Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 48(1), 2021, Pages 159–166
ISSN: 1223-6934

A statistical framework for evaluating convolutional neural networks. Application to colon cancer

Liliana Popa

Abstract. *Purpose*: Explore the efficiency of two convolutional neural networks in helping physicians in establishing colon cancer diagnosis from histopathological image scans.

Methods: The dataset used in this study contains 357 histopathological image slides that ranged from benign cases to colon cancer grade three. The slides were collected by doctors at the Emergency Hospital of Craiova, Romania. The study proposes a statistical framework that studies the performances of two convolutional neural networks AlexNet and GoogleNet.

Results: AlexNet has revealed a competitive accuracy in comparison with GoogleNet. To prove the robustness of the AlexNet in fair terms, we have performed a thorough statistical analysis of its performance.

Conclusions: On this particular dataset which contains histopathological image scans regarding colon cancer, the convolutional neural network AlexNet proved to be superior to GoogleNet.

2020 *Mathematics Subject Classification.* Primary 68T07; Secondary 62L99.
Key words and phrases. convolutional neural networks; statistical analysis, colon cancer; t-test, normality tests, equality of variances.

1. Introduction

Colon cancer represents the fourth most diagnosed cancer around the globe, and also it occupies the fifth place in cancer deaths, 5.8%. Lifestyle factors such as sedentarism, red meat consumption, alcohol, smoking, etc. contribute on colon cancer?s incidence. Besides lifestyle choices, a hereditary predisposition must be also taken into account [1]. Colon cancer appears when epithelial cells from the large intestine suffer mutations [2]. Its incidence and mortality is split into three categories around the world: the low incidence and low mortality group, which comprises the United States of America, Iceland, Japan, and France due to prevention and treatment; the high incidence and low mortality group, which comprises Canada, the United Kingdom, Denmark, and Singapore, due to improved treatment; and lastly, the high incidence and low mortality group, which comprises Brazil, Russia, China, Latin America, the Baltics, and Philippines [3]. According to GLOBOCAN 2018, in Romania, colon cancer ranks the second place in cancer rates (13.3%).

However, in recent years, deep learning has been intensively used in cancer diagnosis, showing that it is superior to other state-of-the-art machine learning methods when dealing with images. In [4], the diagnosis of skin cancer was achieved by using a convolutional neural network enhanced by an improved whale optimization algorithm

Received January 31, 2021. Accepted March 21, 2021.

based on Lvy flight. Deep learning was used to improve the detection of breast cancer, [5], and prostate cancer, [6]. A hybrid spiral optimization intelligent-generalized rough set approach improved a deep neural network in order to automatically detect lung cancer from CT images, [7]. In what regards colon cancer diagnosis, we mention the following studies: in [8] the authors used a deep learning algorithm to analyze whole-slide images and provide an automatic prognostic biomarker for primary colon cancer; in [9], deep learning was used to predict the survival rate of patients using histology slides; in [10] both a convolutional neural network (CNN) and a recurrent neural network were used to classify biopsy histopathology whole-slide images of stomach and colon cancers.

In this paper, our aim is to investigate two types of convolutional neural networks: AlexNet and GoogleNet. The statistical analysis regarded power analysis, normality tests, equality of variances, t-test for independent samples

The paper is organized in 5 sections: section 2 briefly describes the design and implementation of the two CNNs; section 3 the benchmarking dataset and the statistical framework used for assessment; section 4 the presents the experimental results and corresponding discussions. The paper ends with section 5 which deals with conclusions.

2. Method

CNN architecture

CNNs have been developed by Yan LeCun et al. [11]-[12] -[13]. They are similar to other NNs, containing neurons that are trained by updating their weights. CNNs have a grid-like topology. They have multiple layers of neurons that can process complex features. By using convolutions, which are a special linear operation, makes the CNNs just like NNs that use convolutions instead of matrix multiplication [14].

A layer in a CNN has the neurons arranged 3-dimensionally (width, height and depth i.e. the 3 color channels, red, green and blue), and not all the neurons are interconnected. There are 3 types of layers: convolutional, pooling and fully-connected layer. These layers transform a 3-dimensional input into a 3-dimensional output using differential activation function.

