Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2025

Detection and Segmentation of Abnormalities in VCE Images

José Castro, Tiago Costa, Marta Salgado, Antonio Cunha*

Universidade de Trás-os-Montes e Alto Douro, Quinta de Prados, Vila Real 5000-801, Portugal

Abstract

Video Capsule Endoscopy (VCE) is a pivotal technology in modern gastroenterology, offering a non-invasive method to visualize the entire small bowel. However, the clinical application of VCE is hampered by the extensive review time required, as specialists must manually analyze thousands of images from each procedure. This process is not only laborious and costly but also prone to diagnostic errors due to fatigue and the subtle nature of some abnormalities.

While convolutional neural networks have been proposed to automate VCE image analysis, single-backbone models may not fully capture the diversity of lesion types and anatomical variations present in real-world data. Recent work suggests ensemble methods like Mixture of Experts (MoE) can improve visual recognition performance in general image domains. However, to our knowledge, no published study has applied a MoE or Hierarchical MoE (HMoE) architecture to multiclass classification of VCE or even endoscopic images.

This paper explores the potential of more sophisticated ensemble-based deep learning strategies to overcome these limitations. We propose a comparative study evaluating three distinct approaches for the automated classification of VCE images. The objective is to determine whether complex expert-based systems offer tangible benefits over high-performing individual models. Our methodology is centered on the Kvasir-Capsule dataset, focusing on 12 distinct classes of gastrointestinal findings. We first establish a performance baseline by training and evaluating four individual deep learning models: InceptionNeXt, EfficientViT, ConvNeXtV2, and DeiT3. Subsequently, the two top-performing architectures, ConvNeXtV2 and DeiT3, are used as the foundation for two advanced systems. The first is a Mixture of Experts (MoE) model, which employs a gating network to dynamically route images to specialized expert instances of a single base architecture. The second is a Hierarchical Mixture of Experts (HMoE) model, which automatically learns a binary tree structure based on class similarity to create a data-driven classification pathway.

Results

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: author@institute.xxx

Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2025

Keywords: Video Capsule Endoscopy, Deep Learning, Mixture of Experts, Hierarchical Mixture of Experts, Medical Multiclass Classification, Ensemble Models

1. Introduction

Video Capsule Endoscopy (VCE) represents a significant advancement in gastrointestinal tract diagnosis, enabling visualization of previously inaccessible areas through traditional endoscopy. This minimally invasive technology involves ingesting a capsule equipped with a micro camera that captures thousands of images throughout the digestive tract over a period of 8 to 12 hours. However, the clinical utility of VCE is constrained by the time intensive nature that makes it susceptible to errors due to data volume, subtle lesion visibility, inter observer variability, and fatigue.

Deep learning has shown promise in automating classification tasks in medical imaging. Using a single model, however, may not fully capture the variety of visual information present in VCE images, which include multiple lesion types and anatomical variations. One approach to address this is to use a Mixture of Experts architecture, which uses a gating network to select the best model for a given image. A further step could be a Hierarchical Mixture of Experts design, which organizes the models into groups to handle different sub tasks.

This paper explores these different strategies for VCE image classification. We first compare the performance of four individual deep learning models: InceptionNeXt, EfficientViT, ConvNeXtV2, and DeiT3. Following this initial comparison, the two best performing models were selected for further study. Each of these top models was then used separately as the foundation for both a standard Mixture of Experts system and a Hierarchical Mixture of Experts architecture. In these designs, multiple instances of a single top performing model act as the experts, managed by a lightweight gating network.

Our work examines the effectiveness of these distinct approaches for classifying eight types of gastrointestinal tissues and lesions from VCE images. We first evaluate how the individual models perform on their own to identify the top two candidates. We then assess the Mixture of Experts and Hierarchical Mixture of Experts systems built from each of these two selected models. This allows for a direct comparison between a single model's performance and the performance of MoE and HMoE architectures that use that same model as their expert base. The comparison focuses on classification accuracy and other metrics to understand the strengths of each method.

By comparing these different computational approaches, this study aims to identify effective strategies for developing a computer aided tool for gastroenterologists. The goal is to help reduce review time, improve diagnostic consistency, and provide a scalable system for analyzing VCE footage.

