Utilizing Deep Convolutional Neural Networks and Hybrid Classification for Gastrointestinal Disease Diagnosis from Capsule Endoscopy Images

Ehsan Roodgar Amoli (PhD Candidate)[1] , Amin Amiri Tehranizadeh (PhD)[2], Hossein Arabalibeik (PhD)[1]*

[1]Department of Biomedical Systems & Medical Physics, Tehran University of Medical Sciences, Tehran, Iran
[2]Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran

ABSTRACT

Background: Wireless Capsule Endoscopy (WCE) is the gold standard for painless and sedation-free visualization of the Gastrointestinal (GI) tract. However, reviewing WCE video files, which often exceed 60,000 frames, can be labor-intensive and may result in overlooking critical frames. A proficient diagnostic system should offer gastroenterologists high sensitivity and Negative Predictive Value (NPV) to enhance diagnostic accuracy.

Objective: The current study aimed to establish a reliable expert diagnostic system using a hybrid classification approach, acknowledging the limitations of individual deep learning models in accurately classifying prevalent GI lesions. Introducing a hybrid classification framework, ensemble learning techniques were applied to Deep Convolutional Neural Networks (DCNNs) tailored for WCE frame analysis.

Material and Methods: In this analytical study, DCNN models were trained on balanced and unbalanced datasets and then applied for classification. A model scoring hybrid classification approach was used to create meta-learners from the DCNN classifiers. Class scoring was utilized to refine decision boundaries for each class within the hybrid classifiers.

Results: The VG_BFCG model, constructed on a pre-trained VGG16, demonstrated robust classification performance, achieving a recall of 0.952 and an NPV of 0.977. Tuned hybrid classifiers employing class scoring outperformed model scoring counterparts, attaining a recall of 0.988 and an NPV of 1.00, compared to 0.979 and 0.989, respectively.

Conclusion: The unbalanced dataset, with a higher number of Angiectasia frames, enhanced the classification metrics for all models. The findings of this study underscore the crucial role of class scoring in improving the classification metrics for multi-class hybrid classification.

Introduction

Gastrointestinal (GI) tract diseases stand as a predominant contributor to fatal cancers globally, characterized by elevated mortality and incidence rates [1]. The advent of Capsule Endoscopy (CE) [2] has caused a transformative shift in diagnosing Small Bowel

(SB) pathology. Due to recent clinical strides, CE has emerged as the predominant diagnostic modality for SB issues, primarily attributed to its non-invasive nature and commendable outcomes [3-5]. The CE in early screenings can lead to saving lives [6]. The CE recordings yield approximately 60,000 image frames for 8-12 hours [7].

The CEs are recommended for detecting Obscure Gastrointestinal Bleeding (OGIB), iron deficiency anemia, suspected Crohn's disease, and other pathologies in patients, who have undergone upper endoscopy and colonoscopy [3-5].

The CEs have challenges and limitations, including extended and labor-intensive reviewing times, as well as the potential for overlooking lesions, which are frequently reported as primary issues [7-11]. The review of a CE video file typically demands 30 to 180 minutes with a heightened risk of diagnostic errors [12]. Additionally, diagnostic software, such as Suspected Blood Indicators (SBI) has demonstrated suboptimal performance in terms of sensitivity and specificity [13-16].

Multiple studies have focused on developing computer-aided diagnosis systems to analyze CE videos, with a primary emphasis on utilizing deep learning techniques. However, the focus was on detecting bleeding, subsequent efforts also explored the classification of other abnormalities, such as polyps and tumors [17-20]. For instance, Tsuboi et al. [21] implemented a Deep Convolutional Neural Network (DCNN) to develop a diagnostic tool capable of discerning between type 1a and 1b small-bowel Angioectasia in CE images, with notable detection rates.

Hajabdollahi et al. [22] introduced a network designed for the identification of multiple abnormalities utilizing a bifurcated structure with 97.5% sensitivity, 99.3% specificity, and 99.0% accuracy. Leenhardt et al. [23] achieved notable outcomes with 100% sensitivity, 96% specificity, and 100% Negative Predictive Value (NPV) using their DCNN model for Angiectasia detection. However, important factors, such as pathology type and size were overlooked. In a comparison of deep learning architectures focusing on frames with normal features and frames with Angiodysplasia, Valeria et al. [12] discovered that DenseNet-161 exhibited superior performance, with a precision of 94% and a recall of 93%. However, it is worth noting that preprocessing had a negative impact on the overall performance of DenseNet-161.

Based on deep learning techniques, erosions and ulcers in CE images, were identified with 88.2% sensitivity and 90.8% specificity [24]. The DCNN model was also implemented on an augmented dataset comprising 10,000 CE frames for automated bleeding detection with a sensitivity of 99.20% and precision of 99.90% [25]. It is worth noting that the study [25] did not provide information regarding the utilization of validation data.

Fonseca et al. [26] evaluated three DCNN-based models for classifying an imbalanced dataset sourced from the Kvasir-Capsule database [27], encompassed categories, such as Angiectasia, Normal, Polyp, and fresh blood. They also grouped Angiectasia, Blood-Fresh, and Polyp into the "not-normal" class and demonstrated the capacity of a DCNN model to categorize small segments of data extracted from video capsule endoscopies. The pre-trained ResNet50 network, in particular, achieved noteworthy results with 99% sensitivity and 69% specificity.

Xception was employed to classify polyps and lesions in a 3-class dataset comprising normal, P1P, and P2P with 95.9% sensitivity, 95.7% NPV, and 97.1% specificity for each pair of classes [28].

The current research concentrates on specific abnormalities and frequently incorporates augmentation techniques. The NPV has not been investigated enough for the multi-class classification of CE images. The NPV is considerably important for the clinical application of DCNN models, especially when

combined with metrics, such as the F1-score. Further investigation and analysis can affect the improvement of the clinical utility of these models.

According to Saurine's classification [29], vascular lesions, such as Angiectasia and Angiodysplasia are designated as high-risk lesions [30], whereas inflammatory lesions are frequently observed in individuals with Inflammatory Bowel Diseases (IBD) and may contribute to the development of colon cancer [31]. Given the significant implications of these pathologies, early detection and treatment play a crucial role in reducing the risk of mortality in patients. Therefore, the current study aimed to develop a Computer-aided Diagnosis (CAD) system, accurately classifying these pathologies using advanced deep learning techniques.

A comprehensive four-class dataset, consisting of Angiectasia, Inflammatory, Normal, and Angiodysplasia was constructed to ensure accurate classification. Subsequently, an expert DCNN was developed and trained on this dataset, with color and textural patterns specific to high-risk pathologies. The first goal of the present study is to achieve high recall and a strong NPV, recognizing the importance of these metrics for physicians relying on AI-based classifier systems.

The compilation of our dataset involved leveraging three public databases, enabling the creation of a robust DCNN model. Our performance objectives include achieving an NPV greater than 0.99, an overall accuracy exceeding 0.975 and recalls for each class surpassing 0.95 in the context of the four-class image classification task.

