

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Digital recognition of breast cancer using takhisnet: An innovative multi-head convolutional neural network for classifying breast ultrasonic images

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Nanni L., Lumini A., Maguolo G. (2023). Digital recognition of breast cancer using takhisnet: An innovative multi-head convolutional neural network for classifying breast ultrasonic images. Hershey : IGI Global [10.4018/978-1-6684-7544-7.ch066].

Availability:

This version is available at: <https://hdl.handle.net/11585/959223> since: 2024-02-19

Published:

DOI: <http://doi.org/10.4018/978-1-6684-7544-7.ch066>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

Digital Recognition of Breast Cancer Using TakhisNet: An Innovative Multi-Head Convolutional Neural Network for Classifying Breast Ultrasonic Images

Loris Nanni

Università di Padova, Italy

Alessandra Lumini

 <https://orcid.org/0000-0003-0290-7354>

University of Bologna, Italy

Gianluca Maguolo

University of Padova, Italy

ABSTRACT

In this chapter, the authors evaluate several basic image processing and advanced image pattern recognition techniques for automatically analyzing bioimages, with the aim of designing different ensembles of canonical and deep classifiers for breast lesion classification in ultrasound images. The analysis starts from convolutional neural networks (CNNs) in a square matrix that is used to feed other CNNs. The novel ensemble, named TakhisNet, is the combination by sum rule of the whole set of the modified CNNs and the original one. Moreover, the performance of the system is further improved by combining it with some handcrafted features. Experimental results obtained on the well-known OASBUD breast cancer dataset (i.e., the open access series of breast ultrasonic data) and on a large set of bioimage classification problems show that TakhisNet obtains very valuable results and outperforms other approaches previously tested in the same datasets.

INTRODUCTION

Breast cancer is one of the main causes of death for women living in western countries (Wild & Stewart, 2014). The diagnosis and the detection of a breast lesion relies upon ultrasound images. This technique is safe, low cost and let radiologists discriminate between benignant and malignant lesions with a very high accuracy. However, the need of an expert doctor to read ultrasound images increases the cost of screening and makes the diagnosis process operator-dependent (Byra, 2018).

In order to support radiologists, the recent innovation in the analysis of digital images led researchers to propose automatic classifiers with the purpose of discriminating between benignant and malignant tumors. E.g. automatic systems based on manually engineered features extracted from ultrasound images and fed into machine learning classifiers like Support Vector Machines (SVMs) are proposed in (Cheng, Shan, Ju, Guo, & Zhang, 2010). However, the recent rise of deep learning lead to the first attempts to use convolutional neural networks to recognize and classify malignant lesions in medical images (Saikia, Bora, Mahanta, & Das, 2019).

Convolutional neural networks (CNNs) are a class of neural networks designed to perform image classification, image segmentation and object recognition (Krizhevsky, Sutskever, & Hinton, 2012). One of the first successful attempt to use CNNs for image classification can be found in (Krizhevsky et al., 2012), where the authors designed a CNN able to outperform any previous classifier on the ImageNet challenge 2012. Since then, every winner of the ImageNet challenge was a CNN. Nowadays, modern CNNs obtain superhuman accuracies on ImageNet (He, Zhang, Ren, & Sun, 2015).

CNNs have already been used on several medical datasets reaching very high performance. In (Pereira, Pinto, Alves, & Silva, 2016) authors used deep CNNs with very small kernels to perform brain tumor segmentation from MRI images; in (Esteva et al., 2017) it is shown that a single CNN could detect keratinocyte carcinomas and malignant melanomas with the same accuracy as expert dermatologists; in (Chi et al., 2017) a fine-tuned version of GoogleNet (Szegedy et al., 2015) is proposed trained to classify thyroid nodules from ultrasound images; the work in (Lakhani & Sundaram, 2017) presented a system to detect pulmonary tuberculosis from radiographies using AlexNet (Krizhevsky et al., 2012) and GoogleNet (Szegedy et al., 2015). In (Byra, 2018) authors proposed a deep learning based method to classify breast cancer from ultrasound images. They used the OASBUD (Open Access Series of Breast Ultrasonic Dataset), a publicly available dataset containing 52 benignant lesions a 48 malignant lesions (Piotrkowska-Wróblewska, Dobruch-Sobczak, & Byra Michałand Nowicki, 2017). Their approach consisted in combining Fisher Linear Discriminant Analysis (Welling, 2005) and neural style transfer (Gatys, Ecker, & Bethge, 2016).

In this chapter we propose a novel method for the digital classification of breast cancer based on CNNs. The architecture of CNN consists of several convolutional and pooling layers stacked at the beginning of the network. These layers are usually followed by one or more fully connected layers. While the first layers of a CNN can be interpreted as trained feature extractors, the fully connected layers can be thought as a classifier. In particular, the last fully connected layer, after a softmax normalization, returns a probability distribution of the predicted label over the set of the possible classes. Our idea consists in replacing a pooling layer with other CNNs architectures. To be more precise, we add a reshape layer that takes the output of a pooling layer and randomly puts its entries in a square matrix. If the number of features is not a perfect square, the matrix can be padded with some zeros. After this reshape layer, we feed other CNNs with the new matrix. The CNNs on top are trained end-to-end with the first CNN and each of them return its own set of probability distributions. The final ensemble, resulting from the

sum rule of such probability distribution is named TakhisisNet, since it is a multi-head convolutional neural network.

Finally, to improve the performance of the system we have experimented the combination by sum rule of TakhisisNet with a set of Support Vector Machine (SVM) trained on handcrafted features.

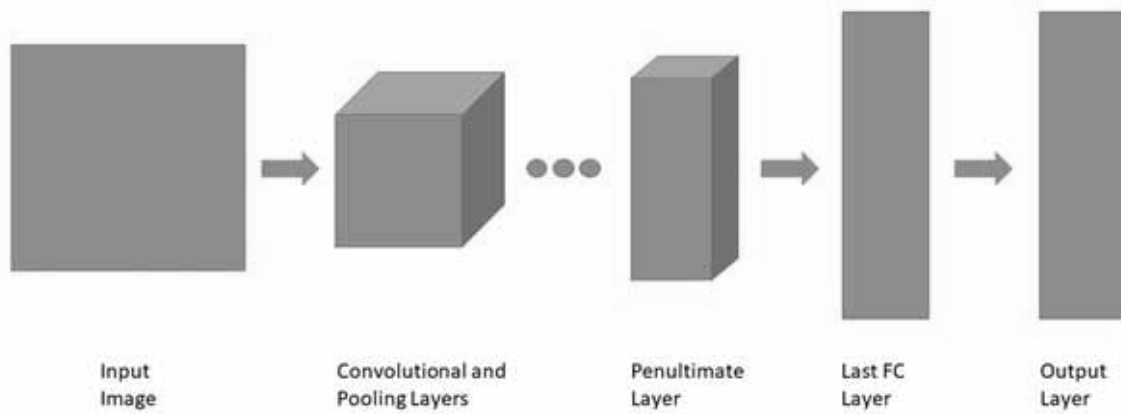
The chapter is organized as follows. The background describes the prior art for CNNs and the architecture of the two models used in this work: Vgg16 and GoogleNet. After that, the authors describes the proposed method, including the architecture of the TakhisisNet. In the Materials section the authors describe the Breast Ultrasound dataset and other medical datasets used during experimentation, and explain the details of the experimental setup. After that, the authors present the classification results along with a discussion of the findings. Finally, the conclusions for the study are presented.

