An Ensemble of Deep CNNs for Classification of Breast Histopathology Images

P Dhiyya
Department of Electronics & Communication Engineering
Velammal College of Engineering & Technology, Madurai, India.

S Vasuki
Department of Electronics & Communication Engineering
Velammal College of Engineering & Technology, Madurai, India.

Abstract
In recent years, deep learning is growing to be propitious in diagnosis of tumor thereby assisting the radiologist in the process of treatment planning. This paper presents a novel approach for identification and classification of tumor in breast histopathology image by fine tuning the pre-trained deep CNNs. Also an ensemble classification is adopted for an increased performance in accuracy. The proposed method is experimented over BreakHis and ICPR2012 datasets. An accuracy of 96.4% and 97.2% is obtained for BreakHis and ICPR 2012 dataset respectively which outperformed the other state-of-the art models.

Keywords: Breast; histopathological image; deep CNNs; ensemble classification.

I. INTRODUCTION
Breast cancer, a common type of malignancy found mostly in women. It is also identified to be the cause of increased death rate in women population, worldwide according to the statistics [1] and it is also observed that the risk of prognosis in breast cancer increases with age. An increase in death rate, nearly 14%, due to breast cancer is recorded on a time period of 2008 to 2012 [2]. A gold standard for identifying these breast cancers are the histopathology image. Also, an evolving trend is seen in histopathology image analysis. Nowadays, it is possible to digitize the images of histopathology specimen and store them in the form of a digital image. The histopathology tissue patterns when used with computer-aided image analysis, enhances the process of disease classification. This exhibits the development in archiving digitized histological studies. Computer assisted diagnosis (CAD) helps pathologists in making informed decisions by performing detection, diagnosis and prognosis of a disease.

The disease identification in a histopathology image is considered as an image classification problem which is usually overcome by training the classifier with robust feature representation techniques considering features like color, texture and shape. However, this becomes a difficult task when these histopathology images are acquired at different magnification factor.

The computational medical imaging field mostly uses raw images which consist of complex data. Deep Convolutional Neural Networks (CNN) is proven to be successful for unsupervised learning of imaging features of those complex data [3]. The deep convolutional neural network (CNN) performs tasks like feature extraction, feature selection, classification and binds them into a single unit. It can automatically extract discriminating features from the labelled images [4]. Thus, in most image classification problem a deep CNN exhibits an incredible performance. However, CNN requires large amount of data which stands as its limitation. So, to get rid of that, transfer learning is used [5]. The transfer learning approach is a tricky way of initializing the network weights. It is done by initializing the network weights with the learned weights of a CNN trained on another dataset. A CNN trained on the Imagenet classification challenge is generally used for that purpose. In case of mammogram based breast cancer screening, deep learning methods show better classification accuracy when compared to conventional multistep computational imaging methods [6].

In this paper, ensemble of deep convolutional neural networks is used to classify the breast histopathology images as malignant or benign. As deep CNNs have large number of parameters it requires a large dataset for training it. Medical imaging dataset are not large enough to train deep CNNs from scratch. This leads to the exploration of using transfer learning for medical imaging. In transfer learning, knowledge is transferred between small target domains and large source [10]. In case of CNNs, pre-training is done using the source dataset and later target dataset is used for re-training parts of the model. Three popular deep learning architectures namely- LeNet[6], AlexNet[4] and VGGNet-16[7] are fine-tuned to predict the disease class. Based on the predicted class probability maps obtained from the trained classification networks, a majority-voting is applied.

II. RELATED WORK AND MOTIVATION
Usually, the breast tissues are visually examined by the pathologist which is a time consuming and non-trivial process. This problem can be overcome by the process of automating the classification of these histopathology images. Earlier, the above mentioned problem was overcome by using a classifier that is trained with handcrafted features. But in recent years, according to studies, deep CNN models outperform other image classification techniques.

