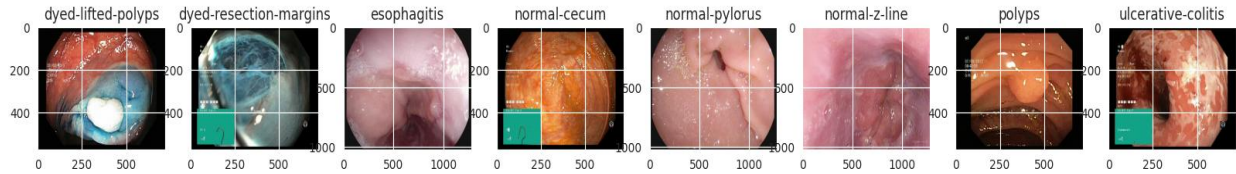


Fig. 2. Sample of collected images.

4.2. Data Preprocessing

This preprocessing chain improves polyp images for classification by combining various image processing methods. The input image is initially resized to a standard shape and subsequently may undergo optional Gaussian blurring to filter out noise. CLAHE is employed to enhance the local contrast, particularly in areas with non-uniform lighting [25]. Edge features are then extracted employing the Sobel operator, which emphasizes edges and textures relevant to model learning. Lastly, the image is normalized with min-max scaling such that pixel values range between [0, 1] to provide data that can be used for training deep learning models. These preprocessing methods are essential for image preparation prior to classification [26]. Resizing regulates image size, enabling the images to be compatible with CNN inputs. Gaussian blur assists in eliminating random noise, enhancing image sharpness without sacrificing significant features. CLAHE enhances localized contrast, rendering finer structures such as polyps more noticeable, particularly in conditions of uneven lighting. Sobel filtering improves edge details, accentuating contours and edges that help the model identify shape-based patterns. Finally, min-max normalization scales pixel intensities between 0 and 1, promoting faster and more stable training of deep learning models. The sample collection of pre-processed images is shown in Fig.3.

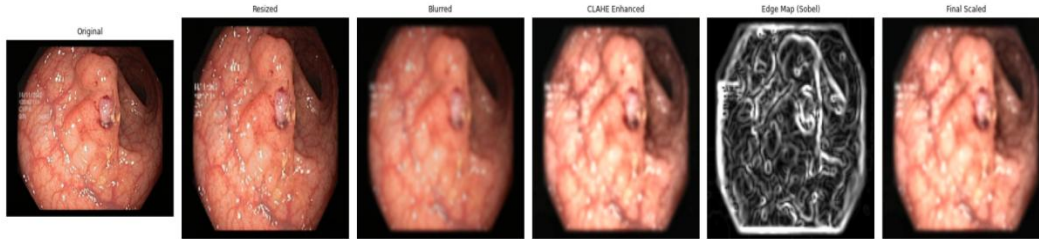


Fig. 3. Sample pre-processed images of polyp.

4.3. Data Augmentation

Data augmentation [27] is essential for improving the accuracy and generalization of deep learning models, especially in medical image classification tasks like polyp detection. Due to the limited number of annotated medical images, augmentation helps to synthetically expand the dataset by applying various realistic transformations [28]. Techniques such as rotation, flipping, cropping, zooming, adjusting brightness and contrast, and adding noise introduce variability that allows the model to better handle differences in angle, lighting, and image quality as shown in Algorithm 1. In this work, we implemented these seven augmentation techniques on polyp images. Fig. 4 displays visual examples of each method applied during preprocessing.

Algorithm 1. Image Augmentation Process

Require: Image dataset D

Ensure: Augmented dataset D'

```

1:  Load image dataset D
2:  for each image I in D do
3:      Apply Rotation to I by a random angle  $\theta \in [-30^\circ, 30^\circ]$ 
4:      Apply Flipping (horizontal and/or vertical) to I
5:      Apply Cropping to a random region of I
6:      Apply Zooming with a random scale factor  $s \in [0.8, 1.2]$ 
7:      Adjust Brightness and Contrast of I randomly
8:      G: Add Gaussian Noise with mean  $\mu = 0$  and variance  $\sigma^2$ 
9:      Add the augmented image I' to dataset D'
10: end for
11: Return D'
    
```

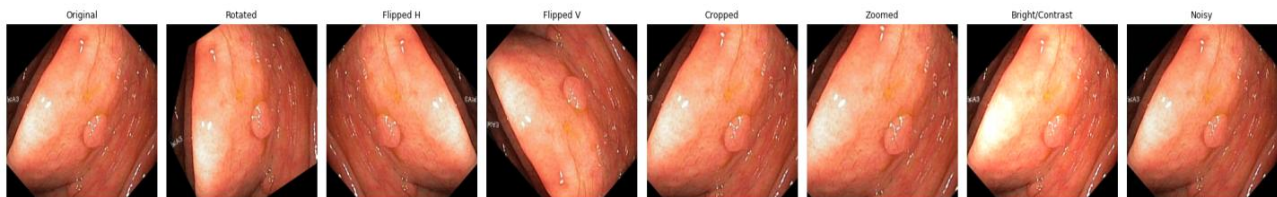


Fig. 4. Sample Augmented Images of Polyps.

4.4. StyleGAN2

After data collection and preprocessing, StyleGAN2 [29] is leveraged to augment the dataset by generating high-quality synthetic images that closely mimic the original samples. This advanced GAN model is particularly effective at capturing fine details and intricate patterns, making it well-suited for generating realistic polyp images in medical imaging tasks. By creating diverse, high-resolution images for each class, StyleGAN2 [30] helps improve data diversity and overcome limitations in sample size. The primary objective of utilizing StyleGAN2 is to enhance the model's ability to generalize, prevent overfitting, and improve its performance on previously unseen gastrointestinal images. Table 1 illustrates the model's performance over multiple epochs, showcasing important metrics such as discriminator accuracy, discriminator loss, generator loss, FID score, and inception score (IS). The StyleGAN2 model was used to generate 4,000 synthetic images, accounting for approximately 40% of the final training dataset. These synthetic samples were uniformly distributed across all federated clients and were utilized consistently in both centralized and federated training experiments. The quantitative impact of StyleGAN2-based data synthesis is analyzed through an ablation study presented in Table 8.

Table 1. Performance of the StyleGAN2 model on our dataset.

