Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 49(2), 2022, Pages 411–421, DOI: 10.52846/ami.v49i2.1641
ISSN: 1223-6934

A statistical evaluation of the preprocessing medical images impact on a deep learning network's performance

Renato Constantin Ivanescu

Abstract. The aim of this paper is to explore the efficiency of preprocessing medical images before applying a deep learning algorithm to classify the data. The study uses a statistical framework that establishes the fact that depending on the dataset used, image preprocessing indeed decreases the computational time, without having a dropdown in performance. The dataset used in this study regard colon cancer, lung cancer, and fetal brain ultrasound scans. The study proposes a statistical performance that studies the performances of the ResNet50 deep learning network in different preprocessing scenarios.

2010 Mathematics Subject Classification. Primary 60J05; Secondary 60J20.
Key words and phrases. deep learning, statistical learning, statistical analysis, colon cancer, lung cancer, fetal brain.

1. Introduction

Artificial intelligence is widely used and has a major impact in many fields of study, contributing to improving the quality of life and activity in many industries. A key research direction is image classification whose purpose is to extract features to categorize images. This direction, which had a rapid development and is advancing day by day, proved to be extremely useful, especially in the medical field. Along with technology, it provides efficient and early diagnosis and treatment to reduce the risk of a disease being fatal.

In this paper we are going two investigate whether preprocessing of medical images enhances the classification performance of different deep learning algorithms. We have tackled two types of medical issues: cancer and fetal anomalies.

Nowadays, the most common disease, with its various types, is cancer. Lung cancer is a type of cancer with a high rate of increase in cases and a mortality rate of over 1.7 million people in 2020. The chance of survival in lung cancer is 19%. The second most deadly cancer in Western Europe is colorectal cancer, but under 10% of the cases are uncured and this percent may remain low only if the disease is detected on time [3]. Lung and colon cancer exceed 25% of all cancer cases [9]. Fortunately, chances of survival for lung and colon cancer patients may increase with the effective deep learning methods that have been developed, which helps in a faster diagnosis of the disease. Regarding the maternal-fetal care, the importance of an accurate interpretation of the ultrasound, permits a priori discussion with the parents regarding the fetus' outcome (long-term mortality, morbidity, etc.). The inconsistency between the pre- and post-natal diagnosis of fetus anomalies obtained after a human

Received May 31, 2022. Accepted June 24, 2022.

performed ultrasound reach a sensitivity that ranges between 27.5% and 96%, [11]. Many factors contribute to this difference, from which we enumerate: time pressure, lack of experience, fetal movement, maternal characteristics, etc.

In the healthcare sector, deep learning, with a board application prospect, represents an important technical means for improving classification accuracy. It is also successfully applied in other fields like speech recognition or automated vehicular driving and due to its great performance, it is believed that in the future it can take place of many human-made activities [14]. Deep learning helps in extracting information from digital media, offering a key Machine learning subset for understanding and training data. At the core of deep learning is a neural network arranged in more than three layers (including input and output) [13]. One well-known architecture of deep learning, which is most likely to solve image classification problems, is convolutional neural networks.

In this paper, our aim is to investigate how the performance of a deep learning method changes if we preprocess the images before the training starts. To compare the results, we have performed a thorough statistical analysis which included power analysis, normality tests, equality of variances, and one-way ANOVA together with post-hoc Tukey test.

The paper is organized in 5 sections: section 2 describes the design and implementation of the convolutional neural network used in this study; section 3 presents the datasets used, whereas section 4 the experimental results together with the corresponding discussions. The paper ends with section 5 that discusses the conclusions.

2. Method

The convolutional neural network (CNN) is a multi-layer structure with a reduced number of weights and low complexity. The network structures consist of connected layers such as Convolution, Activation, Pooling, and fully connected layer. A similarity between convolutional neural networks and traditional neural networks refers to the fact that both of them have trainable weights. The difference is that in a CNN the neurons are not fully connected [13]. The topology of a CNN is grid-like. They are training using the backpropagation algorithm, that tunes the weights so the optimal solution is found. The activation function is the rectified linear unit or ReLU. ReLU's formula is, [5]:

$$f(x) = max(x, 0).$$

In an CNN a layer is arranged 3-dimensionally, having a width, a height, and a depth (the 3 color channels).

