Histopathological image analysis for breast cancer diagnosis by ensembles of convolutional neural networks and genetic algorithms

Miguel A. Molina-Cabello, José A. Rodríguez-Rodríguez, Karl Thurnhofer-Hemsi, Ezequiel López-Rubio
Department of Computer Languages and Computer Science. University of Málaga
Bulevar Louis Pasteur, 35, 29071, Málaga, Spain
Instituto de Investigación Biomédica de Málaga (IBIMA)
C/ Doctor Miguel Díaz Recio, 28, 29010, Málaga, Spain
Emails: {miguelangel,karlkhader,ezeqlr}@lcc.uma.es; joseantoniorodriguez@uma.es

***Abstract*—One of the most invasive cancer types which affect women is breast cancer. Unfortunately, it exhibits a high mortality rate. Automated histopathological image analysis can help to diagnose the disease. Therefore, computer aided diagnosis by intelligent image analysis can help in the diagnosis tasks associated with this disease. Here we propose an automated system for histopathological image analysis that is based on deep learning neural networks with convolutional layers. Rather than a single network, an ensemble of them is built so as to attain higher recognition rates, which are obtained by computing a consensus decision from the individual networks of the ensemble. A final step involves the optimization of the set of networks that are included in the ensemble by a genetic algorithm. Experimental results are provided with a set of benchmark images, with favorable outcomes.**

***Index Terms*—convolutional neural networks, image classification, breast cancer, medical image processing, genetic algorithms**

I. Introduction

The impact of automated digital image processing is pervasive in most medical fields. Different kinds of images are acquired by various procedures such as X-ray imaging, ultrasound imaging, and resonance imaging. The goal of automated medical image processing often involves their enhancement so as to facilitate the inspection of the images by medical staff. Moreover, computer aided diagnosis techniques can also be employed to help practitioners to diagnose a disease [1]. Therefore, an automated intelligent system is able to determine whether a patient is likely to suffer from the disease by analyzing the acquired medical image data. The role of medical image analysis is widely recognized in pathology detection [2].

Breast cancer stands as the most invasive cancer type for women. This leads to a high mortality rate. Nowadays, one of the approaches used to its diagnosis is based on histopathological analysis. Therefore, computer aided diagnosis by the analysis of histopathological images can be integrated into the health workflow in order to support the work of health professionals in their efforts to diagnose breast cancer reliably in a reduced amount of time [3]. The BreakHis dataset [4] is recognized as a very significant breakthrough in these efforts, since it comprises over 7,900 samples of histopathological images. This large sized dataset overcomes the limitations of previous ones, which hampered the performance of earlier intelligent recognition systems for histopathology images [5]. Visual feature descriptors have been proposed to detect patterns in the analyzed image [4]. State of the art deep learning neural networks can be employed for object detection and image classification. This means that this kind of network, and in particular convolutional neural networks (CNNs), are amenable to their application for histopathological image analysis, for example, [6]. A particular kind of CNN, named AlexNet [7], was used in that work in order to carry out such analysis. This kind of neural architecture has recently been proposed to solve different computer vision tasks such as the classification of vehicle types in traffic video footage [8] and the classification of blood cells [9].

Our proposal draws on the reference neural architecture proposed in [6]. Our enhancements to their proposal include the construction of an ensemble of neural networks in order to outperform the diagnosis performance of individual networks. Therefore, the outputs of many networks are combined by a consensus procedure in order to obtain an ensemble output that is more accurate [8]. The set of networks that comprise the ensemble is selected by optimization of the diagnosis performance. This task is carried out by a genetic algorithm. The objective of the presented work is to enhance this system, which is described in a previous work [10].

The structure of this paper is detailed next. First, the related work is reported in Section II. Then, the proposed methodology is presented in Section III, including the specification of the considered ensemble kinds and the genetic algorithm to select the most accurate option for the set of networks that

This work is partially supported by the following Spanish grants: TIN2016-75097-P, RTI2018-094645-B-I00 and UMA18-FEDERJA-084. All of them include funds from the European Regional Development Fund (ERDF). The authors acknowledge the funding from the Universidad de Málaga. Karl Thurnhofer-Hemsi is funded by a Ph.D. scholarship from the Spanish Ministry of Education, Culture and Sport under the FPU program (FPU15/06512).

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

form the ensemble. The experimental results are reported in Section IV, where the optimization of the parameters and the attained performance of the ensembles are shown. Conclusions are extracted in Section V.

II. Related work

The motivations behind the current interest in the application of deep learning techniques to breast cancer diagnosis are multiple. First of all, classic Computer Aided Diagnosis (CAD) systems exhibit significant limitations [11]. Secondly, early detection of this kind of cancer brings a much better prognosis. And thirdly, false detections must be reduced at all costs due to the adverse impact of them on the patients.

Deep learning, and especially CNNs, are particularly suitable to process two and three dimensional medical images. Therefore, CNNs are employed to analyze different types of images that are relevant to breast cancer diagnosis, such as mammographies [12], magnetic resonance images [13] and histopathology images [14].

The goal to be attained varies depending on the kind of image. For mammographies and magnetic resonance images, the challenge is to detect potentially malignant lesions in the breast, with the difference that magnetic resonance images are usually three dimensional. Radiomics approaches aim to extract significant features from radiographic images [15]. On the other hand, for histopathological images, the system must distinguish between pathological and healthy tissues. This involves, for example, the evaluation of the human epidermal growth factor receptor 2 (HER2) protein status [16] and the detection of mitoses [17].

