Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(7), 189-220
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

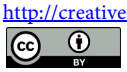
Multiclass Lesion Classification in Wireless Capsule Endoscopy: An Interpretable Deep Learning Framework

Moteb Alghamdi[1], Mohamed Elsersy[2], Salah Abdel-Mageid[1]

[1]Department of Computer Engineering, College of Computer Science and Engineering, Taibah University, Madinah, KSA
[2]Computer Information Systems Department, Higher Colleges of Technology, Al Ain, UAE
Email: mgamdi@taibahu.edu.sa, m.elsersy@hct.ac.ae, sabdelmgeid@taibahu.edu.sa

How to cite this paper: Alghamdi, M., Elsersy, M. and Abdel-Mageid, S. (2026) Multiclass Lesion Classification in Wireless Capsule Endoscopy: An Interpretable Deep Learning Framework. *Journal of Computer and Communications*, **14**, 189-220.
https://doi.org/10.4236/jcc.2026.147010

Received: June 9, 2026
Accepted: July 25, 2026
Published: July 28, 2026



Abstract

Wireless Capsule Endoscopy (WCE) has become an important tool in gastrointestinal diagnostics, yet the manual review of extensive video data remains labor-intensive and subjective. While deep learning has shown promise for automating this task, existing approaches are largely limited to binary bleeding classification and lack robustness, interpretability, and multiclass lesion analysis. This paper introduces a fully automated, CNN-based framework for robust multiclass lesion classification in WCE imagery, addressing bleeding, ulcers, and arteriovenous malformations (AVMs). Leveraging the KAUHC dataset, a novel repository of 3301 annotated small-bowel endoscopic images from Saudi Arabia comprising Normal (2156), AVM (673), and Ulcer (472) frames, our method employs a fine-tuned VGG16 architecture with advanced data augmentation and preprocessing to handle class imbalance, illumination variance, and anatomical complexity. The system achieves strong diagnostic performance, with a precision of 0.97, recall of 0.97, and an F1-score of 0.97, outperforming baseline models such as InceptionV3. Furthermore, we integrate Explainable AI (XAI) techniques (including SHAP and LIME) to provide interpretable decision support and enhance model transparency for clinical interpretation. By bridging the gap between experimental AI and real-world usability, this work offers a reliable, scalable, and interpretable tool with the potential to support computer-assisted gastrointestinal diagnostics and accelerate AI-driven automation in clinical workflows.

Keywords

Convolutional Neural Networks (CNNs), Wireless Capsule Endoscopy, Image Classification, Deep Learning, Medical Imaging

1. Introduction

Wireless Capsule Endoscopy (WCE) has revolutionized gastrointestinal diagnostics by enabling noninvasive, full-length visualization of the digestive tract through a swallowable camera [1]-[3]. This technology is indispensable for detecting critical conditions such as Crohn's disease, ulcers, polyps, and gastrointestinal bleeding [4] [5]. However, a major bottleneck arises post-procedure: each WCE examination generates tens of thousands of images, necessitating exhaustive manual review by clinicians, a process that is labor-intensive, time-consuming, and prone to diagnostic variability, especially for subtle or intermittent bleeding regions that are easily overlooked [6]-[8].

The clinical urgency for efficient and accurate bleeding detection is underscored by its substantial global and regional health burden. Gastrointestinal diseases, where bleeding is a frequent and serious complication, affect millions worldwide, leading to high rates of morbidity, mortality, and healthcare expenditure. According to the World Health Organization (WHO), digestive diseases constitute a significant portion of the global disease burden, with acute bleeding contributing notably to emergency hospitalizations and mortality [9]. This challenge is particularly acute in regions such as the Kingdom of Saudi Arabia (KSA), where gastrointestinal disorders rank among the leading causes of hospital admissions. Recent reports from the Saudi Ministry of Health indicate a rising prevalence of conditions like peptic ulcer disease and inflammatory bowel disease—common sources of gastrointestinal bleeding—thereby escalating the demand for reliable and efficient diagnostic tools [10]. The difficulty in localizing obscure bleeding sources further amplifies the need for technological solutions that can enhance early detection, expedite clinical intervention, and mitigate associated clinical and socioeconomic impacts both globally and within the Kingdom.

Beyond WCE-specific research, a wide body of work has investigated AI-based techniques for medical imaging across multiple modalities, including intracranial hemorrhage (ICH), diabetic retinopathy (DR), and stroke detection. These studies provide useful methodological insights and highlight the diversity of DL strategies employed in clinical diagnostics. Recent advances demonstrate several pivotal trends, including the architectural shift from standard CNNs toward hybrid and transformer-based models that better capture spatial dependencies [11] [12]; the adoption of data-centric innovations such as augmentation and synthetic data generation to address class imbalance and scarcity; the integration of interpretability tools like Grad-CAM and SHAP to enhance clinical transparency [13]; and the development of lightweight, real-time frameworks suitable for edge deployment [14]-[16]. In neuroimaging, genetic algorithms combined with BiLSTM networks have been applied for early stroke detection [17], while CNN-ResNet fusion has been used for hemorrhage and infarct classification [18]. Similarly, in ophthalmology, hybrid DL models with optimization techniques like APSO have been developed for DR severity classification [19], and EfficientNet-based frameworks have been proposed for early DR detection [20]. Despite these advances in adja-

cent fields, their systematic application to WCE (particularly for multi-class lesion detection and explainable diagnostics) remains limited. This disconnect highlights an opportunity to adapt proven methodologies from broader medical AI research to address the unique challenges of gastrointestinal image analysis.

To advance automated WCE analysis, the availability of high-quality, annotated datasets is essential. Recently, the authors in [21] introduced the King Abdulaziz University Hospital Capsule (KAUHC) dataset, a novel annotated repository of small-bowel endoscopic images from Saudi Arabia, comprising labeled frames across three categories: Normal, Arteriovenous Malformations (AVM), and Ulcer.

This paper extends beyond these baselines by employing deep learning architectures, systematic data augmentation, and integrated explainable AI. Therefore, this work is motivated by specific research gaps in current WCE analysis. Most existing solutions are confined to binary bleeding classification and lack validation across diverse lesion types commonly encountered in practice, such as ulcers and arteriovenous malformations (AVMs). Few models are designed to handle the inherent variability in WCE imaging conditions, including class imbalance, lighting inconsistencies, and anatomical complexity. Moreover, the "black-box" nature of deep learning models hinders clinical trust and adoption, with the systematic integration of Explainable AI (XAI) into WCE analysis remaining rare. While AI methods have matured in other medical imaging domains, their tailored application to WCE (especially for multi-class, explainable lesion detection) is still underdeveloped.

To address these gaps, we introduce a robust, interpretable, and clinically adaptable CNN-based framework for automated multi-class lesion detection in WCE, with a primary focus on bleeding, ulcers, and AVMs. Our contributions are threefold:

1) Development of an Interpretable Deep Learning Framework: A fine-tuned CNN-based architecture is designed and evaluated for multi-class WCE image analysis, achieving strong diagnostic performance (e.g., VGG16 F1-score: 0.97), supplemented with XAI modules for explainable predictions.

2) Data-Centric Optimization for Improved Robustness A comprehensive preprocessing and augmentation pipeline is implemented to address class imbalance, imaging artifacts, and clinical variability, thereby enhancing model generalizability across diverse datasets.

3) End-to-End System-Oriented Evaluation: A practical, deployable system design is emphasized (focusing on stability, scalability, and seamless integration into clinical screening workflows) to bridge the gap between experimental validation and real-world application.

The remainder of this paper is organized as follows. Section 2 reviews related work and identifies persistent research gaps. Section 3 details the proposed methodology, including dataset description, preprocessing pipeline, architectural design, and training protocol. Section 4 presents experimental results and analysis, encompassing model performance evaluation, comparative architecture assessment, and explainability analysis. Finally, Section 5 concludes the paper.

2. Related Work

Recent advances in artificial intelligence for medical imaging have demonstrated significant potential across a wide spectrum of clinical applications. While these developments have primarily focused on modalities such as computed tomography (CT) for intracranial hemorrhage (ICH), retinal fundus imaging for diabetic retinopathy (DR), and magnetic resonance imaging (MRI) for stroke, they offer valuable methodological insights that remain largely untapped in Wireless Capsule Endoscopy (WCE) analysis. This section critically examines these cross-domain innovations, evaluates their limitations, and identifies transferable strategies that inform the development of a more robust, interpretable, and clinically applicable WCE diagnostic system.

In the domain of neuroimaging, substantial progress has been made in automating the detection of ICH and stroke through sophisticated hybrid architectures. For instance, the authors in [17] combined genetic algorithms with Bidirectional Long Short-Term Memory (BiLSTM) networks for early stroke detection, achieving high accuracy but lacking interpretability features essential for clinical adoption. Similarly, the work in [18] employed CNN-ResNet fusion for hemorrhage and infarct classification, demonstrating strong binary classification performance but neglecting multiclass diagnostic challenges. Further innovations include ML models using ResNet50 for hemorrhage classification [22], which achieved modest accuracy but lacked integration with explainable AI, and DL approaches using basic CNNs for ICH detection [23], which demonstrated high sensitivity but offered limited architectural novelty and multiclass extension. These studies reveal a common pattern in neuroimaging research: while technical performance metrics are often impressive, there remains insufficient attention to model transparency, generalizability across heterogeneous data, and integration into real clinical workflows.

Further innovations in ICH detection include the use of windowed CT slices with architectures such as InceptionV3, EfficientNet, and ResNext101 [24], though with moderate accuracy and limited explainability. Hybrid models have also gained traction, such as the combination of improved AlexNet with Inception-v4 via transfer learning [25], and the integration of CNNs with metaheuristic optimization algorithms like Rider Optimization (ROA) for hyperparameter tuning [26]. These approaches show gains in accuracy but often at the cost of increased complexity and limited clinical interpretability. More advanced solutions include few-shot learning with U-Net architectures [27] and two-stage approaches with YOLO and 3D-CNN for cerebral microbleed detection [28], yet these still struggle with integration into real-time clinical workflows.

The field has witnessed a notable architectural evolution beyond standard CNNs. Transformer-based models, as introduced for ICH detection in [11], leverage self-attention mechanisms to capture complex spatial dependencies in CT images, reporting improved accuracy over conventional CNNs. Parallel developments include lightweight hybrid frameworks such as SINO-CT-FusionNet [14],

designed for efficient real-time detection, and 3D CNN approaches [29] that incorporate volumetric context for more reliable cross-dataset performance. Empirical analyses using 3D CT images for ICH detection have further underscored the importance of volumetric data and slice-sequence modeling [29], though these approaches are not directly applicable to 2D WCE frames. Multi-label classification of hemorrhage subtypes has also been explored using CNN ResNet architectures [15], while temporal modeling has been incorporated via CNN and BiLSTM fusion for non-contrast CT volumes [30]. Research also extends to automated brain hemorrhage detection through end-to-end DL approaches [31], which demonstrate automation capabilities but lack interpretability and clinical integration details. Despite their demonstrated effectiveness in CT and MRI domains, these advanced architectures have not been systematically adapted to WCE's unique challenges, including variable illumination, motion artifacts, and subtle mucosal patterns.

In ophthalmology, automated DR screening has driven innovations in both efficiency and comprehensiveness. Research ranges from hybrid methods combining CNN-based feature extraction with classifiers like SVM [32] to more streamlined, efficient architectures. An EfficientNet-based framework was developed in [20] that achieves high sensitivity and specificity with reduced computational demands, making it suitable for resource-constrained screening environments. More comprehensive diagnostic approaches, such as the APSO-GResNet hybrid model proposed in [19], combine features from multiple pretrained networks optimized through Adaptive Particle Swarm Optimization to detect a wider spectrum of DR lesions. Other works include feature extraction and classification using CNNs with traditional classifiers like NB, SVM, and KNN for microaneurysm detection [33], and the application of deep medical learning models for ovarian cancer detection [34]. Additional contributions include surveys reviewing DR detection using fundus images [35], which provide comparative insights but offer no technical innovations, and CNN-based approaches for early detection and staging of diabetic retinopathy [36], though these often neglect explainability and deployment constraints. These methodologies highlight the importance of balancing accuracy with computational efficiency and expanding detection scope beyond binary classification—objectives equally critical yet underexplored in WCE research.

