Journal of Digital Imaging (2021) 34:667–677
https://doi.org/10.1007/s10278-021-00442-5

ORIGINAL PAPER

HER2 Molecular Marker Scoring Using Transfer Learning and Decision Level Fusion

Suman Tewary[1,2] · **Sudipta Mukhopadhyay**[3]

Received: 10 May 2020 / Revised: 13 January 2021 / Accepted: 1 March 2021 / Published online: 19 March 2021

Abstract

In prognostic evaluation of breast cancer, immunohistochemical (IHC) marker human epidermal growth factor receptor 2 (HER2) is used for prognostic evaluation. Accurate assessment of HER2-stained tissue sample is essential in therapeutic decision making for the patients. In regular clinical settings, expert pathologists assess the HER2-stained tissue slide under microscope for manual scoring based on prior experience. Manual scoring is time consuming, tedious, and often prone to inter-observer variation among group of pathologists. With the recent advancement in the area of computer vision and deep learning, medical image analysis has got significant attention. A number of deep learning architectures have been proposed for classification of different image groups. These networks are also used for transfer learning to classify other image classes. In the presented study, a number of transfer learning architectures are used for HER2 scoring. Five pre-trained architectures viz. *VGG16*, *VGG19*, *ResNet50*, *MobileNetV2*, and *NASNetMobile* with decimating the fully connected layers to get 3-class classification have been used for the comparative assessment of the networks as well as further scoring of stained tissue sample image based on statistical voting using mode operator. HER2 Challenge dataset from Warwick University is used in this study. A total of 2130 image patches were extracted to generate the training dataset from 300 training images corresponding to 30 training cases. The output model is then tested on 800 new test image patches from 100 test images acquired from 10 test cases (different from training cases) to report the outcome results. The transfer learning models have shown significant accuracy with *VGG19* showing the best accuracy for the test images. The accuracy is found to be 93%, which increases to 98% on the image-based scoring using statistical voting mechanism. The output shows a capable quantification pipeline in automated HER2 score generation.

Keywords Deep learning · Image analysis · Immunohistochemical (IHC) analysis · HER2 molecular marker · Transfer learning

Introduction

Breast cancer is second leading cause of mortality and most commonly diagnosed cancer among women [1]. In regular clinical settings, the diagnostic and prognostic evaluation of cancer is manual observation of stained tissue samples under a microscope by expert pathologists. The progression of cancer is highly associated with change in molecular expression observed as histological features. The visual assessment of histological features of stained tissue samples is normally slow, and often results in inaccurate and irreproducible in diagnostic evaluation of cancer [2]. The prognostic evaluation of cancer is assessed by immunohistochemistry (IHC). Ki-67, oestrogen receptor (ER), progesterone receptor (PR) and HER2 molecular markers are regularly used in therapy planning of the breast cancer patients [3]. HER2 molecular marker consists of four tyrosine kinase receptors placed in cell membrane, which stimulate growth of cells when activated. HER2 is overexpressed in approximately 15% of all primary breast cancers which activates uncontrolled cell proliferation [4].

✉ Sudipta Mukhopadhyay
smukho@ece.iitkgp.ac.in

[1] School of Medical Science and Technology, Indian Institute of Technology Kharagpur, Kharagpur, India

[2] Computational Instrumentation, CSIR-Central Scientific Instruments Organisation, Chandigarh, India

[3] Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India

In regular pathological evaluation, the assessment of HER2-stained tissue sample consists of observing the histological features such as stained cells and cell membrane under microscope having 40× objective and 10× eyepiece lens. These histological features consist of two components: hematoxylin (blue coloured negatively stained cell) and diaminobenzidine or DAB (brown coloured often encircling the tumour cells with variation in intensity). DAB-stained region of cell membrane is important feature for the assessment, which involves the continuity of cell membrane as well as intensity of staining. High intensity cell membrane having a continuity in encircling the tumour cells is scored higher as compared with a low intensity or the broken cell membrane region. In HER2 scoring, the stained tissue samples are scored as negative (0/1+), equivocal (2+) and positive (3+) based on the amount of complete or broken cell membrane and the stain intensity of scoring [5]. The equivocal cases are assessed further for gene amplification through fluorescence in situ hybridization (FISH). The recommendation is shown in Table 1.

For the prognostic evaluation of breast cancer patients, it is important to have accurate assessment in spite of heterogeneity of HER2-stained tissue samples. However, this histological feature assessment is subjective often semi-quantitative to assess the expression level of HER2 protein in stained breast cancer tissue sample [6]. These histological feature assessments, when performed manually, by expert pathologists, are highly subjective, tedious and quite often lead to inter-observer variation. These evaluation results in error-prone decision are unreliable and affect the therapeutic decision making for cancer patients. Often, the HER2 scoring is inaccurate as high as in 20% cases of breast cancer [4]. Also, the heterogeneity in staining from different laboratories and the optics involved in imaging the stained tissue samples can add more errors in terms of inter-observer variability. Automated scoring can overcome the challenges in manual approaches as the automation is not prone to subjective bias along with precise quantitative analysis that can assist pathologists in providing reproducible score [7].

The heterogeneity in HER2 membrane is a challenge for automated image analysis. A number of commercial softwares are available for HER2 scoring such as Automated Cellular Imaging System III (*ACIS III*) (*Dako*), *HER2-CONNECT* (*Visiopharm*) and *SlidePath* Tissue Image Analysis system (*Leica*) [8]. *ACIS III* was evaluated for HER-2 image analysis in gastroesophageal (GE) adenocarcinomas where overall correlation of 84% was achieved between manual HER-2 scoring and the *ACIS III* [9]. *HER2-CONNECT* has shown agreement of 92.3% between the software and the score by pathologists [10]. *SlidePath* Tissue Image Analysis system has shown 91% agreement between manual and Digital Image Analysis for HER2 scoring competing other commercially available image analysis software [11]. All these commercial applications depend on specific materials and are costly for any pathological centre [8]. Under such situation, it is essential to have automated image analysis for the histological evaluation in HER2 scoring for quantitative evaluation. By using automated approaches for HER2-stained tissue samples, the errors in subjective evaluation can be reduced.