BACKGROUND

The topology of CNN is divided into multiple learning stages composed of a combination of several types of layers: i.e. convolutional layers, non-linear processing units, subsampling layers, fully connected layers (Liu et al., 2017).

A convolutional layer takes a multi-channel image as an input and outputs a new multi-channel image, which is the discrete convolution between the input and a bank of convolutional kernels (filters). This kind of layers mimics the visual systems of a human eye. Depending on the weights of the kernel, convolutional layers can highlight specific geometric features of the input image like corners and boundaries. Besides, they can interpret the color of a pixel depending on the surrounding context. For example, a little square containing alternated red and yellow pixels will look orange: that is the founding principle of many painting techniques like pointillism. This means that the human eye cannot see the difference between images that might be far from each other in terms of matrix distance. Convolutional layers with suitable weights can compute an approximation of the average of the pixel colors in a small square and mimic this property of human vision, in particular when the input image is noisy. Output of the convolutional kernels is given to non-linear processing units, which helps in learning abstraction and embeds non-linearity in the feature space. This non-linearity produces different patterns of activations for dissimilar responses and thus facilitates in learning of semantic differences in images. Convolutional layers do not decrease the dimension of their input, as a neural network layer is often supposed to do. Hence, they are often followed by a pooling layer. A standard pooling layer is a subsampling unit that divides every channel of the input image in little squares with potential overlap and outputs a new image that has only one pixel for every square of the original image. Since a square has a larger dimension than a single pixel, these layers perform dimensionality reduction. The aim of subsampling is not only to summarize the results but also to make the input invariant to geometrical distortions. The last layers of a convolutional network are usually one or more fully connected layers, which are trained to perform a classification task. Unlike pooling and convolution, a fully connected layer performs a global operation: it takes input from the previous layer and makes a non-linear combination of selected features, which are used for the classification of data. The general architecture of a standard CNN can be found in Figure 1.

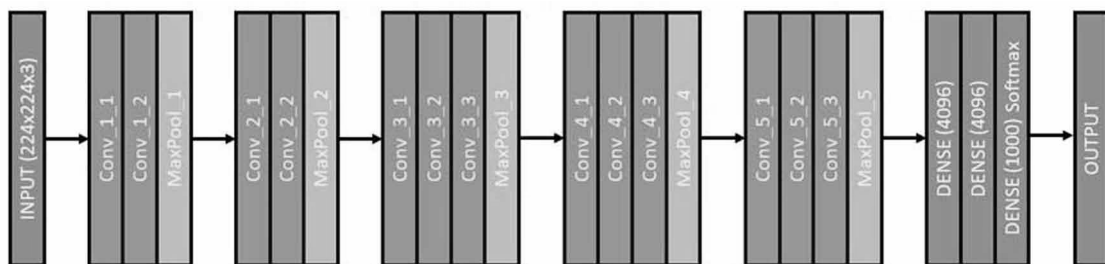
Figure 1. Standard CNN architecture



We performed our experiments using two different CNN architecture: VGG16 and GoogleNet.

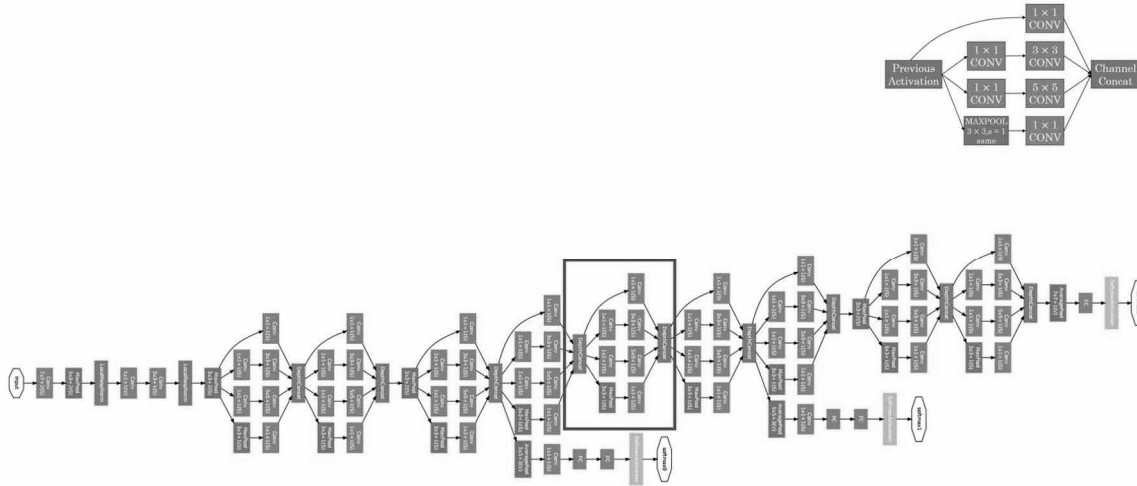
VGGnet (Simonyan & Zisserman, 2015) is a very heavy network that was runner-up in the ImageNet Challenge in 2014. VGGnet (figure 2) is one of the most used architectures for feature extraction from images, due to the simplicity of its uniform architecture. VGGNet is a linear combination of convolutional layers and pooling layers followed by 3 fully connected layers. Each Convolutional layer has the same configuration including a different number of small size filters (3×3 kernel). Each Max Pooling is designed to half the size of the image. The main limitation of VGG is its high computational cost: despite of the use of size filters, VGG require about 140 million parameters, most of which are given by fully connected layers. VGG16 is a 16 layers deep VGGNet.

Figure 2. VGGNet architecture



GoogleNet (Szegedy et al., 2015) is a light network that won the ImageNet Challenge in 2014. GoogleNet implements a novel element, named inception module (figure 3), which is a subnetwork consisting of parallel convolutional filters whose outputs are concatenated. GoogleNet uses batch normalization, image distortions and RMSprop optimizer. The use of inception modules that strongly reduce the number of parameters, therefore GoogleNet architecture consists of 22 layers but requires only 4 million parameters.