Such a convolutional neural network-based architecture is ResNet50. ResNet50 is the short name for Residual Network 50. In 2015, ResNet50 won the ImageNet competition. It consists of 48 convolution layers together with one Max Pooling layer and one Average Pooling layer. The advantage of using this residual network is the reduced training errors due to the addition of shortcut connection and usage of the residual functions [4].

The model training was conducted using Python3 and Keras, a high-level neural networks *Application Programming Interface*. Python notebook was run in Google Collaboratory, a *Software as a Service* provided by Google which provides powerful GPUs and High RAM Memory. The architecture used is ResNet50 along with the

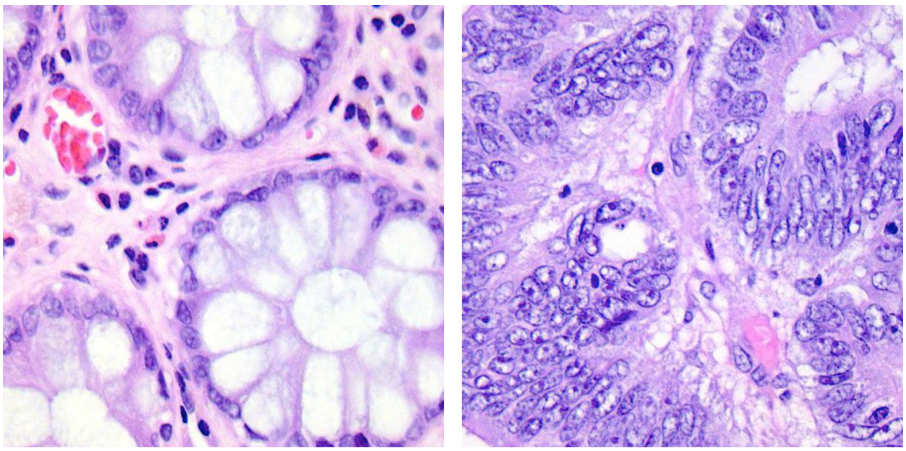

FIGURE 1. (a) Benign colon tissue, (b) Adenocarcinoma colon tissue.

specific configuration parameters. Ten epochs were chosen and a batch size of 128 is used for this work. We considered two scenarios for the weights: one with random initialization and one with imagenet (pre-training on ImageNet). Three input shapes for the image dataset were tested: 128x128, 224x224, 512x512. ResNet was run at first on color images, and secondly on gray-scale images.

3. Datasets

3.1. Description of the datasets. The dataset used for experimentation originates from the publically available data set at the Kaggle website: https://www.kaggle.com/andrewmvd/lung-and-colon-cancer-histopathological-images. It contains 25.000 histopathological images in jpeg format. The data set initially contained 1250 images collected by the authors (250 images of each type) but was expanded to 25000 images by using image augmentation techniques. After applying this technique, images were cropped from 1024x768 pixels to a square size of 768x768 pixels. There are five classes for classification divided into two folders: colon (Cc) (colon adenocarcinoma, colon benign tissue) and lung (Lc) (lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma).

The second dataset (Fetal) used contained ultrasound images of the fetal brain (https://www.kaggle.com/datasets/rahimalargo/fetalultrasoundbrain). The ultrasounds were performed by sonographers with different experience. The images were obtained using Voluson E6 (GE Medical Systems, Zipf, Austria), Voluson S8, Voluson S10, and Aloka (Aloka Co., Ltd., Tokyo, Japan). The images were gathered from 1394 patients that underwent routine fetal ultrasound. The images were used to determine the view plane which was divided in three decision classes trans-thalamic, trans-ventricular, and trans-cerebellum.