In the past few years, many efforts have been made to apply CNNs on the histopathological imaging modality. Ciresan *et al* were the first to apply them to the task of mitosis counting for primary breast cancer grading [18]. Close in time, Cruz-Roa *et al* trained a convolutional network to recognize primary breast cancer [19]. Then, Spanhol *et al* included a further and deeper research in [20] where an evaluation of the combination of six different feature sets and four base classifiers is conducted, and the final system is defined by the combination that produces the best results in the validation set. Meanwhile, Bayramoglu *et al* [21] proposed two different CNN architectures to classify breast cancer histopathological images independently of their magnifications: the single task CNN used to predict malignancy, and the multi-task CNN used to predict both malignancy and image magnification level simultaneously. In addition, Nawaz *et al* [22] presented a *DenseNet* based model for multiclass breast cancer classification to predict the subclass of the tumors, while Motlagh *et al* [23] used the pre-trained model of *ResNet*_$V1$_152 to perform diagnosis of benign and malignant tumors as well as diagnosis based on multiclass classification of various subtypes of histopathological images of breast cancer in BreakHis.

The base of the presented approach is a previous work [10], where the breast cancer diagnosis is attained by classifying histopathological images. First of all, that work proposes an optimization of the architecture of the neural network presented in [6]. This is accomplished by a fine-tuning process of the parameter configuration of the trained model openly provided by the authors on [24], noted as *Reference*. This fine-tuning process is divided into two phases. Each phase tests the performance of a network by training a network considering different parameters. On the one hand, the weight decay and the base learning rate are tuned in the first phase. On the other hand, the solver type and the number of fully-connected layers are tuned in the second phase. The performance of the reference model is improved after the first phase, while the second phase performs the parameter tuning on the model achieving the highest mean accuracy. Figure 1 reports a schema of this operation. As it can be observed, the reference CNN model [24] is provided to the fine-tuning process as input, while the final candidate model [10] is the output of this process. The final candidate model is the base CNN model of this work, noted as *Base*. The parameter values of this neural network model are tuned as follows: base learning rate is set to 10^{-3}, the weight decay is $4 \cdot 10^{-3}$, the solver type is Adam and the number of fully-connected layers is 3.

After that, the performance of the system is increased by using a well-known technique that has been used in machine learning: ensemble learning. This technique has been widely employed to improve the performance over a single estimator. In order to achieve this improvement, the predictions of several classifiers are combined by using a consensus function or algorithm [8]. A schema of this method is described in Figure 2. As can be observed, the input image is provided to each one of the possible considered CNN ($MODEL_i$), which builds its own prediction ($prediction_i$). These predictions are supplied to a consensus function where the prediction of the system is calculated. This output is computed in a different way depending on the selected kind of consensus function. Additionally, a genetic algorithm enhances the performance of the consensus function by choosing the best CNNs of the consensus that performs the output.

As it can be deduced from Figure 2, the performance of the ensemble is directly related to the kind of aggregation function which is used to compute the consensus prediction. However, the previously presented work also considers the use of a genetic algorithm to enhance the ensemble. This proposed genetic algorithm is applied in order to reduce the number of networks that comprise the ensemble and to improve its performance. The evolution of a population of a certain number of individuals over a number of generations is carried out in order to achieve that enhancement. Each one of the considered individuals represents an array of boolean elements. The number of these Boolean elements is the number of neural networks that can belong to the ensemble, where the neural network represented by the element is left out of the ensemble when its boolean value is 0, while a value of 1 means that the neural network is within the ensemble. The fitness function is the mean accuracy, which measures the performance of the ensemble that the individual represents. The objective of the evolution in genetic algorithms is to maximize the fitness value. In the case that an individual yields

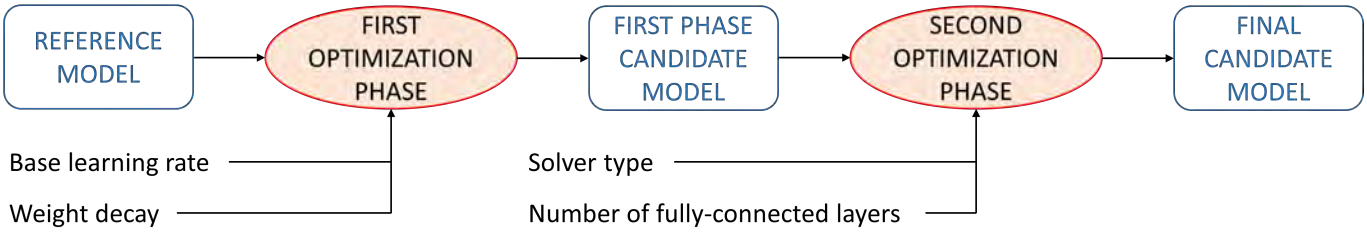


Fig. 1: Schema of the fine-tuning process operation. A reference CNN model is enhanced by a fine-tuning process that is divided into two phases. The first phase optimizes tunes the base learning rate and the weight decay, while the second phase tunes the solver type and the number of fully-connected layers.