Six models, trained on both balanced and unbalanced datasets, were devised to optimize model performance. In the present study, model and class scoring approaches were implemented by employing ensemble learning techniques. Weights were assigned to individual models and classes, respectively, and determined the optimal weights through a full factorial experiment design. Moreover, the current study represents the first investigation to implement hybrid classification for CE images, with a particular focus on metrics such as NPV, overall accuracy, and recall. Based on the obtained results, improving model performance could be accomplished by integrating both model scoring and class scoring mechanisms. This integration effectively modified class distributions within decision boundaries, resulting in enhanced classification results.

Material and Methods

Dataset

This analytical study was conducted based on public databases. Wireless Capsule Endoscopy (WCE) video files for individual patients typically comprise only a limited number of abnormal frames. Consequently, the annotation of these frames is considered a time-intensive task for gastroenterologists, leading to the dataset specifically focused on vascular lesions and inflammatory lesions, which were selected based on their prevalence and potential risk. Some unbalanced datasets were experimented to identify the most suitable one. In the end, an unbalanced dataset was utilized that featured a higher number of Angiectasia frames.

The current study was conducted based on an independent test set, which was completely separated from the data used for training the model, resulting in ensuring the robustness and generalizability of the findings. By using an independent test set, the performance of our model was assessed on unseen data; the effectiveness of our approach was validated in real-world scenarios.

Diverse datasets were employed to develop DCNN models. The datasets utilized encompass GIANA [32], KID series [33], and Kvasir-Capsule [27]. The inclusion of diverse frames from these datasets plays a crucial role in enhancing the robustness of the classifier models. By incorporating a wide range

of samples, the models become more resilient to conditional variations, such as artifacts and variations in the surrounding environments of the capsule endoscopy images. This diversity in data sources contributes to the development of a more comprehensive and adaptable model for effectively analyzing and classifying capsule endoscopy images.

The dataset was segmented into three distinct categories: training, validation, and testing sets, using a random patient-based splitting approach. The allocation of samples followed a specific distribution, with 70% of the data assigned to the training set, 15% to the validation set, and 15% to the testing set. This segmentation ensured that the sets were non-overlapping and met the defined constraints as specified in Equations 1, 2, and 3. In these equations, T, V, and S represent the sets of training, validation, and test data, respectively. The patient sets designated for training, validation, and testing were labeled as P_T, P_V, and P_S, respectively.

$$T \cap V = \emptyset, P_T \cap P_V = \emptyset \qquad \text{Equation 1}$$

$$T \cap S = \emptyset, P_T \cap P_S = \emptyset \qquad \text{Equation 2}$$

$$V \cap S = \emptyset, P_V \cap P_S = \emptyset \qquad \text{Equation 3}$$

This study employed two datasets to examine the influence of additional Angiectasia frames on the overall classification performance (Table 1). The unbalanced dataset featured a higher number of Angiectasia frames compared to the balanced dataset. To ensure an equitable assessment of the model's classification performance, more Angiectasia test frames were incorporated into the unbalanced dataset. All images were in the 24-bit PNG format. Figure 1 illustrates examples of the CE images.

We intentionally did not exclude frames containing artifacts from our dataset. The decision to include such frames was driven by our goal to develop a more resilient model, ultimately enhancing its classification performance.

Data augmentation techniques were employed [34, 35] to address the challenge of limited labeled data and mitigate the risk of

Table 1: Dataset description (balanced and unbalanced)

Classes	Balanced				Unbalanced			
	Train	Augmented	Validation	Test	Train	Augmented	Validation	Test
Angiectasia	420	6300	90	90	600	9000	150	150
Inflammatory	420	6300	90	90	420	6300	90	90
Normal	420	6300	90	90	420	6300	90	90
Angiodysplasia	420	6300	90	90	420	6300	90	90

Angiectasia   Inflamatory   Normal   Angiodysplasia
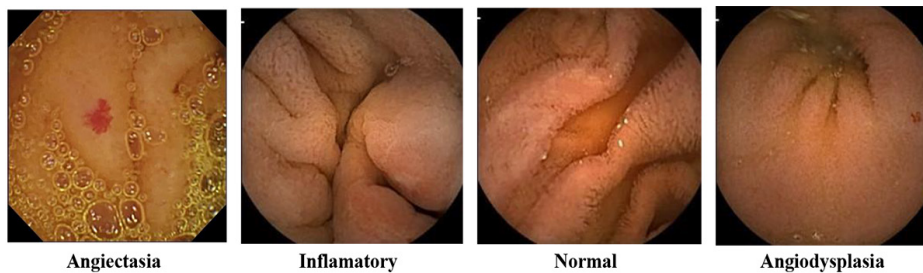
Figure 1: Examples of the pathologies used in developing the Deep Learning (DL) models in this study. The images are sourced from [27, 32], with [27] being a publicly available dataset and access to [32] granted by permission from the corresponding author.

overfitting. This approach involved introducing variations to the training data by applying transformations, such as shifting, horizontal and vertical flipping, rotation, zooming, and luminance adjustment. Data augmentation was exclusively applied to the training data, imparting invariance to brightness changes, scale, and rotation.

Each frame in the training dataset was augmented with three different brightness levels and five rotation levels, resulting in a 15-fold increase in the dataset size. This process causes the model to learn from a more diverse range of examples (Table 1).

Deep learning

Deep Learning (DL) systems have demonstrated superior performance compared to traditional shallow machine learning algorithms, particularly in applications with extensive datasets. The successes of DL models include pattern recognition tasks, such as image classification [36, 37], natural language processing [38], object detection [39, 40], and video analysis [41, 42].

In medical applications with limited datasets, DL models have effectively employed transfer learning, yielding impressive outcomes in tasks, such as classification, localization, detection, and segmentation of medical images. The DCNNs stand out as one of the most commonly utilized DL architectures for medical image analysis.

Deep convolutional neural networks

The DCNNs are commonly composed of two main sections: the ConvNet and the fully connected sections. The ConvNet section is responsible for automatically extracting features of increasing complexity from the input data. Also, this section typically consists of convolutional and pooling layers organized into modules.

In addition to convolutional and pooling layers, these modules may incorporate other techniques, such as batch normalization and dropout layers to improve regularization and prevent overfitting. Batch normalization helps normalize the inputs between layers, while dropout randomly deactivates some neurons during training, forcing the network to learn more robust features.

The overall architecture of a DCNN involves stacking convolutional and pooling modules, followed by a series of fully connected layers. This structure enables the network to learn and interpret the extracted features for accurate decisions.

Figure 2 shows a visual representation of the DCNN structure.

Ensemble learning

In various machine learning frameworks, leveraging a combination of multiple expert decision- makers is a common strategy to enhance performance in challenging situations. This approach is particularly useful when dealing with complex data distributions, class imbalances, and risk management. Figure 3 demonstrates the fusion of base-learners to create a meta-learner, which forms the core of ensemble methods.