Researchers across the world have developed a number of computer vision and machine learning algorithms for the automated extraction of cell membrane in scoring the HER2-stained tissue samples. For the automated image analysis, one of the most important parameter is to separate the stains by extracting the hematoxylin, eosin and DAB components from colour image of RGB colour information using colour deconvolution [12]. *ImmunoMembrane*, a web-based application, is proposed for assessment of HER2-stained tissue using colour deconvolution followed by series of steps for segmentation of cell membrane and classification of HER2 score as 0/1+, 2, and 3+ [13]. Earlier with this application, user could upload IHC images and get the HER2 score, and now, this application is available as *ImageJ* plugin. Along with RGB colour space, various other colour spaces were utilized by researchers to develop algorithms. Cell membrane extraction with modified active contour and particle swarm optimization for parameter tuning is used on Y channel, of CMYK colour space, followed by localizing the nucleus of cell using B channel of RGB colour space [14]. In another work, stain separation is proposed using Y channel from the CMYK colour space [15]. In segmentation and quantification of membrane structures, a fuzzy decision tree is proposed for accurate HER2 scoring [14]. In another automated approach, grey-level hit-or-miss transform-based hourglass shapes rank is used in extracting

Table 1 Recommendation for scoring the HER2-stained tissue sample [5]

Score = 0/1+	Score = 2+	Score = 3+
No membrane staining or incomplete membrane staining in < 10% of invasive tumour cells (0+) or faint/barely perceptible or weak incomplete membrane staining in > 10% of tumour cells (1+) *Clinical Significance—negative*	A weak to moderate complete membrane staining is observed in > 10% of tumour cells or strong complete membrane staining in ≤ 10% of tumour cells *Clinical significance—equivocal or borderline*	A strong (intense and uniform) complete membrane staining is observed in > 10% of invasive tumour cells *Clinical significance—positive*

stained membrane and cell nucleus to find the membrane staining continuity for HER2 scoring [16]. A variety of features such as colour, texture and morphological features have been used by researchers for classification. Stain intensity features and HER2 membrane completeness were used for classification [17]. In another work, combination of features viz. local binary patterns (LBP), histogram of oriented gradients (HOG) and Haralick were used for classification in HER2 scoring [18]. In a recent work, colour and texture features with comparative evaluation of different machine learning approaches were presented in HER2 assessment for image patch and patient level scoring with higher accuracy of 94.2% for 0/1+, 2+, and 3+ cases [8]. In another work, a number of features such as connectedness in uniform local binary pattern, characteristic curves, entropy and energy features with logistic regression and SVM classifier to score HER2-stained tissue samples [19]. In the recent work, 3-class HER2 scoring is done using colour space-based membrane extraction followed by SVM classifier [20].

Deep learning has made major advances in addressing the challenges for artificial intelligence community and turned out to discover efficiently the intricate features in high-dimensional data in different domains of science [21]. Deep learning has shown good usage in microscopic image analysis [22, 23]. With the efficient computation power and advanced learnings, researchers have come up with different deep learning approaches for HER2 scoring. Convolutional neural networks (CNN) have been used for directly feeding the HER2 image patches 128×128 pixels for multi-class classification in HER2 scoring that achieved 97.7% testing accuracy [24]. CNN were used in assessing the cancer cell types and generate accurate HER2 score as compared with the clinical scores [25]. In another work, convolutional neural network architecture *Her2Net* was proposed for segmentation and labelling the HER2-stained tissue samples for HER2 scoring that achieved 98.33% accuracy [26]. In a recent work, stain intensity and HER2 membrane completeness features were used where super-pixel-based tissue region segmentation is performed followed by SVM to distinguish epithelial and stromal region, which are scored using modified *UNet* model [27]. The heterogeneity in HER2-stained tissue samples is very high and a very recent work has targeted this challenge by proposing a deep learning approach to predict where to see in the tissue sample [28]. This approach learns the discriminative features by processing recurrent and residual convolution networks for HER2 scores followed by predicting next location for sample under observation without processing all sub-image patch scores.

The heterogeneity in HER2-stained tissue samples is very complex like the natural images. The 1000-class database ImageNet is very popular for different deep learning architectures [29]. This dataset has been reported very good for transfer learning [30]. Transfer learning is reported for medical image classification of cancer assessment in pathological images [31]. In abdominal ultrasound images, transfer learning has been used for effective classification of organs [32]. Tuning the pre-trained models on ImageNet could help in generating accurate results for the classification of HER2-stained tissue samples. Pre-trained networks *ResNet50* [33], *MobileNetV2* [34], *NASNetMobile* [35] and *VGG* networks *VGG16* and *VGG19* [36] have resulted good accuracy in benchmark dataset.

In this work, transfer learning with modified output layers by removing the last fully connected (FC) layer and collective voting scheme is presented for HER2 screening. Also, a comparative assessment is presented for selection of best CNN architecture. The approach can be a great advance in accurately scoring the heterogeneously stained HER2 samples. The methodology and results are described in the subsequent sections.

Materials and Methods

In this section, the details on data preparation, developed methodology for training and testing on HER2-stained tissue samples, will be discussed.

Data Preparation

In this presented work, the dataset consisted of 172 whole-slide images (WSIs), corresponding to 86 cases of breast cancer patients, in Nano-zoomer Digital Pathology (NDPI) format [7]. A total of 52 cases (13 cases for 0, 1+, 2+, and 3+ class) are labelled by expert pathologists, and for our proposed approach, 40 cases are selected. The reason for selection of three class instead of four, as provided in database, is that the convention for negative (0/1+), equivocal (2+) and positive (3+) is widely used by pathologists using the protocol as shown in Table 1. *ImageJ*-based software *ImmunoMembrane* also provides three-class classification [13]. Following the same protocol, in another recent work, three-class problem is presented [20]. In the current work, to address most of the cases provided in the dataset, all 13 equivocal cases and 13 positive cases are used and the rest 14 cases are negative cases (0/1+) having equal weightage for each class. For all 40 cases, 10 random images are selected corresponding to the label provided by pathologist making total 400 images. For heterogeneity, 30 cases are selected for training and the rest 10 cases are used for testing to have 75:25 ratio in train/test without any overlapping of cases. 40× magnification is selected for the development of the database. The images are opened in *Sedeen* viewer and multiple snapshots are taken for

the preparation of the dataset. The acquired image size is 946 × 1920 pixels which means that the resolution has aspect ratio of approximately 1:2. The image is resized to 448 × 896 pixels and then cropped into 8 sub-image patches to make every patch of 224 × 224 pixels. A total 2130 image patches are generated from the 30 train cases by randomly selecting in total 300 images (2400 patches) followed by cropping each image into 8 sub-image patches.

