electronics

MDPI

Article

Diagnosis Aid System for Colorectal Cancer Using Low Computational Cost Deep Learning Architectures

Álvaro Gago-Fabero [1], Luis Muñoz-Saavedra [1], Javier Civit-Masot [1,2], Francisco Luna-Perejón [1,2], José María Rodríguez Corral [3] and Manuel Domínguez-Morales [1,2,*]

[1] Robotics and Technology of Computers Research Group (TEP-108), Architecture and Computer Technology Department, Escuela Técnica Superior de Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Seville, Spain; lmsaavedra@us.es (L.M.-S.); mjavier@us.es (J.C.-M.); fluna1@us.es (F.L.-P.)

[2] Computer Engineering Research Institute (I3US), Escuela Técnica Superior de Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Seville, Spain

[3] Applied Robotics Research Group (TEP-940), School of Engineering, University of Cadiz, Avda. Universidad de Cádiz, 10, 11519 Puerto Real, Spain; josemaria.rodriguez@uca.es

* Correspondence: mjdominguez@us.es

Abstract: Colorectal cancer is the second leading cause of cancer-related deaths worldwide. To prevent deaths, regular screenings with histopathological analysis of colorectal tissue should be performed. A diagnostic aid system could reduce the time required by medical professionals, and provide an initial approach to the final diagnosis. In this study, we analyze low computational custom architectures, based on Convolutional Neural Networks, which can serve as high-accuracy binary classifiers for colorectal cancer screening using histopathological images. For this purpose, we carry out an optimization process to obtain the best performance model in terms of effectiveness as a classifier and computational cost by reducing the number of parameters. Subsequently, we compare the results obtained with previous work in the same field. Cross-validation reveals a high robustness of the models as classifiers, yielding superior accuracy outcomes of 99.4 ± 0.58% and 93.2 ± 1.46% for the lighter model. The classifiers achieved an accuracy exceeding 99% on the test subset using low-resolution images and a significantly reduced layer count, with images sized at 11% of those used in previous studies. Consequently, we estimate a projected reduction of up to 50% in computational costs compared to the most lightweight model proposed in the existing literature.

Keywords: colorectal cancer; hystopathology; medical imaging; diagnosis-aid system; deep learning; artificial intelligence

check for updates

Citation: Gago-Fabero, Á.; Muñoz-Saavedra, L.; Civit-Masot, J.; Luna-Perejón, F.; Rodríguez Corral, J.M.; Domínguez-Morales, M. Diagnosis Aid System for Colorectal Cancer Using Low Computational Cost Deep Learning Architectures. *Electronics* **2024**, *13*, 2248. https://doi.org/10.3390/electronics13122248

Academic Editors: Gabriella Olmo and Prasan Kumar Sahoo

Received: 15 March 2024
Revised: 29 May 2024
Accepted: 6 June 2024
Published: 7 June 2024

1. Introduction

Colorectal cancer primarily affects older individuals, with the average age of onset between 70 and 71 years, and the majority of patients being older than 50 years at the time of diagnosis. However, it also occurs in younger individuals, with a higher incidence in men than in women. According to a 2020 study by the Global Cancer Observatory (GLOBOCAN) [1], colorectal cancer is the third most common type of cancer in the world, with approximately 1,931,590 cases among men and women in 2020, representing 10% of all cancers diagnosed. This places it behind breast cancer (11.7%) and lung cancer (11.4%), followed by prostate cancer (7.3%) and stomach cancer (5.6%). The World Health Organization (WHO) estimates that it is the third leading cause of death worldwide, with almost 1 million deaths per year and almost 2 million cases diagnosed [2]. Within colorectal cancers, several types can be found, such as carcinoid tumors, gastrointestinal stromal tumors, lymphomas, sarcomas, and adenocarcinomas. According to the American Cancer Society (ACS), adenocarcinoma cases represent around 96% of all colorectal cancers detected [3]. To better understand what an adenocarcinoma is, according to the ACS, it is a type of cancer that begins in the cells that form the glands that produce mucus to lubricate the interior of the colon and rectum.

Detecting adenocarcinoma requires a biopsy (removal of a small piece of tissue). These samples are then analyzed in the laboratory by a pathologist, who will indicate whether adenocarcinoma is present in the cells of the sample or not. This detection requires an anatomical pathologist with between 11 and 14 years of experience [4]. According to the study performed by Fromer [5], the reported frequency of errors committed by anatomical pathologists ranges from 1% to 43%, and for oncology, it is between 1% and 5%, as these are very delicate cases. However, the saturation of emergency services and the lack of health professionals means that the time needed to perform the autopsy and obtain a diagnosis is longer than ideal: according to the Spanish Ministry of Health, in its 2019 report [6], there were fewer than 1200 anatomical pathologists in the country, and with more than 1300 cases per year on average per pathologists. With these numbers, each one can spend less than an hour and a half on each case (and cases are increasing year after year at a higher rate than pathologists).

In this context, the use of techniques that help to reduce the time needed for diagnosis and the burden on medical professionals is essential. Not surprisingly, in recent years, there has been a proliferation in the use of Deep Learning techniques to design systems to aid diagnosis with medical images. Thus, we can find works analyzing skin images [7], various types of cancers [8], glaucoma [9,10], or even X-ray images [11,12], all of them using Deep Learning approaches with classification results above 90%. In addition, in recent years, several studies have applied the principles of AI to the detection of colorectal cancer using medical imaging. Among these works are those developed by Hamida et al. [13] and Singh et al. [14]. These two works (and others) will be compared with our work in the Results section.

