Data Fusion of Histological and Immunohistochemical Image Data for Breast Cancer Diagnostics using Transfer Learning

Pranita Pradhan[1,2][a], Katharina Köhler[3,4], Shuxia Guo[1,2][b], Olga Rosin[3,4], Jürgen Popp[1,2][c], Axel Niendorf[3,4] and Thomas Wilhelm Bocklitz*[,1,2][d]

[1]*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University, Helmholtzweg 4, Jena, 07743, Thüringen, Germany*

[2]*Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, Jena, 07745, Thüringen, Germany*

[3]*MVZ Prof. Dr. med. A. Niendorf Pathologie Hamburg-West GmbH, Lornsenstraße 4-6, Hamburg, 22767, Hamburg, Germany*

[4]*Institute for Histology, Cytology and Molecular Diagnostics,Lornsenstraße 4, Hamburg, 22767, Hamburg, Germany*

Keywords: Breast Cancer, Transfer Learning, Histology, Immunohistochemistry.

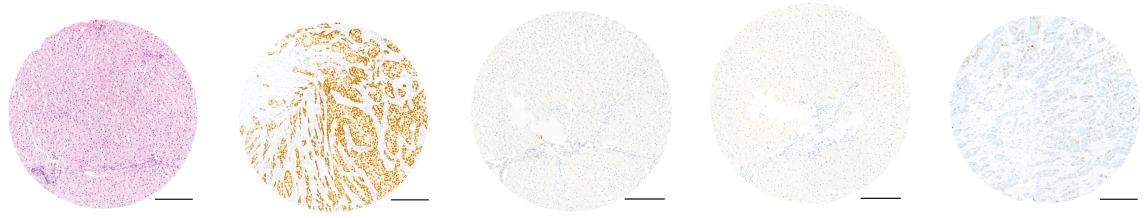
Abstract: A combination of histological and immunohistochemical tissue features can offer better breast cancer diagnosis as compared to histological tissue features alone. However, manual identification of histological and immunohistochemical tissue features for cancerous and healthy tissue requires an enormous human effort which delays the breast cancer diagnosis. In this paper, breast cancer detection using the fusion of histological (H&E) and immunohistochemical (PR, ER, Her2 and Ki-67) imaging data based on deep convolutional neural networks (DCNN) was performed. DCNNs, including the VGG network, the residual network and the inception network were comparatively studied. The three DCNNs were trained using two transfer learning strategies. In transfer learning strategy 1, a pre-trained DCNN was used to extract features from the images of five stain types. In transfer learning strategy 2, the images of the five stain types were used as inputs to a pre-trained multi-input DCNN, and the last layer of the multi-input DCNN was optimized. The results showed that data fusion of H&E and IHC imaging data could increase the mean sensitivity at least by 2% depending on the DCNN model and the transfer learning strategy. Specifically, the pre-trained inception and residual networks with transfer learning strategy 1 achieved the best breast cancer detection.

1 INTRODUCTION

Breast cancer is one of the most prevalent cancers among women. It is diagnosed by a routine procedure which is based on morphological tissue features in hematoxylin and eosin (H&E) stained tissue sections (figure 1a). The morphological tissue features include tumour size and type, which are regularly documented to assess the histological grade of breast cancer tissue (Webster et al., 2005). These morphological tissue features are also used to prevent recurrence risk of breast cancer and prescribe personalized therapies. Breast cancer is additionally verified by other staining technique called the immunohistochemical (IHC) staining technique. The IHC staining technique uses antibodies to highlight specific antigens in the tissue region (Veta et al., 2014), and includes estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor-2 (Her2) (figure 1b-d). Studies have shown that the IHC examination with ER, PR, Her2 and Ki-67 can detect five molecular breast cancer sub-types to provide adequate personalized therapies (Perou et al., 2000; Sørlie et al., 2001; Cheang et al., 2009). However, none of the studies report a combination of histology (H&E) and IHC staining techniques (ER, PR, Her2 and Ki-67) for breast cancer diagnosis. Therefore, in this work, an integration of IHC imaging technique i.e. hormone receptors including ER, PR, Her2 and Ki-67 nuclear protein stained images with H&E stained images is proposed to gain new insights into breast cancer biology (Elledge et al., 2000; Damodaran and Olson, 2012). The combination of histology and IHC stain-

