An Interpretable Deep Learning Framework for Gastrointestinal Abnormality Detection in Capsule Endoscopy

Tejashwini K[1], Aravind Naik[2]
[1]PG Student, Department of Computer Science and Engineering, Srinivas Institute of Technology (SIT), Valachil, Mangaluru,574143, Karnataka, India
[2]Professor, Department of Computer Science and Engineering, Srinivas Institute of Technology (SIT), Valachil, Mangaluru,574143, Karnataka, India.
doi.org/10.64643/IJIRTV12I12-206800-459

***Abstract*—Explainable Artificial Intelligence (XAI) has gained significance in modern medicine, especially in medical imaging applications where transparency and credibility are critical considerations. Gastrointestinal diagnosis using capsule endoscopy yields a massive volume of images; therefore, an automated analysis process would be extremely helpful in this field. This paper proposes a deep-learning-based XAI model for detection of gastrointestinal abnormalities in capsule endoscopy images. The suggested model employs the technique of knowledge distillation along with explainability methods to ensure high prediction accuracy and transparency in decision making. The proposed framework uses a teacher-student network based on ResNets to distill the knowledge of a larger, more complicated teacher model into a smaller, interpretable student model. Grad-CAM is used to make this model interpretable and provides information about the parts of the image that have the greatest impact on the model's prediction. This can help the physician understand the model's decisions. The performance of the suggested model is evaluated using the Capsule Vision 2024 challenge dataset. It comprises different categories of gastrointestinal abnormalities and shows a significant imbalance between classes. The experiment reveals that the suggested distillation model outperforms other baseline approaches and achieves an accuracy of 96.52% and weighted F1-score of 96.37%. Grad-CAM also proves that the model focuses on meaningful regions consistently.**

***Index Terms*—Capsule Endoscopy, Explainable AI, Knowledge Distillation, Grad-CAM, ResNet, Medical Imaging Technology, Deep Learning, GI Anomalies, Class Imbalance, CNN understanding**

I. INTRODUCTION

The introduction of AI to medical diagnostics over recent years has been revolutionary. The use of DL and CV in conjunction has brought about an increased efficiency in the analysis of images for medical uses. In this context, DL and CV are very helpful in practice due to their precision and fast image analysis. One of the main DL techniques is Convolutional Neural Networks (CNNs). This is because such neural networks can extract relevant visual patterns and hierarchical features directly from images without the need for extensive feature engineering. Thus, fewer manual interventions are required, making the resulting image analysis reliable and accurate enough for healthcare applications. More specifically, such advances are contributing to the emergence of new diagnostic support technologies by making clinical image analysis more scalable, reliable, and objective [1].

It is important to note that deep learning has not yet displaced traditional clinical approaches but is increasingly used in medical imaging analysis. Such approaches can be efficient, cost-effective, and precise enough in practice. The use cases for deep learning in medical diagnosis involve segmentation of anatomical structures in imaging modalities, including CT, MRI, and endoscopy, anomaly detection, and disease classification. AI systems can help improve access to expert-level diagnosis in resource-constrained settings, lower workload, and decrease interpretation time [2].

In particular, the proliferation of thousands of images generated through gastrointestinal endoscopy examination makes this field appropriate for AI automation, which can help with quicker and more precise medical analysis [3].

Deep learning techniques have been successfully applied to disease detection and classification in medical imagery, providing superb accuracy. Regarding capsule endoscopy, deep learning models are used to detect diseases and conditions including erosions, bleeding, polyps, and other mucosal lesions [4]. Convolutional neural networks (CNN), such as ResNet, DenseNet, and EfficientNet, demonstrated impressive ability to learn complex visual features in medical data labeled through gastrointestinal datasets [5]. Large amounts of curated data and worldwide AI competitions played a major role in advancement of this field [6].

Deep CNN trained on domain-specific Data are well suited for such purposes, which require contextual understanding and precise visual analysis. Although there have been advancements, several problems exist regarding DL models that prevent them from being used in routine clinical practice. The first one is due to the interpretability problem, whereby many models work as black boxes and the rationale behind their decisions is not clear to the clinicians [7]. Second, in the case of abnormalities in the gastrointestinal tract, there is an imbalance in the dataset, since more majority images (that are normally labeled) are present than minority classes, including worm infection or foreign objects [8].

As a result, the goal of this paper is to design an XKD method for detection of gastrointestinal abnormalities from capsule endoscopic images using an expressive teacher model and a light student model [9]. To improve transparency, the approach utilizes Grad CAM for visualizing model predictions, which allows doctors to check whether the model's results correlate with medically relevant areas [10]. The proposed method was tested using the Capsule Vision 2024 Challenge dataset with ten gastrointestinal findings, ranging from common to rare cases [11].

In order to provide a just and accurate model, class imbalance was tackled using the stratified resampling technique, data augmentation, and smart sampling techniques. Visualization from Grad CAM also validated that the model reliably attended to clinically important regions of interest. These improvements help ensure accuracy, reliability, and fairness, thereby providing an effective framework for deployment.

II. RELATED WORKS

Following 2015, the process of gastrointestinal abnormality detection from capsule endoscopy has evolved from the conventional machine learning-based models to more advanced deep learning models. In the works done by Liu and Yuan [12] and Bordbar et al. [13], it was shown how effective the automated diagnosis can be when performing detection based on hand-crafted features as well as the earliest types of CNN architectures. However, these studies faced some limitations, especially in scalability and effectiveness for large datasets. In the following studies, efforts were made toward increasing robustness and clinical relevance. In particular, multi-task learning models, as seen in the paper published by Mayya et al. [14] offered simultaneous classification and segmentation capabilities.