Epoch	D. Acc	D. Loss	G. Loss	FID Score	Inception Score (IS)
10	78	0.38	1.1	115.2	2.15 ±0.05
20	81.5	0.43	0.75	75	2.45 ±0.07
40	84.5	0.49	0.48	40.8	2.70 ±0.08
60	87.2	0.53	0.24	24.1	2.90 ±0.10
80	89	0.55	0.14	15.3	3.05 ±0.11
100	90.5	0.57	0.06	12.8	3.35 ±0.09

4.5. Voting-Based Model Selection

In the initial phase of this study, we conducted a comparative evaluation as shown in Fig. 5, of five popular convolutional neural network (CNN) architectures: MobileNetV2, ResNet50, ResNet152, VGG16, and VGG19. These models were assessed based on their performance [31] on polyp image classification tasks using centralized training. The comparison considered key findings are accuracy, precision, recall, and F1 findings, as well as computational efficiency. Among the evaluated models, MobileNetV2 consistently outperformed the others, offering a favorable balance between lightweight design and predictive performance. Its low memory footprint and fast inference time made it particularly well-suited for deployment in a federated learning (FL) environment, where client devices may have limited computational capabilities. Based on these observations, MobileNetV2 was selected as the base model for further development and integration into the FL framework. The comparison results are given in the Results and Discussion section. Algorithm 2 given below presents a structured symbolic pseudo-code for the complete pipeline, combining DL model selection via performance voting and classification refinement through Transformer integration.

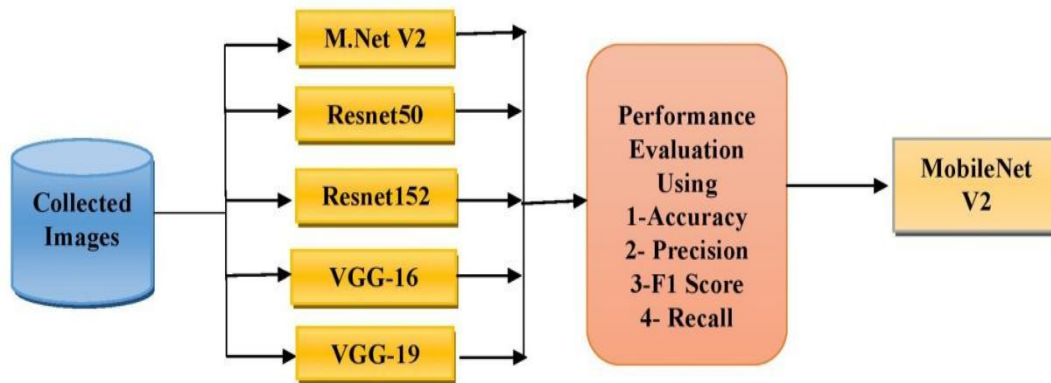


Fig. 5. Voting-based model selection architecture.

$$M_j = \sum_{i=1}^k W_i P_{ij} \quad (1)$$

Where: M_j is Aggregated score for class j

W_i = Weighted assigned to model i , $\sum W_i = 1$

P_{ij} = Probability score for model i for class j

Algorithm 2. Voting and Transformer Integration for Polyp Detection

Require: $D, M = \{M1, M2, ..., Mn\}, T$ //D: Image Dataset
Ensure: y //M: DL Models
 1: $R \leftarrow \emptyset$ //R: Metric Results
 2: **for** $M_i \in M$ **do**
 3: $M_i \leftarrow \text{Train}(M_i, D)$
 4: $(\alpha, \beta, \gamma, \delta) \leftarrow \text{Evaluate}(M_i)$ // $\alpha, \beta, \gamma, \delta$: Performance Matrices
 5: $R \leftarrow R \cup \{(\alpha, \beta, \gamma, \delta)\}$
 6: **end for**
 7: $M^* \leftarrow \text{argmax}(\text{Vote}(R))$ //M*: Selected Model
 8: $I \leftarrow \text{Input}(D)$
 9: $\Phi \leftarrow M^*(I)$ // Φ : Feature map
 10: $\{\phi_1, \phi_2, ..., \phi_k\} \leftarrow \text{Flatten}(\Phi)$ // Φ_i : Flattened patch
 11: $\{\phi_1 + P, ..., \phi_k + P\}$ //P: Position Encoding
 12: $\theta_T \leftarrow \text{Init}()$ // θ_T : Transformer parameters
 13: **repeat**
 14: $\{\phi' i\} \leftarrow \text{MultiHead}(\{\phi_i + P\}, \theta_T)$
 15: $\{\phi'' i\} \leftarrow \text{FFN}(\{\phi' i\})$
 16: **until** Converged = True
 17: $G \leftarrow \text{GAP}(\{\phi'' i\})$ //G: Global pooled vector
 18: $y \leftarrow \text{ReLU}(W \cdot G + b)$ //W, b: FC layer weights, bias
 19: **return** y

4.6. Transformer Integration

To further enhance the model's ability to capture global contextual relationships, especially important in medical imaging tasks involving subtle or spatially dispersed features, we introduced a Transformer [32] module into the architecture, as shown in Fig. 6 below. The motivation for this integration stemmed from the limitations observed in CNNs, which, although effective at local feature extraction, struggle with modeling long-range dependencies. By combining MobileNetV2 with a Transformer block [33], we aimed to improve the model's capacity for detecting complex polyp structures without significantly increasing its computational load. MobileNetV2 serves as the feature extractor, extracting hierarchical features from input colonoscopy images using its depth wise separable convolutions and inverted residual blocks. Feature maps from the final bottleneck layer are used as input to the attention mechanism [34].

The Transformer module integrated with MobileNetV2 consists of two lightweight encoder layers with an embedding dimension of 128 and four multi-head self-attention heads. Each encoder layer includes a positional encoding stage, a multi-head attention block, and a feedforward network with a hidden dimension of 256, followed by layer normalization and residual connections. A dropout rate of 0.1 is applied after the attention and feed forward layers to mitigate overfitting. Learnable positional embeddings are used to preserve spatial information after feature map flattening. The complete Transformer hyperparameter configuration is summarized in Table 2. This module allows the model to reason about spatially distant features, particularly useful for improving the detection of small or irregular polyps. The whole working process of transformer learning using polyp data is given in Fig. 7. Classification Head: The enhanced feature representation is processed through global average pooling and passed to a fully connected layer with ReLU activation to classify the input into one of the predefined polyp categories. This transformer-enhanced MobileNetV2 architecture combines efficiency, robustness, and scalability, making it highly suitable for privacy-preserving training in federated setups. The decision to integrate the transformer module was guided by empirical results and architectural constraints, ensuring that improvements in accuracy do not come at the cost of client-side deployability.

$$\text{Attention}(A, B, C) = \text{softmax}\left(\frac{AB^T}{\sqrt{d_k}}\right)C \quad (2)$$

Where: $A = XW^a, B = XB^b, C = XW^c$: Query, Key, and Value Projections

W_a, W_b, W_c : Learnable weighted matrices

d_k : Dimension of the key vectors

The final refined prediction $\hat{Y}_{\text{transformer}} = \arg \max (\text{MLP}(\text{Transformer}(M)))$