In general, the classical backpropagation (BP) algorithm is used for training a CNN [15], [16]. Training a CNN implies tuning the weights in such a manner that an optimal solution is produced for a specific problem, technically regulating them to minimize the error function, frequently the sum of squared errors (SSE), computed as the difference between the network`s output and the ground truth, summed over all the output nodes and training data. Evaluating the derivatives of the SSE with respect to the weights, we can minimize the error function by using some optimized values of the weights, that correspond to the errors`backpropagation through the network.

As activation function the rectified linear unit (ReLU) is preferred, to other well-known functions. ReLU formula is given by [17]

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

In our study we have considered AlexNet [18] and GoogleNet [19]. AlexNet is a pretrained convolutional neural network. The dataset that it has been trained on

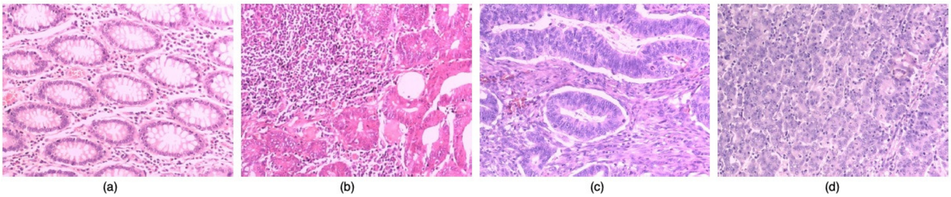

FIGURE 1. (a) healthy tissue; (b) grade I cancer; (c) grade II cancer; (d) grade III cancer.

is ImageNet, and it is publicly available at www.image-net.org. Its architecture contains 8 layers and can classify objects from 1000 classes. The input is an image of 227x227 pixels, with a 32-bit RGB color space. In our study we have replaced the last layers of the network and also resized the original images in order to classify the objects from our dataset. The second neural network, GoogleNet, has 22 layers. Just like AlexNet, GoogleNet is a pretrained network. The training is performed also in ImageNet. Different from AlexNet, it uses 224x224 pixel input images. Just like in the first case, we have resized and replaced the last layers in the network so that we could classify our objects.

3. Benchmarking dataset and statistical framework

3.1. The dataset. The two models were evaluated on a real-world medical test case. The dataset contains 357 histopathological slides, that range from benign to grade three cancer [20], [21]. All the images have been resized. The data is distributed as follows: 62 benign cases, 96 grade I cancer, 99 grade II cancer, and 100 grade III cancer. The data can be found at: https://doi.org/10.6084/m9.figshare.4508672.v1.

Figure 1 depicts a sample of each cancer grade.

3.2. Statistical performance assessment. Both networks are stochastic algorithms. Therefore, in order for the reported results to be effective and robust, the algorithms must be run a certain number of runs. To compute the correct number of runs, the sample size, we have used statistical power analysis. By independently running the algorithm 106 times, we have achieved a suitable statistical power (two-tailed type of null hypothesis with default statistical power goal $P \geq 95\%$ and type I error $\alpha = 0,05$ as the level of significance). The standard 10-fold cross-validation has been used [22]. We have computed the average accuracy (ACA) as the percentage of correctly classified cases during the testing phase. Besides the ACA, the standard deviation (SD) of the ACA and the 95% confidence interval were computed also. We wanted to see whether the models offer or not omnibus robustness.

For comparing their performances, we first checked whether the sample of ACAs of both algorithms have normal distributions or not. Therefore, the *Kolmogorov-Smirnov & Lilliefors* and *Shapiro-Wilk W* tests were performed. We have used the *Kolmogorov-Smirnov & Lilliefors* because we can compute the mean and the standard deviation from the actual data, and the Shapiro-Wilk test because it has better power properties.