Further, those images having only background of tissue, region without any staining and very small regions showing staining components are removed for the training dataset generation purpose. At the end, we kept nearly equal weightage of each class and the final dataset contains 2130 image patches with corresponding levels. For the testing process in this work, the input test images 100 images or 800 image patches are different from the training data for robust testing. In Fig. 1, the sample image of HER2-stained tissue image and corresponding patches are shown. The image patch size 224 × 224 pixels are chosen due to the fact that the commonly used deep learning architectures use default input image size of 224 × 224 pixels.

Proposed Methodology

In this work, the proposed methodology for HER2 scoring is patch-based scoring followed by voting scheme to level the image-based scoring. The selection of deep learning classifier for the prediction of patch-based HER2 score and complete image-based scoring using voting scheme will be discussed in the subsequent sub-sections.

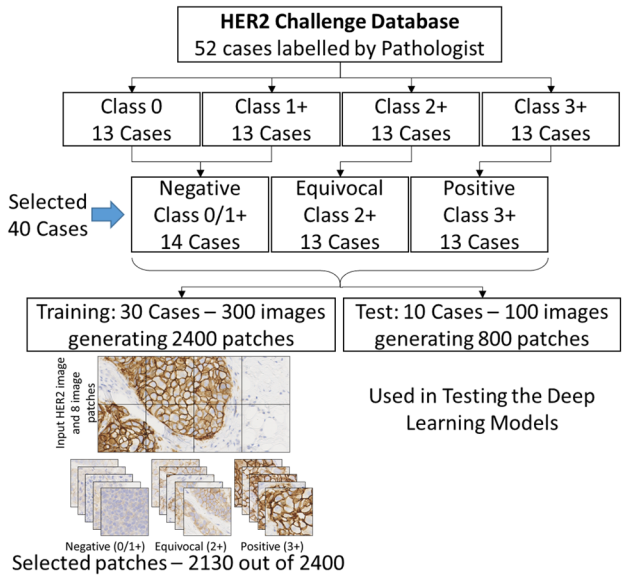

Fig. 1 HER2-stained image and cropped image patches for three classes from the benchmark HER2 Challenge Database

Data Augmentation

Often, small dataset in training causes overfitting, and to generate a robust training model, the training data should be large enough to have variability of data within acceptable limit. To get this additional modification data, augmentation is needed. Data augmentation lets the user increase the training data set as well as reduce over-fitting problems as shown for mitosis detection for breast cancer H&E images and in the recent work on detection and classification of benign and malignant cells in the breast cytology images [37, 38]. In this work, before feeding the data for training, data augmentation is applied. The augmentation involves width and height shift, shear, horizontal and vertical flip and rotation. This step helps in increasing the data for proving a robust training for the heterogeneous data. This step is common for all the training steps as described in the next sub-section.

Transfer Learning—Patch Classification

The most robust CNN architectures have shown significant accuracy for different medical images. A comparative assessment of transfer learning has shown good performance in heterogeneous pathological images [31]. In this work, the well-known deep learning architectures viz. *VGG16*, *VGG19*, *ResNet50*, *MobileNetV2* and *NASNetMobile* have been used for the comparative assessment of the networks. All these pre-trained networks from *Keras* library takes input image size as 224 × 224 × 3 and output shape as fully connected layer of 1000 classes. Here in this study, the pre-trained networks are used to train the last layer with additional fully connected layers to classify into three output classes to predict the score as negative, equivocal and positive. This is formed by three fully connected layers with the *softmax* function defined as:

$$f_{\mathrm{i}}(z) = \frac{e^{z_{\mathrm{i}}}}{\sum_{\mathrm{k}} e^{z_{\mathrm{i}}}} \tag{1}$$

where f_{i} is the i-th value in the class scores vector f and z is the vector of arbitrary real value scores that are squashed in the range *[0, 1]* with summation to one representing the probability distribution. The rectified linear unit (*ReLU*) is used in the additional FC layers which is defined as

$$f(x) = \max(0, x) \tag{2}$$

where the function f thresholds the activations at zero. The pre-trained networks with the additional layers are described in this section.

VGG Network Architectures—VGG16 and VGG19

VGG networks are very deep convolutional networks (up to 19 weight layers) developed for large-scale image classification where the best performing architectures are *VGG16* and

VGG19 [36]. *VGG16* consists of 13 convolution blocks (convolution, rectification, pooling) and 3 fully connected layers. *VGG19* consists of 16 convolution blocks (convolution, rectification, pooling) and 3 fully connected layers. The convolution network uses 3×3 window size filter and 2×2 pooling network. VGG performs better as compared with other networks due to its simple deep architecture. *VGG16* has 16 weight layers and *VGG19* has 19 weight layers in their architecture. In our work, both of the pre-trained models have been used. All the network layers are set as non-trainable with additional fully connected layers to the 3-class output layer.