Spanhol [8], used AlexNet, a pre-trained CNN model, to classify breast tumor into benign and malignant classes from the H&E stained histopathology images of BreakHis dataset. For training, random patches were extracted using sliding window mechanism. A classification accuracy of 79.85% was achieved. The CNN introduced by LeCun [6], works exceptionally well for pattern recognition problems like digit classification [12]. However, its performance in histopathology image was considerably less, achieving only 72% accuracy [8]. In many computer vision tasks, very deep CNN models like VGG, ResNet have achieved state-of-the-art results. Fine-tuning of the VGG model, achieved a classification accuracy of 83.5% over H&E stained breast histopathology images.
In this paper, an approach is proposed for combining different CNN architectures for exploring the strengths and reducing the weaknesses of a given architecture [15]. The main goal of the proposed ensemble system is to produce a powerful image classification system that is able to work at its best for breast histopathology image classification problem.

### III. PROPOSED METHODOLOGY

Deep learning using CNN is a multi-layered image classification technique that incorporates spatial context and weight sharing between pixels [9]. An effective representation of the original image is adopted by a CNN to learn the optimal image features for a specific image classification problem. CNNs are stacks of different types of interconnected specialized layers. The weights of these layers are trained using backpropagation algorithm. The problem overfitting is avoided by training CNNs with large number of labelled data. In case of general pattern recognition tasks, a more generalizable and accurate models are produced using a CNNs which achieves the state-of-the-art performance. LeNet [6], the first CNN proposed to classify handwritten digits; AlexNet [4], a deep network designed for image classification are some of the examples.

Ensemble learning is used to develop a stronger classifier, by combining the predicted scores of multiple weak classifiers. The outcome of this approach is a very high performing system that outperforms the single best CNN that is trained on the given dataset. The one setback observed in ensemble of CNNs is the requirement of high computation power due to large size of the network set. Hence, this approach is appropriate only for cases where computation time is not decisive.

In this paper, we used ensemble of three trained convolutional neural network models namely LeNet, AlexNet and VGGNet-16, to identify the lesions in histopathological images of breast cancer.

LeNet [6]: proposed by Yann LeCun. A traditional CNN used for the handwritten character recognition with high accuracy.

AlexNet [4]: proposed by Alex Krizhevsky. Best performed model for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2012 (ILSVRC12). In binary classification of breast histopathology image it achieved 83% accuracy.

VGGNet [7]: Second best performed model in ILSVRC 2014 challenge. Both VGG-16 and VGG-19 are deep models with 16 and 19 wt layers, respectively, are available as pre-trained models. Each model includes 16 CONV/FC layers. The CONV layers use very small (3 3) convolution filters and are extremely homogeneous. After two or three CONV layers a POOL layer is inserted.

In this work 10% of the training images are used as validation images. These validation images are used for deciding the training hyper parameters. Before fine-tuning a pre-trained model, first the softmax layer is removed and node count of last layer is set to number of classes in the dataset. Initially, all the layers are held fixed except the last layer. The initial learning rate of the network is 0.01 for 20 epochs with early stopping having a patience of 5. Then, all the layers are fine-tuned with a learning rate of 0.001 for 100 epochs with early stopping having a patience of 10.

The learning effectiveness of a CNN depends on the availability of large training data. Data augmentation is one effective way to expand training data when necessary and to reduce overfitting during CNN training by artificially expanding the training set using perturbations of individual images [4]. Data augmentation applies transformations and deformations to the labeled data, thus producing new samples as additional training data. A key attribute of the data augmentation process is that the labels remain unchanged after applying the transformations. In this work we perform random data augmentation with horizontal and vertical flipping, rotation in a range of 100, translation of a maximum of five pixels, and scaling in a range of [1,2].

As already mentioned, a large training data is required to enhance the learning effectiveness of a CNN model. An effective method to expand training data and to also reduce overfitting is data augmentation. Here, the training sets are artificially expanded by applying deformations and transformations to the labeled data [4] as a result of which a set of new samples are produced as training data without any change in the labels. In this work, a random data augmentation with vertical and horizontal flipping, scaling in a range of [1,2], rotation in a range of 100 and translation of five pixels (maximum) is adopted.

