Computers and Electrical Engineering 118 (2024) 109414

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Explainable AI for gastrointestinal disease diagnosis in telesurgery Healthcare 4.0

Meet Patel [a], Keyaba Gohil [b], Aditya Gohil [b], Fenil Ramoliya [b], Rajesh Gupta [b], Sudeep Tanwar [b,*], Zdzislaw Polkowski [c], Fayez Alqahtani [d], Amr Tolba [e]

[a] Deutsche Bank, Pune, India
[b] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, 382481, India
[c] WSG University, Bydgoszcz, Poland
[d] Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia
[e] Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia

ARTICLE INFO

Keywords:
X-AI
Telesurgery
Gastrointestinal diseases
CNN
IG
LIME

ABSTRACT

The escalating prevalence of gastrointestinal disorders, spanning a spectrum from polyps to tumors, underscores the imperative for advanced diagnostic and interventional methodologies. This paper addresses this exigency by introducing telemedicine into the healthcare landscape, presenting an innovative approach for remote patient surgery. Telemedicine harnesses technological advancements to extend the reach of surgical expertise, transcending geographical limitations and augmenting patient-centric care strategies. Within the domain of healthcare 4.0, the integration of Explainable Artificial Intelligence (X-AI) is paramount in the telemedicine sector. This arises from the need to elucidate the decision-making processes of intricate models, particularly in the realm of gastrointestinal disease detection. We introduce *TeleXGI*, a sophisticated model leveraging Convolutional Neural Network (CNN) architectures ResNet50 and MobileNetV2 with X-AI techniques for accurate gastrointestinal disease classification using the Kvasir-V2 dataset. Advanced X-AI techniques such as Grad-CAM, saliency maps, integrated gradients, attribution heatmaps, and Local Interpretable Model-Agnostic Explanations (LIME) are employed to enhance interpretability. These methodologies unveil the inner workings of CNNs, highlighting impacted regions and augmenting predictive transparency. Our performance assessment validates *TeleXGI*'s excellent diagnostic accuracy at 98.8% with ResNet50 and 98.5% with MobileNetV2, faster inference with lightweight models, and precise explanation heatmaps essential to the real-time and high-throughput data-processing requirements of Telesurgery. The credibility of the explanation heatmaps is verified against the ground truth masks of Polyps images from the Kvasir-SEG dataset. This research establishes a benchmark for pivotal healthcare applications, showcasing the potential of X-AI in enhancing the interpretability of CNN models.

* Corresponding author.
E-mail addresses: meetpatel96301@gmail.com (M. Patel), 21bce076@nirmauni.ac.in (K. Gohil), 21bce011@nirmauni.ac.in (A. Gohil), 21bce244@nirmauni.ac.in (F. Ramoliya), rajesh.gupta@nirmauni.ac.in (R. Gupta), sudeep.tanwar@nirmauni.ac.in (S. Tanwar), zdzislaw.polkowski@byd.pl (Z. Polkowski), fhalqahtani@ksu.edu.sa (F. Alqahtani), atolba@ksu.edu.sa (A. Tolba).

https://doi.org/10.1016/j.compeleceng.2024.109414
Received 4 January 2024; Received in revised form 16 May 2024; Accepted 19 June 2024
Available online 8 July 2024

1. Introduction

Healthcare is one of the most important sectors in the world, as it directly impacts all other human activities. It is responsible for human health and well-being, ensuring that human life can be as long and ailment-free as possible. It focuses on treating diseases, illnesses, and disorders and doing relentless research and exploration to improve the quality of life, enhance curability, optimize treatments, and extend longevity. In recent years, healthcare has witnessed increased integration with technology and has transitioned into smart healthcare. Various technological tools and services have been incorporated into different facets of healthcare. Smart healthcare has undergone extensive integration with technology [1] and has gone from Healthcare 1.0 to Healthcare 4.0. Healthcare 1.0 includes the initial stages when medical professionals manually record patient data. With further development, Healthcare 2.0 constituted the digital storage of health records through Electronic Health Records (EHR) and digital imaging. This was followed by the use and integration of wearable devices, data analytics, and interlinked networks and devices. The latest stage in this journey is Healthcare 4.0, which incorporates a more direct approach to the use of technology, robotic tools, and services to achieve a much more robust contribution of technology in traditional healthcare compared to its predecessors [2].

Modern healthcare is witnessing the incorporation of various new revolutionary technologies. The development of imaging systems has enhanced diagnostic and interventional techniques, namely colonoscopy, endoscopy, angioplasty, ophthalmoscopy, and necroscopy. Robotic tools have facilitated the conduction of numerous high-risk surgeries and procedures with unprecedented ease and precision, especially in cardiology and neurology. Procedures like open-heart surgeries can now be performed by robots using small incisions, thus greatly reducing their invasiveness. Brain surgeries have also been automated by robotic techniques to better perform standard procedures with minimal human involvement and risk. 3-D printing techniques and mechanical advancement have resulted in far superior tools and equipment that ensure precision, accuracy, and efficiency.

One of the most potent applications of the aforementioned advancements can be found in gastroenterology. Clinically, the applications include the identification of malignancy and premalignancy in lesions, detecting lesions, and developing scoring systems about risk stratification, disease prognosis prediction, and treatment response. Additionally, predicting which treatments to administer to which patients based on previous data is also a budding new development, along with metrics like bowel preparation score and quality of endoscopic examination [3].

The aforementioned domain of gastrointestinal disorders and its integration with technological tools has witnessed a lot of research. Shin et al. [4] used a region-based Convolutional Neural Network (CNN) method for automatically detecting polyps in colonoscopy examination images. They proposed two post-learning methods for incorporation with a region-based detection system. Fiaidhi et al. [5] proposed an X-AI approach to explain the reasoning behind Deep-learning CNN (DCNN) in terms of particular ulcerative colitis scores about specific frames. They proposed methods like summarization and automatic caption generation. Sutton et al. [6] proposed a weakly supervised approach using deep learning algorithms to distinguish ulcerative colitis from other gastrointestinal disorders and grade its endoscopic severity. They used Grad-CAM as their X-AI model of choice. Chierici et al. [7] proposed a prototype based on Deep Learning (DL) for identifying disease patterns. They achieve this through three binary classification tasks. Ge et al. [8] put forward a five-category classification model based on the Los Angeles grade concept for gastroesophageal reflux disease based on DL and X-AI. Gangrade et al. [9] proposed a system to diagnose gastrointestinal disorders using five deep CNN techniques. Sadeghnezhad et al. [10] proposed an approach using transfer learning and the Inception-ResNet model to deal with the problems of random selection of weights and the need for rich vectors in vector-based approaches. In [11], Mukhtorov et al. propose an explainable DL method based on ResNet 152 in combination with grad-CAM for endoscopy image classification. Gunasekaran et al. [12] proposed an ensemble of pre-trained models for the accurate classification of endoscopic images affiliated with gastrointestinal (GI) diseases. Noor et al. [13] proposed a GI tract disease classification technique using an optimized brightness-controlled contrast-enhancement method to improve WCE image contrast. Ayidzoe et al. [14] proposed the use of visualizations to enhance performance verification, improve monitoring, ensure understandability, and improve interpretability using a CapsNet model. In [15], Alhajlah et al. propose a Mask Recurrent CNN and ResNet-based technique for feature extraction. Arthy and Prasanth proposed [16] a smart heart disease prediction system using IoT and adaptive deep CNN. Kavitha et al. [17] provided a systematic view and impact of AI in smart healthcare systems, principles, challenges and applications.

In the domain of healthcare, AI is considered a 'Black Box' due to the lack of knowledge about its inner workings. Since human lives are at stake, medical professionals need to understand the technicalities of different models. Addressing the aforementioned issue is crucial as the explainability of such frameworks is instrumental for their accountability in the medical domain. While [9,10,12,15,18] of the aforementioned research works do not address 'Black Box' problem and [11,14] overlooked the security and privacy concerns of patients' sensitive data. We propose *TeleXGI* using eXplainable Artificial Intelligence (X-AI) in integration with Healthcare 4.0. *TeleXGI* uses CNN architectures ResNet50 and MobileNetV2. To overcome the black-box aspect of the implementation, *TeleXGI* uses X-AI algorithms, namely saliency maps, Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients (IG), and Local Interpretable Model-agnostic Explanation (LIME). The objective is to design an Explainable Artificial Intelligence (X-AI) system that analyzes high-resolution gastrointestinal endoscopic images remotely, employing Deep Learning and X-AI techniques to predict potential health risks for patients after initial pre-processing and storing this data in a decentralized manner using blockchain technology [19]. The integration of blockchain is a novel addition to such an approach as it greatly enhances the security during storing such data and its transmission. We aim to improve the accuracy of such diagnoses and bolster transmission and storage security. Also, the inner workings are adequately highlighted with the use of X-AI.

Table 1
Symbol table.

Notation	Description	Notation	Description
$\mathcal{P}$	Patients	S	Gastrointestinal diseases
I_{acq}	Scans	τ	Communication protocol
Ω	CNN architectures	F	Feature vectors
Π	XAI techniques	ϱ	Multi-slice scanning
f	Parameter selection function	K	Kernel
ε	Edge computing	ϕ	Data encryption protocols
Θ	Output	d_i	Perturbed instance
$L(g)$	Weighted loss function	$\Omega(g)$	Regularization term
$\mathcal{N}$	Normalization	$\mathbf{W}_i$	Initial weights
W_i'	Normalized weights	$y_{\text{residual}}, y_{\text{depthwise}}, y_{\text{pointwise}}$	Output of convolution
σ	Sigmoid activation function	ZA_{output}	Weights for the output layer
(a_c^k)	Importance weights	(y^c)	Class score
Γ	Attribution score	w_i	Instance-specific weights

1.1. Research contributions

- *TeleXGI* integrates advanced CNN architectures (ResNet50, MobileNetV2) with Explainable AI (X-AI) techniques, enhancing model accuracy and interpretability for transparent remote surgical interventions in gastrointestinal disease detection.
- *TeleXGI* employs Grad-CAM, saliency maps, integrated gradients (IG), and LIME for model interpretability and explainability. Affected areas in the explanation heatmaps are verified against ground truth masks for the "Polyps" category using Intersection over Union (IoU), while cosine similarity is used for other disease categories. This approach ensures transparent and reliable CNN interpretations in telemedicine.
- *TeleXGI* utilizes a decentralized blockchain-based system for secure communication and storage of sensitive information between the remote site and the telesurgery central system, providing transparency and traceability to health records.

