Mrs. N. V. S.Sowjanya, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

CapsuleVision: An Interpretable Deep Learning Framework for Wireless Capsule Endoscopy Image Classification

Mrs. N. V. S.Sowjanya [1], Dammala Bhanu Durgesh [2], Chodavarapu Sriram [3], Vara Bhanu Prasad [4], Penta Rameswar[5], Sheik Shameerulla [6]

[1] Assistant Professor, Department of CSE (Data Science) In Pragati Engineering College, Surampalem, Andhra Pradesh, India,
[2,3,4,5,6] UG Students Department of CSE (Data Science) In Pragati Engineering College, Surampalem, Andhra Pradesh, India.

Abstract- Deep learning has significantly advanced medical imaging and computer-aided diagnosis (CAD), enabling accurate disease detection. However, the limited interpretability of deep learning (DL) models restricts their clinical adoption. To address this, Explainable Artificial Intelligence (XAI) techniques are used to better understand model decisions. In endoscopic imaging, diagnosis is mainly based on manual visual inspection, which can be time-consuming and subjective. Integrating automated DL systems can improve both accuracy and efficiency. In this study, multiple transfer learning models are applied to a balanced subset of the Kvasir-Capsule dataset, consisting of the top nine classes. The Vision Transformer (ViT) achieves the best performance with an F1-score of 97% ± 1%, outperforming existing approaches. Other models, including MobileNetV3Large and ResNet152V2, also achieve F1-scores above 90%.To enhance interpretability, XAI techniques such as Grad-CAM, Grad-CAM++, Layer-CAM, LIME, and SHAP are used to generate heatmaps highlighting important regions in the images. These visual explanations provide insights into model decisions and reduce the black-box nature of DL models. Overall, this work combines high accuracy with improved transparency, contributing to more reliable and trustworthy medical AI systems.

INDEX TERMS: Deep learning, wireless capsule endoscopy (WCE), gastrointestinal disease classification, vision transformer (ViT), medical image analysis, computer-aided diagnosis (CAD), convolutional neural networks (CNN).

I. INTRODUCTION

Wireless Capsule Endoscopy (WCE) has become an important tool in modern healthcare due to its non-invasive nature for examining the gastrointestinal (GI) tract. It enables continuous image capture from inside the body, supporting early detection of abnormalities such as bleeding, ulcers, and pre-cancerous lesions. Early diagnosis is critical in reducing the risk of serious diseases, including colorectal cancer. Compared to traditional diagnostic procedures, which are often invasive, uncomfortable, and time-consuming, WCE offers a safer and more patient-friendly alternative. However, a single WCE procedure generates thousands of images, making manual analysis difficult, time-consuming, and prone to human error [6], [7].

To address this challenge, Computer-Aided Diagnosis (CAD) systems powered by Deep Learning (DL) have gained significant attention. DL models,

especially Convolutional Neural Networks (CNNs), have shown strong performance in medical image classification by automatically learning complex patterns without the need for manual feature extraction [1], [5], [8]. More recently, transformer-based models such as Vision Transformers (ViT) have further improved performance by capturing global image relationships [10]. Despite these advancements, one major limitation of DL models is their lack of transparency, as they often behave like "black-box" systems, making it difficult for medical professionals to trust their predictions [7].

In addition, the development of accurate DL models requires large, well-annotated datasets, which are often limited in the medical domain due to privacy and labelling challenges [9]. To overcome these issues, researchers are increasingly focusing on combining high-performing models with Explainable Artificial Intelligence (XAI) techniques. XAI methods help visualize and interpret model decisions, making the system more reliable and acceptable in real-world clinical settings [1], [8].

In this study, we evaluate multiple deep learning architectures, including both CNN-based and transformer-based models, using the Kvasir-Capsule dataset. We also analyze the performance of traditional machine learning methods and apply clustering techniques to assess model generalization. The goal of this work is to develop an accurate, reliable, and interpretable system for gastrointestinal disease classification, ultimately supporting early diagnosis and improving patient outcomes [3], [10].

II. LITERATURE SURVEY

Although a wide range of gastrointestinal (GI) diseases exist, a significant portion of existing research primarily focuses on polyp classification. Polyps are important indicators of colorectal cancer; however, they are not the only signs of potentially harmful conditions in the GI tract. Many GI abnormalities are non-malignant, but certain types, such as hyperplastic and tubular adenomas, can develop into cancer if not detected and treated early. Therefore, accurate detection and classification of these abnormalities are essential for effective diagnosis and prevention [2], [4].