The training dataset suffers from data imbalance. This problem is tackled by back propagating the weighted loss from the loss layer. Classifier model construction is done by fine-tuning the pre-trained weights of these models separately. Finally, to decide the class labels of the test images the average predicted class probabilities obtained from these trained networks are used. Based on the predicted class probability maps obtained from the trained classification networks, a majority-voting is applied.

Voting based methods operate only on labels, where $S_{i,j}$ is 1 or 0 depending on whether classifier $r$ chooses i, or not, respectively. The ensemble then chooses class $I$ that receives the largest total vote.

$$\sum_{r} R_{r} = 1d_{i}/x = \max_{I=1,\ldots,C}\sum_{r=R_i}R_{r} = 1S_{i,j}$$

Under the condition that the classifier outputs are independent, it can be shown the majority voting combination will always lead to a performance improvement for sufficiently large number of classifiers. If there are a total of $R$ classifiers for a two-class problem, the ensemble decision will be correct if at least $[R/2+1]$ classifiers choose the correct class. Assuming that each classifier has a probability $p$ of making a correct decision then, the ensemble’s probability of making a correct decision has a binomial distribution, specifically, the probability of choosing $k > [R/2+1]$ correct classifiers out of $R$ is,

$$P_{out} = \sum_{k} R_{k} = (R/2)!1(Rk)pk(1-p)^{R-k}$$

Then,

$$P_{out} \rightarrow 1, \text{ as } R \rightarrow \infty \text{ if } p > 0.5$$

$$P_{out} \rightarrow 0, \text{ as } R \rightarrow \infty \text{ if } p < 0.5$$

The requirement of $p > 0.5$ is necessary and sufficient for a two class problem, whereas it is sufficient, but not necessary for multi class problems.
IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets

In this paper, the proposed method is experimented on BreakHis [8], ICPR2012 (http://ludo17.free.fr/mitos_2012/download.html) and ICPR2014 (https://mitos-atypia-14.grand-challenge.org/dataset/) dataset for training and testing.

B. Evaluation on BreakHis dataset

BreakHis Dataset is a very challenging large-scale dataset containing 7909 images with eight sub-classes of breast cancers. It is broadly divided into two classes namely benign and malignant tumors under four different magnification factors: 40X, 100X, 200X, and 400X. Both benign and malignant breast tumors are sorted into different types by pathologists based on the aspect of the tumor cells under microscopes. Hence, the dataset currently contains histopathology images of four distinct types of malignant breast tumors and four types of benign breast tumors as shown in table I. Images are of three channels RGB with a depth of eight bit in each channel, and a size of 700 x 460. Figure 1 shows examples of the breast cancer subclasses.

Table II reports the accuracy of our ensemble models with the different magnification factors of the BreakHis dataset. Compared to the mentioned CNN models, our proposed ensemble model achieved the highest classification accuracy of 96% as average.

C. Evaluation on ICPR2012 dataset

ICPR2012 dataset consists of 50 HPF images acquired by the widely used Apero-XT scanner and stained with H&E as shown in figure 2. Each image is 2084x2084 pixels RGB. A total count of 326 mitotic nuclei, whose centroids are used as ground truth, are manually annotated by an expert pathologist. The number of HPF and mitoses in this dataset is reported in Table II.

Table IV reports the accuracy of our ensemble models of the ICPR2012. Compared to the mentioned CNN models, our proposed ensemble model achieved the highest classification accuracy of 97.2%.

Table I: Image Distribution in BreakHis Dataset

<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass</th>
<th>Magnification factors</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>40x</td>
<td>100x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>444</td>
<td>114</td>
</tr>
<tr>
<td>Benign</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>1014</td>
<td>253</td>
</tr>
<tr>
<td></td>
<td>TA</td>
<td>453</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>569</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>7909</td>
<td>1995</td>
</tr>
</tbody>
</table>