1.2. Organization of the paper

The subsequent sections of the paper are organized as follows: Section 2 introduces the system model and problem formulation of the proposed approach. Section 3 delves into a comprehensive explanation of the proposed scheme. The evaluation of the approach's performance is addressed in Section 4. Section 5 presents potential limitations and crucial factors to consider when employing the proposed *TeleXGI* approach. Finally, Section 6 presents the paper's conclusion and outlines potential avenues for future research. Table 1 provides an overview of the notations, variables and symbols used in the paper along with their corresponding descriptions.

2. System model and problem formulation

2.1. System model

To examine a patient ($\mathcal{P}$) located remotely from the medical officials, telesurgery centers ($\mathcal{T}$) are set up at various places. For this, every $\mathcal{P}$ is assigned a medical imaging device(E) that captures higher-resolution gastrointestinal endoscopic images(GI) of the patients. The crucial data captured is to be transmitted at the remotely located telesurgery center ($\mathcal{T}$) for analysis by using various communication protocols (τ) and data encryption protocols (ϕ). The images captured ($\mathcal{I}_{\text{acq}}$) also require some prepossessing for better results from the DL model Ω. This pre-processed data ($X_{\text{pre-processed}}$) sent to Ω results in the classification of this data into one of the eight classes(c) namely, dyed lifted polyps (c_1), dyed resection margins (c_2), esophagitis (c_3), normal cecum (c_4), normal pylorus (c_5), normal z-line (c_6), polyps (c_7), and ulcerative colitis (c_8). This classification is based on the class score y^c output, obtained from the weights (w_{i}) of the trained model Ω. The output of this layer is θ from Ω is sent to the next layer Π.

To draw more details from the result Θ, *TeleXGI* uses X-AI techniques like Integrated Gradients (IG), Local Interpretable Model-agnostic (LIME), Grad-CAM, and Saliency maps. The input Θ given to Π_i. Π_1 consists of Saliency maps, which back-propagate concerning the input image and generate images that highlight critical features relevant to a specific class prediction. Π_2 involves Grad-CAM highlighting class-specific influential regions within an image. This means that Grad-CAM uses the gradient of the classification score with convolutional features to determine which feature is the most important for classification. IG constitutes Π_3, wherein the influence of individual pixels on the overall class prediction of the model can be attributed. LIME is used at the conclusive stage Π_4. This clarifies the predictions of the neural network by generating locally faithful explanations. The net results of all the Pi_i constitute the combined effect of the X-AI techniques, $H_{\text{X-AI}}$. This $H_{\text{X-AI}}$ is passed on to the predictive analysis and attention layer, where the final prediction by the model is made. This sensitive patient data is securely transmitted between remote sites and telesurgery centers, leveraging the properties of blockchain. Using blockchain ensures that data is stored in a decentralized manner and is not tampered with.

2.2. Problem formulation

This section gives insight into the problem formulation process of the *TeleXGI* approach. *TeleXGI* is an efficient approach to classify various gastrointestinal conditions. It utilizes state-of-the-art CNN architectures ResNet50 and MobileNetV2 to achieve the benchmark result in classification accuracy. *TeleXGI* includes integrating various X-AI techniques, including LIME, Integrated Gradients, Grad-CAM and Saliency Maps for better interpretability of the results from the CNN outputs. The initial and crucial phase of this approach includes remote capturing of gastrointestinal (GI) endoscopic images of the patient ($\mathcal{P}$) by medical imaging devices (E) as shown in Eq. (1).

$$GI = E(\mathcal{P}) \tag{1}$$

The images (GI) are applied with various techniques γ_i to improve their quality to achieve a more detailed visualization of GI. These techniques include CLAHE (Contrast Limited Adaptive Histogram Equalization)(γ_1), gaussian blur (γ_2), median filtering (γ_3), unsharp masking (γ_4), high-pass filtering (γ_5), color balancing (γ_6), spatial alignment (γ_7). High-resolution images captured by (E) are sent to remotely located telesurgery center ($\mathcal{T}$) by using various communication protocols (τ) and data encryption protocols (ϕ) as shown in Eq. (2).

$$GI_{\text{transmitted}} = \tau(\phi(\gamma(\mathcal{P}))) \tag{2}$$

When $GI_{\text{transmitted}}$ are received at $\mathcal{T}$, they are first pre-processed to get results with higher accuracy when fed to the CNN model (Ω). The pre-processing process includes manipulating some features like angles, brightness, zoom, and shear. α denotes this pre-processing of the images. Ω is a DL model capable of performing multiclass classification of the given images. It classifies the input image into one of the classes $c_i \in \{c_1, c_2, \ldots, c_8\}$. We get the final output of this layer as Θ as shown in Eq. (3).

$$\theta = \Omega(\alpha(GI_{\text{transmitted}})) \tag{3}$$

The output Θ is passed to the X-AI model (Π) to get the explanation of the image classification done by Ω. Kvassir uses four X-AI techniques, i.e., Integrated Gradient (IG)(Π_{IG}), Local Interpretable Model-agnostic (LIME)(Π_{LIME}), Grad-CAM(Π_{GradCAM}) and Saliency maps (Π_{SM}). Each X-AI model takes two inputs, i.e., the CNN model (Ω) and Θ. Based on these two inputs, it generates a visual representation $\Theta_{\text{visualized}}$ that is easily interpretable by medical officials to explain the prediction of the model Ω as shown in Eq. (4)

$$\Theta_{\text{visualized}} = \Pi(\Theta, \Omega) \tag{4}$$

The combined information from all the X-AI techniques $H_{\text{X-AI}}$ is passed to the predictive analysis and attention layer, giving valuable insights based on these outputs. This information in images is also stored in the blockchain network to ensure its security and integrity. A smart contract is deployed on the blockchain to manage transactions occurring in the system. Image $\Theta_{\text{visualized}}$ and θ are first converted into a unique code using Interplanetary File System (IPFS), and this unique code is stored on the blockchain via the smart contract ($\mathcal{O}$) as shown in Eq. (5).

$$IPFS(\Theta_{\text{visualized}}, \theta) \xrightarrow{\mathcal{O}} Blockchain \tag{5}$$

3. The proposed scheme

Fig. 1 depicts the proposed *TeleXGI* approach. It consists of five primary layers: a data acquisition and transmission layer responsible for data collection, an AI layer for data processing, an *TeleXGI* layer for analysis, a predictive analysis and attention layer, and a blockchain layer. (few icons used in Fig. 1 were referenced from an online website [20].)

3.1. Data acquisition and transmission layer

Obtaining high-resolution gastrointestinal (GI) endoscopic images is paramount for accurate diagnosis and classification of GI diseases. This multifaceted process involves various key components and advanced techniques to ensure top-tier image quality and secure transmission for medical applications.

Central to this process is the specialized endoscopic equipment (E) designed for GI imaging. These endoscopes are equipped with cutting-edge technology, ensuring optimal visualization of anatomical landmarks and pathological findings. Advanced imaging techniques, such as contrast agents δ_{contrast} and image resolution enhancement ϵ, contribute to image clarity and detailed visualization of the GI tract. Real-time image enhancement techniques γ further enhance image quality, providing valuable diagnostic insights. Wireless telemetry (η) is seamlessly integrated into the endoscopic equipment for capturing gastrointestinal (GI) images. This feature allows wireless transmission of GI images to remote locations, enabling expert analysis and diagnosis by professionals regardless of their physical location. This innovation enhances telemedicine for GI diseases, ensuring timely assessments and facilitating collaborative, interdisciplinary approaches to diagnosis.

In conjunction with these established methods, recent advancements in gastrointestinal imaging have introduced techniques such as virtual chromoendoscopy $\phi_{\text{virtual-chromo}}$ for digital enhancement of mucosal and vascular patterns without traditional dye application. Additionally, confocal laser endomicroscopy (CLE) χ_{CLE} provides microscopic tissue imaging during endoscopy, offering

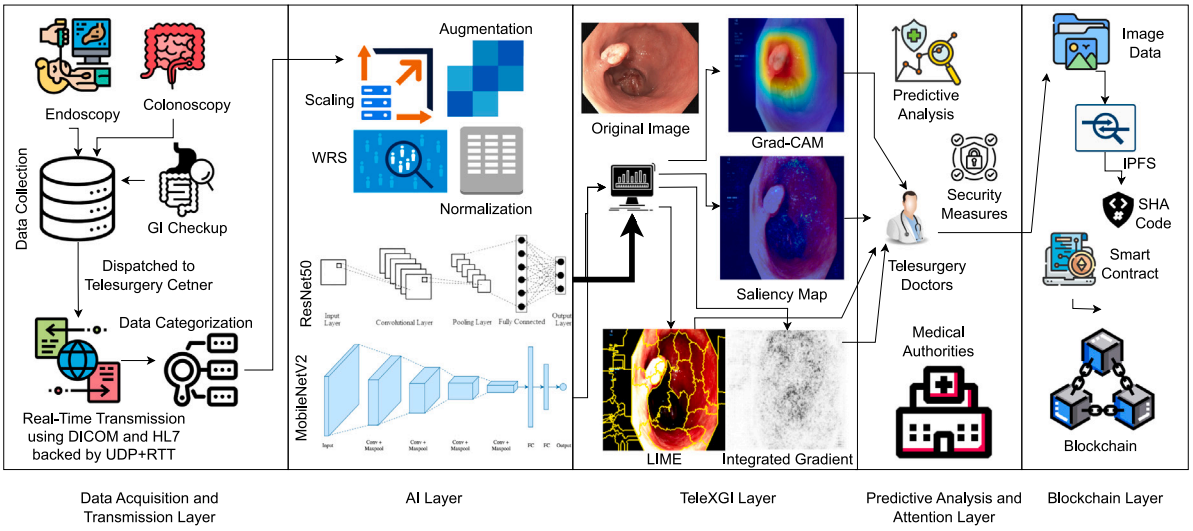


Fig. 1. *TeleXGI* system model.

real-time histological insights. Traditional endoscopy, denoted as $\phi_{\text{traditional}}$, remains a fundamental approach. These techniques augment the data acquisition equation, reflecting the evolving landscape of advanced endoscopic technologies.

$$I_{\text{acq}} = \int (\sum_{i=1}^{n} \eta(\gamma(\epsilon(\delta_{\text{contrast}}(\phi_{\text{virtual-chromo}} | \chi_{\text{CLE}} | \phi_{\text{traditional}}(I_i)))))) \tag{6}$$

Eq. (6) represents a fundamental mathematical representation that characterizes the data acquisition process. Following the initial data collection from endoscopic procedures, the next crucial step involves transmitting this data for further analysis.

