Sequential Modeling of Deep Features for Breast Cancer Histopathological Image Classification

Vibha Gupta, Arnav Bhavsar
School of Computing & Electrical Engineering, Indian Institute of Technology Mandi, India
gupta85vibha@gmail.com, arnav@iitmandi.ac.in

Abstract

Computerized approaches for automated classification of histopathology images can help in reducing the manual observational workload of pathologists. In recent years, like in other areas, deep networks have also attracted attention for histopathology image analysis. However, existing approaches have paid little attention in exploring multi-layer features for improving the classification. We believe that considering multi-layered features is important as different regions in the images, which are in turn at different magnifications may contain useful discriminative information at different levels of hierarchy. Considering the dependency exists among the layers in deep learning, we propose sequential framework which utilizes multi-layered deep features that are extracted from fine-tuned DenseNet. A decision is made by layer for a sample only if it passes a pre-defined cut-off confidence for that layer otherwise, the sample is passed on to next layers. Various experiments on publicly available BreaKHis dataset, demonstrate the proposed framework yields better performance, in most cases, than typically used highest layer features. We also compare results with the framework where each layer is treated independently. This indicates that low-mid-level features also carry useful discriminative information, when explicitly considered. We also demonstrate an improved performance over various state-of-the-art methods.

1. Introduction

Breast cancer (BC) is one of the most prevalent types of cancer in women and it is also one with the highest mortality of all cancer deaths amongst women worldwide [1] [2]. Historically, there has been a rising trend in breast cancer cases globally in the last half century with the incidence especially increasing in recent years. Delay in diagnosis is one of the major reason for high level of mortality in breast cancer cases. Hence, early detection and correct assessment of breast density, which seems to have correlation with breast cancer development, are of utmost importance in providing better screening, and effective and efficient treatment to increase survival rates.

The manual classification of breast cancer histopathological images is fatigue, expensive and time-consuming. Hence, there is a pressing need of computer-aided diagnosis (CADx) systems to relieve the pathologist's workload so that attention can be focused on the most suspicious cases. The CADx systems also help in overcoming the subjectivity in interpretation to achieve more reliability of the obtained results. Being a second opinion system, the CADx systems reduce the workload of specialists, contributing to both diagnosis efficiency and cost reduction.

Recently, deep learning based solutions yield state-of the-art performance for various applications which include object detection, face and speech recognition, action recognition, semantic segmentation, medical imaging etc. However, there are limitations in adopting deep learning in the medical imaging due to lack of publicly sufficient labeled database. Hence, it makes difficult to train the model from scratch due to over-fitting problems. To overcome the such limitation, many transfer learning-based methods [3] have been proposed in various medical imaging applications. At present, recent works have indicated that the existing deep learning models pre-trained on large data can be transferred to other recognition tasks [4] [5] [6].

The good performance of deep learning (DL) based framework can be attributed to its ability to learn hierarchical level abstraction of input data by encoding input data on different layers. However, considering the relatively recent applications of DL in the medical analysis domains, existing research has not extensively explored the benefits of explicitly considering deep multilayer features to improve the classification performance. Traditionally, such applications only utilize the final layer features for classification.

We believe that more detailed evaluation should be undertaken as different regions in the images may contain useful discriminative information at different levels of hierarchy. Thus, information in different layers can potentially be used to improve the discrimination capability in a classification task. Histopathological images possess much heterogeneity in appearance (see Figure 1). In this context, we consider that such variability (in terms of color and texture variation) can be better captured by considering the representative information from different layers.

Also, a physical indication which supports such a hypothesis is that the variations in histopathology images are often captured at different optical magnification levels, where each magnification can represent different information. The lowest magnification captures the larger region of interest (ROI), while other magnifications captures the zoomed-in view of tissue inside the initial ROI. While, explicitly using different magnifications can also potentially yield varying discriminative information for the classifica-

