Carcinoma type classification from high resolution breast microscopy images using hybrid ensemble of deep convolutional features and gradient boosting trees classifiers

Ritabrata Sanyal*, Devroop Kar, and Ram Sarkar, *Senior Member, IEEE*

Abstract—Breast cancer is one of the main causes behind cancer deaths in women worldwide. Yet, owing to the complexity of the histopathological images and the arduousness of manual analysis task, the entire diagnosis process becomes time consuming and the results are often contingent on pathologist's subjectivity. Thus developing an automated, precise histopatholgical image classification system is crucial. This paper presents a novel hybrid ensemble framework consisting of multiple fine tuned convolutional neural network architectures as supervised feature extractors and eXtreme gradient boosting trees (XGBoost) as top level classifiers, for patch wise classification of the high resolution breast histopathology images. Due to semantic complexity of the patch images, a single CNN network can not always extract high quality features, and the traditional softmax classifier does not provide ideal results for classifying the CNN extracted features. Thus we aim to improve patch wise classification by proposing a hybrid ensemble model which incorporates different discriminating feature representations of the patches, coupled with XGBoost for robust classification. Experimental results show that our proposed method outperforms other existing state of the art methods to the best of our knowledge.

Index Terms—Breast cancer, histopathology, deep learning, convolutional networks, ensemble learning, xgboost

1 Introduction

Breast cancer is one of the leading cancer-related death causes in women worldwide [1]. According to a WHO report in 2013, around half a million women worldwide succumbed to breast cancer in 2011. The mortality rate of breast cancer is very high compared to other types of cancer [2]. Breast cancer is also rising at a dire rate in India, causing the most common cancer and the second most common disease among women in urban and rural areas respectively. The women of a lower age (25-40) group are generally more susceptible to breast cancer. According to Surakasula et al. [3], the risk of breast cancer increases precariously until menopause, then decreases gradually. Breast cancer diagnosis consists of a screening test like mammography or ultrasound whenever a lump is detected, followed by a biopsy and histopathological examination to make a definite diagnosis of any malignant growth in the breast tissue. The tissues can either be normal, benign (non-malignant) or malignant. Malignant lesions can be classified as in situ, where the cells are restrained inside the mammary ductal-lobular system or invasive, where the cells spread beyond that structure. Breast cancer can be fatal if it's detected late, yet early detection can reduce the mortality rate substantially as more and more early stage treatment options are now becoming available. So proper differentiation between normal, benign, malignant lesions is critical. Currently the method of histopathological examination of biopsy slides is based on manual qualitative analysis by pathologists. But this manual inspection technique is fraught with difficulties. The process is time consuming and the accuracy of the diagnosis is commensurate to the pathologist's professional expertise and diagnostic experience. The pathologist's subjective interpretation has often lead to diagnostic inconsistencies. Moreover owing to the complexity of histopathological images and the arduous manual analysis task, the pathologists are often prone to fatigue and inattentiveness which can also hamper the diagnosis. The average diagnostic concordance between pathologists is around 75% [4]. Thus this subjectivity of diagnostic interpretation necessitates the use of CAD (Computer Aided Diagnosis) systems for automated histopathology image classification. The primary task of automated diagnosis of breast cancer entail in classifying histopathology images into four classes namenly normal, benign, insitu carcinoma or invasive carcinoma. Traditionally this was done by constructing handcrafted features based on nuclei segmentation, mitosis detection to name a few, from the histopathology image and classifying them with traditional machine learning classifiers. But recently with the advent of deep learning and especially convolutional neural networks in computer vision, researchers have started to use these ideas and techniques for breast histopathology image classification and have achieved state of the art results. In this paper, we have reviewed some of the state of the art

- *Ritabrata Sanyal is associated with the Department of Computer Science and Engineering, Kalyani Government Engineering College. E-mail: sritabrata@gmail.com*
- *Devroop Kar is associated with Department of Computer Science and Engineering, Jadavpur University. E-mail: kardevroop@gmail.com*
- *Ram Sarkar is associated with Department of Computer Science and Engineering, Jadavpur University. E-mail: rsarkar@ieee.org*
- **Corresponding author.*

methods, and to address their challenges, we have proposed a novel hybrid ensemble framework constituting of multiple deep convolutional networks and XGBoost classifiers for breast histopathology image classification.

1.1 Related Works

Automatic breast cancer detection from histopathological images has been a widely researched area for many years. The first published work on image processing based breast cancer detection dates back to more than 40 years [5]. Yet this topic is still extensively studied due to the complexity of images that need to be analyzed. The current state-of-the-art can be broadly categorized into two most common approaches for designing image based recognition systems: i) using visual feature descriptors or "handcrafted features", and ii) deep learning based methods using convolutional neural networks (CNNs).

The traditional approach involves using handcrafted features for segmentation of nuclei and cells from the breast histology image for extracting discernible features to distinguish between malignant and non-malignant tissues. These handcrafted features revolve around using active contours, thresholding, graph cuts, watershed segmentation, pixel wise classification/clustering or a combination of these. Most of these works primarily focused on a 2-class classification task of malignant or non-malignant([6], [7], [8], [9], [10], [11]). Some others studied a more complex 3-class problem of normal, insitu and invasive carcinoma([12], [13]). But these handcrafted feature engineering techniques suffer from some major drawbacks. Most of these methods were carried out on low resolution histopathology images at different magnifications, thus could not generalize properly on high resolution images. In addition, these techniques require a lot of domain knowledge and it is highly arduous to come up with good discriminating high quality features which limits the efficacy of the classifiers. Also, the extracted handcrafted features may have some biases to the dataset being used, hence the approaches can not generalize well on diverse histopathology image datasets.

Later, with the rise in available computing power, deep learning based systems powered with CNNs, started gaining traction in the domain of histopathology imaging and achieved state of the art results ([14], [15], [16]). Contrary to handcrafted manual feature engineering, CNNs automatically learn high quality features from the training data by optimization of a loss function. One of the first notable works based on this approach is by Spanhol et al. [17]. The authors released a dataset called BreakHis [18] consisting of microscopic histopathological breast images captured at different magnifications. They used the AlexNet architecture for classifying images as benign or malignant, and reported a 6% increase in classification accuracy than other traditional methods, yet their classification performance degraded with increasing magnification. Bayramoglu et al. [19] proposed a CNN based magnification independent approach on the BreakHis dataset. All these approaches discussed till now studies either a 2-class or 3-class problem. Araujo et al. [20] first considered a 4-class classification problem on the Bioimaging 2015 dataset [20] which consists of high resolution (2048×1536) breast histopathology images captured at same magnification, with 4 classes *viz.* benign, insitu, invasive, normal. The authors used a network similar to AlexNet [21].

In the past year, several CNN based methods were proposed for automatic breast histology image classification for the ICIAR 2018 challenge [22]. The dataset provided for this challenge was an extension of the Bioimaging 2015 dataset. The key methods presented in this challenge are : at first the high resolution images are preprocessed and enhanced, then they are divided into patches and each patch is given as input to a CNN for classification or feature extraction. The classification label of an image is determined by fusing the posterior probabilities of the patches by majority voting or the extracted features and classifying it with a standard machine learning classifier like SVM. Golatkar et al. [23] proposed a nuclei based patch extraction strategy and fine-tuned a pretrained Inception-v3 network to classify the patches. They used majority voting on the patch predictions to classify the images. Rakhlin et al. [24] proposed a transfer learning approach without fine tuning, based on extracting deep convolutional features from pretrained networks. They used pretrained CNNs to encode the patches to obtain sparse feature descriptors of low dimensionality, which were trained using a LightGBM classifier to classify the histopathology image. Vang et al. [25] fine tuned a pretrained Inception-v3 network on the extracted patches. The patch wise prediction was ensembled using majority voting, logistic regression and gradient boosting trees to obtain image wise prediction. Cao et al. [26] extrcated feature descriptors from multiple pretrained CNN architectures without fine tuning and trained the descriptors using a random forest dissimilarity based learning framework. Awan et al. [27] finetuned ResNet network for patch wise classification. Then the trained ResNet model was used to extract deep feature representations of the patches. After that, the flattened features of 2×2 overlapping blocks of patches were used to train a SVM classifier, followed by a majority voting scheme for image wise classification.