Various secure protocols are employed to ensure the confidentiality of patient data during the transmission of gastrointestinal (GI) images. Widely used standards such as DICOM (Digital Imaging and Communications in Medicine) P_{DICOM} and Health Level Seven (HL7) P_{HL7} guarantee data integrity and patient privacy. Additionally, secure communication channels, including Virtual Private Networks (VPNs), are also established to create encrypted connections between medical facilities and remote experts, safeguarding data from unauthorized access. Secure file transfer protocols, such as Secure File Transfer Protocol (SFTP) P_{SFTP} and HTTPS P_{HTTPS}, are utilized for the safe transfer of GI images to remote servers, enhancing the privacy and security of data transmission. To optimize data transmission efficiency, a combination of User Datagram Protocol (UDP) and Real-Time Transport Protocol (RTT) $P_{\text{UDP+RTT}}$ can be implemented, offering reduced delay and latency for real-time applications while ensuring maximum security. These robust protocols and communication strategies contribute to the improvement of patient care and outcomes in the realm of GI disease classification.

$$I_{\text{transmitted}} = \text{VPN}(P_{\text{UDP+RTT}}(P_{\text{SFTP|HTTPS}}(P_{\text{DICOM|HL7}}(I_{\text{acq}})))) \tag{7}$$

Eq. (7) illustrates the storage, transfer, and reconstruction process of the obtained GI images, ensuring secure and efficient transmission for accurate diagnosis and classification.

3.2. AI layer

The AI layer comprises three primary components: dataset details, post-transmission data preprocessing, and CNN-based classification. These components collectively form an extensive workflow for data analysis and modeling.

3.2.1. Dataset description

For the Kvasir dataset, researchers focused on developing systems to enhance healthcare practices in gastrointestinal disease detection through computer-aided analysis of endoscopic videos. The dataset was collected at Vestre Viken Health Trust in Norway, with images obtained using endoscopic equipment from the gastroenterology department at Bærum Hospital. Carefully annotated by medical experts from Vestre Viken and the Cancer Registry of Norway, the Kvasir dataset includes images classified into three key anatomical landmarks and three clinically significant findings. Anatomical landmarks such as the Z-line, pylorus, and cecum are represented alongside pathological findings like esophagitis, polyps, and ulcerative colitis. Additionally, the dataset includes images related to removing lesions, including "dyed and lifted polyps" and "dyed resection margins". The dataset, meticulously sorted and annotated, serves as a valuable resource for both single- and multi-disease computer-aided detection research. It comprises high-resolution images, ranging from 720 × 576 to 1920 × 1072 pixels. Overall, the Kvasir dataset's comprehensive coverage and medical expert annotation make it instrumental for advancing research at the intersection of computer science and gastroenterology.

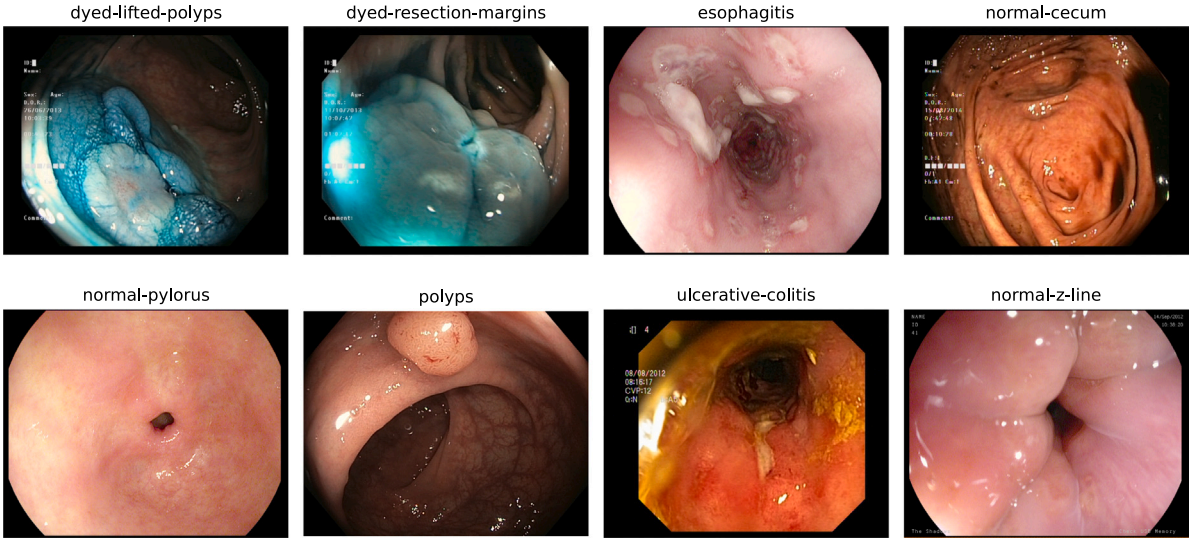


Fig. 2. Sample images from Kvasir-v2 dataset.

The Kvasir v2 dataset used in this analysis consists of 1000 images per class, totaling 8000 images across eight disease categories: dyed lifted polyps, dyed resection margins, esophagitis, normal cecum, normal pylorus, normal z-line, polyps, and ulcerative colitis. Fig. 2 portrays one sample image from each of the 8 categories present in the Kvasir-v2 dataset.

3.2.2. Data pre-processing

Our meticulous pre-processing pipeline involves a series of essential steps to optimize the Kvasir V2 dataset for subsequent deep-learning analysis. This encompasses tasks such as data normalization, resizing, augmentation, and the incorporation of weighted random sampling, collectively enhancing the dataset's suitability for the targeted gastrointestinal disease classification task with 8 classes.

The initial step in pre-processing involves resizing images to a specified input size ($\mathcal{W} = 224$ and $\mathcal{H} = 224$) to align with the model architecture. Our data augmentation strategy, implemented through PyTorch's transformations, aims to diversify the dataset. Specifically, we employ transformations such as random horizontal flip, random vertical flip, color jitter, random resized crop, Gaussian blur, random rotation up to 45 degrees, random perspective with a distortion scale of 0.5, and shear affine. The application of these transformations introduces variability in angle (θ), zoom ($\mathcal{Z}$), shear (S), translation ($\mathcal{T}$), brightness ($\mathcal{B}$), and contrast (C) parameters, enhancing dataset diversity and improving model robustness. The order of these transformations is randomized during training, contributing to the generation of varied and comprehensive training dataset. Normalization ($\mathcal{N}$) of images involves transforming each channel's mean and standard deviation to 0.0 and 1.0, respectively, contributing to noise reduction and improved feature extraction. The resulting pre-processed images are then ready for further analysis in the context of our gastrointestinal disease classification task.

The final pre-processing stage introduces Weighted Random Sampling (WRS) to address potential imbalances in the dataset. Biases introduced by imbalanced class distributions can impact training, particularly affecting classes with fewer instances. To mitigate this, Weighted Random Sampling is employed. Let $\mathcal{I}$ denote the set of images resulting from the normalization, augmentation, and resizing procedures. To counteract bias, we compute initial weights (ω_i) and normalized weights (ω_i') for each of the C classes ($c_1, c_2, \ldots, c_8$) with respective samples ($n_1, n_2, \ldots, n_8$) as expressed in Eq. (8). The normalized weights (ω_i') are then employed for sampling a balanced batch of images.

$$\omega_i = \frac{1}{n_i(\mathcal{I})}, \quad \omega_i' = \frac{\omega_i}{\sum_{j=1}^{C} \omega_j} \tag{8}$$

3.2.3. CNN models

The foundation of our proposed methodology, named *TeleXGI*, lies in the utilization of advanced CNN architectures for the specific task of gastrointestinal disease classification. This method amalgamates cutting-edge image analysis capabilities to provide a robust solution for accurate disease categorization.