Figure 3. GoogleNet architecture



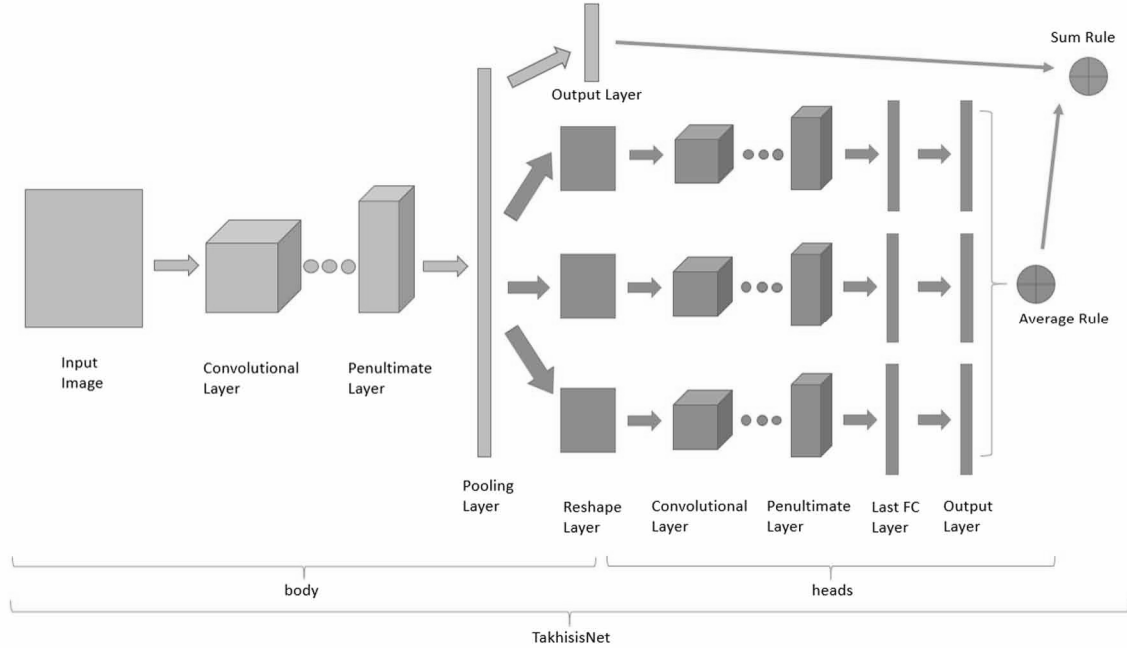
RESEARCH METHODOLOGY: TAKHISISNET ARCHITECTURE

In this chapter we propose a novel CNN architecture based on the fusion of several network: TakhisisNet is a multi head CNN which owes its name to the fictional character of Takhisis, a five-headed dragon. The architecture of the proposed ensemble is shown in figure 4. The “body” of the network is a standard CNN architecture (VGG16 or GoogleNet in this work), from which we extract the output of the Global Average Pooling layer for GoogleNet and the output of the last max pooling for Vgg16.

To be more precise, we add a set of reshape layers that take input from Global Average Pooling layer for GoogleNet and from last max pooling for Vgg16, and randomly reshape it into a square matrix (if the number of features is not a perfect square, the matrix is padded with some zeros). After the reshape layers, we feed other “head”-CNNs with the new matrices. The “head”-CNNs are trained end-to-end with the “body” CNN and each of them returns its own set of probability distributions. Then, we make our predictions by averaging the probability distributions of all the “head”-CNNs and summing it to the probability distribution of the body-CNN.

Besides, it creates an ensemble of neural networks trained end-to-end. In the experiments we used architectures with ten “heads”, each head is a different CNN whose input is the output of a reshape layer. Besides, we used the same architecture for the “body” CNN and for the “heads”. In the tests reported in this chapter we used 10 “heads” which are finally fused, by weighted sum rule, to the “body” to give the response of TakhisisNet. Before the fusion the scores of both “body” and “heads” are normalized to mean 0 and standard deviation 1 (before the normalization the scores of the heads are combined by sum rule); moreover, the weight of “body” is 2 and the weight of “heads” is 1.

Figure 4. Proposed TakhisisNet architecture



In our experiments we evaluated two different implementations of the reshape layers: the first one is based on a simple randomization of the position of the features inside the output matrix, the second is based on pre-processing the output vector according to a continuous wavelet transform before reshaping.

The Wavelet Transform is signal decomposition onto a set of basis function, which is derived from a single prototype wavelet by scaling (dilations and contractions) as well as translations (shifts).

In this work we use the Meyer mother wavelet (Daubechies, 1992), an orthogonal wavelet that is indefinitely differentiable with infinite support. The number of scales has been set to the length of the feature vector and in the case of feature vector larger than 5000, we reduce it to 5000 to avoid huge computational time. Then the output of the wavelet is resized to the size required by the CNN.

Materials

In order to validate our architecture, we performed experiments on several well-known medical datasets and we test our TakhisisNet on a Breast Ultrasound dataset.

The medical datasets used for preliminary validation, see Table 1, include: the Chinese hamster ovary cells (a proteomic profiling dataset), the 2D HELA, a dataset of cell images from fluorescence microscope acquisition, Locate Endogenous and Locate Transfected, two database of pattern for subcellular location and 6 biological dataset from the IICBU Biological Image Repository (FlyCel, Terminal bulb aging, Lymphoma, Muscle aging, Liver gender, Liver aging) <https://ome.grc.nia.nih.gov/iicbu2008/>, and three cancer datasets: the Human colorectal cancer dataset, the breast grading carcinoma and the Laryngeal dataset (Moccia et al., 2017). The 13 datasets are summarized in table 1, where the number of classes and samples for each dataset are reported.

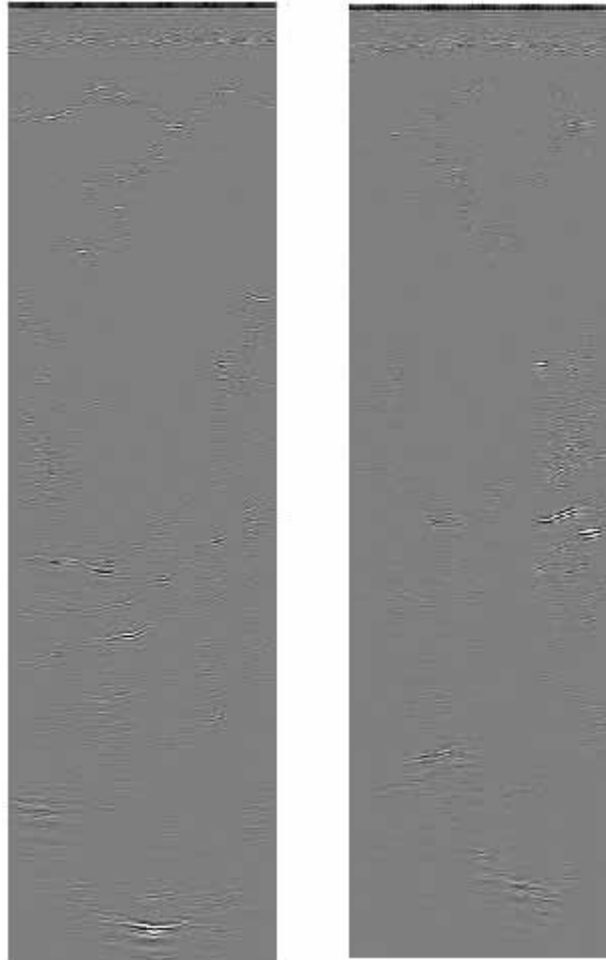
The protocol used in our experiments is five-fold cross-validation for 12 of 13 datasets. Since the Laryngeal dataset (Moccia et al., 2017) came already divided in training and testing, we used the suggested testing protocol in that dataset.