Addressing the problem of data availability and imbalance, Amoli et al. [15] used transfer learning and data augmentation techniques to achieve impressive results despite small amounts of annotated images. Similarly, Kulkarni et al. [16] proposed an approach combining CNNs with EWT to enhance the feature representation and multiclass classification performance. Further research like GastroNet [17] proposed an end-to-end deep learning model that achieved frame-level classification with a high level of accuracy, while not relying on hand designed features. Moreover, pediatric capsule endoscopy has experienced the advancements in terms of deep learning models, further reducing the time spent on the procedure and increasing lesion detection accuracy [18]. Nevertheless, there are still some issues regarding the understanding of the model, the level of uncertainty, and generalization.

2.1 Evolution of explainable AI: recent breakthroughs

Emerging developments in XAI research have led to the realization that models need to be developed not only with accuracy in mind but also with clear explanations for the decision-making process. In medical imaging, explanation transparency is necessary for gaining clinician trust and ensuring safe adoption into clinical practice.

According to the research work of Hou et al. [19] and Purwono et al. [20, despite the advancements that have been made in using visualization techniques for explaining visual data, there are still problems in terms of reliability, robustness, evaluation measures, and real-world applicability.

Table 1 Recent Advances in GI Abnormality Classification Using Capsule Endoscopy

Author name & Year	Methodology	Results	Remarks
Mohapatra et al. [32]	Empirical wavelet transforms with CNN	Delivered effective multi-class GI abnormality classification.	Enhance performance through attention mechanisms and multi-centre validation.
Amoli et al. [15]	Deep CNN with data augmentation	Achieved strong performance on small and imbalanced GI datasets.	Improve generalization across diverse populations and reduce false positives.
Aggarwal et al. [17]	End-to-end CNN (GastroNet V1)	Enabled accurate automated frame-level classification.	Improve model interpretability for better clinical trust.
Huang et al. [33]	Deep learning for pediatric VCE	Provided rapid and reliable multiclass lesion detection in children.	Incorporate explainability and uncertainty estimation for clinical use.

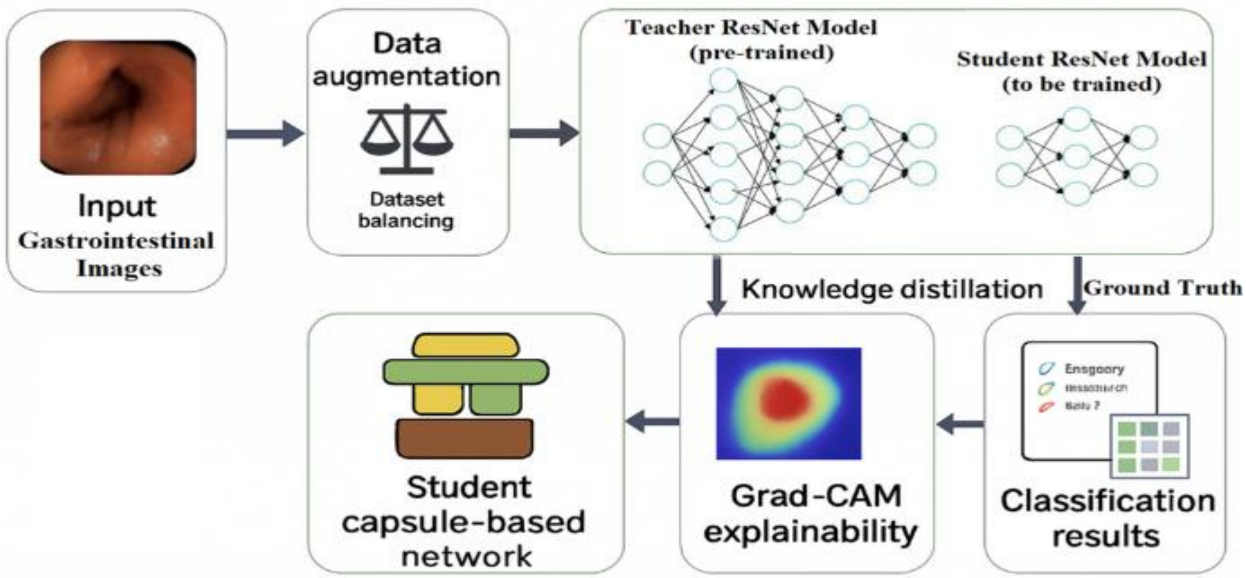

Fig. 1 System Architecture of the Proposed KD-ResCaps Framework

2.2 Knowledge distillation frameworks for medical image interpretation

Knowledge distillation has been proven to be a promising methodology for the creation of efficient and accurate AI models in medicine. Through knowledge transfer from large teachers to smaller students, scientists have managed to achieve higher efficiency without considerable losses in accuracy. Use cases range from segmentation tasks for medical images [21] to MRI reconstruction [22] and CT-based disease classification [23]. In more recent studies, scientists have also attempted to improve the interpretability of models through knowledge distillation by transferring attention maps along with predictions [24]. Nevertheless, many state-of-the-art methodologies are still computationally heavy and specialized to certain imaging modalities. Future research will need to address several challenges, such as maintaining explainability post-compression, dealing with rare diseases, and ensuring model fairness in class-imbalanced datasets [25].

2.3 Progress in Grad-Cam and ResNet for interpretable image analysis

Grad-CAM is now one of the most popular methods for analyzing the reasoning behind deep learning models in medical imaging. With the help of advanced networks like ResNet and Inception-ResNet, it aids in determining the areas of the image affecting the model's output. It has been found to perform well in the classification of breast cancer, providing clinically relevant visualizations [26]. Although it is effective across different networks, Grad-CAM requires further enhancement to detect smaller abnormalities effectively [27].