Ensemble methods are typically categorized into three main classes: Bagging, Stacking, and Boosting. In the Bagging approach, multiple models are trained on randomly sampled data, and the final predictions are obtained by averaging the outputs of these models. On the other hand, Stacking involves training multiple models on the entire dataset and subsequently using a fusion mechanism to derive the final prediction. In contrast, the Boosting mechanism involves sequentially operating ensemble models on the misclassified predictions of prior models to improve overall performance.

In the current study, the Stacking approach was selected as the ensemble method. The classification performance was improved by combining multiple models using Stacking, causing the models to learn from different perspectives and capture diverse aspects of the
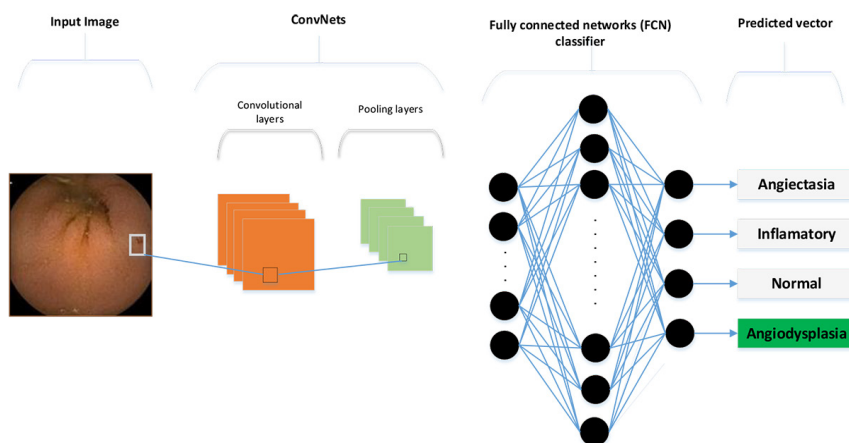
Figure 2: The schematic representation of deep convolutional neural networks (DCNNs).
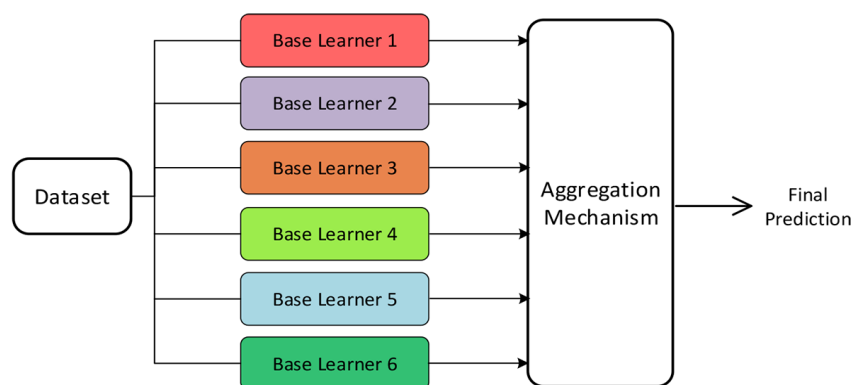
Figure 3: Schematic of Ensemble learning

data. This process results in improved accuracy and robustness in the final predictions.

Proposed method

A total of six well-established DCNN architectures were implemented as the convolutional section of the network to construct the classifier models. Subsequently, a fully-connected classifier network that adjusts was devised, according to the architecture of the convolutional part. The proposed classifier was designated with the suffix BFCG (=batch normalization + fully connected + global average pooling) and integrated batch normalization, fully connected layers, and global average pooling. The architecture of the proposed DCNN model is illustrated in Figure 4.

The final DCNN classifier is composed of the base convolutional part and the designed fully connected part. We denoted these classifiers as VG_BFCG (= VGG16 [43] + BFCG), DN_BFCG (= DenseNet-201 [44] + BFCG), IRN_BFCG (= Inception-ResNet-v2 [45] + BFCG), MN_BFCG (= MobileNetV2 [46] + BFCG), RN_BFCG (= ResNet152V2 [47, 48] + BFCG) and X_BFCG (= Xception [49] + BFCG). A basic schematic of the proposed structure is provided in Figure 4.

The proposed fully connected classifier was developed through empirical trials, incorporating a Batch Normalization (BN) layer and an l1-regularizer to normalize the layer's out-
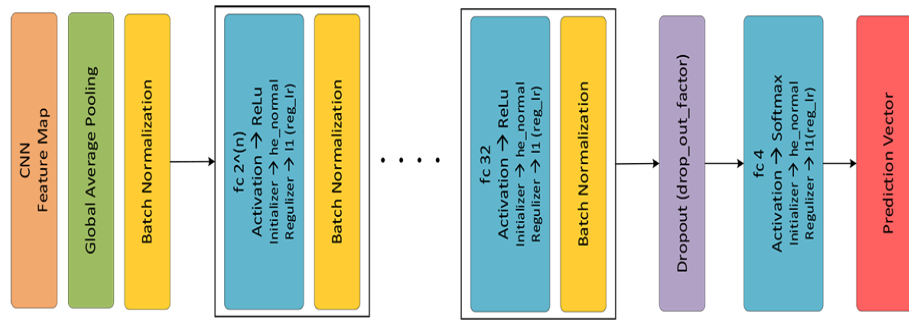
Figure 4: The general structure of the proposed deep convolutional neural networks (DCNN) model. The hyperparameters are defined over feature extractor and classifier blocks. (CNN: Convolutional Neural Network)

put. In the present study, the number of blocks in the classifier significantly influences the model's classification performance. After experimentation, the final block should consist of 32 fully connected neurons. Within each block, the number of neurons is halved in comparison to the fully connected nodes in the preceding block. Consequently, the number of nodes specified in the first block is regarded as a hyperparameter termed the "number of dense nodes".

A grid search was performed to find the best combination of hyperparameters that met the criteria of validation accuracy >0.95, test accuracy >0.95, and NPV >0.95 to identify the optimal model for each architecture. If these criteria were not met, we reported the hyperparameters that yielded the best performance regardless.

The optimal hyperparameter values varied depending on whether the dataset was balanced or unbalanced. For the fully connected classifier, the depth was determined by the number of nodes in its initial layer. Additionally, a regularization coefficient was applied during the training of the fully connected layers as another hyperparameter for the classifier block.

In training the convolutional part of the model, several hyperparameters were tuned, including the batch size, freezing depth (the number of layers to be frozen during training), input size, learning rate, and dropout

coefficient. For data preparation and model execution, the current study utilized Google Colaboratory with the Keras library, taking advantage of cloud GPUs to enhance computational capabilities, leading to efficient processing of the data and training the models in a resource-efficient manner.

Elu [50] activation function and Adadelta [51] optimizer were consistently employed across all models and trials. To enhance training efficiency, an early stopping mechanism was implemented, terminating the training process for hyperparameter configurations, in which the training accuracy did not demonstrate improvement over 15 epochs.

