RESEARCH ARTICLE

VOL. 46 | **SPECIAL ISSUE SI TEC 2023 2025** | PP 1-17

Revista Mexicana de Ingeniería Biomédica

https://dx.doi.org/10.17488/RMIB.46.SI-TAIH.1526

E-LOCATION ID: e1526

Classification of Small Bowel Lesions in Video Capsule Endoscopy Images Using Enhanced Deep Learning Model

Clasificación de Lesiones del Intestino Delgado en Imágenes de Endoscopia por Videocápsula Utilizando un Modelo Mejorado de Aprendizaje Profundo

Erik Orlando Cuevas Rodríguez[1], *Carlos Eric Galván-Tejada*[1], *Rafael Magallanes Quintanar*[1], *Jorge Issac Galván Tejada*[1], *José María Celaya Padilla*[1], *Juan Rubén Delgado Contreras*[1]

[1] Universidad Autónoma de Zacatecas, Zacatecas - México

ABSTRACT

Video capsule endoscopy (VCE) is a noninvasive procedure for diagnosing small bowel (SB) lesions. During VCE procedures, the miniature camera captures thousands of images, requiring considerable time and effort from healthcare professionals to examine each image for abnormalities. This study aims to develop and evaluate a modified DenseNet-201 convolutional neural network (CNN) model to assist in the automatic classification of SB lesions in VCE images. A representative dataset of 5,899 images was created by merging two state-of-the-art datasets, including six types of lesions and one class without lesions. A synthetic oversampling approach generated 1,273 synthetic images for the underrepresented class. Synthetic image quality was evaluated using texture metrics. The experiment was conducted on a set of 7,172 images using data partitions of 70 %, 15 %, and 15 % for training, validation, and testing, respectively, with a Monte Carlo cross-validation to verify the consistency of the experimental results. The model achieved an accuracy of 89.78 %, a precision of 89.66 %, a sensitivity of 89.69 %, a specificity of 89.72 %, and an F1 score of 89.67 % in the test set. The modified DenseNet-201 CNN model could be a valuable tool for diagnosing SB conditions and improving patient outcomes.

KEYWORDS: **convolutional neural network, deep learning, small bowel lesions, synthetic oversampling, video capsule endoscopy**

RESUMEN

La endoscopia por videocápsula (VCE, por sus siglas en inglés) es un procedimiento no invasivo para diagnosticar lesiones del intestino delgado (SB). Durante los procedimientos de VCE, la cámara en miniatura captura miles de imágenes, lo que requiere mucho tiempo y esfuerzo por parte de los profesionales de la salud para examinar cada imagen en busca de anomalías. El objetivo de este estudio es desarrollar y evaluar un modelo modificado de red neuronal convolucional (CNN) DenseNet-201 para ayudar en la clasificación automática de las lesiones del intestino delgado en imágenes de VCE. Se creó un conjunto de datos representativo de 5,899 imágenes mediante la combinación de dos conjuntos de datos del estado del arte, incluyendo seis tipos de lesiones y una clase sin lesiones. Un método de sobremuestreo sintético generó 1,273 imágenes sintéticas para la clase infrarrepresentada. La calidad de las imágenes sintéticas se evaluó mediante métricas de textura. El experimento se llevó a cabo con un conjunto de 7,172 imágenes utilizando particiones de datos del 70 %, 15 %, y 15 % para entrenamiento, validación y prueba, respectivamente, con una validación cruzada Monte Carlo para verificar la coherencia de los resultados experimentales. El modelo alcanzó una exactitud del 89.78 %, una precisión del 89.66 %, una sensibilidad del 89.69 %, una especificidad del 89.72 %, y una puntuación F1 del 89.67 % en el conjunto de pruebas. El modelo CNN DenseNet-201 modificado podría ser una valiosa herramienta para diagnosticar afecciones del SB y mejorar los resultados de los pacientes.

PALABRAS CLAVE: aprendizaje profundo, endoscopia con videocápsula, lesiones del intestino delgado, red neuronal convolucional, sobremuestreo sintético

Corresponding author

TO: CARLOS ERIC GALVÁN TEJADA

INSTITUTION: UNIVERSIDAD AUTÓNOMA DE ZACATECAS

ADDRESS: CARRETERA ZACATECAS-GUADALAJARA KM. 6, EJIDO
LA ESCONDIDA. ZACATECAS, C.P. 98160, MÉXICO.

EMAIL: ericgalvan@uaz.edu.mx

Received:

21 February 2025

Accepted:

07 April 2025

Published:

21 September 2025

INTRODUCTION

The small bowel (SB) is an organ of the gastrointestinal (GI) tract that can be affected by various chronic diseases such as celiac disease, Crohn's disease (CD), occult gastrointestinal bleeding, small bowel neoplasms, and hereditary polyposis syndromes[1]. Traditional endoscopy and colonoscopy are commonly used diagnostic procedures to investigate the GI tract. Despite the relative effectiveness of these methods, in some cases, lesions in less accessible areas can sometimes go undetected. Studies have shown that in 77 % of these cases, the lesions are found in the SB, a region difficult to reach with traditional endoscopic techniques[2].

In 2001, the Food and Drug Administration (FDA) approved the use of an endoscopic videocapsule to diagnose small intestinal lesions[3]. Since then, videocapsule endoscopy (VCE) has become a first-line diagnostic procedure in the study of SB disorders[4]. VCE is an accurate clinical tool for diagnosing and monitoring various SB lesions[5]. It involves a small, swallowable capsule with a miniature camera that captures thousands of images as it passes through the digestive system, providing a solution to the limitations of traditional endoscopy methods. These images should be studied by a healthcare professional to identify any abnormalities or lesions that require immediate medical attention. The VCE procedure presents several challenges. First, the volume of images can be overwhelming, as the capsule typically captures approximately 50,000 to 60,000 images during its journey. This requires significant time and effort from healthcare professionals to meticulously examine each image for possible abnormalities. Second, the quality of images can be affected by factors such as rapid movement, inadequate lighting, or the presence of food residue, which can complicate the visibility of the mucosal surface. Additionally, interpreting these images requires expertise and experience, as subtle lesions can be easily missed[6]. To address these challenges, researchers and developers in medical imaging and healthcare technology have explored ways to automate and streamline the analysis process.

An effective alternative to traditional manual image analysis is the use of artificial intelligence (AI). These advanced technologies can quickly process large volumes of images, identify abnormalities with high precision, and reduce the workload of healthcare professionals. In addition, it can provide consistent results without the variability that can occur with human analysis, reducing human error and possible oversight. Deep learning (DL), a branch of AI, has shown effective results in medical image analysis, leading to the widespread adoption of convolutional neural network (CNN) algorithms due to their simplicity in handling medical images[7][8][9][10]. These algorithms achieve excellent performance in the detection of a variety of VCE pathologies, including ulcers, polyps, celiac disease, and bleeding. For example, Tariq Rahim *et al.*[11] and Andrea Caroppo *et al.*[12], developed methods to identify ulcers and SB bleeding by extracting features from color maps. Xinle Wang *et al.*[13], classified celiac disease using a combination of CNN algorithms. Younghak Shin *et al.*[14], Alba Nogueira Rodriguez *et al.*[15], and Meryem Souaidi *et al.*[16] have all worked on detecting and classifying polyps in endoscopic images. In addition, several works[17][18][19][20][21], focus on the detection and classification of different pathologies, such as intestinal bleeding, angioectasias, polyps, and inflammatory lesions.

