MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220145

The Use of Convolutional Neural Networks in the Prediction of Invasive Ductal Carcinoma in Histological Images of Breast Cancer

Érika G de Assis, Zenilton K. G. do Patrocínio, Cristiane Neri Nobre

Pontifical Catholic University of Minas Gerais
Institute of Exact Sciences and Informatics, Graduate Program in Informatics
Belo Horizonte, MG, Brazil

Abstract

This paper presents a deep learning approach for automatic detection and visual analysis of Invasive Ductal Carcinoma (IDC) tissue regions. The method proposed in this work is a convolutional neural network (CNN) for visual semantic analysis of tumor regions for diagnostic support. Detection of IDC is a time-consuming and challenging task, mainly because a pathologist needs to examine large tissue regions to identify areas of malignancy. Deep Learning approaches are particularly suitable for dealing with this type of problem, especially when many samples are available for training, ensuring high quality of the learned features by the classifier and, consequently, its generalization capacity. A 3-hidden-layer CNN with data balancing reached both accuracy and F1-Score of 0.85 and outperforming other approaches from the literature. Thus, the proposed method in this article can serve as a support tool for the identification of invasive breast cancer.

Keywords:

Deep learning, convolutional neural network, invasive ductal carcinoma.

Introduction

Breast cancer (BCa) is more common among women worldwide, and, in Brazil, after nonmelanoma skin cancer, it accounts for 25% of new cases each year. In Brazil, this percentage is 29% [4]. BCa also affects men, but in this case, it is infrequent, representing only 1% of the total cases of the disease [8].

Over the world, around 2,088,849 new cases are estimated for women in the year 2020, and in 2040 they will be 3,059,829. In the year 2020, circa of 655,690 female deaths are expected related to BCa, and in 2040 the mortality will be around 991,904 [5].

In Brazil, roughly 59,700 new cases of BCa are estimated for each year of the 2018-2019 biennium, with a calculated risk of 56.33 cases per 100,000 women. This type of cancer is also the most frequent in women in all regions of Brazil [4].

Although BCa is very common, it is highly heterogeneous. It presents an extensive range of biological behaviors, from relatively inert or harmless disease to very aggressive, fast-growing malignancy with a short survival rate. Despite this variety of behavior, most tumors (75 to 80%) are categorized into a single class, invasive ductal carcinoma of no particular type (IDC-NST) [10].

During the last decade, significant advances in computational power and improvements in image analysis algorithms have enabled the development of robust computer-aided analytical approaches to radiological data. With the recent advent of digital whole slide scanners, tissue histopathology slides can now be digitized and stored in digital image format. Consequently, data from digitized tissue histopathology became suitable for the application of computerized image analysis and machine learning techniques [6].

This work aims to identify the IDC when it is present in histopathological images labeling them either as the absence of metastasis or as the presence of invasive ductal carcinoma. More specifically, this study presents a deep learning approach for automatic detection and visual analysis of tissue regions of IDC.

Detecting IDC is a time-consuming and challenging task, mainly because a pathologist needs to examine large regions to identify areas of malignancy.

Deep Learning approaches are particularly suitable for dealing with this type of problem, especially when many samples are available for training, ensuring high quality of the learned features by the classifier and, consequently, its generalization capacity.

The relevance of this work is the use of automated methods to reduce time and error of diagnosis, identifying BCa, and categorizing it more accurately.

Methods

The dataset consists of 277,524 histological breast images (50 x 50 x 3), available on Janowczyk's website[1] and used in [7].
The dataset consists of 50 x 50-pixel RGB digital image samples derived from mammographic histopathological examples of 162 digitized WSIs [2].
The file name of each patch is of the format "u-xX-yY-class*C*.png.'' For example, in the filename "10253-idx5-x1351-y1101-class0.png'', *u* is the patient ID (10253.idx5), *X* is the coordinate of where this patch *Y* is the coordinate from where this patch was taken, and *C* indicates the class where 1 and 0 are, respectively, the presence or absence of IDC.