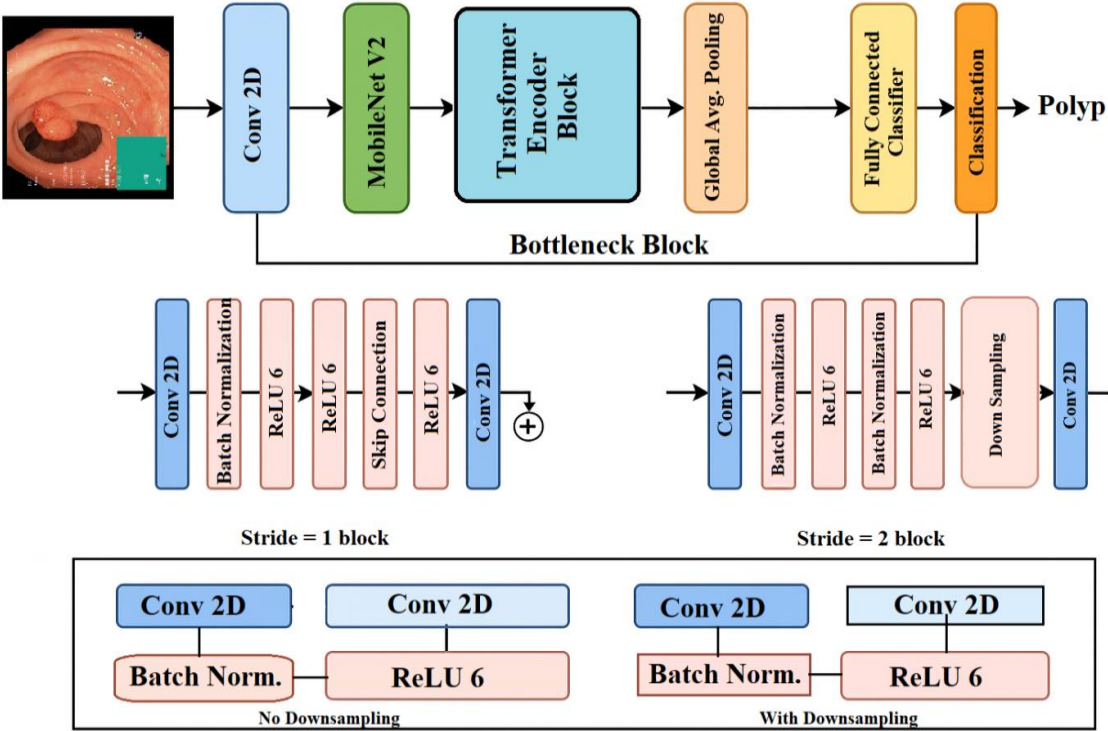


Fig. 6. Detailed architecture of the Transformer-enhanced MobileNetV2, illustrating feature map flattening, positional encoding, multi-head self-attention, feedforward layers, and dimensional flow across encoder blocks.

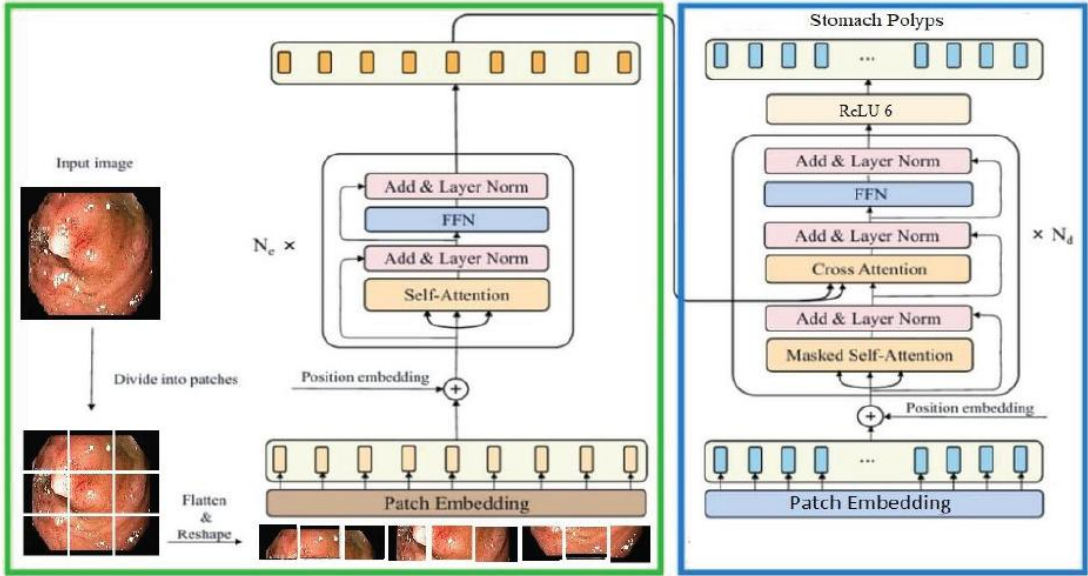


Fig. 7. Working process of the transformer on flattened data.

Table 2. Transformer module hyperparameter configuration.

Parameter	Value
Number of encoder layers	2
Embedding dimension	128
Number of attention heads	4
Feedforward hidden size	256
Dropout rate	0.1
Positional encoding	Learnable
Activation function	ReLU
Normalization	Layer Normalization

4.7. Federated Learning Setup

To ensure patient data privacy while enabling collaborative model training, we adopted a federated learning (FL) [35] strategy in which multiple simulated clients locally trained a shared model on their respective data subsets as shown in Fig. 8. This decentralized framework eliminates the need to transfer raw medical data to a central repository, thereby aligning with privacy regulations such as HIPAA and GDPR.

$$\theta^{t+1} = \sum_{i=1}^k \frac{n_i}{n} \theta_i^t \quad (3)$$

Where: θ^{t+1} : Global Parameter after t

θ_i^t : Model Weight on client i

n_i : Data sample on client i, $n = \sum_{i=1}^k n_i$

In our experimental setup, we simulated 15 clients, each representing an independent medical center with its own subset of polyp image data. The overall dataset was partitioned across these clients in a non-IID Label Skew manner to better mimic real-world scenarios where medical data can vary due to demographic, clinical, and device-based differences. For instance, some clients received more samples of specific polyp types, while others had varying image qualities and resolutions. The training process proceeded in multiple communication rounds, where each round involved the following steps:

1. The central server sent the current global model parameters to all clients.
2. Each client performed local training on its own dataset for a fixed number of epochs.
3. The updated model parameters from all clients were sent back to the server.
4. The server aggregated these updates to refine the global model using the Federated Averaging (FedAvg) algorithm.