From the above discussion it can be said that the classification of breast histopathology images using deep learning broadly fall in any of the two categories : i) finetuning a pretrained CNN network for patch wise classification followed by fusion of posteriors/feature vectors of the patches for image wise classification, ii) using transfer learning without finetuning to obtain patch encodings from multiple pretrained CNNs which are again trained using standard machine learning classifiers for image wise classification. The drawback of i) is that owing to the semantic complexity of a patch, fine tuning only a single CNN does not generate good discriminative feature representation, which could have been got if multiple CNNs were used to generate features. The drawback of ii) is that since the CNNs are not fine tuned, they can not learn patch specific features. Hence the features produced are of low quality, which does not give ideal results when they are fed into traditional machine learning classifiers for image level classification. In this paper, we address these drawbacks, by bridging the gap between the two key methods. We observe that image wise classification performance is directly related to patch wise performance, which is in fact quite intuitive. With this line of thought, we aim to boost patch wise classification accuracy

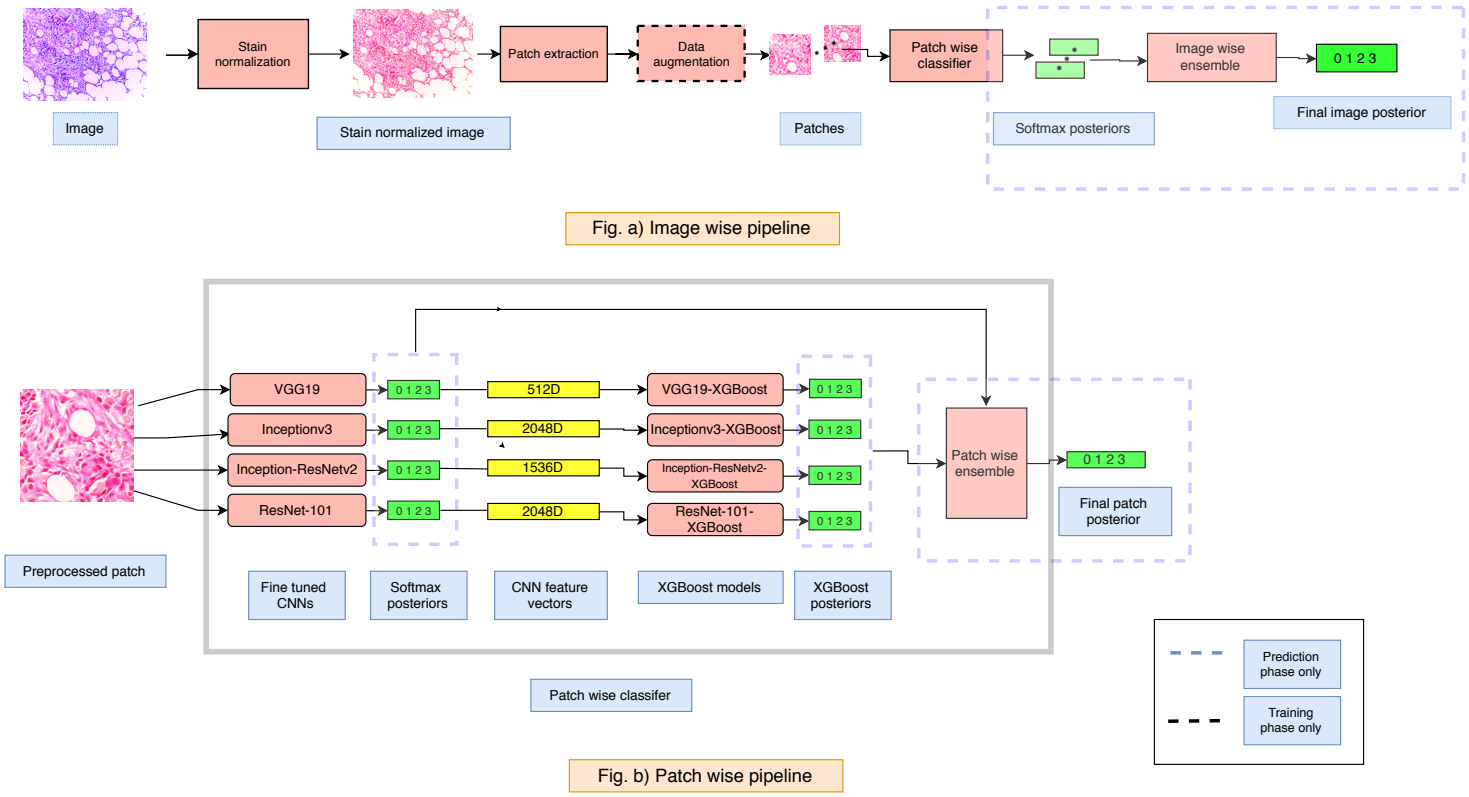

Fig. 1: Workflow of entire process

as much as possible, followed by a simple probabilistic fusion technique to predict the class of the entire image. This would obviate the need for using a machine classifier (like SVM) for image wise classification. We hypothesize that our approach can achieve better classification performance than the previous techniques if we fine tune multiple pretrained CNNs for patch wise classification since it helps us to incorporate complementary information and enables the CNNs to learn target task specific features. We also observe that for classification of extracted features, XGBoost is more robust and perform better than the traditional softmax classifier. Hence XGBoost is used alongside softmax for classification and their predictions are then ensembled to predict the class of a patch. It is to be noted that no previous work used XGBoost as a patch level classifier.

Rest of the paper is organized as follows. Section 2 describes the model architecture and the methodologies employed in this paper in detail. In section 3 we discuss the experimental results and compare our model with the existing state-of-the-art. Finally the paper is concluded in section 4.

2 Methodologies

2.1 Schematic representation

Figure 1 depicts the high level workflow of the entire pipeline. The intricacies of each component of this described system are going to be expounded in the subsequent sections.

2.2 Dataset Overview

The dataset used in this work is provided by the BACH Grand Challenge [22] which was organized as a part of ICIAR 2018 conference (15th International Conference on Image Analysis and Recognition). The dataset consists of 400 Hematoxylin and Eosin (H&E) stained breast histology microscopy images each having dimensions of 2048 x 1536. All the images have been acquired under the same conditions, with magnification of 200x and a pixel size of 0.42 μm 0.42 μm. They have been annotated into 4 different classes viz Benign, Normal, Insitu carcinoma and Invasive carcinoma, where each class label is indicative of the type of cancer present in the image Figure 2. The dataset has a balanced distribution of images, with 100 images per class. The image level annotations are provided by two medical experts. The goal of this challenge was to automatically classify each input image.

2.3 Preprocessing

Preprocessing the histology images is a critical step for development of an automated system for histopathological image analysis [28]. To examine histopathology slides, the contrast between different histology structures especially the nuclei and cytoplasm is enhanced by using stains, which facilitate their manual inspection under a microscope. The most commonly used stains in histopathological slides is the hematoxylin and eosin (H&E) stain. Hematoxylin colors the nuclei with a dark purplish hue whereas eosin gives a light pink color to the cytoplasm. H&E stained images are susceptible to unwanted color variations since different slide