Preprocessing

The dataset consists of 277,524 histological breast images: 198,738 related to the IDC(-) class; and 78,786 to the IDC(+) class.
The difference between the number of samples belonging to the classes can influence the performance of a classification model.

[1] http://www.andrewjanowczyk.com/use-case-6-invasive-ductal-carcinoma-idc-segmentation/

When this difference is significant, the learning systems may find it difficult to induce the concept related to the minority class [3].

One way of counterbalancing this disproportion between classes is by data balancing. Data balancing is a preprocessing technique that aims to change training data distribution to increase model accuracy. There are two main approaches [9] (i) *undersampling*: this strategy eliminates cases of the majority class; and (ii) *oversampling*: this strategy replicates cases of minority classes.

Both undersampling and oversampling have their issues. In the undersampling, one can eliminate potentially relevant examples; while using oversampling, one can increase the probability of overfitting since most oversampling approaches make exact copies of minority class samples [1].

In this work, the use of an oversampling approach was chosen to minimize the amount of potentially valuable data discarded.

Proposed architectures

Initially, in this work, a 3-hidden-layer CNN is proposed using 32, 64, and 128 neurons for the first and second convolutional layers and the fully connected layer, respectively. The first and second convolutional layers used convolutional kernels of size 3 x 3, and they are followed by a max-pooling layer of size 2 x 2 (see Figure 1).

The proposed 3-hidden-layer CNN architecture was implemented as a sequential model using the Keras library.

A dropout layer was used shortly after the max-pooling layer (during the training, it eliminates a random set of activation units by setting them to zero, thereby forcing the network to be "redundant'') to attenuate the overfitting**.**

At the output, a softmax function was used to normalize the results to estimate the probability of a given input image belonging to the IDC(-) class or the IDC(+) class.

The training was performed using the cross-entropy loss function with the Adadelta optimization method. Adadelta [11] is a more robust extension of Adagrad that adjusts learning rates based on a moving window of gradient updates, rather than accumulating all previous gradients. In this way, Adadelta continues to learn even when many updates have been made.

In addition, to this 3-hidden layer architecture (described above), two others were proposed and tested in this work. The first one is an architecture with 7-hidden layers based on the previous one in which two blocks were added, each one with two convolution layers (with a similar number of neurons as in the 3-layer network). Both blocks were followed by dropout layers, but an additional max-pooling layer was only used after the second block. Finally, the other proposed architecture is an 8-hidden layer CNN very similar to the one with 7-hidden layers, but with an additional 128-neuron fully connected layer just before the output (see Figure 2).

Figure 1 - Proposed architecture for 3-layer CNN.

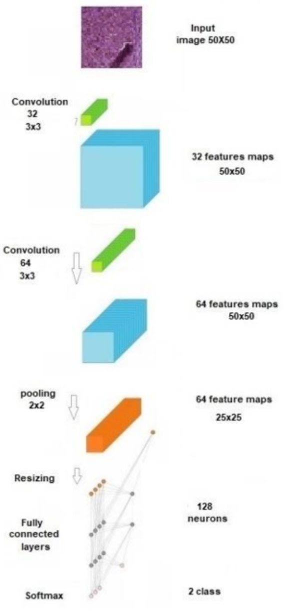
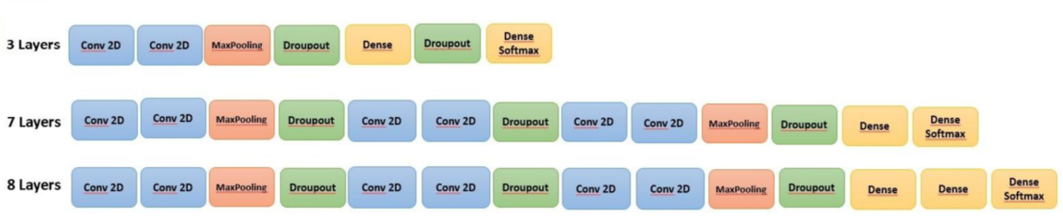


Figure 2 – Scheme of all proposed CNN architectures: 3-hidden layers, 7-hidden layers and 8-hidden layers



In this work, the following performance evaluation measures were used: accuracy[2], Precision[3], Recall[4], e F1-Score[5].
All experiments are performed using 10-fold cross-validation.