The model selection process involves the integration of two prominent CNN architectures: ResNet50 (Ω_{ResNet}) and MobileNetV2 ($\Omega_{\text{MobileNet}}$). The dataset is divided into five folds, with each fold containing 80% of the data for training and 20% for testing. Additionally, 20% of the training data is allocated for validation purposes. To ensure dataset balance and evaluation fairness, the data split into train, validation, and test sets is done in such a way that each set contains a commensurate proportion of the images from each class. The training process incorporates the adaptive cosine learning rate decay method over 30 epochs as illustrated

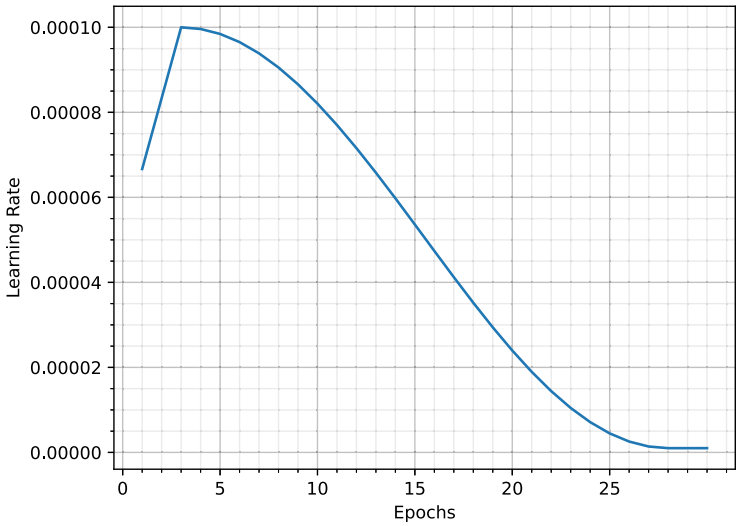


Fig. 3. Cosine annealing learning rate decay scheduler.

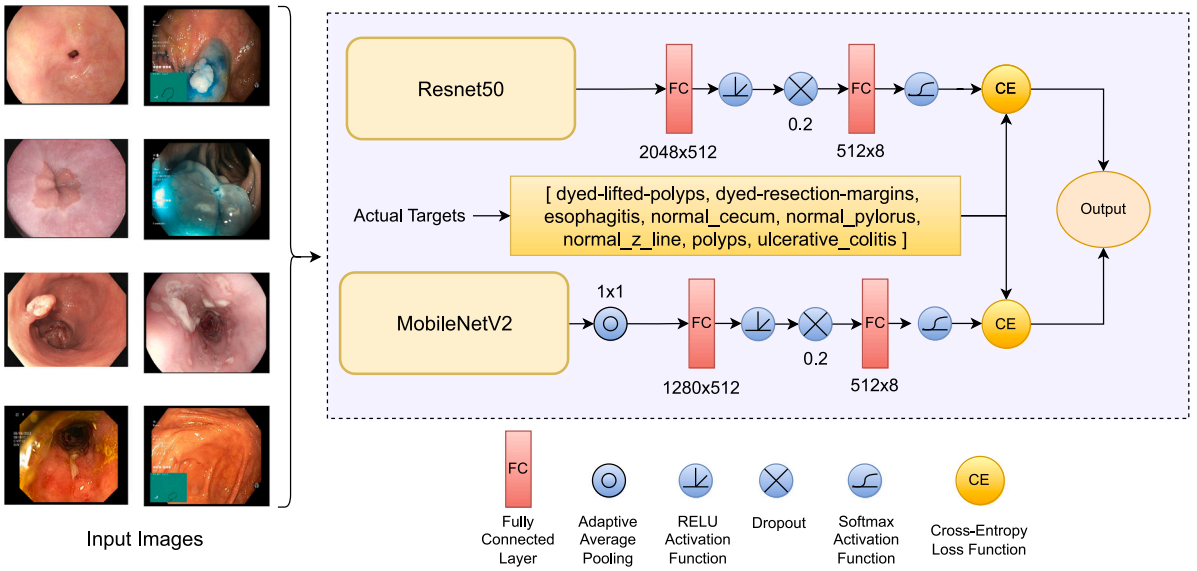


Fig. 4. *TeleXGI*'s model architecture.

in Fig. 3. Learning rate starts at $5 \cdot 10^{-5}$ and goes up to 10^{-4} during the first 3 warmup epochs. After that, it gradually decreases with a minimum learning rate of 10^{-6}. The chosen CNN architectures, ResNet50 and MobileNetV2, are selected based on their documented effectiveness in image classification tasks and their suitability for remote gastrointestinal disease quality assessment. These architectures excel in capturing both local and global image features. The models employ the Adam optimizer and incorporate Rectified Linear Unit (ReLU) activation functions at all intermediate levels. The softmax activation function is applied at the final layer to align with the specific requirements of the task. The illustration of CNN model is illustrated in Fig. 4.

Ω_{ResNet} integrates ResNet-50's deep residual learning architecture for effective feature extraction. The model combines the strengths of ResNet-50 for feature extraction and a Multilayer Perceptron (MLP) for accurate classification. The pre-trained ResNet-50 is initialized, and its final classification layer is replaced with a custom MLP featuring two linear layers with ReLU activation. This integration optimizes the model's capacity to discern intricate patterns in input data, enhancing its classification performance. The

Table 2
Hyperparameters used in training *TeleXGI*'s CNN models.

Parameter	Value	Parameter	Value
Batch size	64	Initial learning rate	$5 \cdot 10^{-5}$
Warmup epochs	3	Base learning rate	10^{-4}
Total epochs	30	Minimum learning rate	10^{-6}
Optimizer	Adam	Weight decay	10^{-5}
Hidden layer #Neurons	512	Classification head dropout rate	0.2

ultimate output Θ_{ResNet} from Ω_{ResNet} is computed as shown in Eq. (9).

$$\begin{aligned} y_{\text{residual}} &= \text{ReLU}\left(F(\mathbf{X}_{\text{preprocessed}}), \{Z_i\}\right) + \mathbf{X}_{\text{preprocessed}} \\ \Theta_{\text{ResNet}} &= \sigma(ZA_{\text{output}} \cdot y_{\text{residual}}) \end{aligned} \tag{9}$$

In this context, y_{residual} represents the output of the residual block, and F denotes a sequence of convolutional and activation layers with trainable weights represented as Z_i.

$\Omega_{\text{MobileNet}}$ distinguishes itself with an efficient and lightweight design, employing depthwise separable convolutions to reduce computational complexity while preserving representational capacity. This design involves a depthwise convolution followed by a pointwise convolution, capturing spatial and channel-wise information. Integrated with an MLP, $\Omega_{\text{MobileNet}}$ utilizes a pre-trained MobileNet V2 backbone for effective feature extraction. The MLP, with 1280 input neurons and 512 hidden layer neurons, adapts to MobileNet V2 output, ensuring precise classification through feature extraction, spatial reduction, and flattening in the forward pass. The ultimate output, $\Theta_{\text{MobileNet}}$, from $\Omega_{\text{MobileNet}}$ is calculated as illustrated in Eq. (10).

$$\begin{aligned} y_{\text{depthwise}} &= \text{DepthwiseConv}(\mathbf{X}_{\text{preprocessed}}, \{Z_i\}) \\ y_{\text{pointwise}} &= \text{PointwiseConv}(y_{\text{depthwise}}, \{Z_i\}) \\ \Theta_{\text{MobileNet}} &= \sigma(ZA_{\text{output}} \cdot y_{\text{pointwise}}) \end{aligned} \tag{10}$$

Here, $y_{\text{depthwise}}$ and $y_{\text{pointwise}}$ denote the outputs after the depthwise and pointwise convolutions, respectively. Furthermore, it is important to note that σ represents the sigmoid activation function, and ZA_{output} pertains to the weights for the output layer.

Both ResNet50 and MobileNetV2 are trained with consistent setting of hyperparameters. Table 2 summarizes the tuned values of hyperparameters used during training.

Both the models are trained on the kvasir-v2 dataset independently. During inference, the softmax logits from both models can be combined using the average function as part of ensemble learning. Additionally, one of the two models can be deployed based on the computing capabilities of the underlying device or the required speed. The ResNet50 model used in this study requires 24.5M trainable parameters, while the MobileNet V2 achieves a comparable classification performance with just 2.8M trainable parameters, providing a 10x reduction in model size. However, as explained in the results section, the explanation heatmaps of the ResNet50 are more precise and smaller in pointing out the problematic regions. Thus, a model choice would be a trade-off between the model size/speed and the explanation capabilities.

3.3. TeleXGI layer

This layer provides insights into the model's predictions, facilitating transmission to the telesurgery team through edge devices. To unravel the decision-making processes, our *TeleXGI* layer employs various X-AI techniques, including Saliency Maps, Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients, and Local Interpretable Model-agnostic Explanations (LIME).

3.3.1. Saliency maps

Saliency Maps, also known as the Vanilla Gradient method, represents a pioneering attribution technique. Unlike traditional methods that back-propagate through the network's layers, Saliency focuses on back-propagating directly concerning the input image, generating images highlighting critical features relevant to a specific class prediction. The saliency map value (Π_{SM}) is defined by Eq. (11).

$$\Pi_{SM} = \sum_{i=1}^{N} \frac{dP_c}{dI_i} \tag{11}$$

In this equation:

- N is the total number of pixels in the input image.
- $\frac{dP_c}{dI_i}$ represents the partial derivative of the predicted class probability P_c with respect to the ith pixel of the input image I_i.

The gradient $\frac{dP_c}{dI_i}$ quantifies how much the predicted probability P_c changes concerning a small change in the intensity of the ith pixel. By computing this gradient for each pixel, the saliency map provides a pixel-wise importance score, highlighting the regions of the input image that influence the network's prediction for a particular class.

The simplicity of the Saliency Maps makes them computationally efficient and easy to interpret. However, it is essential to note that they may not capture complex interactions and dependencies in the network, leading to the development of more sophisticated attribution methods. Nonetheless, Saliency Maps remains a valuable tool for gaining insights into model decision-making processes.

3.3.2. Grad-CAM

Grad-CAM, an acronym for Gradient-weighted Class Activation Mapping [21,22], stands out as an effective method for elucidating the influential regions within an image that contribute significantly to a model's classification decision. In contrast to more elementary approaches like Saliency Maps, Grad-CAM operates at a deeper convolutional layer, allowing for a nuanced understanding of the neural network's decision-making process.