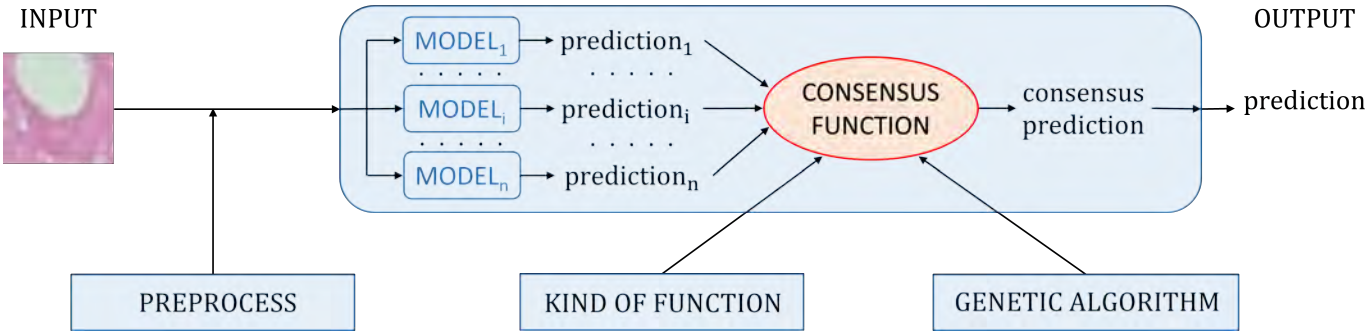


Fig. 2: Schema of the ensemble operation. An image is provided to the system as input. After that, several models receive this input and offer their predictions. Then, all the predictions generated by each model which compose the ensemble are given to a consensus function, which computes the output of the system.

the maximum possible fitness value (1 is its maximum value), the evolution stops.

A fixed number of individuals (noted as *Population size*) are generated randomly in order to compose the first generation. After this initialization, certain actions are carried out in each generation. First of all, each pair of individuals within the population may be crossed over with a certain probability. Furthermore, each individual may mutate by flipping some of its elements with a certain probability. Then, each evolved individual is evaluated by the fitness function. Finally, the resulting individuals compose the next generation that starts if the maximum possible fitness value is not achieved, and the maximum number of generations has not been surpassed. This system is noted as *ConsensusGA*.

III. Methodology

Next the proposed classification methodology for breast cancer histopathological images is detailed. It is based on an ensemble of Convolutional Neural Network (CNN) classifiers. Let M be the number of classes and N be the number of CNNs. That is, N CNNs must be combined. Also, the output vector of the i-th CNN is noted $\mathbf{y}_i$, where $i \in \{1, ..., N\}$. In other words, y_{ij} stands for the predicted score associated to the j-th class by the i-th CNN, for $j \in \{1, ..., M\}$. The ensemble of classifiers can be built by selecting a subset $\mathcal{S} \subseteq \{1, ...N\}$ of the overall set of CNNs. In this work we consider several kinds of neural ensembles:

- Maximum (Max) ensemble. The maxima of the scores associated to a certain class is provided as the output score for that class:

$$\mathbf{z}_{Max} = \max \{\mathbf{y}_i \mid i \in \mathcal{S}\} \tag{1}$$

- Mean ensemble. The arithmetic mean of the scores associated to a class is employed as the final score of the class:

$$\mathbf{z}_{Mean} = \text{mean} \{\mathbf{y}_i \mid i \in \mathcal{S}\} \tag{2}$$

- Median ensemble. The median of the scores associated to a class is employed as the final score of the class:

$$\mathbf{z}_{Median} = \text{median} \{\mathbf{y}_i \mid i \in \mathcal{S}\} \tag{3}$$

- Voting ensemble. The count of times that a class ranks highest among the scores obtained by a CNN is provided as the final score for that class:

$$\mathbf{z}_{Voting} = \left(\left| \left\{ i \in \mathcal{S} \mid j = \arg \max_{k \in \{1,...,M\}} \{y_{ik}\} \right\} \right| \right) \tag{4}$$

where $|\cdot|$ notes the number of elements of a set and $j \in \{1, ..., M\}$.

Once the final scores produced by the ensemble are calculated, the predicted class is that associated to the highest of such final scores. There are many options to choose the optimal subset of CNNs $\mathcal{S}$ which are used to build the ensemble. We propose to solve this optimization problem by a a genetic algorithm. Every individual of the genetic algorithm has a

chromosome consisting of N Boolean variables. Each variable says whether a particular CNN is used in the ensemble. The performance of the ensemble is employed as the fitness function of the generic algorithm. This performance is computed over a validation set which is disjoint with the training set employed to train the CNNs. Also, the architecture of the employed CNNs is subject to optimization, as detailed in Section IV. It must be noted that, in order to improve the results of the recognition system, a preprocessing step has been added to enhance the input image. The preprocessing consists in apply a sharpening filter, which is equivalent to subtracting a Gaussian low pass filtered version of the image from the original image. Therefore, the spatial high frequency components of the input image are given more importance, i.e., the edges and the small details of the image. This way, it is easier for the CNNs to detect the relevant features.

Regarding the computational complexity of the proposed approach, it is linear with respect to the number of CNNs which belong to the ensemble N, so that, $O(N)$, except for the consensus function which is based on the median, which have complexity $O(N \log(N))$. However, an insignificant amount of the overall runtime is taken by the computation of the median, so in practice the complexity is linear with respect to N. Additionally, the convergence of this methodology is achieved due to the genetic algorithm stops at most when the specified maximum number of generations is achieved.

IV. Experiments results

The experiments that we have carried out are described in this section. It is structured as follows. First of all, the hardware and software that have been used in this work are depicted in Subsection IV-A. Secondly, Subsection IV-B presents the image dataset employed in the experiments, while the results are shown in IV-C.

A. Methods

The *Reference* model [24], the *Base* neural network model and the *ConsensusGA* system [10], and the proposed method were built by using the framework Caffe (Convolutional Architecture for Fast Feature Embedding, [25]) which is an open source deep learning framework. Caffe provides a repository of well-knwon pre-trained models such as AlexNet [7].

