Breast Cancer Classification on Histopathological Images Affected by Data Imbalance Using Active Learning and Deep Convolutional Neural Network

Bogdan Kwolek[1,2(✉)], Michał Koziarski[1,2], Andrzej Bukała[2], Zbigniew Antosz[2], Bogusław Olborski[2], Paweł Wąsowicz[2], Jakub Swadźba[2,3], and Bogusław Cyganek[1,2]

[1] Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, 30 Mickiewicza, 30-059 Kraków, Poland
{bkw,cyganek}@agh.edu.pl
[2] Diagnostyka Consilio, 93-357 Łódź, Poland
sekretariat.consilio@diag.pl
[3] Department of Laboratory Medicine, Faculty of Medicine, Andrzej Frycz Modrzewski Kraków University, 30-705 Kraków, Poland

Abstract. In this work, we propose an algorithm for training deep neural networks for classification of breast cancer in histopathological images affected by data unbalance with support of active learning. The output of the neural network on unlabeled samples is used to calculate weighted information entropy. It is utilized as uncertainty score for automatic selecting both samples with high and low confidence. A number of low confidence samples that are selected in each iteration is manually labeled by pathologist. A threshold that decays over iteration number is used to decide which high confidence samples should be concatenated with manually labeled samples and then used in fine-tuning of convolutional neural network. The neural network can optionally be trained using weighted cross-entropy loss to better cope with bias towards the majority class.

1 Introduction

In the last decade, a significant effort has been put forth for breast cancer recognition from histological slides. Histological slides allow the pathologist to distinguish between the normal tissue, non-malignant (benign) tissue, and malignant lesions. Currently, substantial efforts are devoted to recognize the two fundamental types of breast cancer with Computer Aided Diagnosis (CAD) [1]. Various computer-based approaches for analysis of histological images have been proposed to support pathologists in quantifying morphological features [2], detecting malignant lesions [3], and predicting prognosis for breast cancer [1].

Prior to a visual analysis by pathologist the tissue samples are collected during biopsy and then stained with Hematoxylin and Eosin (H&E). Afterwards,

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11731, pp. 299–312, 2019.
https://doi.org/10.1007/978-3-030-30493-5_31

in traditional approach the pathologist examines microscopic images of the tissue samples from the biopsy with different magnification factors. To make the correct diagnosis, the doctor investigates various image features including patterns, textures, and different morphological properties [4]. Different magnification factors are inherent in analysis of histological images, and require panning, zooming, focusing, and the whole diagnosis process is very time consuming and tiresome. The diagnosis results are influenced by many subjective factors. As a consequence, such manual process sometimes leads to erroneous or insufficient diagnosis for breast cancer identification. Manual classification of histological images is laborious for pathologists, prone to inconsistencies, expensive and time-consuming. In some cases, detailed analysis of a single case could require several slides with multiple stainings. Moreover, pathologists undergo pressure to handle large volumes of cases while providing a larger amount of information in the pathology reports. In order to minimize risks associated with improper diagnosis of cancer, as well as to provide a support for pathologists in preparing reports, various image processing and recognition techniques have been elaborated for analyzing pathological images at microscopic resolution [5]. Unfortunately, traditional computer signal processing and computer vision techniques are not able to meet the requirements and to fulfill the expectations of clinicians.

Owing to advancement of digital imaging techniques, a remarkable progress in histological image processing and recognition has been made [1]. Modern whole slide image (WSI) scanners can process entire tissue slices and deliver high-resolution images. WSIs are very large in size and contain huge information. Such images are characterized by small inter-class variance and large intra-class variance. Moreover, features extracted from similar histological images with unlike magnification are usually very different. Thus, automatic classification of breast cancer pathological images is a challenging task.