[a] https://orcid.org/0000-0002-0558-2914
[b] https://orcid.org/0000-0001-8237-8936
[c] https://orcid.org/0000-0003-4257-593X
[d] https://orcid.org/0000-0003-2778-6624

Pradhan, P., Köhler, K., Guo, S., Rosin, O., Popp, J., Niendorf, A. and Bocklitz, T.
Data Fusion of Histological and Immunohistochemical Image Data for Breast Cancer Diagnostics using Transfer Learning.
DOI: 10.5220/0010225504950506
In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2021)*, pages 495-506
ISBN: 978-989-758-486-2



(a) (b) (c) (d) (e)

Figure 1: Five stain type images. (a) Hematoxylin and eosin (H&E), (b) Estrogen receptor (ER), (c) Progesterone receptor (PR), (d) Human epidermal growth factor-2 (Her2) and (e) Ki-67 protein are shown. Scale bar is 200 μm.

ing technique is referred to as 'Data fusion' approach.

Data fusion approach by combining the histological and IHC stained images can provide various tissue features associated with the disease stage and relapse of breast cancer. However, visual inspection of all five stained images is a tedious process which can prolong the diagnosis. Therefore, automation of breast cancer detection based on the combination of histological and IHC imaging data is needed. In this regard, researchers (Pham et al., 2007; Dobson et al., 2010) used computer-assisted image analysis techniques to automatically monitor changes in the tissue features of histological and IHC stained images separately. However, computer-assisted image analysis can be limited due to the need for specific software systems or the need for user-specific input to analyze the images. This slows down the process of analyzing images and providing personalized therapies to the patients. To increase the analysis speed and reduce human intervention, this work proposes machine learning (ML) instead of computer-assisted image analysis techniques.

Conventional ML methods can automatize breast cancer detection based on the fusion of histological and IHC imaging data in the following way. First, the features (e.g. color, shape and texture features) from the five stain type of imaging data (H&E, ER, PR, Her2 and Ki-67) can be extracted using image analysis methods. The feature extraction step in the conventional ML method is subjective and requires the effort of an image analyst. Based on the extracted features, a classification, or a regression model can be constructed. Subsequently, the classification or the regression model can be used to make 'predictions' (i.e. to predict a class like tumour or normal) on a new or unseen dataset. Thus, the extracted features affect the predictions made by the ML model. However, recently developed ML methods are capable of performing automatic feature extraction for classification or regression purpose. These self-learning methods are categorized into a broad family of ML called 'Deep learning' (DL). The DL models can have many types of network architectures. Widely used DL model for images is the deep convolutional neural network (DCNN) and its numerous applications are reported in the field of digital pathology (Liu et al., 2017); for example, cell segmentation or detection (Chen and Chefd'Hotel, 2014), tumour classification (Cireşan et al., 2013; Wang et al., 2016) and carcinoma localization (Janowczyk and Madabhushi, 2016; Coudray et al., 2018; Khosravi et al., 2018; Sheikhzadeh et al., 2018). Nevertheless, a bottleneck for DL models is the requirement of huge dataset during training, which is difficult to acquire, particularly in the medical imaging field. In such cases, 'transfer learning' methods for DCNNs can be applied for improving the model performance (Tajbakhsh et al., 2016).

Transfer learning is the transfer of knowledge learned on a source task using a source dataset to improve the performance on a target task using the target dataset (Torrey and Shavlik, 2010). Transfer learning using any DL model like DCNN can be performed by three strategies. First, a pre-trained DCNN can be used as a feature extractor. In this strategy, features for the target dataset are extracted using a DCNN trained on different or similar source dataset. The second strategy is fine-tuning the weights of the last layers of a pre-trained DCNN, and the third strategy is fine-tuning the weights of all layers of a pre-trained DCNN. In the second and third fine-tuning strategies, the weights of specific layers of a DCNN trained on a source dataset are further optimized based on the target dataset. The three transfer learning strategies like using a DCNN as a feature extractor or fine-tuning of a DCNN, requires adequate knowledge of the size and type of the source and the target dataset (Pan and Yang, 2010). Transfer learning, if used appropriately, can improve the initial and final performance of the DL model on the target dataset. It can also reduce the total training time of the DL model on the target dataset. Different transfer learning strategies acquire different results based on the source and target dataset which is evident in the next section.