The proposed algorithm utilized the Stacking approach and investigated two ensemble learning methods: model scoring and class scoring. Figure 5a illustrates the model scoring hybrid classification, in which the predicted output of each model was multiplied by its respective weight and aggregated across classes. On the other hand, Figure 5b depicts the tuned hybrid classification, where the aggregated predicted class probabilities from the base learners were weighted. The intention behind this approach was to enhance the margin in the vicinity of the decision boundaries, improving the algorithm's overall performance.

The mathematical expressions for hybrid classification, using model scoring and class scoring for m classes and K classifiers, are

given by Equations 4 and 5, respectively. Weight normalization was conducted by dividing the weights by the total sum of weights. The scores (α_i) are organized in a diagonal matrix (α), as expressed in Equation 6.

$$C(x) = \underset{i=1,2,...,m}{\mathrm{argmax}} \sum_{j=1}^{K} \beta_j P_j(x) \qquad \text{Equation 4}$$

$$C(x) = \underset{i=1,2,...,m}{\mathrm{argmax}} \; \alpha \sum_{j=1}^{K} \beta_j P_j(x) \qquad \text{Equation 5}$$

$$\alpha = \begin{bmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_m \end{bmatrix} \qquad \text{Equation 6}$$

The label assigned by the hybrid classifier for input x is denoted by C(x) and calculated using either model scoring or class scoring, as represented by Equations 4 and 5,

respectively. As shown in Equation 6, the weights of the classes are represented by a diagonal matrix, denoted by α. In both approaches, the class with the highest probability is selected as the assigned label.

In Equations 4 and 5, $P_j(x)$ represents the prediction vector (four rows, one column) generated by the j-th model, and β_j and α_i represent the weights for the j-th model and i-th class model, respectively. In both approaches, the class with the highest probability determines the assigned label.

The optimal weights for the two approaches were determined using a brute force algorithm that systematically explores all possible combinations until a solution is found. The time complexity of this algorithm is typically proportional to the input size.

To obtain the optimal weights, a factorial
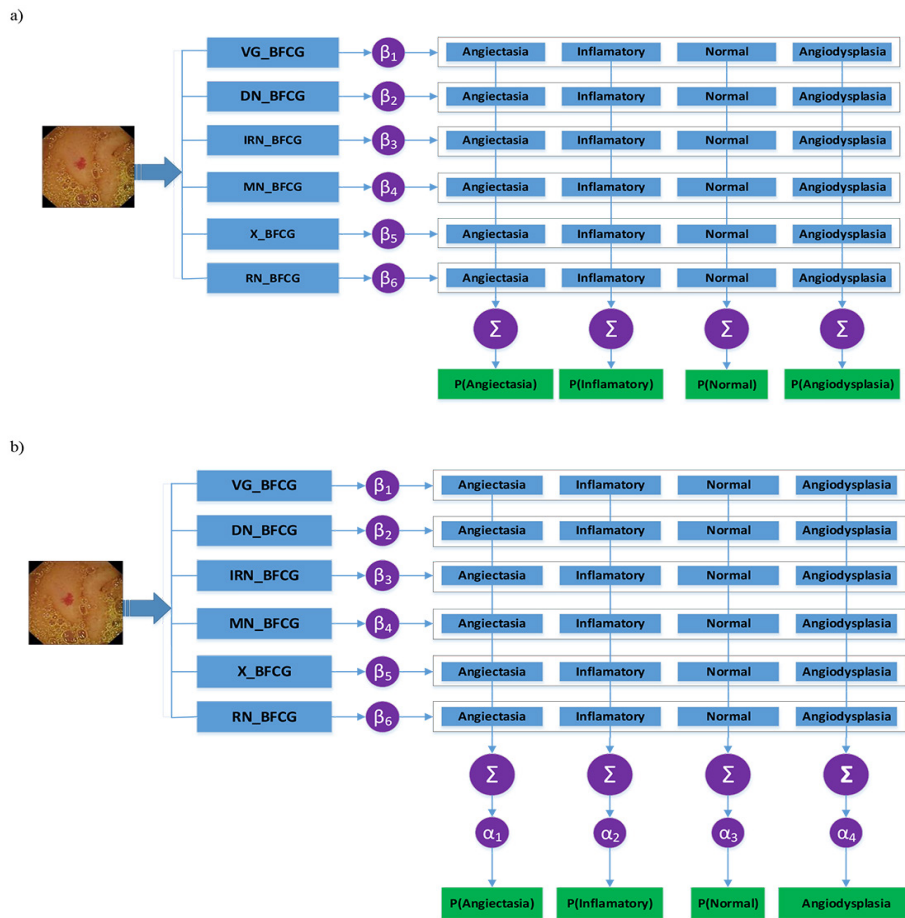
Figure 5: Block diagram of the proposed hybrid classification. **a)** Model scoring, **b)** Tuned hybrid classification using class scoring

design of experiments was employed. The weights of models (β) and classes (α) were defined within the range of [0, 1]. The resolution in the designed experiment was kept consistent for all factors (classes or models). However, due to computational constraints, the search was conducted with different levels of parameters: 16, 20, and 25, resulting in resolutions of 0.0625, 0.05, and 0.04, respectively. For example, with a β resolution of 0.0625, there were 17 weights for each model, leading to a total of 17^6 experiment settings. The search duration on Google Colab for hybrid classification using model scoring was approximately 45 minutes.

In both the model scoring and class scoring approaches (Figure 5a and b), the prediction vector generated by each model is multiplied by its corresponding weight (β) for each class. The modified vectors are then aggregated to produce the final modified predicted vector, and the assigned label (class) is determined as the one with the highest value. The defined weights are utilized during this operation. Additionally, in the tuned hybrid classification using class scoring (Figure 5b), an additional step is introduced by defining a class coefficient (α) to enhance discrimination between classes.

Results

This section presents the performance results of the developed models in the previously introduced four-class classification task. The hyperparameters of the developed models were optimized using grid search and full factorial design of experiments.

Target

This part aimed to identify models that satisfy the following criteria:

i. Recall for each class >0.95
ii. Accuracy >0.975
iii. NPV >0.99.

The performance of both the developed models and the hybrid models were evaluated

based on the following metrics:

$$Recall = \frac{TP}{TP + FN} \qquad \text{Equation 7}$$

$$Precision = \frac{TP}{TP + FP} \qquad \text{Equation 8}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad \text{Equation 9}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad \text{Equation 10}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \qquad \text{Equation 11}$$

TP, TN, FP, and FN denote the quantities of true positives, true negatives, false positives, and false negatives, respectively. Given the limited access to a large medical dataset, the dataset is not balanced. However, the F1-score should be considered (Equation 10) when evaluating the performance of a classifier model on an imbalanced dataset. The F1-score provides a balanced metric by assigning equal importance to precision and recall. This approach is recommended in situations where the class distribution is imbalanced, as it accounts for both the ability of the model to correctly identify positive instances (recall) and its ability to avoid false positives (precision) [15].