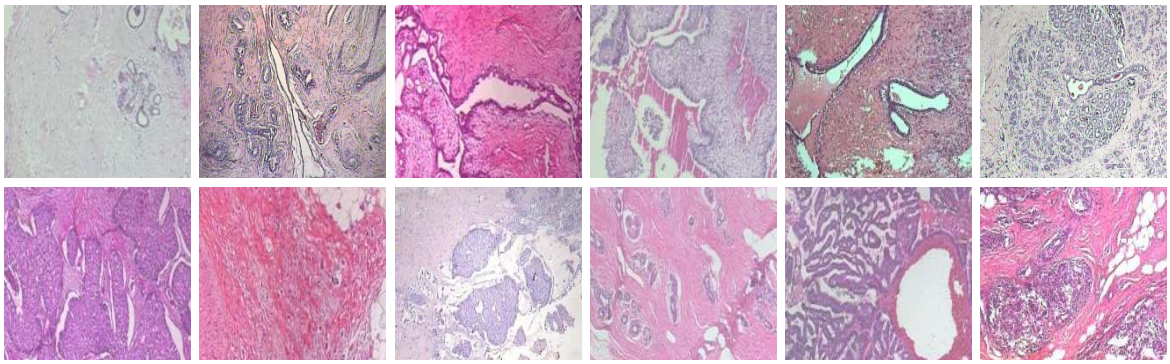

Figure 1. Heterogeneity in histopathological samples(first row: benign tumor, second row: malignant tumor) from BreakHis dataset at magnification factor of 40X.

tion task, as it also suggests such representation at different scales which can also be captured via multi-layered features. Thus, we believe that, such variations in color-texture can be better captured by hierarchy of features that are learnt at different stages of deep learning.

In light of the above discussion, in this work, we have utilized pre-trained CNN (DenseNet after transfer learning) for feature extraction and, pay more attention to exploit the multi-level information, where the representation strength of features from each layer can possibly be different across images. Thus, the hypothesis considers that the low- mid- and high-level information extracted from different layers can all be helpful to improve the discrimination of feature representation, rather than considering only high-level features as is done traditionally.

However, there can be various possible ways to utilize such multi-layerd features. Motivating by the fact that each layer in deep learning is build on top of previous layers, we propose a sequential framework which considers both dependency among layers and their ability to capture heterogeneity in appearance. The proposed framework works in such a way that, in a multi-stage setting, the labels for samples which have a very high classification confidence at a particular stage, will be decided at that stage. Otherwise, the sample will pass on to the next stage. At every step of sequential framework, different features with different cut-off values for the classification confidence are used.

Hence, it is interesting to note that using multi-layered features in proposed framework can prove to be better that just using the high-level features, as is done traditionally, which provides an indication that such deep multilayer features may provide useful information for classification.

Note that models such as RNN and LSTM are also used for sequential processing. However, their architecture is quite different and are used to learn the temporal structure from data containing temporal dependencies (e.g. videos). In this work, there is no temporal dependency in the data (histopathology images), and our premise is to consider the difference in low-mid-high level features due to their sequential dependency in CNN.

The idea of features at multi-level features is also used for localization / segmentation application in the U-net [7]. While the overall task and the classification framework is quite different than that proposed in this work, it still highlights the usefulness of considering mulit-level features.

The main contributions are listed as follows: (1) A framework which uses the multi-layered deep features in a sequential manner for classification of breast cancer histopathology images. (2) This model utilizes the low, mid and high level features obtained by transfer learning, in conjunction with XGBoost classifier. (3) We compare with some competitive frameworks, as well as various state-of-the-art approaches, and clearly demonstrate a positive comparative performance for the proposed method.

2. Related work

In this section, we discuss some previous works, including state-of-the-art methods, aiming to automate the diagnostic procedure in context of breast cancer histopathology. We also discuss various works that have been carried out in the context of transferability of knowledge embedded in the pre-trained deep models (CNNs). Hence, convolutional neural network based frameworks have successfully been applied to analyzing various visual imagery.

Zhang et al. [9] utilized multiple image descriptors along with random subspace ensembles and proposed two-stage cascade framework with a rejection option. In another work [10], an ensembles of one-class classifiers were assessed by the same authors using same dataset. Bahlmann et al. [11], color transformed the RGB patch into two channels, called H and E that intensify the hematoxylin (eosin) at the same time suppressing eosin (hematoxylin) stain. The feature vectors of dimension 22 are extracted and classified using linear classifier to diagnose relevant or irrelevant regions. In [12], approach same as [11], was applied for segmentation and classification. Linder et al. [13] extracted