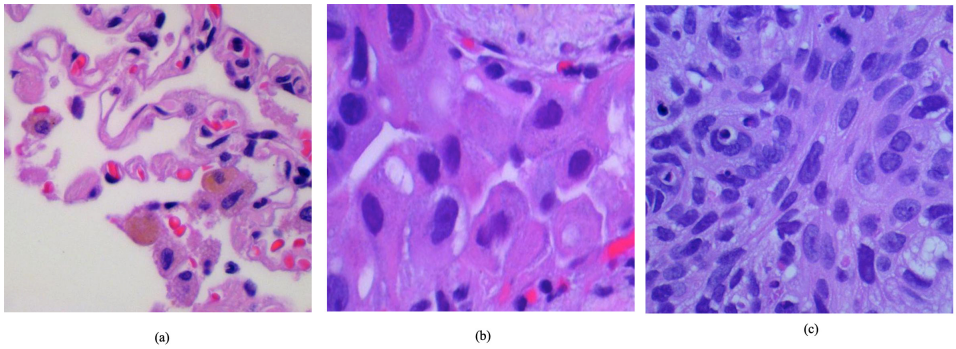

FIGURE 2. (a) Benign lung tissue, (b) Adenocarcinoma lung tissue, (c) Squamous cell lung tissue.

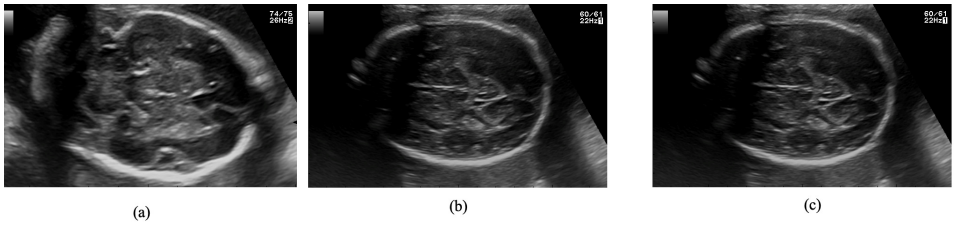

FIGURE 3. (a) trans-cerebellum, (b) trans-ventricular, (c) trans-thalamic.

3.2. Literature Review. Sanidhya Mangal proposed a computer-aided diagnosis system using a shallow neural network for classifying lung and colon cancers, recording accuracy of 97% and 96% respectively, [12]. The training and evaluation strategy proposed, allows the use of high-resolution textured images without having to convert them. The pictures from the dataset were resized to 150x150 pixels containing 3 color channels. Randomized share, zoom transformation, and normalization were also applied. The CNN was trained for 100 iterations and constructed with the next layers: input, convolution, pooling, flatten, fully connected layer or dense layer and dropout layer.

Jiatai Lin et al. proposed a plug-and-play module (Pyramidal Deep-Broad Learning (PDBL)) for improving classification performance for any well-trained classification backbone, [8]. It was tested on the lung and colon cancer histopathological image dataset with three popular architectures: ResNet50, ShuffLeNetV2, EfficientNet2 to evaluate the efficiency. It was demonstrated that tissue-level classification can be improved for any CNN backbones, especially when giving a small number of training samples in the case of lightweight models.

Mehedi Masud et al. proposed a framework capable of identifying various types of lung and colon cancer with a maximum of 93% accuracy, [9]. The classification was performed using a multi-channel CNN. A feature set was formulated from the features extracted from the image data set. Two image transformation techniques were used for feature extraction: two-dimensional Discrete Fourier Transform (2D-DFT) and Single-level discreet two-dimensional wallet transform (2D-DWT). The

Unsharp Masking method was applied for image contrast enhancement before feature extraction. The experiment was carried out for 500 epochs.

Neha Baranwal uses 4 different CNN architectures (ResNet50, VGG-19, Inception_Resnet_V2, and DenseNet) for the tri-category classification of lung cancer images to compare which architecture gives better performance for the dataset, [10]. Increased accuracy was shown by all four CNN models, over 90%, when using images resized to 124x124 with 3 channels and a batch size of 16.

Xie et al. used deep learning algorithm to classify the fetal brain ultrasound as normal or abnormal and obtained 97.9% precision, and 90.9% recall. The overall accuracy was 96.3%, [15]. Gofer et al. used statistical region merging and trainable weka segmentation on the fetal brain images and obtained a mean absolute percentage error of 1.71%.

