# Detection of Angioectasias and Haemorrhages Incorporated into a Multi-Class Classification Tool for the GI Tract Anomalies by Using Binary CNNs

**Christos Barbagiannis[1], Alexios Polydorou[1], Michail Zervakis[1], Andreas Polydorou[2], Eleftheria Sergaki[1]** 

[1]School of Electrical and Computer Engineering (ECE) School, Technical University of Crete, Chania, Greece;
[2]Department of Surgery, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece

**Correspondence to:** Sergaki Eleftheria, elefsergaki@ece.tuc.gr

## ABSTRACT

The proposed deep learning algorithm will be integrated as a binary classifier under the umbrella of a multi-class classification tool to facilitate the automated detection of non-healthy deformities, anatomical landmarks, pathological findings, other anomalies and normal cases, by examining medical endoscopic images of GI tract. Each binary classifier is trained to detect one specific non-healthy condition. The algorithm analyzed in the present work expands the ability of detection of this tool by classifying GI tract image snapshots into two classes, depicting haemorrhage and non-haemorrhage state. The proposed algorithm is the result of the collaboration between interdisciplinary specialists on AI and Data Analysis, Computer Vision, Gastroenterologists of four University Gastroenterology Departments of Greek Medical Schools. The data used are 195 videos (177 from non-healthy cases and 18 from healthy cases) videos captured from the PillCam[(R)] Medronics device, originated from 195 patients, all diagnosed with different forms of angioectasia, haemorrhages and other diseases from different sites of the gastrointestinal (GI), mainly including difficult cases of diagnosis. Our AI algorithm is based on convolutional neural network (CNN) trained on annotated images at image level, using a semantic tag indicating whether the image contains angioectasia and haemorrhage traces or not. At least 22 CNN architectures were created and evaluated some of which pre-trained applying transfer learning on ImageNet data. All the CNN variations were introduced, trained to a prevalence dataset of 50%, and evaluated of unseen data. On test data, the best results were obtained from our

CNN architectures which do not utilize backbone of transfer learning. Across a balanced dataset from no-healthy images and healthy images from 39 videos from different patients, identified correct diagnosis with sensitivity 90%, specificity 92%, precision 91.8%, FPR 8%, FNR 10%. Besides, we compared the performance of our best CNN algorithm versus our same goal algorithm based on HSV colorimetric lesions features extracted of pixel-level annotations, both algorithms trained and tested on the same data. It is evaluated that the CNN trained on image level annotated images, is 9% less sensitive, achieves 2.6% less precision, 1.2% less FPR, and 7% less FNR, than that based on HSV filters, extracted from on pixel-level annotated training data.

# 1. BACKGROUND

## 1.1. Computer Aided Diagnosis Based on Expert Systems, ML, AI

The last decade, artificial intelligence (AI) has gone through rapid development, it seems that AI could provide opportunities for better health care, so it is a challenge to apply AI algorithms in order to create CAD tools for automated detection of anomalies and non-healthy deformities in medical images (MRI, CT, endoscopy, etc.) and assisted tools for disease diagnosis. The automated medical decision-making on non-predefined criteria is focusing on the advantage of AI systems that can avoid the typical fallacies of human psychology (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc.), and the widespread human inability to process statistical data. However, we must have in mind that algorithmic decisions may also be mistaken or discriminatory, reproducing human biases and introducing new ones. Since 2018 the European Union General Data Protection Regulation (GDPR) has required that AI or other systems CAD tools should be able to explain their decisions. The future challenge will consist in finding the best combination between human and automated intelligence, taking into account the capacities and the limitations of both [1].

Deep learning is a specific example of AI. The power of deep learning is that by back-propagation its layers can learn from patterns in the data by discovering implicit features that are not part of human language. CNN algorithms look a great tool to assist endoscopists in identifying images of endoscopic capsule including anomalies of different forms. In early 1950s, the first neural network machine was developed based on ideas developed in the 1940s. The first CNN which is a major form of deep learning, was developed in the late 1970s.

The GARVAN-ES1 expert system is an history example of expert system that since 1984 was introduced into routine use at the Garvan Inst. of Medical Research at St Vinncent's Hospital in Sydney to provide clinical interpretations for reports from a diagnostic laboratory which measures thyroid hormone levels. The program was written in C, to produce reports that are 99% correct, with higher quality product at a lower cost [2].

The recent years, different Machine Learning (ML) and AI techniques are applied to the multi-class endoscopic images that could reliably be used as CAD assisted tools in clinics. Focusing on AI diagnosis of gastrointestinal (GI) diseases, over a period of three consecutive years, the Medico GI challenges, called researchers to design classification algorithms on 16 classes of GI real-world dataset. Algorithms on AI diagnosis of gastrointestinal (GI) are developed and evaluated for the challenges: Medical Multimedia Task at MediaEval 2017, Medico Multimedia Task at MediaEval 2018, and BioMedia ACM MM Grand Challenge 2019 [3]. The goals of these challenges were to evaluate multi class-classification methods for classifying about 16 classes, such as anatomical landmarks (e.g. z-line, pylorus, cecum), as pathological findings (e.g. esophagitis, polyps, ulcerative colitis), as polyp removal cases (e.g. dyed and lifted polyps, dyed resection margins), and normal and regular cases (e.g., normal colon mucosa, stool, instrument etc.) inside the GI tract.