These approaches use robust algorithms to analyze endoscopic images for tasks such as lesion detection, segmentation, and classification, significantly improving the accuracy and efficiency of diagnosis. Despite these advances, challenges remain, particularly in the availability of large datasets that accurately represent the wide range of pathologies associated with SB diseases. Some efforts have been made to achieve this. Pia H. Smedsrud *et al.*[22], labeled and medically verified 47,238 VCE images from fourteen classes of findings, including anatomical landmarks, luminal, and pathological findings in the Kvasir-Capsule dataset. In addition, Vallée *et al.*[23], provided CrohnIPI, a labeled and medically verified dataset with 3,498 VCE images that contain various CD-associated lesions. Unfortunately, the number of CrohnIPI images is limited when automatic CD lesion classification is developed.

As far as we know, no study has developed a multiclass classification model on the CrohnIPI dataset, because of the lack of images in certain classes. Handling multilabel models and training them to detect and identify SB lesions presents challenges and opportunities in medical image analysis. The complexity arises from different healthcare experts providing feedback, and there could be uncertainty or ambiguity when labeling certain lesions.

The present study implements a modified DenseNet-201 CNN model to classify different SB lesions extracted from the Kvasir-Capsule and CrohnIPI datasets. This research aims to create a reliable model that accurately classifies various SB lesions on VCE images. By achieving high accuracy, this model could be integrated into a diagnostic system that can assist endoscopists in accurately identifying various SB lesions.

The rest of the work is structured as follows: Section 2 discusses the methodology, detailing the techniques, tools, and resources used. Section 3 contains the results. Section 4 discusses the results. Section 5 presents the conclusions. Section 6 describes future work, and finally, section 7 reports acknowledgments.

MATERIALS AND METHODS

The methodology (Figure 1) consists of three stages. Data preparation is the first stage, which includes a brief description of the database elements, the approach to working with the data, the preprocessing steps, and the applied data augmentation techniques. The second stage involves training and validation of the DenseNet-201 architecture. This stage contains different strategies for classifying SB-associated lesions, and the configuration of hyperparameters, such as optimizer, learning rate, batch size, and number of epochs, along with the tools used to implement the CNN model. Finally, in the last stage, the performance of the model is evaluated using various metrics.

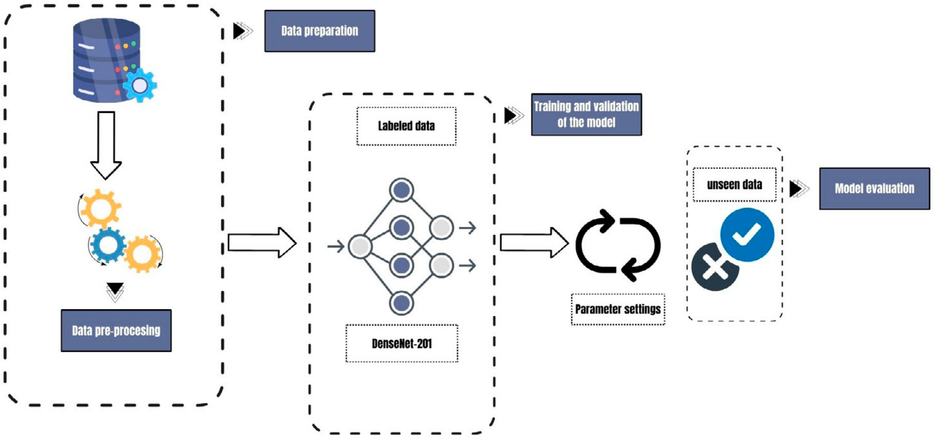


FIGURE 1. **Diagram of the methodology.**

Data preparation

The data preparation stage was divided into four phases (Figure 2). The first phase reviews the databases from which the images were obtained. The second phase addresses the techniques used to split the data. The third phase implements techniques to manage data imbalance. In the fourth and final phase, various transformation techniques are used to enhance the generalization of the architecture.

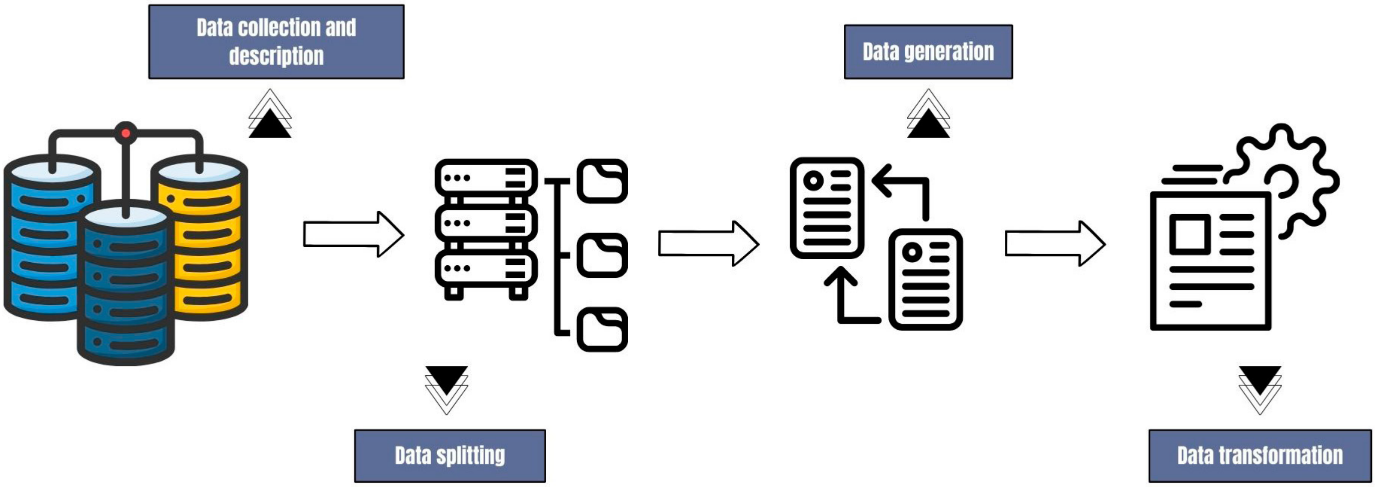
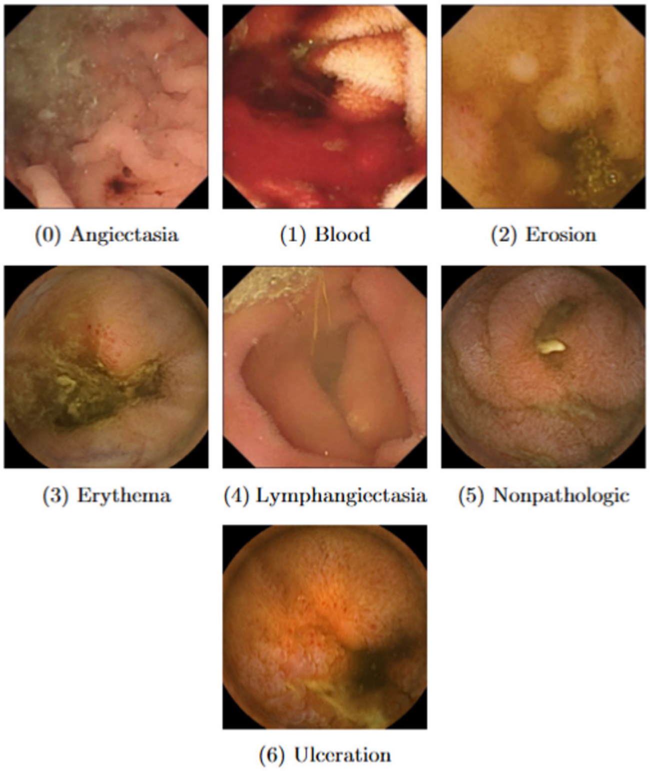