We used a batch size of 32, an initial learning rate of 0.001, and 100 communication rounds, with each client executing three local epochs every round, as shown in Table 3. Because of its ease of use and efficiency in situations with moderate client heterogeneity, the FedAvg [36] algorithm was chosen. To simulate realistic inter-institutional data heterogeneity, a label-skew non-IID partitioning strategy was adopted. Specifically, data were distributed across clients using a Dirichlet distribution, where each client was assigned a dominant subset of polyp classes. The degree of

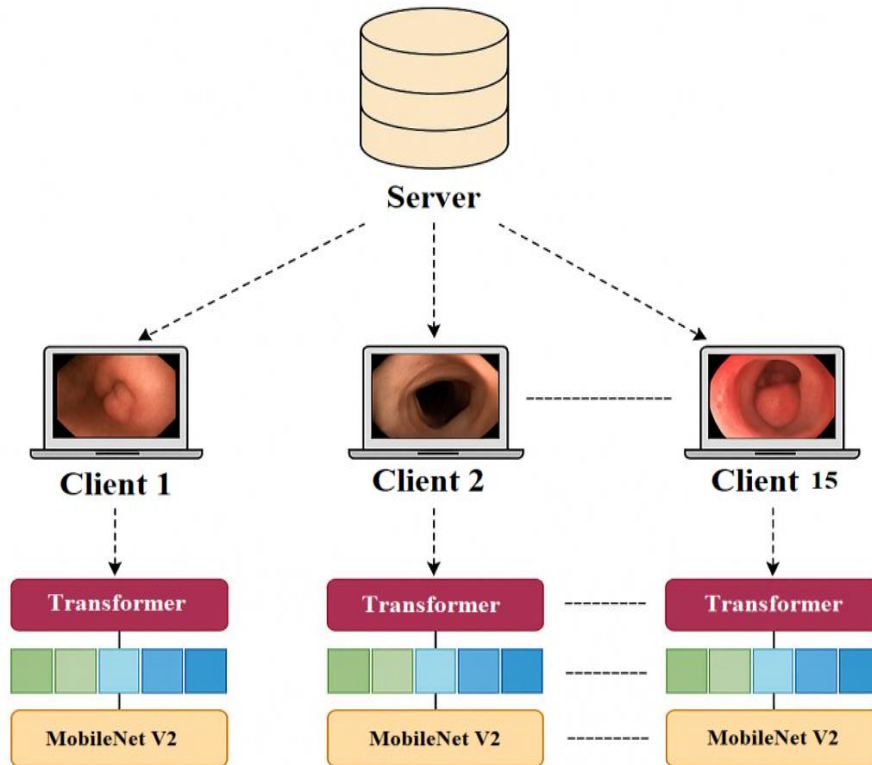


Fig. 8. Architecture of integrating client with server using FL.

heterogeneity was controlled using the Dirichlet concentration parameter α , with lower values indicating higher data imbalance. In our experiments, α was set to 0.5, representing a moderate level of heterogeneity commonly observed across medical center, FedAvg was selected as the aggregation strategy in this study due to its efficiency and widespread adoption in medical federated learning. Investigating alternative aggregation methods such as FedProx and FedDyn under heterogeneous clinical data settings is an important direction for future work. This federated setup ensured that the Transformer-Enhanced MobileNetV2 model could be collaboratively trained while respecting data locality, thus offering a practical solution for privacy-preserving medical image analysis.

Table 3. Experiment setup for communication between client and server.

Parameter	Value
Number of Clients	15
Data Distribution	Non-IID (Label skew)
FL Algorithm	FedAvg
Learning Rate	0.001
Local Epochs per Round	5
Batch Size	32
Communication Rounds	100

5. Result and Discussion

For this work, we used a system whose configuration includes a GPU-enabled workstation (NVIDIA RTX 3080 or higher), 64GB RAM, Intel i9 CPU, running Python 3.10 with TensorFlow/PyTorch, and FL frameworks like Flower. Clients simulate federated nodes; server handles model aggregation and evaluation. The original Kvasir dataset comprises 4,000 images evenly distributed across eight gastrointestinal categories. To enhance data diversity while avoiding excessive synthetic bias, 2,000 additional samples were generated using conventional augmentation techniques, and 4,000 high-quality synthetic images were produced using the StyleGAN2 framework. As a result, the final dataset used for model training consists of 10,000 images, combining real, augmented, and synthetic samples, as summarized in Table 4. The complete dataset was then divided into training (64.5%, 6,450 images), testing (18.9%, 1,890 images), and validation (16.6%, 1,660 images) subsets, as illustrated in Fig. 9. This structured data composition and splitting strategy ensures robust learning, improved generalization, and reliable performance evaluation across all categories. Before presenting the experimental results, the complete set of training and optimization hyperparameters used in this study is summarized in Table 5, to ensure clarity and reproducibility.

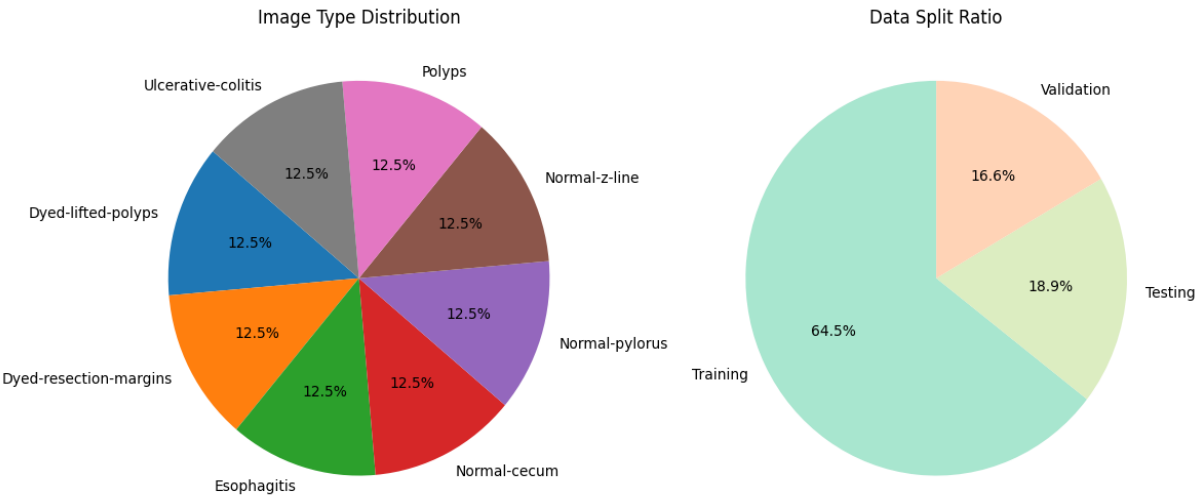


Fig. 9. Total images and splitting ration.

Table 4. Performance of DL models using Voting technique.

Dataset Stage	Description	Number of Images
Original Dataset	Kvasir dataset (8 classes × 500 images)	4,000
Conventional Augmentation	Rotation, flipping, cropping, zooming, brightness/contrast, noise	2,000
StyleGAN2 Synthetic Images	GAN-generated realistic polyp images	4,000
Final Dataset Used for Training	Combined real, augmented, and synthetic images	10,000

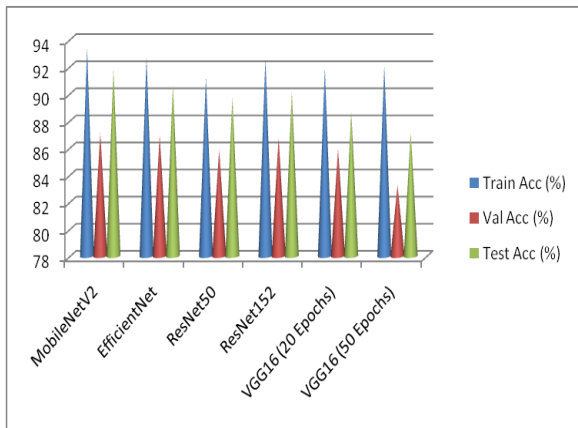
Table 5. Summary of training hyperparameters.