Classification results

Table 2 presents the 5-fold classification performance of the developed DCNN models on both balanced and unbalanced datasets.

The models trained on the balanced dataset were excluded from the current investigations to maintain conciseness and focus. Henceforth, the models trained on the unbalanced dataset were exclusively referenced. Among these models, VG_BFCG exhibited the best performance; however, it did not meet all three of the specified criteria mentioned above.

Grad-Cam

Heatmaps were generated using GraD-CAM [52] to validate the model's predictions concerning the location of abnormalities. In

these heatmaps, each pixel's abnormality score is depicted by a color that reflects the gradient and prediction score. Figure 6 illustrates the heatmaps generated by VG_BFCG.

Ensemble learning

Various combinations of the developed models trained on the unbalanced dataset were investigated to identify the most suitable set of coefficients (weights) for these models. To determine the optimal combinations, various Designs of Experiments (DOE) were utilized at different levels. The qualified combinations that achieved an accuracy >0.975 are outlined in Table 3. The coefficients (β_j) for the models are reported in tuples, corresponding to the ordered list of models presented in Table 2.

Considering the high computational cost associated with using all models, a full factorial

Table 2: 5-fold Classification performance of the developed deep convolutional neural networks (DCNNs) models over balanced and unbalanced dataset

| Architecture | Balanced Dataset | | | | Unbalanced Dataset | | | |
| | 5-fold accuracy | | 5-fold NPV | | 5-fold accuracy | | 5-fold NPV | |
	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)
VG_BFCG	94.60	1.42	98.30	0.55	95.23	2.07	97.70	0.65
DN_BFCG	84.20	2.05	87.70	3.21	92.7	2.39	94.38	2.14
IRN_BFCG	86.20	0.37	92.50	1.14	90.07	1.08	97.02	1.03
MN_BFCG	89.00	1.70	92.10	1.79	88.40	1.66	95.08	1.59
X_BFCG	82.90	1.60	88.10	1.22	86.60	1.30	94.76	1.22
RN_BFCG	86.60	1.61	90.50	2.17	85.90	1.28	83.34	2.77

NPV: Negative Predictive Value, Std: standard deviation
BFCG (=batch normalization + fully connected + global average pooling) represents the structure of our proposed classifier.
VG_BFCG (= VGG16 + BFCG), DN_BFCG (= DenseNet-201 + BFCG), IRN_BFCG (= Inception-ResNet-v2 + BFCG), MN_BFCG (= MobileNetV2 + BFCG), X_BFCG (= Xception + BFCG), RN_BFCG (= ResNet152V2 + BFCG)
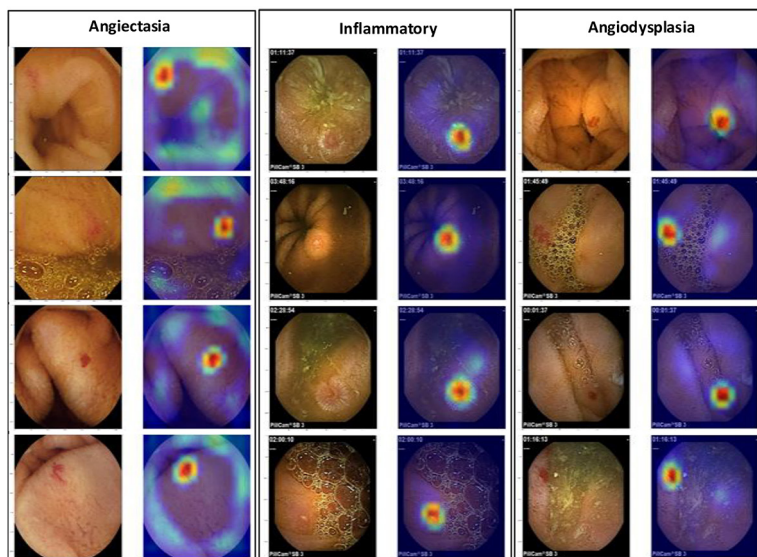
Figure 6: Grad-Cam Heat map visualization provided by the best-performed model; VG_BFCG. (BFCG (=batch normalization + fully connected + global average pooling) represents the structure of our proposed classifier. VG_BFCG (= VGG16 + BFCG))

DOE was initially performed with 9 levels for each factor (step size of 0.125) to identify the most significant models that produced the best accuracy. Subsequently, DOE was employed with finer distinction levels while utilizing only a selected set of models. The following combinations yielded the best results:

1. (VG_BFCG, DN_BFCG, IRN_BFCG, RN_BFCG)
2. (VG_BFCG, DN_BFCG, IRN_BFCG)

Class scoring was implemented over the top-qualified combinations listed in Table 3. For the four factors (classes), a full factorial DOE was performed with a step size of 0.0625. Table 4 details the best weights for the models and classes, with each 4-tuple representing the weights associated with Angiectasia, Inflammatory, Normal, and Angiodysplasia, respectively. The qualitative comparison provided in Table 5, proves the efficacy of our proposed tuned hybrid classifier.

Discussion

The incorporation of more Angiectasia frames in the unbalanced dataset (Table 1) resulted in improved classification metrics for all classes. The unbalanced dataset notably contributed to enhancing the NPV. For IRN_BFCG, it facilitated a more effective distinction between the Normal class and other pathologies. VG_BFCG, trained on the unbalanced dataset, showed the best classification performance, while models with more parameters, such as RN_BFCG, exhibited comparatively poorer performances, showing a complex DCNN is not imperative for lesion classification.

As illustrated in Figure 6, the highlighted regions demonstrate distinct prominence even in the presence of background areas exhibiting similar textures, showing VG_BFCG has been effectively trained to capture the crucial features specific to each pathology. The VG_BFCG achieves accurate classification of the Angiectasia and Angiodysplasia classes, as well as samples without any disease (Normal). The NPV assumes particular importance in the risk management of decision systems, as

Table 3: Hybrid classification performance on the test dataset in detail: weights of models in the top qualified combinations

No.	Coefficients	Acc	NPV	Precision	Recall	F1-score
1	(0.4, 0.2, 0.4, 0, 0, 0.12)	0.979	0.989	(1,0.97,0.99,0.95)	(0.99,0.97,1,0.96)	(0.99,0.97,0.99,0.95)
2	(0.625, 0.3125, 0.625, 0, 0, 0.1875)	0.979	0.989	(1,0.97,0.99,0.95)	(0.99,0.97,1,0.96)	(0.99,0.97,0.99,0.95)
3	(0.875, 0.4375, 0.875, 0, 0, 0.25)	0.979	0.989	(1,0.97,0.99,0.95)	(0.99,0.97,1,0.96)	(0.99,0.97,0.99,0.95)
4	(0.4, 0.15, 0.475, 0, 0, 0)	0.976	1.0	(1,0.96,1,0.93)	(0.99,0.97,0.99,0.96)	(0.99,0.96,0.99,0.95)
5	(0.45, 0.1, 0.525, 0, 0, 0)	0.976	1.0	(1,0.96,1,0.93)	(0.98,0.97,1,0.96)	(0.99,0.96,1,0.95)