The proposal has been written in Python, while the reported experiments have been carried out on a 64-bit Personal Computer with two Intel E5-2670 CPU with eight cores, 2.60 GHz per core, 32 GB RAM, Nvidia GeForce GTX 1080 Ti as GPU and standard hardware.

B. Dataset

A well-known dataset has been chosen to test the performance of the proposed approach. The selected dataset is the BreakHist image dataset [20][1].

This dataset is composed of 7909 breast histopathological images, which are divided into benign and malignant tumors, acquired on 82 patients. Both benign and malignant breast tumors can also be organized into distinct kinds depending on the aspect of the tumoral cells under the microscope. The original BreakHist dataset contained four different histological types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant tumors (breast cancer): ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The images of the dataset have been acquired in 3-channel RGB (Red-Green-Blue) TrueColor (24-bit color depth, 8 bits per color channel) color space using magnifying factors of $40\times$, $100\times$, $200\times$ and $400\times$.

In order to carry out the designed experiments of this work, a subset composed of images acquired at 40X magnification has been considered. In this subset, the samples are divided into two possible classes: benign and malignant tumors. Additionally, 1,000 random 64x64 patches are generated for each image by employing the strategy #4 defined in [6] in order to establish a fair comparison with the *Reference* model. Some examples of the original dataset and the considered experimental images are shown in Figure 3. After that, the resulting dataset is randomly organized into training and test set by accounting for 65% and 35% of the images, respectively.

Additionally, different training subsets have been employed by using a k-fold strategy in order to obtained different trained *Base* method networks. This way, the ensemble of the *ConsensusGA* and the proposed methods are formed by distinct networks. Respecting the test set that the methods have been carried out to establish a comparison between them, that test set is formed by 745 images at 40X magnification, where 255 are benign tumor images, and 490 are malignant tumor images. Therefore, the test set is composed of 74,500 patches of images.

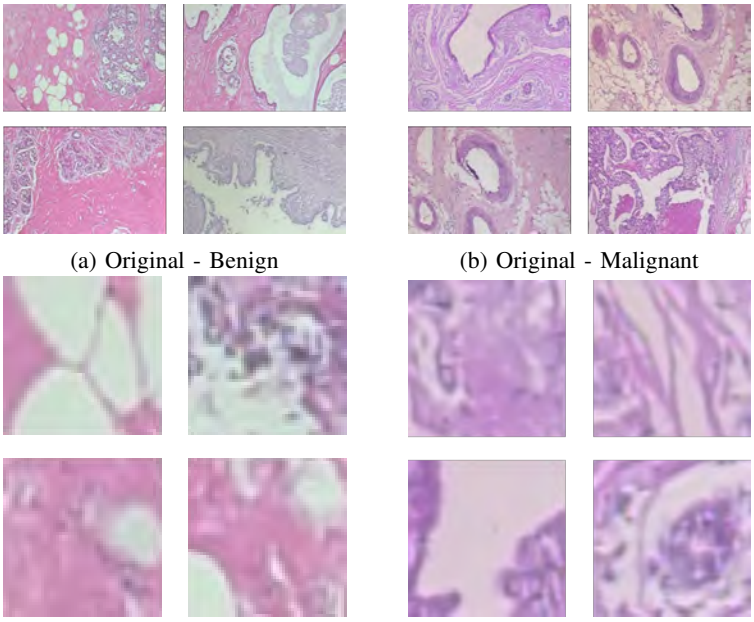
C. Results

From a quantitative point of view, we have selected some different well-known measures in order to compare the performance of the detection and the classification for breast cancer diagnosis. In this work, the spatial accuracy (S), the Accuracy (Acc), and the F-measure (Fm) have been considered. All these measures provide values in the interval $[0, 1]$, where higher is better, and represent the percentage of hits of the system. True positives or number of hits (TP), true negatives or correct rejections (TN), false negatives or misses (FN), false positives or false alarms (FP), the precision (PR), the recall (RC), the specificity (SP), the false positive rate (FPR) and the false negative rate (FNR) are also used in this work. Among all these measures, the spatial accuracy, the accuracy and the F-measure provide a good overall evaluation of the performance of a given method, while FN must be considered against FP (lower is better), PR against RC (higher is better) and FPR against FNR (lower is better). Their definitions are as follows:

$$\text{Fm} = 2\frac{PR * RC}{PR + RC} \qquad S = \frac{TP}{TP + FN + FP} \tag{5}$$

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \qquad SP = \frac{TN}{FP + TN} \tag{6}$$

[1] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/



(a) Original - Benign (b) Original - Malignant

(c) Patches - Benign (d) Patches - Malignant

Fig. 3: First row exhibits several samples from the BreakHis dataset, while second row shows some 64x64 pixel sample patch images (which have been generated from random BreakHis images) used in the experiments. First column shows benign tumors, while second column shows malignant tumors. Patch images from (c) are generated from top left image of (a), while patch images from (d) are generated from top left image of (b).