The approaches to automatic classification of breast caner pathological images can be divided into two groups: methods based on feature engineering and classical machine learning, and methods based on feature/deep learning. In contrast to traditional approaches, which rely on hand-crafted features, recent algorithms learn useful features directly from the training image patches by the optimization of the loss function. The history of extracting handcrafted features for breast cancer recognition image classification is a long one [2]. In [6], a multiple magnifications-based framework for breast cancer on histopathological image classification has been proposed. The authors utilized various joint color-texture features and classifiers, and demonstrated that suitable feature-classifier combinations can largely outperform relevant methods. Before 2017, the system proposed in [6] outperformed in terms of recognition accuracy all machine learning based approaches. A review of current deep learning based approaches to histopathological image analysis can be found in [7]. Several methods were proposed for classification of histopathological images, and they mainly differ in architectures of convolutional neural networks, data augmentation, etc.

As noted in a recent survey [7], the problem of insufficient labeled images is very important in the area of histopathological image analysis. Most of the

approaches to cope with this issue fall into one of the following categories: (1) increasing the number of examples with annotations, (2) utilization of models/parameters for/from other tasks, or (3) exploitation of weak label or unlabeled data. Some research has been done on methods belonging to first two categories. In particular, several data augmentation and transfer learning based method were elaborated to improve learning of neural networks for histological image classification. However, little work has been done on exploitation of weak label or unlabeled data. One of the main reasons is that for histopathological image classification this is very hard and challenging task. One of the reasons that significant effort has only been devoted to classification of histopathological images is that almost all research group are working on freely available datasets, like BreaKHis [8], and thus they underestimate the costs and difficulties associated with data annotation. As previously mentioned, WSIs are very large in size, contain huge information and manual labeling of such images, which requires a highly qualified pathologists, is very time consuming. In [9] an interactive machine-learning system for digital pathology has been proposed. The proposed framework utilizes active learning to direct user feedback, making classifier training efficient and scalable in datasets containing huge amount of histologic objects.

In this work, we propose an algorithm for training deep neural networks for classification of breast cancer in histopathological images with support of active learning (AL). Instead of random selection, AL methods typically actively select samples with lowest confidence as the most valuable samples to add them to the query and finally train the model incrementally [10]. Randomly selecting samples instead of actively choosing samples establishes a lower bound. In the proposed method, both samples with high and low confidence are included in the query. We utilize information entropy as uncertainty score for automatic selection of both samples with high and low confidence. A threshold that decays over iteration number is used to decide which high confidence samples should be concatenated with manually labeled samples and then used in fine-tuning of a convolutional neural network. A *weighted* entropy [11] is calculated on the basis of prediction of CNN, which is tuned in every iteration in such a way. The pool of labeled samples is updated with newly labeled samples by the pathologist. Such high confidence samples for the labeling are selected automatically by the algorithm on the basis of weighted entropy after selecting the uncertain samples.

The contribution of this work is as follows: first, we propose an improved deep convolutional neural network model to achieve accurate and precise classification or grading of breast cancer pathological images. Meanwhile, online data augmentation, transfer learning and fine-tuning strategies are employed to avoid model overfitting effectively. Second, we propose an active learning scheme for fine-tuning a deep residual convolutional neural network on unbalanced data. Finally, experimental results based on freely available pathological image dataset show that the performance of our method is better or at least comparable with recent state-of-the-art methods for breast cancer classification on histopathological images, with good robustness and generalization ability.

2 Breast Cancer Classification on Histopathological Images

During analysis of the stained tissue, pathologists examines overall architecture of tissue, along with nuclei layout, density and variability. The diagnosis process using H&E stained biopsies is not trivial, and the average diagnostic compatibility between pathologists is about 75% [12]. Most of the WSI scanners that are presently in the use carry out slide scanning at $\times 20$ or $\times 40$ magnification with spatial resolution in order of $0.5\,\mu$/pixel and $0.25\,\mu$/pixel, respectively. One of the major difficulties in breast cancer histopathology image analysis is variability of appearance, which is mostly the result of variations in the conditions of the tissue preparation and staining processes. The color appearance can significantly vary due to differences in fixation and in staining processes. A typical histopathology slide comprises a tissue area of about 15×15 mm. Considering the resolutions on which the slides are scanned, the scanned images have size of up to several gigapixels. Taking into consideration that classification of larger images requires far larger number of parameters, as well as that typical WSI image can consists of as many as tens of billions of pixels, the WSI scans are divided into patches of size a few hundred pixels times a few hundred pixels, see also Fig. 1, which are then analyzed independently. The mentioned figure depicts sample histopathological images [8] with two fundamental types of breast cancer: benign and malignant.