The pivotal step in Grad-CAM involves the calculation of importance weights a_c^k for each neuron k in the selected convolutional layer, corresponding to a specific class c. These weights quantitatively express the contribution of individual neurons to the overall class score, computed as follows:

$$a_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \tag{12}$$

Here, Z is a normalization term ensuring that the importance weights collectively sum to 1, and $\frac{\partial y^c}{\partial A_{i,j}^k}$ represents the gradient of the predicted class score y^c with respect to the activation $A_{i,j}^k$ of neuron k.

The derived importance weights a_c^k serve as the foundation for constructing the Grad-CAM map Π_{GradCAM}. This map is formed by executing a weighted sum of the feature maps A^k in the chosen convolutional layer, defined as:

$$\Pi_{\text{GradCAM}} = \text{ReLU}\left(\sum_k a_c^k \cdot A^k \right) \tag{13}$$

The incorporation of the Rectified Linear Unit (ReLU) activation function emphasizes positive contributions, effectively highlighting regions of salience.

Grad-CAM, with its focus on deeper convolutional layers, provides a fine-grained visualization of class-specific discriminative regions. This depth-centric approach captures complex hierarchical features, offering superior localization compared to methods operating at the input layer. The interpretability provided by Grad-CAM enhances our understanding of the neural network's attentional focus during the classification process, making it a valuable tool in the analysis of deep neural networks.

3.3.3. Integrated gradients

Integrated Gradients (IG) [23,24] emerges as a profound method for attributing the influence of individual pixels on the model's class prediction. In contrast to other attribution techniques, IG introduces a baseline comparison, providing a rigorous approach to quantifying pixel contributions. The attribution score, denoted as Γ and computed through Eq. (14), delineates the importance of each pixel in shaping the model's decision.

$$\Gamma = \Pi_{\text{IG}(I)} = (I - I_{\text{baseline}}) \cdot \int_{\alpha=0}^{1} \nabla F(I + \alpha \cdot (I_{\text{baseline}} - I))\, d\alpha \tag{14}$$

Here, α represents the integration parameter, ∇ denotes the gradient, F represents the model's prediction function, I signifies the input image, and I_{baseline} represents the baseline image, often set as a black image or the image with minimal intensity. The integral term in the equation computes the cumulative gradients along the path from the baseline to the input image, providing a nuanced measure of each pixel's impact.

To elaborate further, the baseline image serves as a reference point to assess the change in model prediction as pixels are incrementally included from the baseline to the actual input image. The integral term encapsulates the cumulative effect of these incremental changes, effectively capturing the contribution of each pixel to the final prediction.

Mathematically, the process can be conceptualized as the integration of gradients along the path from the baseline to the input image. This integration, scaled by the difference between the input image and baseline, yields the attribution score Γ. The division by n represents the discretization of the integration path into n steps.

While the original definition of Integrated Gradients relies on an incalculable integral, its practical implementation approximates this integral with summation as shown in Eq. (15), ensuring its feasibility in real-world applications.

$$\Gamma = \Pi_{\text{IG}(I)} = \left(I - I_{\text{baseline}}\right) \cdot \frac{1}{n} \sum_{k=1}^{n} \nabla F\left(I_{\text{baseline}} + \frac{k}{n} \cdot \left(I - I_{\text{baseline}}\right)\right) \tag{15}$$

Integrated Gradients builds upon the foundation laid by Grad-CAM, combining gradient-based techniques with the notion of a baseline to provide a more comprehensive understanding of pixel-wise contributions. This method offers interpretability and introduces a level of rigor in attributing the model's predictions to specific pixels, making it a valuable addition to the repertoire of interpretability techniques.

3.3.4. Local interpretable model-agnostic explanations

In the culminating phase of our analysis, we employ the Local Interpretable Model-agnostic Explanations (LIME) methodology [25], a potent framework designed to elucidate the predictions of complex neural networks by generating locally faithful explanations. LIME strategically crafts a simplified and interpretable surrogate model g to approximate the nuanced behavior of the original model f, thereby offering transparency into the intricate decision-making processes within the model.

At the core of LIME's operation lies the perturbation of the input image, introducing controlled variations to generate a dataset of perturbed instances denoted as $\mathcal{D}$. Predictions from the original model f for these perturbed instances are then collected. LIME computes instance-specific weights w_i for each perturbed instance d_i based on their similarity to the original input I using a Gaussian kernel function K, as defined in Eq. (16).

$$w_i = K(I, d_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d(I, d_i)^2}{2\sigma^2}\right) \tag{16}$$

Here, $d(I, d_i)$ signifies the dissimilarity between the original input I and the perturbed instance d_i, while σ governs the extent of similarity. The Gaussian kernel elegantly captures instance similarity, assigning higher weights to closer instances due to the exponential decay.

Subsequently, the surrogate model g is trained using a weighted loss function $L(g)$, minimizing the dissimilarity between $f(I)$ and $g(I)$ as formulated in Eq. (17).

$$L(g) = \sum_i w_i \cdot [f(d_i) - g(d_i)]^2 + \lambda \cdot \Omega(g) \tag{17}$$

In this expression, $\Omega(g)$ acts as a regularization term promoting simplicity in the surrogate model, and λ controls the regularization strength. The optimization process, often employing techniques like Ridge regression, balances fidelity to the data with model simplicity.

The resultant surrogate model g serves as an interpretable proxy for the original model f, amenable to in-depth analysis for insights into image classification decisions. LIME's model-agnostic nature underscores its versatility, making it applicable across diverse neural network architectures.

3.4. Predictive analysis and attention layer

In the realm of Explainable Artificial Intelligence (X-AI), providing valuable insights concerning expected outcomes is a standard healthcare practice. These insights facilitate informed decision-making in patient care, adhering to stringent measures for data security and confidentiality.

The entire process of data transmission and decision-making is vigilantly monitored to uphold the highest standards of patient care and privacy protection. Integrating X-AI in healthcare enhances diagnostic precision while adhering to ethical and legal principles governing patient data management. This conscientious approach contributes significantly to improved healthcare practices by ensuring responsible handling of patient information.

3.5. Blockchain layer

The information received from the *TeleXGI* Layer is crucial and sensitive and is to be stored in a secured environment where its integrity can be maintained. It is also important to make the information tamper-proof. *TeleXGI* uses blockchain technology to neutralize the aforementioned issues. Inside a blockchain, the information is stored similarly to a ledger.

$$B_1, \ldots, B_i, \ldots B_n \in B \tag{18}$$

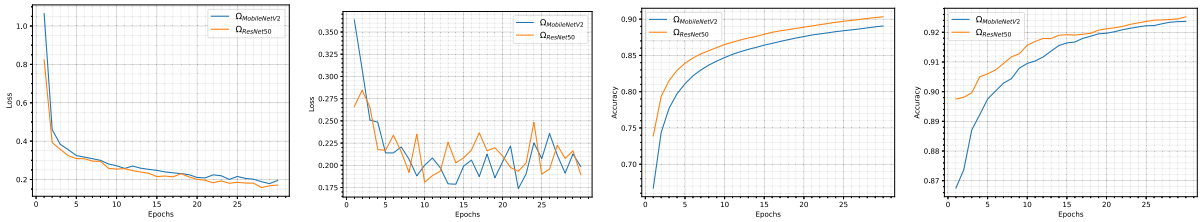
$$L_i = Hash(B_i) \tag{19}$$

Details of every transaction occurring on the blockchain are stored as a record. Furthermore, there are blocks B_i Eq. (18), that store these records, and after the block reaches its full capacity a new blockchain is generated. The blocks are connected using a link hash address. All the data in the form of blocks is shared with every user in the system. Any change done to the information changes the link hash address (L_i), Eq. (19), of the block, resulting in the changes in the whole link blocks and since the data is shared, this change is eventually detected, making the system secure. All the data transmission that occurs in the proposed approach is done securely using the blockchain network. This includes data being transmitted from the telesurgery devices to the medical officials and transmission of captured GI images to the telesurgery center. Using blockchain ensures that this sensitive data is not tampered with during transmission. Smart contracts are used as an interface between users and blockchain for easy and fast storage and retrieval of image data.

$$File \xrightarrow{IPFS} SHA\_code \tag{20}$$

The Interplanetary File System (IPFS) Eq. (20) is first used to get the unique SHA_code of each file. Further, the file in this unique code format is stored on the blockchain using the smart contract deployed Eq. (21).

$$SHA\_code \xrightarrow{SmartContract} Blockchain \tag{21}$$



(a) Training Loss (b) Validation Loss (c) Training Accuracy (d) Validation Accuracy

Fig. 5. (a) Loss curve for different models during the training, (b) Loss curve for different models during the validation, (c) Accuracy curve for different models during the training, (d) Accuracy curve for different models during the validation.

4. Performance evaluation

4.1. AI-based results

The assessment of model performance plays a pivotal role in selecting and refining deep learning models for effective validation of their capabilities. In classifying gastrointestinal images into eight categories, ResNet50 and MobileNetV2 were utilized as the deep learning models, both trained under identical computational settings on an NVIDIA GPU P100 [26]. During the training process, optimization techniques such as Adaptive Moment Estimation (Adam) and Sparse Categorical Cross-Entropy were applied.

The model architectures encompass various layers, including global average pooling, dropout layers, and batch normalization, aimed at enhancing training stability. Additionally, an output layer was integrated for multiclass classification. To leverage insights from the ImageNet dataset, the models were initialized with pre-trained weights. The presence of adjustable trainable layers facilitated effective fine-tuning, ensuring a consistent and unbiased training process for dependable results.

Fig. 5(a) illustrates the training loss, a metric gauging how well the model fits the training data. Lower training loss values signify superior model performance, with ResNet50 achieving the least training loss of 0.1713 by the end of the training period. Fig. 5(c) showcases the training accuracy of ResNet50 and MobileNetV2 across 30 epochs. The graph provides insights into the models' performance on the training data. As the epochs progress, both models initiate with low accuracy, indicative of initial random predictions. Notably, ResNet50 achieves the highest training accuracy of 0.9032, while MobileNetV2 exhibits a slightly lower accuracy of 0.8905 by the conclusion of the training period.