2.4 Recent works on Gastro Intestinal (GI) diseases

In recent times, deep learning approaches have made substantial progress in the field of automated GI pathology detection using capsule endoscopy. Models based on EfficientNets, like CapsuleNet, have performed well in terms of classification accuracy for several types of GI diseases [28Progressive training methods, along with data augmentation techniques, have further helped enhance the accuracy of anomaly detection methods in the capsule endoscopy video frames [29]. Vision transformer models, like FasterViT, have been shown to perform with faster speeds and efficiency while still retaining accuracy [30]. Furthermore, it has been observed that the use of CNNs along with transformers is another promising approach due to their ability to extract both local as well as global features from images, thereby helping to detect complex abnormalities [31]. However, issues like class imbalance, similarity among lesions, and real-time implementation require further attention.

III. METHODOLOGY

The proposed architecture for the classification of GI disorders based on images obtained by capsule endoscopy is an integration of explainable deep learning and knowledge distillation. As seen in Figure 1, the methodology starts with the process of capturing the images, followed by image preprocessing, data balancing, data augmentation, model training, and knowledge transfer. A teacher network based on ResNet serves to train a student network through both soft predictions and hard labels. In addition, the use of Grad-CAM facilitates clinical interpretability by highlighting the significant areas that impact each prediction made by the model.

3.1 Dataset Description and Challenges

The dataset used in this research work was sourced from the Capsule Vision 2024 Challenge [34]. The data comprises images from capsule endoscopies belonging to ten clinically important categories, such as Angio ectasia, Bleeding, Erosion, Erythema, Foreign Body, Lymphangiectasia, Normal, Polyp, Ulcer, and Worms. These categories include several types of gastrointestinal ailments, ranging from mild to severe illnesses, making this dataset a potentially important resource for diagnostic uses. However, the main problem associated with this type of data lies in its imbalance across several categories. For instance, while the Normal category has 12,287 images, the uncommon but clinically significant categories of Worms and Ulcers have only 68 and 286 images, respectively. Another potential problem with the analysis of capsule endoscopy images lies in their poor quality, including motion blur, illumination variation, resolution problems, and different angles of view.

3.2 Class Balancing Strategy

A dual approach to balance the dataset was adopted with the goal of achieving class balance without losing diagnostic diversity as well as equalizing representation across the entire set.

- Class balancing via random up sampling: Rarer classes, such as Worms (158 occurrences), were subjected to oversampling based on random sampling from replicated bootstraps. It helped the learning model encounter rarer classes and learn appropriate boundaries between classes without the danger of overfitting.
- Majority classes down sampling: Very common classes, including those under the Normal class category (28,663 occurrences), underwent under-sampling based on random sampling without replacement.

Through this technique of balancing both under sampled and oversampled classes, equal distribution was ensured across all ten classes.

3.3 Data augmentation and preprocessing transformations

Various techniques for augmenting our training data were used to enhance model robustness and achieve generalizability within various clinical and imaging settings. Variations in capsule endoscopy images come naturally, and this is actually an advantage because it allows our model to learn more robust features without relying on only a few variations.

Variation enhancement for real-world training data

To increase the robustness and ability of the model to generalize, we used several methods of data augmentation in the training process. The transformations reflect possible changes of appearance that happen in capsule endoscopy images. Hence, training on such data teaches the network not only to distinguish between lesions but also be less sensitive to changes that may occur in reality.

1. Random Horizontal Flip:
Depending on the direction of moving through the gastrointestinal tract, lesions might appear in the image from different sides and positions. Hence, flipping the images horizontally increases the chances for successful detection despite variations of position and orientation.

2. Small Angle Rotation (+/-15°):
The camera might move continuously in the gastrointestinal tract, resulting in different orientations of tissue. By applying small angle rotations to training images, we train the network not to rely too much on the orientation of the object.

3. Random Brightness and Contrast (Color Jitter):
Due to the irregularity of lighting and reflections in the gastrointestinal tract, the images might change in brightness and contrast. Training on such images helps the model to ignore illumination change when performing inference.

4. Affine Translation (+/- 10%):
In order to simulate the small positional displacements caused by the shifting of capsules, small translations were added. This ensures that the model is robust to spatial variance and can detect any abnormalities regardless of their location in the image.

5. Image Resizing (224 x 224):
All images were resized to 224 x 224 pixels in order to comply with the input size requirements for the ResNet network.

Table 2 Class-wise Distribution of Training and Validation Images in the Original Dataset

Class Label	Training Samples	Validation Samples
Angioectasia	1154	497
Bleeding	834	359
Erosion	2694	1155
Erythema	691	297
Foreign Body	792	340
Lymphangiectasia	796	343
Normal	28663	12287
Polyp	1162	500
Ulcer	663	286
Worms	158	68

Ensuring Fair and Reliable Evaluation
In regard to the testing set, however, only resizing and normalization were performed without any augmentation. Thus, the original image features were maintained in order to ensure an unbiased evaluation. Table 2 shows the composition of the dataset prior and after preprocessing and augmentation.
The composition of the dataset shown in Table 2 has not changed from the initial one. As the dataset had a large number of minority classes, such as Worms class, augmentation was applied exclusively to the training set. No augmentation has been done for the validation set to ensure unbiased assessment.