Explainable AI (XAI) has emerged as a crucial component in building clinical trust, particularly in high-stakes diagnostic applications. Grad-CAM visualizations were integrated into a multi-label DR classification model in [13] to provide transparent decision support. Further advancements include two-stage DL approaches with Grad-CAM for DR [13] and the use of residual attention networks for stroke detection [37]. Web-based DL frameworks have also been proposed for precise brain hemorrhage detection from CT scans [38], demonstrating practical deployment potential but limited validation in clinical workflows. However, despite these advances, XAI remains conspicuously absent from WCE literature. To our knowledge, no prior study has systematically incorporated explainability tech-

niques such as SHAP or LIME into a multiclass WCE lesion detection framework, representing a significant barrier to clinical translation.

Within WCE-specific research, existing approaches demonstrate notable limitations. In [39], AlexNet was fine-tuned for binary bleeding detection but did not address ulcers or arteriovenous malformations (AVMs), nor were explainability mechanisms incorporated. Similarly, a CNN-Vision Transformer fusion (Conv-ViT) was proposed in [12] for gastrointestinal abnormality detection but with limited validation on multiclass datasets. Other notable contributions include focal modulation-driven CNNs for GI anomaly detection [40], though with relatively low performance metrics, and studies on DL with transfer learning using VGG16 and windowing techniques for ICH detection [41]. Beyond GI applications, research extends to gynecological imaging with innovative machine learning approaches for fibroid detection in uterine ultrasound [42], demonstrating the adaptability of AI across organ systems but remaining disconnected from WCE challenges. These works remain confined to a narrow diagnostic scope, lack robustness to WCE's inherent variabilities, and fail to leverage methodological innovations from adjacent medical imaging domains.

Beyond these specific applications, broader methodological surveys and reviews provide valuable context for AI in medical imaging. Comprehensive surveys on AI-driven biomedical image processing for hemorrhage detection [43] and AI applications in stroke imaging [44] offer useful comparative insights into methodological trends and clinical deployment considerations, though they propose no novel technical models. Research also extends to dual detection approaches for brain tumors and ICH [16], automated brain hemorrhage detection in smart IoT environments [45], and DL with optimization techniques for ICH detection [46]. While these studies provide valuable insights into architectural hybridization, efficiency optimization, and deployment strategies, their direct application to WCE remains limited, and they often lack the depth required for clinical translation in gastrointestinal diagnostics.

Most recently, the authors in [21] introduced the King Abdulaziz University Hospital Capsule (KAUHC) dataset (annotated WCE datasets from the Middle East region). The dataset comprises 3301 labeled frames across three categories (Normal: 2156; AVM: 673; Ulcer: 472) curated from 86 WCE studies conducted at King Abdulaziz University Hospital between December 2019 and December 2023. Images were acquired using the OMOM capsule system at 2 frames per second, with each frame at 512 × 512-pixel resolution. The authors validated annotation quality through multiple methods: 1) three medical experts with consensus-based labeling (agreement by ≥2 experts), 2) Cohen's Kappa inter-rater agreement (κ = 0.81 - 0.83, indicating near-perfect agreement), and 3) baseline machine learning classifiers (DT, RF, KNN, LR, NB, SVM). While binary classification achieved accuracies up to 100%, multiclass performance showed greater variability (57% - 98% accuracy across classifiers) [21]. Crucially, the study does not explore deep learning architectures, lacks advanced data augmentation strategies

to handle class imbalance, and offers no integration of explainable AI (XAI) techniques (all critical components for developing clinically deployable diagnostic systems). Furthermore, validation primarily focuses on binary classification scenarios, with multiclass analysis receiving limited attention despite the dataset's three-class structure.

The limitations observed across these studies reveal several persistent gaps in automated WCE analysis. First, there is a disconnect between methodological advances in broader medical AI and their application to WCE challenges. Second, existing WCE approaches typically focus on binary detection rather than clinically relevant multiclass classification. Third, robustness to imaging variabilities is often addressed through isolated techniques rather than integrated pipelines.

Table 1. Critical summary of Related AI studies in medical imaging.

Ref.	Approach	AI Methods Used	Modality/Target	Key Results	Limitation
[17]	Genetic algorithm + BiLSTM for early stroke detection	GA, BiLSTM	Stroke/ Neuroimaging	Acc: 96.45%, F1: 96%.	Limited explainability; not validated on volumetric or temporal medical data.
[18]	Automated detection using DL	CNN-ResNet	Brain hemorrhage and infarction	Acc: 95%, F1: 96%	Binary focus; lacks multiclass lesion differentiation.
[22]	ML model for hemorrhage classification	ResNet50, ML	ICH/CT	Acc: 91.46%, F1: 94%.	Modest accuracy; no integration with explainable AI.
[23]	DL for ICH detection	CNN	ICH/CT	Sensitivity: 98%, Specificity: 95%	Lacks architectural novelty; no multiclass extension.
[24]	CNN-based binary classification using windowed CT slices	InceptionV3, EfficientNet, etc.	ICH/CT	Acc: 84.03%	Moderate accuracy; no integration of explainability methods.
[25]	Combined AlexNet and Inception-v4	AlexNet, Inception-v4	ICH/CT	Acc: 94.54%, F1: 0.938	No discussion of generalizability to other modalities.
[26]	DL with Rider Optimization	ICHDC-RODL, XCS-LBP, BiLSTM	ICH/CT	Acc: 98.34%, F1: 98.56%	Complex optimization; limited clinical interpretability.
[27]	Few-shot learning with U-Net architecture	Adversarial learning, Multi-scale CNN	ICH/CT	Not specified	Experimental; lacking large-scale validation.
[28]	Two-stage DL with YOLO + 3D-CNN	YOLOv2, 3D-CNN	Cerebral microbleeds/MRI	Sensitivity: 94.32%	Low precision (61.94%); not real time.
[11]	Transformer-based ICH detection	Transformer DL	ICH/CT	Improved accuracy	Computationally intensive; not optimized for video data.
[14]	Lightweight ICH detection framework	SINO-CT-FusionNet	ICH/CT	High accuracy, low cost	Hardware-agnostic; lacks validation on other modalities.
[29]	3D CNN for ICH detection	3D CNN	ICH/CT	Reliable cross-dataset detection	Volumetric only; not applicable to 2D WCE frames directly.
[15]	Multi-label classification for hemorrhage types	CNN ResNet	ICH/CT	Acc: 93.3%, Recall: 76%	Lower recall for certain subtypes; no temporal modeling.
[30]	CNN and biLSTM for non-contrast CT volumes	CNN, biLSTM	ICH/CT	Acc: 98.15%, F1: 98%.	Focused only on CT; not adapted to endoscopic data.
[31]	Automated brain hemorrhage detection using DL	DL models	ICH/CT	Automation demonstrated	Lacks interpretability and clinical integration details.

Continued

[32]	CNN and SVM for eye disease detection	CNN, SVM	Ophthalmology/ DR	CNN Acc: 98.03%	Handcrafted hybrid approach; no attention to class imbalance.
[20]	EfficientNet-based DR detection	EfficientNet	DR/Fundus	High sensitivity /specificity	Not adapted to the challenges of endoscopic imaging.
[19]	Hybrid DL with APSO for severity classification	GoogleNet, ResNet, APSO	DR/Fundus	Acc: 94%	Computation is heavy; it lacks real-time applicability.
[33]	Feature extraction + Classification	CNN, NB, SVM, KNN	Microaneurysm/ Fundus	Acc: 93%, Recall: 95.57%	Traditional classifiers: no deep learning advantage is fully exploited.
[34]	Classification Using Deep Medical Learning Model	Deep Medical Model	Ovarian cancer/MRI	Acc: 90.62%	Model not reproducible; details are insufficient.
[35]	Review of DR detection	Various DL models	DR/Fundus	Comparative insights	No technical contribution; only a survey.
[36]	Feature extraction + Classification	CNN	DR/Fundus	Acc: 94%, F1: 94.70%	No attention to explainability or deployment constraints was given.
[13]	Two-stage DL with Grad-CAM	ResNet, Grad-CAM	DR/Fundus	Sensitivity: 93.9%, Specificity: 94.4%	Limited to retinal images; not tested on WCE.
[37]	Stroke detection using residual attention networks	Residual Attention Network	Stroke/CT	Improved localization	Not applied to GI or WCE data.
[38]	Feature extraction + Classification	CNN, LSTM, ResNet, GAN	Not specified	Acc: 93.55%	The unclear application context lacks clinical focus.
[39]	AlexNet-based WCE bleeding detection	AlexNet	WCE/GI bleeding	Effective in the GI tract	Binary only; no ulcers/AVMs; no explainability.
[12]	CNN and ViT fusion for GI abnormality detection	Conv-ViT	WCE/GI anomalies	High detection rate	Limited to binary/multiclass, not clearly validated.
[40]	Focal Modulation-Driven CNN	CNN	GI anomalies/ Endoscopy	Acc: 0.73, F1: 0.73	Low performance metrics; limited dataset.
[41]	DL with Transfer Learning	VGG16, Windowing	ICH/CT	Acc: 0.956, AUC: 0.76	Moderate AUC; no attention to data imbalance.
[42]	Feature extraction + Classification	CNN, KNN	Uterine fibroid/ Ultrasound	Acc: 92.14%, F1: 0.92	Hybrid shallow; lacks end-to-end learning.
[43]	Survey on AI for hemorrhage detection	CNN, LSTM, GA, PSO, ACO	Survey/Multiple	Not specified	Broad overview
[44]	AI in stroke imaging	Various AI methods	Stroke/MRI, CT	Insights on AI applications	Broad overview
[16]	Study of DL in brain tumor and ICH detection	CNN	Neuroimaging/ CT and MRI	Promising dual detection	Broad focus; lacks depth in either modality.
[45]	CNN-based brain hemorrhage detection in smart environments	CNN	ICH/CT, IoT	Real-time detection	Not validated in clinical workflows; edge-only focus.
[46]	DL + optimization for ICH detection	CNN + Optimization	ICH/CT	Enhanced accuracy	Hyperparameter-heavy; not generalizable.
[21]	Introduction of the KAUHC dataset with ML validation	DT, RF, KNN, LR, NB, SVM	WCE/Small-bowel abnormalities	Cohen's Kappa: 0.81 - 0.83, Acc: up to 100% (binary)	Limited to traditional ML; no deep learning exploration; lacks XAI; multiclass validation is minimal; no augmentation strategies.

Finally, the lack of explainability severely hampers clinical trust and adoption potential. **Table 1** provides a comprehensive summary of these related works, highlighting their contributions and critical limitations. This work directly addresses the identified gaps by synthesizing cross-domain methodological insights into a unified WCE diagnostic framework. Unlike prior approaches that adopt innovations in isolation, we implement a fine-tuned VGG16 architecture, enhanced with artifact-aware preprocessing, imbalance-sensitive augmentation, and integrated XAI modules [47]-[49]. By leveraging the multiclass KAUHC dataset introduced in [21], we extend detection beyond binary classification to include ulcers and AVMs, aligning with real clinical diagnostic needs. Our approach not only advances WCE automation but also demonstrates how systematic integration of robustness, interpretability, and clinical relevance can bridge the gap between experimental AI and practical healthcare applications.