Fig. 5(b) illustrates the loss as a function of epochs for validation data, with MobileNetV2 and ResNet50 displaying the least loss of 0.1986 and 0.1897, respectively. Fig. 5(d) presents the validation accuracy of ResNet50 and MobileNetV2. Validation is crucial for evaluating the models' performance on unseen data. ResNet50 attains the highest validation accuracy of 0.9252, followed by MobileNetV2 with a validation accuracy of 0.9237.

To comprehensively assess model performance, key metrics such as precision (α), recall (β), and F1 score ($\mathcal{F}$) Eq. (22) is computed. These metrics offer insights into the models' predictive power, especially in scenarios featuring an uneven distribution of classes. Precision quantifies the accuracy of positive predictions, recall assesses the models' ability to include all relevant data points, and the F1 score provides a balanced evaluation, considering both false positives and false negatives.

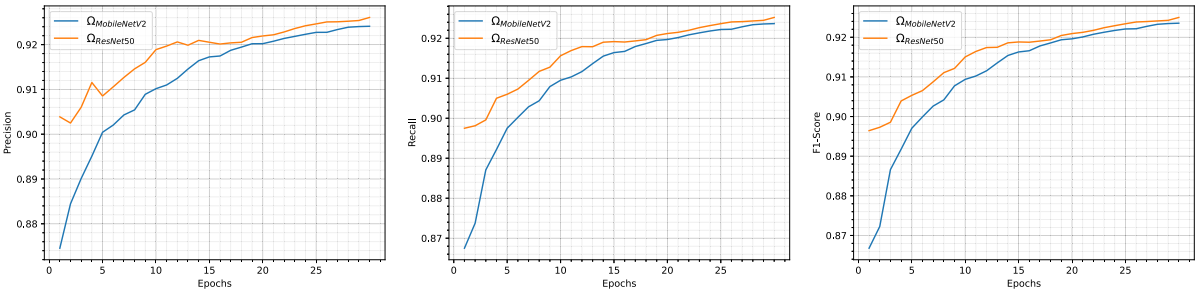
$$\beta = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \alpha = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathcal{F} = 2 \times \frac{\alpha \times \beta}{\alpha + \beta} \tag{22}$$

In the aforementioned equation TP, FN and FP denote true positives, false negatives and false positives, respectively.

Fig. 6 illustrates the comparison of precision, recall, and F1 score for ResNet50 and MobileNetV2, presenting a visual representation of how these metrics evolve throughout epochs and showcasing the models' improvement over time in the task of classifying gastrointestinal images into eight categories.

The harsh repercussions of missing a true positive case necessitate the evaluation of the models based on their ability to identify true positive instances out of all positive instances correctly. The sensitivity (recall) metric is used to verify this capability. Sensitivity measures the proportion of the true positive cases the model correctly identifies. On the test dataset, ResNet50 and MobileNetV2 trained as per the proposed *TeleXGI* architecture and tuned hyperparameter values achieved a remarkable 98.87% and 98.50% recall, respectively on the test dataset. Additionally, it is essential to consider other performance metrics such as precision, F1 score, and accuracy to understand the model's effectiveness comprehensively.

While accuracy is important, it should be considered alongside other metrics, especially in imbalanced datasets that might not adequately represent model performance. Precision represents the proportion of correctly predicted positive cases out of all instances predicted as positive. Both ResNet50 and MobileNetV2 exhibit high precision values, with ResNet50 at 98.93% and MobileNetV2 at 98.53%. This metric is valuable as it helps to assess the model's ability to avoid false positives. F1 score combines precision and recall into a single metric, providing a balanced assessment of a model's performance. Both ResNet50 and MobileNetV2 achieved high F1-score values at 98.88% and 98.51%, respectively, indicating a harmonious balance between precision and recall on the test dataset.



(a) Precision curve per epoch (b) Recall curve Per epoch (c) F1 score curve per epoch

Fig. 6. (a) Precision curve for different models during the validation stage, (b) Recall curve for different models during the validation stage, (c) F1 score curve for different models during the validation stage.

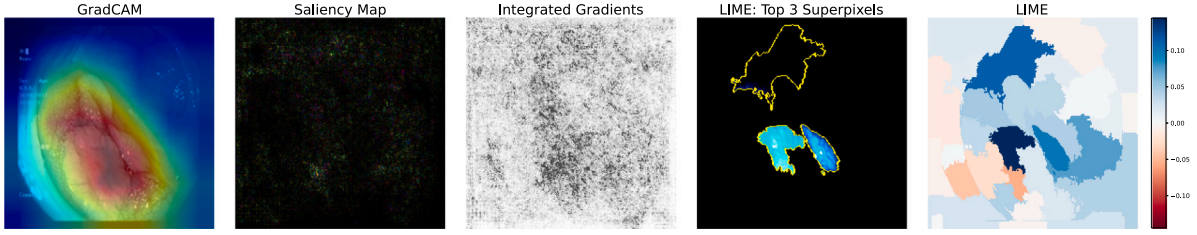


Fig. 7. ResNet50 explanation heatmaps on dyed-lifted-polyps Fig. 2.

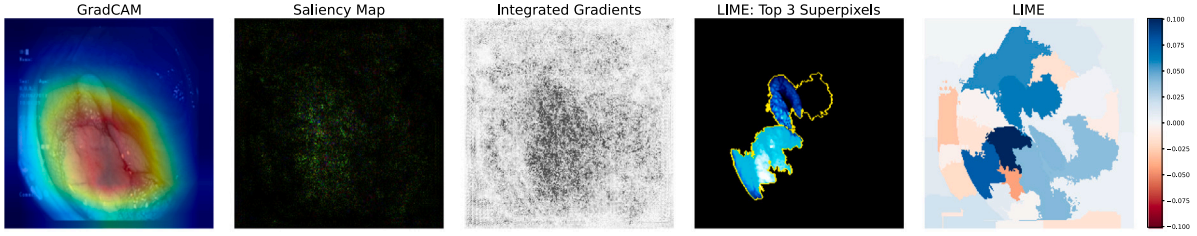


Fig. 8. MobileNetV2 explanation heatmaps on dyed-lifted-polyps image from Fig. 2.

4.2. X-AI based results

As per the classification confidence analysis presented above, it is clear that the ResNet50 model performs better than MobileNetV2, though not by a significant margin. This section explains the models' prediction based on the 4 XAI techniques employed in *TeleXGI*'s XAI layer. To obtain the GradCAM heatmaps, the last ReLU activation layer before the classification head is used from both models.

Figs. 7 and 8 presents the explanation heatmap obtained from each XAI technique with ResNet50 and MobileNetV2 respectively for the dyed-lifted-polyps class. Though the classification accuracy of MobileNetV2 is comparable to ResNet50, it is evident that its explanation heatmaps from each algorithm have a wider spread with many pixels highlighted to explain the predicted class. Comparatively, ResNet50 can locate the abnormal segment precisely. In the case of GradCAM, the reddish pixels have the most influence on a given prediction. For each of the other heatmaps, pixels with the higher value represent a positive influence on the predicted class. Note that LIME begins by dividing the image into small, non-overlapping regions called superpixels. During the training with perturbed instances (as explained in the XAI layer), LIME assigns importance weights to the superpixels based on the contribution of each superpixel to the predictions. Furthermore, these figures also depict top S most important superpixels selected by LIME, where S is configured to be 3.

To substantiate the explanation capability of the trained models, a comparison of the heatmaps with ground truth masks is required. For this purpose, the Kvasir SEG (segmentation) dataset is used. It contains the images pertaining to polyps category along with its ground truth segmentation masks manually annotated and verified by an experienced gastroenterologist. To obtain an explanation binary mask for each of the XAI techniques, the top k pixels with the highest explanation intensity are used. Here, k represents the number of pixels in the highlighted region of the ground truth segmentation mask.

Figs. 9 and 10 depicts the original polyps images, its ground truth, explanation binary mask and an explanation heatmap by the XAI technique with the highest IoU from both the models used. The intersection over union (IoU) metric measures the spatial

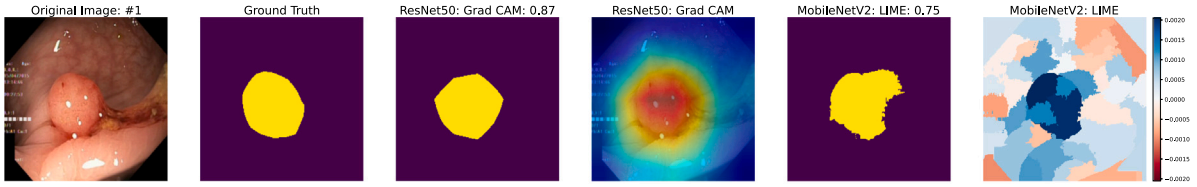


Fig. 9. Highest IoU among 4 XAI techniques for Polyps category — ResNet50: GradCAM, MobileNetV2: LIME.

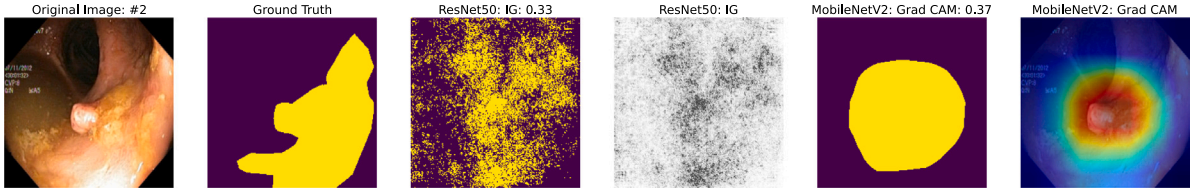


Fig. 10. Highest IoU among 4 XAI techniques for Polyps category — ResNet50: IG, MobileNetV2: GradCAM. Note: IG stands for Integrated Gradients.