The last years, researchers are exploring Machine Learning (ML) algorithms for binary classification

models like logistic regression and SVM that do not support multi-class classification natively and require meta-strategies and researchers that explore CNNs for direct multi classes classifications, or CNNs splitting the classification in binary classifiers. The CNNs "One-versus-Rest" strategy splits a multi-class classification into one binary classification problem per class. As it is mentioned in [4], the "One-versus-All the Rest" approach (also called "One-vs-All"), in which we train C binary classifiers, fc(x), where the data from class c is treated as positive, and the data from all the other classes is treated as negative.

The endoscopic image classification ML is a high interest field of research because it approaches the missed classification of diseases that subsequently ensure effective early disease detection. The ML algorithms can improve the images reviewing from inexperienced examiners (endoscopists, radiographers, etc.) and can reduce the reviewing time in cases such as sequences of images in a video stream.

## 1.2. Diagnosis of Angioectasia and Haemorrhages into Gastrointestinal Tract

Gastrointestinal (GI) disease types of angioectasia and haemorrhages traces refer to any in the gastrointestinal tract, in the upper part: oesophagus, stomach and first part of duodenum and the low part: small intestine to large intestine colon. Statistically the small bowel bleeding traces is only the 5% of all in GI. Abnormal blood vessels, usually veins, is the main cause of bleeding, it is being a common situation in the 30% to 40% of population over the age of 50.

Small bowel capsule endoscopy (SBCE) is a major clinical diagnostic practice, as the next step after a standard gastroscopy for the upper GI endoscopy and a colonoscopy for the lower GI endoscopy has failed to do, because it is able to look at the whole small bowel. Once capsule swallowed, as travelling inside the GI tract, the capsule is capturing minimum two pictures per second (it depends from the model), and is transmitting these images by radiofrequency, to a receiver worn by the patient (except in the case of the capsule from CapsoCam® where the data are stored within the capsule). The capsule's travel often last from 4 to 8 hours, the capsule's batteries offer a lifespan of 10 to 15 hours, the capsule during its travel takes for a total of about 55,000 images. It is expected that when it will be technologically possible to control the mobility and the power to feed more sensors, robotic arms and actuators integrated into capsule's small volume, the robotic endoscopic capsule will be used therapeutically in the GI tract [5, 6].

Until 2021, the way the doctors review the CE video has not changed much since SBCE introduction from 2000. Until now, the evaluation of image sequences aiming the localization and classification of lesions in small intestine area, is based on the personal experience of doctors and lacks of automation based on specific scores to standardize the description of lesions, quantify objectively the lesions in capsule endoscopy reports, with potential prognostic and therapeutic impact. The companies in the Endoscopic Capsule market tends to support their products with proprietary software for the doctors to use in order to automate some aspects of reviewing procedure, mainly to reduce the long review time. This is done by adjusting the number of frames viewed simultaneous (dual view or quad view, as either sequential or overlapping images), adjusting the speed at which frames are presented, and eliminated the similar consecutive snapshots. Moreover, there is the software tool in the RAPID Reader v8.3 program, "suspected blood indicator" (SBI) function which has been developed and evaluated as a computer aided tool for automatic detection of bleeding traces into the GI tract [7].

AI algorithm trained among 281 patients, 10.3% presented with active haemorrhage while 28.9% presented lesions with high bleeding potential (angioectasias, ulcers and tumours), the SBI software tool achieves a 96.6% sensitivity for active small bowel bleeding, with a 97.7% negative predictive value. Regarding high bleeding potential lesions, the SBI displayed an overall sensitivity of 39.5%, being highest for ulcerated neoplasias (100%), but significantly lower for angioectasias (38.5%) or ulcers (20.0%) [8].

It has been shown that as blood can vary in color, the SBI software tool is likely unable to detect all shades of red and so it is a reliable tool to exclude active bleeding and/or major lesions, so the reviewing of the CE video by an expert is still important for the detection of lesions responsible for past bleeding [9, 10].

It has been shown that in the cases of patients with few or minor lesions, the diagnosis by reviewing

the CE videos by doctors, is dependent from the endoscopist experience [11].

At the International Endoscopic Capsule Conference in 2006, there was a consensus that the fastest projection rate proved to be acceptable was 15 fps [12]. It is suggested that the novice examiner should read the video at 12 - 15 fps [13]. Conversely, an experienced examiner can choose a faster rate (20 fps) if Single View mode is used, while selecting Multi View mode can reduce reading time by 30%.

It has been evaluated the effectiveness of the QuickView Rapid® 5 software: 1) to reduce the video stream reading time, 2) to evaluate the number of false negatives in cases of not well experienced endoscopists, and determine the learning curve for reviewing CE images [14-16]. It has been evaluated that the Olympus algorithm called Omni mode, the Omni mode may be appropriate only for detecting major lesions [16]. Moreover, it has been noticed that "Quickview" and "Quadview" modes are still unacceptable because some significant lesions were not detected, with a diagnostic miss rate ranging from 6.5% to 12% [14, 16, 17].

Concerning the methods that technically maximize color differences in images captured by a capsule, are categorized during the process phase either by decomposing the images using simulated blue filter, or by decomposing the images using a simulated "chromoendoscopy" algorithm. Fuji Intelligent Color Enhancement (FICE®) technology decomposes images by wavelengths, and as a result maximizes image differences on patterns, such as vascular and mucosal patterns.

It has been shown that RAPID® 5 Access improves diagnostic yield in the detection rate of denuded redness, reducing reading time, as much in the "auto mode", as in "displaying a single image at 12 fps". RAPID(®) 5 Access improves diagnostic yield, reducing reading time; however, it is still unacceptable because of the diagnostic miss rate and may be useful as an ancillary reading tool [18].

The capsule's images blue filter decomposition versus the FISE decomposition has been compared and evaluated for all lesion categories as follows: 1) that blue filter decomposition provided image improvement (compared to white light raw images) in 83%, 2) that blue filter offers better image enhancement than FISE in capsule endoscopy [19].