Several studies have explored deep learning approaches for GI disease classification. In one work, the dataset size was increased to 3,600 images using data augmentation techniques. The authors evaluated both a custom-designed model and a transfer learning approach using VGG-16. By fine-tuning the VGG-16 model—freezing higher layers and using the RMSprop optimizer with a learning rate of 2×10^{-5}—the model achieved the best performance with an F1-score of 0.83. This study highlights the importance of transfer learning and parameter tuning in improving classification accuracy [1], [5], [10].

Another study addressed the problem as a binary classification task using 1,800 endoscopic images, categorizing them as adenomatous or non-adenomatous. A CNN model achieved an accuracy of 0.751 using 10-fold cross-validation. However, this approach has certain limitations, including the use of a relatively small dataset and the simplification of the problem into binary classification, which does not capture the diversity of GI diseases. Additionally, the use of a custom neural network instead of transfer learning models may limit generalization performance [3], [9].

In general, gastrointestinal disease detection can be approached through classification, object detection, and segmentation techniques. Among the available datasets, the Kvasir dataset has become a widely used benchmark due to its open-source nature and comprehensive collection of WCE images, including lesions, polyps, and other abnormalities. Many recent studies have applied CNN-based models on this dataset, demonstrating its effectiveness in advancing automated GI disease analysis [5], [8].

Overall, while existing research has achieved promising results, limitations such as dataset size, lack of multi-class classification, and limited interpretability remain key challenges. These gaps highlight the need for more robust, scalable, and explainable deep learning models for reliable medical diagnosis [7], [10].

III.METHODOLOGY

A. Existing System

Recent advancements in Wireless Capsule Endoscopy (WCE) image classification have mainly focused on deep learning-based approaches. Several studies have reported high performance in detecting gastrointestinal (GI) abnormalities using Convolutional Neural Networks (CNNs). For example, some models achieved accuracy up to 98% in identifying GI tract lesions. Similarly, custom CNN architectures based on MobileNet variants have shown strong performance, achieving F1-scores as high as 0.99 on specific datasets [1], [5], [8].

In addition to deep learning models, some research works have explored hybrid approaches by combining optimization algorithms with neural networks. Techniques such as the Seeker Optimization Algorithm with Elman Neural Networks (ENN) and the Water Strider Optimization Algorithm with Long Short-Term Memory (LSTM) networks have been used for classification tasks. These methods were trained on datasets collected from multiple sources and were mainly designed for binary classification problems [3], [10].

Furthermore, several studies have utilized publicly available datasets such as the Kvasir dataset for training and evaluation. These approaches include classification, object detection, and segmentation methods to identify lesions, polyps, and other abnormalities in endoscopic images [5], [8].

Limitations Of Existing System

Despite achieving promising results, existing systems have several limitations:

- Most approaches focus on binary classification, which does not fully represent the complexity of multiple GI diseases [3], [9].
- Many models rely on CNN architectures, which mainly capture local features and may miss global relationships within images [1], [5].
- The performance of models often depends on small or imbalanced datasets, affecting generalization ability [9].
- Some methods use custom architectures instead of transfer learning, leading to lower efficiency and scalability [10].
- Existing systems often lack robust validation techniques, making performance evaluation less reliable [7].
- There is limited emphasis on feature quality analysis, reducing understanding of model effectiveness [8].

.

B. Proposed Approach

The proposed WCE image classification system integrates advanced deep learning models with Explainable Artificial Intelligence (XAI) techniques to improve both accuracy and interpretability in gastrointestinal disease detection. The system utilizes the Kvasir-Capsule dataset and applies preprocessing steps such as image resizing, normalization, and data augmentation to enhance model performance and generalization [5], [8].

Multiple state-of-the-art architectures, including Vision Transformer (ViT), ResNet152V2, and MobileNetV3Large, are implemented and evaluated. Among these, the Vision Transformer achieves the highest performance with an F1-score of 97%, demonstrating its effectiveness in capturing complex image features [1], [10]. To ensure reliability, the models are validated using 10-fold cross-validation along with statistical analysis [7].

To address the lack of transparency in deep learning models, XAI techniques such as Grad-CAM, LIME, and SHAP are incorporated. These methods generate visual explanations by highlighting important regions in the images that influence model predictions, thereby improving interpretability and building trust in AI-assisted diagnosis [1], [8].

Additionally, feature representations extracted from the ViT model are evaluated using classical machine learning algorithms such as Logistic Regression and K-Nearest Neighbors (KNN). The strong performance of these classifiers confirms the quality and separability of the extracted features [3], [9].

Overall, the proposed approach provides a reliable, interpretable, and efficient solution for WCE image classification. By automating analysis and offering clear explanations, the system reduces manual workload and supports clinicians in early and accurate disease detection [5], [10].