Therefore, the main objective of this work is to design, develop, and test a diagnostic aid system for colorectal cancer using Deep Learning techniques, specifically Convolutional Neural Networks (CNNs). Due to the limitation of the dataset used, the classification will be between benign tissue and adenocarcinoma, but the process followed here can be applied to other datasets. The approach of this work involves leveraging CNNs with a reduced number of parameters and lower image resolution. This strategic reduction aims to facilitate the implementation of the system in computing environments constrained by limited computational resources, making the technology more accessible and practical in diverse clinical settings. Additionally, there is a concerted effort to enhance the models not only in terms of computational efficiency, but also in their capacity as effective classifiers, improving the classification rate of health professionals and other previous work in this area. This combination aligns with the overarching goal of developing a more practical and effective diagnostic tool in the field of oncology.

The rest of the manuscript is structured as follows: First, the methods used to develop and test the diagnosis-aid system are presented in Section 2, including the description of the dataset. The accuracy results obtained after testing the classifier are detailed and discussed in Section 3, and also the comparison with previous works is included here. Finally, in Section 4, the final conclusions of this work and future research lines are detailed.

2. Materials and Methods

In this section, the dataset used for implementing the diagnosis-aid system is presented; next, the process to obtain the best classifier is detailed and, finally, the evaluation metrics used for it are shown. The full work performed is summarized in Figure 1, and it will be detailed step-by-step in the next subsections.

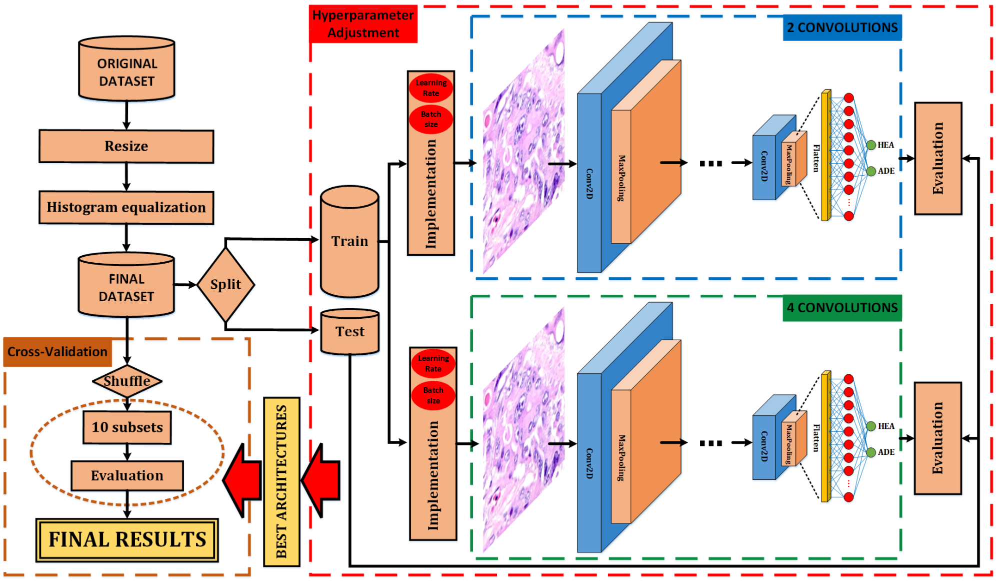


Figure 1. Work graphical abstract.

2.1. Dataset

In this work, a publicly available and pre-tagged dataset is used. It is called LC25000 [15]. It contains lung and colon tissue images, but this work has focused only on the colon cancer images contained in the dataset (10 thousand images). Those images represent zoomed sections of biopsied tissue observed under a microscope. All malignant tumor cases in this dataset belong to the adenocarcinoma class.

As commented before, the adenocarcinoma class represents 96% of all colorectal cancers. It is therefore interesting to develop a reliable and rapid early detection mechanism. This dataset includes a two-class fully balanced division: 5 thousand adenocarcinoma tissue images and 5 thousand healthy tissue images.

All images have a resolution of 768 × 768 pixels in jpeg format. The distribution of images for each tagged class and the subsets' size used to train and test the classifiers are detailed in Table 1.

Table 1. Dataset used in this work and subsets division.

Class	Train (70%)	Validation (10%)	Test (20%)	TOTAL
Healthy	3500	500	1000	5000
Adenocarcinoma	3500	500	1000	5000
TOTAL	7000	1000	2000	10,000

In our initial tests with full-size images, we noticed prolonged training and classification durations without performance improvement. Hence, we prioritized efficiency for deployment in low-resource settings and used resolutions 120 × 120, 90 × 90 and 60 × 60.

Looking at some selected images from the dataset, some differentiating parameters can be detected between the classes (though not in all of them). See Figure 2.

As commented before, the images of the dataset are used to train the classifiers, using 70% of the images for training, 10% for validation, and the last 20% for testing.

Next, the classifiers trained in this work, and the optimization process performed to obtain the best candidate are detailed.

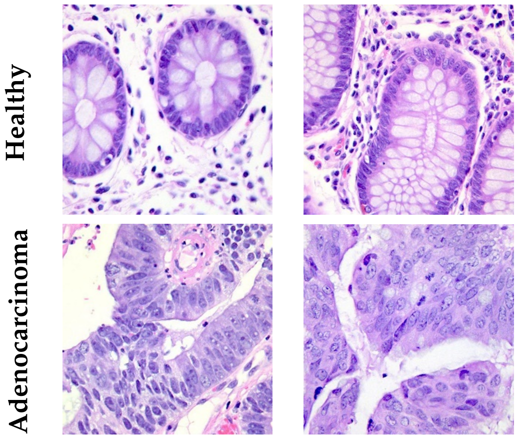


Figure 2. Image examples from LC25000 dataset. (**up**) Healthy tissue; (**down**) Adenocarcinoma.