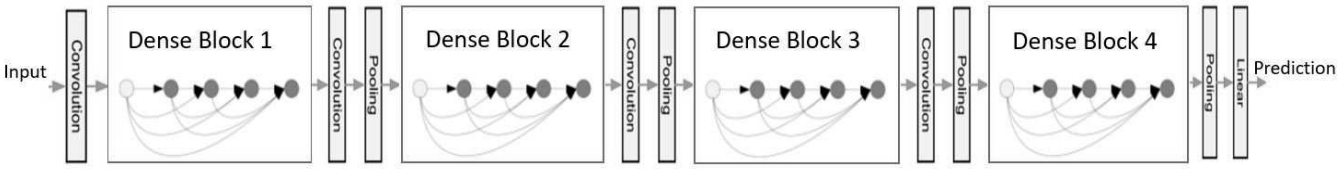

Figure 2. DenseNet-169 architecture [8]

the local binary pattern combined with a contrast measure (LBP/C) and performance evaluated using support vector machine (SVM). However, we note that these methods use an independent dataset (not public). In [14], various experiments that involved the state-of-art texture descriptors such as Local Binary Pattern (LBP), Completed LBP (CLBP), Threshold Adjancey Statistics (PFTAS), Grey-Level Co-occurence Matrix (GLM) etc. and four classifiers are performed and were evaluated at patient level.

In [15], pre-trained Alexnet [16] used for extracting features and classification. Han et al. [17] proposed a multi-classification model to identify subordinate classes of breast cancer (eight classes) using deep learning. They proposed an efficient distance constraint of feature space to formulate the feature space similarities by leveraging intra-class and inter-class labels of breast cancer as prior knowledge in deep learning framework. Song et al. [18] presented a classification model by combining convolution neural network with supervised intra-embedding of Fisher vectors. Song et al. [19] computed image features by FV encoding of CNN-based local features based on the VGG-VD model that is pretrained on ImageNet. In additional, they designed a new adaptation layer to further transform the FV descriptors for higher discriminative space. They perform classification using linear SVM. Color-texture features followed by various contemporaneity classifier are used by [20].

3. Proposed approach

3.1. Feature learning

Here, we discuss the architecture chosen for proposed study, the feature learning which involves feature extraction, and their dimensionality reduction and classification using XGboost.

3.1.1 DenseNet architecture [8]

DenseNet is a network architecture, where within each dense block, layers are directly connected in a feed-forward fashion. The layer is designed in such a way, so that the activation maps of all preceding layers are considered as separate inputs whereas its own activation maps are passed on as inputs to all subsequent layers. DenseNet is mainly composed of a convolution layer, a Dense block, a transition layer, and a classifier after the global average pooling at the input end. In proposed study we utilize DenseNet-169 which consists four dense blocks, and total 169 layers (165-conv+3-transition+1-classification). Figure 2 shows the architecture of DenseNet-169.

We perform transfer learning of pre-trained DenseNet for breast histopathology image classification. We freeze starting 30 layers as they learn generic features, which are common for most of the applications, and retrain remaining layers according to specific application. To fine-tune the DenseNet, we resize the image depending on the input size of the DenseNet. As a common practice, the fully connected layer of the pre-trained network is replaced with a new fully connected layer that has, as many neurons as in the final layer. Here, we add one fully connected layer which subsequently reduces the feature dimension of last fully connected layer (1664-1000-2). During retraining of DenseNet, we add one dropout layer of 0.4 between average pooling and dense layer. Dropout is a regularization technique in which filters are randomly turned off during training which is especially important to avoid, in the case of low training data.

To extract features, we chose all convolution layers of DenseNet. These convolution layers are corresponded to filter size 1*1 and 3*3.

3.1.2 Classification with layer-wise features

The features from different convolutional layers are of very high dimension. Hence, commonly, before using these in other classification frameworks [21] [22], principal component analysis (PCA) is used for dimensionality reduction. In the proposed study, we use XGboost [23] due to its various advantages for dimensional reduction and classification.