Experiments

Computational experiments consisted of creating several Convolutional Neural Networks (CNN) for the identification of Invasive Ductal Carcinoma (IDC).

The following CNNs were implemented with 3, 7, and 8 hidden layers without and with data balancing. The implemented architectures have been previously described and are illustrated in Figure 2.

At each step of the experiment, metrics are evaluated: accuracy, precision, recall, and F1-Score. The models were trained and validated during 08 epochs, and graphs of accuracy over time are also presented.

The dataset is unbalanced and contains more images belonging to the IDC(-) class. Therefore, data balancing through an oversampling approach was adopted. The Random Over Sampling function of Python was used, making equal the number of instances of the majority and minority classes.

Results

Results for 3-hidden layer CNN without and with data balancing

Results for 3-hidden layer CNN with and without data balancing for identification of IDC presented F1-Score of 0.85 for both IDC(-) and IDC(+) classes, as shown in Table 1.

Table 1 - Results for 3-hidden layer CNN without and with data balancing.

	Without data balancing			With data balancing		
Class	P	R	F1	P	R	F1
IDC(-)	0.84	0.93	0.88	0.87	0.82	0.85
IDC(+)	0.75	0.56	0.64	0.83	0.87	0.85
Avg	0.80	0.75	0.76	0.85	0.85	0.85

*P= Precision, R=Recall and F1= F1-Score

According to the results, for the unbalanced database, the recall rate was 93% and 56% for IDC(-) and IDC(+) classes, respectively. This corresponds to a high false-negative rate for IDC(+) of 44%. Precision rates were 84% and 75% for IDC(-) and IDC (+), respectively.

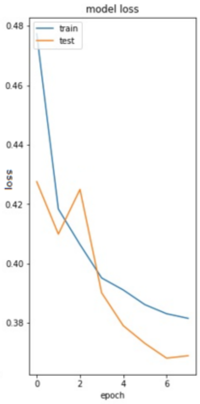
In all evaluated metrics, 3-hidden layer CNN without data balancing presented worse results when compared with the results obtained when data balancing was used. However, for the balanced database, the proposed model successfully classified 82% of the images belonging to the IDC(-) class and 87% of the images classified as IDC(+), corresponding to a false negative rate of 18% and 13% for the IDC(-) and IDC(+) classes, respectively.

The training loss function after 8 epochs was 0.38 and the validation loss was 0.37 (see Figure 3). This indicates that the training procedure was able to minimize the loss function.

Analyzing the confusion matrix (Figure not shown), 960 cases were classified as IDC(-) but were IDC(+). This expressive amount generates an illusory and dangerous result for the patient mistakenly classified as having no invasive neoplasia.

One thousand three hundred twenty-nine (1,329) cases were classified as IDC(+), but they belonged to the IDC(-) class. This situation probably creates a nuisance for the patient classified (at least temporarily) incorrectly as ill.

Figure 3– Loss for 3-hidden layer CNN with data balancing



Results for 7-hidden layer CNN without and with data balancing

According to the results presented in Table 2, the precision rate obtained by the 7-hidden layer CNN without data balancing was 81% for the IDC(-) class and 82% for the IDC(+) class. However, when the images are IDC(-), the 7-hidden layer CNN without data balancing correctly identifies 96%, whereas, for IDC(+), only 43% are correctly classified. On the other hand, the results obtained after the balancing became much worse, especially for the IDC (+) class in which the sensitivity was very low, with a recall of 0.03, for example.

Table 2 - Results for 7-hidden layer CNN without and with data balancing.