2.2. Classifiers

The classifiers used for this work all consist of 2 or 4 convolutional layers, followed by 2/4 MaxPolling layers and, after all of them, a flattening layer and 2 or 4 dense layers. The generic model of these classifiers can be seen in Figure 3. Likewise, as our intention is to test several alternatives (with different number of convolutional layers and dense layers), the exact architecture will not be known until these tests are finished. As can be observed, after the flatten layer and after each dense layer, a dropout layer is included.

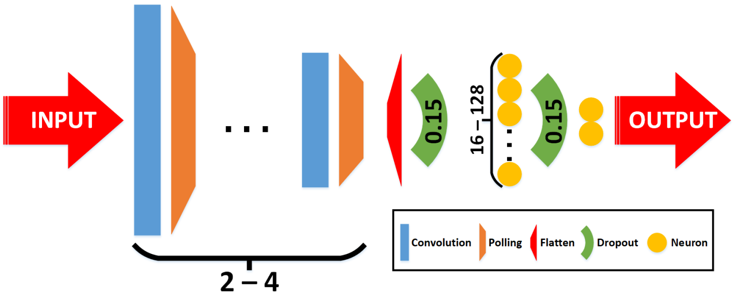


Figure 3. Generic CNN architecture.

To find the best classifier for this problem, a two-stage search process is carried out:

- Phase 1—grid-search: Multiple combinations of architectures and hyperparameters are used to train the classifiers. In this case, 54 combinations are trained, varying the following parameters:
  - Convolutional layers: The number of convolutional layers included in the CNN. This parameter varies with 2 and 4 layers (including a max pooling layer after each one, and a final flatten).
  - Image size (pixels): three resolution reductions were tested, specifically 60 × 60, 90 × 90, and 120 × 120.
  - Learning rate: Step size at each epoch during the training process to update the connection's weights. This parameter varies with 1×10^{-3}, 1×10^{-4}, and 1×10^{-5}.
  - Batch size: The number of training samples in one forward/backward pass. This parameter varies with 10, 20, and 30.

These parameters are summarized in Table 2, including those parameters fixed to a particular value.

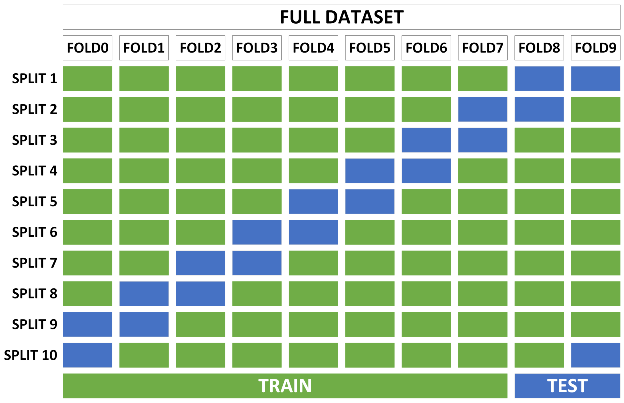
Table 2. Parameters' variations during the grid-search phase.

Parameter	Values
Convolutional Layers	2, 4
Kernel size	3 × 3
Dropout	0.15
Image size	60 × 60, 90 × 90, 120 × 120
Learning rate	$1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}$
Batch size	10, 20, 30
Training epochs	200
Optimizer	Adamax

At the beginning, the parameters' combinations were more than 200 (including intermediate values). However, after some initial tests (with only 50 epochs), no significant improvement was detected when extending the range of values for each parameter. That is why these ranges were significantly reduced for the final grid search.

- Phase 2—cross-validation: With the best candidates from the previous search, robustness tests are performed to determine the adaptability of the classifiers to variations in the training and test sets. This process consists of dividing the complete dataset into several non-overlapping subsets and performing several trainings for each classifier, using a specific number of subsets for training and another for testing (different for each training). The standard deviation in the accuracy of all trainings of the same classifier determines the robustness of the classifier.

For this work, the full dataset is divided into ten folds, using eight of them for training and two for test. This cross-validation split process is summarized in Figure 4.



	FOLD0	FOLD1	FOLD2	FOLD3	FOLD4	FOLD5	FOLD6	FOLD7	FOLD8	FOLD9
SPLIT 1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦
SPLIT 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟩
SPLIT 3	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟩	🟩
SPLIT 4	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟩	🟩	🟩
SPLIT 5	🟩	🟩	🟩	🟩	🟦	🟦	🟩	🟩	🟩	🟩
SPLIT 6	🟩	🟩	🟩	🟦	🟦	🟩	🟩	🟩	🟩	🟩
SPLIT 7	🟩	🟩	🟦	🟦	🟩	🟩	🟩	🟩	🟩	🟩
SPLIT 8	🟩	🟦	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩
SPLIT 9	🟦	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
SPLIT 10	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦

Figure 4. Cross-validation split performed during the second phase.

Summarizing, all of the images are preprocessed offline, reducing their dimensions and applying a histogram equalization process. After that, the optimization algorithm starts with a global search where 54 classifier combinations are tested with 200-epoch trainings (using early-stop option, which stops the training after 10 epochs without any improvement in the validation loss). From this process, the two best models (with the highest test accuracy results) undergo additional robustness testing using a cross-validation technique. Finally, the model with the best mean accuracy and less standard deviation is the final classifier.

Each of these steps is widely known, and is used in previous work, so it is nothing new to make use of them. However, the application of the optimization mechanism composed of these steps in the order indicated above has not been used (except in previous work by this research group) on colorectal histopathological images. Therefore, the novelty presented here is the entire optimization process, as well as the improvement in results that can be observed later (both in terms of accuracy and computational load).

2.3. Evaluation Metrics

To evaluate the effectiveness of the classification systems, it is common to use different and well-known metrics: accuracy (most-used metric), sensitivity (also known as recall), specificity, precision, and F1$_{score}$ [16].