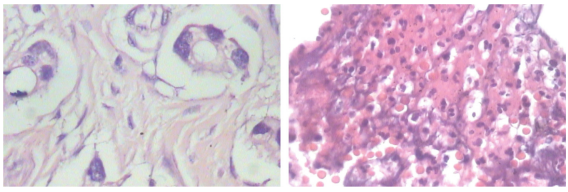

Fig. 1. Sample histopathological images with two fundamental types of breast cancer: benign (left) and malignant (right).

In the area of digital pathology, there are available some freely available datasets that contain hand-annotated histopathological images and corresponding labels. In [13], a performance comparison between four machine learning algorithms, including decision tree (DT), Naïve Bayes (NB), k-nearest neighbors (KNN) and support vector machine (SVM) on the Wisconsin Breast Cancer dataset [14], which consists of 699 instances (458 benign and 241 malignant cases). Experimental results demonstrated that the SVM classifier achieves the highest accuracy of 97.13% with 10-fold cross-validation. BreaKHis dataset [8] contains 7 909 histopathological images of breast cancer from 82 patients. The authors employed six different feature descriptors and four different classical machine learning methods, including 1-NN (1 Nearest Neighbor), QDA

(Quadratic Discriminant Analysis), RF (Random Forest), and SVM with the Gaussian kernel function to carry out classification of benign and malignant tumors. The classification accuracy is between 80 and 85% in 5-fold cross-validation. Although classical machine learning methods demonstrated great usefulness in digital pathology, present deep learning-based methods outperform traditional ones. In [15] a modified AlexNet [16] convolutional neural network improved classification accuracy by 4–6% on BreaKHis dataset. A CNN-based method proposed in [17] can classify breast cancer histopathological images independently of their magnifications. Two different architectures were studied: a single task CNN to predict malignancy, and a multi-task CNN to predict both malignancy and image magnification level simultaneously. Evaluations were carried out on the BreaKHis dataset, and the experimental results were competitive with state-of-the-art results achieved by classical machine learning methods. In [18] a pre-trained ResNet_V1_152 [19] has been applied to carry out diagnosis of benign and malignant tumors on BreaKHis as well as multi-class classification of various subtypes of histopathological images of breast cancer. This deep CNN achieved an accuracy of 98.7% and 96.4% for binary classification and multi-class classification, respectively. It is worth noting that although there are 7 909 histopathological images in the BreaKHis dataset, the number of images is far from enough for effectively using current deep learning techniques.

Another difficulty in breast cancer classification on histopathological images is class imbalance [20]. It can be observed in several histopathological benchmark datasets, including BreaKHis. As underlined in [20], it is largely unclear to what extent the data imbalance affects the performance of deep learning algorithms for histopathological image analysis, and what techniques [21] should be applied to learn from unbalanced data. Overall, class unbalance is very important problem, which is, however, frequently neglected in many evaluations on BreaKHis dataset. Moreover, several studies reports only the classification accuracy, which is sometimes badly chosen as the only metric to judge the classification performance on the imbalanced data.

Significant effort has been devoted to classification of histopathological images. Several research group performed evaluations on freely available datasets with histopathological images, and thus they were not involved in manual data annotation, which is very costly and time consuming task. Active learning is a machine learning technique, which is generally effective when the acquisition cost of label data is substantial [7]. It can be used to support supervised learning for automatic choosing the most valuable unlabeled sample(s) (i.e. the one(s) that could led to improved classification performance when labeled properly and included in training data) and display it for manual labeling by pathologists. Most active learning focuses on selecting examples from a so-called area of uncertainty, i.e. space that is nearest to the model's decision boundary, which for a binary classification problem can be expressed as: $x' = \arg\max_x \min_y P(y|x), y \in \{0, 1\}$. However, simple selecting samples from an unlabeled pool with considerable data imbalance may pose some practical difficulties. The greater fraction of examples in the majority class may result

in a model preferring one class over another. If labels of samples selected by an active learning algorithm are considered as random variables, the class imbalance would result in preference for majority examples in the training dataset subset, i.e. over-representation. Unless properly treated such an over-representation may result in a model with predictive preference for the majority class when labeling. This important issue motivated us to explore active learning on unbalanced data for cancer classification on histopathological images by the use of convolutional neural networks.