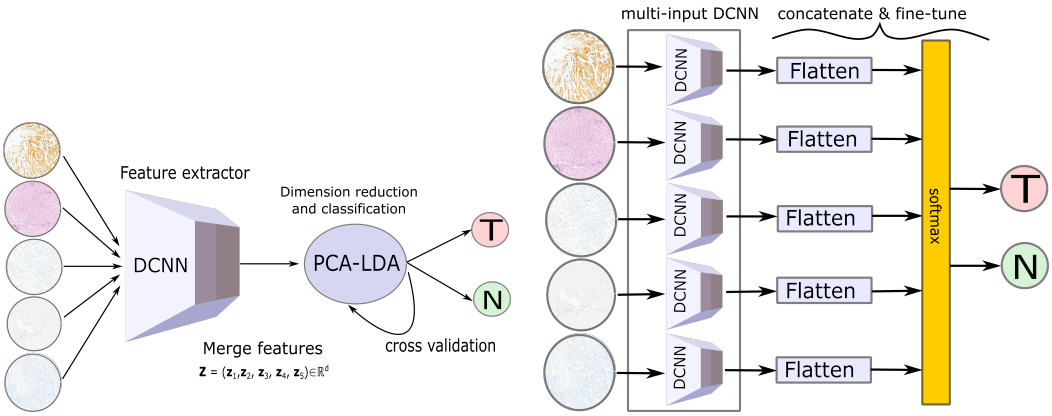


Figure 2: (a) visualizes transfer learning strategy 1 for data fusion approach where a pre-trained DCNN is used as a feature extractor. The features extracted from a pre-trained DCNN for all five stain type images are merged and classified into tumour and normal using the PCA-LDA model. (b) shows transfer learning strategy 2 for data fusion approach where fine-tuning of the last layer of a pre-trained multi-input DCNN is performed. The five DCNNs are pre-trained models like the VGG16, the Inceptionv3 or the ResNet50, each having a stain type as its input.

2 RELATED WORK

Transfer learning in medical imaging can be achieved by training a DCNN on a large medical or non-medical dataset, and transferring its knowledge to the target medical dataset (Bayramoglu and Heikkilä, 2016; Tajbakhsh et al., 2016). A recent study used a large non-medical dataset like the ImageNet dataset (Russakovsky et al., 2015) to pre-train a DCNN and transfer its off-the-shelf features to investigate two computer-aided detection (CADs) problems namely thoracoabdominal lymph node detection and interstitial lung disease detection (Shin et al., 2016). In their work, three different DCNNs including the CifarNet (Krizhevsky and Hinton, 2009), the AlexNet (Krizhevsky et al., 2012) and the GoogleNet (Szegedy et al., 2015) were evaluated with three transfer learning strategies. Similarly, a recent publication (Mormont et al., 2018) compared various transfer learning strategies based on pre-trained DCNNs using eight classification datasets in digital pathology. Their results showed that fine-tuning the ResNet (He et al., 2016) and the DenseNet (Huang et al., 2017) models outperformed the other tested models in the morphological classification task. Similar findings were observed in other references (Antony et al., 2016; Kieffer et al., 2017; Ravishankar et al., 2016).

In contrast to the previously mentioned applications where fine-tuning of a DCNN achieved the best performance, several other applications using a DCNN as feature extractor achieved significant performance on binary and multi-class classification tasks. These applications included prediction of morphological changes in cells in microscopic images (Kensert et al., 2018), classification of colon polyps in endoscopic images (Ribeiro et al., 2016), identification of mammographic tumours (Huynh et al., 2016) and detection of pulmonary nodules in computed tomography scans (Van Ginneken et al., 2015). It is clear from the previous researches that transfer learning techniques are data-dependent, and a generalization of the above-mentioned results is not feasible, especially in the medical imaging field (Litjens et al., 2017). Therefore, no consensus of the proper application of transfer learning in the medical imaging field is established. Likewise, the application of transfer learning, especially for medical imaging data requires utmost care and further investigations.