In the beginning of the statistical assessment process, we focused on the stability and robustness of the algorithm. As we proceed further, we are interested in the more complex statistical tests that are able to discriminate between the two neural networks. Comparing only the ACAs would be subjective and simplistic. Hence, both algorithms were run in the same conditions: 106 computer runs in a complete 10-fold cross-validation cycle. Obviously, we are dealing with two independent groups of samples, thus our concern refers to the mean difference between the performances of the two algorithms, because there is an obvious variability between the 106 computer runs. Thus, we have performed the t-test for independent samples to compare the difference in means. Besides the t-test we were also interested to see the equality of variances, so we performed the Levene and the Brown-Forythe tests. Recall that if we use a statistical test such as the t-test, we first must verify the following assumptions: the two populations must have an approximately Gaussian distribution, and their variances have to be equal. If the variances are unequal, then we are dealing with the heteroscedasticity phenomenon, which might lead to false positives affected by the Type I error rate.

The results of all these tests and their corresponding discussion are presented in the next section.

4. Results and discussion

The AlexNet and GoogleNet performance results in terms of ACA over 106 computer runs, stability (SD) and 95% confidence interval (CI) are presented in the Table 1.

Table 1. Testing the normality of the AlexNet and GoogleNet ACAs.

Variable	ACA (%)	SD	95% CI
AlexNet	89.53	0.28	(89.10 , 90.59)
GoogleNet	85.62	3.67	(79.29 , 90.98)

We can see from Table 1 that on average the AlexNet performs better than Google Net on this dataset, almost 5% gain in average accuracy. Regarding the stability of the model, the SD for AlexNet is 0.28, whereas for the GoogleNet is 3.67, proving the fact that both models are omnibus. As we have mentioned before, we performed two normality tests, the *Kolmogorov-Smirnov & Lilliefors* test and *Shapiro-Wilk W test*, to verify the assumption that the sample data has a Gaussian distribution. This gives us insights regarding the existence of outliers, which affect the results. The results of the two tests are presented in Table 2.

Table 2 reveals interesting results. While the *Kolmogorov-Smirnov & Lilliefors* test indicates that the two samples are normally distributed, ($p - level > 0.05$), the *Shapiro-Wilk W* test indicates that both samples are not governed by the Normal distribution. The *Kolmogorov-Smirnov* test is quite sensitive to extreme values, but the *Lilliefors* corrects this [22], and also it has been reported that this test's power is not high enough [23]. The *Shapiro-Wilk W* test provides better power because it used the correlation between the data and the corresponding normal scores. Hence,

Table 2. AlexNet and GoogleNet performance indicators (ACA, SD, 95% CI).

	Kolmogorov-Smirnov		Shapiro-Wilk W CI	
Variable	K-S max D	Lilliefors p	S-W W	p-level
AlexNet	0.088	0.15	0.939	0.0006
GoogleNet	0.579	0.2	0.947	0.002

it is preferred. Nevertheless, to tackle this issue, we have used the *Central Limit Theorem*, that states that if the sample size increases above 30, then the sample distribution becomes approximately normal [24]. Since the sample size is 106, then, we can presume the normality of data.

Next, we have used the *Levene's* and *Brown-Forsythe* test to verify the equality of variances. The results are depicted in Table 3.

Table 3. Testing the equality of variances of the eCNN and CNN ACAs.

Variable	AlexNet vs. GoogleNet
Levene($F(1, df)/p-level$)	198.17 / 0.000
Brown-Forsythe($(1/df)/p-levelCI$)	197.91 / 0.000

Table 3 reveals the fact that the two algorithms do not have equal variances, so they do not have similar behaviors ($p-level < 0.05$). Still, this issue can be avoided, becoming less problematic due to the fact that we are using samples that have the same number of observations. We can thus proceed with the t-test for independent samples to compare the difference in means between the two competitors. The results shown in Table 4, that indeed there are significant differences in means ($p-level < 0.05$) between AlexNet and GoogleNet concerning the testing performances, hence the power of classifying correctly colon cancer from histopathological images of AlexNet over GoogleNet has been proven. It is only fair to sustain, that the 5% accuracy gain of the AlexNet is not statistically insignificant.

Table 4. Comparing testing performances (t-test).