$$RC = \frac{TP}{TP + FN} \qquad PR = \frac{TP}{TP + FP} \tag{7}$$

$$FNR = \frac{FN}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \tag{8}$$

In this work, an analysis of the behavior of the system respecting its possible parameter values has been studied. The selected parameters in this study are the type of the neural ensemble (T), the number of networks considered to belong to the ensemble (N), the population size (P), and the maximum number of generations of the genetic algorithm (G). The use of the preprocess step is also analyzed. The impact of all these parameters is measured with the number of networks that belong to the optimal subset of CNNs that are used to build the ensemble (noted as N_{within}) and the performance achieved by that optimal subset. The rest of the parameters are fixed: the fitness function is the F-measure, the objective value is 1.0, the probability of the crossover of two individuals is set to 0.5, while the probability for mutating an individual is fixed to 0.2. The selected configurations to be tuned in this analysis are reported in Table I. Note that in our previous work [10], the parameters of the proposed method *ConsensusGA* were fixed ($N = 30, P = 300, G = 1000, preprocess = False$) except the ensemble type. Additonally, in that work, the measurement used as a criterion of the fitness function in the genetic algorithm was the accuracy. However, that measure is invalidated since it will possibly be biased towards the positive class if there is no other bias control mechanism due to the dataset is sensibly unbalanced. This way, the F-measure has been selected as a measure criterion because it can give a better classifier's performance understanding. The experiments have been repeated 10 times in order to obtain a realistic average of the performance of the system.

Table II reports the performance of the best tuned configuration. As it is shown, the use of an ensemble of components improves the performance of only one component (*Base*). This consideration is usually expected. The more interesting is to observe how the use of the genetic algorithm (GA) is suitable to enhance the performance of an ensemble. In this study, the four considered ensemble types improve their performance when the genetic algorithm is applied, even better than the *ConsensusGA* method. Moreover, the use of the preprocess step also enhances the performance of the system. Going deeper, Table III exhibits the performance of the best configuration of each ensemble type according to its yielded F-measure. As can be observed, Mean ensemble obtains the best results in the measures which provide an overall performance of the system (*Fm*, *Acc* and *S*). It is interesting to see how Median ensemble has a low rate of false negatives, which improves its *FNR* and *RC*. It also has the lowest N_{within}.

The performances obtained by each ensemble type for each tuned configuration can be studied into more detail. Figure 4 exhibits the F-measure yielded by each ensemble in terms of the number of neural networks which can belong to the ensemble, the number of generations and the population size. As it can be observed, the selected kind of ensembles yield a similar performance in terms of Fm. However, the use of the preprocess step has a positive impact in the performance. Mean ensemble seems to be the best of the considered types of

TABLE I: Considered parameters and their possible values to study the performance of the system.

Parameter		Value
Type of ensemble, T	=	{Voting, Max, Mean, Median}
Population size, P	=	$\{10, 30\}$
Individual length (Number of neural networks), N	=	$\{1, 2, 3, 5, 8, 10, 13, 15, 20, 25, 30\}$
Maximum number of generations, G	=	$\{1, 2, 10, 50, 100, 200, 300, 500\}$
Preprocess, $preprocess$	=	{True, False}

TABLE II: Performances of best tuned configurations according to their F-measure yielded for each ensemble type. Best results are highlighted in **bold**. Best results by ensemble type are highlighted in *italic*.