To apply them, the classification results obtained for each class must be tagged individually as "True Positive" (TP; belonging to a class and classified as the same class), "True Negative" (TN; belonging to another class and classified as that class), "False Positive" (FP; belonging to another class and classified to the evaluated class), or "False Negative" (FN; belonging to the class and classified as another class). According to them, the high-level metrics are presented in the following equations:

$$Accuracy = \sum_c \frac{TP_c + TN_c}{TP_c + FP_c + TN_c + FN_c}, c \in classes \tag{1}$$

$$Specificity = \sum_c \frac{TN_c}{TN_c + FP_c}, c \in classes \tag{2}$$

$$Precision = \sum_c \frac{TP_c}{TP_c + FP_c}, c \in classes \tag{3}$$

$$Sensitivity = \sum_c \frac{TP_c}{TP_c + FN_c}, c \in classes \tag{4}$$

$$F1_{score} = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \tag{5}$$

About those metrics:

- Accuracy: all samples classified correctly compared to all samples (see Equation (1)).
- Specificity: proportion of "true negative" values in all cases that do not belong to this class (see Equation (2)).
- Precision: proportion of "true positive" values in all cases that have been classified as it (see Equation (3)).
- Sensitivity (or recall): proportion of "True Positive" values in all cases that belong to this class (see Equation (4)).
- F1$_{score}$: This considers both the precision and the sensitivity (recall) of the test to compute the score. It is the harmonic mean of both parameters (see Equation (5)).

There are other commonly used metrics, but not all works use them. However, the Receiver Operating Characteristic (ROC) curve [17] is of particular interest in diagnostic systems, because it is the visual representation of the True Positives Rate (TPR) versus the False Positives Rate (FPR), as the discrimination threshold is varied. When using the ROC curve, the area under the curve (AUC) is used as a value of the system's goodness-of-fit.

3. Results and Discussion

This section will present the results of this work sequentially: we will start with a summary of the best results obtained in phase 1, the best candidates will be re-evaluated using the cross-validation technique in phase 2 and, finally, the final classifier and the results obtained will be presented. Additionally, a search for similar previous work will be carried out, and we will compare ourselves with them.

For this task, Visual Studio Code was utilized on a computer equipped with an Intel Core i9 13th generation processor, boasting 32 GB of RAM, and housing a GeForce 3080Ti graphics card.

3.1. Phase 1—Grid-Search

As previously indicated, more than 50 combinations of all modifiable parameters have been trained for this phase. Table 3 shows the accuracy results for the test subset obtained from all of the evaluated candidates. These results can also be seen in Figure 5.

Table 3. Summary of the results obtained in Phase 1.

Convolutions	Image Size	Batch Size	Learning Rate	Test Accuracy
2	60 × 60	10	1×10^{-3}	99.57%
			1×10^{-4}	95.28%
			1×10^{-5}	92.99%
		20	1×10^{-3}	83.00%
			1×10^{-4}	99.36%
			1×10^{-5}	94.00%
		30	1×10^{-3}	97.64%
			1×10^{-4}	93.20%
			1×10^{-5}	92.85%
	90 × 90	10	1×10^{-3}	97.64%
			1×10^{-4}	99.07%
			1×10^{-5}	95.28%
		20	1×10^{-3}	32.76%
			1×10^{-4}	84.12%
			1×10^{-5}	93.92%
		30	1×10^{-3}	96.57%
			1×10^{-4}	73.39%
			1×10^{-5}	94.13%
	120 × 120	10	1×10^{-3}	96.47%
			1×10^{-4}	98.71%
			1×10^{-5}	94.61%
		20	1×10^{-3}	81.00%
			1×10^{-4}	89.74%
			1×10^{-5}	95.11%
		30	1×10^{-3}	94.22%
			1×10^{-4}	71.14%
			1×10^{-5}	93.72%
4	60 × 60	10	1×10^{-3}	98.57%
			1×10^{-4}	99.93%
			1×10^{-5}	89.91%
		20	1×10^{-3}	99.64%
			1×10^{-4}	99.86%
			1×10^{-5}	92.56%
		30	1×10^{-3}	99.93%
			1×10^{-4}	99.07%
			1×10^{-5}	99.21%
	90 × 90	10	1×10^{-3}	96.72%
			1×10^{-4}	99.11%
			1×10^{-5}	65.82%
		20	1×10^{-3}	99.49%
			1×10^{-4}	99.73%
			1×10^{-5}	67.10%
		30	1×10^{-3}	98.72%
			1×10^{-4}	99.68%
			1×10^{-5}	73.95%

Table 3. *Cont.*

Convolutions	Image Size	Batch Size	Learning Rate	Test Accuracy
4	120 × 120	10	1×10^{-3}	94.42%
			1×10^{-4}	99.00%
			1×10^{-5}	47.21%
		20	1×10^{-3}	99.30%
			1×10^{-4}	99.93%
			1×10^{-5}	40.00%
		30	1×10^{-3}	98.07%
			1×10^{-4}	99.79%
			1×10^{-5}	41.85%