Variable	t-test / p-level
AlexNet vs. GoogleNet	-9.709 / 0.000

Since AlexNet proved to be superior to GoogleNet in this particular case, we have continued the statistical benchmark only on AlexNet. Therefore, we provided also the corresponding confusion matrix and three other classification parameters: precision or positive predictive value, recall or sensitivity, and F1-score (Dice similarity coefficient). The confusion matrix gives valuable information regarding misclassification. Each cell of the matrix contains a number that shows how many cases were indeed classified correctly from each class. Precision is the ratio of cases with 'positive' results that are correctly classified. Recall measures the proportion of 'true positives' that are correctly classified as such. The F1-score is the harmonic mean of the precision and recall [25].

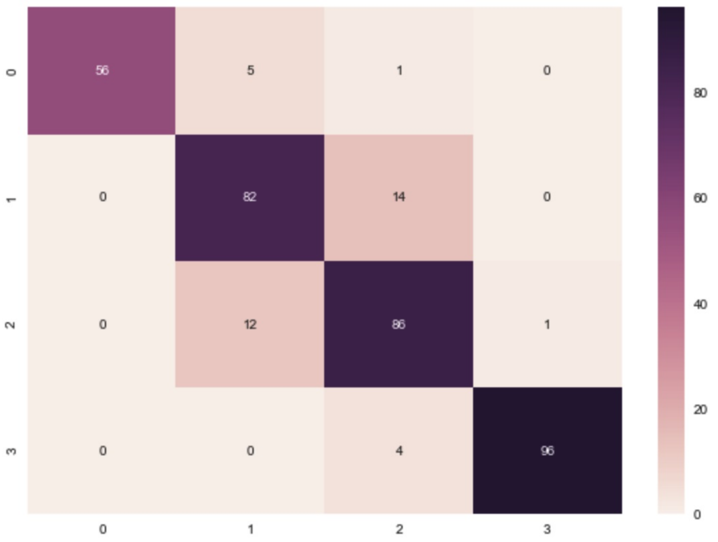

FIGURE 2. Confusion matrix heatmap.

The confusion matrix heatmap is presented in Figure 2.

$$Precision = \frac{Truepositives}{Truepositives + Falsepositives} \tag{1}$$

$$Recall = \frac{Truepositives}{Truepositives + Falsenegatives} \tag{2}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

Since in general the three parameters are computed for binary classes, we have presented their value for each class in particular, as well as the overall value. The values are depicted in Table 5 and Table 6.

TABLE 5. Precision and recall per class (AlexNet).

	Grade 0	Grade 1	Grade	Grade 3
Precision	1	0.82	0.81	0.92
Recall	0.90	0.85	0.86	0.96

From both tables, we can see that the values of these statistical parameters are close to 0.90, which means that the classifier indeed performs very well.

5. Conclusions and future work

Two convolutional neural networks are statistical benchmarked in order to see which one of them performs better at classifying colon cancer from histopathological images.

Table 6. Overall precision, recall, and F1-score (AlexNet).

	Colon cancer
Precision	0.90
Recall	0.89
F1-score	0.89

According to our statistical analysis, AlexNet surpassed GoogleNet on this particular dataset, providing a significantly increased gain in accuracy.

Our study regarded a statistical framework for determining which convolutional neural network (AlexNet or GoogleNet) performs better when applied on a colon cancer dataset which contains histopathological images. The benchmarking results in terms of ACA, SD, 95% CI, confusion matrix, precision, recall, F1-score, Kolmogorov-Smirnov, Shapiro-Wilk, Levene, Brown-Forsythe and t-test, showed a significance gain in terms of accuracy of AlexNet, when compared with GoogleNet.

References

[1] P. Rawla, T. Sunkara, and A. Barsouk, Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors, *Prz Gastroenterol.* **14** (2019), no. 2, 89–103.

[2] I. Ewing, J.J. Hurley, E. Josephides, and A. Millar, The molecular genetics of colorectal cancer. Frontline Gastroenterol, *Frontline Gastroenterol.* **5** (2014), 26–30.

[3] [DOI: 10.1136/gutjnl-2015-310912] M. Arnold, M.S Sierra, M. Laversanne, et al., Global patterns and trends in colorectal cancer incidence and mortality, *Gut.* **66** (2017), 683–691.