3 The Algorithm

The aim of active learning algorithms is to attain best possible performance of the learned model with as few labeled samples as possible [10]. Standard AL algorithms run in an iterative manner and in each learning step usually select only a few of the most informative samples, i.e. samples that have quite low prediction confidence, and frequently engage the user to label the recommended data. The algorithm stops when a predefined stopping criteria is met. In case of unbalanced data, conventional classification algorithms are often biased towards the majority class because their loss functions attempt to optimize quantities, which do not take into account the data distribution. In the worst case, particularly when the dataset is severely unbalanced, minority examples can be treated as outliers of the majority class and ignored. The learning algorithm simply produces a trivial classifier with tendency to classify every example as the majority class. An approach proposed in [22] performs active learning using both majority & clearly classified samples and minority & most informative samples. However, it is unable to cope with unbalanced data due to reasons mentioned above.

We use weighted information entropy [11] as uncertainty score. Weighted entropy is a measure of information supplied by a probabilistic experiment whose elementary events are characterized both by their objective probabilities and by some qualitative (objective or subjective) weights associated with the events. We characterize each event x_i by $\{p_i, w_i\}$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} p_i = 1$, $w_i \leq 0$, where p_i is the probability of the event x_i and weight w_i quantifies qualitative aspect of x_i. The weighted entropy of $\{p_i, w_i\}$, $i = 1, \ldots, n$ is defined as:

$$H_w(w_1, w_2, \ldots, w_n; p_1, p_2, \ldots, p_n) = -\sum_{k=1}^{n} w_k p_k \log p_k \tag{1}$$

In our approach the weights w_i are determined on the basis of class probabilities. We give higher weight to minority class and lower weight to majority class.

The pseudo-code of the algorithm is listed below. The input arguments are as follows: `max_it` – maximum number of iterations, `unc_samp_size` – number of the most uncertain samples that are selected in each iteration for manual labeling by the pathologist, `delta` – initial threshold value that is used in selection of the most confident samples, `delta_decay` – smaller than zero factor to decrease the value of delta if the classifier performs better, `x_init, y_init` –

Algorithm 1. Active learning on unbalanced data

```
Input: max_it, unc_samp_size, delta, delta_decay, x_init, y_init, x_pool, -, W
1: PL = x_init, y_init
2: uratio = num(y_init==0) / num(y_init==1)
3: PU = x_pool, -
4: train(PL, W, uratio)
5: W = load_weights
6: for i in range(max_it):
7:     y_pred_prob = predict(PU, W)
8:     un_idx = get_uncertain_samples( y_pred_prob, unc_samp_size, uratio)
9:     y_pool[un_idx]                # perform manual labeling of PU[un_idx]
10:    PL = append(PL, PU[un_idx])
11:    hc_idx = get_high_confid_samples(y_pred_prob, delta, uratio)
12:    hc = hc_idx - un_idx # remove samples also selected as uncertain
13:    PH = PU[hc]
14:    ptrain = concatenate(PL, PH)
15:    train(ptrain, W)          # optionally: train(ptrain, W, uratio)
16:    PU = delete(PU[un_idx])
17:    W = load_weights
18:    delta = delta * delta_decay
19:    uratio = num(y_pred_prob==0) / num(y_pred_prob==1)
20:    acc = evaluate(x_test, y_test, W)
21: return W
```