ResNet Architecture—ResNet50

ResNet is a very deep network by learning from the residual representation functions and won the *ImageNet* challenge

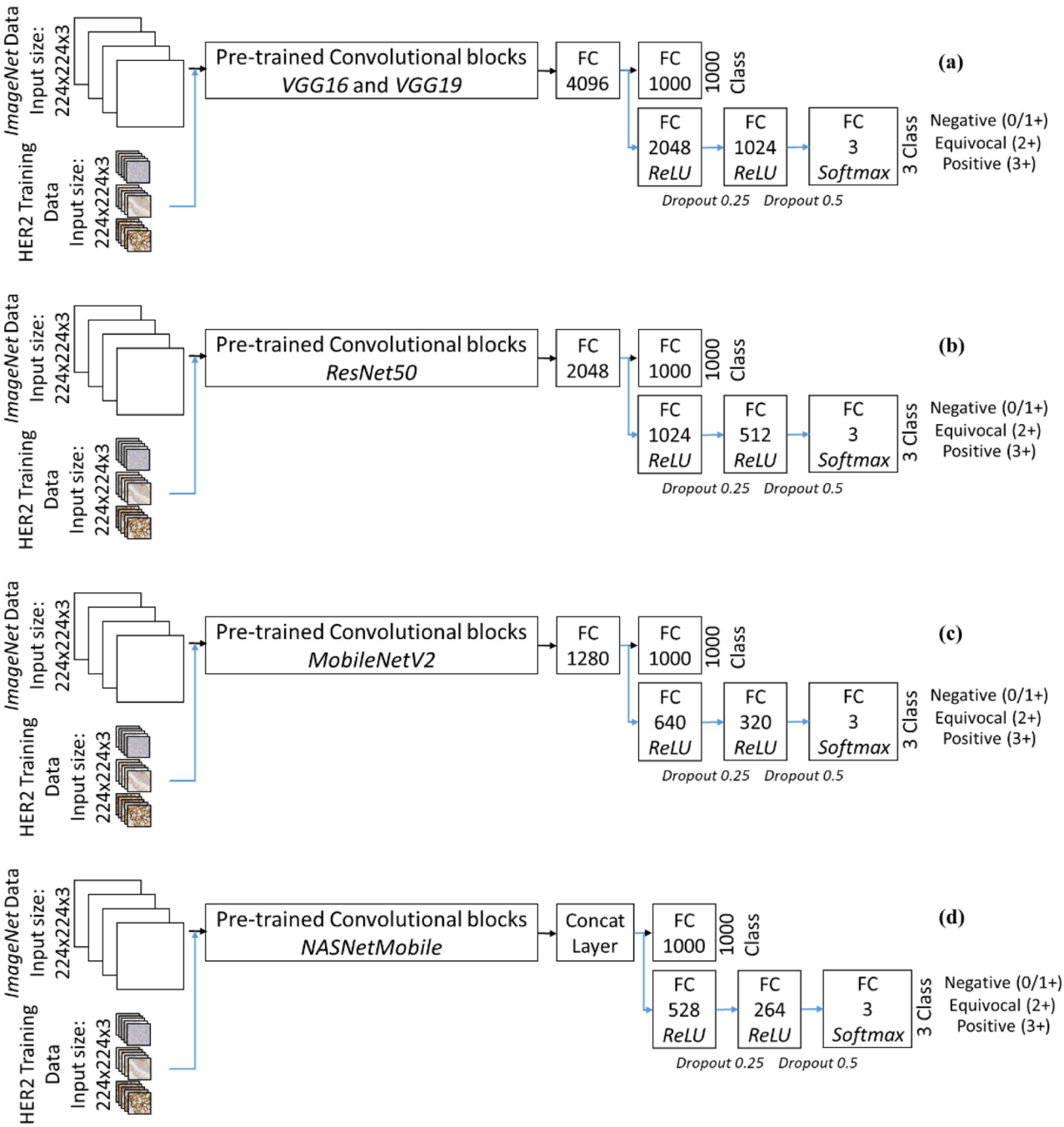

Fig. 2 The proposed transfer learning scheme with different pre-trained classifiers (**a**) *VGG16* and *VGG19*, (**b**) *ResNet50*, (**c**) *MobileNetV2* and (**d**) *NASNetMobile*

in 2015 [33]. It combined convolution filters of multiple size to manage the degradation problem as well as reducing the training time. The network uses shortcut connection for deep residual learning, to overcome the vanishing gradient problem of plain networks, with bottleneck design to reduce time complexity. There are multiple variants available, and in our proposed study, *ResNet50* pre-trained model is used. The training weights are not fixed with updating the batch normalization by setting the layers trainable with additional fully connected layers to the 3-class output layer.

MobileNet Architecture—MobileNetV2

For resource constraint environment, the *MobileNet* architecture was proposed having inverted residual with linear bottleneck. The convolutional blocks permit to isolate the network expressiveness, represented by expansion layers, from its capacity, represented by bottleneck inputs [34]. This module takes input as a low-dimensional compressed depiction that first expanded to high dimension and then filtered with a lightweight depth-wise convolution. In our work, *MobileNetV2* pre-trained model is used. Similar to the *ResNet50*, this network's training weights are not fixed with updating the batch normalization by setting the layers trainable and additional fully connected layers to the 3-class output layer.

NASNet Architecture—NASNetMobile

In this architecture, the architectural building block is searched on a small dataset followed by transferring it to a larger dataset. The search space called as neural network search (*NASNet* search space) enables transferability in the architecture [35]. In this work, the mobile version of the network *NASNetMobile* pre-trained model is used. Here also, the layers are set as trainable with additional fully connected layers to the 3-class output layer. The three levels of data having total 2130 patches with data augmentation are fed to the pre-trained models with modified last layers.

The schematic of the developed transfer learning framework is shown in Fig. 2. Dropout is added in these layers. The pre-trained networks are modified and re-trained with the added fully connected dense layers is trained. In the proposed scheme, the *VGG* networks have similar fully connected layers, and due to this, both *VGG16* and *VGG19* are shown in one block as Fig. 2a. The weights not changed on the pre-trained *VGG* networks and the only modification in terms of weights are due to the additional fully connected layers. On the other hand, the rest of the pre-trained networks were re-trained along with the additional fully connected layers due to the presence of batch normalization layers which need the update in training weights. All these networks are trained for 25, 50, 75 and 100 epochs with batch size 32. The learning rate is set at 0.0001 with *adam* optimizer. The training is done with robust testing on separate image data.

Statistical Voting for HER2 Scoring

Using the trained classifiers, the performance of the same can be tested on the image patches. These patches are only sub-images of the whole image. Hence, a voting scheme is required for decision level fusion to generate the overall score of the image. For this purpose, statistical operator *Mode* is used. *Mode* operator from the eight patch-based scores will give the most frequently occurring score in the image. Overall, this will generate the HER2 score.