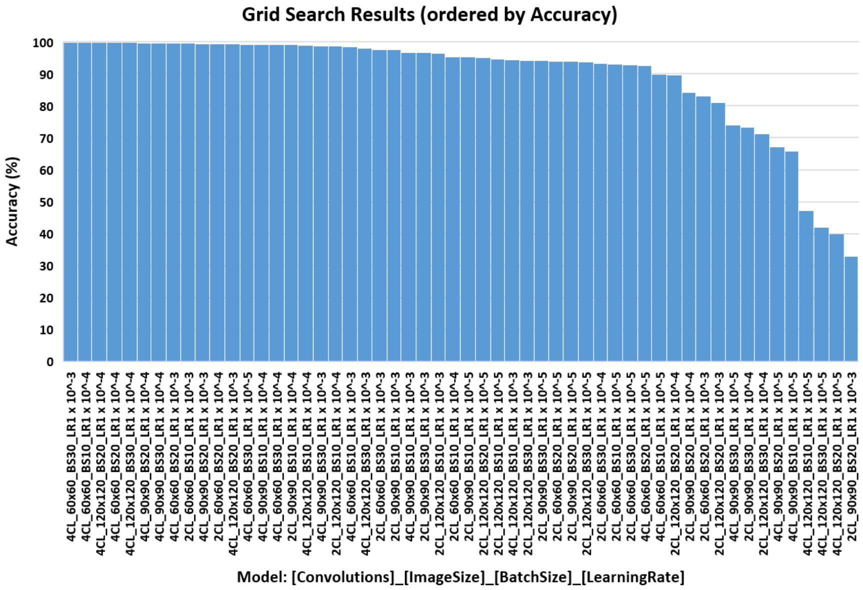


Figure 5. Grid-search results ordered by accuracy.

Among all the trained classifiers, we will look for the best candidate of each architecture to be tested for robustness, i.e., we will look for the best model with two convolutions and the best model with four convolutions.

Among the top ten models, 60 × 60 images dominate (5 of 10), followed by 90 × 90 (3 of 10) and 120 × 120 (2 of 10). However, this does not suggest more information in smaller images or less useful details in larger ones. The complexity lies in finding optimal weights for larger images, requiring more epochs for convergence.

Since, in some cases, the accuracy results are similar, a convergence study is performed during training. The convergence results yield interesting data: in the two-convolution model, the highest results suffer from sudden drops in accuracy during training, suggesting that the final results may not be very robust. For this reason, the selected two-convolution candidate is not among the most accurate, but it is the one with the best convergence.

On this basis, the following two candidates are obtained (marked in red in Table 3):

- Candidate 1: Two convolutions, image size 60 × 60 pixels, batch size 10 and learning rate 1×10^{-4}. Test accuracy results: 95.28%.
- Candidate 2: Four convolutions, image size 60 × 60 pixels, batch size 10 and learning rate 1×10^{-4}. Test accuracy results: 99.93%.

In both cases, the final selected hyperparameters are the same. If we look at the best accuracy result for the two-convolution model (marked in blue in Table 3), the only variation from the selected candidate is the learning rate (which is 10 times higher). Because of this, convergence is faster but, as noted above, there are concerns about sudden drops in convergence (as can be seen in Figure 6).

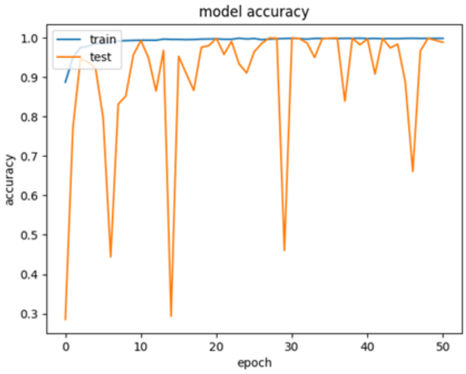


Figure 6. Candidate discarded due to sudden drops in accuracy during training.

This trend contrasts with the convergence found in the proposed candidates. This fact can be observed by looking at Figure 7.

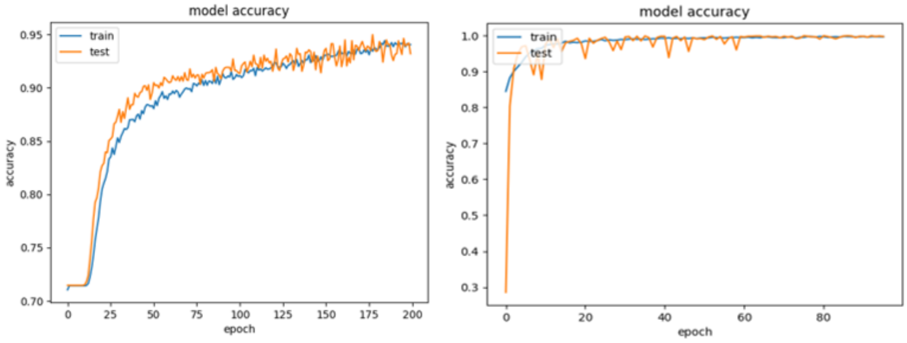


Figure 7. Convergence tendency of the selected candidates: (**left**) model with 2 convolutions; (**right**) model with 4 convolutions.

For the selected candidates, after the processing pipeline, the number of extracted features with 60 × 60 images are: 13 × 13 for the two-convolution classifier, and 2 × 2 for the four-convolution classifier. As observed, for convolution layers, the image information is compressed in only four features for a 60 × 60 resolution, which is why we will also perform the comparisons using the two-convolution model, in order to include a classifier that use more than 100 features. However, as will be observed next, it seems that the four-convolutions system correctly extracts the most important features from the images, as the results obtained are better.

3.2. Phase 2—Cross-Validation

As previously indicated, the two chosen candidates are further tested for robustness by applying the cross-validation technique. The results for each of the folds are shown in Table 4 for candidate 1 and Table 5 for candidate 2.

Table 4. Cross-validation results for candidate 1.

FOLD	1	2	3	4	5	6	7	8	9	10	Average	STD
Accuracy (%)	96.04	91.43	92.14	91.79	94.29	92.14	94.29	92.86	94.29	92.81	93.2	1.46
Loss	0.1736	0.214	0.2379	0.2125	0.1843	0.2069	0.2041	0.1873	0.1919	0.2347	0.204	

As can be seen in the tables above, both candidates have high robustness, with a standard deviation in the accuracy results of less than 1.5% for the first candidate and less than 0.6% for the second candidate. Furthermore, we can see that the average accuracy is higher than 93% for candidate 1 and higher than 99% for candidate 2.