2. Methodology

Our experimental methodology, as depicted in Figure 1, is structured to compare three distinct architectural approaches: individual deep learning models, a MoE system, and a HMoE system. The pipeline consists of several key stages, starting with data acquisition and preprocessing, followed by model training and a final evaluation. The preprocessing stage prepares the Kvasir Capsule dataset by performing image resizing, pixel normalization, data augmentation, and partitioning into training, validation, and test sets. The central part of our study is a comparative analysis. We begin by establishing a performance baseline through the training and evaluation of the individual models. Following this, the top-performing models are selected to serve as the foundation for the MoE and HMoE architectures, which are then built and tested separately. This structure allows for a direct comparison between the performance of a single model and the more complex expert-based systems derived from it.

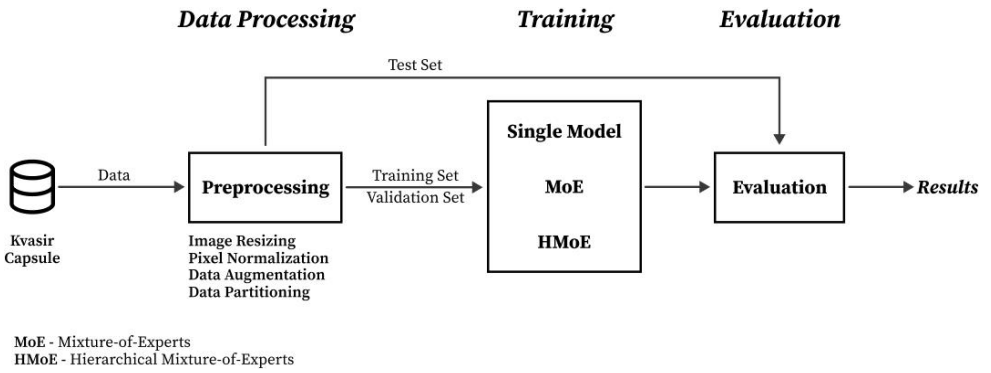

Fig. 1. Diagram illustrating the overall methodological pipeline, from data acquisition and preprocessing to model training and evaluation

2.1. Dataset

This study utilizes the Kvasir Capsule dataset, a public collection of images from gastrointestinal video capsule endoscopy procedures. The full dataset contains approximately 47,000 images across 14 classes, all labeled by experienced gastroenterologists. For our experiments, we used a subset of this data. Two classes, "Ampulla of vater" and "Blood - hematin", were excluded due to having a very low number of images. This resulted in a final dataset composed of 12 classes: Angiectasia, Blood - fresh, Erosion, Erythema, Foreign body, Ileocecal valve, Lymphangiectasia, Normal clean mucosa, Polyp, Pylorus, Reduced mucosal view, and Ulcer.

2.2. Preprocessing

To prepare the dataset for training, a consistent preprocessing pipeline was applied across all experiments. All images were resized to a resolution of 224 by 224 pixels and normalized using the standard ImageNet mean and standard deviation values. The dataset was then partitioned into training, validation, and test subsets using a 70, 15, and 15 percent split, respectively, with a fixed random seed to ensure reproducibility.

Dynamic data augmentation was applied exclusively to the training set to improve model generalization. This included random horizontal and vertical flips, rotations, affine transformations, color jittering, and random autocontrast. To address the notable class imbalance in the dataset, a weighted random sampler was used for the training data loader. This sampler gives more weight to images from minority classes, ensuring each class is represented more equitably during training. The validation and test data loaders did not use augmentation or sampling to provide a consistent and unbiased evaluation.

For the Hierarchical Mixture of Experts model, an additional preprocessing step was performed after the initial data loading. A feature extractor was used to compute an average feature vector, or prototype, for each class from the training set. These prototypes were then used to automatically build a binary tree structure based on class similarity, which dictates the routing logic for the HMoE model during training and inference.

2.3. Training

The training methodology was designed to fairly compare the performance of three distinct approaches. First, to establish a performance baseline, four different models were trained individually: InceptionNeXt, EfficientViT, ConvNeXtV2, and DeiT3. Each model was trained for 10 epochs using the AdamW optimizer with a learning rate of 1e-4 and a cosine annealing learning rate scheduler. Training was guided by a standard cross-entropy loss function.

Based on the results of this initial comparison, the two best-performing model architectures were selected to serve as the foundation for the MoE and HMoE systems. Each of these two architectures was then used to build and train its own separate MoE and HMoE models.

The MoE models were trained in a multi-stage process. Initially, the gating network was trained for several epochs while the expert parameters remained frozen. This allows the gating network to learn effective routing before the experts' weights are updated. Subsequently, the entire system, including the experts and the gating network, was fine-tuned end-to-end with a lower learning rate. The training objective for the MoE models combined a cross-entropy loss with a load-balancing loss to encourage the use of all experts.