Table 3
ResNet50: IoU between polyps ground truths and explanation binary masks. Image No. 1 and 2 refers to the Image #1 and Image #2 from Figs. 9 and 10.

Image No.	XAI technique	IoU	Image No.	XAI technique	IoU
1	GradCAM	**0.86**	2	GradCAM	0.32
1	Saliency map	0.08	2	Saliency map	0.23
1	Integrated gradients	0.16	2	Integrated gradients	**0.33**
1	LIME	0.56	2	LIME	0.31

Table 4
MobileNetV2: IoU between polyps ground truths and explanation binary masks.

Image No.	XAI technique	IoU	Image No.	XAI technique	IoU
1	GradCAM	0.68	2	GradCAM	**0.37**
1	Saliency map	0.16	2	Saliency map	0.27
1	Integrated gradients	0.21	2	Integrated gradients	0.32
1	LIME	**0.75**	2	LIME	0.21

overlap between the region highlighted by the explanation heatmap and the ground truth mask. The models were trained on the classification task. The IoU metric between the explanation's binary masks and ground truth segmentation mask is shown to validate the accuracy of the highlighted regions. Since the models were not primarily trained on the segmentation task to predict the mask, it is safe sometimes to ignore the low IoU. The primary focus of the explanation heatmap is to emphasize on the affected regions rather than exactly pinpoint the mask. From Figs. 9 and 10 it can be verified that either one of GradCAM, Integrated Gradients and LIME can perform most effectively to explain the region of focus behind the predicted class. This variation can be attributed to the size of the affected region present in the image or the type of model used for predicting the disease. For instance, the polyps image is shown in Fig. 10 contains only a small portion of the image where the polyps can be seen. However, the ground truth mask marks a big portion of the image as affected. In this case, Integrated gradients and GradCAM provide a higher IoU when using ResNet50 and MobileNetV2, respectively. While LIME, which takes only the top 3 superpixels in its segmentation mask, has understandably lower IoU when compared to the huge ground truth region. Moreover, this analysis bolsters *TeleXGI*'s proposed use of multiple XAI techniques to scrutinize the explanation of the predicted class. For instance, the polyps prediction for the image shown in Fig. 9 is best explained by GradCAM and LIME when the underlying CNN model used for prediction were ResNet50 and MobileNetV2, respectively. Similarly, the larger affected region is shown in Fig. 10, which is best explained by Integrated Gradients and GradCAM in the case of ResNet50 and MobileNetV2, respectively. In each scenario, the IoU scores between ground truth and the binary explanation mask extracted via all four XAI techniques are displayed in Tables 3 and 4 for ResNet50 and MobileNetV2, respectively.

The images and IoU analysis depicted in Figs. 9 and 10 shows some of the best and worst cases identified during experiments. To substantiate the accurate explanations by employing XAI techniques, Fig. 11 presents a comprehensive set of examples having different sizes for the affected regions. As shown, *TeleXGI*'s XAI techniques can precisely explain heatmaps across various polyps' sizes, shapes, and locations in the image.

For the classes other than Polyps, a dataset with ground truth segmentation masks annotated by expert medical professionals is not available. According to the above inspection with Polyps ground truths, it was observed that the GradCAM was the best-performing technique with a high IoU in the majority of cases. Thus, the explanation heatmap of GradCAM is utilized as a ground

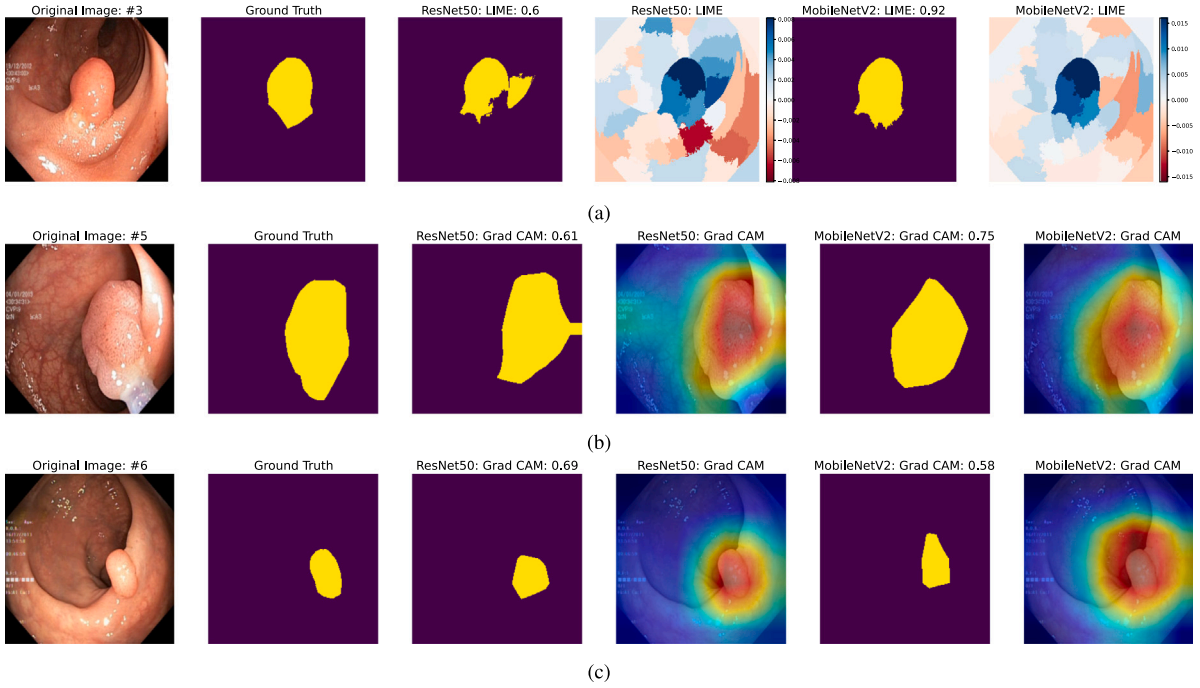


Fig. 11. Highest IoU among 4 XAI techniques for Polyps images with different shape, size, and location.

Table 5
Cosine similarity values between explanation heatmaps of GradCAM and other XAI techniques. SM: Saliency Map, IG: Integrated Gradients, CS: Cosine Similarity.

Class	XAI pair	CS:ResNet50	CS:MobileNetV2
dyed-lifted-polyps	GradCAM-SM	0.61	0.74
dyed-lifted-polyps	GradCAM-IG	0.72	0.77
dyed-lifted-polyps	GradCAM-LIME	0.83	0.82
dyed-resection-margins	GradCAM-SM	0.66	0.73
dyed-resection-margins	GradCAM-IG	0.68	0.77
dyed-resection-margins	GradCAM-LIME	0.79	0.85
esophagitis	GradCAM-SM	0.72	0.75
esophagitis	GradCAM-IG	0.73	0.78
esophagitis	GradCAM-LIME	0.85	0.87
normal-cecum	GradCAM-SM	0.63	0.66
normal-cecum	GradCAM-IG	0.64	0.69
normal-cecum	GradCAM-LIME	0.82	0.92
normal-pylorus	GradCAM-SM	0.70	0.75
normal-pylorus	GradCAM-IG	0.73	0.76
normal-pylorus	GradCAM-LIME	0.84	0.73
normal-z-line	GradCAM-SM	0.73	0.75
normal-z-line	GradCAM-IG	0.74	0.76
normal-z-line	GradCAM-LIME	0.72	0.75
polyps	GradCAM-SM	0.61	0.65
polyps	GradCAM-IG	0.69	0.69
polyps	GradCAM-LIME	0.86	0.80
ulcerative-colitis	GradCAM-SM	0.69	0.70
ulcerative-colitis	GradCAM-IG	0.70	0.76
ulcerative-colitis	GradCAM-LIME	0.80	0.63

truth for further analysis. Since the GradCAM does not produce a binary mask, a cosine similarity metric is used to quantify the relative performance of other XAI methods. The cosine similarity value considers the overall similarity in the spatial patterns, regardless of specific locations. Before calculating the cosine similarity, both explanation heatmaps — one from GradCAM and the other from LIME/Integrated Gradients/Saliency Map, are normalized using min–max normalization. It ensures that the values across the heatmaps are in a similar range. Table 5 shows each class's cosine similarity scores between explanation heatmaps. From this table, it is evident that the explanation heatmaps of X-AI techniques LIME, Integrated Gradients, and Saliency Maps possess high cosine similarity scores to the GradCAM heatmap, in that order.

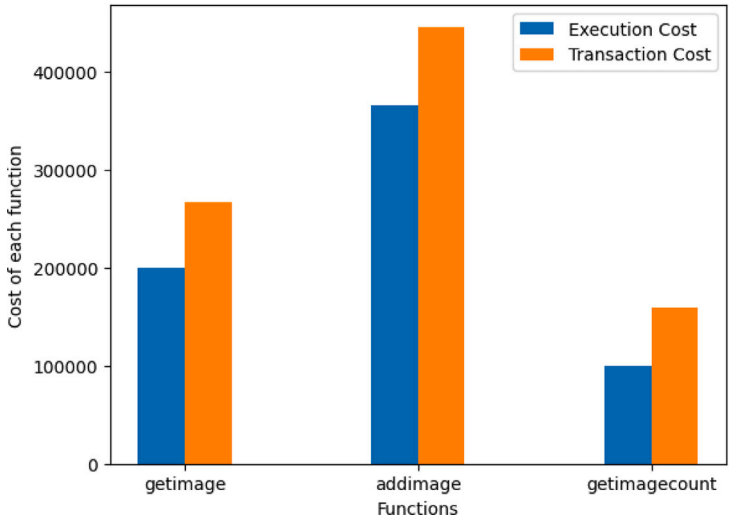


Fig. 12. Cost graph for the smart contract.