XGboost [23]: XGBoost is short for Extreme Gradient Boosting. It is very popular due to its high efficiency and performance. It is an additive tree classification model where each new tree is added to compliment the already built Trees. The final decision is the weighted sum of all predictions made by each individual trees. It is a scalable implementation of gradient boosting machines. It use a regularized model formalization to control over-fitting, which gives it better performance. It also provides score corresponding to each feature, thus enabling the reduction in the feature dimensionality. In the proposed study, we used

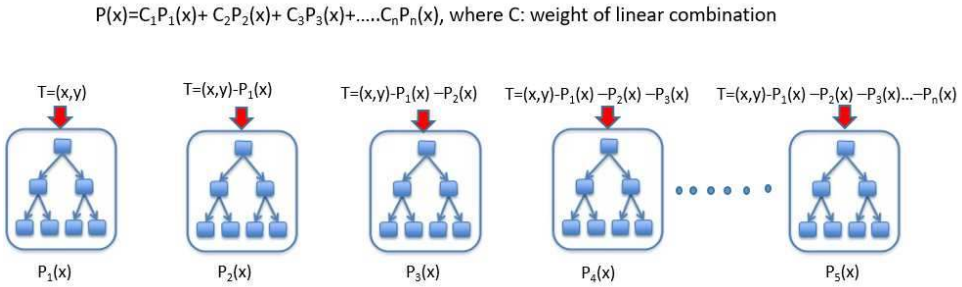

Figure 3. XGboost architecture

score to chose top performing features (<3000). Detailed architecture of XGboost is given in Figure 3. In Xgboost, new trees are added in such a manner that it complement the already-built ones (reduce the residual). Hence reduces the bias and possibly variance compared to base learners. Default values of parameters are used (max depth:2, number of trees: 250) while performing experiment.

3.2. Sequential framework:

This subsection discusses the sequential framework in detail. The pictorial representation of framework is shown in Figure 4. The flow of procedure is described below:

- The classifier at each layer is trained on features extracted from that individual convolution layer of DenseNet.
- For each section, we use a some cut-off for the classification confidence (probability) to decide whether the sample will pass through the next layer or not.
- As the XGboost provides probabilities of classifying samples in given classes (in this case, two classes), we have two values for each samples. To get one score for each sample on which threshold / cut-off would be applied, we calculate difference of probabilities. If the difference is high, such layer has high confidence for such samples, and if it crosses the cut-off for some layer, the decision for sample will be made at that layer itself. Hence, sample will not propagate further. This reduces the confusion at subsequent layers.
- The cut-off point for each section is decided in a such a way that sections corresponding to low-level features have high cut-off and those for high-level features have a low cut-off. The motivation behind fixing such cut-off is that low level feature do not capture any specific pattern, and hence a sample being classified at these layers should be clearly discriminative even with such generic features. To choose cut-off points for 165 layers, we pick suitable first and second value and divide it into 165 equal parts. However, these first and second value is chosen empirically using validation data (for which higher accuracy is produced at low ambiguity).
- For samples which do not have enough confidence to cross level of any layers, decisions are made by two ways:
  - Average of probabilities: Probabilities of samples is calculated from each layer and then average out. The class which has higher probability will be the decision.
  - Maximum Voting: Vote based on the probabilities is made by each layer. The class which obtain higher votes is assigned as final class.

3.3. Baseline DenseNet classification framework

For comparison, we also calculate the baseline accuracy which is produced by pre-trained Densenet. As baseline, features of last fully connected layer is used on the trained neural network.

4. Results & discussion

We now discuss our experimentation, and then provide results for different scenarios, along with various comparisons.

4.1. Experimental protocol

This subsection discusses the dataset, training-testing protocol and evaluation metrics in detail.

4.1.1 Dataset description

We utilize break-His [14] dataset to validate the effectiveness of proposed framework. All the images (7909) in the dataset, collected from 82 different patients out of which 24 for benign and 58 for malignant. Four different magnifications were utilized to capture the images. Each individual class such as benign and malignant has four sub-category. The information about the distribution of images is given in Table 1. Figure 1 shows the samples images of benign and malignant tumor at lowest magnifications.

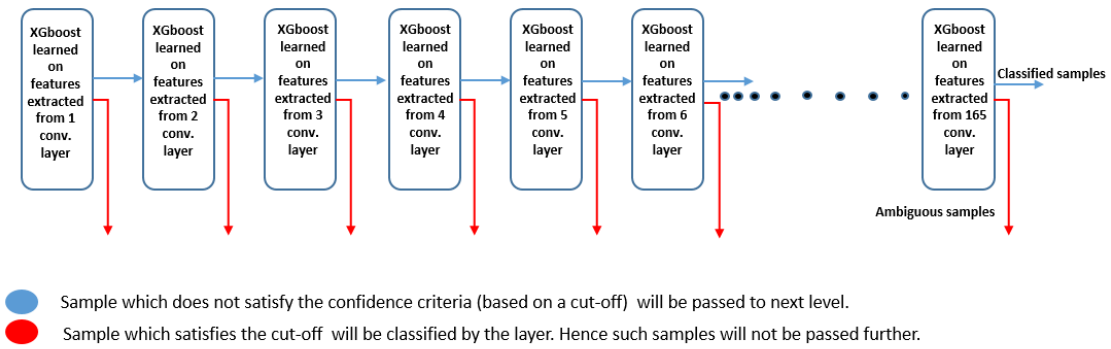

Figure 4. Sequential modeling of multi-layered features.