initial training pool consisting of samples and corresponding labels, `x_pool` – unlabeled pool of data samples, `W` – weights of the pre-trained neural network. The factor `uratio`, see line #2 in Algorithm 1, is used to express unbalance in dataset, and it can be determined as ratio of number of samples in each class. The output of the predictor, see line #7 in Algorithm 1, is used to calculate the weighted entropy, which in turn is used to determine the most uncertain samples, see line #8, as well as high confidence samples, see line #11. The number of uncertain samples selected in each iteration from the unlabeled pool `PU` depends on value of the predefined constant `unc_samp_size`. After selecting such samples, see index `un_idx`, a pathologist manually assigns labels to the recommended samples. The labeled pool `PL` is progressively updated in each iteration by `PU[un_idx]`, see line #10. The discussed pool `PL` is concatenated with high confidence pool `PH` and then used to fine-tune the neural network, i.e. to update weights `W`. This means that the samples from the unlabeled pool `PU` are progressively fed into the convolutional neural network. Depending on the option, it can be trained using commonly used categorical cross-entropy or weighted categorical cross-entropy in order to better cope with unbalanced data. The indexes of samples from `PU` to be included in `PH` are calculated on the basis of the method `get_high_confid_samples`, which selects samples whose weighted entropy is smaller than `delta`, see line #11. The `delta` variable is updated in every iteration, see line #18. In this way, more and more confident samples are

selected in subsequent iterations. The `uratio` factor is calculated on the basis of the predictions `y_pred_prob`. Since the `get_high_confid_samples` method can select samples that were previously selected by `get_uncertain_samples` method, the algorithm excludes samples with indexes `un_idx`, which were labeled by the pathologist. During training of the neural network in a predefined number of epochs, the best weights are stored, and then loaded before evaluation of the accuracy, as well as calculating the prediction in the next iteration. The pool of unlabeled samples `PU` is updated at the end of each iteration. To overcome the influence from the imbalanced histopathological images in subclasses, the minority class can be additionally balanced by turning images up and down, right and left, and rotating them counterclockwise by 90 and 180°.

4 Experimental Results and Discussion

We investigated the performance of various convolutional neural networks in breast cancer classification on histopathological images. It is well known that learning hyperparameters have a great influence on the performance of the trained CNN model, particularly the learning rate. Thus, in addition to investigations on various models of neural networks for active learning on unbalanced data, we devoted a considerable attention to selecting the learning hyperparameters for training. Due to limited amount of training images with breast cancer and model over-fitting risks we investigated techniques for reducing the number of CNN parameters as well as various data augmentation methods. The evaluations were realized on BreaKHis dataset consisting of 2480 images in benign class and 5429 images in malignant class. We randomly selected 6418 images for training subset, 802 images for validation and 689 images for test subset.

4.1 Breast Cancer Classification on Histopathological Images Using Deep Convolutional Neural Networks

At the beginning we investigated transfer learning of pre-trained VGG16 convolutional neural network for breast histopathology image classification. We utilized VGG16 with weights learned on imageNet dataset without the top layer. We extracted the features delivered by the VGG without the top layer and stored them for the future use. Next, on features with shape $(7, 7, 512)$ we trained a convolutional neural network consisting of 256 filters in the first layer and relu activation, dropout layer and output layer with sigmoid activation. The network has been trained using binary cross-entropy loss and RMSprop algorithm with learning rate set to 2e-5. The batch size has been set to 32 and training was in 200 epochs. The classification accuracy on test data was equal to 87.16%. Afterwards, we investigated fine-tuning of VGG16 with on-line data augmentation. The weights of VGG16 network without the top layer were frozen. After flattening the last layer of such a base network, we added a dense layer consisting of 256 neurons with relu activation, dropout set to 0.5 and an output neuron with activation set to sigmoid. The network has been trained using binary cross-entropy

loss and RMSprop algorithm with learning rate set to 2e-5. During training we executed online data augmentation (rotation, horizontal/vertical shift, image flip). The batch size has been set to 32 and training was in 30 epochs. The classification accuracy on the test data was equal to 85.7%. Finally, the `block5_conv1` layer has been set as trainable and complementary training of the network with such a trainable layer has been done in 100 epochs. The classification accuracy on the test dataset improved to 96.5%.