Testing Methodology

The proposed methodology is tested on a new data set, different from the training data. From the 10 test cases, a total of 100 images are used for the testing. These 100 images are different than training. These images are fed to the developed methodology with patch classification using transfer learning to generate the patch-based score followed by statistical voting using the mode operator. This voting scheme will generate the final image-based score for the 100 test images. The overall testing scheme is shown in Fig. 3. From

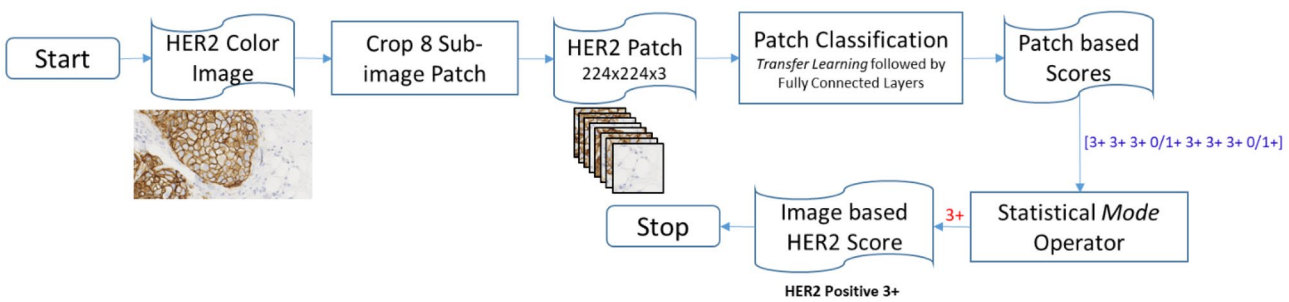

Fig. 3 Overall schematic for the score generation in the proposed approach

these 100 test images, there are 800 image patches generated for further statistical voting. For comparative assessment, both patch-level scoring and image-based scoring will be described in the next section.

Implementation Details

The proposed approach is developed using *Keras* library with *TensorFlow* backend on *Python 3.7* platform. The pre-trained models are available under *Keras* applications library. The *Keras* library allows defining the deep convolutional networks and modifying the last layers for the dataset used in this study. Along with *Keras*, other python libraries such as *OpenCV*, *NumPy*, *Scikit-learn* and *Matplotlib* are used for the implementation. All the training is performed on a HP Z6 G4 Workstation having Xeon 4114 processor, 8 GB *NVIDIA* P4000 graphics card, 32 GB RAM and the Windows operating system. A solid-state drive was used for storing the training data and trained models. This fast storage makes the approach faster for training as well as testing.

Performance Metrics

The proposed approach is assessed by different performance metrics to select the most suitable architecture for the dataset used in this study. The most widely used training accuracy curves for all five folds are used to see the variation of training among different training folds. Along with the training curves, the confusion matrix for the best performing fold is used to compare different CNN architectures. On the testing data of 100 images having 800 sub-image patches, the classification report is used for the comparative assessment. Matrices used in classification report are precision, recall, F1-score, accuracy, macro accuracy and weighted accuracy. For N test samples, the true positive (TP), false negative (FN), false positive (FP) and true negative (TN) are calculated to generate these performance metrics using the following formula [39] for l class problem with individual class weightage, w_c:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$F1\text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{Accuracy} = \frac{\sum_{c=1}^{l} \text{TP}_c}{N} \tag{6}$$

$$\textit{Macro Average} = \begin{cases} \frac{\sum_{c=1}^{l} Precision}{l} \\ \frac{\sum_{c=1}^{l} Recall}{l} \\ \frac{\sum_{c=1}^{l} F1score}{l} \end{cases} \tag{7}$$

$$\text{Weighted Average} = \sum_{c=1}^{l} w_c \times \begin{cases} \frac{\sum_{c=1}^{l} \text{Precision}}{l} \\ \frac{\sum_{c=1}^{l} \text{Recall}}{l} \\ \frac{\sum_{c=1}^{l} F1\text{score}}{l} \end{cases} \tag{8}$$

Finally, the improvement in accuracy for patch-based scoring to image-based scoring using the proposed voting scheme the normalized confusion matrix for each architecture is used. The results obtained from the proposed approach and the test schematic will be discussed in the next section.

Results and Discussion

The proposed approach is developed on 2130 image patches having three levels (0/1+, 2+ and 3+). These images are extracted from 30 different cases of HER2-stained sample of breast cancer patients. *ImageNet* data-based pre-trained models are trained with the prepared image patches and modified fully connected layers as discussed in the previous section. These five networks are training using the 2130 image patches with data augmentation to increase this data in multiple folds. All these image data have roughly equal contribution in the dataset to have a balanced distribution of data. As the testing data is separate than the training data, hence there is no division in terms of *train/test* for the data. All image patches are used in the training for different number of epochs.

All the five models have shown on average more than 90% training accuracy. The accuracy for the five deep learning classifiers is represented under different epochs. The training is performed for 25, 50, 75 and 100 epochs with batch size 32. Five networks, viz. *VGG16*, *VGG19 ResNet50*, *MobileNetV2* and *NASNetMobile*, are trained and the training accuracies and loss are shown in Fig. 4. These models are then tested on new set of test data. In test dataset, a total of 100 images were selected from the 10 cases of HER2-stained tissue samples. All these 100 images are fed to the network, where each image is cropped into 8 sub-image patches, i.e. 800 patches.