Table 1. Detailed description of BreaKHis dataset [14].

	Magnifications				Total	Patient
	40x	100x	200x	400x		
Benign	625	644	623	588	2480	24
Malignant	1370	1437	1390	1232	5429	58
Total	**1995**	**2081**	**2013**	**1820**	**7909**	**82**

4.1.2 Training & Testing Protocol

In our experiments, we have randomly chosen 58 patients (70%) for training/validation and remaining 25 for testing (30%). We train the classifiers using images for the chosen 58 patients, and, perform three trial for magnification-specific study. Each trial is a random selection of training-testing data. The trained models across each trial are tested using images of the remaining 25 patients. We also employ data augmentation to increase the data size. For augmentation, we adopt rotation, flip, height shift, width sift and translation. After augmentation, we have six times the original training data.

4.1.3 Evaluation metrics

In this work, Patient recognition rate (PRR) is used as an evaluation metric for the study, to make it comparable with other existing methods. PRR is a ratio of correctly classified images to total images of cancer images. The definition of patient recognition rate is given as follows:

$$PRR = \frac{\sum_{i=1}^{N} PS_i}{N}, PS = \frac{N_{rec}}{N_P} \quad (1)$$

where N is the total number of patients (available for testing). N_{rec} and N_P are the correctly classify images and total cancer images of patient P respectively.

4.2. Sequential framework with deep features

In this subsection we discuss the results obtained using proposed sequential framework under various situations which handle the ambiguous samples.

We take subsections of sequential framework increasing in sets of 10 and calculate accuracy till the final layer in each set. We start making such subsections from the last layers as higher layers are richer in terms of high-level information.. To get more clarity on how accuracy varies when more number of layers are added to framework, we provide the graph in Fig. 5. The vertical line shows an approximate point after which the accuracy does not improve further for most cases. The results corresponding to individual subsections are reported in 2, 3, 4. Table 5 shows the behavior of some random convolution layers and last dense layer. In table 5, L1-L10 is corresponds to convolution layers and L11 represents the fully connected layer.

Table 2 illustrates the results of sequential framework without considering the methods which resolves the ambiguous samples. Each individual entry in table is represented by two values. First value is corresponds to patient score while second represent the percentage of ambiguous samples. The results correspond to the maximum voting and average probability are reported in 3, 4. In all the tables, values are presented in percentage (%).

Sequential framework with average voting: Under this, we discuss the results which obtained in situation where ambiguous samples are classified by calculating average of probabilities assigned by layers and then classified in the class which has larger average probability. The detailed results are illustrated in table 3.

Sequential framework with maximum voting: Under this, we discuss the results which obtained in situation where ambiguous samples are classified by calculating votes given by layers and then classified in the class which has larger number of votes. The detailed results are illustrated in table 4.

Observation based on Tables 2, 3, 4, 5:

- For 40x, the lower level layers contribute significantly to the classification performance.

Table 2. Performance of sequential framework with ambiguity. In the representation x-y, x denotes the performance, and y denotes the ambiguity (%).