The HMoE models also underwent a unique training process. First, a feature extractor based on the selected top-performing architecture was used to compute class prototypes from the training data. These prototypes were then used to automatically construct a binary tree structure that groups visually similar classes together. The complete HMoE model, including the shared feature extractor and the binary decision gates at each node of the tree, was then trained end-to-end. The loss function for the HMoE was a sum of the cross-entropy losses at each decision node in the tree, guiding the model to learn both correct routing and final classification.

2.4. Evaluation

The performance of all three approaches, individual models, MoE, and HMoE, was assessed on the held-out test set. Standard classification metrics, including overall accuracy, precision, recall, and F1-score, were calculated to measure both general and per-class effectiveness. Confusion matrices were also generated for each model to visualize misclassification patterns.

For the individual model approach, the four architectures were first compared to establish a performance baseline and identify the top-performing architectures.

For the MoE systems, evaluation included an expert utilization analysis. This was done by examining the average gating weights assigned to each expert, both across the entire test set and on a per-class basis, to understand how the model learned to specialize. For the HMoE systems, a key evaluation metric was routing accuracy. This measured how effectively the binary gates at each node of the learned tree directed samples to the correct subsequent node or leaf.

Finally, the MoE and HMoE systems were directly compared against each other, as both configurations were trained using the same two expert models. This allowed for a clear assessment of how hierarchical gating strategies compare to flat ensemble routing in terms of classification performance and expert usage.

3. Results

3.1. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3060 GPU with 12GB of VRAM. Model development and training were carried out using the PyTorch deep learning framework, with pre-trained backbone models sourced from the timm library. These included ConvNeXtV2 and DeiT3, both originally trained on the ImageNet dataset.

To ensure reproducibility, a fixed random seed was applied to all stochastic operations, including dataset partitioning and model initialization. Training efficiency was enhanced using PyTorch's automatic mixed precision (AMP), which reduced memory usage and accelerated GPU computation. Additionally, gradient checkpointing was employed, where supported, to further reduce memory overhead.

3.2. Performance Analysis

The initial evaluation showed that all architectures achieved very high performance on the classification task. ConvNeXtV2 emerged as the top performing model, reaching an overall test accuracy of 99.25 percent. EfficientViT and DeiT3 both delivered strong and identical results, each achieving an accuracy of 99.12 percent. InceptionNeXT also demonstrated robust performance with an accuracy of 98.87 percent.

Table 1. Performance evaluation of each model in the Individual Training

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
InceptionNeXt	0.9887	0.86	0.85	0.86
EfficientViT	0.9912	0.86	0.85	0.85
ConvNeXtV2	0.9925	0.58	0.42	0.37
DeiT3	0.9912	0.59	0.65	0.60

Based on this analysis, two models were selected to serve as the architectural foundation for the subsequent MoE and HMoE systems. ConvNeXtV2 was chosen due to its superior overall accuracy. DeiT3 was also selected, not only for its high performance, which was tied for second place, but also for its efficiency as the smallest model among the top contenders.

Tabelas MoE e HMoE

4. Discussion

The integration of Deep Learning into healthcare has demonstrably advanced diagnostic automation and enhanced clinical outcomes. This project specifically addressed the challenging domain of automatic abnormality detection in VCE images, characterized by significant visual variability, uncontrolled illumination, and diverse gastrointestinal morphologies.

Our approach leveraged the DenseNet121 architecture [13], renowned for its densely connected layers that promote efficient feature reuse and robust information propagation throughout the network. This structural advantage proved highly effective in discerning spatial and structural patterns within VCE images, despite their inherent visual complexity. The model was rigorously trained using transfer learning [1], incorporating partial fine-tuning and stratified validation, leading to robust and consistent performance across all eight analysed clinical classes.

To situate our findings within the broader scientific landscape, we compared the performance of our implemented models against established benchmarks from the literature. Recent studies have shown promising results for AI-based polyp detection in colonoscopy [6][7][8], with deep learning approaches achieving high accuracy rates. Real-time polyp detection systems [9] have demonstrated the feasibility of implementing such technologies in clinical practice, while comprehensive reviews [10] have highlighted the potential for deep learning in diagnosis of precancerous lesions in upper gastrointestinal endoscopy. Clinical evaluations of AI-based polyp detection systems [11] have shown encouraging results in multicenter studies, supporting the clinical viability of such approaches.