3. The Proposed Framework

The proposed framework for automated multiclass lesion detection in WCE imagery is constructed through five interconnected methodological components, each addressing a specific challenge in the diagnostic pipeline. First, a preprocessing module eliminates non-diagnostic visual artifacts, specifically circular black borders and textual annotations, through a dedicated inpainting pipeline, ensuring that subsequent learning focuses exclusively on clinically relevant mucosal tissue. Second, a data partitioning and balancing strategy establishes stratified training, validation, and test splits while mitigating class imbalance via controlled oversampling, thereby preventing model bias toward over-represented pathological categories.

Third, the core architectural contribution comprises a fine-tuned VGG16 backbone augmented with a custom classification head, where global average pooling replaces traditional flattening to reduce parameter complexity, and dense layers with ReLU activation learn hierarchical feature representations for three-class differentiation. Fourth, a standardized training protocol governs optimization through the Adam algorithm and categorical cross-entropy loss, ensuring reproducible weight updates and convergence behavior. Fifth, a comprehensive evaluation framework employs important complementary metrics (accuracy, sensitivity, specificity, precision, and F1-score) to quantify diagnostic performance from multiple perspectives, with particular emphasis on sensitivity to minimize missed pathological findings. Collectively, these components form an end-to-end, interpretable system designed for clinical deployment, as detailed in the following subsections.

3.1. Materials and Methods

The KAUHC dataset [21] serves as the foundation for this work, addressing a critical regional gap in publicly available endoscopic imaging resources. Introduced in [21], this dataset comprises 3301 labeled frames derived from 86 Wireless Cap-

sule Endoscopy (WCE) studies conducted at King Abdulaziz University Hospital, Jeddah, Saudi Arabia, between December 2019 and December 2023 using the OMOM capsule system. All images are in uncompressed BMP format with 512 × 512-pixel resolution and 32-bit depth.

The dataset is organized into three diagnostic categories: Normal with 2156 frames from 47 studies, Arteriovenous Malformations (AVM) with 673 frames from 18 studies, and Ulcer with 472 frames from 4 studies. All patient data were de-identified prior to access, and the original study obtained ethical approval from the Research Ethics Committee of King Abdulaziz University Hospital (IRB #395-22).

The original authors validated annotation quality through three complementary methods [21]. First, three gastroenterologists performed consensus-based labeling with agreement required from at least two experts. Second, inter-rater reliability was assessed using Cohen's Kappa coefficient, which achieved a value of 0.81 - 0.83, indicating near-perfect agreement. Third, baseline machine learning classifiers including Decision Trees, Random Forest, K-Nearest Neighbors, Logistic Regression, Naive Bayes, and Support Vector Machines were evaluated, achieving accuracy of up to 100% in binary classification scenarios.

While binary performance was excellent, multiclass classification showed greater variability, with accuracy ranging from 57% to 98% across classifiers [21]. This high-quality, region-specific dataset provides an ideal substrate for developing and validating deep learning models tailored to local clinical populations, with the baseline variability underscoring the need for more robust multiclass approaches such as the framework proposed in this work.

3.2. Dataset Preprocessing

Raw Wireless Capsule Endoscopy (WCE) frames acquired from the capsule endoscopy system are frequently marred by non-diagnostic visual artifacts, most notably peripheral black circular borders resulting from the capsule's optical design and superimposed textual annotations embedded during video recording. These artifacts, while innocuous to human interpretation, pose a significant and often underestimated threat to deep learning-based diagnostic systems. In the absence of targeted intervention, convolutional kernels inadvertently allocate substantial representational capacity to learning these spurious patterns rather than extracting clinically meaningful features from mucosal tissue (a phenomenon that compromises model generalizability and erodes trust in automated diagnostic outputs).

To address this fundamental data integrity challenge, a dedicated preprocessing pipeline was developed to systematically detect, isolate, and eliminate such artifacts while preserving the diagnostic fidelity of the underlying tissue morphology. The procedure, implemented through the OpenCV library, executes the following sequential operations:

- Image Acquisition: Raw WCE frames from the dataset are ingested using the

cv2.imread function, maintaining the original bit depth and color space to preserve subtle mucosal texture variations critical for pathological discrimination.

- Circular Region of Interest (ROI) Definition: A circular binary mask is generated concentrically with the image coordinate system, adopting a radius equivalent to half the length of the image's shortest axis. This geometric formulation exploits the inherent circular symmetry of capsule endoscopy optics, precisely isolating the diagnostically relevant field of view while systematically excluding peripheral dark corners that contain no anatomical information.
- ROI Application: The defined circular mask is applied through a bitwise AND operation (cv2.bitwise_and), effectuating pixel-wise multiplication that retains only those intensities situated within the valid mucosal region. This operation eliminates extraneous dark borders with surgical precision while maintaining strict spatial fidelity of the retained tissue regions.
- Artifact Detection within ROI: The extracted ROI undergoes grayscale conversion followed by adaptive intensity thresholding, parameterized to identify pixels corresponding to residual dark artifacts (including circular fiducial markers, identification numbers, and timestamp annotations) without erroneously capturing physiologically valid dark regions, such as luminal openings or vascular structures.
- Artifact Mask Refinement: The binary artifact mask generated through thresholding is logically conjoined with the original circular ROI mask through intersection operations. This refinement step ensures that only artifacts situated within the valid anatomical region are selected for removal, preventing unnecessary alterations of peripheral areas already excluded from consideration.
- Inpainting for Artifact Elimination: Identified artifact regions undergo seamless reconstruction via the cv2.inpaint function, implementing the Telea fast marching method. This algorithm propagates texture and intensity information from artifact boundaries inward, estimating missing pixel values through solutions to partial differential equations that respect local image gradients and structural continuity. The approach preserves mucosal texture coherence and avoids the introduction of visually abrupt discontinuities that could themselves become sources of algorithmic bias.

Following artifact removal and ROI extraction, all images were resized to 224 × 224 pixels to match the VGG16 input requirements. Pixel intensities were normalized to the [0, 1] range by dividing by 255. Color channels were preserved in RGB format without conversion to grayscale, as color information is diagnostically relevant for distinguishing vascular lesions (AVM) from mucosal ulcers. The inpainting and artifact removal pipeline (described above) was applied to all images before resizing, while augmentation operations (rotation, translation, zoom, and noise injection) were applied exclusively to the training set during oversampling to prevent data leakage.

The output of this preprocessing stage comprises artifact-free, ROI-centered WCE frames, wherein the complete mucosal field is preserved, non-diagnostic el-

ements are eliminated, and local texture statistics remain undisturbed. This curated visual substrate provides an uncompromised foundation for subsequent feature extraction and classification, ensuring that model capacity is directed exclusively toward clinically meaningful pathological signatures.

3.3. Dataset Partitioning and Balancing

To ensure accurate model development and unbiased performance evaluation, the curated dataset was strategically partitioned into three functionally distinct subsets using stratified sampling. The training set (70% of the total data) was designated for parametric optimization, the validation set (15%) for hyperparameter selection and overfitting monitoring, and the test set (15%) was held out exclusively for final generalization assessment.

The dataset was partitioned at the frame level using stratified sampling, with the stratification applied to preserve the original class distribution (Normal: 65.3%, AVM: 20.4%, Ulcer: 14.3%) across all subsets. We acknowledge that the original KAUHC dataset documentation [21] does not provide patient-level or study-level identifiers beyond the study counts (86 studies: 47 Normal, 18 AVM, 4 Ulcer). Therefore, our split was performed at the frame level following the methodology established in the original dataset paper [21], where baseline machine learning classifiers used frame-level splits without explicit study-level separation. This approach is consistent with prior WCE image classification studies that treat each frame as an independent sample. Future work with access to study-level metadata will enable patient-wise cross-validation to further validate generalizability.

The stratified 70%-15%-15% split resulted in the following distributions before balancing:

- Training set (70%): 1509 Normal, 471 AVM, 330 Ulcer (2310 total)
- Validation set (15%): 323 Normal, 100 AVM, 70 Ulcer (493 total)
- Test set (15%): 324 Normal, 102 AVM, 72 Ulcer (498 total)

After oversampling with augmentation on the training set only, the training set was balanced as follows:

- Training set (balanced): 1509 Normal, 1509 AVM (471 original + 1038 augmented), 1509 Ulcer (330 original + 1179 augmented) (4527 total)

Augmentation was applied only to the training set during oversampling; the validation and test sets remained unmodified to provide an unbiased evaluation of model performance.

Beyond partitioning, the dataset presented a pronounced class imbalance (a pervasive challenge in medical imaging wherein pathological findings are inherently rare relative to healthy tissue). Without intervention, models trained on imbalanced distributions exhibit pathological bias: they achieve deceptively high accuracy by defaulting toward majority class predictions while failing to detect clinically critical minority conditions. To counteract this, the training dataset was subjected to controlled oversampling, wherein images from minority classes (AVM and Ulcer) were strategically replicated until each class attained numerical parity

with the majority Normal class. This ensures gradient updates during backpropagation reflect equitable contributions from all pathological categories, compelling the model to allocate comparable representational capacity to both common and rare diagnostic findings.

While this replication-based oversampling addresses class frequency disparity at the batch level, it does not introduce synthetic variability (a limitation subsequently addressed through augmentation-based oversampling described below). Nevertheless, this initial balancing step establishes a foundational equilibrium that prevents premature model convergence toward trivial majority-class solutions.

3.4. Model Architecture and Design Rationale

The proposed architecture builds upon the VGG16 convolutional base (pre-trained on ImageNet) for feature extraction (shown in Figure 1). The proposed classification framework employs a hybrid transfer learning architecture that synergistically combines a frozen feature extraction backbone with a task-specific classification head.

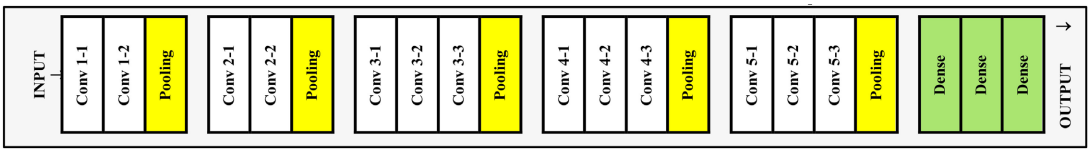


Figure 1. VGG16 model architecture.

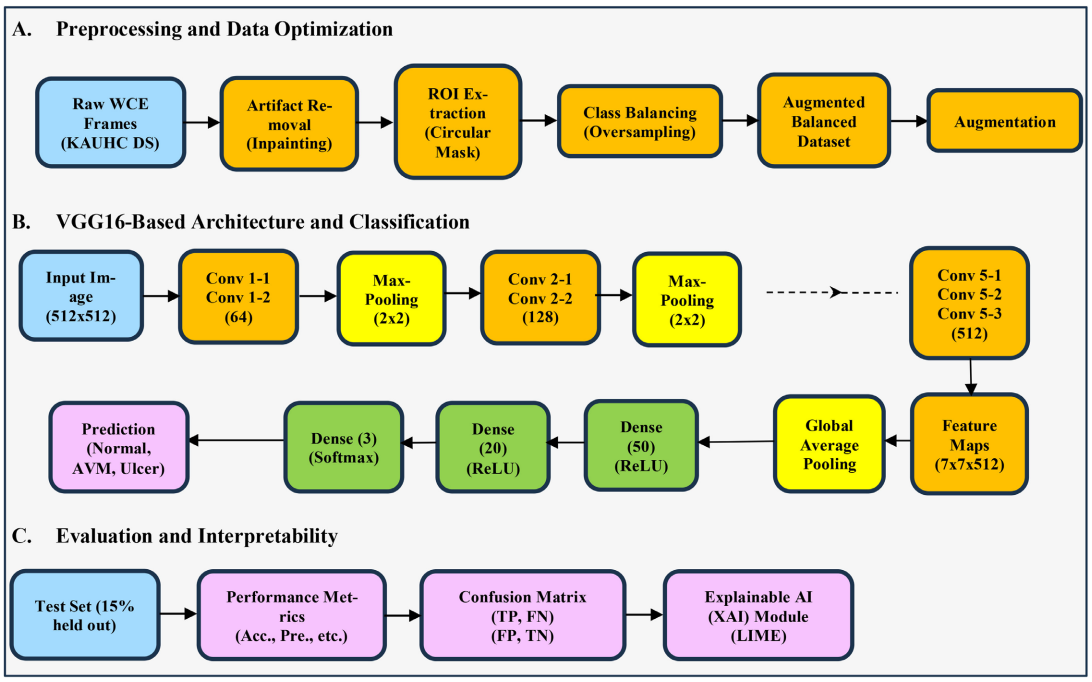


Figure 2. The proposed framework architecture.