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.001
LR Scheduler	Cosine Annealing
T_max	50
eta_min	1.00E-06
Batch size	32
Weight decay	1.00E-04
Momentum	0.9
Early stopping	Patience = 10 epochs
Augmentation probability	0.5 per operation
Voting weights (Section 4.5)	Uniform (Wi = 1)
FL local epochs	5

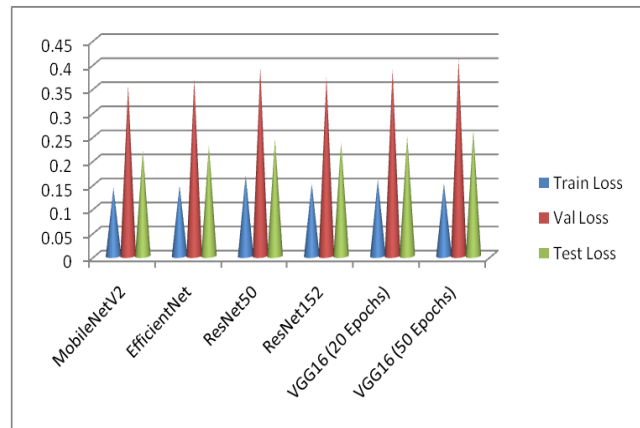
Based on the performance results given in Table 6 and graphically shown in Fig. 10(a, b) for all models evaluated via the voting technique, MobileNetV2 was selected as the base architecture due to its proven efficiency and strong performance on edge devices. It utilizes depth wise separable convolutions and inverted residual connections to reduce computation while preserving accuracy. In this architecture, MobileNetV2 serves as the core feature extractor, processing the input colonoscopy image into a compact and informative feature map. To enhance model performance, a Transformer block was integrated into MobileNetV2, permitting global context and dependence on long-range relationships to be captured by the network. This hybrid architecture significantly improved results, achieving a Test Accuracy of 98.37% and lower loss than the original model. The integration demonstrates the effectiveness of combining MobileNetV2 efficiency with Transformer attention capabilities in medical image classification. The comparison of MobileNetV2 and integration of transformer with MobilenetV2 are given in Table 7.

Table 6. Performance of DL models using Voting technique.

Model	Training Loss	Training Accuracy (%)	Validation Loss	Validation Accuracy (%)	Testing Loss	Testing Accuracy (%)	MCC Score	Epoch Time (min:sec)
MobileNetV2	0.1423	93.39	0.3555	87.13	0.2198	91.86	0.907	03:16
EfficientNet	0.1468	92.7	0.3682	86.92	0.2299	90.65	0.8932	04:20
ResNet50	0.1683	91.24	0.3921	85.87	0.2435	89.72	0.8821	05:15
ResNet152	0.1511	92.65	0.3728	86.75	0.2372	90.15	0.891	07:43
VGG16	0.1589	91.89	0.389	85.92	0.2493	88.7	0.8703	05:40
VGG19	0.1517	92.05	0.4101	83.25	0.2617	87.15	0.8511	04:22



(a)



(b)

Fig. 10. (a) Graph of test, train, and validation accuracy; (b) Graph of test, train, and validation loss.

Standard measures including precision, accuracy, and recall were used to evaluate the model's performance, alongside advanced evaluation criteria including MCC, IoU, and Dice loss. Focal and Jaccard losses were also employed to address class imbalance and improve segmentation reliability.

$$\text{Accuracy}(\alpha_i) = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (4)$$

$$\text{Precision}(\beta_i) = \frac{TP_i}{TP_i + FP_i} \quad (5)$$

$$\text{Recall}(\gamma_i) = \frac{TP_i}{TP_i + FN_i} \quad (6)$$

$$F1 - \text{Score}(\delta_i) = \frac{2 \cdot \beta_i \cdot \gamma_i}{\beta_i + \gamma_i} \quad (7)$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

$$\text{FocalLoss}(\mathcal{F}_{\text{focal}}) = -\alpha_t(1 - \rho_t)^\gamma \log(\rho_t) \quad (9)$$

Where: ρ_t is the predicted probability of the true class

α_t is a weighting factor

γ is the focusing parameter

$$\text{Jaccard Loss} = 1 - \left| \frac{P \cap G}{P \cup G} \right| = 1 - \frac{\sum_i P_i G_i}{\sum_i P_i + \sum_i G_i - \sum_i P_i G_i} \quad (10)$$

Where: $P_i \in P$: Prediction and $G_i \in G$: Ground Truth

Table 7. Performance table of transformer integration with MobileNetV2.

Model	Test Loss	Test Accuracy (%)	Precision	Sensitivity	Specificity	MCC	Focal Loss	Jaccard Loss	IoU	Dice Loss
MobileNetV2	0.2198	91.86	0.905	0.912	0.918	0.907	0.1842	0.1628	0.851	0.179
MobileNetV2 + Transformer	0.0876	98.37	0.981	0.986	0.988	0.9764	0.0723	0.0584	0.943	0.062

$$\text{Intersection over Union (IoU)} = \left| \frac{P \cap G}{P \cup G} \right| = \frac{\sum_i P_i G_i}{\sum_i P_i + \sum_i G_i - \sum_i P_i G_i} \quad (11)$$

$$\text{Dice Loss} = 1 - \frac{2 \cdot \sum_i P_i G_i + \epsilon}{\sum_i P_i + \sum_i G_i + \epsilon} \quad (12)$$

Where ϵ is a small constant to prevent deviation by zero

To systematically evaluate the contribution of conventional data augmentation and StyleGAN2-based synthesis, a progressive ablation study was conducted using three training configurations. As summarized in Table 8, conventional augmentation improves performance over the original dataset, while the inclusion of StyleGAN2-generated synthetic images yields a substantial gain across all evaluation metrics, particularly in sensitivity and MCC, highlighting improved robustness and generalization. The confusion matrix given below Fig. 11, shows strong classification performance, with most classes achieving high correct predictions along the diagonal.

Table 8. Ablation study on the effect of data augmentation and StyleGAN2-based synthesis.