Blue filter decomposition has been used as Blue Mode (BM) modality for PillCam® and it has been achieved detection improvement [20].

During an experiment, comparing the diagnosis results of reviewing the same data by two groups of endoscopists: one group of experts versus one inexperienced, in each group, the results varied depended on the findings, even between experts. In both groups, the accuracies were lower in cases with subtle mucosal lesions, such as erosion, angioectasia, and divertic ulum, and more higher in cases with more prominent intraluminal changes (e.g. active small-bowel bleeding, ulcer, tumor, stenotic lumen) [21]. That means that the use of computer aided diagnosis tools will assist to eliminate the factor of doctor's reviewing experience.

The algorithm proposed by our team, based on pixels statistical analysis features, using colorimetric image filters in the HSV color space extracted of manually pixel-level annotations of regions of interest (ROI) can achieve sensitivity of 99% [22].

About the performance of CNN architecture when pretrained with AlexNet, it is shown that using a dataset including 8200 non-healthy and 40,000 healthy frames (there is no mention from how many different videos or different patients), can be achieved detection of sensitivity 92.2% (there is no information for precision and FPR) [23].

Another proposed methodology is first step to classify the bleeding samples into active and inactive subgroups based on the statistical features derived from the histogram probability of the color space. Then for each subgroup, we highlight the blood regions via fully convolutional networks (FCNs) [24, 25].

## 2. AIMS

The aim of our team is to develop a multi-class CAD tool for diagnosing different GI tract diseases, consisting of a number of binary CNNs, each one for a different disease. That, because of the benefits of training binary CNNs instead of a multi-class CNN are a quite important (less time intensive, less compu-

tational complexity, thus achieving better values for sensitivity and accuracy). The proposed algorithm is to aid inexperienced endoscopists to identify haemorrhage state or non-haemorrhage state into frames of endoscopic videos. Angiectasias are the most common lesions diagnosed in patients with medium GI bleeding. Each one of the CNN binary classifiers belonging to this CAD tool is trained to detect a specific disease. The input data for each binary CNN are the frames of endoscopic videos showing frames of healthy states, and/or frames of different known diseases (such as polyps, angiectasias, etc.). For example, the task of another binary classifier from the same system could be to identify polyps, of another could be to classify the polyps to adenomatous polyps (adenomas) from non-neoplastics (under publication), and so on. Because our proposed CNN models are designed to classify a given image only in two classes, the output layer has two neurons. The last fully connected layer, which is a two-dimensional feature vector, is given as an input to softmax classifier, which makes the final prediction whether there is an "angiectasia" or "not". The frames classified as "not" can be healthy states or frames showing other diseases than "angiectasia".

Moreover, we compare the performance of our present proposed CNN for identifying "angiectasia" or "not", using training data on image level annotation versus our recently published CAD algorithm based on HSV colorimetric lesions features extracted of manually pixel-level annotations of regions of interest (ROI). We trained and tested both algorithms on the same data.

## 3. DATA AND METHODS

The proposed CAD algorithm is based on a CNN, trained on dataset annotated on image level (not pixel), on the same dataset that used by the authors, in their published algorithm based on pixels statistical analysis features, using colorimetric image filters in the HSV color space [22]. The data used, are not yet publicly available for the research community, consist of 195 capsule endoscopy procedures from patients that were referred for obscure gastrointestinal bleeding or for suspected Crohn's disease. Abnormalities were found in 177 (90.7%) of cases, including dark hemorrhage and angiodysplasia(s) in 31 cases (15.9%), active bleeding in 8 cases (4.1%)). The completely healthy cases were 18 (9.3%) of cases. All the cases, were treated in five Greek national university hospitals. The videos are anonymized, and the confidentiality protocol applied complies with current European Union legal regulations.

All the videos used are taken by the endoscopic capsule Pillcam$^{TM}$ SB videos (previously known as Given Imaging, Yokneam, Israel), which has the ability to capture $250 \times 250$ pixels, the exported images are of $576 \times 576$ pixels through the camera's own extrapolation algorithm (see samples in Figure 1). All the patients are diagnosed by the attending expert doctors of four University Gastroenterology Departments of Medical Schools: 1) of National and Kapodistrian University of Athens, 2) Attikon Hospital and 3) Laikon Hospital of Athens, and the 4) Aristotle University of Thessaloniki. All the video streams were reviewed manually by experts and automated by the capsule's own diagnostic tool SBI$^{TM}$ (Suspected Blood Indicator) of Rapid$^{TM}$ READER and were annotated by surgeons gastroenterologists with more than 35 years experience. The team's expert doctor, in collaboration with the team's data analysis engineer, eliminated images taken from the same area by applying the capsule's own software Quickview$^{TM}$. The metadata from the manually checked annotation process are: 1) lesion ID, 2) size of lesion in mm, 3) lesion diagnosis, 4) number of lesions per image.

The training and testing datasets are balanced, each one consists of 50% video frames showing conditions of GI haemorrhaging and GI angioectasias, and 50% non-haemorrhaging, see Set1 in Table 1. Frames identified as taken from the same area of lesion or from the same healthy area, are excluded. There is a number of less than ideal video frames (*i.e.* including foaming and bubbles, processed food, out of focus imagery, highly light reflections, out of bound lighting conditions).

For the training procedure are used 130 videos. In detail, 92 videos showing other diseases than haemorrhage (non-haemorrhage) cases of 92 patients, 26 videos of 26 patients showing haemorrhage cases and 12 videos of complete healthy cases of 12 patients. Frames from the same patient are not used for both sets, there is no overlap in content between the training set and the testing set, see Set 2 in Table 1.