ACC: Accuracy, NPV: Negative Predictive Value

Table 4: Tuned hybrid classification: coefficients of models and weight of classes yielding the highest evaluation metrics on the test dataset

No.	Coefficients	Class weight	Acc	NPV
1	(0.4, 0.2,0.4, 0.12)	(0.75, 0.25, 0.375, 0.4375)	0.9881	1.00
2	(0.625, 0.3125, 0.625,1875)	(0.75, 0.25, 0.375, 0.4375)	0.9881	1.00
3	(0.875, 0.4375, 0.875, 0.25)	(0.8125, 0.25, 0.375, 0.4375)	0.9881	1.00

ACC: Accuracy, NPV: Negative Predictive Value

false predictions of abnormal frames can lead to detrimental outcomes in clinical routines. Among the classification tasks examined in the present study, Angiodysplasia proved to be the most challenging pathology.

The hybrid classification using model scoring has shown the effectiveness of the developed models in accurately labeling frames. As evident in Table 3, the highest accuracy (greater than 0.978) was achieved. Complex DCNNs like Xception and MobileNet did not contribute significantly to the proposed hybrid classification. In Table 4, it is observed

that the accuracy was significantly enhanced by employing the class scoring mechanism. The accuracy increased from 0.978 to 0.9881, and the NPV increased from 0.989 to 1.00 using the class scoring technique. These results signify that the use of class scoring improved the performance metrics by enabling better distinction among different classes.

Using the combination of VG_BFCG, DN_BFCG, IRN_BFCG, and RN_BFCG yielded the best classification performance in hybrid classification as listed in Table 3. RN_BFCG and X_BFCG displayed poor classification

Table 5: A comprehensive comparison between our approach and related works.

Year	Study	Model	Dataset	Abnormality	Performance evaluation	Process Time/frame (s)
2021	Houdeville et al. [3]	CNN-based model	More than 1200 Pill-cam and MiroCam still frames (with or without angiectasias)	Angiectasias	Sensitivity = 97.4% Specificity = 98.8% NPV = 97.6%	0.021
2021	Caroppo et al. [53]	Three pre-trained DCNN (VGG19, InceptionV3 and ResNet5) + SVM	KID dataset (KID I + KID II)	Bleeding	Average accuracy I = 97.65% Average accuracy II = 95.70%	-
2022	Alam et al. [54]	CNN based architecture (RAt-CapsNet)	Kvasir Capsule	Ulcer + Erosion + Blood + Lymphangiectasis + Angiectasia	Accuracy (binary class) = 98.51% Accuracy (multi class) = 95.65%	-
2022	Vani et al. [55]	State-of-the-art CNN	Endoatlas and GastroLab database (2019)	Ulcer	Precision = 97.8% Recall = 97% Accuracy = 96.68% ROC = 0.84	-
2022	Vats et al. [56]	Multi-channel encoder-decoder network	Kvasir-Capsule + Computer Assisted Diagnosis for Capsule Endoscopy	Nine different pathologies	Sensitivity >94% Specificity >97% AUC >97% Accuracy >98% For nine abnormality classes	-
2023	Ours	Six pre-trained DCNNs, Hybrid classification (model scoring and class scoring)	GIANA + KID + Kvasir-Capsule	Angiectasia, Angiodysplasia, Inflammatory	Accuracy = 98.8% NPV = 100%	0.007

CNN: Convolutional Neural Network, NPV: Negative Predictive Value, DCNN: Deep Convolutional Neural Network, KID: KID series [33] SVM: Support Vector Machine, ROC: Receiver Operating Characteristic Curve, AUC = Area Under The ROC Curve, GIANA: GIANA [32]

performances. Consequently, these two classifiers do not contribute complementary distinction powers compared to the other four models. The combination of only the best classifiers can attain accuracy>0.975.

Ensemble learning has significantly improved the classification metrics (Tables 2-4). The implementation of class scoring in the top hybrid classifiers is remarkably effective in constructing an accurate decision boundary that aligns with hybrid classification tasks (Table 3).

The major contribution of this study lies in the introduction of a novel hybrid classification approach for a four-class task, with a specific focus on evaluating the NPV and F1-score. This aspect of the research has not been previously explored, adding novelty to the study. To further validate the performance of the models, heatmap visualization was conducted, providing compelling evidence of the models' robust feature learning and high accuracy.

The results obtained in this study highlight a notable difference from prior works, which have predominantly relied on private datasets and focused on binary classifications (Table 5). This limitation poses challenges in terms of generalizability. In contrast, the proposed algorithm demonstrates satisfactory performance and holds promise for future studies. However, it is important to acknowledge the limitations of this research. Firstly, the proposed DCNN models were trained using static frames, which may limit the applicability of our findings in clinical practice. To extrapolate the obtained results to real-world scenarios, further evaluation using full-length video data is necessary. This would provide a more comprehensive understanding of the algorithm's performance in dynamic contexts. Secondly, the use of a publicly available dataset restricted the ability to validate the developed models on a larger dataset with a greater number of patients. While the dataset used in this study was valuable for initial experimentation, future work should aim to validate the algorithm's performance on diverse datasets to ensure its robustness and generalizability.

Conclusion

This study presents an effective DCNN model designed to accurately differentiate diseases from CE images. Three public databases were leveraged, and among the models developed, VG_BFCG trained on the unbalanced dataset exhibited superior performance. Grad-Cam heat maps illustrated that VG_BFCG is particularly adept at extracting key features for each pathology.

Hybrid classification utilizing model scoring achieved an accuracy of 0.979 and NPV of 0.989. In the current study, while the six models introduced performed well individually, the fine-tuned ensemble structure using the class scoring mechanism led to increased accuracy by adjusting the class probabilities within the decision boundaries. The performance of the tuned hybrid classifier developed surpassed our goal, with an accuracy of 0.9881 and an NPV of 1.00. The approach in this work exhibits several notable highlights, as follows: 1) our dataset encompasses both major forms of lesions - vascular lesions and protruding lesions - providing a more comprehensive understanding of disease classification. This inclusion contributes to a more robust and accurate classification system and 2) we conducted a thorough evaluation of the decision boundaries drawn for the Normal class, specifically considering color and textural patterns. This analysis enabled the identification of areas with potential for improvement, thus enhancing the accuracy and reliability of our classification process. Lastly, through the use of class scoring, significant enhancements were achieved in all classification performance metrics. By assigning weights to each class during the decision-making process, we improved the overall performance of our hybrid classifier.