FIGURE 2. **Diagram of data preparation.**

Data collection and description

The Kvasir-Capsule database contains 47,238 labeled images extracted from videos collected from clinical examinations conducted in the Department of Medicine, Bærum Hospital, Vestre Viken Hospital Trust in Norway, using the Olympus EC-S10 endocapsule with a resolution of 336 × 336 pixels. It has two main categories with their corresponding classes according to the Minimal Standard Terminology of the World Endoscopy Association[22].

The category of anatomy contains anatomical landmarks characterizing the GI tract. The category of luminal findings includes the content of the bowel lumen, the aspect of the mucosa, and pathological findings. The Crohn IPI database contains images captured via PillCam3 from 63 patients from Nantes University Hospital[23]. The resolution of these images is 320 × 320 pixels. In total, 3,484 images were annotated, each image was reviewed independently of the previous and following images without any idea of the patient undergoing the examination. The integration of these two datasets allows for the exploitation of their complementary strengths, thereby enhancing the accuracy and reliability of the classification model. Figure 3 shows an example of each database class used: Angiectasia, blood, erosion, erythema, lymphangiectasia, ulceration, and nonpathological.



(0) Angiectasia (1) Blood (2) Erosion

(3) Erythema (4) Lymphangiectasia (5) Nonpathologic

(6) Ulceration

FIGURE 3. **Database classes**

The two databases were merged to create a new database that exclusively consists of SB lesions with 5,899 images. The selection process involved choosing the most representative classes with SB lesions to increase the variety of lesions. Table 1 presents the number of images per database class.

TABLE 1. **Class distribution.**

Class	Number of images	
Angiectasia	866	
Blood	446	
Erosion	506	
Erythema	284	
Lymphangiectasia	592	
Nonpathological	2124	
Ulceration	1081	
	Total	5899

Data splitting

The dataset, consisting of 5,899 images available, was divided into training, validation, and test sets according to the distribution shown in Table 2. Each image was randomly assigned to one of the three sets, ensuring consistency of images across sets.

TABLE 2. **Data splitting details.**

<table>
<tr><th>Set</th><th>Percentage of the images (%)</th><th colspan="2">Total number of images</th></tr>
<tr><td>Training</td><td>70</td><td colspan="2">4129</td></tr>
<tr><td>Validation</td><td>15</td><td colspan="2">885</td></tr>
<tr><td>Testing</td><td>15</td><td colspan="2">885</td></tr>
<tr><td></td><td></td><td>Total</td><td>5899</td></tr>
</table>

The partition followed a 70:15:15 ratio for the training, validation, and test sets. Furthermore, repeated random subsampling (Monte Carlo cross-validation) was used for validation. This process was repeated three times with each iteration involving a different random split.

Data generation

Before handling data imbalance, a resizing method is implemented to ensure that the synthetic samples are generated in the resized feature space. The bilinear interpolation method[24] is used to resize the original images to 224 × 224 × 3 pixels, a size congruent with the DenseNet-201 model[25]. To address the class imbalance, the Adaptive Synthetic (ADASYN) sampling approach[26] was applied. For each minority class data example x_i generate s_i synthetic data examples by randomly selecting one minority data example x_{zi}, from the K nearest neighbors for data x_i via Equation 1.

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \tag{1}$$

where (x_{zi}-x_i) is the difference vector in n-dimensional spaces, and λ is a random number: $\lambda \in [0,1]$. The result of applying ADASYN using a minority sampling strategy generated a new training dataset with 5,402 images. The augmented dataset result of the implemented ADASYN technique is shown in Table 3.

TABLE 3. **Class distributions in the training set after the ADASYN oversampling method.**

<table>
<tr><th>Class</th><th colspan="2">Number of images</th></tr>
<tr><td>Angiectasia (class 0)</td><td colspan="2">568</td></tr>
<tr><td>Blood (class 1)</td><td colspan="2">316</td></tr>
<tr><td>Erosion (class 2)</td><td colspan="2">357</td></tr>
<tr><td>Erythema (class 3)</td><td colspan="2">1480</td></tr>
<tr><td>Lymphangiectasia (class 4)</td><td colspan="2">412</td></tr>
<tr><td>Nonpathological (class 5)</td><td colspan="2">1501</td></tr>
<tr><td>Ulceration (class 6)</td><td colspan="2">768</td></tr>
<tr><td></td><td>Total</td><td>5402</td></tr>
</table>

With this method, 1,273 synthetic images were created for erythema, the class with the lowest representation of the dataset. Figure 4 shows three examples of synthetic images generated from the minority class.

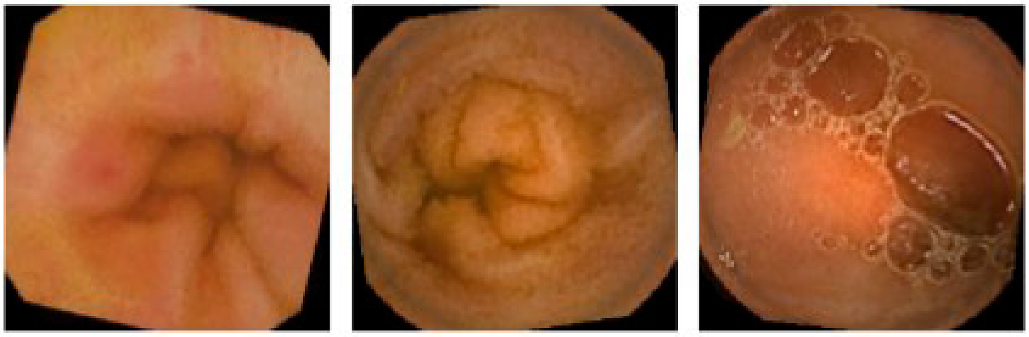


FIGURE 4. **Synthetic images from the minority class using the ADASYN method.**

Data transformation

In addition, various data augmentation techniques were applied. The list below shows the various augmentation techniques, and their parameters used in this work.