3.4 Custom Dataloader
Efficient Data Pipeline for Capsule Endoscopy Analysis
In order to deal with the complicated nature of the Capsule Vision 2024 dataset, a unique Dataset object was designed using PyTorch. Our pipeline is based on the structure of the dataset, making the process of preparing data during the training phase very convenient. Some tests were performed on the Capsule Vision 2024 Challenge dataset. The Capsule Vision 2024 dataset consists of ten classes of abnormality. The classes are not balanced. With this problem the student model that we distilled did a great job: it was accurate 96.52 percent of the time and it had a weighted F1-score of 96.37 percent.
The Capsule Vision 2024 Challenge dataset results were actually, better than other approaches that we tried. To get an idea of how the model makes its decisions we Grad-CAM visualizations. These visualizations show that the model pays attention to the parts of the image that're important for medical reasons. This makes the model more transparent which is really helpful especially when doctors are using it.
Here is what we did to train the model and handle the data:

- We used batches of 16 samples at a time, which we ran through PyTorch DataLoader. This kept the training process efficient. It did not use too much memory.

- We randomized the order of the training data so that the model would not be biased towards the order of the samples. However, we did not randomize the test data so that the results would be fair.
- We made a tweak to improve performance: we used pinned memory and we set up multi-worker loading, which made it faster to transfer data and it made better use of the GPU.

The whole model pipeline is also very flexible. This means that we can easily add classes or try new things at test time or even customize the model for specific patients, in the future. The Capsule Vision 2024 Challenge dataset model pipeline is very easy to modify.

3.5 Model Architecture

This section presents the design and implementation of an explainable GI abnormality classification framework that integrates a ResNet-based teacher student architecture with Grad-CAM for visual interpretability. The proposed system is developed to efficiently handle capsule endoscopy data while ensuring transparency in predictions. Capsule endoscopy is an extensively used, minimally invasive process that is used in visualizing the gastrointestinal tract. It has been effective in providing visualization of the digestion system, thus making it easier to diagnose any issues within the digestion system, such as bleeding, ulcers, erosions, and polyps. Nonetheless, thousands of images are obtained during one test session; this makes the analysis of these images laborious, difficult, and full of inconsistencies. This emphasizes the need to have efficient and automated diagnostics. Deep Learning process has proved quite successful in the analysis of medical images Specifically, the Residual Neural Network (ResNet) has shown great success because of its hierarchical feature learning property, which allows it to analyze intricate visual patterns. This work proposes a unified framework based on Explainable AI (XAI) and Knowledge Distillation (KD), focusing on both performance and interpretability.

Main Features of the Suggested Framework

- Learning Effectively with Knowledge Distillation

The distilled model has been developed by learning efficiently from the knowledge of the teacher by training it using both soft targets provided by the teacher network and hard targets from ground truth data. This will enable the student network to attain results equivalent to the teacher with reduced computational efficiency, thus suitable for practical applications in the clinical setting.

- Transparent Decision-Making Using Grad- CAM

Grad-CAM is applied to reveal salient areas of each image input in the capsule image dataset that affect decision-making. This visualization technique helps clinicians understand the rationale of decisions made.

- Optimal Accuracy with Practicality

The suggested distilled model shows high performance in the test set with an accuracy of 96.52% and weighted F1-score of 96.37%.

Teacher Student Model Design

- Teacher Model (Knowledge-Rich ResNet18):

The teacher model is based on a pre-trained ResNet18 trained on large-scale datasets such as ImageNet. This provides strong general feature extraction capabilities. The final fully connected layer is replaced with a custom classification layer for 10 GI abnormality classes, allowing adaptation to capsule endoscopy data.

- Student Model (Lightweight and Deployable Design):

This particular design student model is basically an efficient one because it is designed to learn to behave like the teacher via knowledge distillation. However, it is still practical enough to be deployed in low-resource scenarios such as the implementation of the model in embedded medical devices. More so, this model is not structurally simplified, although the term “lightweight” pertains to its practicality, stable convergence, and clinical application potential.

3.6 Model Training Strategy

1. Training of the Teacher Model

We have started from ResNet18 as a backbone, which is pre-trained on ImageNet and fine-tuned on our GI diagnosis problem via a balanced capsule endoscopy dataset. Specifically, we exchanged the original classification layer for an alternative that allows for ten different GI anomalies detection. Optimizer is Adam with a learning rate of 1e-4 and cross-entropy loss. The latter allows for overcoming the limits of pattern recognition by building a powerful feature representation.

2. Transfer of Knowledge to the Student Model

Having achieved desired performance from the teacher model, we fixed its weight for further reference. As we mentioned above, our student model will consist of two parts convolutional base, inherited from teacher, and classifier. For that purpose, the student model has been trained according to the knowledge distillation approach. Combining cross-entropy loss with KL-divergence has provided the student model with the possibility to learn both target labels and soft outputs generated by the teacher model. Thus, the model acquired deep insight into class correlations allowing for reproducing the teacher's accuracy without significant resources consumption.

3. Training Strategy and Monitoring

For the sake of simplicity, we have demonstrated our experiment over only one epoch. However, our implementation is scalable and can be extended to larger amounts of data and epochs. We have continuously monitored batch-wise loss in order to observe the course of training in real-time via tqdm.

4. Fair and Consistent Learning Setup

In order to demonstrate our idea fairly, we used equal conditions for both models in terms of training. In particular, we trained both models using the same balanced dataset with identical optimizer parameters. While the former one serves as the main knowledge source, which should build up the sophisticated representation, the latter acts as compression of the information acquired.