Authors' Contribution

In this study, E. Roodgar Amoli made significant contributions to the conceptualization, algorithm design, and simulation of the study. Additionally, E. Roodgar Amoli conducted data analysis and interpretation and also wrote the initial draft of the manuscript. A. Amiri Tehranizade contributed to the algorithm design and participated in the interpretation of the data. H. Arabalibeik played a supervisory role in this study and contributed to the editing of the manuscript. All authors, including E. Roodgar Amoli, A. Amiri Tehranizade, and H. Arabalibeik, actively participated in reading, modifying, and approving the final version of the manuscript. These contributions highlight the collaborative effort and expertise of the authors in executing and completing the study.

Ethical Approval

The Ethics Committee of Tehran University of Medical Sciences approved the protocol of the study (Ethic cod: IR.TUMS.MEDICINE.REC.1397.737).

Funding

Conflict of Interest

None

References

1. Ward EM, Sherman RL, Henley SJ, Jemal A, Siegel DA, Feuer EJ, et al. Annual Report to the Nation on the Status of Cancer, Featuring Cancer in Men and Women Age 20-49 Years. *J Natl Cancer Inst.* 2019;**111**(12):1279-97. doi: 10.1093/jnci/djz106. PubMed PMID: 31145458. PubMed PMCID: PMC6910179.

2. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. *Nature.* 2000;**405**(6785):417. doi: 10.1038/35013140. PubMed PMID: 10839527.

3. Houdeville C, Souchaud M, Leenhardt R, Beaumont H, Benamouzig R, McAlindon M, et al. A multisystem-compatible deep learning-based algorithm for detection and characterization of angiectasias in small-bowel capsule endoscopy. A proof-of-concept study. *Dig Liver Dis.* 2021;**53**(12):1627-31. doi: 10.1016/j.dld.2021.08.026. PubMed PMID: 34563469.

4. Nam JH, Hwang Y, Oh DJ, Park J, Kim KB, Jung MK, Lim YJ. Development of a deep learning-based software for calculating cleansing score in small bowel capsule endoscopy. *Sci Rep.* 2021;**11**(1):4417. doi: 10.1038/s41598-021-81686-7. PubMed PMID: 33627678. PubMed PMCID: PMC7904767.

5. Spada C, McNamara D, Despott EJ, Adler S, Cash BD, Fernández-Urién I, et al. Performance measures for small-bowel endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *Endoscopy.* 2019;**51**(6):574-98. doi: 10.1055/a-0889-9586. PubMed PMID: 31075800.

6. Guo X, Yuan Y. Semi-supervised WCE image classification with adaptive aggregated attention. *Med Image Anal.* 2020;**64**:101733. doi: 10.1016/j.media.2020.101733. PubMed PMID: 32574987.

7. Xiao Z, Feng LN. A study on wireless capsule endoscopy for small intestinal lesions detection based on deep learning target detection. *IEEE Access.* 2020;**8**:159017-26. doi: 10.1109/ACCESS.2020.3019888.

8. Karargyris A, Bourbakis N. Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. *IEEE Trans Biomed Eng.* 2011;**58**(10):2777-86. doi: 10.1109/TBME.2011.2155064. PubMed PMID: 21592915.

9. Lee YG, Yoon G. Real-time image analysis of capsule endoscopy for bleeding discrimination in embedded system platform. *World Acad Sci Eng Technol.* 2011;**59**:2526-30.

10. Ghosh T, Fattah SA, Wahid KA, Zhu WP, Ahmad MO. Cluster based statistical feature extraction method for automatic bleeding detection in wireless capsule endoscopy video. *Comput Biol Med.* 2018;**94**:41-54. doi: 10.1016/j.compbiomed.2017.12.014. PubMed PMID: 29407997.

11. Liaqat A, Khan MA, Shah JH, Sharif M, Yasmin M, Fernandes SL. Automated ulcer and bleeding classification from WCE images using multiple features fusion and selection. *Journal of Mechanics in Medicine and Biology.* 2018;**18**(04):1850038. doi: 10.1142/S0219519418500380.

12. Valério MT, Gomes S, Salgado M, Oliveira HP, Cunha A. Lesions multiclass classification in endoscopic capsule frames. *Procedia Computer Science.* 2019;**164**:637-45. doi: 10.1016/j.procs.2019.12.230.

13. Park SC, Chun HJ, Kim ES, Keum B, Seo YS, Kim YS, et al. Sensitivity of the suspected blood indicator: an experimental study. *World J Gastroenterol.* 2012;**18**(31):4169-74. doi: 10.3748/wjg.v18.i31.4169. PubMed PMID: 22919250. PubMed PMCID: PMC3422798.

14. Buscaglia JM, Giday SA, Kantsevoy SV, Clarke JO, Magno P, Yong E, Mullin GE. Performance characteristics of the suspected blood indicator feature in capsule endoscopy according to indication for study. *Clin Gastroenterol Hepatol.* 2008;**6**(3):298-301. doi: 10.1016/j.cgh.2007.12.029. PubMed PMID: 18255353.

15. Ghosh T, Chakareski J. Deep Transfer Learning for Automated Intestinal Bleeding Detection in Capsule Endoscopy Imaging. *J Digit Imaging.* 2021;**34**(2):404-17. doi: 10.1007/s10278-021-00428-3. PubMed PMID: 33728563. PubMed PMCID: PMC8290011.

16. Al Mamun A, Em PP, Ghosh T, Hossain MM, Hasan MG, Sadeque MG. Bleeding recognition technique in wireless capsule endoscopy images using fuzzy logic and principal component analysis. *International Journal of Electrical and Computer Engineering (IJECE).* 2021;**11**(3):2688-95. doi: 10.11591/ijece.v11i3.pp2688-2695.

17. Yuan Y, Meng MQ. Deep learning for polyp recognition in wireless capsule endoscopy images. *Med Phys.* 2017;**44**(4):1379-89. doi: 10.1002/mp.12147. PubMed PMID: 28160514.

18. Jia X, Xing X, Yuan Y, Xing L, Meng MQ. Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proceedings of the IEEE.* 2019;**108**(1):178-97. doi: 10.1109/JPROC.2019.2950506.

19. Rustam F, Siddique MA, Siddiqui HU, Ullah S, Mehmood A, Ashraf I, Choi GS. Wireless capsule endoscopy bleeding images classification using CNN based model. *IEEE Access.* 2021;**9**:33675-88. doi: 10.1109/ACCESS.2021.3061592.

20. Vallée R, De Maissin A, Coutrot A, Normand N, Bourreille A, Mouchère H. Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. 21st International Workshop on Multimedia Signal Processing (MMSP); Kuala Lumpur, Malaysia: IEEE; 2019. p. 1-5.

21. Tsuboi A, Oka S, Aoyama K, Saito H, Aoki T, Yamada A, et al. Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images. *Dig Endosc.* 2020;**32**(3):382-90. doi: 10.1111/den.13507. PubMed PMID: 31392767.