In the next stage we investigated resNet neural networks. Similarly to experiments with the VGG network, the fully connected layer of the pre-trained network has been replaced with a new fully connected layer. The network was pre-trained using the same parameters on the same training data as the VGG16 neural network. Next, the `res5a_branch2a` layer has been set as trainable and complementary training of the network with such a trainable layer has been done in 60 epochs. During training the network achieved 99.0% accuracy on the validation data. On test data the accuracy was equal to 97.8%.

In the last stage of this part of experiments we investigated pre-trained resNet18 neural network with weights learned on the imageNet dataset. The fully connected layer of the network has been replaced with a new fully connected layer with 256 neurons and the network has been pre-trained on the same training data using identical parameters and online data augmentation. The best classification accuracy that we obtained was about 75%. Finally, the `zero_padding2d_18` layer has been set as trainable and complementary training of the network with such a trainable layer has been done in 60 epochs. On the test data the classification accuracy improved to 91.5%.

The above experiments demonstrated that by setting in the base neural network the last layer as trainable and then extending such a base network about a dense layer, pre-trained in advance and fine-tuning the neural network in several epochs can lead to substantial improvement in classification performance. Bearing in mind that resNet18 neural network has far smaller number of training parameters in comparison to VGG16 and resNet50 neural networks, and thus the fine-tuning can be done in far shorter time, and particularly that our focus was on developing best strategies for active learning for breast cancer classification on histopathological images affected by data unbalance and not just experimenting with neural network architectures to obtain improvements in classification accuracy, the resNet18 neural network has been selected for further investigations on active learning algorithms.

4.2 Breast Cancer Classification on Histopathological Images Using Active Learning and Deep Convolutional Neural Network

In the next stage of the experiments we investigated active learning algorithms for breast cancer classification on histopathological images affected by data unbalance. The active learning algorithm was based in resNet18 neural network with weights learned on the imageNet dataset. The weights of the network without the top layer were frozen. The fully connected layer of the network has been replaced with a new fully connected layer consisting of 256 neurons with

relu activation and subsequent dropout set to 0.5. The output of the network was softmax layer. This means that the `zero_padding2d_18` layer that in the previous experiment was set as trainable, and thanks to which the classification accuracy considerably improved, in the discussed evaluations has been frozen. From the training data, twenty percent of training samples were selected for initial training of the neural network. The neural network has been initially trained in five epochs, and size of batch with data shuffling equal to 32. To better cope with imbalanced data, it has been trained using *weighted* categorical cross-entropy and RMSprop algorithm with learning rate set to 2e-5. The class weights were determined on the basis of data labels in the initial data. The best weights obtained in the training were stored and then utilized to initialize the network in the subsequent iterations, see line #4 – 5 in Algorithm 1. The active learning has been performed in nine iterations. In each iteration a complementary training of the network in five epochs has been carried out. The number of uncertain samples, i.e. samples labeled in each iteration by pathologist has been set to 300.

Table 1 presents results that were achieved in experimental evaluations. The presented results are averages of scores from ten independent runs of each considered algorithm with unlike weights initializations. First row contains results that were achieved by a baseline active learning algorithm with samples selected randomly for annotation by a pathologist. The discussed algorithm does not use the high-confident samples as pseudo-annotated data. Second row contains scores that were obtained by active learning algorithm using samples that were selected on the basis of weighted categorical cross-entropy. In a similar way to the previous algorithm, it does not use the high-confident samples as pseudo-annotated data. As we can observe, this algorithm achieves better classification performance. Third row contains scores that were obtained by algorithm using samples that were selected on the basis of categorical cross-entropy, and in which high-confident samples were utilized as pseudo-annotated data. The accuracy, recall and F1-scores are better in comparison to scores achieved by previously discussed algorithms. The last row contains results achieved by the proposed algorithm, which employs samples that are selected on the basis of weighted categorical cross-entropy, and in which high-confident samples are utilized as pseudo-annotated data. The accuracy, recall and F1-scores are superior to scores achieved by algorithms discussed previously. Particularly, owing to using the

Table 1. Classification performance on BreaKHis dataset using active learning: rs - random sampling, hc - high confidence samples, wce - weighted cross-entropy.