- Random rotation of 15 degrees.
- Vertical and horizontal flipping with a probability of 0.5.

These techniques help to improve generalization, robustness to orientation variations, and the ability of the model to handle data in real-world conditions, which is key in VCE image classification. Finally, the dataset has a total of 7,172 images, of which 5,402 are for training, and 885 for validation and testing purposes, respectively.

Training and validation of the model

A CNN model is composed of several fundamental layers, each with a specific function in image processing. The basic structure is illustrated in Figure 5.

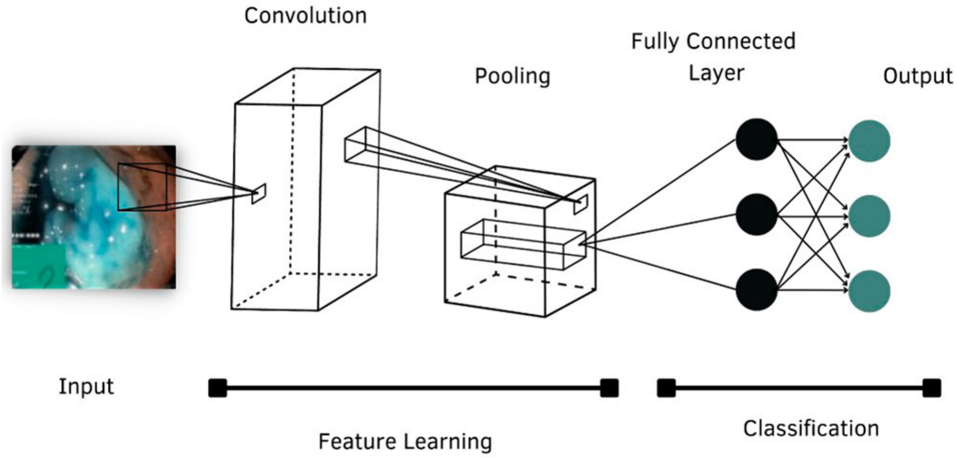


FIGURE 5. **Architecture of a convolutional neural network [27].**

The convolution layer performs a dot product between the input x and the weights W^k and adds a bias b^k. Next, a nonlinearity or activation function is applied to the output of the convolution layer, resulting in k feature maps h^k as shown in Equation 2.

$$h^k = f(W^k \cdot x + b^k) \tag{2}$$

The pooling layer downsamples each feature map in the subsampling layers, reducing network parameters, accelerating training, and preventing overfitting. The mathematical expression of using max pooling can be observed in Equation 3.

$$O^k = max(W^k \cdot x + b^k) \tag{3}$$

Finally, in the classification stage, a loss function is used in the output layer to calculate the prediction error over the training samples. This error reveals the difference between the actual and expected outputs. The Softmax function estimates the probability of belonging to a class, and its output is the probability of $p \epsilon \{0,1\}$ as shown in Equation 4.

$$p_i = \frac{e^{ai}}{\sum_{k=1}^{N} e_k^a} \tag{4}$$

where e^{ai} represents the unnormalized output of the previous layer, while N represents the number of neurons in the output layer. Finally, the mathematical representation of the cross-entropy loss function is Equation 5.

$$H(p, y) = -\sum y_i \ log(p_i) \tag{5}$$

where y_i is the true label for class i. Cross-entropy is a key metric in DL to measure how well a model classifies samples. It is commonly used in neural networks with Softmax in the output.

DenseNet-201 model

This architecture is 201 layers deep and works with twenty million parameters. Fewer parameters are needed than conventional CNNs because they do not need nonessential feature maps[28]. To preserve the feedforward nature, each layer obtains additional inputs from all preceding layers and passes its feature maps to all subsequent layers [25]. This is called dense connectivity, where direct connections from any layer to all subsequent layers are produced. Figure 6 schematically illustrates the layout of the resulting dense connectivity.

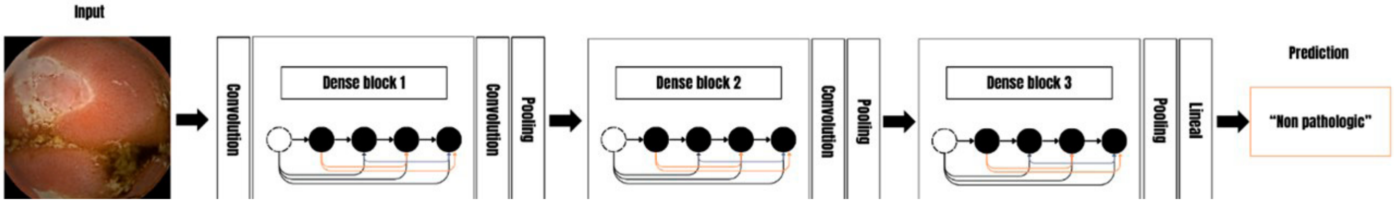


FIGURE 6. Deep DenseNet with three dense blocks.

Hyperparameter configuration

The following parameters were used for training the network: 200 epochs; batch size 32; Stochastic Gradient Descent with Momentum (SGDM) optimization with a learning rate of 0.001. The model has been pre-trained on ImageNet[29]. In addition, the DenseNet architecture was modified by appending additional layers to its classifier module. Specifically, inserting a fully connected layer with 1024 neurons followed by a LeakyReLu activation function with a negative slope of 0.001 to introduce nonlinearity. Dropout regularization with 0.2 dropout probability was applied to prevent overfitting. Finally, another fully connected layer was added to produce the predictions corresponding to the number of classes in the classification task. Figure 7 shows the custom architecture.

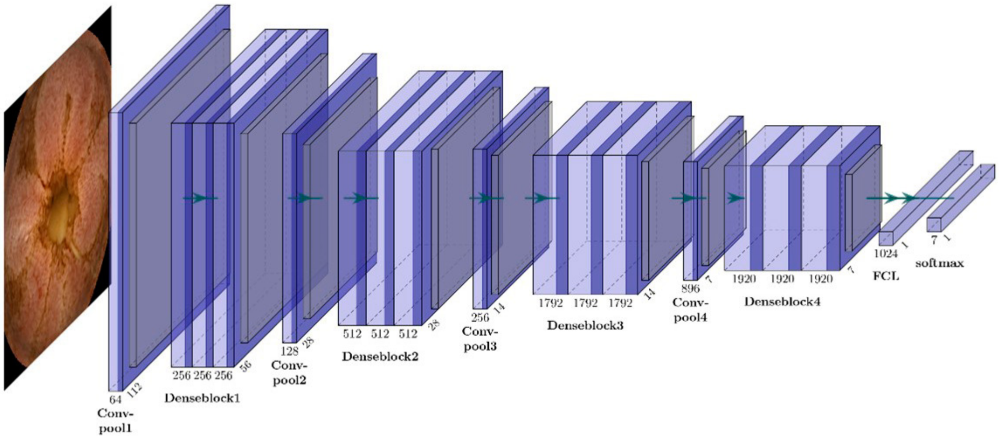


FIGURE 7. **Custom DenseNet-201 architecture.**

Figure 7 illustrates how the spatial dimensions of each layer decrease while the number of feature maps increases. The total number of feature maps F_l in each block at the corresponding l^{th} output layer is calculated with Equation 6.

$$F_l = F_0 + L_l \times k \tag{6}$$

where F_0 is the initial number of feature maps, L_l is the number of layers in each block, and k is the growth rate. This equation highlights how the network becomes deeper by adding new feature maps at each layer, contributing to the total F_l. The growth rate k=32 is constant, but the total number of feature maps increases as layers are added within each dense block and decreases by half in each transition layer.