Training Data Configuration	Number of Images	Accuracy (%)	F1-score	Sensitivity	MCC
Original Kvasir dataset	4,000	95.74	0.941	0.948	0.912
Original + Conventional Augmentation	6,000	96.82	0.953	0.961	0.931
Original + Augmentation + StyleGAN2 (Proposed)	10,000	98.37	0.971	0.986	0.976

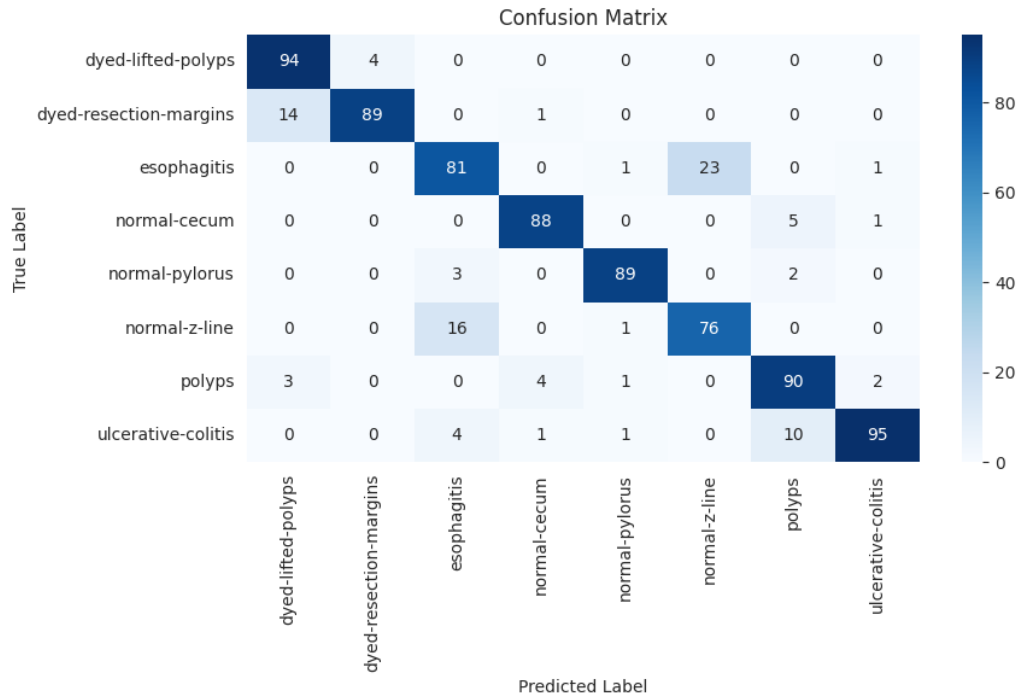


Fig. 11. Confusion Matrix of proposed model.

To support efficient training in a federated learning environment where computational resources and bandwidth can vary across clients, we use our proposed model. This architecture balances computational efficiency with enhanced feature learning, making it ideal for decentralized medical image analysis tasks such as polyp detection. While standard evaluation metrics such as F1-score and accuracy are important for measuring predictive performance, Federated Learning (FL) systems must also be assessed in terms of communication efficiency, client-side cost, scalability, and robustness under data heterogeneity. These additional metrics offer deeper insight into the system's feasibility in real-world decentralized environments like hospitals or edge devices. Below, we report and analyze key FL-specific metrics obtained from our experiments using the MobileNetV2 + Transformer architecture in a federated setup. The convergence rate given in Table 9., indicates how many communication rounds (global epochs) were required for the global model to achieve stable, optimal accuracy. A faster convergence implies better communication efficiency and lower overall resource usage. We measured the total amount of data exchanged i.e., Communication Cost stored in Table 10, between clients and the server during training. This includes model weight uploads and global model downloads. Despite slightly higher model size due to the Transformer block, the communication overhead remained lightweight, making it suitable for edge deployment. The computational efficiency on client side stored in Table 11 was profiled in terms of training time per epoch per client and estimated memory consumption during local updates. The added Transformer block slightly increased memory usage and training time but stayed within practical limits for modern clinical edge hardware. We tracked the percentage of clients participating in each communication round (performance given in Table 12) in each communication round and measured how the model performed under partial participation scenarios.

Table 9. Convergence Rate

Model	Accuracy (%)	Convergence Rounds
MobileNetV2 (FL)	91.86	38
MobileNetV2 + Transformer (FL)	98.37	27

Table 10. Communication Cost

Model	Total Upload (MB)	Total Download (MB)	Avg Per Round (MB)
MobileNetV2 (FL)	12.6	13	0.65
MobileNetV2 + Transformer	13.8	14.2	0.75

Table 11. Client-Side Computation Cost

Model	Avg. Epoch Time (s)	Peak RAM Usage (MB)	Estimated Energy per Epoch (Wh)
MobileNetV2 (FL)	2.4	720	0.018
MobileNetV2 + Transformer	2.8	810	0.021

Table 12. Client Participation Rate Impact

Participation Rate (%)	Accuracy (%)
100	98.37
80	97.6
60	97.1

Our model showed robustness even under reduced client participation, making it resilient in practical non-synchronous environments. To test the effect of data imbalance, we compared model performance under IID (random data split) and non-IID (realistic clinical-like skewed) distributions. Data Heterogeneity (IID vs. non-IID) impact shown in below Table 13. Under non-IID conditions, performance slightly declined and convergence was slower, but the model still maintained high accuracy and stable learning demonstrating good generalizability. To account for the stochastic nature of federated learning, all experiments were repeated five independent times using different random seeds affecting client sampling and local training initialization. Performance metrics are therefore reported in Table 14, as mean $\pm$ standard deviation. Furthermore, statistical significance tests were conducted to compare federated and centralized training results. A paired statistical significance analysis was performed between centralized and federated training results. A paired t-test ($p < 0.05$) confirmed that the performance improvements achieved by the proposed federated framework are statistically significant across all reported metrics. These results in Table 15, indicate that the proposed Transformer-enhanced MobileNetV2 framework maintains stable performance under varying degrees of data heterogeneity, demonstrating suitability for real-world federated medical imaging environments. After evaluating the proposed framework under various federated settings, we further compare its performance with recent state-of-the-art federated and centralized polyp detection methods, as summarized in Table 16, to highlight its relative advantages.

Table 13. Data Distribution Impact.

Data Split Type	Accuracy (%)	Convergence Rounds
3	98.37	27
5	97.9	29
10	97.4	31
15	96.8	33

Table 14. Statistical performance comparison over multiple runs (mean $\pm$ std)

Training Setting	Accuracy (%)	F1-score	Dice
Centralized	96.91 $\pm$ 0.42	0.958 $\pm$ 0.006	0.931 $\pm$ 0.008
Federated (Proposed)	98.37 $\pm$ 0.31	0.971 $\pm$ 0.004	0.943 $\pm$ 0.006

Table 15. Impact of non-IID data heterogeneity on federated learning performance

Dirichlet α	Data Distribution	Accuracy (%)	Convergence Rounds
∞ (IID)	Uniform	98.37	27
1	Mild non-IID	97.9	29
0.5	Moderate non-IID	96.8	33

Table 16. Comparative performance with state-of-the-art methods

Method	Learning Setting	Dataset	Accuracy (%)	F1-score	Communication Rounds	Model Size
FedADT [22]	Federated	MNIST / Fashion-MNIST	96.2	—	40	—
GastroNet [7]	Centralized	Kvasir	96.1	0.958	—	14.8 MB
NSPNet [14]	Centralized	CVC-ClinicDB	95.4	0.949	—	—
Fed-Polyp (Cao et al.) [35]	Federated	Kvasir-SEG	95.8	0.951	35	—
Proposed Method	Federated	Kvasir	98.37	0.971	27	9.4 MB