**Table 1.** Data used: 195 videos* from endoscopic capsule pillcamtm sb separated into two subsets: set 1 = {130 videos}, set 2 = {65 videos}. All as positive haemorrhage diagnosed).

**Set 1 (training set):** Selected 3200 video frames from 26 videos showing hemorrhage traces and 3200 relevant video frames (in terms of location in GI tract) from 92 videos showing other **diseases** than hemorrhage cases, and 12 videos showing complete healthy cases. Set 1 used only for training **purposes** of the proposed algorithms.

| Diagnosis | Men | Women |
|---|---|---|
| Haemorrhage (Small** & Medium***) | 23 | 3 |
| Non-haemorrhage (healthy) | 7 | 5 |
| Non-haemorrhage (Polyps/tumors) | 16 | 8 |
| Non-haemorrhage (Scars) | 14 | 7 |
| Non-haemorrhage (Ulcer(s)) | 10 | 12 |
| Non-haemorrhage (Other non-healthy) | 9 | 14 |

**Set2 (testing set):** Selected 1000 frames from 13 videos showing hemorrhage traces and 1000 relevant video frames (in terms of location in GI tract) from 46 videos showing other **diseases** than hemorrhage cases, and from 6 videos showing complete healthy cases. Set 2 used only for testing **purposes**.

| Diagnosis | Men | Women |
|---|---|---|
| Haemorrhage (Small** & Medium***) | 10 | 3 |
| Non-haemorrhage (healthy) | 3 | 3 |
| Non-haemorrhage (Polyps/tumors) | 15 | 6 |
| Non-haemorrhage (Scars) | 4 | 6 |
| Non-haemorrhage (Ulcer(s)) | 4 | 5 |
| Non-haemorrhage (Other non-healthy) | 3 | 3 |

*Videos come from Gastroenterology Departments of three Greek public hospitals: Aretaieio Hospital, Atticon Hospital, Laiko Hospital, **Small: continuous surface of 800 to 1000 pixels, ***Medium: continuous surface of 1000 to 3000 pixels. The Extensive bleeding areas, defined by us with more than 3000 pixel, they are not considered due that are obviously obvious.

The training dataset includes of 6200 video frames (50% of these frames are showing haemorrhage conditions. Totally, 3100 video frames are showing non-haemorrhage state (healthy and other diseases) are relevant in terms of location in GI tract. All the frames selected manually by expert doctors, in a way to be inclusive of different shapes, dimension, active and past hemorrhage traces, see Table 1). Figure 1 illustrates a sample of our data.

Similarly, in order to control the evaluation procedure, to ensure that precision does not depend on prevalence, the precision is normalized to a prevalence of 50%. Thus, is introduced the ratio of the number of cases in the haemorrhage control group and the number of cases in the non-haemorrhage group, used to establish the Negative Predictive Value (NPV) and the Positive Predictive Value (PPV) or Precision, is equal to the Prevalence of the diseases in the studied population. The testing group includes total 1000 frames, of 500 non-haemorrhage frames (healthy or other diseases than haemorrhage) and of 500 haemorrhage frames, from patients that do not belong in the training dataset. The training dataset, is named
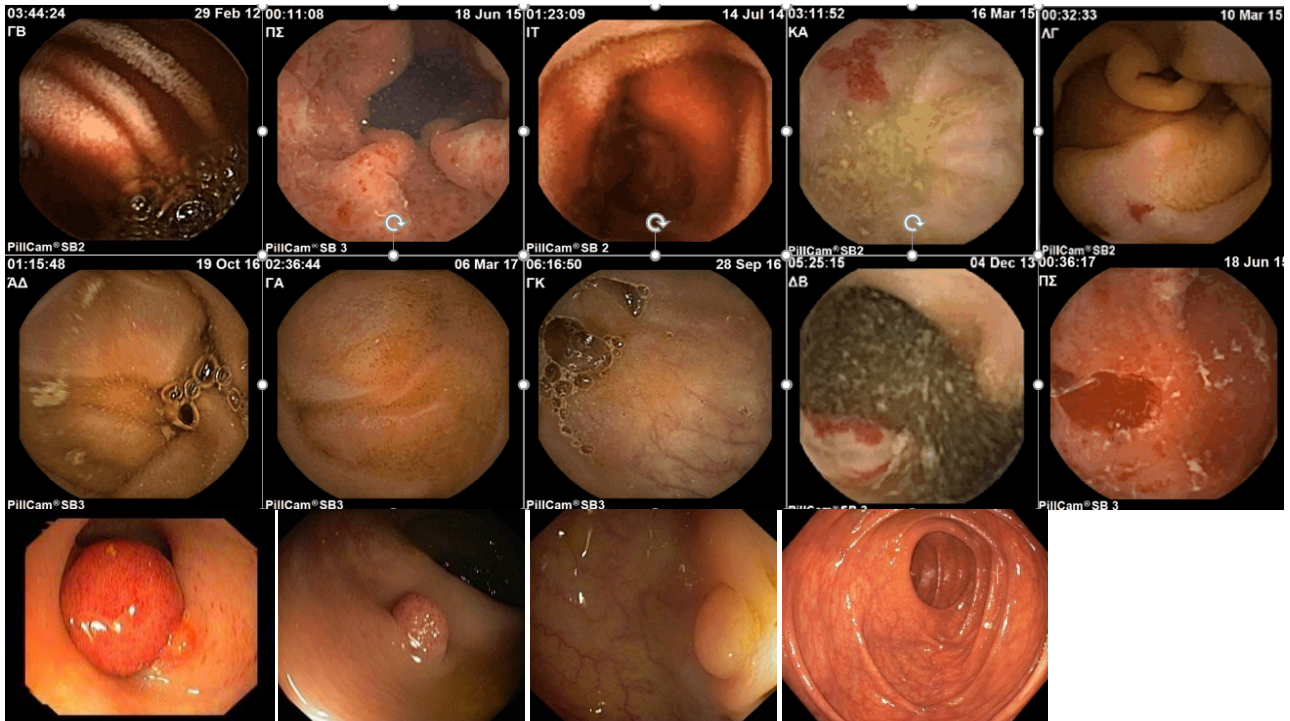
**Figure 1.** Sample of our CE captured frames: 1st row, from left to right: all cases of bleeding and angioectasia. 2nd row, from left to right: three normal cases, the last two angioectasia. 3rd row, from left to right: cases of polyps, the last is normal.