3.7 Knowledge Distillation for Efficient Medical Ai Models

Knowledge distillation (KD) is a model compression approach where a large, high-performing neural network (teacher) transfers its learned knowledge to a smaller, lightweight model (student). This is particularly crucial for the application of the system in real-time and constrained medical environments, where the model should continue performing well but at the same time utilize less parameters and less computational power. In this study, a combined KD loss function was applied to train the network using both soft and hard labeling approaches:

1) Learns from Ground Truth (Cross-Entropy Loss): In this case, the student will be able to learn labeled data from the training dataset. The model will compare its output with the true class label and punish it in cases of mistakes, thus learning good boundaries for classes.
2) Learns from Teacher Knowledge (KL Divergence Loss): Through this process, the student will be able to imitate the teacher using probability distributions rather than only labels. The softmax function with temperature scaling ($T = 2.0$) enables the student to understand the complex relationships between classes.

The total loss during training is defined as $L=\alpha\cdot \text{KLDivLoss}(\text{softmax}(s),\text{softmax}(t))+(1-\alpha)\cdot \text{CrossEntropyLoss}(s,y)$

where:

- s: logits from the student model
- t: logits from the teacher model
- y: ground-truth label
- $\alpha = 0.5$: weight to balance soft vs. hard targets
- $T = 2.0$: temperature parameter for softening logits

This approach is important in practice because the student model not only learns the correct predictions but also benefits from the teacher model’s confidence across all classes. It helps in recognizing small visual similarities that can be found in ulcers and erosion, where it becomes very difficult to distinguish between the two. Rather than concentrating solely on labels, the model finds it easier to generalize when it learns such complex visual relationships. At the same time, it does not become bulky and cumbersome. Thus, the model becomes practical in application as well. Overall, the inclusion of both cross-entropy and KL-divergence leads to finding an appropriate balance between precision and efficiency.

IV. OBSERVED RESULTS

In this chapter, the proposed framework for knowledge distillation from the balanced set of capsule endoscopy image data was comprehensively evaluated. The student network, which was developed using the pre-trained ResNet18 teacher model, was evaluated by means of the classification metrics, confusion matrix, Grad-CAM visualization technique, and comparisons with classical deep learning methods. Overall, the results show that the student network demonstrates great diagnostic accuracy but remains lightweight and interpretable.

4.1 Evaluation metrics and grad-cam visualization

These measures included accuracy, precision, recall, and F1-score both macro average and weighted average.

In comparison to the macro average, which treats all classes equally, the second method takes into account the number of instances for each class. This approach solves the issue of imbalance in class distribution. The performance measures for different classes are provided in Table 5 below.

Table 3 Classification Performance of the Model on the Original Dataset

Class	Precision	Recall	F1-Score	Support
Angioectasia	0.99	0.77	0.87	497
Bleeding	0.87	0.85	0.86	359
Erosion	0.95	1.00	0.97	1155
Erythema	0.99	1.00	0.99	297
Foreign Body	0.83	0.89	0.86	340
Lymphangiectasia	0.99	0.64	0.78	343
Normal	0.98	0.99	0.98	12287
Polyp	0.96	1.00	0.98	500
Ulcer	0.78	0.68	0.73	286
Worms	0.88	0.66	0.76	68
Accuracy	0.9652			
Macro Average	0.92	0.85	0.88	16132
Weighted Average	0.97	0.97	0.96	16132

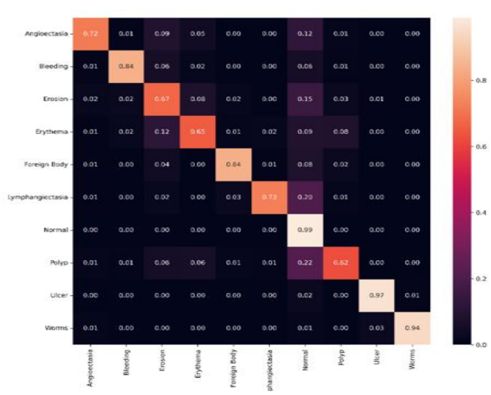

Fig. 2 Confusion Matrix for Validation Samples

Table 4 Comparative analysis of baseline and various models

Model	Accuracy (%)	Remarks
DaViT model [39]	85.92	Shows limited generalization and robustness, indicating that performance in real clinical settings still needs improvement.
CapsoNet model [35]	86.34	Needs better optimization for real-time use, especially with improved generalization and data augmentation strategies.
ViT-CNN Hybrid [33]	89.79	Introduces class-wise anomaly handling, but this increases model complexity and reduces efficiency.
VGG16 and ResNet [32]	91	Performance drops for minority classes even after augmentation, affecting overall test accuracy.
FasterViT model [34]	94.94	Achieves a good trade-off between accuracy and efficiency, though further architectural refinement is still required.
ResNet18 (no KD)	94.8	Strong baseline performance, but limited by higher inference time and larger model size.
KD Student Model	96.52	Delivers the best balance among accuracy, compactness, and speed, making it highly suitable for real-time clinical applications.

The other tool used for evaluation purposes was a normalized confusion matrix (figure 2). All the cells show the number of instances when one class was predicted as another class, which helps understand how the classifier commits errors even when working with unbalanced datasets. This information allows us to understand the strengths and weaknesses of our model. Our predictions were evaluated on the original distribution of data in order to ensure that the results will be balanced for both common and rare gastrointestinal diseases. As a result, the distilled student model produced 96.52% accuracy and a weighted F1 score of 96.37%. This model showed great precision and recall.