Analysis of the classification errors indicated that misclassifications predominantly occurred between classes exhibiting subtle visual similarities. Noteworthy instances included confusions between "Normal-z-line" and "Esophagitis," and between "Dyed-resection-margins" and "Dyed-lifted-polyps." These patterns underscore the intrinsic difficulties in differentiating ambiguous visual characteristics, such as minor variations in texture, illumination, or the presence of non-pathological artifacts.

The comprehensive evaluation on an 800-image test set, derived from a stratified 15% split of the original dataset, confirmed DenseNet121's superior performance. The model achieved an overall accuracy of 88%, with corresponding precision and recall values also reaching 88%, thus demonstrating a well-balanced sensitivity and specificity across all classes.

Beyond quantitative metrics, the qualitative visual analysis, utilizing correctly classified and incorrectly classified examples augmented by Grad-CAM activation maps, provided valuable insights. This analysis corroborated the model's capacity to focus on clinically relevant regions for its predictions. However, in instances of misclassification, the Grad-CAM maps occasionally highlighted ambiguous regions or visual artifacts - such as reflections or non-clinical borders - suggesting areas where the model could be misled.

5. Conclusion

This work successfully developed an automated system for detecting gastrointestinal abnormalities in VCE images, leveraging advanced Deep Learning techniques. Addressing the time-consuming and error-prone nature of manual VCE analysis, our approach focused on training CNNs to classify eight distinct mucosa categories.

The methodology involved comprehensive data processing, including careful organization, preprocessing, and strategic data augmentation to enhance model robustness. A comparative analysis of various prominent CNN architectures conducted using transfer learning and partial fine-tuning.

DenseNet121 emerged as the superior model, demonstrating remarkable performance with an overall accuracy, precision, recall and F1-score of 88%. This robust performance underscores its strong capacity for generalization and effective discrimination across diverse abnormality types, even amidst visual variability, inconsistent illumination, and the presence of artifacts inherent in real-world endoscopic examinations. It was particularly noteworthy that DenseNet121 outperformed EfficientNetB7 and ConvNeXtBase, which was somewhat surprising given EfficientNetB7's reputation for state-of-the-art performance and ConvNeXtBase's modern design. This unexpected outcome could be attributed to EfficientNetB7's greater complexity relative to the dataset size, its potential sensitivity to hyperparameter choices, or a sub-optimal fine-tuning strategy for this specific application. Furthermore, the interpretability afforded by Grad-CAM activation maps confirmed the model's ability to focus on clinically relevant image regions, enhancing trust and understanding of its predictions.

The findings highlight the significant potential of AI to streamline and enhance VCE diagnostics, offering a valuable CAD tool. By automating the initial screening process, this system can reduce the workload on healthcare professionals, improve diagnostic accuracy, and ultimately lead to more efficient patient care.

References

[1] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning (ICML), PMLR, pp. 6105–6114, 2019.

[3] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.

[4] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," arXiv preprint arXiv:1712.04621, 2017.

[5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[6] G. Urban, et al., "Deep learning for identifying and locating colorectal polyps," Gastroenterology, vol. 154, no. 6, pp. 1632–1639, 2018.

[7] I. Barua, et al., "Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis" Gastrointestinal Endoscopy, vol. 89, no. 2, pp. 255–264, 2021.

[8] K. Keshtkar, et al., "A systematic review and meta-analysis of deep learning for detection of polyps in colonoscopy," Gastrointestinal Endoscopy, vol. 91, no. 6, pp. 1304–1313, 2023.

[9] D. Jha, et al., "Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning" Endoscopy, vol. 53, no. 1, pp. 45–56, 2021.

[10] Z. Wang, et al., "Deep learning for diagnosis of precancerous lesions in upper gastrointestinal endoscopy: a review," World Journal of Gastroenterology, vol. 27, no. 20, pp. 2531–2548, 2021.

[11] S. Y. Quan, et al., "Clinical evaluation of a real-time artificial intelligence-based polyp detection system: a US multi-center pilot study," Clinical Gastroenterology and Hepatology, vol. 20, no. 4, pp. 883–891, 2022.

[12] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," CVPR 2016.

[13] G. Huang et al., "Densely Connected Convolutional Networks," CVPR 2017.

[14] Z. Liu et al., "A ConvNet for the 2020s," CVPR 2022.

[15] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510-4520.