For all these patches, no manual score is provided. The overall score for the case is used for the level of these patches. Hence, it may be possible that the testing accuracy will be lower as all 8 patches will have the same level. The final score is based on the 8 scores generated from each patch and the statistical mode operator is used for the overall image level score. The image patch based (P based) and

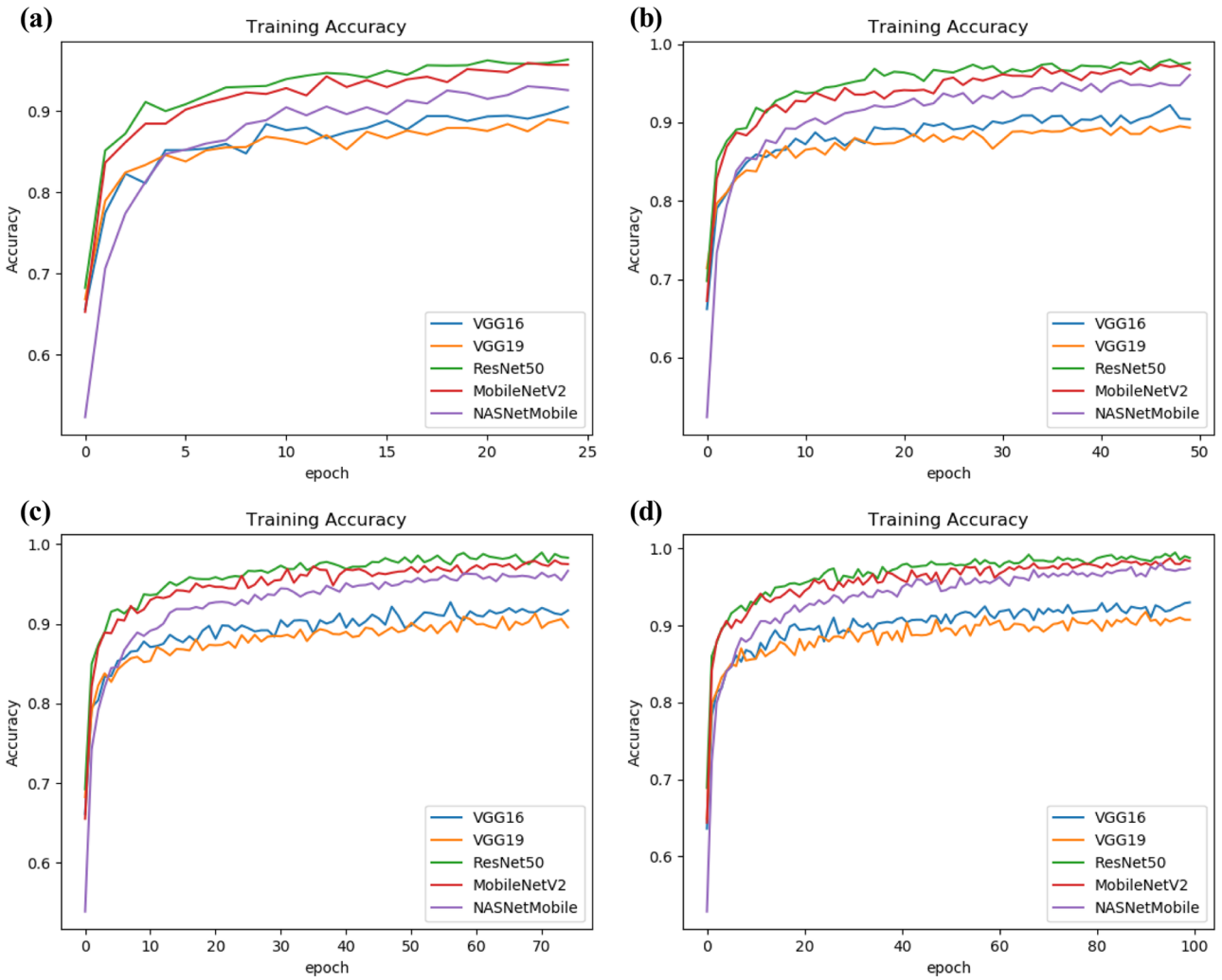

Fig. 4 Training accuracy reported in (**a**) 25, (**b**) 50, (**c**) 75 and (**d**) 100 epochs for training using *VGG16*, *VGG19*, *ResNet50*, *MobileNetV2* and *NASNetMobile*

overall image based (*I* based) using the statistical voting is testing scheme is shown. This network performance in a glance is shown in Table 2.

As the best accuracy is achieved in *VGG19* network for 100 epochs, the same is used for the comparative assessment for all the five networks. This will emphasize on the improvement of scoring based on statistical voting-based decision. It can be noted that all the test patches are not labelled manually. Each patch is labelled as per the ground truth labelling for the whole image, which means each patch

Table 2 Training and testing performance for the pre-trained networks with fully connected layers

Transfer learning pre-trained model with FC layers	25 Epochs			50 Epochs			75 Epochs			100 Epochs		
	Training	Testing		Training	Testing		Training	Testing		Training	Testing	
		P based	*I* based		*P* based	*I* based		*P* based	*I* based		*P* based	*I* based
VGG16	0.91	0.90	0.96	0.90	0.92	0.97	0.92	0.90	0.95	0.93	0.88	0.95
VGG19	0.89	0.93	0.96	0.89	0.88	0.94	0.90	0.92	0.98	0.91	0.93	0.98
ResNet50	0.96	0.87	0.94	0.98	0.84	0.94	0.98	0.81	0.87	0.99	0.81	0.90
MobileNetV2	0.96	0.81	0.94	0.97	0.81	0.95	0.98	0.79	0.90	0.98	0.79	0.88
NASNetMobile	0.93	0.73	0.81	0.96	0.75	0.82	0.97	0.75	0.80	0.97	0.75	0.81

Table 3 Classification report for testing of transfer learning in HER2 scoring using *VGG19* architecture for 100 epochs

Confusion matrix for the pre-trained *VGG19* network with added FC layers

VGG19—patch-based scoring

		Predicted Score		
		0/1+	2+	3+
Ground truth	0/1+	298 (93.1%)	22 (6.9%)	0 (0.0%)
	2+	19 (7.9%)	221 (92.1%)	0 (0.0%)
	3+	5 (2.1%)	14 (5.8%)	221 (92.1%)

Classification report for the pre-trained *VGG19* network with added FC layers

VGG19—patch-based scoring

Class	Precision	Recall	F1-score	Support
0/1+	0.93	0.93	0.93	320
2+	0.86	0.92	0.89	240
3+	1.00	0.92	0.96	240
Accuracy			0.93	800
Macro accuracy	0.93	0.92	0.93	800
Weighted accuracy	0.93	0.93	0.93	800