Experimental setup

The model was implemented in Python version 3.11.2, utilizing PyTorch version 2.2.1 with CUDA support and torchvision version 0.17.1 for deep learning and computer vision tasks. An AMD Ryzen 9 processor was employed, and a single GeForce RTX 3060 served as the graphics processing unit.

Evaluation of model performance

To comprehensively evaluate the performance of the model on unseen data, a range of metrics was used, including micro averaging, macro averaging, and weighted averaging. In addition, the confusion matrix was analyzed to assess the accuracy of the model by determining the proportion of correctly classified instances for each class.

Finally, the area under the receiver operating characteristic (ROC) curve was calculated as a key metric to evaluate the overall performance of the model.

RESULTS AND DISCUSSION

The model was trained using the augmented training set. By plotting the loss curve over 200 epochs, the training and validation progress of the model was monitored. Figure 8 shows the evolution of the loss function in each epoch of the training process.

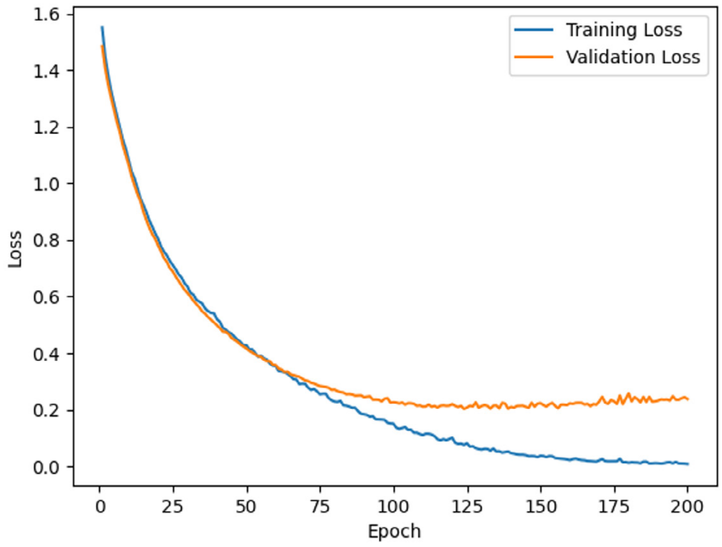


FIGURE 8. **Evolution of the loss curve in the training and validation sets for each training epoch.**

After epoch 100, the validation loss exhibited minor fluctuations, and finally, at epoch 125, the minimum value of 0.3759 was reached. The results obtained on test set images are presented in Table 4. The micro, macro, and weighted averages were calculated.

TABLE 4. **Performance metric results.**

Metric	Micro average	Macro average	Weighted average
Accuracy	0.9016	0.9016	0.8978
Precision	0.9016	0.8907	0.8966
Recall	0.9016	0.8695	0.8969
Specificity	0.9016	0.8788	0.8972
F1-score	0.9016	0.8889	0.8967

The micro average results for all the metrics, including accuracy, precision, recall, specificity, and F1-score, were all 0.9016. For the macro average, the accuracy remained at 0.9016. However, the precision decreased slightly to 0.8907, the recall to 0.8695, specificity to 0.8788, and F1-score to 0.8889. The weighted average results indicate an accuracy of 0.8978, precision of 0.8966, recall of 0.8969, specificity of 0.8972, and F1-score of 0.8967. The model was evaluated regarding the area under the curve (AUC), accuracy, sensitivity, and specificity using the test set with a Monte Carlo cross-validation with three iterations. In addition, measures of central tendency are used, such as mean and standard deviation. Table 5 shows the results of this trial.

TABLE 5. **Monte Carlo cross-validation results.**

Iteration	AUC	Accuracy	Sensitivity	Specificity
1	0.95	0.8926	0.8925	0.8926
2	0.96	0.8994	0.8966	0.8994
3	0.97	0.9016	0.9016	0.8996
Mean	0.96	0.8978	0.8969	0.8972
Standard deviation	0.0082	0.0038	0.0037	0.0032

The mean values of AUC, accuracy, sensitivity, and specificity were 0.96, 0.8978, 0.8969, and 0.8972, respectively. The standard deviation values for AUC, accuracy, sensitivity, and specificity were 0.0082, 0.0038, 0.0037, and 0.0032, respectively. A confusion matrix was constructed to analyze the performance of the model in more detail. Figure 9 illustrates the confusion matrix that resulted before and after applying the ADASYN method.