4. Statistical performance assessment

ResNet50 is a stochastic algorithm. Thus, to truly establish if the results are robust, we need to run the algorithm a certain number of times. We have achieved a suitable statistical power (two-tailed type of null hypothesis with power goal p$\geq$95% and type I error α=0.05) for 100 computer runs. We have recorded the average accuracy (ACA) and running time. Besides the ACA, we have also computed the standard deviations of the accuracies (SD) together with the 95% confidence interval. Through this we wanted to demonstrate whether the model offers *omnibus* robustness, [16].

In order to apply one-way ANOVA together with post-hoc Tukey, we needed first to verify the normality of the sample of ACAs. We have applied the *Kolmogorov-Smirnov* & *Lilliefors* and *Shapiro-Wilk W* tests, [6][7]. We have run ResNet50 on preprocessed data in the same conditions: 100 independent computer runs in a complete 10-fold cross-validation cycle. The algorithm was run on the dataset with images sizes of 512×512, 224×224, and 128×128.

Besides the data normality, we also checked whether the sample ACAs had equal variances. If the samples do not have equal variances, then we are dealing with the heteroscedasticity phenomenon. This phenomenon produces the Type I error, that is creates false positives, [1][2]. Because all the sample sizes are equaling 100, we can presume that they variances are equal.

The performance results of the ResNet50 over 100 computer runs, in terms of average accuracy, stability (SD), 95% confidence interval (CI) accuracy, and time are displayed in Table 1 for the Cc dataset, Table 2 for the Lc dataset, and Table 3 for Fetal dataset.

Dataset	ACA%	SD	95%	Time (s)
color 512x512 random	0.661	0.098	(0.64, 0.68)	5151.822
color 224x224 random	0.659	0.084	(0.64, 0.67)	1927.072
color 128x128 random	0.628	0.106	(0.60,0.65)	1305.306
color 512x512 ImageNet	0.817	0.013	(0.81,0.82)	5012.837
color 224x224 ImageNet	0.764	0.008	(0.763, 0.766)	1786.683
color 128x128 ImageNet	0.706	0.012	(0.703, 0.708)	1328.68
gray-scale 512x512 random	0.508	0.017	(0.504,0.511)	5047.25
gray-scale 224x224 random	0.528	0.059	(0.516,0.54)	1624.79
gray-scale 128x128 random	0.573	0.108	(0.55,0.59)	895.152
gray-scale 512x512 ImageNet	0.819	0.018	(0.81,0.82)	4890.078

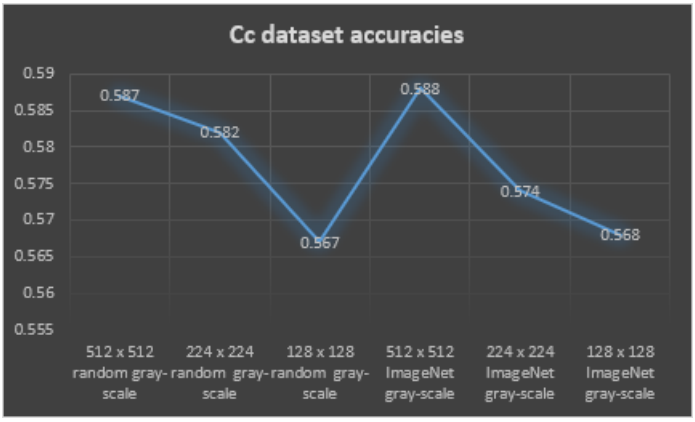

FIGURE 4. Cc accuracies.

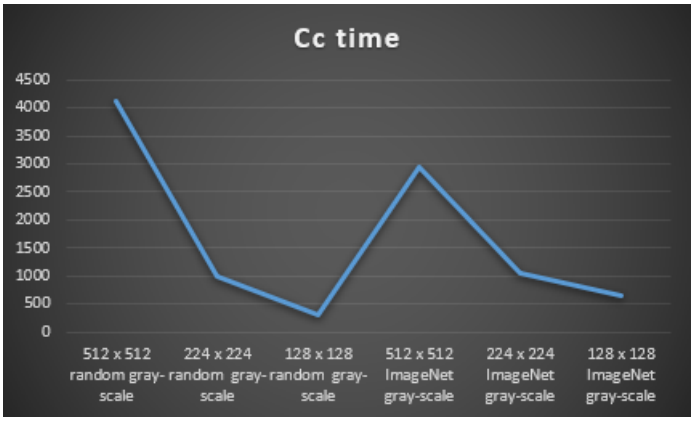

FIGURE 5. Cc times.

gray-scale 224x224 ImageNet	0.769	0.008	(0.76, 0.77)	1474.152
gray-scale 128x128 ImageNet	0.708	0.011	(0.70,0.71)	996.903

Table 1. ResNet50 performance indicators for Cc dataset.