VGG19 – Image based scoring with voting

		Predicted score		
		0/1+	2+	3+
Ground truth	0/1+	40 (100.0%)	0 (0.0%)	0 (0.0%)
	2+	1 (3.3%)	29 (96.7%)	0 (0.0%)
	3+	0 (0.0%)	1 (3.3%)	29 (96.7%)

VGG19—image-based scoring with voting

Class	Precision	Recall	F1-Score	Support
0/1+	0.98	1.00	0.99	40
2+	0.97	0.97	0.97	30
3+	1.00	0.97	0.98	30
			0.98	100
Macro accuracy	0.98	0.98	0.98	100
Weighted accuracy	0.98	0.98	0.98	100

is labelled the same irrespective of the actual label. This is the reason for lower accuracy in image patch-based classification. On applying the statistical voting scheme to classify the image-based accuracy from the eight patch-based classification scores, the overall accuracy increases as shown in Table 3 by confusion matrix and classification report. The *VGG* networks are showing better testing accuracy per set of epochs; their performance is much higher than other three networks. *VGG19* with 100 epochs is showing the best performance in test data with 93% accuracy in patch-based scoring and this accuracy increases to 98% in case of overall image-based scoring. It can be observed that the test accuracy has improved for all the five classifiers. The time and space requirement for the training are shown in Table 4.

For the 100 epochs, the transfer learning using *VGG19* takes about 42 min. Though the model size and weights are large, the performance of *VGG* networks is better than other networks for the HER2 Challenge dataset. A comparative assessment is performed and the accuracy is very high. Further, it can be noted that the approach is fully automatic. The user needs to feed the image and the approach automatically generates score for the 8 sub-patches followed by final scoring. To test the significance, one-way ANOVA (analysis of variance) was performed between patch-based score and image-based score showing p value 0.04, i.e. less than 0.05. This suggests that there is a significant difference between the outcome of patch-based and image-based scores and the proposed approach is significant. The proposed approach is also compared with other deep learning-based approaches for a qualitative assessment of HER2 in small patches cropped from whole slide images. The comparison is shown in Table 5. The reported literature shows a qualitative comparison on HER2 scoring and the proposed approach is meeting the standard. It should be noted that these methods are on the same benchmark data. Also, there is no step for segmentation of nuclei and membrane. The approach required the input image as a whole and the output is the overall HER2 score.

In this work, there is no similarity in terms of training data preparation as compared with other reported literature. The training data and testing data are generated from the

Table 4 Training parameters for transfer learning networks in HER2 scoring

Transfer learning pre-trained model with FC layers	Model size and weights (in megabytes, MB)	Training time for 100 epochs
VGG16	632 MB, 552 MB	40.84 min
VGG19	652 MB, 572 MB	41.38 min
ResNet50	121 MB, 100 MB	45.24 min
MobileNetV2	21.1 MB, 12.9 MB	45.47 min
NASNetMobile	28.1 MB, 21.7 MB	68.51 min

Table 5 Comparison of proposed approach with state-of-the art methods applied for HER2 scoring

Ref.	Dataset used	Methods/features	Classifiers	Accuracy
Pitkäaho, Lehtimäki et al.	119 core regions extracted from 81 WSIs	Data augmentation, block-based Scoring by CNN	CNN, AlexNet architecture	Accuracy 97.7%
Singh and Mukundan	1345 image patches from 52 WSIs	Intensity and colour features ULBP	Neural network classifier	Accuracy 91.1%
Cordeiro, Ioshii et al.	2580 patches from 86 WSIs	Image patch level and patient level scoring with colour and texture features	SVM, KNN, MLP and decision tree were compared	Best Accuracy 94.2%
Saha and Chakraborty	752 images cropped from 79 WSIs	Image patch-based nuclei detection and cell membrane extraction	Her2net—LSTM recurrent network	Segmentation and classification Accuracy 98.33%
Mukundan	4019 image patches from 52 WSIs	Characteristic curves, ULBP connectedness, entropy	Logistic regression, SVM	Average 91% Maximum 93%
Khameneh, Razavi et al.	127 WSIs	Super-pixel for tissue region extraction Colour and texture features	Modified Unet architecture for tissue classification	Classification Accuracy 87%
Proposed approach—*transfer learning and statistical voting*	2130 training patches and 800 test patches from 40 WSIs	Image patch-based labelling using transfer learning followed by statistical mode for scoring	*VGG19* Architecture followed by fully connected dense layers for 3 class	Test data of 100 images Patch-based accuracy 93% Image-based accuracy 98%

different cases of the HER2 Challenge dataset. The ground truth used in this work is as per the training data provided in the dataset. The score was obtained from hospital records which has at least two experts score as per the routine practice. There was no manual marking over the images in the dataset, only the overall score was provided and the same was used in this study. In future, if the scores can be made available for multiple pathologists and the regions are marked, the results could be improved.

The comparative assessment suggests that the proposed approach of *VGG19* pre-trained model with statistical voting is able to classify the HER2-stained tissue images accurately. The accuracy is competing the state-of-the-art methods reported in the literature. Further, in future, it will be targeted to apply the deep learning to develop a comprehensive software with user-specified training and classification system.

Conclusion

Computer-aided automated image classification can aid the pathologists in biomedical image analysis. An automated HER2 quantification using transfer learning followed by statistical voting is presented. Deep learning provides a better abstract representation compared to conventional machine learning algorithms. Five pre-trained deep learning architectures have been used for this study. In the training, the benchmark deep learning architectures with added fully connected layers have shown significant training accuracy. The training time is also very less for the amount of data with augmentation. On the test dataset of 100 images, the VGG19 has shown best accuracy in both patch level and whole image level accuracy. Also, the statistical voting has improved the accuracy for VGG19 pre-trained model from 93 to 98% using the proposed approach of decision level fusion.

Acknowledgements The authors would like to thank the HER2 Challenge research group in the Department of Computer Science, University of Warwick, UK for the access to Challenge dataset. Suman Tewary is grateful to Director, CSIR-Central Scientific Instruments Organisation, Chandigarh for providing the research facility.