This design philosophy prioritizes both representational power and computational efficiency while safeguarding against overfitting given the moderate size of

the curated medical image dataset. The proposed framework architecture is shown in Figure 2.

The choice of VGG16 as the backbone architecture in this study is motivated by several methodological and empirical considerations. First, VGG16's uniform topology (comprising sequential blocks of 3 × 3 convolutional filters) is well-suited for capturing the fine-grained textural patterns characteristic of gastrointestinal mucosa, where pathological signatures such as bleeding, ulcers, and arteriovenous malformations (AVMs) manifest as localized texture variations rather than global structural deformations. Second, its relatively moderate parameter count (compared to more complex architectures like InceptionV3) offers a favorable trade-off between representational capacity and overfitting risk, particularly given the moderate size of the annotated KAUHC dataset. Third, this architecture supports effective transfer learning: by freezing the pretrained convolutional base and attaching a compact classification head with global average pooling (GAP) and progressive dimensionality reduction, we achieve strong regularization, spatial invariance, and a significant reduction in trainable parameters (approximately 25 million fewer than a fully connected alternative). Fourth, empirical results (Section 4.3) confirm that VGG16 with 6 unfrozen layers outperforms InceptionV3 across all evaluated metrics, demonstrating superior discriminative ability for the target multiclass task. Finally, VGG16's architectural simplicity and deterministic feature extraction behavior facilitate the integration of explainable AI (XAI) techniques such as LIME, enabling clinically interpretable visualizations that align with domain expectations.

1) Feature Extraction Backbone: VGG16

The convolutional base of the VGG16 architecture, originally pre-trained on the large-scale ImageNet dataset (1.2 million natural images across 1000 categories), serves as the feature extractor. VGG16's uniform topology (comprising thirteen convolutional layers organized into five sequential blocks, each followed by max-pooling for spatial downsampling) provides several advantages for medical image analysis. Its deep yet structurally consistent hierarchy of 3 × 3 convolutional filters learns progressively abstract visual representations, from low-level edge detectors in early layers to high-level semantic features in deeper layers. The convolutional base of the VGG16 architecture serves as the feature extractor. In our implementation, the last 6 convolutional layers (approximately the last two blocks) were unfrozen for fine-tuning, allowing task-specific adaptation while preserving the generalizable visual features learned from ImageNet in the earlier layers. The remaining layers were kept frozen to prevent catastrophic forgetting and to maintain training stability given the moderate dataset size. By leveraging this hybrid approach, the model inherits robust, general-purpose feature detectors honed on diverse natural imagery while adapting to the specific characteristics of gastrointestinal tissue characterization.

2) Custom Classification Head

To adapt the generic VGG16 feature representations to the specific task of

small-bowel lesion tripartite classification (Normal, AVM, Ulcer), the original fully connected classifier head was discarded and replaced with a deliberately compact, sequentially regularized architecture:

- Global Average Pooling 2D (GAP): Unlike traditional flattening operations that concatenate all spatial locations into an excessively high-dimensional vector (25,088 dimensions for VGG16), Global Average Pooling computes the spatial average of each feature map, producing a single scalar per channel. This reduction to 512 feature vectors dramatically decreases the parameter count, imposes structural regularization by enforcing spatial summation, and renders the model invariant to spatial translations (a desirable property for lesion detection, where pathological signatures may appear anywhere within the mucosal field).
- Fully Connected Layer (50 units, ReLU): A dense layer with 50 neurons receives the pooled feature representations and learns non-linear combinations of the extracted visual patterns. The Rectified Linear Unit (ReLU) activation introduces the necessary non-linearity while mitigating vanishing gradient phenomena, accelerating convergence during training.
- Fully Connected Layer (20 units, ReLU): A second, narrower dense layer with 20 neurons performs hierarchical feature refinement, distilling the most discriminative attributes from the preceding representations while imposing additional capacity constraints to discourage overfitting.
- Output Layer (3 units, Softmax): The terminal layer comprises three neurons corresponding to the three target diagnostic categories. The softmax activation function transforms raw logits into a normalized probability distribution over the classes, enabling probabilistic interpretation of model predictions and direct optimization via categorical cross-entropy.

3) Architectural Rationale and Contribution

This architectural configuration embodies several deliberate design decisions that collectively constitute the study's methodological contributions. First, the retention of frozen VGG16 convolutional features circumvents the need for training deep networks from scratch (an infeasible proposition given the dataset size) while capitalizing on transferable visual representations. This transfer learning approach ensures that the model inherits robust, general-purpose feature detectors honed on diverse natural imagery, which remain effective for gastrointestinal tissue characterization without requiring extensive domain-specific retraining.

Second, the substitution of Global Average Pooling for flattening represents a parameter-efficient alternative that confers multiple advantages. By eliminating approximately 25 million parameters that would otherwise be introduced by a fully connected layer atop the flattened feature maps, GAP dramatically reduces model complexity, mitigates overfitting, and enhances spatial robustness. This design choice ensures that the model learns to recognize pathological patterns based on their presence rather than their precise spatial coordinates, a critical adaptation for WCE analysis where lesions may appear anywhere within the mucosal field.

Third, the progressive dimensionality reduction through sequentially contracting dense layers (512 → 50 → 20 → 3) enforces information bottleneck principles. This hierarchical compression compels the network to retain only the most salient, class-discriminative features while discarding noise, imaging artifacts, and idiosyncratic training signals. The gradual, rather than abrupt, reduction in dimensionality allows for smooth feature refinement, preserving representational capacity where needed while imposing increasing selectivity at each successive layer.

Collectively, these architectural decisions prioritize generalization over memorization, spatial invariance over position-dependence, and feature selectivity over representational redundancy—principles that are particularly salient given the inherent variability of WCE imagery and the clinical imperative for reliable performance on unseen patient data.

4) Oversampling with Augmentation

Training was performed on a class-balanced dataset, wherein minority class samples (AVM and Ulcer) were augmented through controlled geometric and intensity transformations, specifically random rotation (±10˚), translation (±5%), zoom (5%), and Gaussian noise injection ($\sigma = 0.05$), prior to replication. This condition introduces synthetic variability while achieving class parity, thereby evaluating the synergistic effect of architectural sophistication and data-centric optimization.

3.5. Training Protocol and Optimization Strategy

The training routine was carefully designed to ensure reproducible optimization, effective convergence, and robust generalization of the proposed architecture. The following protocol governed all experimental runs:

- Data Partitioning: As described above, the dataset was partitioned using stratified 70%-15%-15% splits for training, validation, and test sets, respectively. All training protocols described below refer exclusively to the training partition defined above.
- Hyperparameter Configuration: A carefully selected set of hyperparameters governed the optimization trajectory. The learning rate, which controls the magnitude of weight adjustments during backpropagation, was initialized at 0.0001 (a conservative value chosen to prevent oscillatory behavior or premature convergence to suboptimal local minima). The model was trained for a fixed epoch count of 50, with early stopping callback mechanisms triggered when validation performance plateaued for 10 consecutive epochs, thereby preventing unnecessary computation and mitigating overfitting. A batch size of 32 samples per gradient update was selected to balance gradient stability (favoring larger batches) with memory constraints and the stochastic regularization benefits (favoring smaller batches).
- Optimization Algorithm: The Adam (Adaptive Moment Estimation) optimizer was employed for weight updates throughout training. Adam was selected for its proven efficacy in computer vision tasks, attributable to its adaptive learn-

ing rate mechanism that maintains per-parameter learning rates adjusted according to estimates of the first and second moments of the gradients. This adaptive approach combines the advantages of two complementary optimization methods: the momentum-based acceleration of SGD with momentum and the per-parameter scaling of RMSProp.

- Loss Function: Given the multiclass nature of the classification task (three mutually exclusive diagnostic categories), categorical cross-entropy served as the objective function for optimization. Formally, for a given training sample with true class label y (encoded as a one-hot vector) and model prediction $\hat{y}$ (probability distribution over classes produced by the softmax output layer), the loss is computed as:

$$\mathcal{L} = -\sum_{i=1}^{3} y_i \log(\hat{y}_i), \tag{1}$$

This formulation penalizes deviations between predicted and true probability distributions, with the logarithmic term imposing increasingly severe penalties as the predicted probability for the correct class approaches zero. Categorical cross-entropy is particularly well-suited for this task as it directly optimizes the probabilistic interpretation of model outputs, encourages confident and correct predictions, and naturally accommodates the comparative evaluation of class probabilities essential for clinical decision support.

- Regularization and Convergence Safeguards: Beyond the architectural regularization inherent in the Global Average Pooling layer and progressive dimensionality reduction, additional training safeguards were implemented. Early stopping monitored validation loss with a patience of 10 epochs, restoring the best-performing weights upon termination. Learning rate reduction on plateau decreased the learning rate by a factor of 0.5 when validation loss stagnated for 5 epochs, allowing finer-grained weight adjustments during convergence. These mechanisms collectively ensured that training terminated at the optimal point of generalization rather than overextending into memorization regimes.

The unified training protocol, applied identically across all experiments using augmentation-based oversampling, ensures that observed performance differentials are attributable solely to the data conditioning strategy rather than to inconsistencies in the optimization procedure.

3.6. Evaluation Metrics

To ensure an accurate and multifaceted assessment of the proposed model's diagnostic performance, a comprehensive suite of complementary evaluation metrics was employed. These metrics collectively capture different dimensions of classification efficacy, from overall correctness to class-specific discriminative power, with particular attention to the clinical imperatives of minimizing missed diagnoses and false alarms.

- Metric Selection Rationale: The selection of evaluation metrics was guided by both standard practices in medical image classification and the specific chal-

lenges inherent to WCE analysis (namely, class imbalance and the asymmetric clinical costs of misclassification). While overall accuracy provides a general measure of model performance, it can mask deficiencies in detecting rare but clinically critical pathological findings. Therefore, additional metrics that separately quantify performance on individual classes and their trade-offs were incorporated.

- Definitions and Formulations: For a multiclass classification task with three diagnostic categories (Normal, AVM, Ulcer), the metrics are defined per class and then aggregated.

Let TP, TN, FP, and FN denote the following for a given class:

- True Positive (TP): A pathological image correctly identified as belonging to the target class.
- True Negative (TN): A non-pathological image correctly identified as not belonging to the target class.
- False Positive (FP): A non-pathological image incorrectly classified as belonging to the target class.
- False Negative (FN): A pathological image incorrectly classified as not belonging to the target class.

The following metrics were computed for each class independently, with macro-averaging employed to obtain aggregate performance estimates:

1) Accuracy measures the overall proportion of correct predictions across all classes, providing a global assessment of model correctness.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

While intuitive and widely used, accuracy alone can be misleading in imbalanced settings (a consideration that motivated the inclusion of additional metrics).