Table 1. Descriptive Summary of the Medical Datasets: ShortName, Name, the number of classes (#C), number of samples (#S), URL for downloading, Reference.

Dataset	Name	#C	#S	URL for Download
CH	Chinese hamster ovary cells	5	327	http://ome.grc.nia.nih.gov/iicbu2008/hela/index.html#cho
HE	2D HELA	10	862	http://ome.grc.nia.nih.gov/iicbu2008/hela/index.html
LO	Locate Endogenous	10	502	http://locate.imb.uq.edu.au/downloads.shtml
TR	Locate Transfected	11	553	http://locate.imb.uq.edu.au/downloads.shtml
RN	Fly Cell	10	200	http://ome.grc.nia.nih.gov/iicbu2008/rnai/index.html
TB	Terminal bulb aging	7	970	https://ome.grc.nia.nih.gov/iicbu2008
LY	Lymphoma	3	375	https://ome.grc.nia.nih.gov/iicbu2008
MA	Muscle aging	4	237	https://ome.grc.nia.nih.gov/iicbu2008
LG	Liver gender	2	265	https://ome.grc.nia.nih.gov/iicbu2008
LA	Liver aging	4	529	https://ome.grc.nia.nih.gov/iicbu2008
CO	Human colorectal cancer	8	5000	https://zenodo.org/record/53169#.WaXjW8hJaUm
BGR	breast grading carcinoma	3	300	https://zenodo.org/record/834910#.Wp1bQ-jOWU1
LAR	Laryngeal dataset	3	1320	https://zenodo.org/record/1003200#.WdeQcnBx0nQ

As to breast cancer classification is concerned, we used the OASBUD (Open Access Series of Breast Ultrasonic Data) (Piotrkowska-Wróblewska, Dobruch-Sobczak, & Byra Michałand Nowicki, 2017), a freely available breast lesion dataset¹. The dataset includes raw ultrasound data recorded from breast focal lesions corresponding to 52 and 48 scans from malignant and benign lesions of different people, respectively. Two orthogonal scans are available for each lesion, therefore in our system the final score of each lesion is obtained as the sum of the scores of the two orthogonal scans.

Figure 5. Samples form the OASBUD dataset



In figure 5 some samples from the OASBUD dataset are shown. Since the dataset is quite small, we used the leave-one-out cross validation as testing protocol: the test set consisted of 2 images from the same lesion and the training set consisted of all remaining images.

SOLUTIONS AND RECOMMENDATIONS

In order to validate the TakhisisNet architecture we perform experiments in several medical image classification tasks. In all the experiments the class distribution has not been maintained when splitting the datasets between training and testing (according to the k-fold cross validation testing protocol), and all the input images has been evaluated without ad hoc preprocessing. Since CNNs need input images of a fixed size, we have used the following approach: first the input image is padded to square size to maintain the aspect ratio, then it is resized using bilinear interpolation.

The evaluation of the proposed approaches and the comparison with the literature is performed according to two of the most used performance indicators in image classification: accuracy and AUC. Accuracy is the ratio between the number of true predictions and the total number of samples, while AUC, i.e. the Area Under the ROC-curve, is a performance indicator mainly used for 2-class problems which can be interpreted as the probability that the classifier will assign a higher score to a randomly picked positive sample than to a randomly picked negative one (Huang & Ling, 2005). To validate the experiments the Wilcoxon signed rank test (Demšar, 2006) has been used.

In the first experiment we evaluate the proposed architecture on several well-known medical datasets, comparing the performance of TakhisisNet with its “body”-CNN (i.e. GoogleNet or VGG16) and the ensemble of its “heads”-CNNs. For fine-tuning the networks on each image classification problem we have used the following parameter configuration: learning rate 0.0001, miniBatch size 30, twenty-five epochs to train the “body” CNN, and 100 epochs for training each “head” of the TakhisisNet after the reshape layers. During the training phase a simple data augmentation has been performed (only for “body”, not for the “heads”) including reflection and scaling on both x and y axes.

In table 2 the performance, in terms of accuracy, of TakhisisNet and its components is reported: the “body” is the standard CNN model (VGG16 or GoogleNet) fine-tuned on the specific dataset, the “heads” is the fusion by average rule of the 10 heads of the network. From the results in table 2 it is clear that TakhisisNet outperforms both its components in almost all the datasets.

Table 2. Performance (accuracy) obtained by TakhisisNet vs. Body vs. Heads varying the base model (GoogleNet and VGG16) and using the standard random method for the reshape layer.

MODEL	METHOD	DATASET												Avg
		CH	HE	LO	TR	RN	TB	LY	MA	LG	LA	BG	LAR	
GoogleNet	Heads	88.62	83.60	92.00	86.55	44.50	59.18	72.00	60.83	97.33	93.14	80.67	85.76	78.68
	Body	95.08	86.74	93.00	92.73	45.00	57.84	73.33	82.08	91.67	83.43	86.00	85.08	81.00
	TakhisisNet	95.38	87.79	93.80	92.73	49.50	60.10	74.40	83.33	94.00	88.00	87.67	85.83	82.71
VGG16	Heads	94.15	89.88	91.00	85.45	55.50	63.71	78.67	71.67	99.00	96.19	84.00	88.03	83.10
	Body	99.69	93.60	98.20	93.27	69.50	61.44	80.80	85.00	85.33	88.57	93.00	91.44	86.65
	TakhisisNet	99.69	93.72	98.00	93.64	72.50	64.12	83.47	86.25	87.67	91.62	93.00	91.89	87.96

In table 3 the performance obtained varying the reshape method (random or wavelet) is reported. Even if the results in table 2 suggests that VGG16 performs better than GoogleNet in almost all the tested problems, we selected GoogleNet as base model for next experiments since its requirements in terms of memory occupancy and GPU time for training are much lower than for VGG16. Moreover, for a sake of time only a subset of datasets have been evaluated in table 3. For a better comparison some results of table 2 are replicated in table 3. In the last two lines of table 3 we also report the performance of other two ensemble: ENS_HEADS, is the fusion between Heads based on Random and Heads based on Wavelet; ENS_TAK, is the fusion between the two TakhisisNet, the one based on Random and the other based on Wavelet.

Table 3. Performance (accuracy) obtained by TakhisisNet vs. Body vs. Heads varying the reshape method (random and wavelet) and using GoogleNet as base model. The last two rows report the performance of two ensembles.