6. Conclusion and Future Research Directions

In this work, we proposed a deep learning architecture for polyp detection utilizing federated learning, which is both effective and privacy-preserving. We started by analyzing five cutting-edge CNN models, and the best-performing architecture in terms of performance and computational cost was found to be MobileNetV2. To improve its performance further, we added a lightweight Transformer module, which enhanced the ability of the model to learn global contextual features that were vital for precise polyp classification. The hybrid model proposed was used in a federated learning (FL) environment with the FedAvg algorithm while maintaining data privacy in multiple simulated clients. Comprehensive experiments showed that our model exhibited robust performance with minimal communication and computational overhead. In particular, the federated MobileNetV2 + Transformer model resulted in 98.37% accurate classification, F1-score of 0.971, and converged at 27-33 communication rounds. The size of the model

remained compact at around 9.4 MB, which is ideal to deploy on edge devices in real-world clinical settings. Unlike prior studies limited to small-scale federated simulations, this work evaluates federated learning scalability using up to 15 heterogeneous clients, providing a more realistic validation of decentralized medical image analysis. Future directions will include extending the evaluation of the proposed federated framework to additional publicly available colorectal polyp benchmarks, such as CVC-ClinicDB and HyperKvasir, to further investigate cross-dataset generalization and robustness across diverse imaging conditions. In addition, adaptive federated learning strategies and enhanced privacy-preserving mechanisms, including differential privacy and secure aggregation, will be considered to further improve scalability, reliability, and compliance with healthcare data protection regulations.

All the Declarations and Statements

Author Contributions Statement

Kumar Janardan Patra – Conceptualization, Data Curation, Model Development, and Writing – Original Draft: Conducted data collection and preprocessing, developed and deployed the proposed model, and prepared the initial manuscript.

Jibitesh Mishra – Supervision and Methodology: Provided overall guidance, supervised the research work, and contributed to the design and validation of the proposed methodology.

Soumya Ranjan Nayak – Visualization, Formal Analysis, and Writing – Review & Editing: Assisted in result analysis, prepared visualizations, ensured technical accuracy, and contributed to reviewing and refining the manuscript.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

No funding available.

Data Availability Statement

This study analyzed the publicly available Kvasir dataset.

Ethical Declarations

Not Applicable.

Acknowledgments

We would like to express our gratitude to the reviewers for their precise and succinct recommendations that improved the presentation of the results obtained.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, AI-assisted tools were used to improve language clarity and readability. The authors take full responsibility for the content, originality, and integrity of the work.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence

DL - Deep Learning

FL - Federated Learning

CNN - Convolutional Neural Network

ViT - Vision Transformer

MCC - Matthews Correlation Coefficient

IoU - Intersection over Union

Appendix A/B/C..., with appendix title

None.

References

- [1] Patra, K. J., Mishra, J., Dash, S. K., & Jayasingh, S. K. (2024, November). Enhancing Diagnostic Precision for Early Detection of Gastric Polyps with U-Net3+ Segmentation. In *2024 3rd Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)* (pp. 1-5). IEEE.
- [2] Ribeiro, J., Nóbrega, S., & Cunha, A. (2022). Polyps detection in colonoscopies. *Procedia Computer Science*, 196, 477-484.
- [3] Doo, F. X. (2025). Artificial Intelligence (AI) in Abdominal Imaging. In *What Radiology Residents Need to Know: Abdominal Radiology* (pp. 347-359). Cham: Springer Nature Switzerland.
- [4] Hasan, M. M., & Rahman, M. M. (2025). Privacy-preserving polyp segmentation using federated learning with differential privacy. *Smart Health*, 36, 100551.
- [5] Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, 11(5), 7374-7398.
- [6] Devkota, A., Amireskandari, A., Palko, J., Thakkar, S., Adjero, D., Jiang, X., ... & Gyawali, P. K. (2025). Federated Foundation Model for GI Endoscopy Images. *arXiv preprint arXiv:2505.24108*.
- [7] Yasmin, F., Hassan, M. M., Hasan, M., Zaman, S., Bairagi, A. K., El-Shafai, W., ... & Chun, Y. C. (2023). GastroNet: Gastrointestinal polyp and abnormal feature detection and classification with deep learning approach. *IEEE Access*, 11, 97605-97624.
- [8] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.
- [9] Ejiga Peter, O. O., Adeniran, O. T., John-Otumu, A. M., Khalifa, F., & Rahman, M. M. (2025). Text-Guided Synthesis in Medical Multimedia Retrieval: A Framework for Enhanced Colonoscopy Image Classification and Segmentation. *Algorithms*, 18(3), 155.
- [10] de Moura Lima, A. C., de Paiva, L. F., Bráz, G., de Almeida, J. D. S., Silva, A. C., Coimbra, M. T., & de Paiva, A. C. (2023). A two-stage method for polyp detection in colonoscopy images based on saliency object extraction and transformers. *IEEE Access*, 11, 76108-76119.
- [11] Ribeiro, E., Uhl, A., Wimmer, G., & Häfner, M. (2016). Exploring deep learning and transfer learning for colonic polyp classification. *Computational and mathematical methods in medicine*, 2016(1), 6584725.
- [12] Klare, P., Sander, C., Prinzen, M., Haller, B., Nowack, S., Abdelhafez, M., ... & Wittenberg, T. (2019). Automated polyp detection in the colorectum: a prospective study (with videos). *Gastrointestinal endoscopy*, 89(3), 576-582.
- [13] Gross, S., Trautwein, C., Behrens, A., Winograd, R., Palm, S., Lutz, H. H., ... & Tischendorf, J. J. (2011). Computer-based classification of small colorectal polyps by using narrow-band imaging with optical magnification. *Gastrointestinal endoscopy*, 74(6), 1354-1359.
- [14] Li, H., Zhen, G., Chu, C., Ma, Y., & Zhao, Y. (2024). Noise-Perception Multi-Frame Collaborative Network for Enhanced Polyp Detection in Endoscopic Videos. *Electronics*, 14(1), 62.
- [15] Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., ... & Rittscher, J. (2020). An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific reports*, 10(1), 2748.
- [16] Krenzer, A., Heil, S., Fitting, D., Matti, S., Zoller, W. G., Hann, A., & Puppe, F. (2023). Automated classification of polyps using deep learning architectures and few-shot learning. *BMC Medical Imaging*, 23(1), 59.
- [17] Younas, F., Usman, M., & Yan, W. Q. (2023). A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence*, 53(2), 2410-2433.
- [18] Hsu, C.-M., Hsu, C.-C., Hsu, Z.-M., Shih, F.-Y., Chang, M.-L., & Chen, T.-H. (2021). Colorectal Polyp Image Detection and Classification through Grayscale Images and Deep Learning. *Sensors*, 21(18), 5995. <https://doi.org/10.3390/s21185995>
- [19] Korbar, B., Olofson, A. M., Miraflor, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L., ... & Hassanpour, S. (2017). Deep learning for classification of colorectal polyps on whole-slide images. *Journal of pathology informatics*, 8(1), 30.
- [20] Eriyanti, N. A., Sigit, R., & Harsono, T. (2021, September). Classification of colon polyp on endoscopic image using support vector machine. In *2021 International Electronics Symposium (IES)* (pp. 244-250). IEEE.
- [21] Lo, C.-M., Yeh, Y.-H., Tang, J.-H., Chang, C.-C., & Yeh, H.-J. (2022). Rapid Polyp Classification in Colonoscopy Using Textural and Convolutional Features. *Healthcare*, 10(8), 1494. <https://doi.org/10.3390/healthcare10081494>
- [22] Gao, H., Wu, Q., Zhao, X., Zhu, J., & Zhang, M. (2023). FedADT: An Adaptive Method Based on Derivative Term for Federated Learning. *Sensors*, 23(13), 6034. <https://doi.org/10.3390/s23136034>
- [23] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., De Lange, T., Johansen, D., & Johansen, H. D. (2019, December). Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling* (pp. 451-462). Cham: Springer International Publishing.
- [24] Mahmood, F., Yang, Z., Durr, N. J., Chen, R., Xu, W., & Borders, D. (2019, March). Polyp segmentation and classification using predicted depth from monocular endoscopy. In *Medical Imaging 2019: Computer-Aided Diagnosis* (p. 36). SPIE. doi: 10.1117/12.2513117
- [25] Khan, M. A., Sarfraz, M. S., Alhaisoni, M., Albeshier, A. A., Wang, S., & Ashraf, I. (2020). StomachNet: Optimal deep learning features fusion for stomach abnormalities classification. *IEEE Access*, 8, 197969-197981.
- [26] Mohammad, F., & Al-Razgan, M. (2022). Deep feature fusion and optimization-based approach for stomach disease classification. *Sensors*, 22(7), 2801.
- [27] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1), 1-48.
- [28] Mikołajczyk, A., & Grochowski, M. (2018, May). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)* (pp. 117-122). IEEE.