Algorithm	Accuracy	Precision	Recall	F1-score
rs, hc-no	0.9258	0.9416	0.9517	0.9467
wce, hc-no	0.9428	0.9652	0.9517	0.9583
wce-no, hc	0.9467	0.9644	0.9584	0.9613
wce, hc	0.9507	0.9612	0.9678	0.9644

weighted cross-entropy and high-confident samples as pseudo-annotated data our algorithm achieves the smallest number of false negatives, i.e. it achieves the highest recall. This is highly desirable property because practically it is very dangerous and costly to miss an image with malignant while it is positive.

Figure 2 illustrates sample results that were obtained in one of the mentioned above experiments, in which we evaluated the classification performance of active learning using entropy (c.f. green curves) and weighted entropy (c.f. blue curves). Left plot on Fig. 2 depicts the evolution of size of `PL` pool over iteration number as well as evolution of size of `PH` pool (consisting of high confidence samples) vs. iteration number. As we explained in Sect. 3, the concatenated `PL` and `PH` samples were used to train the CNN. As previously mentioned, the number of `PL` samples increases about 300 in each iteration, whereas the number of `PH` samples depends on their uncertainty score referred to the `delta` parameter, and therefore can be different in each iteration, see also green and blue curves on the left plot on Fig. 2. Right plot on Fig. 2 presents classification accuracies vs. iteration number. As we can observe, on the initial training pool consisting of only labeled data, the algorithm based on entropy achieved 86% accuracy, whereas algorithm based on weighted entropy achieved 83.9% accuracy. In seventh iteration the entropy-based algorithm learned on 3035 samples from PL pool and 3146 training samples from PH pool, and achieved on such training data 93.3% classification accuracy, whereas weighted entropy-based algorithm learned on 3035 samples from PL pool and 3335 samples from PH pool and achieved 94.62% classification accuracy. In all remaining experiments, the classification accuracies achieved by weighted entropy-based active learning in iterations #3 – 7 were higher in comparison to accuracies achieved by entropy-based active learning. As we can observe on Fig. 2, in next iterations the increase of the classification accuracy was not so high despite larger number of the `PL` samples and smaller proportion of `PH` data in total training data fed to the CNN. In tenth iteration the classification accuracy on the test data was equal to 95.06% for entropy-based algorithm and 95.21% for weighted-entropy based algorithm. The classification accuracies are far larger than 91.5% accuracy achieved by the resNet18 neural network with `zero_padding2d_18` set as trainable, c.f. results in Subsect. 4.2. Comparing results achieved by resNet18 neural network with the fully connected layer replaced by new fully connected layer and fine-tuned as in most relevant work, c.f. Subsect. 4.2, and results obtained with active learning, we can observe considerable improvement of classification accuracy. The increase of the classification accuracy from about 75% to 95% has been achieved owing to use our techniques for training neural networks on histopathological images affected by unbalanced data.

Since calculations of precision and recall do not make use of the true negatives, precision-recall analysis is useful in cases where there is an imbalance in the samples between the two classes. Figure 3 illustrates the precision-recall that has been obtained on the basis of results produced by weighted entropy based-algorithm. The average precision score is equal to 0.99.

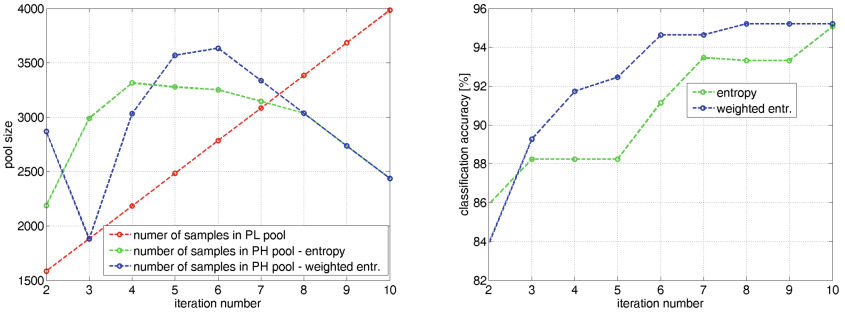

Fig. 2. Number of samples in `PL` and `PH` pool vs. iteration number (left). Classification accuracy vs. iteration number (right). (Color figure online)

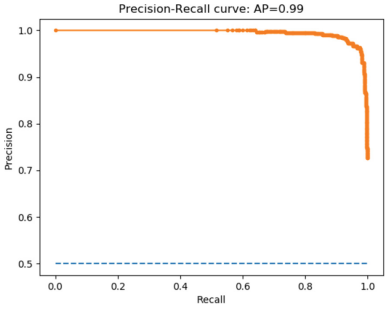

Fig. 3. Precision-recall curve.