Table II: Accuracy of Ensemble Models for BreakHis Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Magnification Factor</th>
<th>40x</th>
<th>100x</th>
<th>200x</th>
<th>400x</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet (1)</td>
<td></td>
<td>47.3</td>
<td>47.4</td>
<td>46</td>
<td>45.2</td>
<td>46.5</td>
</tr>
<tr>
<td>AlexNet (2)</td>
<td></td>
<td>80.5</td>
<td>74.8</td>
<td>74.5</td>
<td>81.9</td>
<td>77.9</td>
</tr>
<tr>
<td>VGGNet-16 (3)</td>
<td></td>
<td>90.4</td>
<td>87.5</td>
<td>85</td>
<td>88.2</td>
<td>87.7</td>
</tr>
<tr>
<td>Ensemble (1,2)</td>
<td></td>
<td>60.4</td>
<td>58.3</td>
<td>57.4</td>
<td>61.7</td>
<td>59.4</td>
</tr>
<tr>
<td>Ensemble (2,3)</td>
<td></td>
<td>85.7</td>
<td>81.6</td>
<td>81.9</td>
<td>85.7</td>
<td>83.7</td>
</tr>
<tr>
<td>Ensemble (1,3)</td>
<td></td>
<td>81.5</td>
<td>83.2</td>
<td>81.3</td>
<td>81.5</td>
<td>81.8</td>
</tr>
<tr>
<td>Ensemble (1,2,3)</td>
<td></td>
<td>94.5</td>
<td>97.6</td>
<td>96.3</td>
<td>95.7</td>
<td>96</td>
</tr>
</tbody>
</table>

Table III: Image Distribution in ICPR2012 Dataset

<table>
<thead>
<tr>
<th>Dataset (HPFs/mitoses)</th>
<th>ICPR 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>35/226</td>
</tr>
<tr>
<td>Testing data</td>
<td>15/100</td>
</tr>
</tbody>
</table>

Table IV: Accuracy of Ensemble Models for ICPR2012 Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet (1)</td>
<td>68.4</td>
</tr>
<tr>
<td>AlexNet (2)</td>
<td>82.7</td>
</tr>
<tr>
<td>VGGNet-16 (3)</td>
<td>91.6</td>
</tr>
<tr>
<td>Ensemble (1,2)</td>
<td>71.8</td>
</tr>
<tr>
<td>Ensemble (2,3)</td>
<td>89.3</td>
</tr>
<tr>
<td>Ensemble (1,3)</td>
<td>85.7</td>
</tr>
<tr>
<td>Ensemble (1,2,3)</td>
<td>97.2</td>
</tr>
</tbody>
</table>

Figure 1: Examples of histopathological images of breast cancer subclasses from BreakHis dataset.

Figure 2: H&E stained HPF examples from the ICPR2012 dataset.
D. Evaluation on ICPR2014 dataset

ICPR2014 dataset consist of 1200 HPF images stained with H&E from Aperio scanner of resolution 0.2456 μm/pixel as illustrated in figure 3. The ground truth of training data is split into 70% training and 30% testing due to unavailability of ground truth in testing data. This dataset is more challenging than ICMR2012 due to variability of tissue appearance. The number of HPF and mitoses in this dataset is reported in table V.

The average accuracy of the CNN models and ensemble models in BreakHis and ICPR2012 datasets are depicted in figure 3.

Table V reports the accuracy of our ensemble models of the ICPR2014. Compared to the mentioned CNN models, our proposed ensemble model achieved the highest classification accuracy of 89.4%. The average accuracy of the CNN models and ensemble models for BreakHis, ICPR2012 and ICPR2014 are depicted in figure 4.

Table VI reports the accuracy of our ensemble models of the ICPR2012. Compared to the mentioned CNN models, our proposed ensemble model achieved the highest classification accuracy of 89.4%.

V. CONCLUSION

In the context of classification, deep convolutional neural networks (CNNs) have been widely proven in the scientific and industrial community. In this work, we investigated the performance of an ensemble method over deep neural network model on a classification task related to breast cancer detection. The ensemble model proves that deep learning model used in natural images processing can achieve high performance in medical images processing. The most important finding of this work is that this simple ensemble outperforms the best stand-alone CNN. In our case we achieved about 96% of accuracy in the multi-class breast cancer classification task. The performance achieved can be improved if we provide more data using larger datasets.

REFERENCES

[7] K. Simonyan, A. Zisserman, Very Deep Convolutional...