MODEL (reshape method)	METHOD	DATASET								Avg
		CH	HE	LO	TR	RN	LG	LA	BG	
GoogleNet (random)	Heads	88.62	83.60	92.00	86.55	44.50	97.33	93.14	80.67	83.30
	Body	95.08	86.74	93.00	92.73	45.00	91.67	83.43	86.00	84.20
	TakhisisNet	95.38	87.79	93.80	92.73	49.50	94.00	88.00	87.67	86.10
GoogleNet (wavelet)	Heads	88.92	84.19	92.40	84.18	34.00	94.00	94.10	72.33	80.51
	Body	95.08	86.74	93.00	92.73	45.00	91.67	83.43	86.00	84.20
	TakhisisNet	95.69	87.67	93.40	93.09	47.50	92.67	88.00	87.33	85.66
ENS_heads		90.77	86.28	94.20	89.27	44.00	97.00	93.71	79.67	84.36
ENS_Tak		95.69	87.91	93.60	93.09	49.50	93.67	88.19	87.00	86.08

From the accuracies reported in Tables 2 and 3 and considering the results of the Wilcoxon signed rank test, the following conclusions can be drawn:

- TakhisisNet outperforms the base CNN architecture (“body”) with a p-value of 0.005 for both GoogleNet and Vgg16 models;
- ENS-heads outperforms with a p-value of 0.1 both heads-Random and heads-Wavelet;
- Unfortunately, there is not a static performance difference between ENS-Tak and TakhisisNet-Random or between ENS-Tak and TakhisisNet-Wavelet.

The second experiment is focused to the main aim of this work, breast ultrasound image classification. For computational reasons, due to the onerous testing protocol used in the breast ultrasound dataset (i.e. leave-one-out) we have run tests using only GoogleNet. In table 4 the performance of TakhisisNet (based on GoogleNet) are compared with other ensembles designed for a sake of comparison.

The first ensemble, named ENS_HAND is the combination of several SVM (Chang & Lin, 2011) classifiers trained with different handcrafted descriptors (L. Nanni, Brahnam, Ghidoni, & Lumini, 2018):

1. Multithreshold LPQ (MLPQ) (L. Nanni, Brahnam, Lumini, & Barrier, 2014) is a multi-threshold approach applied the LPQ descriptor. The MLPQ features used in our experiments are extracted and combined according to the method proposed in (L. Nanni et al., 2014).
2. Gaussian Of Local Descriptors (GOLD) (Serra, Grana, Manfredi, & Cucchiara, 2015) is a descriptor that models the local features distribution with a multivariate Gaussian, without any quantization.
3. Full BSIF (Loris Nanni, Paci, Caetano dos Santos, Brahnam, & Hyttinen, 2016) is an extension of the Binarized Statistical Image Feature (Kannala & Rahtu, 2012) that assigns each pixel of the input image a n-bit label obtained by means of a set of n linear filters. Full BSIF is a multi-threshold extension of BSIFT used to create an ensemble of classifiers.
4. Full RIC, is a multiscale Rotation Invariant Co-occurrence of Adjacent LBP (RIC) (Nosaka & Fukui, 2014). This variant extracts RIC features using different parameter settings, i.e. varying the

LBP radius and displacement among the LBPs. The values used in our experiments are: (1, 2), (2, 4) and (4, 8).

5. LETRIST, the descriptor proposed in (Song, Li, Meng, Wu, & Cai, 2017): a descriptor that explicitly encodes the joint information within an image across feature and scale spaces.

A detailed description of each of these descriptors along with information about the parameter sets used for each descriptor can be found in APPENDIX 1.

The second ensemble, named ENS_DEEP is the combination of several GoogleNets trained with different parameters or data augmentation methods (all the combination have been considered) for a total of 24 CNNs:

1. Batch Size $\in \{10, 30, 50, 70\}$
2. Learning rate $\in \{0.0001, 0.001\}$
3. Data Augmentation $\in \{\text{left-right reflection; 4 side reflection + scaling; 4 side reflection + scaling + roto-translation}\}$

The third ensemble, named ENS_TH is the fusion by the sum rule between TakhisisNet² and ENS_HAND; notice that before the fusion the scores of both TakhisisNet and ENS_HAND have been normalized to mean 0 and standard deviation 1.

The first line of table 4 is the most performing stand-alone GoogleNet belonging to the ensemble ENS_DEEP chosen considering the performance directly on the test set. Even if in many papers the tuning of parameters is performed selecting the best configuration on the test set, this practice can generate overfitting and it useful only to find an over bound to the performance. Therefore, the resulting network is named GoogleNet_OF. The second line is the most performing GoogleNet selected on the base of the performance in the training set. The third line is a VGG16 network fine-tuned on in the OASBUD dataset. The last two lines of table 4 report the best results published in the literature on this dataset.

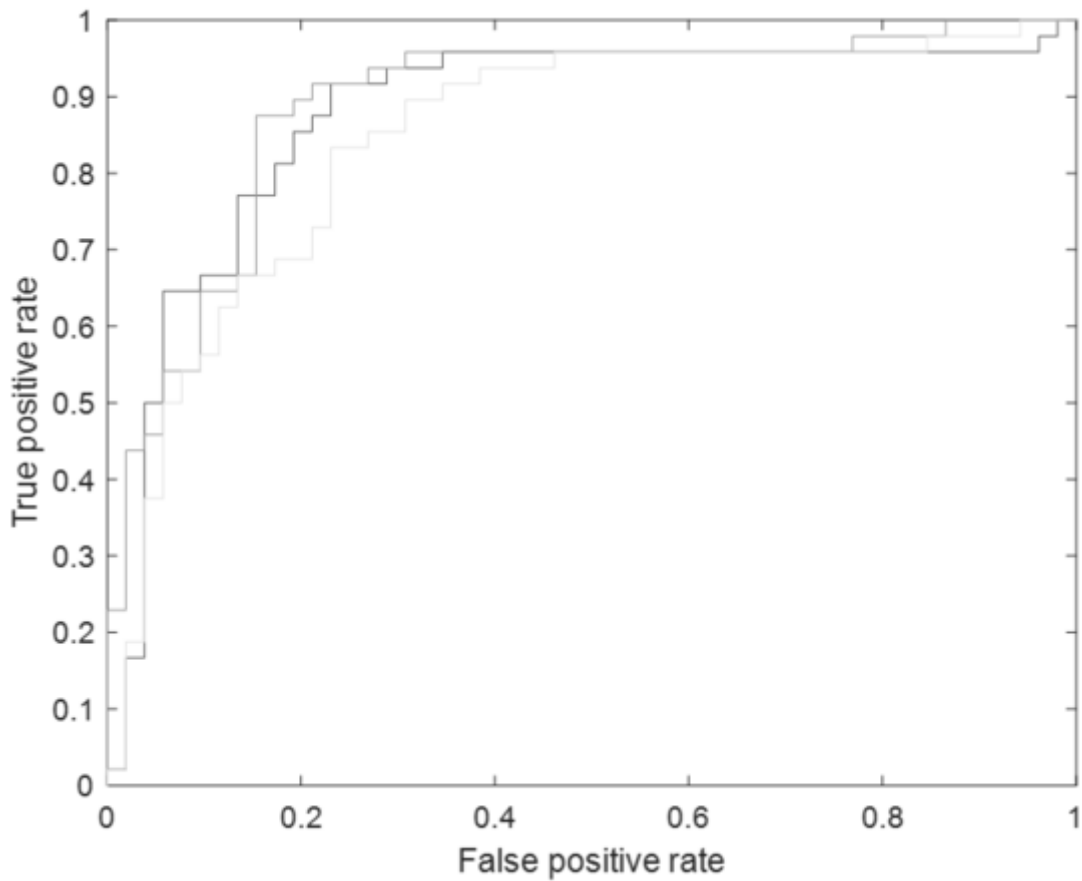
Table 4. Comparisons in the OASBUD dataset in terms of AUC