The algorithm has been implemented in python language using Keras/ TensorFlow frameworks. The training has been realized on TitanX GPU.

5 Conclusions

The proposed AL-based algorithm for breast cancel classification on unbalanced histopathological datasets considerably reduces the label effort from pathologists, without significantly sacrificing the accuracy. Thanks to the use of weighted cross-entropy in the loss function during training the CNN, as well as weighted entropy both for selecting uncertain samples and determining high confidence samples the proposed algorithms achieves high classification accuracy, high average precision score as well as precision-recall tradeoff, and it is less biased towards the majority class.

Acknowledgments. This research was co-financed by the European Regional Development Fund in the Intelligent Development 2014-2020 Programme, within the grant "The system of automatic analysis and recognition of histopathological images" supported by the National Center for Research and Development: grant no. POIR.01.01.01-00-0861/16-00, and Diagnostyka Consilio.

References

1. Chen, J.M., et al.: Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: a review. Tumor Biol. **39**(3) (2017)
2. Veta, M., Pluim, J., van Diest, P., Viergever, M.: Breast cancer histopathology image analysis: a review. IEEE Trans. Biomed. Eng. **61**(5), 1400–1411 (2014)
3. Filipczuk, P., Fevens, T., Krzyzak, A., Monczak, R.: Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. IEEE Trans. Med. Imaging **32**(12), 2169–2178 (2013)
4. Aswathy, M., Jagannath, M.: Detection of breast cancer on digital histopathology images: present status and future possibilities. Inform. Med. Unlocked **8**, 74–79 (2017)
5. McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B., Kovacevic, J.: Automated histology analysis: opportunities for signal processing. IEEE Signal Process. Mag. **32**(1), 78–87 (2015)
6. Gupta, V., Bhavsar, A.: Breast cancer histopathological image classification: is magnification important? July 2017
7. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Comput. Struct. Biotechnol. J. **16**, 34–42 (2018)
8. Spanhol, F., Oliveira, L., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2016)
9. Nalisnik, M., et al.: Interactive phenotyping of large-scale histology imaging data with HistomicsML. Sci. Rep. **7** (2017). Article no. 14588
10. Huang, S., Jin, R., Zhou, Z.: Active learning by querying informative and representative examples. IEEE Trans. PAMI **36**(10), 1936–1949 (2014)
11. Guiasu, S.: Weighted entropy. Rep. Math. Phys. **2**(3), 165–179 (1971)
12. Elmore, J., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA **313**(11), 1122–1132 (2015)
13. Asri, H., Mousannif, H., Moatassime, H.A., Noel, T.: Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. **83**, 1064–1069 (2016)
14. Wolberg, W., Street, W., Mangasarian, O.: Breast Cancer Wisconsin (Diagnostic) Data Set (1993). https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). Accessed 31 Mar 2019
15. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: Breast cancer histopathological image classification using convolutional neural networks, July 2016
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks (2012)
17. Bayramoglu, N., Kannala, J., Heikkilä, J.: Deep learning for magnification independent breast cancer histopathology image classification, December 2016
18. Motlagh, M.H., et al.: Breast cancer histopathological image classification: a deep learning approach. bioRxiv (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, June 2016
20. Koziarski, M., Kwolek, B., Cyganek, B.: Convolutional neural network-based classification of histopathological images affected by data imbalance. In: Bai, X., et al. (eds.) FFER/DLPR-2018. LNCS, vol. 11264, pp. 1–11. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12177-8_1

21. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
22. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE TCSVT **27**(12), 2591–2600 (2017)