References

1. Ma J, Jemal A: Breast cancer statistics: Springer, 2013
2. Mosquera-Lopez C, Agaian S, Velez-Hoyos A, Thompson I, et al: Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. IEEE reviews in biomedical engineering 8:98-113, 2014
3. Joensuu K, Leidenius M, Kero M, Andersson LC, Horwitz KB, Heikkilä P, et al: ER, PR, HER2, Ki-67 and CK5 in early and late relapsing breast cancer--reduced CK5 expression in metastases. Breast cancer: basic and clinical research 7:23, 2013
4. Wolff AC, et al: Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. Archives of Pathology and Laboratory Medicine 138:241-256, 2014

5. Rakha EA, et al: Updated UK Recommendations for HER2 assessment in breast cancer. Journal of clinical pathology 68:93-99, 2015
6. Nitta H, et al: The assessment of HER2 status in breast cancer: the past, the present, and the future. Pathology international 66:313-324, 2016
7. Qaiser T, et al: Her 2 challenge contest: a detailed assessment of automated her 2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 72:227-238, 2018
8. Cordeiro CQ, Ioshii SO, Alves JH, Oliveira LF et al: An Automatic Patch-based Approach for HER-2 Scoring in Immunohistochemical Breast Cancer Images Using Color Features. arXiv preprint https://arxiv.org/1805.05392, 2018
9. Jeung J, Patel R, Vila L, Wakefield D, Liu C et al: Quantitation of HER2/neu expression in primary gastroesophageal adenocarcinomas using conventional light microscopy and quantitative image analysis. Archives of pathology & laboratory medicine 136:610-617, 2012
10. Brügmann A, et al: Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. Breast cancer research and treatment 132:41-49, 2012
11. Dobson L, et al: Image analysis as an adjunct to manual HER-2 immunohistochemical review: a diagnostic tool to standardize interpretation. Histopathology 57:27-38, 2010
12. Ruifrok AC, Johnston DA: Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23:291-299, 2001
13. Tuominen VJ, Tolonen TT, Isola J et al: ImmunoMembrane: a publicly available web application for digital image analysis of HER2 immunohistochemistry. Histopathology 60:758-767, 2012
14. Tabakov M, Kozak P: Segmentation of histopathology HER2/neu images with fuzzy decision tree and Takagi–Sugeno reasoning. Computers in biology and medicine 49:19-29, 2014
15. Pham N-A, et al: Quantitative image analysis of immunohistochemical stains using a CMYK color model. Diagnostic pathology 2:1, 2007
16. Wdowiak M, Markiewicz T, Osowski S, Swiderska Z, Patera J, Kozlowski W, et al: Hourglass shapes in rank grey-level hit-or-miss transform for membrane segmentation in HER2/neu images. Proc. International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing: City
17. Gavrielides MA, Masmoudi H, Petrick N, Myers KJ, Hewitt SM, et al: Automated evaluation of HER-2/neu immunohistochemical expression in breast cancer using digital microscopy. Proc. 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro: City
18. Wan T, Cao J, Chen J, Qin Z, et al: Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. Neurocomputing 229:34-44, 2017
19. Mukundan R: Analysis of Image Feature Characteristics for Automated Scoring of HER2 in Histology Slides. Journal of Imaging 5:35, 2019
20. Tewary S, Arun I, Ahmed R, Chatterjee S, Mukhopadhyay S, et al: AutoIHC-Analyzer: computer-assisted microscopy for automated membrane extraction/scoring in HER2 molecular markers. Journal of Microscopy 281:87-96, 2021
21. LeCun Y, Bengio Y, Hinton G, et al: Deep learning. nature 521:436-444, 2015
22. Huang Y, Zheng H, Liu C, Ding X, Rohde GK, et al: Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE journal of biomedical and health informatics 21:1625-1632, 2017
23. Meng N, Lam EY, Tsia KK, So HK-H, et al: Large-scale multi-class image-based cell classification with deep learning. IEEE journal of biomedical and health informatics 23:2091-2098, 2018
24. Pitkäaho T, Lehtimäki TM, McDonald J, Naughton TJ, et al: Classifying HER2 breast cancer cell samples using deep learning. Proc. Proc Irish Mach Vis Image Process Conf: City
25. Vandenberghe ME, Scott ML, Scorer PW, Söderberg M, Balcerzak D, Barker C, et al: Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Scientific reports 7:45938, 2017
26. Saha M, Chakraborty C: Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Transactions on Image Processing 27:2189-2200, 2018
27. Khameneh FD, Razavi S, Kamasak M, et al: Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. Computers in biology and medicine, 2019
28. Qaiser T, Rajpoot NM: Learning where to see: A novel attention model for automated immunohistochemical scoring. IEEE transactions on medical imaging 38:2620-2631, 2019
29. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L, et al: Imagenet: A large-scale hierarchical image database. Proc. IEEE conference on computer vision and pattern recognition: City,2009
30. Huh M, Agrawal P, Efros AA, et al: What makes ImageNet good for transfer learning? arXiv preprint https://arxiv.org/1608.08614, 2016
31. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I, et al: Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine 27:317-328, 2018
32. Cheng PM, Malhi HS: Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. Journal of digital imaging 30:234-243, 2017
33. He K, Zhang X, Ren S, Sun J, et al: Deep residual learning for image recognition. Proc. Proceedings of the IEEE conference on computer vision and pattern recognition: City
34. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C, et al: Mobilenetv2: Inverted residuals and linear bottlenecks. Proc. Proceedings of the IEEE conference on computer vision and pattern recognition: City
35. Zoph B, Vasudevan V, Shlens J, Le QV, et al: Learning transferable architectures for scalable image recognition. Proc. Proceedings of the IEEE conference on computer vision and pattern recognition: City
36. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. arXiv preprint https://arxiv.org/1409.1556, 2014
37. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J, et al: Mitosis detection in breast cancer histology images with deep neural networks. Proc. International conference on medical image computing and computer-assisted intervention: City
38. Khan S, Islam N, Jan Z, Din IU, Rodrigues JJC, et al: A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters 125:1-6, 2019
39. Sokolova M, Lapalme G: A systematic analysis of performance measures for classification tasks. Information processing & management 45:427-437, 2009