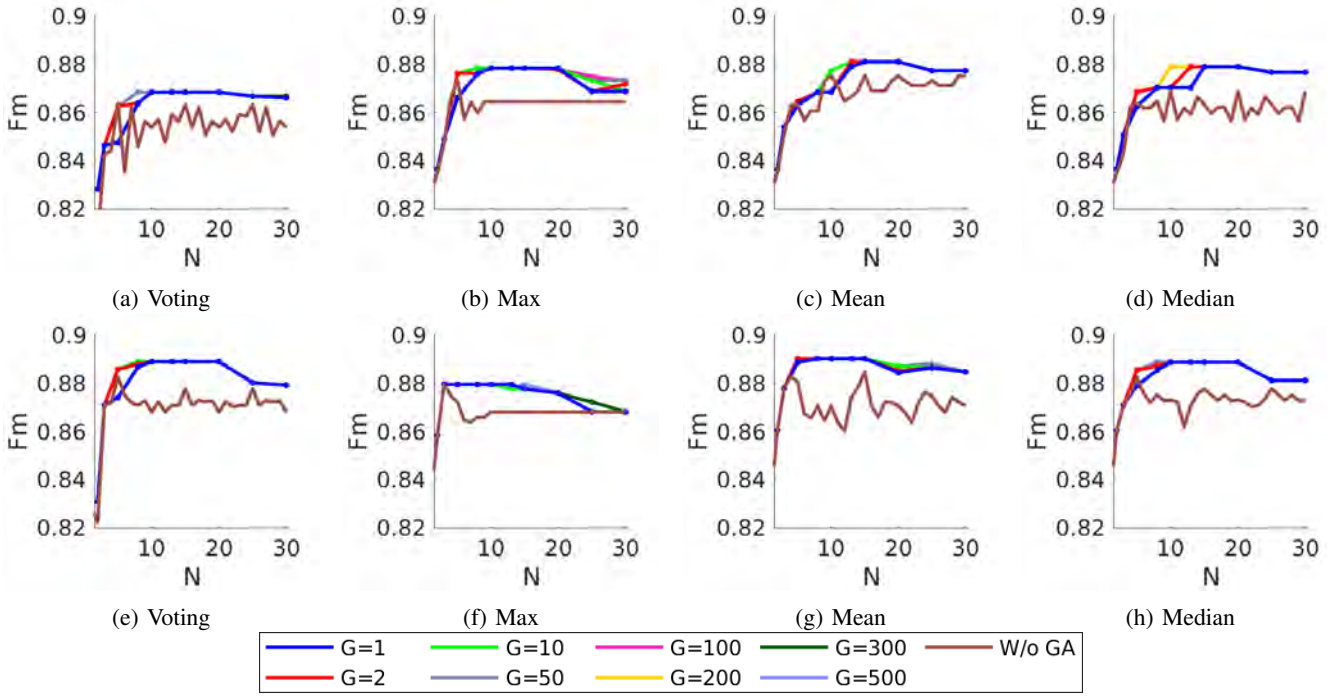
Method	Fm	Acc	PR	RC	N_{within}
Base [10]	0.826 ± 0.005	$\mathit{0.788 \pm 0.015}$	0.863 ± 0.016	0.793 ± 0.013	$\mathbf{1.000 \pm 0.000}$
Base + preprocess	$\mathit{0.830 \pm 0.046}$	0.785 ± 0.052	$\mathit{0.865 \pm 0.030}$	$\mathit{0.800 \pm 0.067}$	$\mathbf{1.000 \pm 0.000}$
Max w/o GA	0.873 ± 0.026	0.840 ± 0.036	0.880 ± 0.026	0.867 ± 0.028	5.300 ± 1.300
Max w/o GA + preprocess	$\mathit{0.880 \pm 0.021}$	0.835 ± 0.029	0.853 ± 0.029	$\mathit{0.909 \pm 0.037}$	$\mathit{3.000 \pm 0.000}$
Max with GA	0.878 ± 0.023	$\mathit{0.845 \pm 0.031}$	0.878 ± 0.025	0.878 ± 0.021	5.300 ± 1.300
Max with GA + preprocess	$\mathit{0.880 \pm 0.021}$	0.835 ± 0.029	0.853 ± 0.029	$\mathit{0.909 \pm 0.037}$	$\mathit{3.000 \pm 0.000}$
Max ConsensusGA [10]	0.878 ± 0.023	$\mathit{0.845 \pm 0.031}$	$\mathit{0.881 \pm 0.026}$	0.874 ± 0.020	8.300 ± 5.300
Mean w/o GA	0.875 ± 0.026	0.843 ± 0.033	0.881 ± 0.025	0.870 ± 0.035	7.800 ± 2.200
Mean w/o GA + preprocess	0.885 ± 0.017	0.840 ± 0.022	0.845 ± 0.018	$\mathit{0.928 \pm 0.026}$	6.800 ± 5.600
Mean with GA	0.881 ± 0.028	0.850 ± 0.036	$\mathbf{0.889 \pm 0.031}$	0.874 ± 0.035	8.300 ± 2.200
Mean with GA + preprocess	$\mathbf{0.890 \pm 0.022}$	$\mathbf{0.850 \pm 0.027}$	0.862 ± 0.015	0.921 ± 0.040	$\mathit{3.800 \pm 1.000}$
Mean ConsensusGA [10]	0.877 ± 0.025	0.845 ± 0.033	0.885 ± 0.025	0.870 ± 0.035	17.300 ± 3.900
Median w/o GA	0.869 ± 0.035	0.835 ± 0.042	0.876 ± 0.029	0.863 ± 0.050	6.000 ± 2.700
Median w/o GA + preprocess	0.883 ± 0.037	0.835 ± 0.053	0.835 ± 0.039	$\mathbf{0.936 \pm 0.038}$	$\mathit{5.800 \pm 1.500}$
Median with GA	0.879 ± 0.031	$\mathit{0.848 \pm 0.035}$	0.884 ± 0.030	0.874 ± 0.043	6.800 ± 2.800
Median with GA + preprocess	$\mathit{0.889 \pm 0.026}$	0.845 ± 0.034	0.847 ± 0.022	$\mathbf{0.936 \pm 0.038}$	$\mathit{5.800 \pm 1.500}$
Median ConsensusGA [10]	0.877 ± 0.030	0.845 ± 0.034	$\mathit{0.884 \pm 0.021}$	0.870 ± 0.047	15.500 ± 3.100
Voting w/o GA	0.864 ± 0.021	0.830 ± 0.022	0.878 ± 0.033	0.851 ± 0.012	$\mathit{5.500 \pm 1.000}$
Voting w/o GA + preprocess	0.883 ± 0.037	0.835 ± 0.053	0.835 ± 0.039	$\mathbf{0.936 \pm 0.038}$	6.000 ± 1.400
Voting with GA	0.868 ± 0.023	0.835 ± 0.024	$\mathit{0.879 \pm 0.034}$	0.858 ± 0.014	$\mathit{5.500 \pm 1.000}$
Voting with GA + preprocess	$\mathit{0.889 \pm 0.027}$	$\mathit{0.847 \pm 0.034}$	0.856 ± 0.014	0.924 ± 0.043	6.000 ± 1.400
Voting ConsensusGA [10]	0.867 ± 0.020	0.833 ± 0.021	0.875 ± 0.028	0.858 ± 0.014	16.500 ± 1.900

TABLE III: Performances of best tuned configurations according to their F-measure yielded for each ensemble type. Best results are highlighted in **bold**.