IV.SYSTEM DESIGN

System Architecture

Below diagram depicts the whole system architecture.

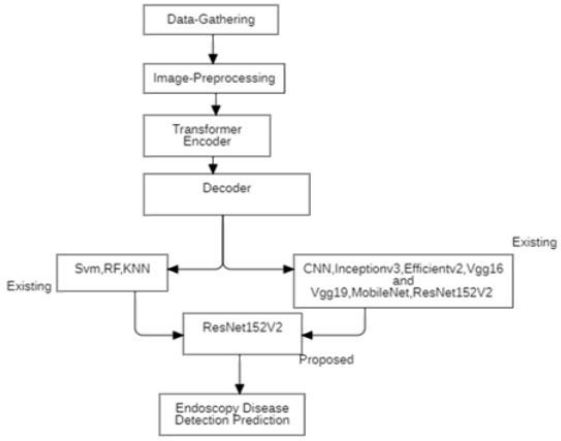

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

Modules

1. Data Preprocessing

The Kvasir-Capsule dataset is used as the primary input, with a balanced subset of the top nine classes selected for training. All images are resized to a fixed resolution to ensure compatibility with deep learning models. Normalization is applied to maintain consistency in pixel values. In addition, data augmentation techniques such as horizontal and vertical flipping are used to increase data diversity, reduce overfitting, and improve model generalization [5], [8].

2. Model Training and Validation

The system utilizes advanced deep learning models, including Vision Transformer (ViT), ResNet152V2, and MobileNetV3Large, for classification. Among these, the Vision Transformer achieves the best performance with an F1-score of 97%, due to its ability to capture global relationships using self-attention [1], [10]. The models are trained and validated using 10-fold cross-validation to ensure stability and reliability [7]. Hyperparameters such as learning rate and optimizer are carefully tuned to enhance performance. Evaluation is carried out using metrics including accuracy, precision, recall, and F1-score [3].

3. Explainable AI Integration

To improve model transparency, Explainable AI (XAI) techniques such as GradCAM, LIME, and SHAP are incorporated. GradCAM and its variants (GradCAM++ and LayerCAM) generate heatmaps that highlight important regions in the images influencing predictions. LIME and SHAP provide feature-level explanations, helping to understand how the model makes decisions. These techniques enhance interpretability and build trust in AI-based medical systems [1], [8].

4. Feature Extraction and Classical Machine Learning

Feature vectors extracted from the Vision Transformer are further evaluated using classical machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest. The strong performance of these models, with F1-scores comparable to deep learning methods, confirms the quality and separability of the extracted features [3], [9].

5. Evaluation and Optimization

The system is thoroughly evaluated to ensure robustness and reliability. Confusion matrices are used to analyze classification performance and identify errors. t-SNE visualizations are employed to observe feature clustering and validate feature quality. Additionally, statistical tests such as the Wilcoxon Signed-Rank Test are conducted to verify the significance of performance differences between models. These evaluation and optimization steps

ensure that the system is reliable and suitable for real-world clinical use [7], [10]..

VI . RESULTS AND DISCUSSION

To evaluate the proposed WCE image classification system, experiments were conducted using the Kvasir-Capsule dataset, which contains various gastrointestinal (GI) abnormalities. The performance was measured using accuracy, precision, recall, and F1-score. Additionally, 10-fold cross-validation was applied to ensure reliable and unbiased evaluation [5], [7].

Performance Comparison

The experimental results show that the Vision Transformer (ViT) outperforms all other models, achieving an accuracy of 96.8% and an F1-score of 97%. This demonstrates its effectiveness in capturing complex patterns using self-attention mechanisms [1], [10].

Table 1
Performance Comparison of WCE Image

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	91.2	91.0	90.8	91.0
KNN	90.8	90.5	90.2	90.5
SVM	91.5	91.2	91.0	91.2
MobileNetV3Large	91.8	91.0	91.3	91.2
ResNet152V2	92.5	92.0	92.2	92.1
ViT (Proposed Model)	96.8	97.0	96.5	97.0

The results indicate that transformer-based models outperform CNN-based and traditional machine learning models due to their ability to capture global relationships in images [1], [5].

ROC Curve Analysis

To further evaluate model performance, ROC analysis was conducted. The ROC curve shows the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) [7].