[4] [doi: 10.1016/j.artmed.2019.101756] N. Zhang, Y-X. Cai, Y.-Y. Wang, Y.-T. Tian, X.-L. Wang, and B. Badami, Skin cancer diagnosis based on optimized convolutional neural network, *Art. Intel. Med.* **102** (2020), 101756.

[5] L. Shen, L. Margolies, J. Rothstein, E. Fluder, R. McBride, and W. Sieh, Deep learning to improve breast cancer detection on screening mammography, *Scientific Rep.* **9** (2019), 12495.

[6] S. Yoo, I. Gujrathi, M. Haider, and F. Khalvati, Prostate Cancer Detection using Convolutional Neural Networks, *Scientific Rep.* **9** (2019), 19518.

[7] M. Shakeel, M.A. Burhanuddin, and M.I. Desa, Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier, *Neural Computing and Applications* (2020), 1–14.

[8] [doi: 10.1016/S0140-6736(19)32998-8] O.J Skrede, S. Raedt, A. Kleppe, T.S Hveem, K. Liestol, and J. Maddison, Deep learning for prediction of colorectal cancer outcome: a discovery and validation study, *The Lancet* **395** (2020), 10221, 350–360.

[9] [doi: 10.1371/journal.pmed.1002730] J.N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. Valous, D. Ferber, L. Jansen, C.C. Reyes-Aldasoro, I. Zorning, D. Jager, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study,*PLoS Med.* **16** (2019), no. 1, e1002730.

[10] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, and M. Tsuneki, Deep learning models for histopathological classification of gastric and colonic epithelial tumors, *Scientific Rep* **10** (2020), 1504.

[11] Y. LeCun, Generalization and network design strategies, Technical Report CRG-TR-89-4, University of Toronto (1989).

[12] [doi: 10.1109/5.726791] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition,*Proceedings of the IEEE* **86** (1998), no. 11, 2278–2324.

[13] [doi: 10.1007/978-3-642-35289-8_3] Y. LeCun, L. Bottou, G. Orr, and K.-L. Muller, Efficient BackProp, In *Montavon G., Orr G.B., Mller KR. (eds) Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science* 7700, Springer, Berlin, Heidelberg, (2012), 9–48.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts, London, 2016.

[15] C. Bishop, *Neural networks for pattern recognition*, Oxford University Press, UK, 1995.

[16] S. Haykin, *Neural networks, a comprehensive foundation(Second Edition)*, Prentice Hall, 1999.

[17] F. Koehler and A. Risteski, Representational Power of ReLU Networks and Polynomial Kernels: Beyond Worst-Case Analysis, arXiv:1805.11405 [cs.LG], (2018).

[18] [DOI: 10.1145/3065386] A. Krizhevsky, I. Sutskever, and G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Association for computing machinery* (ACM) **60** (2017), no. 6, 84–90.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, arXiv:1409.4842 [cs.CV], (2014).

[20] [10.1109/SYNASC.2016.074] C. Stoean, In search of the optimal set of indicators when classifying histopathological images, *18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (SYNASC), IEEE (2016), 449–455.

[21] C. Stoean, R. Stoean, A. Sandita, D. Ciobanu, C. Mesina, and C.L. Gruia, SVM-based cancer grading from histopathological images using morphological and topological features of glands and nuclei. In *Intelligent Interactive Multimedia Systems and Services 2016*, Springer, (2016), 145–155.

[22] J. Peat and B. Barton, *Medical Statistics: A guide to data analysis and critical appraisal*, Blackwell Publishing, Oxford, 2005.

[23] H.J. Thode, *Testing for normality*, Marcel Dekker, New York, 2002.

[24] D.G. Altman, *Practical Statistics for Medical Research*, Chapman and Hall, New York, 1991.

[25] S. Belciug, *Artificial Intelligence in Cancer: Diagnostic to Tailored Treatment*, Elsevier, 2020.

(Liliana Popa)
Department of Computers and Information Technology, Faculty of Automation,
Computers and Electronics, University of Craiova, Blvd. Decebal no. 107, Craiova,
200440, Romania
E-mail address: liliana.popa@edu.ucv.ro