as Set 1, the testing set, named as Set 2, both described in Table 1. For the testing procedure are used 65 videos (46 showing other diseases, 6 showing healthy cases, 13 showing hemorrhage cases).

To create a new artificially training data from the existing training data Set 1, the augmentation technique applied in balanced subset of the data Set 1 (in the training dataset, and not to the validation or test dataset). Figure 2 illustrates a sample of augmented data.

### 3.1. Preparation of Training Data

The pool of images that feed the CNN model is determined by shuffling the video frames of Set 1 (see Table 1).

It is ensured that the aspect ratio for all the images used is equal to 1:1 and same resolution, the latern ensured by cropping useless pixels around the edges. After this procedure, we end up with 6400 images, of $224 \times 224$ pixels.

The chosen color model is RGB, normalized in the range of [0 - 1] (instead of the usual [0 - 255]). From all the above, its obvious that the final payload of each image is $224 \times 224 \times 3$ pixels.

Data augmentation was applied by the ImageDataGenerator class, using Keras deep learning neural network library. The transformations applied upon the dataset are: Rotation: 40, Width Shifting: 0.2, Height Shifting: 0.2, Shear Intensity: 0.2, Zoom: 0.2, Horizontal Flip.

### 3.2. Creation of Deep Learning Models

The CNNs training run entirely in the cloud. The Keras API Python library run on top of Tensorflow numerical platform in Python. The back propagation is automatically generated by TensorFlow, for optimizer we used Adam with learning rate 0.0001. For Loss is used binary cross entropy. Our methodology to create the CNN architecture, was to add the blocks one at a time in order to see the differences in the results as we going deeper.
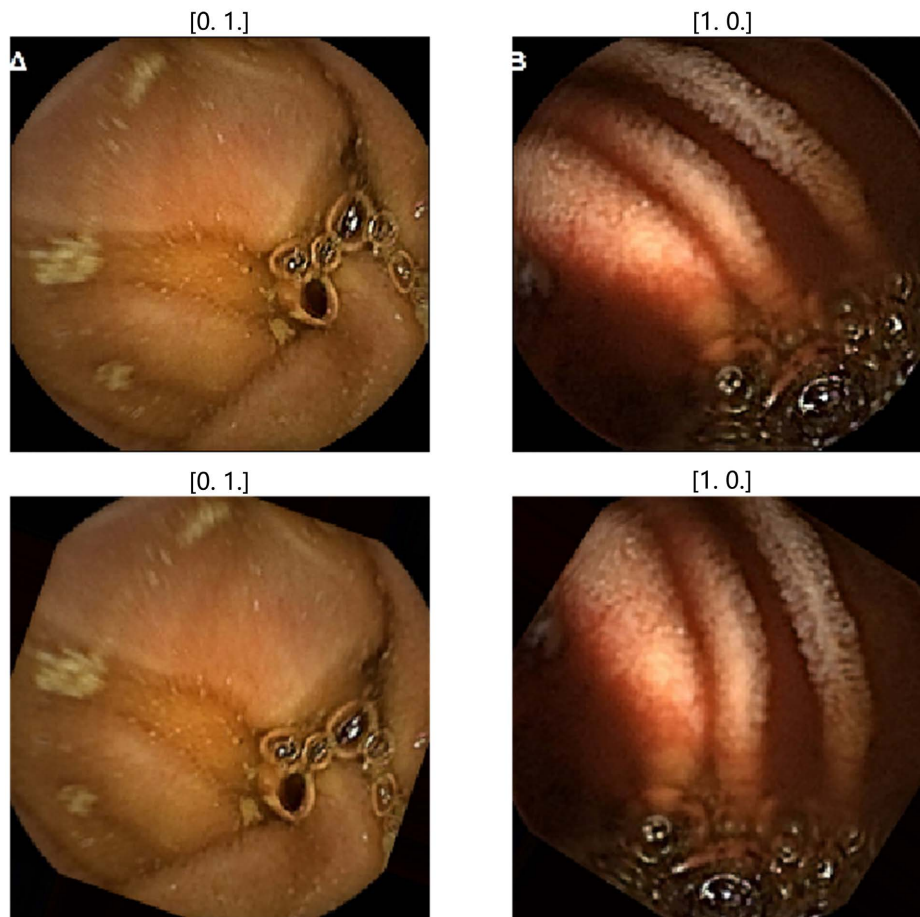
**Figure 2.** By applying augmentation technique transformation of Rotation, our original training data is extended. Left up, is shown an original healthy video frame, left and down is shown the extracted by rotation augmentatio. Similarly, right, is shown a non-healthy case (case showing bleeding trace).