Method	AUC
GoogleNet_OF	0.855
GoogleNet	0.822
VGG16	0.849
ENS_DEEP	0.874
TakhisisNet	0.877
ENS_HAND	0.887
ENS_TH	0.891
(Byra, 2018)	0.847
(Antropova, Huynh, & Giger, 2017)	0.826

From the analysis of table 4 it is clear that the proposed system TakhisisNet obtains performance higher than that reported in the literature (Byra, 2018)(Antropova et al., 2017). Another clear result is

the usefulness of the ensemble: GoogleNet obtains an AUC of 0.822, drastically lower than the 0.877 obtained by TakhisisNet. The highest AUC value is equal to 0.887 and has been obtained by the ensemble ENS_TH designed by the fusion of both deep learning and handcrafted features: this fact highlights the diversity among the two methods that can be exploited to improve the overall performance in an ensemble. The ROC curves calculated for the worst and the best performing classifiers (the stand-alone GoogleNet (yellow curve), TakhisisNet (blue curve) and ENS_TH (green curve)) are shown in figure 6. The difference between AUC values was statistically significant according to the Wilcoxon signed rank test ($p\text{-value} < 0.1$).

Figure 6. ROC curves of the proposed methods: GoogleNet (yellow curve), TakhisisNet (blue curve) and ENS_TH (green curve)

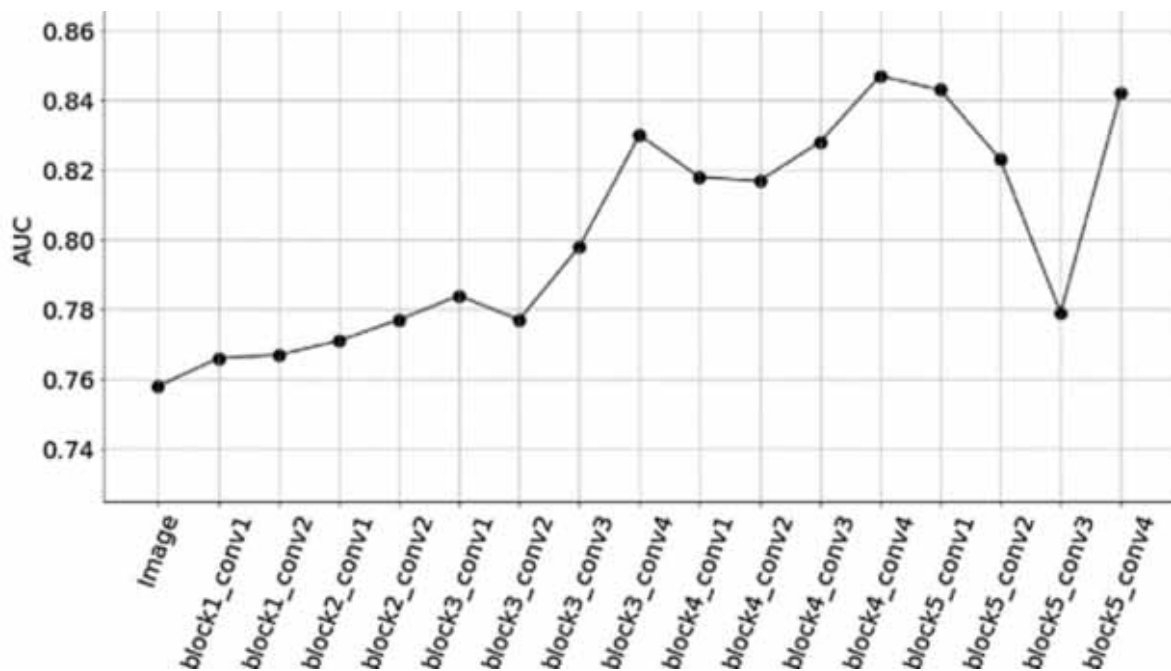


It is interesting to stress the high performance obtained by ensemble of texture descriptors; it outperforms the method based on CNN. As already show in other works, handcrafted descriptors can outperform deep learning approaches if a large training set is not available, as in the problem faced in this work.

Another interesting result is that we apply the same CNN ensemble validated on several medical datasets and we use a set of handcrafted descriptors that we have already proposed in the literature

(L. Nanni, Brahnam, Ghidoni, & Lumini, 2018); we have not optimized our ensemble for the breast ultrasound cancer dataset. That dataset is quite small, so it is risky to optimize a given approach on that dataset, e.g. in (Byra, 2018), see Figure 7, the performance of their approach is not very stable, to choose which layer to feed their classifier considering the performance on the test set could overfit the system.

Figure 7. Performance reported in (Byra, 2018) using different layers of VGG19 for the classification; “image” means that the raw imaged are used for feeding their classifier



As already stressed, the main problem with CNNs is that they need a large training set for the training/tuning step; we expect to obtain better performance when more patients will be available. For example, the method proposed in (Antropova, Huynh, & Giger, 2017) obtains an AUC of 0.826 in the breast dataset tested in this paper, instead in a larger dataset of 1125 breast lesions an higher performance is obtained: AUC of 0.872.

As shown in Figure 6 our system does not obtain a 100% true positive rate (TPR) with an acceptable false positive rate (FPR), the same problem occurs with (Byra, 2018); anyway also considering those performance indicators we obtain state of the art performance:

With a TPR of ~0.65 our approach obtains an FPR of 0.1, instead (Byra, 2018) obtains a TPR ~0.55 with an FPR of 0.1.

FUTURE RESEARCH DIRECTIONS

Usually, neural networks require large amounts of training data in the order of millions of images, especially when the resolution is low. Nonetheless, our experiments based on fine-tuning pre-trained networks and

the fusion of several CNNs shows that very high performance can already be achieved with few training images resulting in performance comparable or even higher than handcrafted descriptors. Although this work used a supervised training approach, i.e. the training images were manually labeled as malignant or benign lesion, there are several attempts in the literature aimed at unsupervised or semi-supervised training of neural networks (Rasmus, Valpola, Honkala, Berglund, & Raiko, 2015)(Erhan, Courville, Bengio, & Vincent, 2010) with some works the medical field (Kallenberg et al., 2016). Therefore, we expect that in the future the size of the training set would not be a problem, given the availability of large datasets of medical images that could be used for training neural networks in an unsupervised, or semi-supervised mode. Finally, we try to couple our idea also to other CNN architectures.