Table 5. Cross-validation results for candidate 2.

FOLD	1	2	3	4	5	6	7	8	9	10	Average	STD
Accuracy (%)	100	99.29	98.21	99.64	99.64	99.64	98.57	100	99.64	99.28	99.4	0.58
Loss	0.0036	0.0179	0.0366	0.0172	0.0076	0.0127	0.0389	0.0063	0.0067	0.0175	0.0165	

In short, they are both good candidates, but the second candidate presents results that are more than 5% better than the first candidate. Therefore, the classifier finally selected will correspond to candidate 2.

Analyzing the detailed results of the final candidate, we can see in Table 6 the results of all the metrics, and its confusion matrix in Figure 8.

Table 6. Metrics results for the final candidate (candidate 2).

Class	Accuracy	Specificity	Precision	Sensitivity	F1-Score
Healthy (HEA)	100%	98.24%	98.27%	100%	99.13%
Adenocarcinoma (ADE)	98.24%	100%	100%	98.24%	99.11%

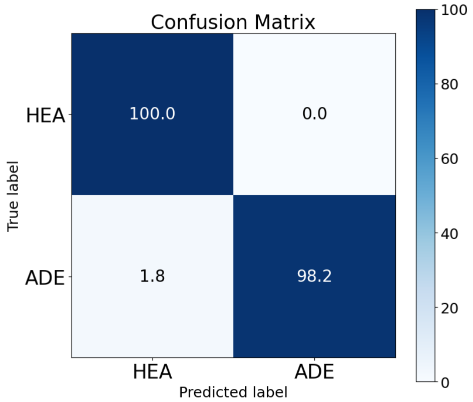


Figure 8. Confusion matrix for the final candidate (candidate 2).

As can be seen, there are no false negatives in the “healthy” class, which means that all cases in this class are correctly classified. However, there is a 1.8% false negative rate for the “adenocarcinoma” class (i.e., 18 cases out of the 1000 present in the test subset for this class).

Finally, the result for the area under the curve (AUC) is 99.4% (see Figure 9).

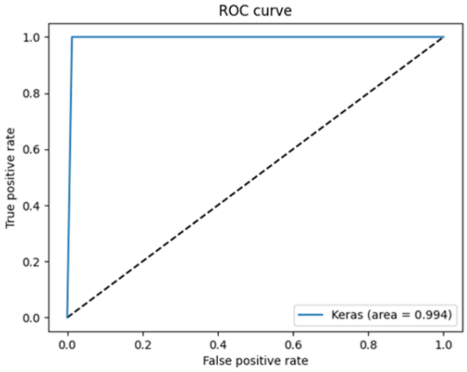


Figure 9. ROC curve and AUC for the final candidate (candidate 2).

3.3. Comparison to Other Works

Due to the relevance of the problem dealt with in this work, there are many articles in which this problem is tackled from different perspectives and providing different solutions, with most of them obtaining good results. So, we performed a deep search using the main search engines (Google Scholar, Scopus, and IEEEXplore) with the search string "((deep learning) OR (machine learning)) AND ((colon) OR (colorectal)) AND ((histology) OR (histopathology))" during the last 3 years. We filtered the results, taking only those published in international journals and those that use CNNs.

The resulting works are presented in Table 7. It is relevant to emphasize that comparisons with several of the identified studies are constrained due to the absence of a test set that would enable the analysis of the model's effectiveness results in a manner not influenced by the observation of validation results during training.

Table 7. Previous works summary and comparison.

Work	CNN Model	Results	Additional Comments
Hamida et al. [13] (2021)	ResNet (18CL, 6PL, 1DL) SegNet (26CL, 10PL, 1DL)	Acc: 96.77–99.98% Acc: 78.39–99.12%	Same dataset (and others)
Masud et al. [18] (2021)	custom (3CL, 3PL, 2DL)	Acc: 96.33%	Same dataset Test subset not considered
Tasnim et al. [19] (2021)	MobileNetV2 (+50CL, 1PL, 2DL)	Acc: 95.48–99.67%	Loss: 0.0124 (best)
Schiele et al. [20] (2021)	InceptionResNetV2 (25CL, 5PL, 5DL)	Acc: 95%	AUC: 84.2%
Babu et al. [21] (2021)	InceptionV3 (+90CL, 15PL, 3DL)	Acc: 96.5–99%	Same dataset (and others)
Sakr et al. [22] (2022)	custom (4CL, 4PL, 2DL)	Acc: 99.5%	Same dataset Imags 180 × 180 Test subset not considered
Ananthakrishnan et al. [23] (2023)	VGG16 (13CL, 5PL, 2DL)	Acc: 98.6%	Same dataset Test subset not considered
Ravikumar and Kumar [24] (2023)	ResNet50V2 (49CL, 2PL, 2DL)	Acc: 99.5%	Same dataset
Singh et al. [14] (2024)	DenseNet201 (49CL, 2PL, 2DL)	Acc: 99.12%	Same dataset Test subset not considered
This work (2023)	custom (2CL, 2PL, 1DL) custom (4CL, 4PL, 1DL)	Acc: 91.43–96.04% Acc: 98.21–100%	

CL: convolutional layers; PL: pooling layers; DL: dense layers; Acc: Accuracy; AUC: area under ROC curve.

As can be seen, the previous work can be divided between those using customized convolutional networks and those using pre-trained networks (which are considerably more computationally expensive).