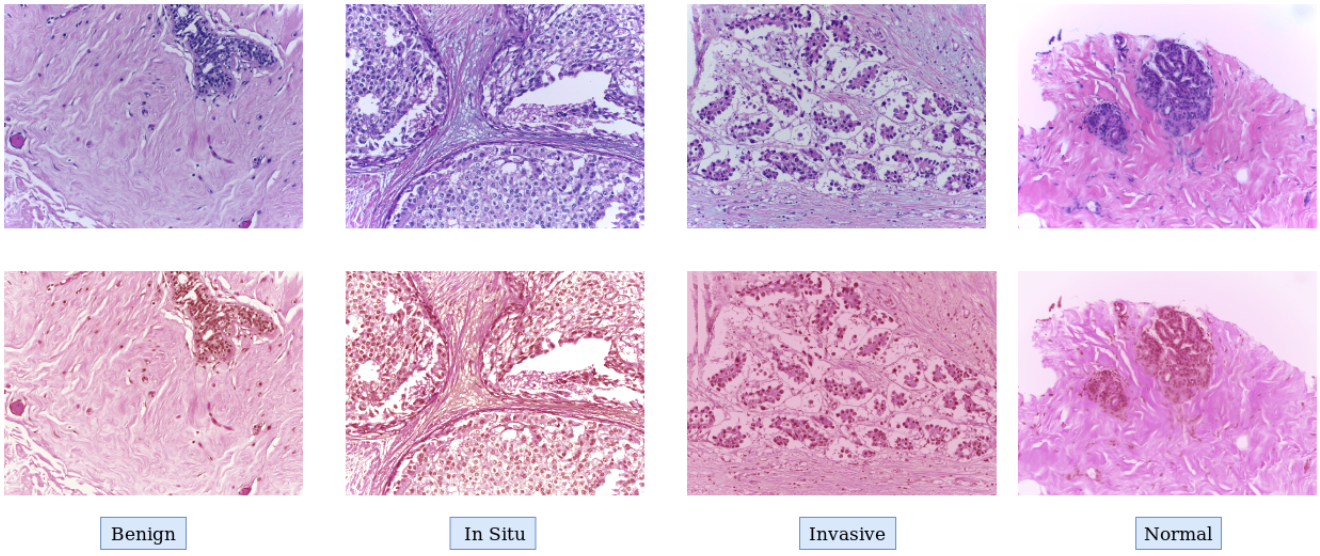

Fig. 2: Sample high resolution images from the ICIAR-2018 dataset are shown in the first row, and their corresponding stain normalized versions are shown in the second row.

scanners respond differently to various colors, and slides prepared using different stain concentrations absorbs different amounts of lights resulting in a variation of appearance. Also various other factors like differences in lab protocols, source manufacturers and staining time [29] can lead to these undesirable variations. Owing to this, a model trained on a particular stain appearance can not generalize well for unseen samples [30]. Thus to reduce such variations, the images must be standardized before feeding them to a model. Stain color normalization is a technique to this end. In our work Vahadane normalization [31], a stain normalization technique based on sparse non-negative matrix factorization (SNMF) is used. The original images and the ones obtained after normalization are shown in Figure 2.

2.4 Patch extraction and Data augmentation

The competence of a CNN network to generalize well on unseen data depends on the number of available training samples. The CNN can learn more discriminating features among various classes if number of training samples available per class is more [32]. The paucity of images in the dataset (400 images for 4 classes) poses a considerable challenge for training deep neural networks [33]. Also, the fact that the images are of high resolution (2048 x 1536), exacerbates the problem [33], because the network can not learn all the features needed to discriminate among images with such large spatial dimensions, from a limited training set. Hence it fails to generalize well on unseen data, making the network prone to overfitting. To address this issue, we divide the high resolution input image into smaller patches. Then we augment the extracted patches to artificially increase the size of training set to prevent overfitting [21].

We extract patches with 50% overlap having spatial dimensions of 512 x 512, from the high resolution (2048 x 1536) input image in a sliding window fashion, described as follows. A window of dimension $k \times k$ with a stride of s is slid over the entire image with width of I_w and height of I_h. The total number of patches extracted will follow from the equation :

$$n_p = \left[1 + \frac{I_w - k}{s}\right] \times \left[1 + \frac{I_h - k}{s}\right] \quad (1)$$

where $I_w = 2048$, $I_h = 1536$, stride $s = 256$, $k = 512$. Thus total number of patches extracted per image, $n_p = 35$. We experiment with patches extracted with 50% overlap ($s = 256$) and with no overlap ($s = 512$) and study the effect they have on the test accuracy of the CNNs. As we can see from Table 1, the former performs better, hence 50% overlap strategy is chosen in our study. A detailed diagnostic information of nuclei and their surrounding tissue structure is critical for classifying histopathological images [34]. Thus, we avoid extracting smaller patches of dimensions 64 x 64 or 128 x 128 because it is very unlikely that they will capture sufficient information so as to distinguish between the different types of carcinoma. This can be also seen from Table 1, where 512×512 patches give the highest test accuracy, compared to patches of lower dimension.

Data augmentation forms a crucial component of any deep learning pipeline since it helps us to prevent overfitting by boosting the size of the training set [21]. This is done by generating artificial training samples by domain specific transformations. We exploit the rotational invariance property of medical images to expand the size of our training set. An extracted patch is rotated in the X-Y plane with variations of $30°$. The vertical axis passing through the centre of an augmented patch subtends an angle of $\frac{k\pi}{6}, k \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ with that of the original patch. One extracted patch results in 11 augmented patches and thus we have a total of 420 augmented 512×512 patches for a 2048×1536 whole slide image.

The features learnt by a deep CNN when trained on a large dataset are transferable to various tasks from different domains [35]. This technique called transfer learning leverages the generic features learnt from natural images that are fundamental to all images and optimises them for the vari-

Patch Size	Test Accuracy(%)			
	VGG19	*Inception-v3*	*ResNet-101*	*Inception-Resnet-v2*
512 x 512 (50% overlap)	**80**	**76**	**78.50**	**79.20**
512 x 512 (no overlap)	77.30	74.50	76	78.40
256 x 256	78.10	73.50	74.5	76
128 x 128	74.80	72	71.40	74
64 x 64	73	71	70.5	70

TABLE 1: Comparative analysis of patch size and test accuracy (patch level) of all the CNNs. The highlighted patch extraction strategy gives the best test accuracy

ety of target domain tasks. With such an approach a CNN is first trained on a large annotated dataset of natural images. The pretrained weights are used to initialise the network and are constantly updated by backpropagation in the fine tuning stage. This enables the network to learn target task specific features, yet with considerably less training data than it would be needed if it was trained from scratch, that is with randomly initialized weights. This is because if the CNN is initialized with random weights and trained, it will need to learn both low level features like edges, corners, curves, local intensity variations etc and high level features (features hierarchically learnt from low level features), to construct a coherent meaning of the data. This demands a lot of labelled training data by the CNN to generalize well on unseen samples. On the other hand, a CNN initialized with weights pretrained on the large dataset of natural images, can already detect these low level features common to any image. So during finetuning a CNN on our breast histology dataset, it does not need to learn every low level feature and just learn the target specific high level features. If a CNN is trained from scratch and number of training samples is less, it will overfit and classification performance will degrade. Thus fine tuning is a very promising approach in the medical imaging domain where deep CNNs can not be trained from scratch due to paucity of extensive labelled datasets. Recent studies([36], [37], [38], [39]) have demonstrated the efficacy of such a fine tuning approach for a variety of medical image classification tasks.

In this study, we first fine-tune five different state-of-the-art CNN architectures that had been pretrained on the ImageNet dataset. The CNNs are fine-tuned on the augmented patch datatset. Each fine-tuned CNN is used in two ways : 1) supervised feature extraction from a patch and using the extracted features to train a Extreme Gradient Boosting (XgBoost) classifier. 2) to output softmax posteriors. All the classifiers are ensembled and the posterior probability of the ensemble model is used to determine the class of the patch. We justify the use of five different CNNs by hypothesizing that due to their structural disparity, they will extract different discriminative features of semantic image representation. Thus their ensemble will enable a higher quality of rich features to be extracted, thereby achieving better classification performance than those achieved by the individual networks. Now we're going to discuss these methods in detail :

2.4.1 Supervised Feature Extraction

2.5 Patch-wise classification

We fine-tune the state of the art CNN architectures namely VGG19 [40], Inceptionv3 [41], Inception-ResNetv2 [42] and ResNet-101 [43], for extracting rich feature representations of a patch. These networks are used because they have achieved state of the art results on large image recognition [44], medical imaging and a variety of other tasks [37], [45]. The fine tuning strategy is same for all the CNN architectures used. A 512×512 patch is given as input to the CNNs. The top fully connected networks of the CNNs are removed and replaced by a Global Average Pooling (GAP) layer [46] to allow the networks to consume images of any arbitrary dimensions. A GAP layer reduces the dimensionality of the feature map from the last convolution block and outputs a fixed size feature vector of length equal to the depth (number of channels) of the feature map. A GAP layer reduces the number of parameters of the network and thus helps in controlling overfitting.. A fully connected layer with softmax activation is then installed on top of the GAP layer to output posterior probabilities of the classes. The loss function of the CNN is a cross-entropy loss which is defined by :

$$L = -\sum_{j=1}^{k} y_j log(\hat{y}_j) \tag{2}$$

where k is the number of classes which in our case is equal to 4 (benign, normal, insitu and invasive), y_j is the j^{th} value of the ground truth class of the image, the patch is a constituent of and $\hat{y}_j$ is the j^{th} value of softmax posterior.

$$\hat{y}_j = g_s(l_j) \tag{3}$$

where l_j is the j^{th} logit value of the CNN and $g_s(.)$ is the softmax activation function. Finally the decision $\hat{D}$ of the CNN is obtained as:

$$\hat{D} = \arg\max_j \hat{y}_j \tag{4}$$

All the CNNs are initialised with weights pretrained on the ImageNet dataset. All the layers are fine tuned for 100 epochs using Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.0001 and a batch size of 8. Dropout regularization is used for decreasing overfitting. After fine tuning the four CNN networks, the feature vector output of the GAP layer is used as the feature representation of a patch.