CONCLUSION

The purpose of the present chapter was to propose a novel CNN architecture named TakhisisNet, based on the fusion of a “body”-CNN of a set of “heads”-CNN trained end-to-end with the body. We have tuned and evaluated TakhisisNet on the digital classification of bio-images, and encouraged by the good accuracy achieved, we applied this promising network to breast ultrasound image. The experimental results demonstrate the promising performance of our classification system with a higher AUC than stand-alone approaches. It was unexpected that in this dataset our ensemble of handcrafted descriptors (ENS_HAND) performs better than other famous deep learning approaches: this fact is probably due to the small dimension of the training set; this fact is exploited to maximize the classification performance in this dataset by fusing the proposed TakhisisNet with the handcrafted ensemble. Our approach outperforms state of the art methods of the same dataset.

Even if the proposed system has shown to be robust also with a small training set, we hope that a further improvement can be obtained using large amounts of unlabeled data made available from the very high number of screening participants. Further study for semi-supervised training of TakhisisNet will be explored.

Finally, we share the MATLAB code of TakhisisNet for research purposes at <https://github.com/LorisNanni>.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation for the “NVIDIA Hardware Donation Grant” of a Titan X used in this research.

REFERENCES

Antropova, N., Huynh, B. Q., & Giger, M. L. (2017). A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical Physics*, 44(10), 5162–5171. doi:10.1002/mp.12453 PMID:28681390

- Byra, M. (2018). Discriminant analysis of neural style representations for breast lesion classification in ultrasound. *Biocybernetics and Biomedical Engineering*, 38(3), 684–690. doi:10.1016/j.bbe.2018.05.003
- Chang, C., & Lin, C. (2011). LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1–39. doi:10.1145/1961189.1961199
- Cox, D., & Pinto, N. (2011). Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011* (pp. 8–15). 10.1109/FG.2011.5771385
- Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision* (pp. 59–74). 10.1234/12345678
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. Ten Lectures on Wavelets; doi:10.1137/1.9781611970104
- Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7, 1–30. doi:10.1016/j.jecp.2010.03.005
- Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*. doi:10.1145/1756006.1756025
- Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*. doi:10.1109/TKDE.2005.50
- Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., & Igel, C. ... Lillholm, M. (2016). Un-supervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Transactions on Medical Imaging*. doi:10.1109/TMI.2016.2532122 PMID:26915120
- Kannala, J., & Rahtu, E. (2012). BSIF: Binarized statistical image features. In *21st International Conference on Pattern Recognition (ICPR)* (pp. 1363–1366). Academic Press.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11–26. doi:10.1016/j.neucom.2016.12.038
- Moccia, S., De Momi, E., Guarnaschelli, M., Savazzi, M., Laborai, A., Guastini, L., ... Mattos, L. S. (2017). Confident texture-based laryngeal tissue classification for early stage diagnosis support. *Journal of Medical Imaging (Bellingham, Wash.)*, 4(3), 34502. doi:10.1117/1.JMI.4.3.034502 PMID:28983494
- Nanni, L., Brahnam, S., Ghidoni, S., & Lumini, A. (2018). Bioimage Classification with Handcrafted and Learned Features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi:10.1109/TCBB.2018.2821127 PMID:29994096
- Nanni, L., Brahnam, S., Lumini, A., & Barrier, T. (2014). *Ensemble of local phase quantization variants with ternary encoding* (Vol. 506). Studies in Computational Intelligence; doi:10.1007/978-3-642-39289-4_8
- Nanni, L., Paci, M., Caetano dos Santos, F. L., Brahnam, S., & Hyttinen, J. (2016). Review on Texture Descriptors for Image Classification. In *Computer Vision and Simulation: Methods, Applications and Technology*. Nova Science Publisher.

- Nosaka, R., & Fukui, K. (2014). HEp-2 cell classification using rotation invariant co-occurrence among local binary patterns. *Pattern Recognition*, 47(7), 2428–2436. doi:10.1016/j.patcog.2013.09.018
- Nosaka, R., Ohkawa, Y., & Fukui, K. (2011). Feature extraction based on co-occurrence of adjacent local binary patterns. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi:10.1007/978-3-642-25346-1_8
- Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987. doi:10.1109/TPAMI.2002.1017623
- Ojansivu, V., & Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5099 LNCS, pp. 236–243). doi:10.1007/978-3-540-69905-7_27
- Piotrkowska-Wróblewska, H., Dobruch-Sobczak, K., & Byra Michał and Nowicki, A. (2017). Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. *Medical Physics*, 44(11), 6105–6109. doi:10.1002/mp.12538 PMID:28859252
- Rasmus, A., Valpola, H., Honkala, M., Berglund, M., & Raiko, T. (2015). Semi-supervised learning with Ladder networks. *Advances in Neural Information Processing Systems*.
- Serra, G., Grana, C., Manfredi, M., & Cucchiara, R. (2015). Gold: Gaussians of local descriptors for image representation. *Computer Vision and Image Understanding*, 134, 22–32. doi:10.1016/j.cviu.2015.01.005
- Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. CoRR, abs/1409.1
- Song, T., Li, H., Meng, F., Wu, Q., & Cai, J. (2017). LETRIST: Locally Encoded Transform Feature Histogram for Rotation-Invariant Texture Classification. *IEEE Transactions on Circuits and Systems for Video Technology*. doi:10.1109/TCSVT.2017.2671899
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). IEEE.

ADDITIONAL READING

- Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 127, 248–257. doi:10.1016/j.cmpb.2015.12.014 PMID:26826901
- Ciampi, F., Chung, K., Van Riel, S. J., Setio, A. A. A., Gerke, P. K., Jacobs, C., ... van Ginneken, B. (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific Reports*, 7(1), 46479. doi:10.1038/rep46479 PMID:28422152

Huynh, B. Q., Li, H., & Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging (Bellingham, Wash.)*, 3(3), 34501. doi:10.1117/1.JMI.3.3.034501 PMID:27610399

Li, H., Giger, M. L., Huynh, B. Q., & Antropova, N. O. (2017). Deep learning in breast cancer risk assessment: Evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *Journal of Medical Imaging (Bellingham, Wash.)*, 4(4), 41304. doi:10.1117/1.JMI.4.4.041304 PMID:28924576

Samala, R. K., Chan, H.-P., Hadjiiski, L., Helvie, M. A., Richter, C., & Cha, K. (2018a). Cross-domain and multi-task transfer learning of deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. In *Medical Imaging 2018* (Vol. 10575, p. 105750Q). Computer-Aided Diagnosis. doi:10.1117/12.2293412

Samala, R. K., Chan, H.-P., Hadjiiski, L. M., Helvie, M. A., Richter, C., & Cha, K. (2018b). Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine and Biology*, 63(9), 95005. doi:10.1088/1361-6560/aabb5b PMID:29616660

KEY TERMS AND DEFINITIONS

AUC: The “Area Under the ROC Curve” (AUC) measures the entire two-dimensional area underneath the entire ROC curve. One way of interpreting AUC is as the probability that the model ranks a random positive sample more highly than a random negative sample.