The results at the level of individual classes demonstrate almost flawless recognition of Erosion, Erythema, and Polyp. It is also worth noting some confusion between Erosion and Ulcer owing to their visually similar appearance, as well as between Angioectasia and Bleeding. To enhance explainability, Grad-CAM was employed to visualize model attention (figure 3).

These heatmaps indicate that the model is concerned with medically-relevant areas instead of the irrelevant background. For instance, activations were appropriately centered in regions such as bleeding, polyps, and ulcers. Furthermore, based on the results of AUC-ROC analysis, all classes were found to possess significant discriminative abilities with AUC-ROC scores ranging between 0.83 and 0.94. Worm (0.94) and Ulcer (0.92) had the highest discrimination abilities; however, even challenging classes like Erythema (0.83) produced stable results. Taken together, these results suggest the clinical applicability and robustness of the presented model. Figure 5 illustrates the classification performance of various models. While VGG16 yields an accuracy rate of 93.20%, ResNet-18 achieves 94.80%.

However, KD-based student model achieves the best results of 96.52% accuracy rates, outperforming VGG16 and ResNet-18 by 3.32% and 1.72%, respectively. Thus, knowledge distillation was effective in transferring useful features of the teacher model into a light student model.

In Table 4, the advantages of using KD-based student networks in terms of computational speed and memory efficiency while maintaining high accuracy are further illustrated. Overall, student model is highly recommended due to its practicability. Student networks are more effective as they offer better accuracy rates and class-wise performances; moreover, they are easier to interpret compared to their counterparts.

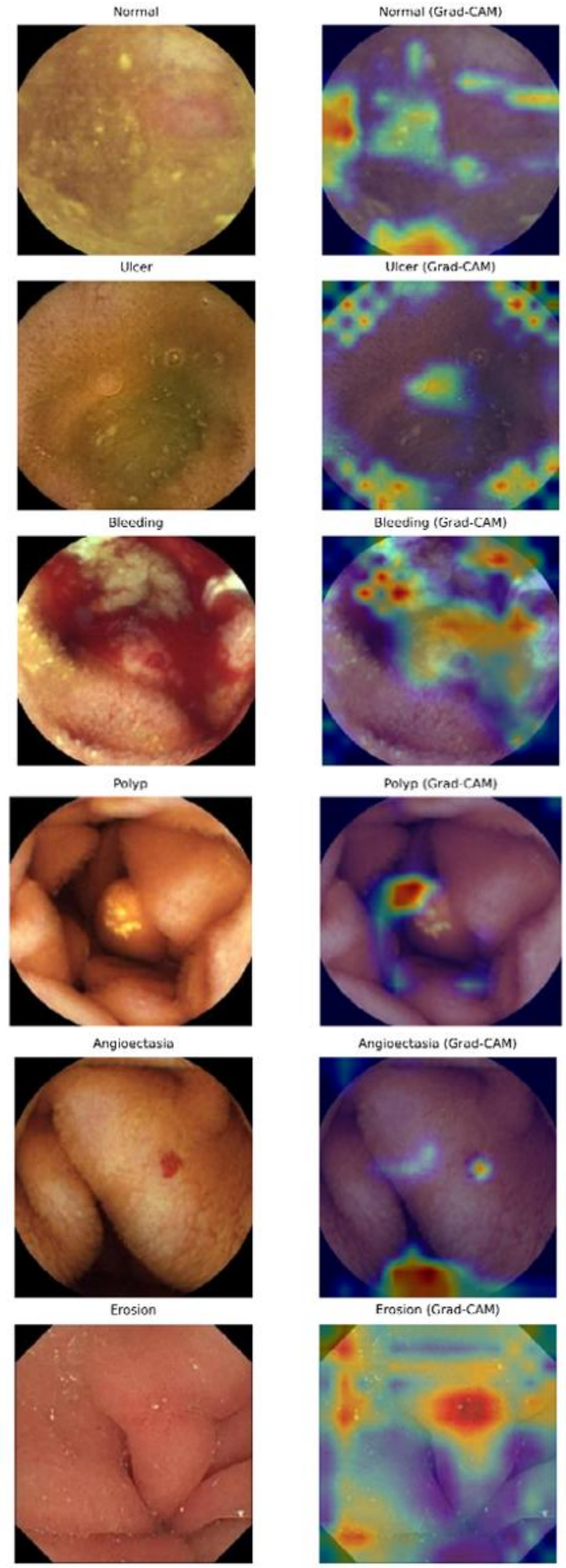

Fig. 3 Grad-CAM overlays highlighting region of interest across test samples

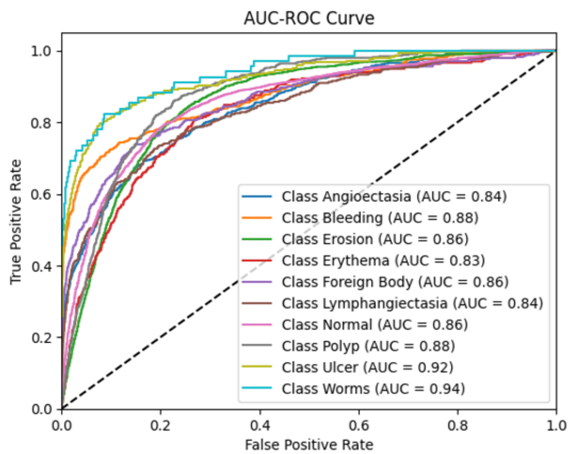

Fig. 4 Multi-class AUC-ROC curves of the model on the original dataset.