In this contribution, data fusion of histological and immunohistochemical imaging data for classifying breast cancer is presented for the first time. Due to our small dataset size, the classification task is performed using two transfer learning strategies. From previous experience, the third transfer learning strategy i.e. the training of a DCNN from scratch is avoided, as it is computationally expensive and may lead to overfitting in the absence of large datasets. The performance of the two transfer learning strategies for the data fusion approach is compared with histological imaging data. Moreover, the two transfer learning strategies are performed using three pre-trained DCNN models like the VGG16 (He et al., 2016), the Inceptionv3 (Szegedy et al., 2016) and the ResNet50 network (Simonyan and Zisserman, 2014). The goal of this study was to verify whether the data fusion approach along with

transfer learning improves the breast cancer diagnosis based on the sensitivity and F1 score metric.

3 MATERIAL AND METHODS

3.1 Sample Preparation

A Tissue Microarray (TMA) with 97 cores representing 23 breast cancer cases (78 tumour cores, 18 non-cancerous tissue cores or the normal breast tissue and one control core of liver tissue) was produced using the Manual Tissue Arrayer MTA-1 by Estigen. The cases were randomly selected out of the daily routine of MVZ Prof. Dr. med. A. Niendorf Pathologie Hamburg-West GmbH and anonymized according to a statement of the ethics committee of the Hamburg Medical Chamber. Core tissue biopsies (1.0 mm in diameter) were taken from individual FFPE (formalin-fixed paraffin-embedded) blocks and arranged within a new recipient block. From the block, 2 μm sections were cut, placed on glass microscope slides and H&E staining (figure 1a) following a standard protocol was performed. Digital images of histology (H&E) slides were obtained at 40$\times$ magnification using the 3DHistech Pannoramic 1000 Flash IV slide scanner with a spatial resolution of 0.24 μm/pixel (.mrxs image file). Subsequently, immunohistochemistry staining (ER, PR, Her2 and Ki-67) (figure 1b-e) was performed on super frost charged glass slides.

3.2 Image Preprocessing

For the analysis, 96 TMAs or scans (78 tumour scans and 18 normal scans) from 23 patients were used, and each TMA had five stain types (H&E, PR, ER, Her2 and Ki-67). The pixel intensity I of each TMA was standardized using a min-max scaling $(I - I_{min})/(I_{max} - I_{min})$, where I_{min} and I_{max} is the minimum and maximum intensity of a pixel in a TMA. The background pixels were cropped manually and non-overlapping patches of size 1024$\times$1024 were extracted from a standardized TMA. This led to 9 patches per TMA (702 tumour and 162 normal patches). The four corner patches including a large number of background pixels were removed, leading to 390 tumour and 90 normal patches. Based on the 480 selected patches, three pre-trained models were used with two transfer learning strategies.

3.3 DCNN Architectures

To check the robustness of the data fusion approach, three DCNNs: the VGG network, the Inception network and the residual network, with unique architectures were chosen. The VGG network is a DCNN that has acquired state-of-the-art performances for image classification tasks. However, the VGG network can exhibit the problem of vanishing gradients with an increasing number of layers (Hanin, 2018). Thus, the residual network which can solve the problem of vanishing gradients by adding the 'shortcut connections' was explored in this work. Furthermore, the inception network that provides width in addition to the depth to a conventional DCNN was utilized. A detailed explanation of the architecture of the three models is given further.

3.3.1 VGG Network

A VGG network is a DCNN with different configurations from 11 to 16 convolutional layers followed by three fully connected layers. The number of convolutional layers increases the depth of the VGG network. It is shown that an increase in the depth of the VGG network decreases the top-5 validation error (He et al., 2016). However, the decrease in the error for the VGG network from 16 to 19 convolutional layers is not significant. Thus, the VGG network with 16 convolutional layers referred to as VGG16 from Keras was used (Chollet et al., 2015). The input to the VGG16 network was an RGB image of size 224$\times$224, and each image was preprocessed by subtracting the mean RGB values computed over the training dataset.