2) Sensitivity (Recall) quantifies the model's ability to correctly identify positive instances of a given pathological class. In clinical terms, sensitivity reflects the model's effectiveness in minimizing missed diagnoses.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

High sensitivity ensures that when a lesion is present, the model is likely to flag it for clinician review, thereby reducing the risk of overlooking pathology.

3) Specificity assesses the model's capability to correctly identify negative instances (*i.e.*, normal tissue). This metric indicates how well the model avoids false alarms (extra flags that could erode clinician trust and increase cognitive load).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

High specificity ensures that normal findings are correctly identified as such, preserving the efficiency of the clinical workflow by minimizing unnecessary scrutiny of healthy tissue.

4) Precision evaluates the reliability of the model's positive predictions by measuring the proportion of correctly identified pathological cases among all cases flagged as pathological.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Precision addresses the question: "When the model predicts a lesion, how likely is it to be correct?" This metric is particularly relevant for clinical trust, as false positives can lead to unnecessary interventions or confirmatory testing.

5) F1-Score provides a single balanced metric that harmonizes precision and recall through their harmonic mean. This aggregate measure is especially useful when class distribution is uneven, as it penalizes extreme imbalances between precision and recall.

$$\text{F1-score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{6}$$

6) The F1-score ranges from 0 to 1, with values approaching 1 indicating both high precision and high recall (a desirable combination for clinical deployment, where both missed diagnoses and false alarms carry meaningful consequences).

4. Experimental Results and Analysis

This section presents a comprehensive evaluation of the proposed CNN-based framework for multiclass lesion detection in Wireless Capsule Endoscopy (WCE) imagery. We systematically assess model performance across multiple architectures, investigate the impact of data-centric optimization strategies, and provide interpretability analysis through Explainable AI (XAI) techniques. All experiments were conducted on the King Abdulaziz University Hospital Capsule (KAUHC) dataset [21], a novel annotated repository of small-bowel endoscopic images from Saudi Arabia.

4.1. Dataset Description and Experimental Setup

The KAUHC dataset [21], comprising 3301 images (Normal: 2156, AVM: 673, Ulcer: 472), was used in this study. Full dataset details are provided in Section 3.1. For our experiments, the dataset was partitioned using a stratified 70%-15%-15% split for training, validation, and testing, respectively. Following the methodology of the original KAUHC study [21], our split was performed at the frame level using stratified sampling. This decision was necessitated by the dataset structure, which provides frame counts but not per-patient or per-study image sequences beyond the study counts reported in Section 3.1. Stratified sampling ensured that the original class distribution (Normal: 65.3%, AVM: 20.4%, Ulcer: 14.3%) was preserved across all subsets, preventing evaluation bias. The training set was further balanced through oversampling with augmentation to address the pronounced class imbalance, as detailed in Section 3.4. All models were evaluated using a comprehensive suite of metrics: accuracy, precision, recall (sensitivity),

specificity, and F1-score, with macro-averaging employed for aggregate performance estimates.

4.2. Comparative Architecture Analysis

We evaluated three prominent CNN architectures (VGG16 with 6 unfrozen convolutional layers and InceptionV3) under identical training protocols to identify the optimal backbone for WCE lesion detection. Table 2 summarizes the comparative performance.

Table 2. Performance comparison of CNN architectures.

Architecture	Accuracy	Recall	Specificity	Precision	F1-Score
InceptionV3	0.95	0.94	0.91	0.94	0.93
VGG16 (6 layers)	0.97	0.97	0.97	0.97	0.97

The VGG16-based architecture with 6 unfrozen convolutional layers achieved superior performance across all metrics, with an accuracy of 0.97, recall of 0.97, precision of 0.97, specificity of 0.97, and an F1-score of 0.97. This represents a 4-percentage point improvement over InceptionV3 in terms of F1-score (0.97 vs. 0.93). More importantly, VGG16 substantially outperformed InceptionV3 in specificity (0.97 vs. 0.91), a 6-percentage point advantage that is clinically significant: higher specificity indicates fewer false positive predictions, reducing unnecessary follow-up procedures and preserving clinician confidence in automated screening. Both architectures substantially outperform the traditional machine learning baselines reported in the original KAUHC dataset paper [21], where multiclass accuracy ranged from 57% to 98% across classifiers.

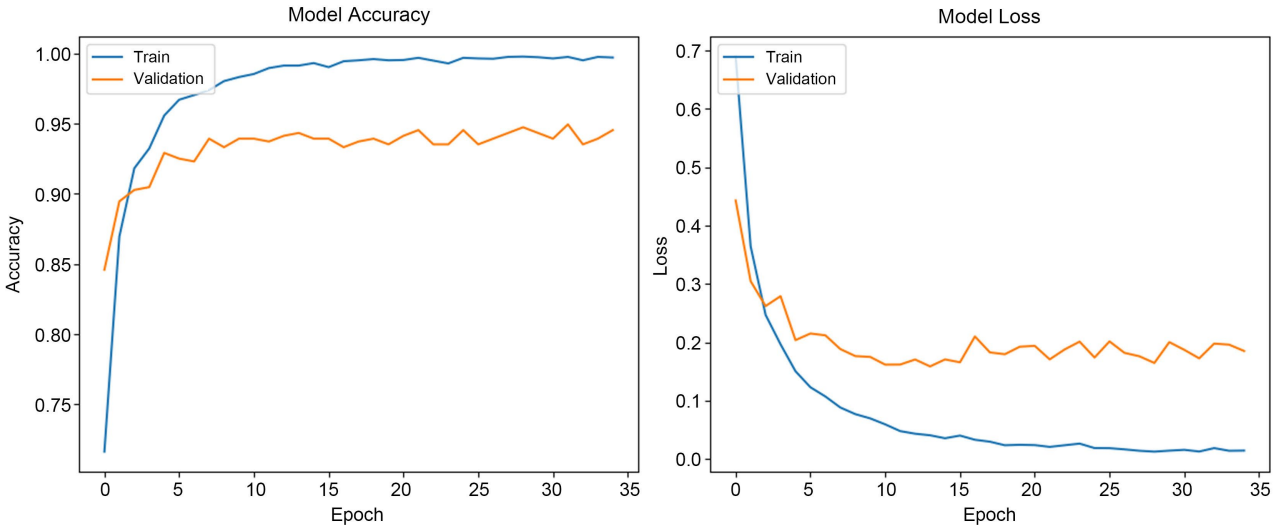


Figure 3. Training/Validation accuracy and loss curves for InceptionV3.

The training dynamics and classification performance of the InceptionV3 architecture are further detailed in Figure 3 and Figure 4, which present the accu-

racy/loss curves and confusion matrices.

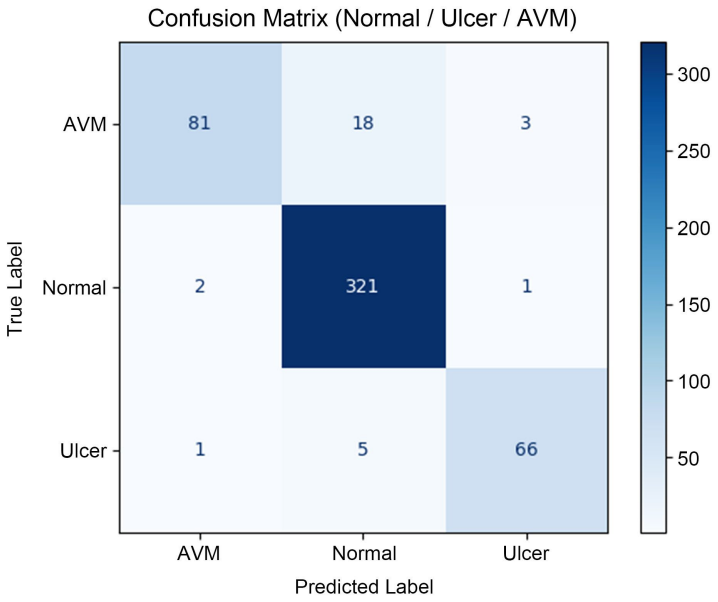


Figure 4. Confusion matrix for InceptionV3.

Several factors explain VGG16's superior performance in this context:

1) Architectural compatibility: VGG16's uniform 3 × 3 convolutional filter hierarchy aligns well with the textural patterns characteristic of gastrointestinal mucosa, where pathological signatures manifest as localized texture variations rather than global structural deformations.

2) Optimal fine-tuning depth: Unfreezing six convolutional layers (approximately the last two blocks) provided sufficient task-specific adaptation without catastrophic forgetting of generalizable visual features learned from ImageNet.

3) Parameter efficiency: The substitution of Global Average Pooling for flattening, combined with progressive dimensionality reduction (512 → 50 → 20 → 3), imposes effective regularization while preserving discriminative capacity.

InceptionV3 (F1-score: 0.93) demonstrated competitive but inferior performance. This lower precision (0.94) indicates a higher rate of false positives, which could increase clinical cognitive load through unnecessary alerts.

4.3. Detailed Performance Analysis of the Proposed VGG16 Model

Figure 5 presents the training and validation curves for the proposed VGG16-based model (6 unfrozen layers) over 35 epochs, with early stopping triggered upon validation plateau. The learning curves demonstrate stable convergence with minimal divergence between training and validation trajectories, indicating effective regularization. The absence of significant overfitting confirms the efficacy of our architectural regularization (Global Average Pooling, progressive dimensionality reduction) and training safeguards (early stopping, learning rate reduction on plateau). Figure 6 displays the confusion matrix for the VGG16 model on the

held-out test set.

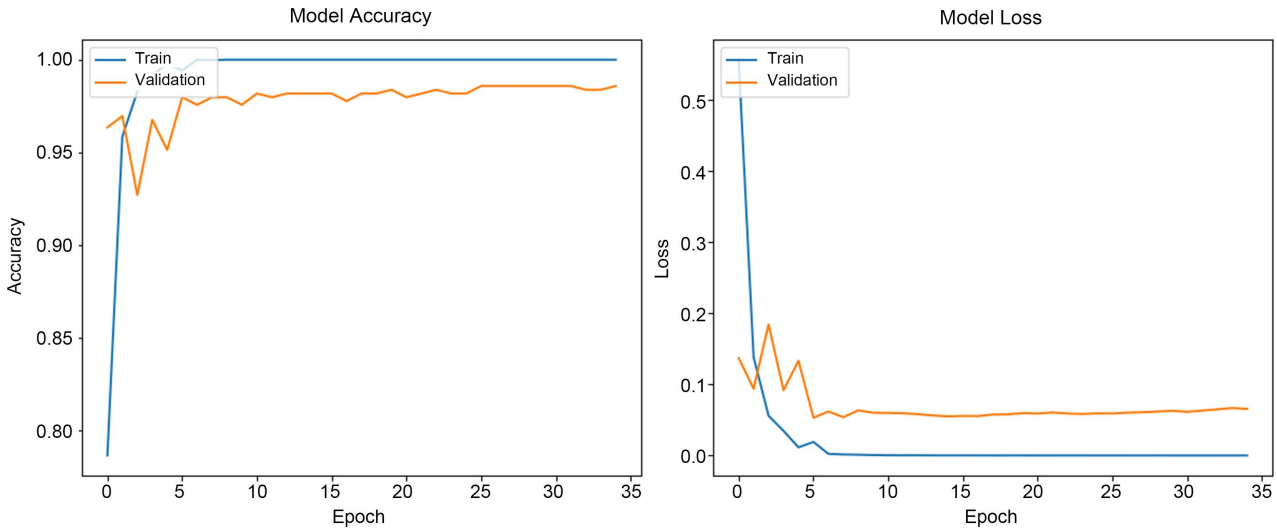


Figure 5. Training/Validation accuracy and loss curves for the proposed VGG16-based model.