From Table 1, Figure 4 and Figure 5, we can see that the highest accuracies are obtained when using images with 512x512 size, color, and ResNet 50 has been trained using ImageNet (81.7%), and when using gray-scale images, 512x512, and ResNet50 pretrained by ImageNet (81.9%). Also, we can see that the time improved by over 1000 seconds when using gray-scale images, and the accuracy improved by 0.2%. The SDs values show that the model is stable in any situation.

Dataset	ACA%	SD	95%	Time (s)
color 512x512 random	0.782	0.035	(0.77, 0.78)	7670.703
color 224x224 random	0.762	0.041	(0.75, 0.77)	2892.145
color 128x128 random	0.762	0.041	(0.75,0.77)	1649.009

color 512x512 ImageNet	0.760	0.040	(0.75,0.76)	8339.095
color 224x224 ImageNet	0.688	0.043	(0.68, 0.69)	2625.344
color 128x128 ImageNet	0.661	0.027	(0.65, 0.66)	1867.016
gray-scale 512x512 random	0.642	0.017	(0.63,0.64)	7834.65
gray-scale 224x224 random	0.505	0.064	(0.49,0.51)	2310.15
gray-scale 128x128 random	0.520	0.082	(0.49,0.53)	1248.57
gray-scale 512x512 ImageNet	0.510	0.065	(0.49,0.52)	7912.199
gray-scale 224x224 ImageNet	0.623	0.035	(0.61, 0.63)	2074.284
gray-scale 128x128 ImageNet	0.580	0.036	(0.57,0.58)	1358.687

Table 2. ResNet50 performance indicators for Lc dataset.

From Table 2, Figure 6 and Figure 7, we can see that the best performances are obtained when using color images, random weights, with different image sizes 512x512, 224x224, and 128x128, and color images, pretrained network and 512x512 size. Interesting enough, is that when we are reducing the images sizes to 128x128,

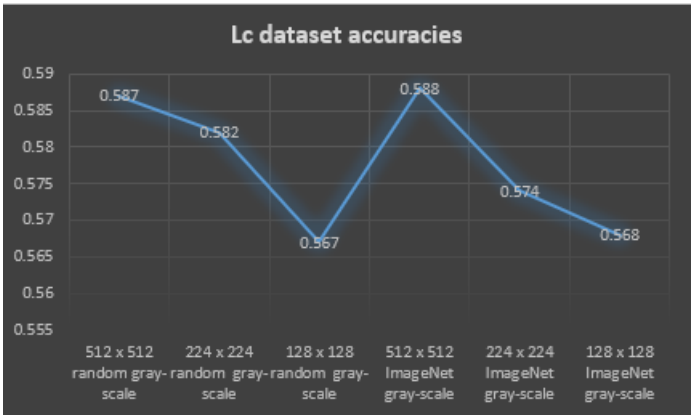

FIGURE 6. Lc accuracies.

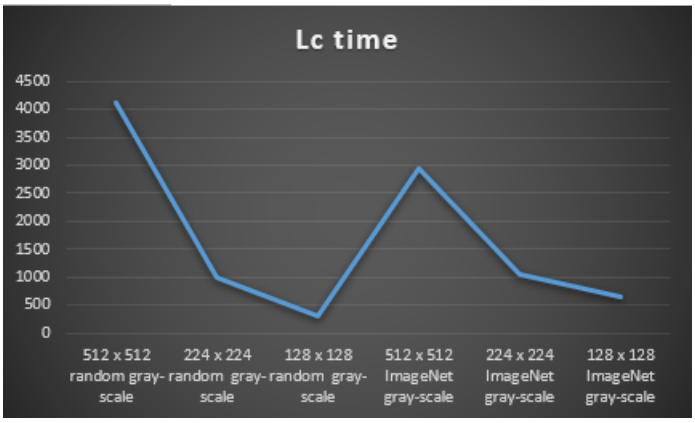

FIGURE 7. Lc time.