Among the papers using pre-trained networks, it can be seen that all of them obtain a value close to or higher than 99% for their best case. However, it should be noted that these networks have a computational cost of between 5 and 10 times the computational cost of the network implemented in this work (and, even so, our work obtains results of 100% in the best case, with an average of over 99%). The models obtained by [13] and Ananthakrishnan et al. [23] have high effectiveness, similar to that of other studies, and have reduced complexity, with fewer than 20 convolutional layers. Nevertheless, they still involve a significantly high computational load compared to customized models, and particularly in comparison with the models proposed in this work.

On the other hand, if we look at those works in which customized convolutional networks are used, both use the same dataset as the one used in our work. Moreover, the computational complexity of both works seems similar to ours.

The work by Masud et al. [18] has fewer convolutional layers and fewer pooling layers compared to our second model (although more compared to our first model). However, the results obtained by Masud et al. [18] are similar to those obtained with our less complex model and, in the case of our more complex model, the results we obtain improve the work performed by Masud et al. [18] by more than 3%.

With respect to the work implemented by Sakr et al. [22], the model used is practically identical to ours (although it has one more dense layer), and obtains similar results. This study also does not use a test set, only a validation set, and does not utilize cross-validation, thereby limiting the comparison with the results of the current study. However, there are two important aspects to highlight: in their work, the best result obtained is 99.5%, while that percentage of accuracy corresponds to the average of our classifier (obtaining 100% for the best case) and, as a second aspect, the work performed by Sakr et al. [22] uses image resolutions of 180 × 180 pixels, while ours uses images of 60 × 60 pixels. This last detail means that the convolutional layers need fewer parameters and are executed in less time (since we are talking about images with a size 11% of that of Sakr's images). All this means that our work is one step ahead of Sakr's work in terms of both performance and computational cost. Focusing on our proposed model with a reduced number of parameters, it is noted that, while the accuracy percentage is 3% fewer, it utilizes less than 50% of the computational resources compared to Sakr's proposal.

The computational cost improvement in the presented classifiers has a major implication for the design of portable rapid diagnostic devices. Such devices may not be of great use in hospitals or medical centers, since they have computers that can perform the calculations in real time without the need for this reduction in complexity.

However, in remote locations that do not have a medical center and/or do not have direct access to the internet, an embedded device that is capable of analyzing these samples and providing a real-time response is very useful. This is especially true in countries with a poor infrastructure system or a highly delocalized population.

Unfortunately, due to incomplete disclosure of execution times, hardware details, and source code in many studies, direct empirical comparisons were unfeasible, limiting testing on uniform hardware configurations. Thus, our evaluation relies on reported specifications and resources from these studies. In situations with ample hardware resources, the decision between simpler, less resource-intensive models and complex, computationally demanding ones may seem less pressing. However, our research underscores the importance of creating accessible models for settings with limited resources. Remarkably, one of our models achieved 99% accuracy, comparable to some highly effective models in the existing literature. This highlights that our simpler models maintain high accuracy while demanding lower computational resources.

The limitations of this study are concentrated on the binary classification design (benign, malignant). Our future work aims to incorporate a tiered classification system, allowing for further classification of malignant tissues into subtypes. Additionally, the computationally efficient design paves the way for integration into an embedded system, to evaluate its accuracy and performance in real-time applications. Moreover, future work will involve applying explainable AI techniques to elucidate the classifier's decision-making process and generate customized reports.

4. Conclusions

This work has presented the incidence of colorectal cancer worldwide, being third in the ranking of deaths. Furthermore, according to the ACS, 96% of the cases are adenocarcinoma.

Due to this and the saturation of public health services, the need to develop a system to aid the diagnosis of colon cancer to differentiate between healthy tissue and tissue with adenocarcinoma is justified.

For this purpose, we start from a public dataset already labeled by health professionals and with 5000 images for each of the classes. With this dataset, we followed a two-stage optimization process to obtain the best possible classifier using customized convolutional neural network architectures.

The results obtained are better than 99.5% accuracy and 99.4% AUC. To the best of our knowledge, the proposed model is the lightest to achieve such high effectiveness rates as a classifier, while also requiring very low image resolution. This makes it viable for execution on lower-performance systems, and its integration into cost-effective embedded systems can be considered.

Compared with previous works, the developed classifier obtains similar (or even better) accuracy results that those obtained by more complex models. And, regarding the model complexity, the classifier developed in this work improves all the previous works. So, the next step would, therefore, be to test the system designed in this work on embedded devices to check its operation in real time.

Author Contributions: Conceptualization: J.C.-M. and M.D.-M.; methodology: J.M.R.C. and M.D.-M.; software: Á.G.-F. and L.M.-S.; validation: J.C.-M., F.L.-P. and M.D.-M.; formal analysis: J.M.R.C. and M.D.-M.; investigation: L.M.-S., J.C.-M. and F.L.-P.; resources: J.M.R.C. and M.D.-M.; data curation: Á.G.-F.; writing—original draft preparation: Á.G-.F. and M.D.-M.; writing—review and editing: L.M.-S., J.C.-M., F.L.-P. and J.M.R.C.; visualization: J.M.R.C.; supervision: J.C.-M. and M.D.-M.; project administration: J.M.R.C.; funding acquisition: M.D.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and code associated with this study are available upon request by contacting the corresponding author.