More formally, let $M : \Omega \rightarrow R^+$ represent a patch where $\Omega = \{(x, y)|x, y \in Z\}$ represent the patch domain and (x, y) is the spatial coordinate of a pixel in the patch domain Ω. A fine-tuned CNN can be defined by the mapping $C_i : R^{W \times H \times D} \rightarrow R^{d_i}$ which takes in a patch $M \in R^{W \times H \times D}$ as input and outputs a feature vector $v_i \in R^{d_i}$. W, H, D are the width, height and number of color channels of the patch respectively. d_i is the dimension of the output feature vector. Let V be the set of feature vectors extracted by the CNNs .Then,

$$v_i = C_i(M) \tag{5}$$

$$V = \{v_i; i = 1, 2, 3, 4\} \tag{6}$$

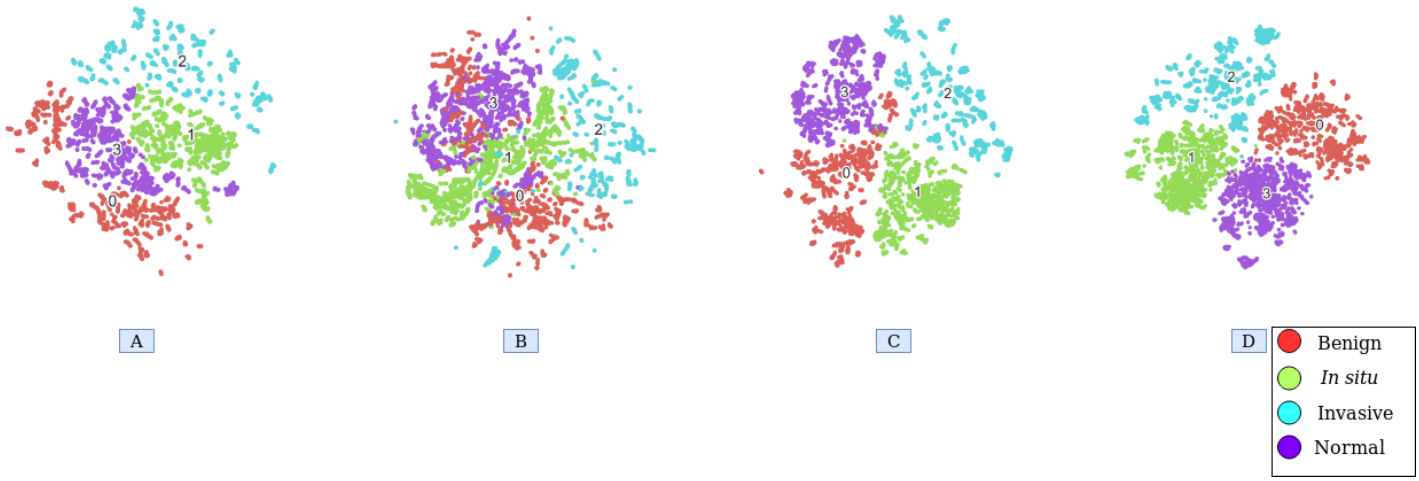

Fig. 3: 2-D projections of the GAP layer feature vector output of the training patches by the 4 fine tuned CNN feature extractors using t-SNE. A) ResNet-101;B) Inception-v3;C) Inception-ResNet-v2;D) VGG-19

where $i = 1$ for VGG19, $i = 2$ for Inception-v3, $i = 3$ for ResNet-101, $i = 4$ for Inception-Resnet-v2. $d_1 = 512$, $d_2 = 2048$, $d_3 = 2048$, $d_4 = 1536$. $W, H = 512$, $D = 3$.

2.5.1 Classification

After supervised feature extraction from a patch, we use the extracted feature representation to classify the patch. We use two types of classifiers in this study:

- **Softmax** : The softmax function maps the non-normalized output of a neural network to a probability distribution over predicted output classes. When a N-dimensional feature vector is given as input, the output of softmax function is a vector of length N representing the posterior probabilities, where the n^{th} element is the likelihood that the vector represents a member of the n^{th} class [47]. The softmax equation is given as:

$$Pr(y = j|\mathbf{x}) = \frac{e^{(\mathbf{w}_j^T \mathbf{x} + b_j)}}{\sum_{k=1}^{K} e^{(\mathbf{w}_k^T \mathbf{x} + b_k)}} \tag{7}$$

- **XgBoost** : We also use a XgBoost or eXtreme Gradient Boosting machine [48] as a patch level classifier on top of the feature vector representation extracted by the CNN. XgBoost is a scalable tree boosting algorithm, which ensembles several weak classifiers into a strong one and has been achieved state of the art performance in various domains([49], [50], [51], [52]). It is more robust than the traditional softmax, for classifying CNN extracted feature representations, as demonstrated in ([53], [54]). Due to these reasons, we hypothesize that XgBoost will achieve better patch classification performance than softmax classifier. We also adduce empirical evidence to justify our hypothesis.

At prediction time, we use a hybrid ensemble strategy. All the softmax and XgBoost posteriors are ensembled to ascertain the class of the patch. We were inspired by this hybrid strategy from [54], where the authors demonstrate that at prediction phase, incorporating both the softmax and XGBoost posteriors in the ensemble boosts the classification performance than ensembling only the XGBoost or softmax posteriors.

More formally, let a $Softmax$ and a $XGBoost$ classifier be represented by the mappings $\alpha_i : R^{d_i} \rightarrow R^n$ and $\beta i : R^{d_i} \rightarrow R^n$ respectively. They take in the feature vector $v_i \in R^{d_i}$ extracted by CNN C_i as input and output posterior probability vectors $p_i^{\alpha} \in R^n$ and $p_i^{\beta} \in R^n$ respectively. For any patch M, if P represents the set of posteriors of the classifiers, $P = \{p_i^{\alpha}, p_i^{\beta}; i = 1, 2, 3, 4\}$. The final posterior probability q_{pa} and class c_{pa} of the patch is determined by the ensembler function Ψ as follows.

$$q_{pa} = \Psi(P) \tag{8}$$

$$c_{pa} = \arg\max_j q_{paj} \tag{9}$$

where $j = \{1, 2, 3, .., n\}$, number of classes $n = 4$. The ensembler function Ψ is described in next section.