Convolutional Neural Networks: A convolutional neural network (CNN) is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data by means of learnable filters.

Deep Learning: Deep learning is a subset of machine learning that models high-level abstractions in data by means of network architectures, which are composed of multiple nonlinear transformations.

Ensembles of Classifiers: An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new samples.

k-Fold-Cross Validation: Cross-validation is a statistical method used to estimate the skill of machine learning models. This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. At each round one different fold is treated at turn as a validation set, and the method is trained on the remaining k-1. Finally, the performance are combined (e.g. averaged) over the rounds.

Leave-One-Out Testing Protocol: Leave-one-out cross-validation (LOOCV) is a particular case of k-fold-cross validation where k is the dimension of the set of observations, therefore each test set includes only one sample.

Machine Learning: Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

ROC Curve: A receiver operating characteristic (ROC) curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate vs. False Positive Rate.

ENDNOTES

¹ Available at <https://doi.org/10.5281/zenodo.545928>

² In this test, each CNN that built TakhisisNet is coupled with different learning rate and batch size (the same values of ENS_DEEP), moreover to boost performance we have used 20 heads instead 10

APPENDIX

In this appendix each of the descriptors used to build the ensemble ENS_HAND is briefly described.

Multithreshold LPQ: MLPQ (L. Nanni et al., 2014) is a multi-threshold approach applied the LPQ descriptor. LPQ (Ojansivu & Heikkilä, 2008) is a texture descriptor that uses the local phase information extracted from the 2-D short-term Fourier transform computed over a rectangular neighborhood of radius R at each pixel position in an image. In (L. Nanni et al., 2014) a multi-threshold version of the simple binary quantizer of LPQ is proposed where the different pixels are encoded with 3 values using a set of thresholds τ around zero:

$$q(x) = \begin{cases} 1, & x \geq \tau \\ 0, & -\tau \leq x < \tau \\ -1 & \text{otherwise.} \end{cases}$$

According to (L. Nanni et al., 2014) the quantization is performed using the following thresholds $\tau \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and concatenating the resulting descriptors. In this study we combine sets of MLPQ extracted by varying the following parameters: the neighborhood size ($R \in \{1, 3, 5\}$), the scalar frequency, ($\alpha \in \{0.8, 1, 1.2, 1.4, 1.6\}$), the correlation coefficient between adjacent pixel values $P \in \{0.75, 0.95, 1.15, 1.35, 1.55, 1.75, 1.95\}$. Each descriptor is used to train a different SVM classifier which are fused by sum rule.

Gaussian Of Local Descriptors: GOLD (Serra et al., 2015) is an improvement of the Bag of Word (BoW) approach (Csurka, Dance, Fan, Willamowski, & Bray, 2004). The original BoW method extracts local features to generate a codebook, which is used to encode the local features into codes that form a global image representation. The codebook generation step is performed through clustering methods on the training set. GOLD is a four-step process that generates the codebook via a flexible local feature representation that is obtained by a parametric probability density estimation. Differently from BoW, GOLD does not requires neither quantization nor a training set. The steps for extracting GOLD features are:

1. Feature extraction: dense SIFT descriptors (Cox & Pinto, 2011) are extracted on a regular grid of the image.
2. Spatial Pyramid Decomposition: the image is decomposed by a multilevel recursive method, and features are softly assigned to regions according to a local weighting.
3. Parametric probability density estimation: for each region local mean and covariance of extracted local descriptors are inferred, according to as a multivariate Gaussian distribution.
4. Projection on the tangent Euclidean space: the final description is the mean concatenated to the projection the covariance matrix on the tangent Euclidean space.

The resulting GOLD descriptor is fed into an SVM classifier with a histogram kernel.

Full BSIF: (Loris Nanni et al., 2016) is an extension of the Binarized Statistical Image Feature (BSIF) (Kannala & Rahtu, 2012) that assigns each pixel of the input image a n -bit label obtained by means of a set of n linear filters. The original BSIF is a local image descriptor constructed by binarizing the responses to n linear filters which, differently to previous binary descriptors, are learnt from natural im-

ages using independent component analysis (ICA). BSIF assigns an n -bit label to each pixel of an image using a set of n linear filters. The BSIF descriptor has two parameters: the filter size d (filters are $d \times d$) and the number n of features extracted. In the original BSIF the binarization of a digit d_i (the result of the application of a filter on a pixel) is performed according to a fixed threshold $\tau=0$:

$$b_i = \begin{cases} 1, & \text{if } d_i > 0 \\ 0, & \text{if } d_i \leq 0 \end{cases}$$

Full BSIF is a multi-descriptor version of BSIF obtained by varying the parameters of filter size ($d \in \{3, 5, 7, 9, 11\}$) and a threshold τ for binarizing ($\tau \in \{-9, -6, -3, 0, 3, 6, 9\}$). The resulting 35 combinations of descriptors are used each to train a different SVM that are combined by sum rule.

Full RIC: this is a multiscale Rotation Invariant Co-occurrence of Adjacent LBP (RIC-LBP) (Nosaka & Fukui, 2014). The RIC-LBP features are variants of the standard LBP designed to simultaneously have characteristics of rotation invariance and a high descriptive ability. LBP (Ojala, Pietikainen, & Maenpaa, 2002) is an operator that describes a local region as a binary pattern obtained by thresholding the difference between a center pixel and its neighboring pixels. LBP represents the magnitude relation of intensities, therefore it is robust against changes in illumination among image patterns, but it does not preserve structural information among binary patterns. The CoALBP (Nosaka, Ohkawa, & Fukui, 2011) uses LBP pairs to keep such structural information: the histogram feature generated from CoALBP contains information on the structure of the image, since it describes the frequency of LBP pairs that are located near each other. Rotation invariant is reached in RIC-LBP by using a map table that maps together all the LBP pairs which have rotation equivalence according to a rotation angle θ . Finally, an RIC-LBP histogram is generated from for the entire image. The multi-scale variant of RIC-LBP extracts RIC-LBP features using different parameter settings, i.e. varying the LBP parameters radius and displacement (P,R). The values used in our experiments are: (1, 2), (2, 4) and (4, 8).). Each resulting descriptor is used to train a different SVM which are combined by sum rule.

LETRIST: (Song et al., 2017) the Locally Encoded TRansform feature hISTogram descriptor is a histogram representation that explicitly encodes the joint information within an image across feature and scale spaces. It consists of the following three major steps: first local texture structures are characterized by a set of transform features their correlation is calculated by applying linear and non-linear operators on the extremum responses of directional Gaussian derivative filters in scale space. Second, these transform features are binarized into by scalar quantization. Third, the discrete texture codes are aggregated into a compact histogram representation using cross-scale joint coding. In this work we used the default parameter proposed in the MATLAB toolbox (<https://github.com/stc-cqupt/letrist>).