Measure	Max	Mean	Median	Voting
N_{within}	$\mathbf{3.000 \pm 0.000}$	3.800 ± 1.000	5.800 ± 1.500	6.000 ± 1.400
S	0.785 ± 0.033	$\mathbf{0.803 \pm 0.035}$	0.800 ± 0.041	0.801 ± 0.043
FPR	0.312 ± 0.079	$\mathbf{0.290 \pm 0.043}$	0.334 ± 0.054	0.305 ± 0.038
FNR	0.091 ± 0.037	0.079 ± 0.040	$\mathbf{0.064 \pm 0.038}$	0.076 ± 0.043
SP	0.688 ± 0.079	$\mathbf{0.710 \pm 0.043}$	0.666 ± 0.054	0.695 ± 0.038
PR	0.853 ± 0.029	$\mathbf{0.862 \pm 0.015}$	0.847 ± 0.022	0.856 ± 0.014
RC	0.909 ± 0.037	0.921 ± 0.040	$\mathbf{0.936 \pm 0.038}$	0.924 ± 0.043
Acc	0.835 ± 0.029	$\mathbf{0.850 \pm 0.027}$	0.845 ± 0.034	0.847 ± 0.034
Fm	0.880 ± 0.021	$\mathbf{0.890 \pm 0.022}$	0.889 ± 0.026	0.889 ± 0.027

ensemble, whose performance is slightly better than the other ones. It is interesting to see how a higher number of networks which can belong to the ensemble does not mean a higher performance of the system. An important analysis which must be highlighted is the impact of the genetic algorithm on the system. Practically all tuned configurations yield better when the genetic algorithm is used. So that, given N networks, an ensemble of them yield better if the genetic algorithm is employed to select which networks belong to the ensemble or not. Additionally, as it can be expected, a higher population size P can improves the performance of the system. It happens the same for the maximum number of generations G: a higher value of this parameter can enhance the performance.

Moreover, the results of the number of neural networks which are within the ensemble (N_{within}) against the number of neural networks which can belong to the ensemble (so that, N) are reported in Figure 5. It must highlighted that the number of networks within the ensemble is very different from each kind of ensemble. In this sense, ensembles which do not need a quite number of networks can be suitable to be used when resources are limited. It seems that the higher value of N the higher value of N_{within}. However, this behavior does not happen in every case. Additionally, higher values of the population size P and the maximum number of generations



(a) Voting (b) Max (c) Mean (d) Median

(e) Voting (f) Max (g) Mean (h) Median

Fig. 4: F-measure obtained by each ensemble (Fm) by considering a number of neural networks which can belong to the ensemble (N) and the number of generations (G). First row reports $preprocess = False$ while second row reports $preprocess = True$. Each subfigure represents a kind of ensemble. Note that the values of each method are connected between them with lines to better compare the results, but this does not mean that the results are related.

G incite a higher value of networks which are within the ensemble N_{within}. However, the most interesting behavior is the use of a lower number of networks when the preprocess step is applied.

V. Conclusions

This work has presented a proposal to detect breast cancer by employing histopathological images. The convolutional neural networks are the base of this approach. Given a model as a reference, several kinds of neural network ensembles have been applied to improve the performance of the convolutional neural network model. Furthermore, a genetic algorithm has also been developed in order to enhance the performance of the ensemble. Moreover, the suitability of a preprocess step of the input image based on a sharpening filter has also been considered. A study of the behavior of the different parameter values has been reported. This study reflects the preprocess and the genetic algorithm improve the performance of the ensemble. Moreover, the number of networks that belong to the ensemble is reduced according to the network selection provided by the genetic algorithm. This way, the required computation time and resources are also benefited. Finally, the proposed approach is appropriate to classify breast cancer images with a high performance according to the obtained results in the experiments.

References

[1] C. Stoean, R. Stoean, A. Sandita, C. Mesina, C. Gruia, and D. Ciobanu, "How much and where to use manual guidance in the computational detection of contours for histopathological images?" *Soft Computing*, vol. 23, no. 11, pp. 3707–3722, 2019.

[2] M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, "Medical image processing, analysis and visualization in clinical research," in *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*. IEEE, 2001, pp. 381–386.

[3] P. Baran, S. Mayo, M. McCormack, S. Pacile, G. Tromba, C. Dullin, F. Zanconati, F. Arfelli, D. Dreossi, J. Fox, Z. Prodanovic, M. Cholewa, H. Quiney, M. Dimmock, Y. Nesterets, D. Thompson, P. Brennan, and T. Gureyev, "High-resolution X-ray phase-contrast 3-d imaging of breast tissue specimens as a possible adjunct to histopathology," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2642–2650, 2018.

[4] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[5] G. Aresta, T. Araújo, S. Kwok, S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. Vu, M. To, E. Kim, J. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar, "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019.

[6] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural network," in *International Joint Conference on Neural Networks (IJCNN 2016) Vancouver, Canada*, 2016, p. 2560–2567.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 2, 2012, pp. 1097–1105.

[8] M. A. Molina-Cabello, R. M. Luque-Baena, E. López-Rubio, and K. Thurnhofer-Hemsi, "Vehicle type detection by ensembles of convolu-