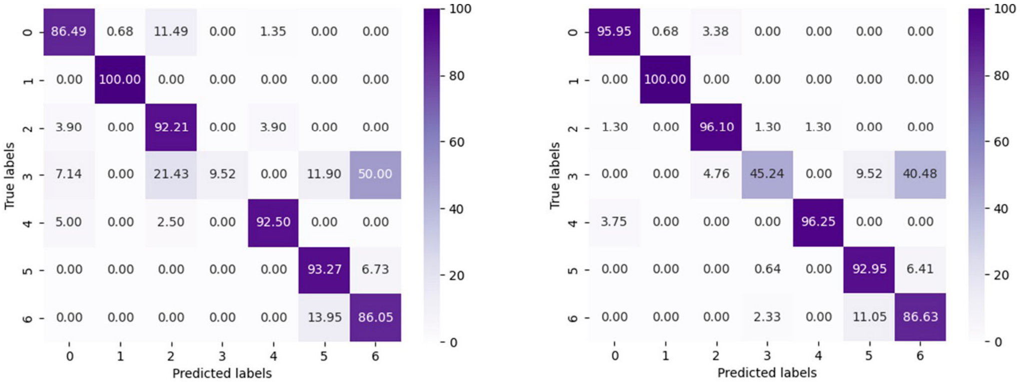


FIGURE 9. **DenseNet-201 model confusion matrix.**

Figure 9 shows the percentage of correct and incorrect predictions for each model class before and after applying the ADASYN method. Before applying the ADASYN method, for class 0 (angiectasia), the model achieved moderate accuracy, correctly classifying 86.49 % of true class 0 instances. Misclassifications were considerable, with 11.49 % predicted as class 2. Class 1 (blood) showed perfect performance with 100 % accuracy. Class 2 (erosion) had 92.21 % accuracy, with minimal misclassifications in classes 0 (3.90 %) and 3 (3.90 %). For class 3 (erythema), the accuracy decreased to 9.52 %, with notable misclassifications into classes 0 (7.14 %), 2 (21.43 %), 5 (11.90 %), and 6 (50.00 %). Class 4 (lymphangiectasia) displayed a high accuracy of 92.50 %, with errors in class 0 (5.00 %) and class 2 (2.50 %). Class 5 (nonpathological) had a correct classification rate of 93.27 %, with some instances misclassified in class 6 (6.73 %). For class 6 (ulceration), the accuracy was 86.05 %, with errors in class 5 (13.95 %). After applying the ADASYN method, for class 0, the model achieved a high accuracy with 95.95 % of true class 0 instances. Misclassifications were minimal, with 3.38 % predicted as class 2. Class 1 also showed strong performance, with 100 % accuracy. In class 2, 96.10 % of the instances were correctly classified, but there were misclassifications in classes 1, 3, and 4, each at 1.30 %. Class 3 had a lower correct classification rate of 45.24 %, with significant misclassifications into classes 2 (4.76 %), 5 (9.52 %), and 6 (44.48 %). Class 4 contained 96.25 % accuracy, with small errors in class 1 (3.75 %). Class 5 also had a high accuracy of 92.95 %, with some misclassifications in class 6 (6.41 %). For class 6, the correct classification rate was 86.63 %, with errors in classes 3 (2.33 %) and 5 (11.05 %). In addition, the area under the ROC curve (Figure 10) was measured.

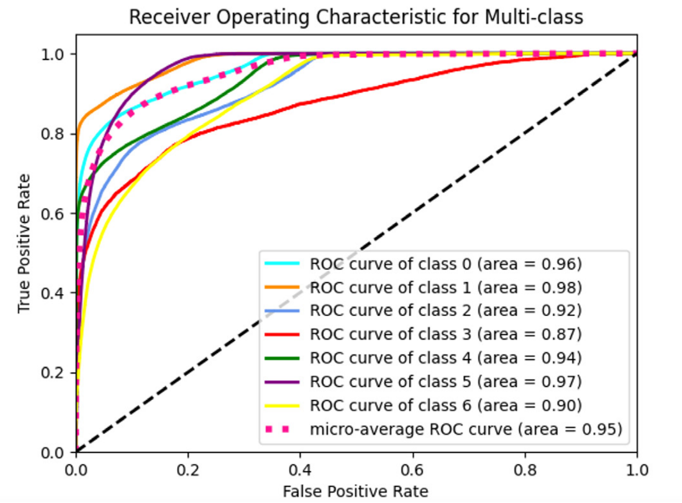


FIGURE 10. **Multiclass ROC curve.**

Class 0 had an AUC of 0.96. Class 1 had an AUC of 0.98. Class 2 had an AUC of 0.92. Class 3 had an AUC of 0.87. Class 4 had an AUC of 0.94. Class 5 had an AUC of 0.97. Class 6 with an AUC of 0.90. The micro average ROC curve had an AUC of 0.95. The quality of synthetic images was evaluated by comparing texture features such as correlation, contrast, and homogeneity between images. Table 6 compares an original image of erythema with a synthetic image.

TABLE 6. **Texture feature comparison results.**

Metric	Correlation	Contrast	Homogeneity
Erythema original image	0.9559	117.9898	0.3368
Erythema synthetic image	0.9588	108.6636	0.3680

The correlation values for the original and synthetic images were 0.9559 and 0.9588, respectively. The contrast of the original images was 117.9898, while the synthetic images exhibited a contrast of 108.6636. The homogeneity for the original image was 0.3368, whereas the synthetic images demonstrated a higher homogeneity of 0.3680. In addition, the color dispersion was analyzed through its graph. Figure 11 shows the color distribution of the original erythema and the synthetic image.

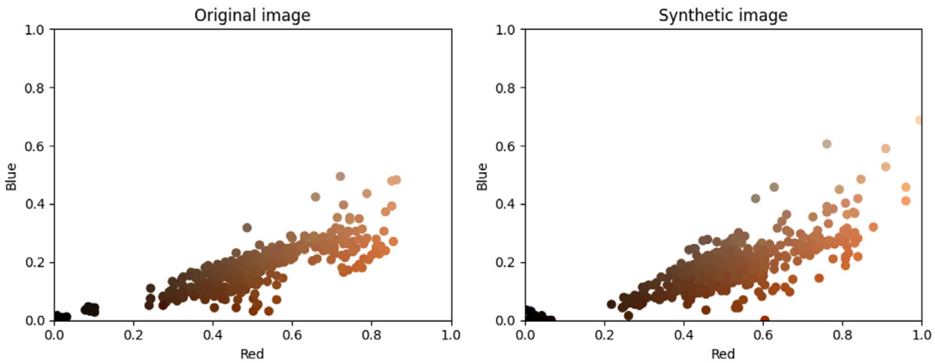


FIGURE 11. **Color scatter plot between erythema images.**

The synthetic image, while similar, showed a slightly wider spread, particularly in the higher-intensity regions.

This work proposed a modified DenseNet-201 for the classification of VCE images to identify various types of SB lesions in a custom dataset. The main advantage of the proposed work and other reference works mentioned above is the lesion-level classification. This is a more accurate way to diagnose VCE lesions. Even with a limited number of images, the model can classify different lesions with 89.78 % accuracy. By merging the two databases mentioned in Section 2, a more diverse dataset featuring various lesions that could be encountered in a VCE examination was developed. An ADASYN oversampling method was implemented to address the lack of images in the minority class. By selectively applying the minority sampling strategy, we focused only on generating synthetic samples for the erythema class, rather than creating synthetic samples for all classes to maintain dataset balance. This approach not only conserves resources but also preserves the natural distribution of the original database, thereby enhancing the validity of our results. This improved the performance of the model in classifying the minority class by 36 %. Figure 9 shows how the ADASYN method helps balance accuracy between classes, leading to more consistent performances in different types of lesions. Since occult bleeding is the main indication for VCE, the model demonstrates exceptional accuracy in classifying this lesion, achieving 100 % correct predictions. This high accuracy highlights the potential utility of the model in clinical settings, particularly in the detection and diagnosis of occult bleeding. However, erythema and ulceration are the two classes with fewer than 90 % correctly classified instances. The model has more difficulty distinguishing these classes than others, possibly due to the subtle and often overlapping characteristics of erythema and ulcerations in VCE images. Erythema usually serves as an early indicator of inflammation. As the inflammatory response progresses, the affected mucosal areas can develop more severe lesions, ultimately leading to ulceration. An image-to-image comparison was conducted to assess the correlation and verify that the synthetic images resembled the original images. The analysis revealed that both the original and synthetic erythema images exhibited high correlation values. This finding indicates that synthetic images maintain a strong linear relationship with the original images in terms of pixel intensity patterns. The lower contrast in the synthetic image indicates that some of the fine details or variations in intensity might have been smoothed out or lost during the synthetic image generation process. The increase in homogeneity for the synthetic image implies that it may have less variation and smoother transitions between different areas, which could be due to the generation process making the image appear more blended or less textured. The synthetic erythema image closely resembles the original in terms of overall structure but may have slightly less contrast, leading to a smoother appearance with fewer distinct features. According to Table 6, the correlation is close to 1 in both images, indicating a high similarity between an original image and a synthetic image. Despite this, the ADASYN method shows several shortcomings: some images show color inconsistencies and do not perfectly match the feature distributions of real images. Nevertheless, the synthetic image has captured the general color distribution of the original image. To compare the performance of our model, some architectures employed in other VCE-related research were tested under the same conditions as mentioned in this study. Marin-Santos *et al.*[4] studied 15,792 images related to CD lesions and developed an architecture with 6 blocks, consisting of convolution, batch normalization, and ReLu layers, which use 3 × 3 kernels. Klang *et al.*[5] implemented an EfficientNet-B5 to classify intestinal strictures as normal or ulcerated mucosa on 27,892 VCE images. Vallée, Rémi *et al.*[30], implemented a ResNet-34 to classify CD lesions on 3,984 VCE images. These architectures were subjected to the same conditions used in the present study, as shown in Table 4, except EfficientNet-B5, where a batch size of four was used. They were trained using the ADASYN augmented dataset and tested on the test set. The dimensions of the images were modified according to the architecture specifications. For the custom CNN, the image was resized to 320 × 320 pixels. With EfficientNet-B5, it was adjusted to 456