- [29] Viazovetskiy, Y., Ivashkin, V., & Kashin, E. (2020). Stylegan2 distillation for feed-forward image manipulation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16* (pp. 170-186). Springer International Publishing.
- [30] Zeng, P., Song, R., Chen, S., Li, X., Li, H., Chen, Y., ... & Chen, Z. (2025). Expert-guided StyleGAN2 image generation elevates AI diagnostic accuracy for maxillary sinus lesions. *Communications Medicine*, 5(1), 1-10.
- [31] Tajbakhsh, N., Gurudu, S. R., & Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2), 630-644.
- [32] Elkarazle, K., Raman, V., Then, P., & Chua, C. (2023). Improved colorectal polyp segmentation using enhanced MA-NET and modified Mix-ViT transformer. *IEEE Access*, 11, 69295-69309.
- [33] Fu, B., Zhang, M., He, J., Cao, Y., Guo, Y., & Wang, R. (2022). StoHisNet: A hybrid multi-classification model with CNN and Transformer for gastric pathology images. *Computer Methods and Programs in Biomedicine*, 221, 106924.
- [34] Bae, S. H., & Yoon, K. J. (2015). Polyp detection via imbalanced learning and discriminative feature learning. *IEEE transactions on medical imaging*, 34(11), 2379-2393.
- [35] Cao, X., Fan, K., & Ma, H. (2024). Federal learning-based a dual-branch deep learning model for colon polyp segmentation. *Multimedia Tools and Applications*, 1-22.
- [36] Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Authors' Profiles



Kumar Janardan Patra is a PhD Scholar in OUTR, India and completed his Master of Technology (M.Tech) degree in 2023 from the Institute of Management and Information Technology (IMIT), Cuttack, a Constituent College of BijuPatnaik University of Technology (BPUT), Odisha, India in Computer Science and Engineering (CSE). He has a deep passion for data analytics, data mining, machine learning, and AI, soft computing, visual computing, demonstrating strong interest and enthusiasm in these fields.



Jibitesh Mishra has interests in research, consultancy, working on new technologies, and projects. He has worked in association with various IT industries such as Infosys and Tech Mahindra. He has written many books and published papers in reputable journals. After completing his PhD in the year 2001, he has worked with State Govt of Odisha in various projects. His research interests are in fractal graphics and software engineering. He started introducing the concept of Web Engineering in India in 2006 after organizing the International Conference on Web Engineering & Applications (ICWA2006). In the social front, he coordinated the training program for SC & ST graduate engineers in the year 2008. He was on Sabbatical leave during 2008 to 2010 at King Khalid University, Abha, Saudi Arabia. Currently, he is a Professor in the School of Computer Sciences and Dean of Students' Welfare at OTR. His current research initiatives are in the area of Data Analytics, Immersive and haptic experience, Engineering the Mobile Apps and Ontology Engineering.



Soumya Ranjan Nayak now holds the position of Assistant Professor-II in the School of Computer Engineering at Kalinga Institute of Industrial Technology, KIIT Deemed to be University, located in Odisha, India. He obtained a Doctor of Philosophy (Ph.D) and Master of Technology (M.Tech) in Computer Science and Engineering under a scholarship provided by the Ministry of Human Resource Development (MHRD) of the Government of India. These degrees were earned at CET, BPUT Rourkela, India. Prior to this, he completed a Bachelor of Technology (B. Tech) and a Diploma in Computer Science and Engineering. He has authored over 180 articles that have been published in reputable international journals and conferences such as IEEE, Elsevier, Springer, World Scientific, IOS Press, Taylor & Francis, Hindawi, Inderscience, IGI Global, and others. These publications have undergone a rigorous peer-review process. In addition to the aforementioned accomplishments, the individual has authored 20 book chapters, published 6 books, and obtained 7 Indian patents (with 4 patents being granted). Furthermore, he has secured 4 International patents, all of which have been granted. The researcher's current areas of focus encompass bio-medical image analysis and classification, machine learning, deep learning, pattern recognition, fractal graphics, and computer vision. The author's writings have garnered over 4000 citations, with an h-index of 34 and an i10-index of 89, as reported by Google Scholar. Dr. Nayak holds the position of an associate editor for several esteemed academic journals, including the Journal of Electronic Imaging (SPIE), Mathematical Problems in Engineering (Wiley), Journal of Biomedical Imaging (Wiley), Applied Computational Intelligence and Soft Computing (Wiley), and PLOS One.

How to cite this paper:

Kumar Janardan Patra, Jibitesh Mishra, Soumya Ranjan Nayak, "Federated Transformer-Enhanced MobileNet V2 for Decentralized Polyp Detection in Colonoscopy Images", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.4, pp. 121-138, 2026. DOI:10.5815/ijigsp.2026.04.07