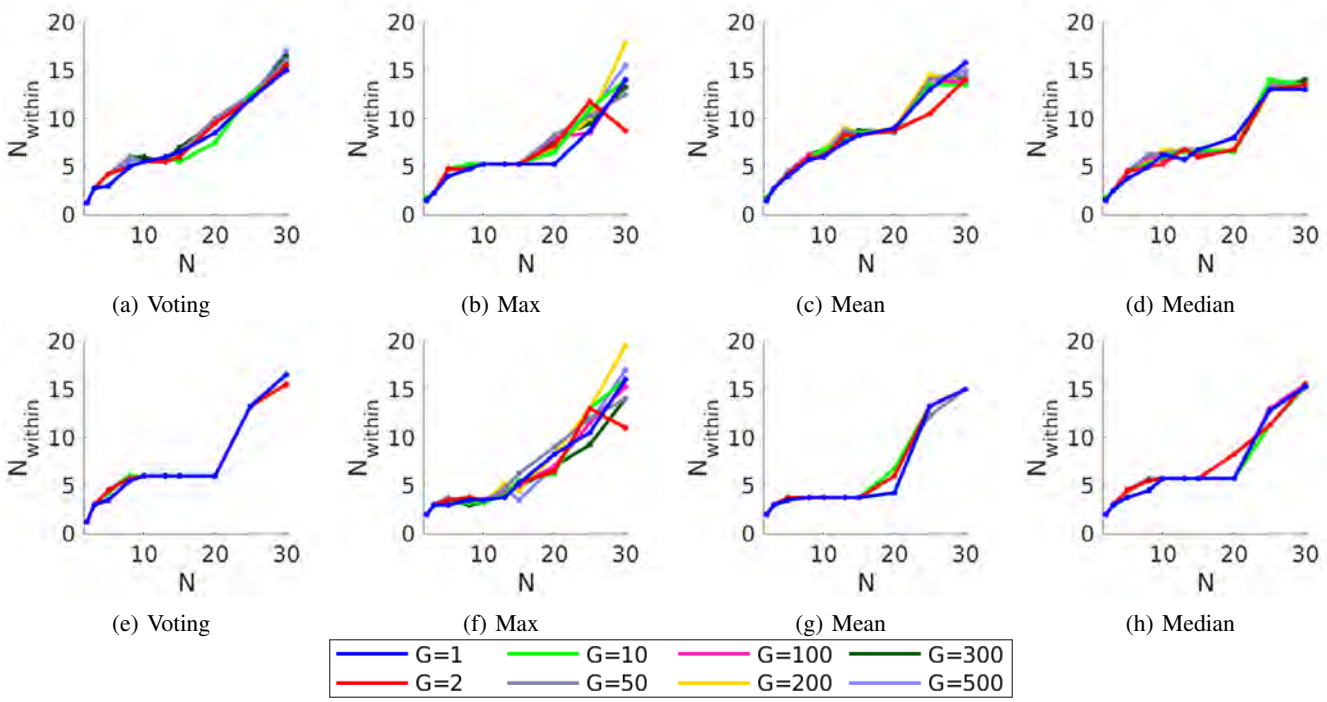


Fig. 5: Number of neural networks which are within the ensemble (N_{within}) by considering a number of neural networks which can belong to the ensemble (N) and the number of generations (G). Each subfigure represents a kind of ensemble. First row reports $preprocess = False$ while second row reports $preprocess = True$. Note that the values of each method are connected between them with lines to better compare the results, but this does not mean that the results are related.

tional neural networks operating on super resolved images," *Integrated Computer-Aided Engineering*, no. Preprint, pp. 1–13, 2018.

[9] M. A. Molina-Cabello, E. López-Rubio, R. M. Luque-Baena, M. J. Rodríguez-Espinosa, and K. Thurnhofer-Hemsi, "Blood cell classification using the hough transform and convolutional neural networks," in *World Conference on Information Systems and Technologies*. Springer, 2018, pp. 669–678.

[10] M. A. Molina-Cabello, C. Accino, E. López-Rubio, and K. Thurnhofer-Hemsi, "Optimization of convolutional neural network ensemble classifiers by genetic algorithms," in *International Work-Conference on Artificial Neural Networks*. Springer, 2019, pp. 163–173.

[11] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, "Deep convolutional neural networks for mammography: Advances, challenges and applications," *BMC Bioinformatics*, vol. 20, 2019.

[12] S. Guan and M. Loew, "Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks," *Journal of Medical Imaging*, vol. 6, no. 3, 2019.

[13] M. El Adoui, S. Mahmoudi, M. Larhmam, and M. Benjelloun, "MRI breast tumor segmentation using different encoder and decoder CNN architectures," *Computers*, vol. 8, no. 3, 2019.

[14] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1–6, 2019.

[15] K. Spuhler, J. Ding, C. Liu, J. Sun, M. Serrano-Sosa, M. Moriarty, and C. Huang, "Task-based assessment of a convolutional neural network for segmenting breast lesions for radiomic analysis," *Magnetic Resonance in Medicine*, vol. 82, no. 2, pp. 786–795, 2019.

[16] F. Khameneh, S. Razavi, and M. Kamasak, "Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network," *Computers in Biology and Medicine*, vol. 110, pp. 164–174, 2019.

[17] N. Wahab, A. Khan, and Y. Lee, "Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images," *Microscopy*, vol. 68, no. 3, pp. 216–233, 2019.

[18] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," *Med Image Comput Comput Assist Interv*, vol. 8150, pp. 411–418, 2013.

[19] A. Cruz-Roa, A. Basavanhally, F. A. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. E. Tomaszeweski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging*, 2014.

[20] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering (TBME)*, vol. 63, no. 7, pp. 1455–1462, 2016.

[21] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2440–2445, 2016.

[22] M. A. Nawaz, A. A. Sewissy, and T. H. A. Soliman, "Multi-class breast cancer classification using deep learning convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 9, 2018.

[23] M. H. Motlagh, M. Jannesari, H. Aboulkheyr, P. Khosravi, O. Elemento, M. Totonchi, and I. Hajirasouliha, "Breast cancer histopathological image classification: a deep learning approach," *bioRxiv 242818*, vol. 242818, 2018.

[24] F. Spanhol, P. Cavalin, L. S. Oliveira, C. Petitjean, and L. Heutte, "Caffe models trained on the images of breakhist acquired with 40x magnification factor." https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ , 2017, accessed: 2017-05-25.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, 2014, pp. 675–678.