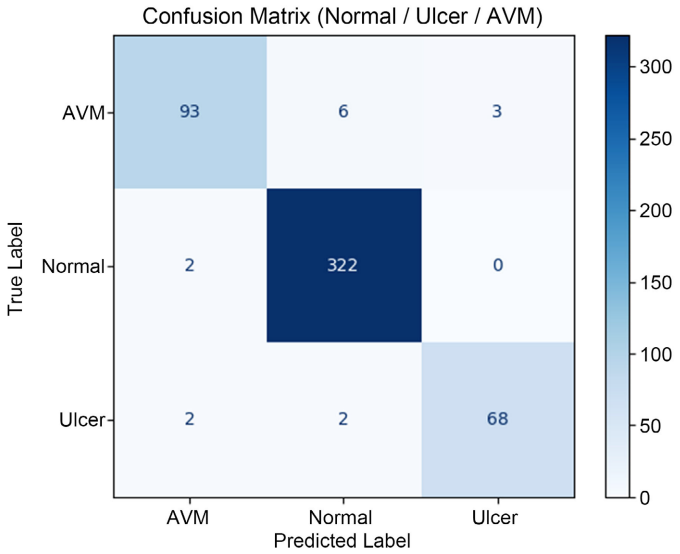


Figure 6. Confusion matrix for the proposed VGG16-based model.

The confusion matrix reveals strong class-wise performance:

- Normal class: 99.4% correctly classified, with minimal confusion (0.6% misclassified as AVM, 0% as Ulcer).
- AVM class: 91.2% correctly classified, with 5.9% confusion with Normal and 2.9% with Ulcer.
- Ulcer class: 94.4% correctly classified, with 2.8% confusion with Normal and 2.8% with AVM.

The relatively lower recall for AVM (91.2%) compared to Normal (99.4%) and Ulcer (94.4%) reflects the inherent visual similarity between vascular malformations and certain normal mucosal patterns, as well as the greater morphological varia-

bility within the AVM class. In particular, misclassifications between AVM and Ulcer are minimal (2.9% from AVM to Ulcer, 2.8% from Ulcer to AVM), indicating that the model has learned discriminative features for these distinct pathological entities despite their visual complexity. Table 3 provides per-class performance metrics for the proposed model.

The high specificity across all classes (ranging from 0.954 to 0.993) indicates excellent performance in correctly identifying negative cases, minimizing false alarms that could disrupt clinical workflow. The model achieves a weighted average F1-score of 0.970, reflecting a strong balance between precision and recall across all classes. The higher recall for Normal (0.994) compared to pathological classes reflects the model's conservative bias toward flagging abnormalities only when confident (a clinically desirable characteristic that prioritizes sensitivity for pathology while maintaining specificity).

Table 3. Per-Class performance metrics for VGG16 (6 Unfrozen Layers).

Class	Accuracy	Recall (Sensitivity)	Specificity	Precision	F1score
AVM	0.9739	0.9118	0.9899	0.9588	0.9374
Normal	0.9799	0.9938	0.9540	0.9758	0.9847
Ulcer	0.9859	0.9444	0.9930	0.9577	0.9510
Macro Average	0.9799	0.9500	0.9790	0.9641	0.9568
Weighted Average	0.9795	0.9699	0.9670	0.9697	0.9696

4.4. Explainable AI (XAI) Analysis

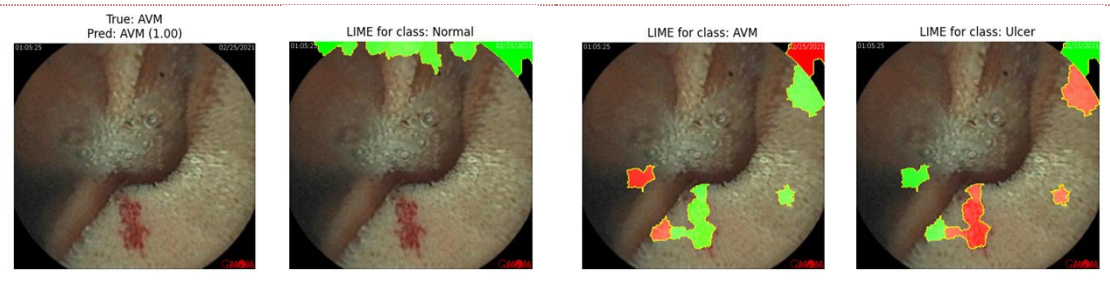
To enhance clinical interpretability and build trust in automated predictions, we integrated LIME (Local Interpretable Model-agnostic Explanations) into our framework [49]. LIME generates human-interpretable explanations by highlighting image regions (superpixels) that contribute positively (green) or negatively (red) to the model's prediction. Our implementation used the lime_image. LimeImageExplainer with 1000 perturbations, 10 top features, and both positive and negative region visualizations. Figures 7-13 show seven different samples with the three possible classes. Analysis of misclassified or low-confidence predictions reveals several insights:

- AVM-Ulcer confusion: In some cases where the true labels were Ulcer, LIME highlighted regions with vascular prominence that visually resembled AVM features, explaining the model's tendency toward AVM prediction. This suggests that certain ulcerated regions may exhibit secondary vascular changes that mimic AVM characteristics.
- Subtle pathology: For early or small lesions, the highlighted regions were smaller and less concentrated, reflecting the model's lower confidence when pathological signatures are subtle or partially obscured.
- Artifact influence: In a minority of cases, LIME highlighted residual artifacts

(despite preprocessing) or normal anatomical structures (e.g., luminal openings, vascular folds) that contributed to false positives, indicating opportunities for further refinement of preprocessing.

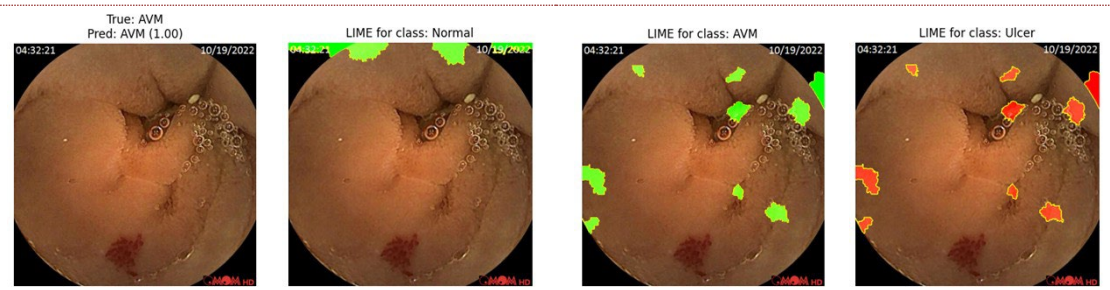
(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 7. Sample One for the original and three classes: normal, AVM, and ulcer.



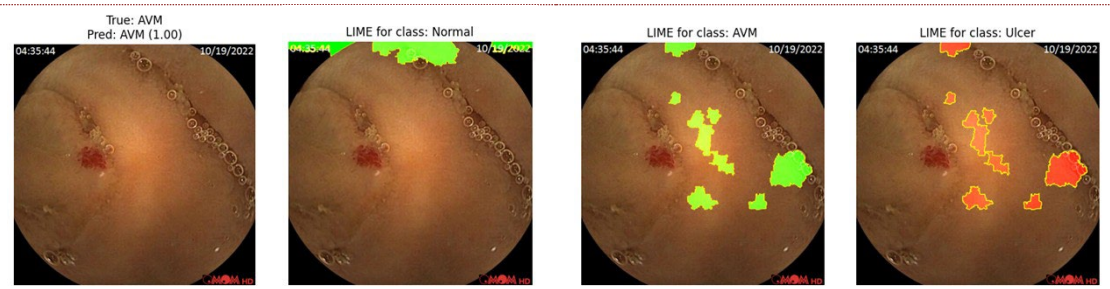
(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 8. Sample Two for the original and three classes: normal, AVM, and ulcers.



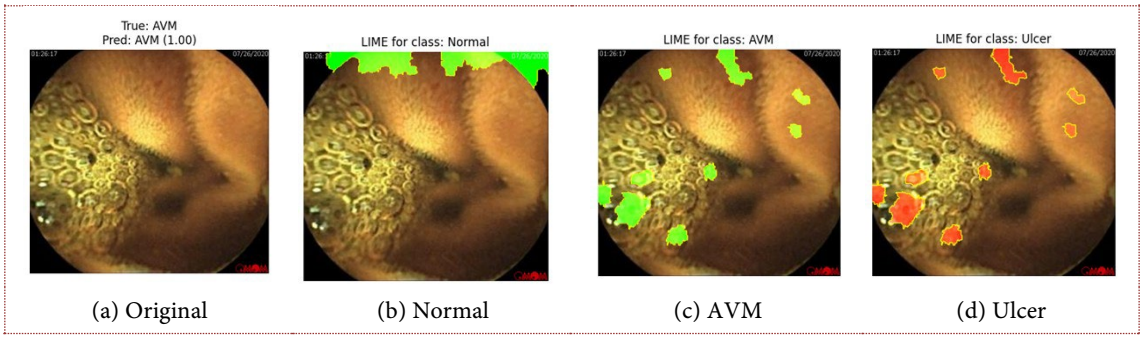
(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 9. Sample Three for the original and three classes: normal, AVM, and ulcer.



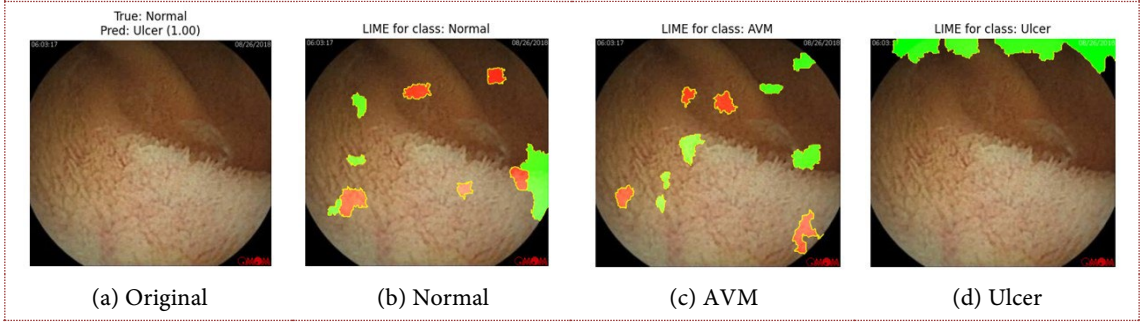
(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 10. Sample Four for the original and three classes: normal, AVM, and ulcer.



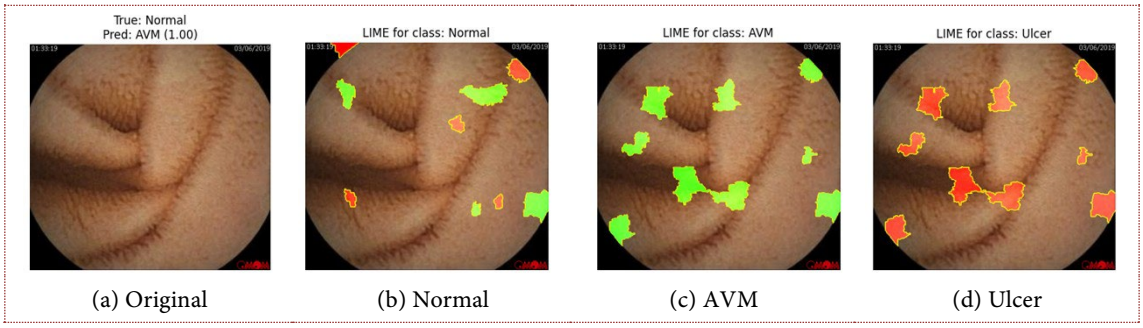
(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 11. Sample Five for the original and three classes: normal, AVM, and ulcer.