2.6 Image Level Classification

To classify an image, we first classify all the patches as described in Section 2.4. Then all the patch wise posteriors are ensembled to get the image level class label. We follow this approach because patches are trained with their image level class labels as ground truth classes. So during prediction, the output posteriors of the constituent patches serve as a good degree of support of the image level class. More formally, let $I : \Phi \rightarrow R^+$ represent an input preprocessed image where $\Phi = \{(x, y)|x, y \in Z\}$ represent the image domain and (x, y) is the spatial coordinate of a pixel in the patch domain Φ. Let $S = \{M_i; i = 1, 2, 3..., n_p\}$ be the set of patches extracted from the image I. Number of patches $n_p = 35$. Let $Q = \{q_{pai}; i = 1, 2, 3, ..., n_p\}$ be the set of posteriors of all the patches in S. The posterior probability q_{im} and class

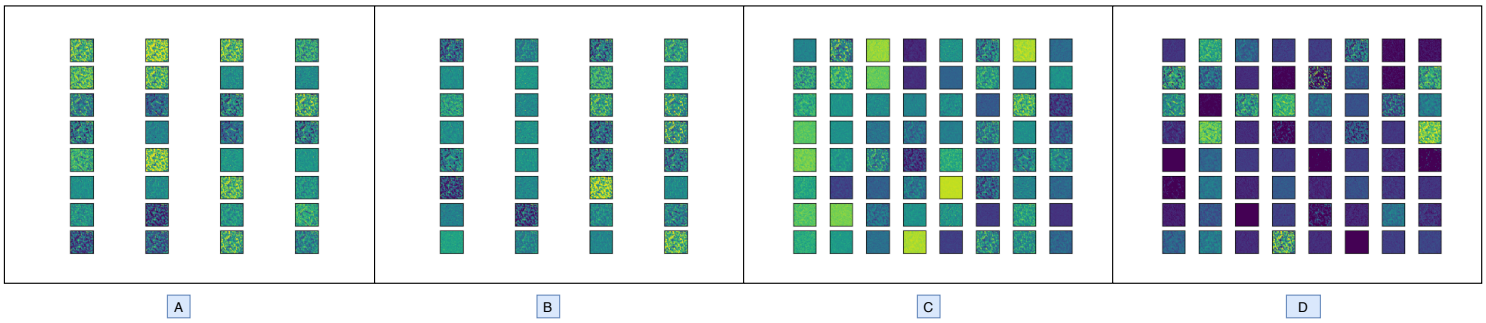

Fig. 4: Filtered images obtained after the first convolutional layer of the four fine tuned CNNs. A) Inception-v3; B) Inception-ResNet-v2; C) ResNet-101; D) VGG19.

c_{im} of the image I is determined by the ensembler function Ψ as follows:

$$q_{im} = \Psi(Q) \tag{10}$$

$$c_{im} = \arg\max_j q_{imj} \tag{11}$$

where $j = \{1, 2, 3, .., n\}$, number of classes $n = 4$. The ensembler function Ψ is described in next section.

2.7 Ensemble Strategy

If P is a set of m posterior probability vectors with n classes, it can be represented as :

$$P = \{p_{ij} \in R^+ \mid 0 \leqslant p_{ij} < 1 \wedge \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = m\} \tag{12}$$

If $\Psi : R^{m\times n} \rightarrow R^n$ be the ensembler function which combines the m posterior probabilities to generate a posterior probability vector $q \in R^n$, representing confidence scores of the $n = 4$ classes, then :

$$q = \Psi(P) = \frac{r(P)}{\eta} \tag{13}$$

$$\eta = \sum_{j=1}^{n} r_j(P) \tag{14}$$

where $r : R^{m\times n} \rightarrow R^n$ is defined as the rule function which can be either be mean, product, majority voting etc. η is the normalizing factor to ensure the sum of confidence scores in posterior q equals 1. The final classification label c is determined by :

$$c = \arg\max_j q_j \tag{15}$$

For patch wise classification, P is the set of posteriors generated by the softmax and XgBoost classifiers, hence $m = 8$ (four CNNs and four XGBoosts). For image level classification, P is the set of posteriors generated by the 35 constituent patches, hence $m = 35$. Number of classes, $n = 4$ in both cases.

In this study, we experiment with 5 rule functions $r(P)$, namely sum rule, where $r(P) = \frac{1}{m}\sum_{i=1}^{m} P_i$, product rule, where $r(P) = \frac{1}{m}\prod_{i=1}^{m} P_i$, their weighted versions, and finally the majority voting rule, where $r(P) = \sum_{i=1}^{m} I(\arg\max_i(P_i) = i)$ (where $I(.)$ is the indicator function: $I(y) = 1$ if y is true and $I(y) = 0$ otherwise).

2.8 Carcinoma vs Non carcinoma Classification

A patch or image is said to be cancerous if it is either insitu or invasive and it is non cancerous otherwise. Let $q \in R^n$ be the final posterior of any patch/image. Then q_j is the probability that class of patch/img is j, where $j = \{1, 2, 3, 4\}$ and $j = 1$ for benign, $j = 2$ for insitu, $j = 3$ for invasive, $j = 4$ for normal. Let the probability that class of the input patch/image is i be $Pr(y = i|x)$, where $i = \{1, 2\}$ and $i = 1$ for carcinoma, $i = 2$ for non-carcinoma.

$$Pr(y = j|x) = q_j \tag{16}$$

$$Pr(y = i|x) = \begin{cases} q_2 + q_3, & \text{if } i = 1 \\ q_1 + q_4, & \text{if } i = 2 \end{cases} \tag{17}$$

3 Experimental Results and Discussion

In this section we evaluate the classification performance of our proposed ensemble framework with metrics such as accuracy, sensitivity, precision, specificity, F1 score.

3.1 Dataset

Conforming to the standard practice, the dataset is split into three parts for training, validation and testing purposes. 70%, 10% and 20% of the images are used for training, validation and testing respectively. This resulted in 280 images for training, 40 for validation and 80 for testing. Each image is divided into 35 patches resulting in 9800, 1400 and 2800 patches for the training, validation and testing sets respectively. Each training patch is then rotated through variations of 30° to generate 11 augmented patches, thus expanding the size of the training set to 117600 patches. The distribution of images and patches are uniform throughout the four classes.

3.2 Baseline Comparison

We employ the following baseline methods to compare our proposed model with:

- *Fine tuned CNNs with XGBoost classifier*
- *Fine tuned CNNs with softmax classifier*
- *Ensemble of fine tuned CNNs with XGBoost classifier*
- *Ensemble of fine tuned CNNs with softmax classifier*
- *Randomly initialized trained CNNs with XGBoost classifier*

	Accuracy(%)	
Ensemble Strategy	*4-class*	*2-class*
Mean	81.36	93
Product	86.50	98.57
Weighted Mean	83	95.63
Weighted Product	85	97.58
Majority Voting	77.84	89.41

TABLE 2: Comparison of different patch level ensemble strategies and their corresponding 4-class and 2-class accuracy. Patch level ensemble refer to the strategy used for ensembling the posterior outputs of the classifiers required for predicting the class of a patch.

	Accuracy(%)	
Ensemble Strategy	*4-class*	*2-class*
Mean	93.75	96.25
Product	91.25	95
Weighted Mean	92.5	95
Weighted Product	90	93.75
Majority Voting	95	98.75

TABLE 3: Comparison of different image level ensemble strategies and their corresponding 4-class and 2-class accuracy. Image level ensemble refer to the strategy used for ensembling the posterior outputs of the component patches for predicting the class of the image. Product rule has been used for patch wise ensembling since it gave highest accuracy among other methods.

- *Randomly initialized trained CNNs with softmax classifier*
- *Transfer learned CNNs with XGBoost classifier*

3.3 Classification Performance of Proposed Model

The comparative analysis of 4-class patch and image wise accuracies of the baseline models and the proposed ensemble model are shown in Table 5,Table 6. It can be easily seen that our proposed ensemble surpasses all other baselines including its componenet classifiers by a large margin in terms of accurcay. An ensemble model performs better than it's component classifiers when they are complementary in nature. This complementary nature of our method can be attributed to: i) using different fine tuned CNN architectures having different generalising capability, helps in extracting different discriminative features from the same patch, ii) using both XGBoost and softmax classifiers on top of the features extracted by the CNNs helps in adding more variations in the classifier predictions. The baseline comparison also corroborates our claim (See Section 2.4) of using a fine tuning strategy rather than training from scratch (initialized with random weights) or a transfer learning approach. Training CNNs initialized with random weights perform worse than fine tuned CNNs because the former demands more training samples to generalize on unseen data, thus making it prone to overfitting (See Section 2.4). Transfer learned CNNs perform conspicuously worst in terms of accuracy. It is because transferring features learnt on ImageNet dataset which consists of day to day natural images may not yield good results on the domain of medical images which are inherently quite different. This sheds light on the importance of the fine tuning approach which enables the CNNs initialized on pretrained Imagenet weights to learn target specific features.