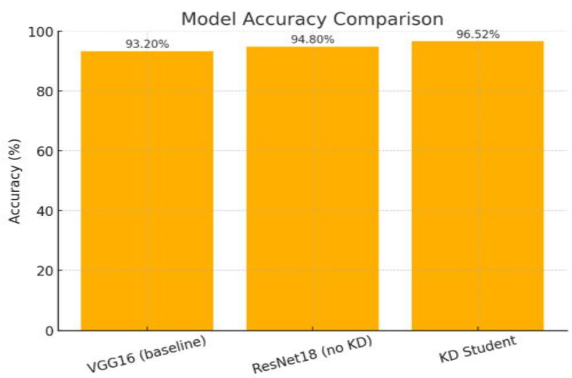

Fig. 5 Comparative accuracy analysis of baseline and knowledge distillation models

Therefore, the proposed approach is suitable for capsule endoscopy and may also be adapted to other medical imaging tasks that require both efficiency and explainability.

V. ABLATION STUDY

To examine how each component of the proposed system contributes to the final results, an ablation study was conducted. The study looks at knowledge distillation, class balancing and Grad-CAM-based interpretability one by one. Also, how they work together. This is really important for datasets that are not balanced where we need to be able to detect the smaller groups to make a correct diagnosis of GI issues. Knowledge distillation has an impact. When we trained the student model on its own it got 94.8% accuracy and 91.7% weighted F1-score. When we used knowledge distillation to train it the performance went up to 96.52% accuracy and 96.37% weighted F1score. The importance of class balancing is clear. We tested the model on the unbalanced dataset.

The performance went down to 89.5% accuracy and 88.0% weighted F1-score.

The decline was particularly pronounced for smaller classes such as worms and ulcers. This demonstrates that it is essential to perform class balancing on the dataset to avoid biases for the model and improve the recognition of the problem. Explainability via Grad-CAM was also critical. Despite the fact that Grad-CAM did not impact classification performance directly, it helped provide transparency and thus increased confidence in the model for clinical practitioners.

Capsule endoscopy creates many images. Analyzing all the images manually is challenging. In these cases, visual explanations help us see which parts of the picture made the model predict what it did. Experts who reviewed the results said they trusted the automated diagnosis more when the Grad-CAM heatmaps matched the locations of the lesions. Also, Grad-CAM still worked well even after we applied knowledge distillation. Since we did not have annotations for each pixel, we could not do an analysis of the locations and that will have to be done in the future. Overall, the study shows that knowledge distillation, class balancing and Grad-CAM each help the final system perform better.

VI. CONCLUSION

The proposed explainable KD-based framework demonstrates that high accuracy and trustworthy decision support can be achieved for GI anomaly detection using deep learning methods.

In particular, the distilled model achieved exceptional results, with an accuracy of 96.52% and a weighted F1-score of 96.37%. In addition, by applying class balancing techniques, we have managed to avoid underrepresentation of rare classes such as ulcers and worms. As for transparency, this KD framework ensures clinical applicability due to the use of Grad-CAM. Moreover, it remains lightweight and modular and, therefore, ready for practical implementation in medicine and flexible enough for other applications.

FUTURE WORK

There are several promising directions for further research on this topic. First of all, there is a need to consider dynamic characteristics of video sequences in addition to static ones. To do this, we will utilize recurrent or transformer-based encoders for detecting motion abnormalities, which could not be caught in a still image. Another direction will consist in using multimodal patient information, including lab results and sensor data in addition to clinical history.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] J. Zhou, N. Hu, Z.-Y. Huang, B. Song, C. C. Wu, F.-X. Zeng, and M. Wu, "Application of Artificial Intelligence in Gastrointestinal Disease: A Narrative Review," *Annals of Translational Medicine*, vol. 9, no. 14, p. 1188, 2021.

[3] P. T. Kröner, M. M. Engels, B. S. Glicksberg, K. W. Johnson, O. Mzaik, J. E. Hooft, M. B. Wallace, H. B. El-Serag, and C. Krittanawong, "Artificial Intelligence in Gastroenterology: A State-of-the-Art Review," *World Journal of Gastroenterology*, vol. 27, no. 40, p. 6794, 2021.

[4] Z. El-Jabbari and J. Boumhidi, "Deep Learning Techniques for Enhanced Classification of Gastrointestinal Images," in *Proc. 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)*, 2024, pp. 1–7.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint* arXiv:1409.1556, 2014.

[6] P. Handa, A. Mahbod, F. Schwarzhans, R. Woitek, N. Goel, M. Dhir, D. Chhabra, S. Jha, P. Sharma, V. Thakur, *et al*., "Capsule Vision 2024 Challenge: Multi-Class Abnormality Classification for Video Capsule Endoscopy," *arXiv preprint* arXiv:2408.04940, 2024.

[7] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[8] M. Buda, A. Maki, and M. A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[9] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional Teacher-Student Learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6445–6449.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[11] S. Pawan and J. Rajan, "Capsule Networks for Image Classification: A Review," *Neurocomputing*, vol. 509, p. 102120, 2022.

[12] J. Liu and X. Yuan, "Obscure Bleeding Detection in Endoscopy Images Using Support Vector Machines," *Optimization and Engineering*, vol. 10, no. 2, pp. 289–299, 2009.

[13] M. Bordbar, M. S. Helfroush, H. Danyali, and F. Ejtehadi, "Wireless Capsule Endoscopy Multiclass Classification Using Three-Dimensional Deep Convolutional Neural Network Model," *BioMedical Engineering OnLine*, vol. 22, no. 1, p. 124, 2023.