The architecture of the CNN model created consists of two parts: a convolutional base and a classifier. The convolutional base contains convolutional layers (Conv2D) and Max Pooling layers. Each model has different number of Conv2D, all convolutional windows are [3 × 3] with filter size (32, 64, 128, 256) and MaxPooling with pooling size (2, 2) (that in order the CNN to be capable of recognizing an hemorrhage as an object even when its appearance varies in some way). At the end the CNN model there is a classifier consisting of a flatten layer and fully connected layers. The first fully connected layers have 128, 256 or 512 neurons. The activation function used are the ReLU and Sigmoid: The ReLU used in all layers except the output layer and returns 0 for every negative value in the input image while it returns the same value for every positive value: $R(z) = \max(0, z)$. The Sigmoid used only in the output layer the main reason why we use sigmoid function is because it exists between (0 to 1) and we have only two states (Non-showing hemorrhage traces and showing hemorrhage traces).

Our proposed CNNs variations are trained: 1) on the original dataset Set 1, and 2) on an artificially extended dataset, as it is created by applying data augmentation technique on Set 1. Our chosen data augmentation is performed by: a) rotation range 40, b) width and height shift range 0.2, c) shear range 0.2, d) zoom range 0.2, e) horizontal flip.

During training phase, dropout technique applied (to randomly and temporarily dropping out units,

hidden and visible, along with their connections, in a neural network). We applied to classifier: Dropout (0.5) and to convolutional base: Dropout (0.2), which means that means that we dropped out randomly the 50% and 20% of the CNN's units.

By implementing of 22 CNN variations, we have been able to choose the two best performers in terms of sensitivity and specificity. Similarly the techniques implemented for those two best performers, where data augmentation and Dropout only in the classifier. The architecture of the CNN used data augmentation (named: Model 1_Ag) consists of: three Conv2D layers with filter size (32, 64, 128), three Max Pooling layers, one flatten layer and two fully connected layers with 128 and 2 neurons respectively. The architecture of the CNN used Dropout only in the classifier (named: Model 3_DPcl) consists of: four double Conv2D layers with filter size (32, 64, 128, 256), four Max Pooling layers, one flatten layer, three fully connected layers with 256, 256 and 2 neurons respectively and two Dropout (0.5) layers.

Some of our CNNs variations were trained using Transfer Learning (TL). The pre-trained models used are the VGG16 and RESNET50 trained on ImageNet dataset and fine-tune them on our training data Set1. Our transfer learning is applied as follows: First we trained only the classifier and kept the convolutional base freeze of the transferring model, finally we Fine-Tune the transferring model by unfreezing the part of convolutional base (block 4 and 5 for the VGG16, and the RESNET50). The RESNET50 performs the initial convolution and max-pooling using $7 \times 7$ and $3 \times 3$ kernel sizes.

Respectively, all the models trained for 10 epochs except in TL for 5 epochs. The CNN variations where TL was used, their tested performance was identical.

The algorithms testing was carried out under the guidance of the supervising of our team's experienced doctor, at "Aretaieion" Greek University Hospital in Athens.

## 3.3. Evaluation Metrics

For the testing procedure, the indicators used to measure the performance of the proposed algorithms are based on statistical metrics of specificity, sensitivity, precision, False Positive Ratio (FPR), False Negative Ratio (FNR). Equations from (1) to (6) were used for all CNNs.

Accuracy is the probability of correctly recognizing a frame as True Positive (TP) or True Negative (TN) among the total number of cases examined. It is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition):

$$\text{Sencitivity} = \frac{TP}{TP + FN} \tag{2}$$

Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

Precision measures the proportion of all positive (TP + FP) identifications was actually correct:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

FPR (False Negative Ratio) or (Type I Error) which is the error of proportion to recognize as bleeding (positives) the images that are actually healthy (negatives). We need that to have very low values. It is calculated as following:

$$FPR = \frac{FP}{TN + FP} \tag{5}$$

FNR (False Negative Ratio) or (Type II Error) is the error of proportion to recognize as healthy (negatives) the images that are actually bleeding (positives). We have to notice that it is very important in medical diagnosis to achieve FNR very low because people that have bleed may not get the proper healthcare because of the results.

$$FNR = \frac{FN}{TP + FN} \tag{6}$$

For the above equations:
TP: Is the state that a bleeding image was correctly predicted as bleeding.
TN: Is the state that a non-bleeding image was correctly predicted as non-bleeding.
FP: Is the state that a non-bleeding image was falsely predicted as bleeding.
FN: Is the state that a bleeding image was falsely predicted as non-bleeding.

## 4. RESULTS

The proposed binary classifier based on CNN architecture using data augmentation, identified the binary classification of class haemorrhage (frames showing cases of GI haemorrhaging and GI angioectasias abnormalities) and class of non-haemorrhage (frames not showing hemorrhage traces) with FNR 10%, FPR 8%, sensitivity 90%, specificity 92%, accuracy 91%, precision 91.8%.

These performance metrics are compared to their respective values as published by our previous work in [22] (CAD method based on color HSV colorimetric filters, trained and tested on pixel-level annotations on the same dataset). We evaluated that our best CNN trained on image level annotated images is by 9% less sensitive, by 2.6% less precision, by 1.2% less FPR, and by 7% less FNR.

We verified that the mean reviewing time per patient by the well-experienced doctor was 29 ± 9 min by handcrafted reviewing and by 6 ± 2 min by the proposed method.

Moreover, we verified that our CNNs variations which were trained using Transfer Learning (TL) achieved the worst performance (specificity 4%, accuracy 50%, sensitivity 98%).