Mag.	Layers																	
	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	165	Base-line
40x	92.39-0.53	92.30-0.29	91.66-0.20	91.04-0.09	91.73-0.0	91.40-0.0	92.06-0.0	91.90-0.0	93.22-0.0	92.72-0.0	92.72-0.0	92.56-0.0	92.89-0.0	93.38-0.0	93.38-0.0	**94.71-0.0**	**94.71-0.0**	84.72
100x	**96.37-3.38**	**96.11-2.09**	95.91-1.77	95.49-1.48	95.25-1.42	95.09-1.41	95.21-1.15	94.99-1.08	96.17-1.00	94.93-1.00	94.78-0.94	94.70-0.94	94.65-0.82	94.75 -0.75	94.75-0.75	94.76-0.72	94.76-0.72	89.44
200x	96.35-0.42	96.05-0.10	96.18-0.08	96.07-0.08	96.08-0.08	95.63-0.05	96.06-0.05	95.52-0.0	96.18-0.0	96.10-0.0	96.28-0.0	96.05-0.0	96.30-0.0	**96.76-0.0**	96.38-0.0	96.30-0.0	96.78-0.0	95.65
400x	92.36-3.48	**92.47-2.64**	91.49-2.10	90.86-1.66	90.44-1.41	90.22-1.24	90.17-1.20	90.19-1.17	90.11-1.10	90.03-1.10	89.93-1.09	89.93 -1.09	89.93-1.09	90.13-1.05	90.14-1.04	90.33-1.04	90.33-1.04	82.65

Table 3. Performance of sequential framework with probability averaging (%).

Mag.	Layers																	
	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	165	Base-line
40x	91.57	92.06	91.57	91.07	91.73	91.40	92.06	91.90	93.22	92.72	92.72	92.56	92.89	93.38	93.38	**94.71**	**94.71**	84.72
100x	92.89	93.09	93.09	93.29	92.96	92.98	93.46	93.29	**95.17**	93.39	94.17	93.25	93.26	93.35	93.35	93.27	93.35	89.44
200x	95.85	95.97	96.00	95.98	96.00	95.64	96.06	95.52	96.18	96.10	96.28	96.05	96.30	**96.76**	96.38	96.30	96.78	95.65
400x	88.86	89.74	**89.81**	89.65	89.35	89.13	88.87	88.96	89.14	88.96	88.86	89.06	89.14	89.32	89.41	89.80	89.80	82.65

Table 4. Performance of sequential framework with maximum voting (in %).

Mag.	Layers																	
	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	165	Base-line
40x	91.57	92.06	91.57	91.07	91.73	91.40	92.06	91.90	93.22	92.72	92.72	92.56	92.89	93.38	93.38	**94.71**	**94.71**	84.72
100x	92.63	**95.90**	93.29	93.28	93.05	93.07	93.46	93.13	94.87	93.13	93.41	93.34	93.26	93.35	93.35	93.35	93.35	89.44
200x	95.77	95.97	96.09	95.98	96.00	95.64	96.06	95.52	96.18	96.10	96.28	96.05	96.30	**96.76**	96.38	96.30	96.78	95.65
400x	86.45	87.85	88.54	88.16	87.95	88.67	88.48	88.49	88.81	88.70	88.70	88.70	89.01	89.25	89.25	**89.52**	89.52	82.65

Table 5. Performance of some random convolution layers and fully connected layers (%).

Magnification-specific study											
Mag.	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
40x	76.27	79.40	79.70	88.82	79.99	83.94	91.68	90.85	92.41	93.72	93.49
100x	79.53	89.03	90.49	87.70	88.56	83.30	93.54	93.76	93.98	93.53	93.91
200x	80.32	90.70	93.22	95.02	93.55	92.91	95.71	95.55	95.08	95.30	94.82
400x	72.69	79.02	80.34	88.56	77.00	79.14	87.76	87.81	88.74	87.88	87.85

Table 6. Performance comparison of sequential framework with independent framework (in %).

Frameworks	Mag.			
	40x	100x	200x	400x
Sequential (with averaging)	94.71	93.57	96.76	92.47
Sequential (with maximum voting)	94.71	95.9	96.76	89.11
Independent model (considering all layers)	91.90	93.64	95.84	90.15

- As 400x is more specific in terms of structures, these are better captured by higher level layers.
- For mid-level magnifications i.e. 100x ad 200x all layers contribute similarly.
- Performance using the features from only one convolutional layer is generally worse than the results produced by modeling of deep multi-layer features.
- Lower convolutional layers perform worse than a deeper convolutional layer in most of the cases.
- Deeper convolutional layers show better performance than the fully connected layers (in most cases). It signifies the role of convolution layers to build more discriminative features for classification.
- Information fused into multi-layered network through Xgboost performs better than baseline DenseNet. The improvement in accuracy over baseline signifies the role of low-mid level features together with high level features in multi-layered framework.
- Considering the ambiguity percentage in Table 2, we note that the fraction of such ambiguous samples which do not satisfy the cut-off for any layer, is very small. In any case, even these are further resolved. The ambiguity is calculated in same fashion as patient score is calculated. However, here instead of taking mean of all patient, we do the addition to know the ambiguity score.