4.3. Blockchain-based results

The proposed approach uses blockchain to store the result from the X-AI Layer to maintain the security and integrity of the crucial data generated. A smart contract was built using Solidity language to demonstrate the effort made in the aforementioned direction. The smart contract was built using the Remix IDE with the compiler version 0.8.18+commit.87f61d96. The smart contract has the following features:

- SHA_codemap: A unique code for each image generated using the IPFS system.
- addimage: A function to add the SHA_code of the image into the blockchain.
- getimage: A function to retrieve the image using its SHA_code of the image.
- getimagecount: A function that returns the total value of images in the blockchain database.

The two costs that are involved in deploying a contract on any blockchain network are transaction costs and execution costs. Transaction costs are estimated based on traffic in terms of several other transactions going on the blockchain at the time our transaction is to be executed. Execution cost is estimated based on several computation operations involved in the transaction. Fig. 12 gives insight into the values of transaction costs and execution costs for each of the aforementioned functions of the smart contract.

4.4. Comparative analysis of TeleXGI with similar works

Table 6 presents a comparative analysis of the various recent works and our proposed approach *TeleXGI*. It highlights the differences based on a variety of factors, including the usage of data augmentation and sampling during training, the usage of X-AI techniques for better interpretability, and considerations for security using blockchain. This analysis effectively elaborates on the novelty of *TeleXGI*.

5. Limitations of *TeleXGI*

Telesurgery process requires fast real-time data transmission to ensure smooth functioning. It necessitates choosing an appropriate model based on the available hardware capacity and execution speed requirements. Such a decision required careful consideration due to the speed-explanation accuracy trade-off. Secondly, it may require the examination of explanation heatmaps from multiple XAI techniques for different sizes of target regions in the given image, as explained via IoU analysis with polyps ground truths. Running all of the proposed XAI techniques may not be feasible. Therefore, it is crucial to select a set of XAI techniques that are most suitable for the task at hand. Third, Blockchain stores data permanently, which can lead to increased storage requirements. Thus, a decision to employ blockchain as part of a secure transmission mechanism should be based on the long-term storage needs for medical data and alignment with the data retention policies of healthcare regulations. Finally, blockchain transactions can introduce latency and may limit throughput, especially in public blockchain networks. In a real-time telesurgery setting, low latency is crucial, requiring an intelligent design of a blockchain solution that minimally impacts data transmission speed.

Table 6
Comparative analysis of *TeleXGI* with existing works.

Literature					Results		
Ref.	Model used	Dataset variant	Data augmentation	Sampling during training	Accuracy	X-AI	Blockchain
[12]	DenseNet201 + InceptionV3 + ResNet201	8000 images - 8 classses (Kvasir-V2)	✓	X	0.95	X	X
[14]	CapsNet	8000 images - 8 classses (Kvasir-V2)	X	X	0.8057	✓	X
[9]	Effecient-NetV2B3	2500 images - 5 classes (A subset of Kvasir-V2)	✓	X	0.993	X	X
[11]	ResNet-152	8000 images - 8 classses (Kvasir V2)	✓	X	0.9828	✓	X
[15]	ResNet-50 + ResNet-152	4500 images - 8 classses (Kvasir-V2 + Hyper-Kvasir)	✓	X	0.9643	X	X
[10]	Inception Capsule with self-Attention	8000 images - 8 classses (Kvasir-V2)	X	X	0.9589	X	X
[18]	Clustered + Ensemble approach	8000 images - 8 classses (Kvasir-V2)	✓	X	0.9388	X	X
TeleXGI	MobileNetV2	8000 images - 8 classses (Kvasir-V2)	✓	✓	0.985	✓	✓
TeleXGI	ResNet50	8000 images - 8 classses (Kvasir-V2)	✓	✓	0.988	✓	✓

6. Conclusion and future scope

In conclusion, our proposed *TeleXGI* methodology, integrating advanced CNN architectures with X-AI techniques, stands as a pivotal advancement in telemedicine for gastrointestinal disease detection using the Kvasir-V2 dataset. We introduced a robust classification system leveraging ResNet50 and MobileNetV2, providing a dependable means to identify various gastrointestinal conditions and augmenting the diagnostic capabilities of telemedicine through the application of diverse X-AI techniques to enhance model interpretability. Blockchain is also used in the *TeleXGI* to build an environment that preserves the security and integrity of data stored within the system. Looking ahead, future endeavors in telemedicine for gastrointestinal disease detection encompass the exploration of multi-modal data fusion to enhance diagnostic accuracy by combining imaging with comprehensive health records. The implementation of edge computing mitigates latency in interventions by processing data in close proximity to its source. The advancement of X-AI techniques deepens model interpretability, fostering trust in AI-assisted healthcare. Investigating human–robot collaboration further refines precision and dexterity, potentially revolutionizing surgical outcomes. These avenues collectively propel us toward more resilient and trustworthy healthcare systems in the realm of gastrointestinal image classification.

Declaration of competing interest

There is no Conflict of Interest

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was funded by the Researchers Supporting Project Number (RSP2024R509), King Saud University, Riyadh, Saudi Arabia.

References

[1] Raparla K, Pandey N, Modh S. Indigenous and disruptive remote patient monitoring devices - A case study on AI in healthcare. SDMIMD J Manage 2023;14:27–34.
[2] Kumar A, Krishnamurthi R, Nayyar A, Sharma K, Grover V, Hossain E. A novel smart healthcare design, simulation, and implementation using healthcare 4.0 processes. IEEE Access 2020;8:118433–71.
[3] Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. World J Gastroenterol 2019;25(14):1666.
[4] Shin Y, Qadir HA, Aabakken L, Bergsland J, Balasingham I. Automatic colon polyp Detection Using Region based deep CNN and post learning approaches. IEEE Access 2018;6:40950–62.
[5] Fiaidhi J, Mohammed S, Zezos P. An xAI thick data assisted caption generation for labeling severity of ulcerative colitis video colonoscopy. In: 2022 IEEE 10th international conference on healthcare informatics. ICHI, 2022, p. 647–52.
[6] Sutton R, Zaiane O, Goebel R, Baumgart D. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. Sci Rep 2022;12:2748.
[7] Chierici M, Puica N, Pozzi M, Capistrano A, Donzella MD, Colangelo A, Osmani V, Jurman G. Automatically detecting crohn's disease and ulcerative colitis from endoscopic imaging. BMC Med Inform Decis Mak 2022;22(6):1–11.
[8] Ge Z, Wang B, Chang J, Yu Z, Zhou Z, Zhang J, Duan Z. Using deep learning and explainable artificial intelligence to assess the severity of gastroesophageal reflux disease according to the los angeles classification system. Scand J Gastroenterol 2023;58(6):596–604, PMID: 36625026.
[9] Gangrade S, Sharma PC, Sharma AK, Singh Y, Salehi AW. Computer-aided polyps classification from colonoscopy using deep learning models. 2023.
[10] Sadeghnezhad E, Salem S. InceptionCapsule: Inception-resnet and CapsuleNet with self-attention for medical image classification. 2024, arXiv preprint arXiv:2402.02274.
[11] Mukhtorov D, Rakhmonova M, Muksimova S, Cho Y-I. Endoscopic image classification based on explainable deep learning. Sensors 2023;23(6):3176.
[12] Gunasekaran H, Ramalakshmi K, Swaminathan DK, Mazzara M. GIT-Net: An ensemble deep learning-based GI tract classification of endoscopic images. Bioengineering 2023;10(7):809.
[13] Nouman Noor M, Nazir M, Khan SA, Song O-Y, Ashraf I. Efficient gastrointestinal disease classification using pretrained deep convolutional neural network. Electronics 2023;12(7):1557.
[14] Ayidzoe MA, Yongbin Y, Mensah PK, Cai J, Bawah FU. Visual interpretability of capsule network for medical image analysis. Turk J Electr Eng Comput Sci 2022;30(3):978–95.
[15] Alhajlah M, Noor MN, Nazir M, Mahmood A, Ashraf I, Karamat T. Gastrointestinal diseases classification using deep transfer learning and features optimization. Comput Mater Contin 2023;75:2227–45.
[16] Arthy S, Prasanth A. A smart heart disease prediction system using iot and adaptive deep convolution neural network. AIP Conf Proc 2024;2802(1):090005, arXiv:https://pubs.aip.org/aip/acp/article-pdf/doi/10.1063/5.0181854/18911646/090005_1_5.0181854.pdf.
[17] Kavitha M, Roobini S, Prasanth A, Sujaritha M. Systematic view and impact of artificial intelligence in smart healthcare systems, principles, challenges and applications. 2022, p. 25–56.
[18] Rashid RB, Alam MJ, Fattah SA. Proximity-linked multi-disease classification in endoscopy: Incorporating deep learning advancements in a cluster-based framework. In: 2023 26th international conference on computer and information technology. ICCIT, 2023, p. 1–5.
[19] Salah K, Rehman MHU, Nizamuddin N, Al-Fuqaha A. Blockchain for AI: Review and open research challenges. IEEE Access 2019;7:10127–49.
[20] Flaticon. 2023, https://www.flaticon.com/, Accessed 12 September 2023.
[21] Moujahid H, Cherradi B, Al-Sarem M, Bahatti L, Eljialy ABAMY, Alsaeedi A, Saeed F. Combining CNN and grad-cam for COVID-19 disease prediction and visual explanation. Intell Autom Soft Comput 2022;32(2).
[22] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision. ICCV, 2017, p. 618–26.
[23] Jha A, K. Aicher J, R. Gazzara M, Singh D, Barash Y. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. Genome Biol 2020;21:1–22.
[24] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017, arXiv:1703.01365.
[25] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier. 2016, arXiv:1602.04938.
[26] NVIDIA Corporation. NVIDIA tesla P100 data sheet. 2016, URL https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-p100/pdf/nvidia-tesla-p100-datasheet.pdf, Accessed 22 April 2024.