(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 12. Sample Six for the original and three classes: normal, AVM, and ulcer.



(a) Original (b) Normal (c) AVM (d) Ulcer

Figure 13. Sample Seven for the original and three classes: normal, AVM, and ulcer.

These XAI insights serve dual purposes: 1) they provide clinicians with transparent decision support, enabling verification of model reasoning; and 2) they guide model improvement by identifying failure modes and highlighting features requiring enhanced representation. This addresses a critical gap noted in the original KAUHC dataset paper [21], which did not incorporate explainability techniques despite their importance for clinical adoption.

4.5. Comparison with State-of-the-Art

Table 4 compares our proposed framework with existing approaches for WCE analysis and related medical imaging studies, with particular emphasis on the KAUHC dataset baselines. Our framework achieves the highest reported F1-score (0.97) for multiclass lesion detection on the KAUHC dataset, substantially outperforming the traditional machine learning baselines reported in [21].

Table 4. Comparison with State-of-the-Art approaches.

Study	Approach	Classes	Dataset	Key Metrics	XAI Integration
[39]	AlexNet	Binary (Bleeding vs. Normal)	Private	Acc: 96.2%	No
[12]	Conv-ViT Fusion	Binary/Multiclass	Kvasir-Capsule	F1: 0.92	No
[40]	Focal Modulation CNN	GI Anomalies	Private	F1: 0.73	No
[21]	Traditional ML (DT, RF, KNN, LR, NB, SVM)	3-class (KAUHC)	KAUHC (3301 images)	Acc: 57% - 98% (multiclass), F1: 0.53 - 0.98	No
Proposed Framework	Fine-tuned VGG16 + XAI	3-class (Normal, AVM, Ulcer)	KAUHC (3301 images)	F1: 0.97, Recall: 0.97, Precision: 0.97	Yes (LIME)

Specifically, while the original KAUHC study achieved multiclass accuracies ranging from 57.33% (NB) to 98.94% (DT) with corresponding F1-scores from 0.53 to 0.99, their results showed high variability across classifiers and required SMOT balancing for optimal performance. In contrast, our deep learning approach delivers consistent, high-performance metrics (F1: 0.94 - 0.98 across classes) without the performance degradation observed in some traditional ML classifiers (e.g., NB at 57% accuracy).

Compared to other deep learning approaches for WCE, our work uniquely combines three critical elements: 1) robust multiclass classification covering clinically relevant lesion types (AVM and Ulcer beyond binary bleeding classification), 2) systematic data-centric optimization addressing class imbalance and imaging artifacts, and 3) integrated XAI for clinical interpretability. All are validated on a region-specific dataset that addresses the scarcity of annotated endoscopic imaging resources in the Middle East [21].

4.6. Limitations and Generalizability

Regarding model generalizability, our framework incorporates multiple strategies to mitigate overfitting and promote robustness: 1) a strict 70-15-15 stratified split with a heldout test set, ensuring unbiased evaluation; 2) artifact-aware preprocessing (circular ROI extraction and inpainting) to remove dataset-specific spurious features; 3) augmentation-based oversampling combined with architectural regularization (global average pooling and progressive dimensionality reduction) to learn discriminative yet generalizable features; and 4) transfer learning from ImageNet, preserving generic visual primitives while fine-tuning only the last six layers. The close alignment between training and validation curves (**Figure 5**) and the clinically meaningful LIME explanations (**Figures 7-13**) further indicate that the model captures genuine pathological patterns rather than memorizing dataset artifacts.

A notable limitation of the current study is the frame-level data split, which may not fully account for potential dependencies between frames from the same WCE study. While this approach follows the methodology of the original KAUHC dataset paper [21] and is standard in frame-based WCE classification, future inves-

tigations should employ patient-wise or study-wise cross-validation when such metadata becomes available. Although the current study uses a single-center dataset, the internal diversity (86 studies, four-year collection) and the high inter-rater agreement in annotations provide a solid foundation. Future work will include multicenter external validation and temporal modeling to fully establish generalizability across different populations and capsule systems.

5. Conclusions

This paper presents a robust, interpretable deep learning framework for automated multiclass lesion detection in Wireless Capsule Endoscopy (WCE) using the KAUHC dataset (a novel repository of 3301 small-bowel images from Saudi Arabia comprising Normal, AVM, and Ulcer categories). Our fine-tuned VGG16 architecture with 6 unfrozen layers, enhanced by artifact-aware preprocessing (circular ROI extraction and inpainting) and augmentation-based oversampling, achieved superior diagnostic performance: precision of 0.97, recall of 0.97, and F1-score of 0.97. This substantially outperformed InceptionV3 (F1: 0.94), as well as traditional machine learning baselines from the original KAUHC study.

Integration of LIME-based explainability revealed that model decisions align with clinically relevant features (vascular clusters for AVM and mucosal disruption for ulcers), addressing the "black-box" concern hindering clinical adoption. While limitations include single-center data and frame-based analysis without temporal context, our framework uniquely combines multiclass classification, robust preprocessing, and integrated XAI.

Besides its clinical contribution, this work introduces a computing-oriented framework for addressing challenges such as model design, data processing, and system-level integration. It offers an end-to-end computational pipeline with deep feature learning, data preprocessing, explanation modules, and optimization strategies within a unified architecture. Furthermore, this work contributes to the design of an intelligent medical image analysis computing system, where key design constraints include robustness, scalability, and algorithmic efficiency. Such a system offers future extensions for real-time deployments, edge-based medical AI applications, and embedded processing.

Future work will pursue multi-center validation, temporal modeling, and prospective clinical studies to evaluate the real-world impact on diagnostic efficiency. This work bridges experimental AI and clinical application, offering a reliable, interpretable tool for computer-assisted gastrointestinal diagnostics.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cao, Q., Deng, R., Pan, Y., Liu, R., Chen, Y., Gong, G., *et al.* (2024) Robotic Wireless Capsule Endoscopy: Recent Advances and Upcoming Technologies. *Nature Communications*, **15**, Article No. 1597. https://doi.org/10.1038/s41467-024-49019-0

[2] Redondo-Cerezo, E., Sánchez-Capilla, A.D., De La Torre-Rubio, P. and De Teresa, J. (2014) Wireless Capsule Endoscopy: Perspectives beyond Gastrointestinal Bleeding. *World Journal of Gastroenterology*, **20**, 15664-15673. https://doi.org/10.3748/wjg.v20.i42.15664

[3] Mehedi, I.M., Rao, K.P., Alotaibi, F.M. and Alkanfery, H.M. (2023) Intelligent Wireless Capsule Endoscopy for the Diagnosis of Gastrointestinal Diseases. *Diagnostics*, **13**, Article 1445. https://doi.org/10.3390/diagnostics13081445

[4] Giordano, A., Escapa, M., Urpí-Ferreruela, M., Casanova, G., Fernández-Esparrach, G., Ginès, À., *et al.* (2023) Diagnostic Accuracy of Artificial Intelligence-Aided Capsule Endoscopy (TOP100) in Overt Small Bowel Bleeding. *Surgical Endoscopy*, **37**, 7658-7666. https://doi.org/10.1007/s00464-023-10273-w

[5] Lei, I.I., Thorndal, C., Manzoor, M.S., Parsons, N., Noble, C., Huhulea, C., *et al.* (2024) The Diagnostic Accuracy of Colon Capsule Endoscopy in Inflammatory Bowel Disease—A Systematic Review and Meta-Analysis. *Diagnostics*, **14**, Article 2056. https://doi.org/10.3390/diagnostics14182056

[6] Oukdach, Y., Garbaz, A., Kerkaou, Z., El Ansari, M., Koutti, L., Papachrysos, N., *et al.* (2025) Vision Transformer Distillation for Enhanced Gastrointestinal Abnormality Recognition in Wireless Capsule Endoscopy Images. *Journal of Medical Imaging*, **12**, Article ID: 014505. https://doi.org/10.1117/1.jmi.12.1.014505

[7] Ito, T., Sekiya, N., Makino, K., Yamaguchi, T., Nagao, M., Kotani, S., *et al.* (2022) Extraction of Non-Diagnosable Images Captured by a Capsule Endoscope and Polyp Detection Using YOLOv5. 2022 *IEEE/SICE International Symposium on System Integration* (*SII*), Narvik, 9-12 January 2022, 742-747. https://doi.org/10.1109/sii52469.2022.9708904

[8] Tan, C.M., Bernstein, M., Raboud, J., Mannino, B. and Tinmouth, J. (2022) Efficiency in the Endoscopy Unit: Can We 'Turn around' Room Turnover? An Observational Quality Improvement Study. *Journal of the Canadian Association of Gastroenterology*, **5**, e57-e64. https://doi.org/10.1093/jcag/gwac005

[9] World Health Organization (2023) Global Burden of Gastrointestinal Diseases. WHO Press. https://cdn.who.int/media/docs/default-source/gho-documents/world-health-statistic-reports/2023/world-health-statistics-2023_20230519_.pdf

[10] General Authority for Statistics, Health Statistics. Kingdom of Saudi Arabia. https://www.stats.gov.sa/en/statistics?index=119025&subindex=123758

[11] ElZemity, A., ElFdaly, M., Abdelfattah, S., Abdelwahab, A., Ramadan, M., Zakzouk, S., *et al.* (2023) A Transformer-Based Deep Learning Architecture for Accurate Intracranial Hemorrhage Detection and Classification. 2023 *International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies* (3*ICT*), Sakheer, 20-21 November 2023, 215-220. https://doi.org/10.1109/3ict60104.2023.10391388

[12] Oukdach, Y., Kerkaou, Z., Ansari, M.E., Koutti, L. and Ouafdi, A.F.E. (2023) ConV-ViT: Feature Fusion-Based Detection of Gastrointestinal Abnormalities Using CNN and ViT in WCE Images. 2023 10*th International Conference on Wireless Networks and Mobile Communications* (*WINCOM*), Istanbul, 26-28 October 2023, 1-6. https://doi.org/10.1109/wincom59760.2023.10322944

[13] Jiang, H., Xu, J., Shi, R., Yang, K., Zhang, D., Gao, M., *et al.* (2020) A Multi-Label Deep Learning Model with Interpretable Grad-Cam for Diabetic Retinopathy Classification. 2020 42*nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (*EMBC*), Montreal, 20-24 July 2020, 1560-1563.

https://doi.org/10.1109/embc44109.2020.9175884

[14] Sindhura, C., Yalavarthy, P.K. and Gorthi, S. (2024) SINO-CT-Fusion-Net: A Lightweight Deep Learning Framework for Detection and Classification of Intracranial Hemorrhages. 2024 *IEEE International Conference on Image Processing* (*ICIP*), Abu Dhabi, 27-30 October 2024, 3179-3185. https://doi.org/10.1109/icip51287.2024.10648215

[15] Karthik, T.S., Hussain, N., Anushkannan, N.K., Pinnamaneni, R., Vijayakrishna Rapaka, E. and Das, S. (2022) Automated Intracranial Haemorrhage Detection and Classification Using Rider Optimization with Deep Learning Model. 2022 *International Conference on Automation, Computing and Renewable Systems* (*ICACRS*), Pudukkottai, 13-15 December 2022, 588-594. https://doi.org/10.1109/icacrs55517.2022.10029294

[16] Zhang, Q. (2023) The Study of Deep Learning in Brain Tumor and Intracranial Hemorrhage Detection. 2023 *IEEE International Conference on Sensors, Electronics and Computer Engineering* (*ICSECE*), Jinzhou, 18-20 August 2023, 1614-1617. https://doi.org/10.1109/icsece58870.2023.10263299

[17] Saleem, M.A., Javeed, A., Akarathanawat, W., Chutinet, A., Suwanwela, N.C., Asdornwised, W., *et al.* (2024) Innovations in Stroke Identification: A Machine Learning-Based Diagnostic Model Using Neuroimages. *IEEE Access*, **12**, 35754-35764. https://doi.org/10.1109/access.2024.3369673