4.3. Comparisons across variants of the multi-layered framework

Here, in Table 6, we first compare the framework considering the different approaches for resolution of ambiguities. We note that the performance is complementary for the 100x and 400x magnification, otherwise both strategies seem to perform similarly. Having discussed the sequential framework, it is natural to compare the performance of

Table 7. Performance Comparison of magnification specific system (in %). For the proposed method, the numbers in bracket provide its rank based on the performance among all approaches.

	Methods	Magnifications (values in percentage (%))			
		40x	100x	200x	400x
Existing works	Spanhol et al. [14]	83.8±4.1	82.1±4.9	85.1±3.1	82.3±3.8
	Spanhol et al. [15]	90.0±6.7	88.4±4.8	84.6±4.2	86.10±6.2
	Bayramoglu et al. [24]	83.08±2.08	83.17±3.51	84.63±2.72	82.10±4.42
	Gupta et al. [20]	86.74±2.37	88.56±2.73	90.31±3.76	88.31±3.01
	Song et al. [18]	90.02±3.2	88.9±5.0	86.9±5.2	86.3±7.0
	Song et al. [19]	90.02±3.2	91.2±4.4	87.8±5.3	87.4±7.2
	Han et al. [17]	97.1±1.5	95.7±2.8	96.5±2.1	95.7±2.2
Proposed	Baseline-accuracy (DenseNet)	84.72	89.44	95.65	82.65
	Independent framework	91.90	93.64	95.84	90.15
	Sequential framework with deep multi-layered features (maximum voting)	94.71±.88(2)	95.9±4.2(1)	96.76±1.09(1)	89.11±0.12(2)

the sequential approach with a scenario where each layer treated independently, and a majority voting rule is followed based on the decisions of the XGBoost classifier for such independent features from different layers. Interestingly, we find that, in most cases, the results from the sequential approach are better or similar than those of the approach that considers independent features. It can be observed that performance at 400x for one case of sequential approach is slightly lower than that of the independent framework. The reason could be the less number of images at 400x compared to other magnifications, due to which some of the classifiers operating on multi-layered features may yield somewhat lesser performance.

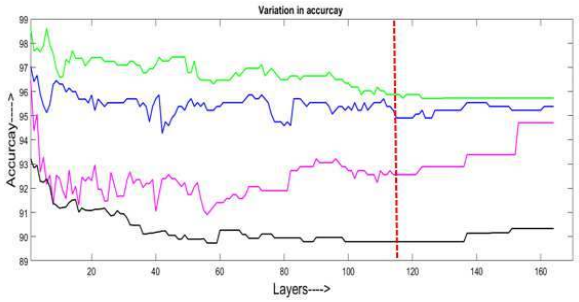

Figure 5. Variation in accuracy with number of layers

4.4. Performance comparison with state-of-art

To validate the effectiveness of proposed framework for magnification-specific study, we compare our results with the state-of-art approaches. The comparative results are provided in Table 7 which depicts the proposed method outperforms all contemporary methods except [17] in some cases. In [17], a class structure-based deep convolution neural network (CSDCNN) is reported which embeds an additional distance constraint. Thus, it non-trivially extends upon an existing deep learning network. Our purpose in this work is to understand the role of multi-layered features from an existing architecture, as it is, with a contemporary classifier. The proposed framework outperforms most state-of-the-art approaches even when fine-tuning an existing model, rather than formulating a new deep architecture. In comparison to [17], the proposed model is quite straightforward. However, it would be interesting to consider an additional distance constraint, to improve the effectiveness of the feature learning.

It is clear that, in order to get higher accuracy for each magnification, the model corresponding to each magnification should be designed individually. It is also intuitive that, due to the variations in textures inherent at different levels of feature representation, discriminative pattern may not be captured by same number of layers for each magnification. Thus, expectedly, the multi-layered sequential modeling yields better results than the others (Table 7).

It can be seen from the second part of Table 7 that information fused in multi-layered sequential network performs better than baseline Densenet. The improvement in accuracy over baseline signifies the role of low-mid level features together with high level features in multi-layered framework. Finally, as indicated earlier, the proposed framework also outperforms the case when considering independent multi-layered features (for most of the cases). In Table 7, along with values, we also provide the rank of proposed approach in terms of the performance when compared with existing frameworks.