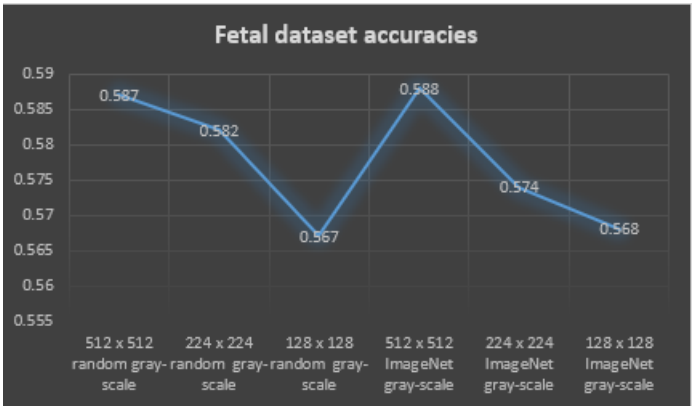

FIGURE 8. Fetal accuracies.

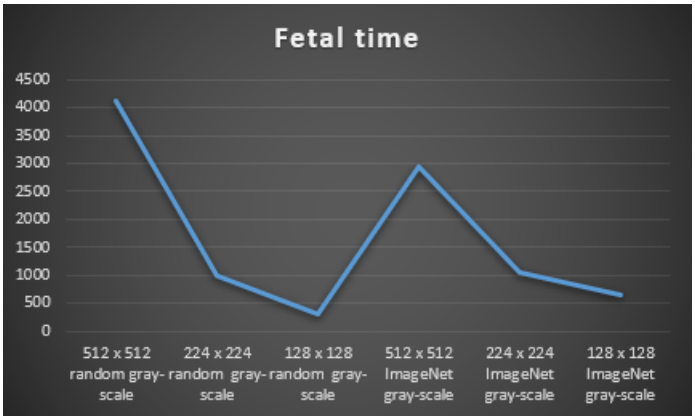

FIGURE 9. Fetal time.

the accuracy is similar, but the time decreases by almost 7000 seconds, proving that indeed preprocessing data is necessary. The SDs values show that the model is stable in any situation.

Dataset	ACA%	SD	95%	Time
gray-scale 512x512 random	0.587	0.0138	(0.587, 0.588)	4109.04
gray-scale 224x224 random	0.582	0.0103	(0.585, 0.589)	982.931
gray-scale 128x128 random	0.567	0.001	(0.5673,0.5688)	294.384
gray-scale 512x512 ImageNet	0.588	0.001	(0.586,0.591)	2953.93
gray-scale 224x224 ImageNet	0.574	0.012	(0.572, 0.576)	1058.15
gray-scale 128x128 ImageNet	0.568	0.011	(0.566, 0.569)	652.886

Table 3. ResNet50 performance indicators for Fetal dataset.

From Table 3, and Figure 8 and Figure 9, we can see that no matter the preprocessing method used, the model performs poorly. The only remarkable difference can be seen in terms of computational time, the biggest difference being around 4000 seconds.

Variable	Kolmogorov-Smirnov		Shapiro-Wilk W	
	K-S max D	Liliefors-p	S-W W	p-level
color 512x512 random	0.192	0.0000	0.847	0.0000
color 224x224 random	0.117	0.0000	0.916	0.0000
color 128x128 random	0.182	0.0000	0.849	0.0000
color 512x512 ImageNet	0.169	0.0000	0.887	0.0000
color 224x224 ImageNet	0.101	0.0000	0.949	0.0000
color 128x128 ImageNet	0.220	0.0000	0.838	0.0000
gray-scale 512x512 random	0.435	0.0000	0.528	0.0000
gray-scale 224x224 random	0.396	0.0000	0.531	0.0000
gray-scale 128x128 random	0.305	0.0000	0.695	0.0000
gray-scale 512x512 ImageNet	0.178	0.0000	0.846	0.0013
gray-scale 224x224 ImageNet	0.118	0.0000	0.918	0.0000
gray-scale 128x128 ImageNet	0.192	0.0000	0.830	0.0000

Table 4. Testing the normality Cc.