	Without data balancing			With data balancing		
Class	**P**	**R**	**F1**	**P**	**R**	**F1**
IDC(-)	0.81	0.96	0.88	0.72	1.00	0.84
IDC(+)	0.82	0.43	0.56	0.79	0.03	0.05
Avg	0.82	0.70	0.72	0.74	0.72	0.62

* P= Precision, R=Recall and F1= F1-Score

Results for 8-hidden layer CNN without and with data balancing

According to the results presented in Table 3, the precision rate obtained by the 8-hidden layer CNN without data balancing was 90% for the IDC(-) class and 77% for the IDC(+) class. However, when the images are IDC(-), the 8-hidden layer CNN without data balancing correctly identifies 91%, whereas, for IDC(+), only 74% are correctly classified. The balancing in this

[2] Accuracy = $\frac{\text{TP+TN}}{\text{TP+ TN+FP+FN}}$

[3] Precision = $\frac{\text{TP}}{\text{TP+FP}}$

[4] Recall = $\frac{\text{TP}}{\text{TP+ FN}}$

[5] F1-Score =2× $\frac{\text{Precison ×Recall}}{\text{Precision+Recall}}$

case, unlike the 7-layer configuration, increased the evaluated metrics by 3 to 4 percentage points.

Table 3 - Results for 7-hidden layer CNN without and with data balancing.

	Without data balancing			**With data balancing**		
Class	**P**	**R**	**F1**	**P**	**R**	**F1**
IDC(-)	0.90	0.91	0.91	0.91	0.91	0.91
IDC(+)	0.77	0.74	0.75	0.76	0.77	0.76
Avg	0.84	0.83	0.83	0.87	0.87	0.87

*P= Precision, R=Recall and F1= F1-Score

Discussion

From the results presented, it is possible to note that the imbalance of the classes affected the quality of the model since the metric (recall and F1-Score) for the minority class IDC(+) were much worse than the balanced model except for the 7-hidden layer CNN. For example, the best accuracy result was found by 8-hidden layer CNN with an accuracy rate of 0.86. However, the average precision rate (0.84) and average F1-Score (0.83) were lower than those found in 3-layer CNN with data balance in which both metrics were 0.85. These results were already expected because the data balancing contributed to improving the classifier's hit rate.

We compared our best-performing 3-hidden layer CNN (with data balancing and considering that the best classifier is the one that best classifies IDC(+) class) to the work of Janowczyk and Madabhushi [7] and evaluated metrics show that our proposed CNN model has performed best.

In Janowczyk and Madabhushi [7], the authors used different methods to tailor their version of AlexNet (with 32 x 32 input) to handle 50 x 50 size images, such as resizing, dropout scaling, cropping, and cropping with rotations (see Table 4 for results obtained in [7]).

Table 4 - Results from [7] for different versions of Alexnet.

Method	**F1-Score**	**Balanced Accuracy**
Resize	0.76	0.85
Resize + Dropout	0.75	0.84
Cropping	0.75	0.84
Cropping + Roration	0.75	0.84

The best results were obtained by resizing in 50k samples, reaching an F1-Score of 0.76 and a balanced accuracy of 0.85 [7].

Table 5 presents a comparison of the results obtained by our best-performing 3-layer CNN (using data balancing and considering a more significant hit in IDC(+) class) with the best approach presented in [7] - the version of AlexNet with resizing. One should notice no difference between balanced accuracy (used in [7] and "regular"/"overall" accuracy used in the experiments with our 3-layer CNN are considering the version in which the dataset was balanced.

Table 5 - Comparision of our best CNN result with other approach

Approach	**F1**	**P**	**Ac**
AlexNet with resize [7]	0.76	-	0.85
Our 3-layer CNN with bal	0.85	0.85	0.85

* F1= F1-Score, P=Precision and Ac = Accuracy

So, regarding the accuracy, our proposal showed 85%, while in [7] the best approach presented the same value and the other variations had an accuracy of 84%.

But, in terms of F1-Score, our proposal presented an improvement of 11.8% concerning the best approach of Janowczyk and Madabhushi [7], indicating a better compromise between precision and recall of our proposal.