From Table 5, Table 6, it can also be seen that XGBoost classifiers perform better by a 3-4% margin than the softmax classifiers on the same CNN extracted feature vector, which is consistent with our claim in Section 2.5.1. From Table 4, we can see that the ensemble of fine tuned CNNs with

	Accuracy(%)	
Baseline ensembles	*Patch wise*	*Image wise*
Fine tuned CNNs + softmax	82.50	90
Fine tuned CNNs + XGBoost	85	93.75
Fine tuned CNNs + softmax + XGBoost (Proposed ensemble)	86.50	95

TABLE 4: Comparison of 4-class patch and image wise accuracies of softmax and XGBoost ensembles with our proposed ensemble model which incorporates both softmax and XGBoost classifiers.

Baseline type		**Architecture**			
Feature extractor	*Classifier*	*VGG19*	*Inception-v3*	*ResNet-101*	*Inception-Resnet-v2*
Fine-tuned CNN	Softmax	81	77	79.50	80.20
Fine-tuned CNN	XGBoost	83	79.80	80.30	81.10
Randomly initialized CNN	Softmax	73.50	70	71.40	72
Randomly initialized CNN	XGBoost	75	7.20	72.80	75.60
Transfer learned CNN	XGBoost	66.50	64.70	61	63.40
Proposed Ensemble Model		86.50			

TABLE 5: Comparison of patch level 4 class accuracy (%) of proposed ensemble model with the baseline models .

Baseline type		**Architecture**			
Feature extractor	*Classifier*	*VGG19*	*Inception-v3*	*ResNet-101*	*Inception-Resnet-v2*
Fine-tuned CNN	Softmax	87.5	82.5	86.25	87.5
Fine-tuned CNN	XGBoost	91.5	86.25	90	90
Randomly initialized CNN	Softmax	80	77.50	78.75	78.75
Randomly initialized CNN	XGBoost	83.75	80	81.25	82.5
Transfer learned CNN	XGBoost	69.50	68.20	64.60	67
Proposed Ensemble Model		95			

TABLE 6: Comparison of image level 4 class accuracy (%) of proposed ensemble model with the baseline models .

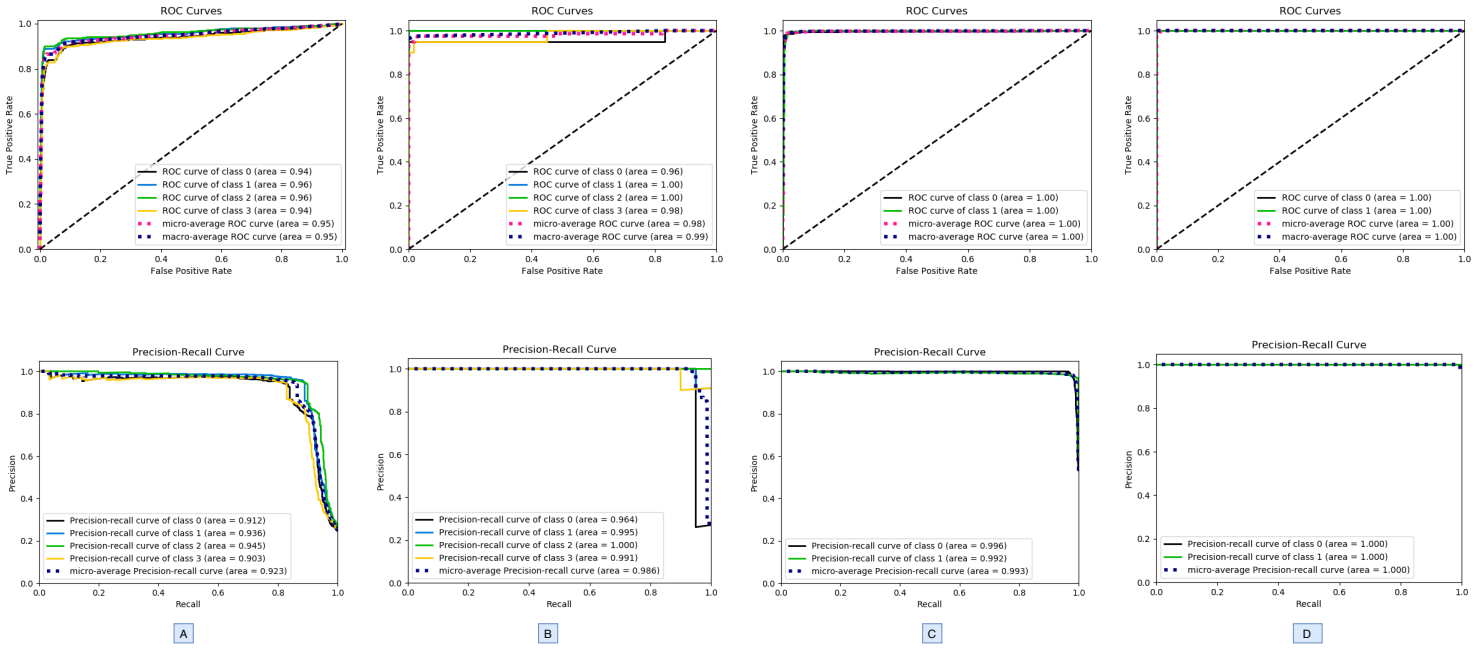

Fig. 5: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves of patch and image wise classification of final ensemble model. A) 4-class patch wise; B) 4-class image level; C) 2-class patch wise; D) 2-class image level. For 4-class: 0: Benign, 1: Insitu, 2: Invasive, 3: Normal. For 2-class: 0: Non-carcinoma, 1: Carcinoma.

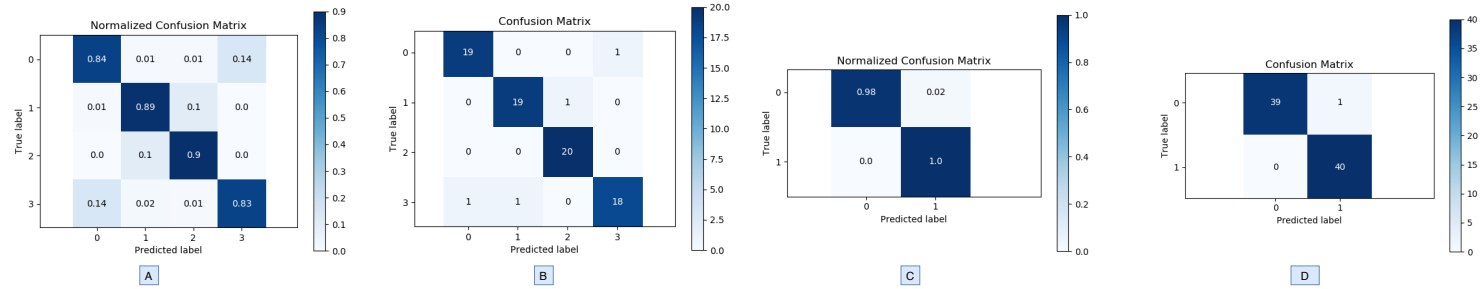

Fig. 6: Confusion matrix of patch and image wise classification of final ensemble model. A) 4-class patch wise; B) 4-class image level; C) 2-class patch wise; D) 2-class image level. For 4-class: 0: Benign, 1: Insitu, 2: Invasive, 3: Normal. For 2-class: 0: Non-carcinoma, 1: Carcinoma.

Metrics(%)	Classes				Avg
	Benign	InSitu	Invasive	Normal	
Accuracy	84	89	90	83	86.50
Precision	85	87.10	88.10	85.60	86.45
Sensitivity	84	89	90	83	86.50
Specificity	95.10	95.60	96	95.30	95.50
F1-Score	84.50	88	89	84.30	86.45

TABLE 7: Four class patch level classification metrics (%) of our proposed ensemble model.

Metrics(%)	Classes				Avg
	Benign	InSitu	Invasive	Normal	
Accuracy	95	95	100	90	95
Precision	95	95	95.24	94.74	95
Sensitivity	95	95	100	90	95
Specificity	98.33	98.33	98.33	98.33	98.33
F1-Score	95	95	97.56	92.31	94.97

TABLE 8: Four class image level classification metrics (%) of our proposed ensemble model.