## 5. CONCLUSIONS

We trained variations of CNN algorithms using image level annotated datasets from endoscopic capsule videos with goal the binary classification, to identify the endoscopic capsule video frames that are showing haemorrhage traces due to GI haemorrhaging and/or angioectasias inside the GI tract. The best performance of our tested CNN variations will be used as a sub-task classifier under the umbrella of our under development CAD tool for different disease in GI tract, among other binary classifiers for other GI diseases (*i.e.* polyps and ulcers.

Our results of the best CNN binary classifier are compared with our recently published CAD method aimed to achieve the same goal, based on feature extraction which requires significantly a more labor intensive approach (manually annotated pixels instead of image level), by creating HSV colorimetric filters. Both methods are controlled over the same data collection, where the training procedure. Our previous proposed method using HSV colorimetric filters provided higher levels of performance metrics than the one currently proposed (CNN CAD) [22]. This can be explained by the fact that the dataset for the manually selected colorimetric features extraction was based on pixel-level and big enough, as compared to the more demanding CNN method of training in terms of absolute number of samples. Both diagnosis methods, proved less accurate than the one exercised by the well-experienced doctor. The proposed CAD diagnosis methods can be an important assisted tool for less experience doctors. In any case, methods that eliminate similar images are necessary to reduce the overview time.

Concerning the technical aspect of the training of a CNN model, our experience can testify for the following: The use of a validation dataset during the training phase is helpful for the train process to en-

sure the case of having underfit or overfit. The training and the testing dataset must be balanced (showing non-angioectasia and non-haemorrhages, and other cases).

The techniques of Dropout and Augmentation improve the CNN training performance. For the Dropout, we had better results when the technique was applied to the classifier. For the Transfer Learning we had overtrained, which depends on many factors, one must be that the CNNs pre-training achieved using images non-similar of our data (images from GI tract).

Moreover, we verified that Deeper CNN model by adding more layers doesn't mean better performance.

## AUTHOR CONTRIBUTIONS

## ACKNOWLEDGEMENTS

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

## REFERENCES

1. Sartor, G. (2020) The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligences. EPRS|European Parliamentary Research Service Scientific Foresight Unit (STOA), PE 641.530, European University Institute of Florence.

2. Compton, P. and Jansen, B. (1988) Knowledge in Context: A Strategy for Expert System Maintenance. *Conference AI* 88: *Proceedings of the* 2*nd Australian Joint Artificial Intelligence Conference*, Adelaide, 15-18 November 1988, 292-306.

3. Jha, D., *et al.* (2021) A Comprehensive Analysis of Classification Methods in Gastrointestinal Endoscopy Imaging. *Medical Image Analysis*, **70**, Article ID: 102007. https://doi.org/10.1016/j.media.2021.102007

4. Murphy, K.P. (2012) Machine Learning: A Probabilistic Perspective. Kindle Edition.

5. Zhang, Q., Prendergast, J.M., Formosa, G.A., Fulton, M.J. and Rentschler, M.E. (2020) Enabling Autonomous Colonoscopy Intervention Using a Robotic Endoscope Platform. *IEEE Transactions on Biomedical Engineering*, **68**, 1957-1968. https://doi.org/10.1109/TBME.2020.3043388

6. Prendergast, J.M., Formosa, G.A., Fulton, M.J., Heckman, C.R. and Rentschle, M.E. (2020) A Real-Time State Dependent Region Estimator for Autonomous Endoscope Navigation. *IEEE Transactions on Robotics*, **37**, 918-934. https://doi.org/10.1109/TRO.2020.3038709

7. Pasha, S.F. and Leighton, J.A. (2017) Evidence-Based Guide on Capsule Endoscopy for Small Bowel Bleeding. *Gastroenterology and Hepatology*, **13**, 88-93.

8. Boal Carvalho, P., Magalhães, J., Dias, D.E., Castro, F., Monteiro, S., Rosa, B., Moreira, M.J. and Cotter, J. (2017) Suspected Blood Indicator in Capsule Endoscopy: A Valuable Tool for Gastrointestinal Bleeding Diagnosis. *Arquivos de Gastroenterologia*, **54**, 16-20. https://doi.org/10.1590/s0004-2803.2017v54n1-03

9. Tal, A.O., Filmann, N., Makhlin, K., Hausmann, J.F., Rust, M., Herrmann, E., Zeuzem, S. and Albert, J.G. (2014) The Capsule Endoscopy "Suspected Blood Indicator" (SBI) for Detection of Active Small Bowel Bleeding: No Active Bleeding in Case of Negative SBI. *Scandinavian Journal of Gastroenterology*, **49**, 1131-1135. https://doi.org/10.3109/00365521.2014.923503

10. Han, S., Fahed, J. and Cave, D. (2018) Suspected Blood Indicator to Identify Active Gastrointestinal Bleeding: A Prospective Validation. *Gastroenterology Research*, **11**, 106-111. https://doi.org/10.14740/gr949w

11. Jensen, M.D., Nathan, T. and Kjeldsen, J. (2010) Inter-Observer Agreement for Detection of Small Bowel Crohn's Disease with Capsule Endoscopy. *Scandinavian Journal of Gastroenterology*, **45**, 878-884. https://doi.org/10.3109/00365521.2010.483014

12. Jang, B.I., *et al.* (2010) Inter-Observer Agreement on the Interpretation of Capsule Endoscopy Findings Based on Capsule Endoscopy Structured Terminology: A Multicenter Study by the Korean Gut Image Study Group. *Scandinavian Journal of Gastroenterology*, **45**, 370-374. https://doi.org/10.3109/00365520903521574