Conclusions

Precise detection of invasive cancer in histological images is critical in digital pathology diagnosis and classification tasks. Convolutional Neural Network (CNN) represents the most popular learning method for computational vision tasks, which has recently been successfully applied in digital pathology, including tumor detection.

In this work, a 3-hidden layer CNN was used to identify Invasive Ductal Carcinoma (IDC) in histological images of breast cancer.

The obtained results indicate the efficiency of the proposed approach, which a precision, F1-Score and accuracy of 85% on average. Other architectural variants (with 7 and 8 layers) were implemented and tested, but the architecture with 3-hidden layers using data balancing achieved the best average results.

Thus, the approach presented in this paper has the potential to serve as a decision support tool to help pathologists accelerate the identification and localization of breast cancer, significantly relieving their workload.

Despite the promising results, the model needs to be refined to increase accuracy and precision, thus reducing its number of false positives and false negatives.

In future works, we suggest further refinement of the prototype model and investigate other machine learning methods to cope with this critical task.

Acknowledgements

This study was financed in part by the Coordination for the Improvement of Higher Education Personnel - Brasil (CAPES) - Finance Code 001. The authors thank the National Council for Scientific and Technological Development of Brazil (CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico) and the Foundation for Research Support of the Minas Gerais State (FAPEMIG). The work was developed at the Pontifical Catholic University of Minas Gerais, PUC Minas in the Applied Computational Intelligence laboratory – LICAP.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer.SMOTE: Synthetic Minority Over-sampling Technique. Jounal of Ar-tificial Intelligence Research, v. 16, pp. 321–357, 2002.

[2] A. Cruz-Roa and A. Basavanhally and F. A. Gonzalez and H. Gilmoreand M. Feldman and S. Ganesan and N. Shih and J. E. Tomaszeweskiand A. Madabhushi. Automatic detection of invasive ductal carcinomain whole slide images with convolutional neural networks. MedicalImaging: Digital Pathology, 2014.

[3] Fawcett and F. J. Provost. Adaptive Fraud Detection. Data Mining andKnowledge Discovery, v. 1, n.3, pp. 291-316, 1997.

[4] Ministério da Saúde, Instituto Nacional de Câncer, Estimativa 2018:Incidência de Câncer no Brasil, INCA, 2018.

[5] International Agency for Research on Cancer. World Hearlth Orga-nization, Globocan 2018: Estimated Cancer Incidence, Mortality andPrevalence Worldwide in 2018-2012, 2018.

[6] M.N. Gurcan MN and L.E. Boucheron and A. Can A and A. Madabhushiand N.M. Rajpoot and B. Yener , Histopathological image analysis: Areview ,IEEE Rev Biomed Eng, vol. 2, pp. 147-171, 2009.

[7] Janowczyk, A and Madabhushi, A. Deep learning for digital pathologyimage analysis: A comprehensive tutorial with selected use cases,JPathol Inform, vol. 7, n.29, pp. 1-17, 2013.

[8] Jbilou J, Halilem N, Blouin-Bougie J, Amara N, Landry R, Simard J.,Medical genetic counseling for breast cancer in primary care: a synthesisof major determinants of physicians' practices in primary care settings,Public health genomics, vol. 17, n. 4, pp. 190-208, 2013.

[9] Oded Maimon and Lior Rokach,Data Mining and Knowledge DiscoveryHandbook, Springer, 2nd, 2010.

[10] T. Takuji and K. Masashi and M. Kengo and D. Fukada and H. Mori and Y. Okano, Centrosomal Kinase AIK1 Is Overexpressed in Invasive Ductal Carcinoma of the Breast, Cancer Research, vol. 59, n.1, pp. 2041- 2044, 1999.

[11] M. D. Zeiler, ADADELTA: an adaptive learning rate method, arXiv preprint arXiv:1212.5701, 2012.

Address for correspondence

Érika G de Assis

Érika G. de Assis - Pontifical Catholic University of Minas Gerais -
Belo Horizonte, MG, Brazil

E-mail: dudabh@gmail.com