XGBoost classifier performs better than that with softmax classifier by 3.50% and 3.75% on 4-class patch wise and image wise classification tasks respectively. It can also be seen that our final ensemble model after incorporating both softmax and XGBoost classifiers perform better than the XGBoost ensemble by 1.50% and 1.25% on 4-class patch wise and image wise classification tasks respectively. These results are also consistent with our claim in Section 2.5.1.

From Table 7, Table 8, Table 9 and Table 10, we can see that our ensemble model achieves a 4-class patch and image level accuracy of **86.50%** and **95%**. The 2-class patch and image level classification accuracy reported by our model is **98.57%** and **98.75%** respectively. We can also see that the

Metrics(%)	Classes		Avg
	Carcinoma	Non-carcinoma	
Accuracy	98	99.14	98.57
Precision	99.64	97.55	98.60
Sensitivity	97.50	99.65	98.58
Specificity	99.64	97.50	98.57
F1-Score	98.55	98.59	98.57

TABLE 9: Two class patch level classification metrics (%) of our proposed ensemble model.

Metrics(%)	Classes		Avg
	Carcinoma	Non-carcinoma	
Accuracy	97.5	100	98.75
Precision	100	97.56	98.78
Sensitivity	97.50	100	98.75
Specificity	100	97.50	98.75
F1-Score	98.73	98.77	98.75

TABLE 10: Two class image level classification metrics (%) of our proposed ensemble model.

sensitivity of the Benign and Normal classes are considerably lower than Insitu and Invasive classes. This can be attributed to the immense structural similarity amongst those two classes. Thus many benign patches get misclassified as normal and vice-versa. But in the case of 2-class classification, benign and normal classes are considered in the same category. Hence, the sensitivity of the non-carcinoma class is more than carcinoma class.

3.4 Ablation Studies

It is of vital importance to conduct ablation studies to study the importance of any particular model component on the classification performance of the entire framework. We study the importance and contribution of the following components to the final classification performance of the ensemble model, by removing them one at a time:

- **Fine tuned Inceptionv3 :** We remove the finetuned Inceptionv3 model and its corresponding softmax and XGBoost classifiers.
- **Fine tuned ResNet-101 :** We remove the finetuned ResNet-101 model and its corresponding softmax and XGBoost classifiers.
- **Fine tuned Inception-Resnet-v2 :** We remove the finetuned Inception-Resnet-v2 model and its corresponding softmax and XGBoost classifiers.
- **Fine tuned VGG19 :**We remove the finetuned VGG19 model and its corresponding softmax and XGBoost classifiers.
- **Softmax classifiers :** We remove the softmax classifiers from our final ensemble and only take into account the classification performance of the ensemble of XGBoost posteriors.
- **XGBoost classifiers :** We remove the XGBoost classifiers from our final ensemble and only take into account the classification performance of the ensemble of softmax posteriors.

For comapring the contribution of the individual CNN architectures, we can see from Table 11, that removing Inceptionv3 has the least impact on the final ensemble model, followed by ResNet-101, Inception-ResNet-v2 an then VGG19. This in fact follows from Table 5, Table 6) where we can see VGG19 gives the highest classification accuracy, followed by Inception-ResNet-v2, ResNet-101 and then Inception-v3. Thus contribution of VGG19 to the final ensemble is the most and that of Inceptionv3 is least. For comapring the choice of classifiers, it can be seen from Table 11, that removing XGBoost classifiers has a worse impact on the final ensemble accuracy, than removing softmax classifiers. Hence contribution of XGBoost is more than that of softmax.

3.5 Comparison with State of the Art

We compare our model with existing state of the art methods. From Table 12, we can see that the patch level accuracies are significantly more than those of the state of the art techniques. Among the methods which used ICIAR 2018 dataset for training and testing, the highest 4-class accuracy was 79%, achieved by Golatkar et al. [23]. Our model reports a 4-class patch wise accuracy more by 7.50%. The highest 4-class image level accuracy achieved by existing methods is 90% . Our model reports 5% more for the same. For 2-class classification, our model achieves patch and image wise accuracy more by 16% and 5% respectively, than the corresponding highest values achieved by existing methods ([55], [24]). Thus it gives support to our main claim that by incorporating different features from multiple fine tuned CNNs and classifying those features by using a robust classifier ensemble of XGBoost and softmax, we can improve patch wise classification performance compared to previous methods. Since patch wise accuracy is boosted beforehand, using a simple majority voting scheme for aggregating the constituent patch predictions suffices to raise the state of the art of 4-class image wise classification by 5%, to the best of our knowledge.

4 Conclusion

In this study, we proposed a novel hybrid ensemble framework for breast cancer histopathology image classification. With the objective of improving patch wise classification, we fine tune different CNN architectures for extracting a variety of discriminative feature representations from an input patch. For classification of the extracted features, incorporating XGBoost classifier results in better classification performance compared to the traditional softmax. For image level classification, we choose the majority class, out of those predicted by the constituent patches, as output. Our method outperforms other existing state of the art techniques, achieving a 4-class patch level accuracy of 86.50% and image level accuracy of 95%. Though our model successfully improves the patch wise and consequently image wise classification performance, one of the major caveats of our approach is that a lot of models needs to be trained separately, thereby increasing the complexity of the system. Hence, for future work, we intend to explore end to end trainable architectures, where a high resolution image can be directly fed as input to the model. In this case, proper

	Accuracy (%)			
Ablation	*4-class Patch Level*	*4-class Image Level*	*2-class Patch Level*	*2-class Image Level*
Fine tuned Inceptionv3	86.30	95	98.30	98.75
Fine tuned ResNet-101	85.50	93.75	96.25	97.50
Fine tuned Inception-Resnet-v2	85.20	93.75	95	96.25
Fine tuned VGG19	84.30	92.50	92.40	93.75
Softmax classifiers	85	93.75	94.80	96.25
XGBoost classifiers	82.50	90	90	92.5

TABLE 11: Comparison of different ablated models and their classification accuracy.

Method	Dataset used	Number of classes	Patch wise accuracy (%)	Image wise accuracy (%)
Araujo et al. [20]	Bioimaging 2015	4	66.7	77.8
Araujo et al. [20]	Bioimaging 2015	2	77.6	83.3
Rakhlin et al. [24]	ICIAR 2018	4	-	87.5
Rakhlin et al. [24]	ICIAR 2018	2	-	93.8
Vang et al. [25]	ICIAR 2018	4	-	87.5
Golatkar et al. [23]	ICIAR 2018	4	79	85
Golatkar et al. [23]	ICIAR 2018	2	-	93
Roy et al. [55]	ICIAR 2018	4	77.4	90
Roy et al. [55]	ICIAR 2018	2	84.7	92.5
Awan et al. [27]	ICIAR 2018	4	-	83
Awan et al. [27]	ICIAR 2018	2	-	87
Ferreira [56]	ICIAR 2018	4	-	90
Iesmantas et al. [57]	ICIAR 2018	4	-	87
Gao et al. [58]	ICIAR 2018	4	-	87.5
Wang et al. [59]	ICIAR 2018	4	-	91
Mahbod et al. [60]	ICIAR 2018	4	-	88.50
Chennamsetty et al. [61]	ICIAR 2018	4	-	87
Vu et al. [62]	ICIAR 2018	4	-	71
Yan et al. [63]	Private	4	82.1	91.3
Proposed	ICIAR 2018	4	86.50	95
Proposed	ICIAR 2018	2	98.57	98.75

TABLE 12: Comparison of proposed approach with state of the art methods

care has to be taken to prevent overfitting. To address the paucity of data, the images can be augmented using general adversarial networks (GANs). Also, to increase classification performance, attention mechanisms can be used to aid the network to look at important regions and ignore less important ones. With these things in mind, we will be working on this domain and make use of the best of the above methods to address the challenges.

Acknowledgments

The authors would like to thank the CMATER Research Laboratory of the Computer Science and Engineering Department, Jadavpur University, India, for providing us the infrastructural support to carry out this research.