Variable	Kolmogorov-Smirnov		Shapiro-Wilk W	
	K-S max D	Liliefors-p	S-W W	p-level
color 512x512 random	0.145	0.0000	0.921	0.0000
color 224x224 random	0.142	0.0000	0.914	0.0000
color 128x128 random	0.142	0.0000	0.916	0.0000
color 512x512 ImageNet	0.201	0.0000	0.885	0.0000
color 224x224 ImageNet	0.089	0.0000	0.960	0.0000
color 128x128 ImageNet	0.120	0.0000	0.931	0.0000
gray-scale 512x512 random	0.136	0.0000	0.942	0.0000
gray-scale 224x224 random	0.116	0.0000	0.941	0.0000
gray-scale 128x128 random	0.138	0.0000	0.917	0.0000
gray-scale 512x512 ImageNet	0.151	0.0000	0.953	0.0013
gray-scale 224x224 ImageNet	0.139	0.0000	0.909	0.0000
gray-scale 128x128 ImageNet	0.191	0.0000	0.857	0.0000

Table 5. Testing the normality Lc.

Variable	Kolmogorov-Smirnov		Shapiro-Wilk W	
	K-S max D	Liliefors-p	S-W W	p-level
gray-scale 512x512 random	0.271	0.0000	0.78	0.0000
gray-scale 224x224 random	0.127	0.0000	0.930	0.0000
gray-scale 128x128 random	0.335	0.0000	0.715	0.0000
gray-scale 512x512 ImageNet	0.189	0.0000	0.868	0.0013
gray-scale 224x224 ImageNet	0.357	0.0000	0.715	0.0000
gray-scale 128x128 ImageNet	0.320	0.0000	0.854	0.0000

Table 6. Testing the normality Fetal.

From Table , we can see that the samples are not governed by the Gaussian distribution. Nevertheless, this issue can be resolved due to the Central Limit Theorem, that states that if the sample size is above 30, then the distribution of the sample is approximately normal. Since our samples have 100 as size, we presume that they are normally distributed.

After applying the One-Way ANOVA technique, we have used the Tukey's honestly significant difference (Tukey HSD) *post-hoc* test, so that we could highlight the statistically significant differences between the ResNet50's performance, on each dataset.

The groups used for ANOVA were the classification accuracies of each ResNet50 obtained in 100 computer runs (complete 10-fold cross-validation cycle) on each dataset. The One-way ANOVA and Tukey HSD were performed using Statistica StatSoft package. The ANOVA output is displayed in Table X, and depict the combined sums of squares (SS), degrees of freedom (df), mean squares (MS), F-value, and p-level (contrasts: quadratic polynomial).

Dataset	SS	df	MS	F-value	p-level
Cc	12.24	11	1.112	298.1	0.0000
Lc	11.922	11	1.083	478.8	0.0000
Fetal	0.045	5	0.009	93	0.0000

Table 7. One-way ANOVA results.

From Table 7, we can see that there are significant differences (p-level < 0.05) between the performances of the ResNet50 depending on the manner the dataset had been preprocessed. To find out between which samples are the differences we have applied Tukey HSD.

The *post-hoc* Tukey HSD test revealed no statistically significant differences in classification performances (p-level < 0.05) in the following cases:

• On the Cc dataset: between the preprocessed data image size 128x128, color, pretrained on ImageNet vs. image size 224x224, gray-scale, random weights, 512x512 color pretrained on ImageNet vs. 512x512 gray-scale pretrained on ImageNet, 224x224 color pretrained on ImageNet vs 224x224 gray-scale pretrained on ImageNet, and 128x128 color pretrained on ImageNet vs 128x128 gray-scale pretrained on ImageNet.

• On Lc dataset: between 224x224 color random vs. 512x512 color random, 224x224 color random vs. 128x128 color random, 512x512 color random vs. 128x128 color random, 128x128 color pretrained on ImageNet vs. 512x512 gray-scale random, 224x224 gray-scale random vs 128x128 gray-scale random, 512x512 gray-scale pretrained on ImageNet vs. 224x224 gray-scale random, and 512x512 vs. 128x128 gray-scale random, 224x224 gray-scale pretrained on ImageNet vs. 512x512 gray-scale random.