[18] Bhole, C., Mhamane, A., Pawar, R., Machagar, A. and Shetty, C. (2024) Automated AI-Driven Detection of Brain Infarct and Hemorrhage via CT Scan and MRI Imaging. 2024 *Second International Conference on Advanced Computing & Communication Technologies* (*ICACCTech)*, Sonipat, 16-17 November 2024, 842-849. https://doi.org/10.1109/icacctech65084.2024.00138

[19] Jabbar, A., Liaqat, H.B., Akram, A., Sana, M.U., Azpíroz, I.D., Diez, I.D.L.T., *et al.* (2024) A Lesion-Based Diabetic Retinopathy Detection through Hybrid Deep Learning Model. *IEEE Access*, **12**, 40019-40036. https://doi.org/10.1109/access.2024.3373467

[20] Arunasakthi, K., Suriya, S., Amrudan, T.S.S., Shabarrish, P., Ragul, A.P.S. and Akilandeswari, V. (2025) EfficientNet-Based Deep Learning Framework for Early Detection of Diabetic Retinopathy in Retinal Fundus Images. 2025 5*th International Conference on Pervasive Computing and Social Networking* (*ICPCSN*), Salem, 14-16 May 2025, 1173-1179. https://doi.org/10.1109/icpcsn65854.2025.11035600

[21] Ghandorh, H., Bali, H.H., Yafooz, W.M.S., Boulila, W. and Alsahafi, M. (2024) King Abdulaziz University Hospital Capsule Dataset: A Novel Small-Bowel Endoscopic Image Repository from Saudi Arabia. *Data in Brief*, **57**, Article ID: 111093. https://doi.org/10.1016/j.dib.2024.111093

[22] Liu, X., Zhu, S., Yang, W. and Zhou, S. (2021) Few-Shot Learning for Auromatic Intracranial Hematoma Segmentation. 2021 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Houston, 9-12 December 2021, 3400-3407. https://doi.org/10.1109/bibm52615.2021.9669426

[23] Babu, P.P.S. and Brindha, T. (2024) CT Scans Reimagined: Early Intracranial Hemorrhage Detection with AI. 2024 7*th International Conference on Circuit Power and Computing Technologies* (*ICCPCT*), Kollam, 8-9 August 2024, 138-143. https://doi.org/10.1109/iccpct61902.2024.10672960

[24] Chakraborty, P., Bandyopadhyay, A., Misra, M., Gupta, P., Sardar, T.H. and Pandey, B. (2024) Automated Detection of Intracranial Hemorrhage Using Convolutional Neural Networks. 2024 *IEEE AITU*: *Digital Generation*, Astana, 3-4 April 2024, 20-26. https://doi.org/10.1109/ieeeconf61558.2024.10585483

[25] Khan, S., Ali, H. and Shah, Z. (2023) Brain Hemorrhage Detection Using Improved AlexNet with Inception-v4. 2023 *IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things* (*AIBThings*), Mount Pleasant, 16-17 September 2023, 1-5. https://doi.org/10.1109/aibthings58340.2023.10292461

[26] Pandiaraja, P., Teja, N.S.M., Reddy, N.J.S.R.K. and Reddy, T.P.K. (2025) Analysis of Biomedical Images through Deep Learning Model for the Identification of Diseases. 2025 *Third International Conference on Augmented Intelligence and Sustainable Systems* (*ICAISS*), Trichy, 21-23 May 2025, 749-755. https://doi.org/10.1109/icaiss61471.2025.11042076

[27] Lewick, T., Kumar, M., Hong, R. and Wu, W. (2020) Intracranial Hemorrhage Detection in CT Scans Using Deep Learning. 2020 *IEEE Sixth International Conference on Big Data Computing Service and Applications* (*BigDataService*), Oxford, 3-6 August 2020, 169-172. https://doi.org/10.1109/bigdataservice49289.2020.00033

[28] Al-masni, M.A., Kim, W., Kim, E.Y., Noh, Y. and Kim, D. (2020) A Two Cascaded Network Integrating Regional-Based YOLO and 3D-CNN for Cerebral Microbleeds Detection. 2020 42*nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (*EMBC*), Montreal, 20-24 July 2020, 1055-1058. https://doi.org/10.1109/embc44109.2020.9176073

[29] Sengupta, J., Alzbutas, R. and Kl, H. (2022) An Empirical Analysis on Detection and Recognition of Intra-Cranial Hemorrhage (ICH) Using 3D Computed Tomography (CT) Images. 2022 *IEEE* 2*nd Mysore Sub Section International Conference* (*MysuruCon*), Mysuru, 16-17 October 2022, 1-5. https://doi.org/10.1109/mysurucon55714.2022.9972614

[30] Wu, K.H., Zeng, K., Shalaginov, M.Y. and Zeng, T.H. (2024) Brain Hemorrhage CT Image Detection and Classification Using Deep Learning Methods. 2024 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Lisbon, 3-6 December 2024, 5322-5326.

[31] Shakina Princess, S. and Vignesh, R. (2025) Automating Brain Hemorrhage Detection through Deep Learning Approaches. 2025 *International Conference on Visual Analytics and Data Visualization* (*ICVADV*), Tirunelveli, 4-6 March 2025, 915-921. https://doi.org/10.1109/icvadv63329.2025.10961684

[32] Anser Pasha, C.A., Bhargav, K.S., Akki, R., Sanath, R. and Vismitha, V. (2024) Automatic Detection of Diabetic Retinopathy Using CNN and SVM. 2024 *Second International Conference on Advances in Information Technology* (*ICAIT*), Chikkamagaluru, 24-27 July 2024, 1-4. https://doi.org/10.1109/icait61638.2024.10690517

[33] Karumanchi, D.S., Iwin Thanakumar Joseph, S., Karunakaran, V. and Velliangiri, S. (2021) Investigation of Intelligent Approaches in Micro Aneurysms Detection. 2021 *International Conference on Computer Communication and Informatics* (*ICCCI*), Coimbatore, 27-29 January 2021, 1-6. https://doi.org/10.1109/iccci50826.2021.9402448

[34] Saravanan, V., Sankaradass, V., Shanmathi, M., Bhimavarapu, J.P., Deivakani, M. and Ramasamy, S. (2023) An Early Detection of Ovarian Cancer and the Accurate Spreading Range in Human Body by Using Deep Medical Learning Model. 2023 *International Conference on Disruptive Technologies* (*ICDT*), Greater Noida, 11-12 May 2023, 68-72. https://doi.org/10.1109/icdt57929.2023.10151103

[35] Ikram, A. and Imran, A. (2023) Review on Detection and Classification of Diabetic Retinopathy Using Fundus Images. 2023 17*th International Conference on Open Source Systems and Technologies* (*ICOSST*), Lahore, 20-21 December 2023, 1-5. https://doi.org/10.1109/icosst60641.2023.10414197

[36] Shah, K., Gupta, M., Mishra, S., Oberoi, S., Agrawal, R. and Jogalekar, U.A. (2025)

Leveraging CNNs for Early Detection and Staging of Diabetic Retinopathy. 2025 3*rd International Conference on Disruptive Technologies* (*ICDT*), Greater Noida, 7-8 March 2025, 665-670. https://doi.org/10.1109/icdt63985.2025.10986718

[37] Larasati, S.S.A., Shalsadilla, S.R. and Utaminingrum, F. (2025) Enhancing Stroke Detection in Brain CT Images with Residual Attention Networks. *IET Conference Proceedings*, **2024**, 574-579. https://doi.org/10.1049/icp.2025.0312

[38] Rooshita, K., Vyshnavi, N.V., Chowdary, N.R.S., Singh, T. and Patra, P. (2024) Deep Learning for Precise Brain Hemorrhage Detection from CT Scans: A Web-Based Approach. 2024 15*th International Conference on Computing Communication and Networking Technologies* (*ICCCNT*), Kamand, 24-28 June 2024, 1-6. https://doi.org/10.1109/icccnt61001.2024.10724023

[39] Sunitha, S. and Sujatha, S.S. (2021) An Improved Bleeding Detection Method for Wireless Capsule Endoscopy (WCE) Images Based on AlexNet. 2021 3*rd International Conference on Signal Processing and Communication* (*ICPSC*), Coimbatore, 13-14 May 2021, 11-15. https://doi.org/10.1109/icspc51351.2021.9451699

[40] Saji, A.J., Pushpan, G., Rajan, R., Nourin Fathima, N. and Anoop, M. (2024) Advanced Endoscopic Precision: Focal Modulation Guided CNNs for Targeted Diagnosis of GI Anomalies. 2024 7*th International Conference on Circuit Power and Computing Technologies* (*ICCPCT*), Kollam, 8-9 August 2024, 341-346. https://doi.org/10.1109/iccpct61902.2024.10673322

[41] Rane, H. and Warhade, K. (2021) A Survey on Deep Learning for Intracranial Hemorrhage Detection. 2021 *International Conference on Emerging Smart Computing and Informatics* (*ESCI*), Pune, 5-7 March 2021, 38-42. https://doi.org/10.1109/esci50559.2021.9397009

[42] Sakthimohan, M., Sabarna, R.C., Varshini, S., Prathap, C., Gopi Krishnan, P. and Elizabeth Rani, G. (2025) Innovative Machine Learning Approaches for Accurate Fibroid Detection in Uterine Ultrasound Imaging. 2025 *International Conference on Electronics and Renewable Systems* (*ICEARS*), Tuticorin 11-13 February 2025, 1880-1885. https://doi.org/10.1109/icears64219.2025.10941032

[43] Raguraman, P., Kumaresan, M. and Ramesh, S. (2024) A Comprehensive Survey of Ai-Driven Biomedical Image Processing for Intracerebral Hemorrhage Detection and Classification: Current Trends, Challenges, and Future Directions. 2024 8*th International Conference on Electronics*, *Communication and Aerospace Technology* (*ICECA*), Coimbatore, 6-8 November 2024, 1649-1653. https://doi.org/10.1109/iceca63461.2024.10801000

[44] Tursynova, A., Omarov, B., Shuketayeva, K. and Smagul, M. (2021) Artificial Intelligence in Stroke Imaging. 2021 11*th International Conference on Cloud Computing*, *Data Science & Engineering* (*Confluence*), Noida, 28-29 January 2021, 41-45. https://doi.org/10.1109/confluence51648.2021.9377102

[45] Tanwar, S., Choudhary, N., Chaudhary, V., Dahiya, A., Kaushik, P. and Rathore, R. (2024) Advancing Healthcare: CNN-Based Brain Hemorrhage Detection in Intelligent Environments. 2023 *International Conference on Smart Devices* (*ICSD*), Dehradun, 2-3 May 2024, 1-6. https://doi.org/10.1109/icsd60021.2024.10751290

[46] Balipa, M., P Kundapur, P. and Adithya (2023) Intracranial Hemorrhage Detection Using Deep Learning and Optimization Techniques. 2023 *Third International Conference on Ubiquitous Computing and Intelligent Information Systems* (*ICUIS*), Gobichettipalayam, 1-2 September 2023, 483-489. https://doi.org/10.1109/icuis60567.2023.00086

[47] de Vries, B.M., Zwezerijnen, G.J.C., Burchell, G.L., van Velden, F.H.P., Menke-van

der Houven van Oordt, C.W. and Boellaard, R. (2023) Explainable Artificial Intelligence (XAI) in Radiology and Nuclear Medicine: A Literature Review. *Frontiers in Medicine*, **10**, Article 1180773. https://doi.org/10.3389/fmed.2023.1180773

[48] Lundberg, S. and Lee, S. (2017) A Unified Approach to Interpreting Model Predictions. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 4768-4777.

[49] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778