5. Conclusion

In this paper we focus on better exploring the potential of fine-tuned pre-trained CNN models in breast cancer histopathology image classification, and presents a sequential model which integrates the features of various layers. Through various experiments, the results demonstrate that the proposed multilayer deep feature fusion in sequential framework indeed outperforms the baseline network, and also the classification using only highest level features. This

indicates that all high-level, mid-level and low-level features can have useful discriminating information, if considered explicitly in an sequential framework. The proposed approach is also shown to outperform most state of-the-art classification methods.

References

[1] American Cancer Society. Breast cancer facts & figures 2011-2012. *American Cancer Society INC.*, 1(34), 2011. 1

[2] Peter Boyle, Bernard Levin, et al. *World cancer report 2008.* IARC Press, International Agency for Research on Cancer, 2008. 1

[3] Hak Gu Kim, Yeoreum Choi, and Yong Man Ro. Modality-bridge transfer learning for medical image classification. *arXiv preprint arXiv:1708.03111*, 2017. 1

[4] Fan Hu, Gui-Song Xia, Jingwen Hu, and Liangpei Zhang. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11):14680–14707, 2015. 1

[5] Keiller Nogueira, Otávio AB Penatti, and Jefersson A dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017. 1

[6] Esam Othman, Yakoub Bazi, Naif Alajlan, Haikel Alhichri, and Farid Melgani. Using convolutional features and a sparse autoencoder for land-use scene classification. *International Journal of Remote Sensing*, 37(10):2149–2167, 2016. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[9] Yungang Zhang, Bailing Zhang, Frans Coenen, and Wenjin Lu. Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles. *Machine vision and applications*, 24(7):1405–1420, 2013. 2

[10] Yungang Zhang, Bailing Zhang, Frans Coenen, Jimin Xiao, and Wenjin Lu. One-class kernel subspace ensemble for medical image classification. *EURASIP Journal on Advances in Signal Processing*, 2014(1):17, 2014. 2

[11] Claus Bahlmann, Amar Patel, Jeffrey Johnson, Jie Ni, Andrei Chekkoury, Parmeshwar Khurd, Ali Kamen, Leo Grady, Elizabeth Krupinski, Anna Graham, et al. Automated detection of diagnostically relevant regions in h&e stained digital pathology slides. In *SPIE Medical Imaging*, pages 831504–831504. International Society for Optics and Photonics, 2012. 2

[12] Eric Cosatto, Matt Miller, Hans Peter Graf, and John S Meyer. Grading nuclear pleomorphism on histological micrographs. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. 2

[13] Nina Linder, Juho Konsti, Riku Turkki, Esa Rahtu, Mikael Lundin, Stig Nordling, Caj Haglund, Timo Ahonen, Matti Pietikäinen, and Johan Lundin. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic pathology*, 7(1):22, 2012. 2

[14] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016. 3, 4, 5, 7

[15] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2560–2567. IEEE, 2016. 3, 7

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[17] Zhongyi Han, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li, and Shuo Li. Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific Reports*, 7, 2017. 3, 7

[18] Yang Song, Ju Jia Zou, Hang Chang, and Weidong Cai. Adapting fisher vectors for histopathology image classification. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 600–603. IEEE, 2017. 3, 7

[19] Yang Song, Hang Chang, Heng Huang, and Weidong Cai. Supervised intra-embedding of fisher vectors for histopathology image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–106. Springer, 2017. 3, 7

[20] Vibha Gupta and Arnav Bhavsar. Breast cancer histopathological image classification: Is magnification important? In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 769–776. IEEE, 2017. 3, 7

[21] Erzhu Li, Junshi Xia, Peijun Du, Cong Lin, and Alim Samat. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017. 3

[22] Hirokatsu Kataoka, Kenji Iwata, and Yutaka Satoh. Feature evaluation of deep convolutional neural networks for object recognition and detection. *arXiv preprint arXiv:1509.07627*, 2015. 3

[23] Tianqi Chen and Carlos Guestrin. Xgboost: reliable large-scale tree boosting system. arxiv. *2016a. ISSN*, pages 0146–4833, 2016. 3

[24] Neslihan Bayramoglu, Juho Kannala, and Janne Heikkilä. Deep learning for magnification independent breast cancer histopathology image classification. 7