[14] V. Mayya, K. Karthik, K. P. Karadka, S. S. Kamath, *et al*., "Multi-Task Deep Neural Network Models for Learning COVID-19 Disease Representations from Multimodal Data," *International Journal of Medical Engineering and Informatics*, vol. 15, no. 6, pp. 501–515, 2023.

[15] E. R. Amoli, A. A. Tehranizadeh, and H. Arabalibeik, "Utilizing Deep Convolutional Neural Networks and Hybrid Classification for Gastrointestinal Disease Diagnosis from Capsule Endoscopy Images," *Journal of Biomedical Physics & Engineering*, vol. 15, no. 4, p. 369, 2025.

[16] S. Kulkarni, G. K. Thakur, K. Bhargavi, C. R. Babu, G. B. Madhavi, and P. C. S. Reddy, "A Hybrid Deep Learning Framework for Gastrointestinal Tract Classification Based on Endoscopic Images," in *Proc. 2nd World Conference on Communication & Computing (WCONF)*, 2024, pp. 1–6.

[17] S. Aggarwal, I. Gupta, A. Kumar, S. Kautish, A. S. Almazyad, A. W. Mohamed, F. Werner, and M. Shokouhifar, "GastroFuse-Net: An Ensemble Deep Learning Framework Designed for Gastrointestinal Abnormality Detection in Endoscopic Images," *Mathematical Biosciences and Engineering*, vol. 21, no. 8, pp. 6847–6869, 2024.

[18] M. Mascarenhas, M. J. Almeida, M. Martins, P.

Cardoso, F. Mendes, H. Cardoso, J. Ferreira, and G. Macedo, "Deep Learning and Capsule Endoscopy: Automatic Panendoscopic Detection of Protruding Lesions," *Journal of Coloproctology*, vol. 44, Suppl. 1, p. 246, 2024.

[19] J. Hou, S. Liu, Y. Bie, H. Wang, A. Tan, L. Luo, and H. Chen, "Self-Explainable AI for Medical Image Analysis: A Survey and New Outlooks," *arXiv preprint* arXiv:2410.02331, 2024.

[20] P. Purwono, A. N. E. Wulandari, and K. Nisa, "Explainable Artificial Intelligence (XAI) in Medical Imaging: Techniques, Applications, Challenges, and Future Directions," *Advanced Mechanical and Mechatronic Systems*, vol. 1, no. 1, pp. 52–66, 2025.

[21] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient Medical Image Segmentation Based on Knowledge Distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[22] B. Murugesan, S. Vijayarangan, K. Sarveswaran, K. Ram, and M. Sivaprakasam, "KD-MRI: A Knowledge Distillation Framework for Image Reconstruction and Image Restoration in MRI Workflow," in *Medical Imaging with Deep Learning (MIDL)*, pp. 515–526, 2020.

[23] H. Chen, Y. Jiang, H. Ko, and M. Loew, "A Teacher-Student Framework with Fourier Transform Augmentation for COVID-19 Infection Segmentation in CT Images," *Biomedical Signal Processing and Control*, vol. 79, p. 104250, 2023.

[24] J. Zheng, K. Xie, D. Zhang, Z. Lv, and X. Yu, "Stepwise Self-Knowledge Distillation for Skin Lesion Image Classification," *Scientific Reports*, vol. 15, no. 1, p. 25238, 2025.

[25] S. Ahmad, Z. Ullah, and J. Gwak, "Multi-Teacher Cross-Modal Distillation with Cooperative Deep Supervision Fusion Learning for Unimodal Segmentation," *Knowledge-Based Systems*, vol. 297, p. 111854, 2024.

[26] S. Liu, G. M. S. Himel, and J. Wang, "Breast Cancer Classification with Enhanced Interpretability: DALAResNet50 and DT Grad-CAM," *IEEE Access*, 2024.

[27] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, A. Bolatkan, N. Shinkai, *et al*., "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024.

[28] H. Singh and A. Kumar, "Capsule Vision 2024 Challenge Report," *Authorea Preprints*, 2024.

[29] A. Bansal, B. Singla, R. R. Wankhade, and N. Patil, "Capsule Vision Challenge 2024: Multi-Class Abnormality Classification for Video Capsule Endoscopy," *arXiv preprint* arXiv:2411.01479, 2024.

[30] M. Saravanan, "Optimizing Image Classification in the Capsule Vision 2024 Challenge Using FasterViT: A Study in Efficient Vision Transformer Adaptation," *Authorea Preprints*, 2024.

[31] A. Samal and R. Batsyas, "Multi-Class Abnormality Classification in Video Capsule Endoscopy Using Deep Learning," *arXiv preprint* arXiv:2410.18879, 2024.

[32] S. Mohapatra, G. K. Pati, M. Mishra, and T. Swarnkar, "Gastrointestinal Abnormality Detection and Classification Using Empirical Wavelet Transform and Deep Convolutional Neural Network from Endoscopic Images," *Ain Shams Engineering Journal*, vol. 14, no. 4, p. 101942, 2023.

[33] Y.-H. Huang, Q. Lin, X.-Y. Jin, C.-Y. Chou, J.-J. Wei, J. Xing, H. M. Guo, Z.-F. Liu, and Y. Lu, "Classification of Pediatric Video Capsule Endoscopy Images for Small Bowel Abnormalities Using Deep Learning Models," *World Journal of Gastroenterology*, vol. 31, no. 21, 2025.

[34] MIAAI, "MISAHUB: Training and Validation Dataset of Capsule Vision 2024 Challenge, Version 3," 2024.

[35] D. R. Verma, V. Saxena, D. Sharma, and A. Gupta, "Multi-Class Abnormality Classification Task in Video Capsule Endoscopy," 2024.