22. Hajabdollahi M, Esfandiarpoor R, Sabeti E, Karimi N, Soroushmehr SR, Samavi S. Multiple abnormality detection for automatic medical image diagnosis using bifurcated convolutional neural network. *Biomedical Signal Processing and Control.* 2020;**57**:101792. doi: 10.1016/j.bspc.2019.101792.

23. Leenhardt R, Vasseur P, Li C, Saurin JC, Rahmi G, Cholet F, et al. A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. *Gastrointest Endosc.* 2019;**89**(1):189-94. doi: 10.1016/j.gie.2018.06.036. PubMed PMID: 30017868.

24. Aoki T, Yamada A, Aoyama K, Saito H, Tsuboi A, Nakada A, et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest Endosc.* 2019;**89**(2):357-63.e2. doi: 10.1016/j.gie.2018.10.027. PubMed MID: 30670179.

25. Xiao Jia, Meng MQ. A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images. *Annu Int Conf IEEE Eng Med Biol Soc.* 2016;**2016**:639-42. doi: 10.1109/EMBC.2016.7590783. PubMed PMID: 28268409.

26. Fonseca F, Nunes B, Salgado M, Cunha A. Abnormality classification in small datasets of capsule endoscopy images. *Procedia Computer Science.* 2022;**196**:469-76. doi: 10.1016/j.procs.2021.12.038.

27. Smedsrud PH, Thambawita V, Hicks SA, Gjestang H, Nedrejord OO, Næss E, et al. Kvasir-Capsule, a video capsule endoscopy dataset. *Sci Data.* 2021;**8**(1):142. doi: 10.1038/s41597-021-00920-z. PubMed PMID: 34045470. PubMed PMCID: PMC8160146.

28. Afonso J, Mascarenhas M, Ribeiro T, Cardoso H, Andrade P, Ferreira JP, Saraiva MM, Macedo G. Deep Learning for Automatic Identification and Characterization of the Bleeding Potential of Enteric Protruding Lesions in Capsule Endoscopy. *Gastro Hep Advances.* 2022;**1**(5):835-43. doi: 10.1016/j.gastha.2022.04.008.

29. Saurin JC, Delvaux M, Gaudin JL, Fassler I, Villarejo J, Vahedi K, et al. Diagnostic value of endoscopic capsule in patients with obscure digestive bleeding: blinded comparison with video push-enteroscopy. *Endoscopy.* 2003;**35**(7):576-84. doi: 10.1055/s-2003-40244. PubMed PMID: 12822092.

30. Mascarenhas Saraiva MJ, Afonso J, Ribeiro T, Ferreira J, Cardoso H, Andrade AP, et al. Deep learning and capsule endoscopy: automatic identification and differentiation of small bowel lesions with distinct haemorrhagic potential using a convolutional neural network. *BMJ Open Gastroenterol.* 2021;**8**(1):e000753. doi: 10.1136/bmjgast-2021-000753. PubMed PMID: 34580155. PubMed PMCID: PMC8477239.

31. Gueye L, Yildirim-Yayilgan S, Cheikh FA, Balasingham I. Automatic detection of colonoscopic anomalies using capsule endoscopy. IEEE international conference on image processing (ICIP); Canada: IEEE; 2015. p. 1061-4.

32. Bernal J, Aymeric H. Gastrointestinal image analysis (giana) angiodysplasia d&I challenge. Available from: https://endovissub2017-giana.grand-challenge.org/home/. 2017

33. Koulaouzidis A, Iakovidis DK, Yung DE, Rondonotti E, Kopylov U, Plevris JN, et al. KID Project: an internet-

based digital video atlas of capsule endoscopy for research purposes. *Endosc Int Open.* 2017;**5**(6):E477-83. doi: 10.1055/s-0043-105488. PubMed PMID: 28580415. PubMed PMCID: PMC5452962.

34. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning [Internet]. arXiv [Preprint]. 2017 [cited 2017 Dec 13]. Available from: https://arxiv.org/abs/1712.04621.

35. Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters.* 2017;**24**(3):279-83. doi: 10.1109/LSP.2017.2657381.

36. Pritt M, Chern G. Satellite image classification with deep learning. Applied imagery pattern recognition workshop (AIPR); Washington, DC, USA: IEEE; 2017. p. 1-7.

37. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision; ICCV; 2015. p. 1026-34.

38. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey. *WIRs: Data Mining and Knowledge Discovery.* 2018;**8**(4):e1253. doi: 10.1002/widm.1253.

39. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2016. p. 779-88.

40. Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision; CVPR; 2015. p. 1440-8.

41. Chen J, Li K, Deng Q, Li K, Philip SY. Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics.* 2019. doi: 10.1109/TII.2019.2909473.

42. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng.* 2018;**2**(10):741-8. doi: 10.1038/s41551-018-0301-3. PubMed PMID: 31015647.

43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Internet]. arXiv [Preprint]. 2014 [cited 20214 Sep 4]. Available from: https://arxiv.org/abs/1409.1556.

44. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2017. p. 4700-8.

45. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17); AAAI; 2017. p. 4278-84.

46. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2018. p. 4510-20.

47. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2016. p. 770-8.

48. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference; Amsterdam, Netherlands: Springer; 2016. p. 630-45.

49. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2017. p. 1251-8.

50. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus) [Internet]. arXiv [Preprint]. 2015 [cited 2015 Nov 23]. Available from: https://arxiv.org/abs/1511.07289.

51. Zeiler MD. Adadelta: an adaptive learning rate method [Internet]. arXiv [Preprint]. 2012 [cited 2012 Dec 22]. Available from: https://arxiv.org/abs/1212.5701.

52. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision; CVPR; 2017. p. 618-26.

53. Caroppo A, Leone A, Siciliano P. Deep transfer learning approaches for bleeding detection in endoscopy images. *Comput Med Imaging Graph.* 2021;**88**:101852. doi: 10.1016/j.compmedimag.2020.101852. PubMed PMID: 33493998.

54. Alam MJ, Rashid RB, Fattah SA, Saquib M. RAt-CapsNet: A Deep Learning Network Utilizing Attention and Regional Information for Abnormality Detection in Wireless Capsule Endoscopy. *IEEE J Transl Eng Health Med.* 2022;**10**:3300108. doi: 10.1109/JTEHM.2022.3198819. PubMed PMID: 36032311. PubMed PMCID: PMC9401095.

55. Vani V, Prashanth KM. Ulcer detection in Wireless Capsule Endoscopy images using deep CNN. *Journal of King Saud University-Computer and Information Sciences.* 2022;**34**(6):3319-31. doi: 10.1016/j.jksuci.2020.09.008.

56. Vats A, Raja K, Pedersen M, Mohammed A. Multichannel residual cues for fine-grained classification in wireless capsule endoscopy. *IEEE Access.* 2022;**10**:91414-23. doi: 10.1109/ACCESS.2022.3201515.