Acknowledgments: We want to thank the research group "TEP108—Robotics and Computer Technology" from University of Seville (Spain).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xi, Y.; Xu, P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl. Oncol.* **2021**, *14*, 101174. [CrossRef] [PubMed]
2. World Health Organization. Colorectal Cancer. 2023. Available online: https://www.iarc.who.int/cancer-type/colorectal-cancer/ (accessed on 12 December 2023).
3. American Cancer Society. Invasive Adenocarcinoma. 2023. Available online: https://www.cancer.org/cancer/diagnosis-staging/tests/biopsy-and-cytology-tests/understanding-your-pathology-report/colon-pathology/invasive-adenocarcinoma-of-the-colon.html (accessed on 12 December 2023).
4. Clement Santiago, A.; Lubelchek, R. What Does a Pathologist Do? 2024. Available online: https://www.verywellhealth.com/how-to-become-a-pathologist-1736292 (accessed on 29 April 2024).
5. Fromer, M.J. Study: Pathology errors can have serious effect on cancer diagnosis & treatment. *Oncol. Times* **2005**, *27*, 25–26.
6. Barber Pérez, P.L.; González-López-Valcárcel, B. Estimación de la Oferta y Demanda de Médicos Especialistas: España 2018–2030. 2019. Available online: https://www.sanidad.gob.es/areas/profesionesSanitarias/profesiones/necesidadEspecialistas/docs/20182030EstimacionOfertaDemandaMedicosEspecialistasV2.pdf (accessed on 29 April 2024).
7. Muñoz-Saavedra, L.; Escobar-Linero, E.; Civit-Masot, J.; Luna-Perejón, F.; Civit, A.; Domínguez-Morales, M. A Robust Ensemble of Convolutional Neural Networks for the Detection of Monkeypox Disease from Skin Images. *Sensors* **2023**, *23*, 7134. [CrossRef] [PubMed]
8. Civit-Masot, J.; Bañuls-Beaterio, A.; Domínguez-Morales, M.; Rivas-Pérez, M.; Muñoz-Saavedra, L.; Corral, J.M. Non-small cell lung cancer diagnosis aid with histopathological images using Explainable Deep Learning techniques. *Comput. Methods Programs Biomed.* **2022**, *226*, 107108. [CrossRef] [PubMed]
9. Corral, J.M.; Civit-Masot, J.; Luna-Perejón, F.; Díaz-Cano, I.; Morgado-Estévez, A.; Domínguez-Morales, M. Energy efficiency in edge TPU vs. embedded GPU for computer-aided medical imaging segmentation and classification. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107298. [CrossRef]
10. Civit-Masot, J.; Domínguez-Morales, M.J.; Vicente-Díaz, S.; Civit, A. Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction. *IEEE Access* **2020**, *8*, 127519–127529. [CrossRef]
11. Civit-Masot, J.; Luna-Perejón, F.; Domínguez Morales, M.; Civit, A. Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images. *Appl. Sci.* **2020**, *10*, 4640. [CrossRef]
12. Muñoz-Saavedra, L.; Civit-Masot, J.; Luna-Perejón, F.; Domínguez-Morales, M.; Civit, A. Does two-class training extract real features? A COVID-19 case study. *Appl. Sci.* **2021**, *11*, 1424. [CrossRef]
13. Hamida, A.B.; Devanne, M.; Weber, J.; Truntzer, C.; Derangère, V.; Ghiringhelli, F.; Forestier, G.; Wemmert, C. Deep learning for colon cancer histopathological images analysis. *Comput. Biol. Med.* **2021**, *136*, 104730. [CrossRef] [PubMed]

14. Singh, O.; Kashyap, K.L.; Singh, K.K. Lung and Colon Cancer Classification of Histopathology Images Using Convolutional Neural Network. *SN Comput. Sci.* **2024**, *5*, 223. [CrossRef]
15. Borkowski, A.A.; Bui, M.M.; Thomas, L.B.; Wilson, C.P.; DeLand, L.A.; Mastorides, S.M. LC25000 Lung and colon Histopathological Image Dataset. *arXiv* **2019**, arXiv:1912.12142. Available online: https://arxiv.org/abs/1912.12142 (accessed on 29 April 2024).
16. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
17. Hoo, Z.H.; Candlish, J.; Teare, D. What is an ROC curve? *Emerg. Med. J.* **2017**, *34*, 357–359. [CrossRef] [PubMed]
18. Masud, M.; Sikder, N.; Nahid, A.A.; Bairagi, A.K.; AlZain, M.A. A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors* **2021**, *21*, 748. [CrossRef] [PubMed]
19. Tasnim, Z.; Chakraborty, S.; Shamrat, F.J.; Chowdhury, A.N.; Nuha, H.A.; Karim, A.; Zahir, S.B.; Billah, M.M. Deep learning predictive model for colon cancer patient using CNN-based classification. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 687–696. [CrossRef]
20. Schiele, S.; Arndt, T.T.; Martin, B.; Miller, S.; Bauer, S.; Banner, B.M.; Brendel, E.M.; Schenkirsch, G.; Anthuber, M.; Huss, R.; et al. Deep learning prediction of metastasis in locally advanced colon cancer using binary histologic tumor images. *Cancers* **2021**, *13*, 2074. [CrossRef] [PubMed]
21. Babu, T.; Singh, T.; Gupta, D.; Hameed, S. Colon cancer prediction on histological images using deep learning features and Bayesian optimized SVM. *J. Intell. Fuzzy Syst.* **2021**, *41*, 5275–5286. [CrossRef]
22. Sakr, A.S.; Soliman, N.F.; Al-Gaashani, M.S.; Pławiak, P.; Ateya, A.A.; Hammad, M. An efficient deep learning approach for colon cancer detection. *Appl. Sci.* **2022**, *12*, 8450. [CrossRef]
23. Ananthakrishnan, B.; Shaik, A.; Chakrabarti, S.; Shukla, V.; Paul, D.; Kavitha, M.S. Smart Diagnosis of Adenocarcinoma Using Convolution Neural Networks and Support Vector Machines. *Sustainability* **2023**, *15*, 1399. [CrossRef]
24. Ravikumar, K.; Kumar, V. CoC-ResNet-classification of colorectal cancer on histopathologic images using residual networks. *Multimed. Tools Appl.* **2023**, *83*, 56965–56989.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.