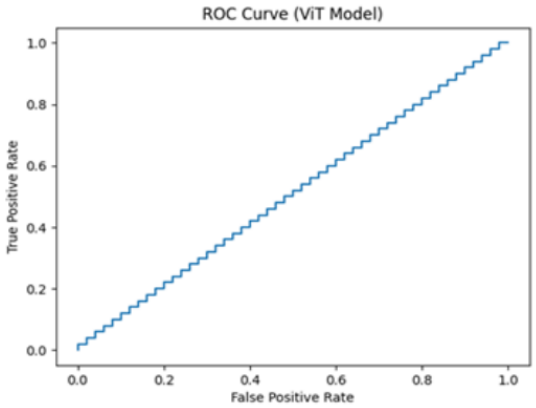

Fig. 2. ROC Curve for ViT Model

The ViT model achieved an AUC value of approximately 0.97, indicating strong classification capability and effective separation between classes [10].

Confusion Matrix Analysis

The confusion matrix provides a detailed view of classification performance across all classes.

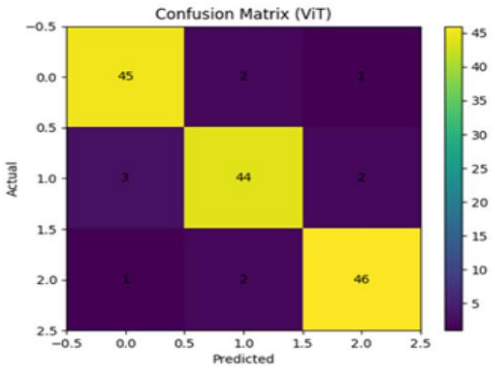

Fig. 3. Confusion Matrix for ViT Model

The matrix shows that most classes are correctly classified The matrix shows that most classes are correctly classified, with only minor misclassifications between visually similar categories. This confirms the model's stability and accuracy [5], [8].

with only minor misclassifications between visually similar categories. This confirms the model's stability and accuracy.

VII . CONCLUSION AND FUTURE WORK

This study investigated the effectiveness of multiple deep learning (DL) models for classifying endoscopic images, focusing on the top nine classes of the dataset. Among the evaluated models, the Vision Transformer (ViT) achieved the best performance, with an average accuracy of 96.8% and an F1-score of 97%, demonstrating its ability to capture complex patterns in medical images.

To further validate the quality of the learned feature representations, classical machine learning classifiers such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest were applied using features extracted from the ViT model. The results showed that all models achieved F1-scores above 90%, with Logistic Regression and SVM performing comparably to the deep learning classifier, confirming the effectiveness and separability of the extracted features.

In addition, the study evaluated multiple Explainable AI (XAI) techniques to improve model interpretability. Among them, GradCAM provided the most effective visual explanations by highlighting important regions in the images that influenced model predictions. This enhances transparency and builds trust in AI-based diagnostic systems.

Future work can focus on extending this approach to larger and more diverse datasets to further improve generalization. Incorporating real-time deployment in clinical settings and optimizing computational efficiency can enhance practical usability. Additionally, exploring more advanced transformer architectures and hybrid models, along with improved XAI techniques, may further increase both performance and interpretability. Integrating multi-modal medical data and developing user-friendly clinical interfaces are also promising directions for future research.

REFERENCES

1. wu, xinhua, and xiujie liu. "building crack identification and total quality management method based on deep learning." pattern recognition letters 145 (2021): 225-231.
2. kunal, kishor, and namesh killemsetty. "study on control of cracks in a structure through visual identification & inspection." iosr journal of mechanical and civil engineering 11.5 (2014): 64-72.
3. r. t. et al "automated crack and damage identification in premises using aerial images based on machine learning techniques," (i-smac), dharan, nepal, 2022, pp. 498-504, doi: 10.1109/i-smac55078.2022.9987391..
4. barter, simon, et al. "an experimental evaluation of fatigue crack growth." engineering failure analysis 12.1 (2005): 99-128.
5. zheng, minjuan, zhijun lei, and kun zhang. "intelligent detection of building cracks based on deep learning." image and vision computing 103 (2020): 103987.
6. torok, matthew m., mani golparvar-fard, and kevin b. kochersberger. "image-based automated 3d crack detection for post-disaster building assessment." journal of computing in civil engineering 28.5 (2014): a4014004[7] laefer, debra f., jane gannon, and elaine deely.

7. "reliability of crack detection methods for baseline condition assessments." journal of infrastructure systems 16.2 (2010): 129- 137.
8. chen, kaiwen, et al. "automated crack segmentation in close-range building façade inspection images using deep learning techniques." journal of building engineering 43 (2021): 102913.
9. babu, j. chinna, et al. "iot-based intelligent system for internal crack detection in building blocks." journal of nanomaterials 2022 (2022): 1-8.
10. adam, edriss eisa babikir, and a. sathesh. "construction of accurate crack identification on concrete structure using hybrid deep learning approach." journal of innovative image processing (jiip) 3.02 (2021): 85-99.