× 456 pixels. For ResNet-34, the size was set to 256 × 256 pixels. Table 7 shows the results of the trial.

TABLE 7. **Performance comparison between state-of-the-art CNNs.**

Model	Accuracy	Precision	Recall	F1-score
Marin-Santos *et al.* [4]	0.7898	0.7841	0.7898	0.7808
Klang *et al.* [5]	0.8734	0.8829	0.8734	0.8634
Vallée, Rémi *et al.* [30]	0.8870	0.8863	0.8870	0.8837
Modified DenseNet-201 (This work)	**0.8978**	**0.8966**	**0.8969**	**0.8967**

The CNN with the lowest performance is the custom CNN by[4], whose accuracy is less than 80 %. Despite this, its training time was 30 minutes, the lowest among the different architectures. This suggests that more epochs are needed to extract relevant features and enhance accuracy. However, for this experiment, only 200 epochs were used for comparison purposes. EfficientNet-B5[5], which took a total of 1808 minutes to train, achieved 87.34 % accuracy. This architecture employs depthwise separable convolutions and integrates squeeze and excitation (SE) blocks, which enhance feature recalibration by learning channel-wise dependencies. It also uses the swish activation function, which has been shown to improve performance compared to ReLu. These characteristics make EfficientNet-B5 competitive among CNNs. However, despite its strong performance in general image classification tasks, the architecture shows some deficiencies when classifying images of the small intestine. These challenges may be due to the intricate and variable nature of small intestine imaging, which requires specialized diligence and feature extraction that EfficientNet-B5 may not fully capture. ResNet-34[30], which took 299 minutes to train, achieved the second-best overall performance with 88.7 % accuracy. This highlights the efficiency of ResNet architecture in processing information. The ResNet residual blocks allow for deeper networks by mitigating the vanishing gradient problem, enabling robust learning and improved accuracy without excessive computational cost. The proposed model in this study (DenseNet-201) achieves the best overall performance with an accuracy of 89.78 %. The training time was 2022 minutes, making it the longest among the architectures evaluated. If this model is integrated into a clinical tool, video processing time could be a limitation. This extensive training time raises potential concerns regarding the feasibility of integrating this model into a clinical tool, particularly for video processing. However, the present work does not implement the model for real-time video analysis. Consequently, the time needed for the model to analyze complete video sequences for lesions was not measured. Although the main contribution of this work is the lesion-level classification, which provides greater precision and detail in the analysis and diagnosis of images. This methodology allows the identification and categorization of specific lesions within medical images, providing valuable information for the treatment and follow-up of diseases. Therefore, the CNN-based multiclassification models are promising for being potentially more valuable than conventional physician SB-VCE readings. However, before being fully introduced in clinical practice, there are still challenges: most studies are conducted under experimental conditions and have a retrospective single-center study design. Additionally, the incorporation of AI into clinical practice and trials requires a significant leap in terms of standardization, quality assessment, reproducibility, and workflow integration to ensure its effective and reliable implementation. So far, CNN-based systems aim to enhance diagnostics by serving as a support tool for specialists, assisting in the identification and analysis of medical images rather than replacing clinical judgment. In summary, an accuracy of 89.78 % was achieved in the classification of different SB lesions with the model implemented in this work. A more diverse dataset of 7,172 images was generated via a minority sampling strategy. The experiment was repeated three times to verify the consistency of the results. In each iteration, the different sets were selected in a random and balanced manner across the classes. In addition, the performance of the model was compared with another VCE-related research. The results obtained confirm the ability of the network to distinguish between SB-related lesions. This work could serve as a tool in the diagnosis of SB-related lesions, reducing human error and improving the detection of lesions, allowing faster patient diagnoses and treatment plans.

CONCLUSIONS

This study combined the Kvasir-Capsule and CrohnIPI datasets to create a more representative dataset with 5,899 images. The ADASYN technique to resample the minority class erythema in the training dataset. As a result, the resampled dataset expanded to 7,172 images, while the validation and testing sets remained at 885 images each. Our findings show that the ADASYN technique improves the accuracy of the custom DenseNet-201 model implemented to classify different SB lesions. The presented methodology demonstrated a high overall accuracy of 89.78 %. However, in this study, some limitations were identified. The model struggles to distinguish between certain classes, primarily because of the lack of images. Although the resampling technique improves performance, it is not sufficient for the DenseNet-201 model to achieve over 90 % accuracy in classifying the erythema class.

CONTRIBUTIONS OF THE AUTHORS

E. O. C. R. conceptualization, investigation, methodology, formal analysis, visualization, software, and writing-review & editing manuscript; C. E. G. T. conceptualization, investigation, formal analysis, visualization, writing of the original draft, and writing-review & editing manuscript; R. M. Q. conceptualization, investigation, formal analysis, visualization, writing of the original draft, and writing-review & editing manuscript; J. I. G. T. formal analysis, visualization, validation, and writing-review & editing manuscript; J. M. C. P. investigation, formal analysis, and writing-review & editing manuscript; J. R. D. C. formal analysis, validation, writing of the original draft, and writing-review & editing manuscript. All authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] M. F. Lynch Mejia, "La cápsula endoscópica como estudio diagnóstico en gastroenterología," Rev. Med. Sinerg., vol. 4, no. 4, pp. 18-25, 2019, doi: https://doi.org/10.31434/rms.v4i4.179

[2] J. A. Urrego, W. O. Regino, and M. G. Zuleta, "Is the videocapsule endoscopy the best option for diagnosis of possible bleeding from the small intestine?," Rev. Colomb. Gastroenterol., vol. 35, no. 2, pp. 196-206, 2020, doi: https://doi.org/10.22516/25007440.262