13. Cave, D.R. (2004) Reading Wireless Video Capsule Endoscopy. *Gastrointestinal Endoscopy Clinics of North America*, **14**, 17-24. https://doi.org/10.1016/j.giec.2003.10.007

14. Saurin, J.C., Lapalus, M.G., Cholet, F., D'Halluin, P.N., Filoche, B., Gaudric, M., Sacher-Huvelin, S., Savalle, C., Frederic, M., Lamarre, P.A. and Ben Soussan, E. (2012) Can We Shorten the Small-Bowel Capsule Reading Time with the "Quick-View" Image Detection System? *Digestive and Liver Disease*, **44**, 477-481. https://doi.org/10.1016/j.dld.2011.12.021

15. Hosoe, N., Rey, J.F., Imaeda, H., Bessho, R., Ichikawa, R., Ida, Y., Naganuma, M., Kanai, T., Hibi, T. and Ogata, H. (2012) Evaluations of Capsule Endoscopy Software in Reducing the Reading Time and the Rate of False Negatives by Inexperienced Endoscopists. *Clinics and Research in Hepatology and Gastroenterology*, **36**, 66-71. https://doi.org/10.1016/j.clinre.2011.09.009

16. Hosoe, N., Watanabe, K., Miyazaki, T., Shimatani, M., Wakamatsu, T., Okazaki, K., Esaki, M., Matsumoto, T., Abe, T., Kanai, T., Ohtsuka, K., Watanabe, M., Ikeda, K., Tajiri, H., Ohmiya, N., Nakamura, M., Goto, H., Tsujikawa, T. and Ogata, H. (2016) Evaluation of Performance of the Omni Mode for Detecting Video Capsule Endoscopy Images: A Multicenter Randomized Controlled Trial. *Endoscopy International Open*, **4**, 878-882. https://doi.org/10.1055/s-0042-111389

17. Shiotani, A., Honda, K., Kawakami, M., Murao, T., Matsumoto, H., Tarumi, K., Kusunoki, H., Hata, J. and Haruma, K. (2011) Evaluation of RAPID (®) 5 Access Software for Examination of Capsule Endoscopies and Reading of the Capsule by an Endoscopy Nurse. *Journal of Gastroenterology*, **46**, 138-142.

https://doi.org/10.1007/s00535-010-0312-7

18. Iakovidis, D.K., Tsevas, S. and Polydorou, A. (2010) Reduction of Capsule Endoscopy Reading Times by Unsupervised Image Mining. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, **34**, 471-478. https://doi.org/10.1016/j.compmedimag.2009.11.005

19. Krystallis, C., Koulaouzidis, A., Douglas, S. and Plevris, J.N. (2011) Chromoendoscopy in Small Bowel Capsule Endoscopy: Blue Mode or Fuji Intelligent Colour Enhancement? *Digestive and Liver Disease: Official Journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver*, **43**, 953-957. https://doi.org/10.1016/j.dld.2011.07.018

20. Abdelaal, U.M., Morita, E., Nouda, S., Kuramoto, T., Miyaji, K., Fukui, H., Tsuda, Y., Fukuda, A., Murano, M., Tokioka, S., Umegaki, E., Arfa, U.A. and Higuchi, K. (2015) Blue Mode Imaging May Improve the Detection and Visualization of Small-Bowel Lesions: A Capsule Endoscopy Study. *Saudi Journal of Gastroenterology*, **21**, 418-422. https://doi.org/10.4103/1319-3767.170954

21. Jang, B.I., Lee, S.H., Moon, J.S., Cheung, D.Y., Lee, I.S., Kim, J.O., Cheon, J.H., Park, C.H., Byeon, J.S., Park ,Y.S., Shim, K.N., Kim, Y.S., Kim, K.J., Lee, K.J., Ryu, J.K., Chang, D.K., Chun, H.J. and Choi, M.G. (2010) Inter-Observer Agreement on the Interpretation of Capsule Endoscopy Findings Based on Capsule Endoscopy Structured Terminology: A Multicenter Study by the Korean Gut Image Study Group. *Scandinavian Journal of Gastroenterology*, **45**, 370-374. https://doi.org/10.3109/00365520903521574

22. Polydorou, A., Sergaki, E., Polydorou, A., Barbagiannis, C., Vardiambasis, I., Giakos, G. and Zervakis, M. (2021) Improving CAD Hemorrhage Detection in Capsule Endoscopy. *Journal of Biomedical Science and Engineering*, **14**, 103-118. https://doi.org/10.4236/jbise.2021.143011

23. Jia, X. and Meng, M.Q. (2016) A Deep Convolutional Neural Network for Bleeding Detection in Wireless Capsule Endoscopy Images. *Proceedings of 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC* 2016), Orlando, 16-20 August 2016, 639-642. https://doi.org/10.1109/EMBC.2016.7590783

24. Jia, X. and Meng, M.Q. (2017) Gastrointestinal Bleeding Detection in Wireless Capsule Endoscopy Images Using Handcrafted and CNN Features. *Proceedings of 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC* 2017), Jeju, 11-15 July 2017, 3154-3157. https://ieeexplore.ieee.org/document/8037526/figures#figures https://doi.org/10.1109/EMBC.2017.8037526

25. Jia, X. and Meng, M.Q. (2017) A Study on Automated Segmentation of Blood Regions in Wireless Capsule Endoscopy Images Using Fully Convolutional Networks. *Proceedings of 2017 IEEE 14th International Symposium on Biomedical Imaging* (*ISBI* 2017), Melbourne, 18-21 April 2017, 179-182. https://ieeexplore.ieee.org/document/7950496 https://doi.org/10.1109/ISBI.2017.7950496