• On the Fetal dataset: between 512x512 gray-scale pretrained with Imaget vs. 512x512 gray-scale random, 224x224 gray-scale pretrained with ImagetNet vs 128x128 gray-scale random.

5. Conclusions

In this study, we were interested in seeing whether preprocessing medical images improves the computational time, without losing to much in performance. We have applied ResNet50 deep learning neural network on three publicly available datasets. At first we wanted to see if the performance is lost if we transform the images from color to gray-scaler, and afterwards if we reduce their sizes. The network used random weights as well as pretrained weights on ImageNet. The statistical analysis showed that in some cases, indeed transforming the images into grayscale does not modify the performance but decreases the computation size (Cc dataset). On the Lc dataset, the computational time decreased by 7000 seconds, while the performance remained the same. On the fetal dataset, we cannot draw important conclusions, since the model performs poorly in any case.

6. Acknowledgment

This work was supported by a grant of the Ministry of Research Innovation and Digitization, CNCS - UEFISCDI, project number PN-III-P4-PCE-2021-0057, within PNCDI III.

References

[1] S. Belciug, *Artificial Intelligence in Cancer: diagnostic to tailored treatment*, Elsevier, 2020.
[2] S. Belciug and F. Gorunescu, A hybrid genetic algorithm-queuing multi-compartment model for optimizing inpatient bed occupancy and associated costs, *Artificial Intelligence in Medicine* **68** (2016), 59–69. DOI: 10.1016/j.artmed.2016.03.001
[3] A. Bour, et al., Automatic colon polyp classification using Convolutional Neural Network: A Case Study at Basque Country, *IEEE International Symposium on Signal Processing and Information Technology* (2019).
[4] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, arxiv.org/abs/1512.0338 (2015).
[5] F. Koehler and A. Risteski, Representational Power of ReLU Networks and Polynomial Kernels: Beyond Worst-Case Analysis, arXiv:1805.11405 (2018).
[6] F. Gorunescu, et al., A statistical framework for evaluating neural networks to predict recurrent events in breast cancer, *International Journal of General Systems* **39** (2010), no. 5, 471–488. DOI: 10.10180/03081079.2010.484282
[7] F. Gorunescu,, et al., An evolutionary computational approach to probabilistic neural network with application to hepatic cancer diagnosis, *18th IEEE Symposium on computer-based medical systems* (2005), 461–466. DOI: 10.1109/CBMS.2005.24
[8] L. Jitai, et al., PDBL: Improving Histopathological Tissue Classification with Plug-and-Play Pyramidal Deep-Broad Learning, *IEEE Trans Med Imaging* (2022).
[9] M. Masud, et al., A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework, *Sensors (Basel, Switzerland)* **21** (2021), no. 3, 748.
[10] B. Neha, et al., Classification of Histopathology Images of Lung Cancer Using Convolutional Neural Network (CNN), arXiv:2112.13553 (2021).
[11] L. Salomon, et al., A score-based method for quality control of fetal images at routine second trimester ultrasound examination, *Prenatal Diagnosis* **28** (2016), no. 9, 822–827.
[12] M. Sanidhya, et al., Convolution Neural Networks for diagnosing colon and lung cancer histopathological images, arXiv:2009.03878 (2020).
[13] M. Sheldon and A. Mukul, A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification, *International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications* (2021).
[14] M. Suganthi and J.G.R. Sathiaseelan, An Exploratory of Hybrid Techniques on Deep Learning for Image Classification, *4th International Conference on Computer, Communication and Signal Processing* (2020).
[15] H.N. Xie, et al., Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal, *Ulstr. Obstet Gynecol.* (2020). DOI: 10.1002/uog.21967
[16] Z.H. Zhang, et al., Variable selection in logistic regression model with genetic algorithm, *Annals of Translational Medicine* **6** (2018), no. 3. DOI: 10.21037/atm.2018.01.15

(Renato Constantin Ivanescu) Department of Computers and Information Technologies, University of Craiova, Blvd Decebal 107, Craiova, 200776, Romania
E-mail address: constantin.ivanescu@edu.ucv.ro