[3] J. Melson et al., "Video capsule endoscopy," Gastrointest. Endosc., vol. 93, no. 4, pp. 784-796, 2021, doi: https://doi.org/10.1016/j.gie.2020.12.001

[4] D. Marin-Santos, et. al., "Automatic detection of Crohn disease in wireless capsule endoscopic images using a deep convolutional neural network," Appl. Intell., vol. 53, no. 10, pp. 12632-12646, 2022, doi: https://doi.org/10.1007/s10489-022-04146-3

[5] E. Klang et al., "Automated detection of Crohn's Disease intestinal strictures on capsule endoscopy images using deep neural networks," J. Crohns Colitis, vol. 15, no. 5, pp. 749-756, 2021, art. no. 103638, doi: https://doi.org/10.1093/ecco-jcc/jjaa234

[6] Ş. Öztürk and U. Özkaya, "Residual LSTM layered CNN for classification of gastrointestinal tract diseases," J. Biomed. Inform., vol. 113, 2021, art. no. 103638, doi: https://doi.org/10.1016/j.jbi.2020.103638

[7] T. Agrawal, R. Gupta, S. Sahu, and C. Espy-Wilson, "SCL-UMD at the medico task-mediaeval 2017: Transfer learning based classification of medical images," in CEUR Workshop Proc., vol. 1984, Dublin, Ireland, 2017, pp. 3-5. [Online]. Available: https://ceur-ws.org/Vol-1984/Mediaeval_2017_paper_21.pdf

[8] Y. Chang, Z. Huang, W. Chen and Q. Shen, "Gastrointestinal tract diseases detection with deep attention neural network," in Proc. of the 27th ACM International Conference on Multimedia, New York, NY, US, 2019, pp. 2568-2572, doi: https://doi.org/10.1145/3343031.3356061

[9] T. H. Hoang, et al., "An application of residual network and faster - RCNN for medico: Multimedia task at MediaEval 2018," in CEUR Workshop Proc, vol. 2283, Sophia Antipolis, France, 2018, pp. 3-5. [Online]. Available: https://ceur-ws.org/Vol-2283/MediaEval_18_paper_11.pdf

[10] Z. M. Lonseko et al., “Gastrointestinal Disease Classification in Endoscopic Images Using Attention-Guided Convolutional Neural Networks,” Appl. Sci., vol. 11, no. 23, 2021, art. no. 11136, doi: https://doi.org/10.3390/APP112311136

[11] T. Rahim, M. A. Usman, and S. Y. Shin, “A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging,” Comput. Med. Imag. Graph., vol. 85, 2020, art. no. 101767, doi: https://doi.org/10.1016/J.COMPMEDIMAG.2020.101767

[12] A. Caroppo, A. Leone, and P. Siciliano, “Deep transfer learning approaches for bleeding detection in endoscopy images,” Comput. Med. Imag.Graph., vol. 88, 2021, art. no. 101852, doi: https://doi.org/10.1016/J.COMPMEDIMAG.2020.101852

[13] X. Wang et al., “Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction,” Comput. Methods Programs Biomed., vol. 187, 2020, art. no. 105236, doi: https://doi.org/10.1016/J.CMPB.2019.105236

[14] Y. Shin, et al., “Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches,” IEEE Access, vol. 6, pp. 40950-40962, 2018, doi: https://doi.org/10.1109/ACCESS.2018.2856402

[15] A. Nogueira-Rodríguez et al., “Deep Neural Networks approaches for detecting and classifying colorectal polyps,” Neurocomput., vol. 423, pp. 721-734, 2021, doi: https://doi.org/10.1016/J.NEUCOM.2020.02.123

[16] M. Souaidi and M. El Ansari, “Multi-scale hybrid network for polyp detection in wireless capsule endoscopy and colonoscopy images,” Diagnostics, vol. 12, no. 8, 2022, art. no. 2030, doi: https://doi.org/10.3390/DIAGNOSTICS12082030

[17] P. M. Vieira, et al., “Multi-pathology detection and lesion localization in WCE videos by using the instance segmentation approach,” Artif. Intell. Med., vol. 119, 2021, art. no. 102141, doi: https://doi.org/10.1016/j.artmed.2021.102141

[18] F. Fonseca, B. Nunes, M. Salgado, and A. Cunha, “Abnormality classification in small datasets of capsule endoscopy images,” Procedia Comput. Sci., vol. 196, pp. 469-476, 2022, doi: https://doi.org/10.1016/J.PROCS.2021.12.038

[19] S. Mahmood et al., “A robust deep model for classification of peptic ulcer and other digestive tract disorders using endoscopic images,” Biomedicines, vol. 10, no. 9, 2022, art. no. 2195, doi: https://doi.org/10.3390/BIOMEDICINES10092195

[20] S. Jain et al., “A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images,” Comput. Biol. Med., vol. 137, 2021, art. no. 104789, doi: https://doi.org/10.1016/j.compbiomed.2021.104789

[21] Y. Ku, H. Ding, and G. Wang, “Efficient Synchronous real-time CADe for multicategory lesions in gastroscopy by using multiclass detection model,” Biomed. Res. Int., vol. 2022, 2022, art. no. 504149, doi: https://doi.org/10.1155/2022/8504149

[22] P. H. Smedsrud et al., “Kvasir-Capsule, a video capsule endoscopy dataset,” Sci. Data, vol. 8, no. 1, 2021, art. no. 142, doi: https://doi.org/10.1038/s41597-021-00920-z

[23] R. Vallée, et al., “CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn’s disease lesions recognition algorithms,” in Proc. SPIE 11317, Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, Houston, TX, US., 2020, pp. 440-446, doi: https://doi.org/10.1117/12.2543584

[24] A. C. Bovik, “Chapter 3-Basic Gray Level Image Processing,” in The Essential Guide to Image Processing, 2nd. Ed., 2009, pp. 43-68, doi: https://doi.org/10.48550/arXiv.1406.2661

[25] E. Mocsari and S. S. Stone, “Densely Connected Convolutional Networks,” 2017, arXiv:1608.06993, doi: https://doi.org/10.48550/arXiv.1608.06993

[26] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in Proc. of the International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 2008, pp. 1322-1328, doi: https://doi.org/10.1109/IJCNN.2008.4633969

[27] E. O. Cuevas-Rodriguez et al., “Comparative study of convolutional neural network architectures for gastrointestinal lesions classification,” PeerJ, vol. 11, 2023, art. no. e14806, doi: https://doi.org/10.7717/PEERJ.14806

[28] T. Chauhan, H. Palivela, and S. Tiwari, “Optimization and fine-tuning of DenseNet model for classification of COVID-19 cases in medical imaging,” Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, 2021, art. no. 100020, doi: https://doi.org/10.1016/J.JJIMEI.2021.100020

[29] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211-252, 2015, doi: https://doi.org/10.1007/s11263-015-0816-y

[30] A. de Maissin et al., “Multi-expert annotation of Crohn’s disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network,” Endosc. Int. Open, vol. 9, no. 7, pp. E1136-E1144, 2021, doi: https://doi.org/10.1055/A-1468-3964