References

[1] R. M. Rangayyan, F. J. Ayres, and J. L. Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," *Journal of the Franklin Institute*, vol. 344, no. 3-4, pp. 312–348, 2007.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *Ca-a Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[3] A. Surakasula, G. C. Nagarjunapu, and K. Raghavaiah, "A comparative study of pre-and post-menopausal breast cancer: Risk factors, presentation, characteristics and management," *Journal of research in pharmacy practice*, vol. 3, no. 1, p. 12, 2014.

[4] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *Jama*, vol. 313, no. 11, pp. 1122–1132, 2015.

[5] B. Stenkvist, S. Westman-Naeser, J. Holmquist, B. Nordin, E. Bengtsson, J. Vegelius, O. Eriksson, and C. H. Fox, "Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations," *Cancer research*, vol. 38, no. 12, pp. 4688–4697, 1978.

[6] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in *2008 19th International Conference on Pattern Recognition*, pp. 1–4, IEEE, 2008.

[7] A. Basavanhally, E. Yu, J. Xu, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Incorporating domain knowledge for tubule detection in breast histopathology using

o'callaghan neighborhoods," in *Medical Imaging 2011: Computer-Aided Diagnosis*, vol. 7963, p. 796310, International Society for Optics and Photonics, 2011.

[8] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 7, pp. 1977–1984, 2011.

[9] S. Ali and A. Madabhushi, "An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery," *IEEE transactions on medical imaging*, vol. 31, no. 7, pp. 1448–1460, 2012.

[10] A. B. Tosun and C. Gunduz-Demir, "Graph run-length matrices for histopathological image segmentation," *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 721–732, 2010.

[11] M. Veta, P. J. Van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in h&e stained breast cancer histopathology images," *PloS one*, vol. 8, no. 7, p. e70221, 2013.

[12] A. Brook, R. El-Yaniv, E. Isler, R. Kimmel, R. Meir, and D. Peleg, "Breast cancer diagnosis from biopsy images using generic features and svms," tech. rep., Computer Science Department, Technion, 2008.

[13] B. Zhang, "Breast cancer diagnosis from biopsy images by serial fusion of random subspace ensembles," in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 1, pp. 180–186, IEEE, 2011.

[14] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 411–418, Springer, 2013.

[15] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, p. 26286, 2016.

[16] K. Sirinukunwattana, S. e Ahmed Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images.," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[17] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 international joint conference on neural networks (IJCNN)*, pp. 2560–2567, IEEE, 2016.

[18] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[19] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd International conference on pattern recognition (ICPR)*, pp. 2440–2445, IEEE, 2016.

[20] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, *et al.*, "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, 2019.

[23] A. Golatkar, D. Anand, and A. Sethi, "Classification of breast cancer histology using deep learning," in *International Conference Image Analysis and Recognition*, pp. 837–844, Springer, 2018.

[24] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *International Conference Image Analysis and Recognition*, pp. 737–744, Springer, 2018.

[25] Y. S. Vang, Z. Chen, and X. Xie, "Deep learning framework for multi-class breast cancer histology image classification," in *International Conference Image Analysis and Recognition*, pp. 914–922, Springer, 2018.

[26] H. Cao, S. Bernard, L. Heutte, and R. Sabourin, "Improve the performance of transfer learning without fine-tuning using dissimilarity-based multi-view learning for breast cancer histology images," in *International Conference Image Analysis and Recognition*, pp. 779–787, Springer, 2018.

[27] R. Awan, N. A. Koohbanani, M. Shaban, A. Lisowska, and N. Rajpoot, "Context-aware learning using transferable features for classification of breast cancer histology images," in *International Conference Image Analysis and Recognition*, pp. 788–795, Springer, 2018.

[28] T. A. A. Tosta, L. A. Neves, and M. Z. do Nascimento, "Segmentation methods of h&e-stained histological images of lymphoma: a review," *Informatics in medicine unlocked*, vol. 9, pp. 35–43, 2017.

[29] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2015.

[30] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[31] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab, "Structure-preserved color normalization for histological images," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 1012–1015, IEEE, 2015.

[32] L. Deng, D. Yu, *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[33] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, W. Xie, G. L. Rosen, *et al.*, "Opportunities and obstacles for deep learning in biology and medicine. biorxiv," 2017.

[34] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: a reviewcurrent status and future potential," *IEEE reviews in biomedical engineering*, vol. 7, pp. 97–114, 2013.

[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, pp. 3320–3328, 2014.

[36] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[37] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 31–40, 2016.

[38] J. Margeta, A. Criminisi, R. Cabrera Lozoya, D. C. Lee, and N. Ayache, "Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 5, no. 5, pp. 339–349, 2017.

[39] A. Kumar, P. Sridar, A. Quinton, R. K. Kumar, D. Feng, R. Nanan, and J. Kim, "Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 791–794, IEEE, 2016.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning. corr abs/1602.07261," *URL http://arxiv. org/abs/1602.07261*, 2016.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition. computer vision and pattern recognition (cvpr)," in *2016 IEEE Conference on*, vol. 5, p. 6, 2015.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[45] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural

networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.
[46] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[47] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
[48] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, ACM, 2016.
[49] M. Luckner, B. Topolski, and M. Mazurek, "Application of xgboost algorithm in fingerprinting localisation task," in *IFIP International Conference on Computer Information Systems and Industrial Management*, pp. 661–671, Springer, 2017.
[50] W. Yu, Z. Na, Y. Fengxia, and G. Yanping, "Magnetic resonance imaging study of gray matter in schizophrenia based on xgboost," *Journal of Integrative Neuroscience*, vol. 17, no. 4, pp. 331–336, 2018.
[51] M. Livne, J. K. Boldsen, I. K. Mikkelsen, J. B. Fiebach, J. Sobesky, and K. Mouridsen, "Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke," *Stroke*, vol. 49, no. 4, pp. 912–918, 2018.
[52] M. Nishio, M. Nishizawa, O. Sugiyama, R. Kojima, M. Yakami, T. Kuroda, and K. Togashi, "Computer-aided diagnosis of lung nodule using gradient tree boosting and bayesian optimization," *PloS one*, vol. 13, no. 4, p. e0195875, 2018.
[53] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, "A novel image classification method with cnn-xgboost model," in *International Workshop on Digital Watermarking*, pp. 378–390, Springer, 2017.
[54] L. Li, R. Situ, J. Gao, Z. Yang, and W. Liu, "A hybrid model combining convolutional neural network with xgboost for predicting social media popularity," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1912–1917, ACM, 2017.
[55] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri, "Patch-based system for classification of breast histology images using deep learning," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 90–103, 2019.
[56] C. A. Ferreira, T. Melo, P. Sousa, M. I. Meyer, E. Shakibapour, P. Costa, and A. Campilho, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," in *International Conference Image Analysis and Recognition*, pp. 763–770, Springer, 2018.
[57] T. Iesmantas and R. Alzbutas, "Convolutional capsule network for classification of breast cancer histology images," in *International Conference Image Analysis and Recognition*, pp. 853–860, Springer, 2018.
[58] Y. Guo, H. Dong, F. Song, C. Zhu, and J. Liu, "Breast cancer histology image classification based on deep neural networks," in *International Conference Image Analysis and Recognition*, pp. 827–836, Springer, 2018.
[59] Y. Wang, L. Sun, K. Ma, and J. Fang, "Breast cancer microscope image classification based on cnn with image deformation," in *International Conference Image Analysis and Recognition*, pp. 845–852, Springer, 2018.
[60] A. Mahbod, I. Ellinger, R. Ecker, Ö. Smedby, and C. Wang, "Breast cancer histological image classification using fine-tuned deep network fusion," in *International Conference Image Analysis and Recognition*, pp. 754–762, Springer, 2018.
[61] S. S. Chennamsetty, M. Safwan, and V. Alex, "Classification of breast cancer histology image using ensemble of pre-trained neural networks," in *International Conference Image Analysis and Recognition*, pp. 804–811, Springer, 2018.
[62] Q. D. Vu, M. N. N. To, E. Kim, and J. T. Kwak, "Micro and macro breast histology image analysis by partial network re-use," in *International Conference Image Analysis and Recognition*, pp. 895–902, Springer, 2018.
[63] R. Yan, F. Ren, Z. Wang, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, and F. Zhang, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, 2019.