3.3.2 Inception Network

Deep networks like VGG network require an appropriate selection of the number of convolution filters and filter sizes. For this reason, the inception network concatenates convolutional layers of different filter size, including the spatial dimension of 1$\times$1, 3$\times$3 and 5$\times$5. This captures information at various scales while increasing the computational complexity. In order to reduce the computational cost, a convolutional layer of 1$\times$1 filter size is applied before each convolutional layer of filter size 3$\times$3 and 5$\times$5. These two salient features of the Inception network reduce the dimensionality in the feature space and thereby allows the network to be deeper and wider. Moreover, the inception network replaces the fully connected layer with global averaging layers which reduces the number of trainable weights, thus reducing over-fitting on the training dataset (Szegedy et al., 2016). The Inceptionv3 implementation from Keras, which has 95 layers and requires an RGB image as input with size 299$\times$299 was used.

Table 1: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual networks using transfer learning strategy 1. Here, two feature sets extracted from pre-trained models are used; one feature set is extracted from H&E images only, while the other feature set is extracted from all the five stain types. All metrics are computed for 96 TMAs by taking majority voting of the predictions acquired for the patches using the PCA-LDA model. N: normal scans, T: tumour scans.

Data fusion (H&E+IHC imaging data)						Only histological imaging data					
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N T	13 11	5 67	79.06	76.24	VGG16	N T	14 13	4 65	80.56	76.61
Inceptionv3	N T	16 8	2 70	89.32	85.47	Inceptionv3	N T	15 5	3 75	88.46	86.97
ResNet50	N T	14 3	4 75	86.97	87.80	ResNet50	N T	14 5	4 73	85.68	84.96

Table 2: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual networks using transfer learning strategy 2. Data fusion approach used multi-input DCNN with the five stain type images as input, whereas a single-input DCNN was used only the H&E image as input. The last layers of both single-input and multi-input DCNNs were fine-tuned. The mean sensitivities are computed for 96 TMAs by taking majority voting of the predictions obtained for the patches. N: normal scans, T: tumour scans.

Data fusion (H&E+IHC imaging data)						Only histological imaging data					
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N T	7 4	11 74	66.88	70.86	VGG16	N T	3 5	15 73	55.13	57.57
Inceptionv3	N T	0 0	18 78	50.00	44.83	Inceptionv3	N T	9 4	9 74	72.44	75.66
ResNet50	N T	0 0	18 78	50.00	44.83	ResNet50	N T	12 3	6 75	81.41	83.78

Table 3: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual network using the two transfer learning strategies. All metrics are computed for 96 TMAs by taking majority voting of the predictions acquired by the models for patches.

Transfer learning strategy 1						Transfer learning strategy 2					
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N T	13 11	5 67	79.06	76.24	VGG16	N T	7 4	11 74	66.88	70.86
Inceptionv3	N T	16 8	2 70	89.32	85.47	Inceptionv3	N T	0 0	18 78	50.00	44.83
ResNet50	N T	14 3	4 75	86.97	87.80	ResNet50	N T	0 0	18 78	50.00	44.83

3.3.3 Residual Network

The configurations of the VGG network show that deep neural networks achieve good top-5 accuracy until a certain depth limit (He et al., 2016). An increase in the network depth causes a problem of vanishing or exploding gradients (Hanin, 2018) which affects the network convergence and degrades the performance (Simonyan and Zisserman, 2014). Therefore, the residual networks are built to solve this degradation problem by adding activations of the top layers into the deeper layers of the network. For instance, in a deep neural network the activation a of the $(l+2)^{th}$ layer with weight w and bias b is given as

$$a_{(l+2)} = f[(w_{(l+2)} \times a_{(l+1)}) + b_{(l+2)}], \quad (1)$$

where f is an activation function like linear rectified unit ($f = \max(a_{(l+2)}, 0)$). However, in a residual block the activation a of the l^{th} layer (or an identity mapping) is added via the 'skip or shortcut connections' (Bishop et al., 1995; Venables and Ripley, 2013) to the $(l+2)^{th}$ layer of the network. Therefore, the activation of the $(l+2)^{th}$ layer in a residual block can be given a

$$a_{(l+2)} = f[(w_{(l+2)} \times a_{(l+1)}) + b_{(l+2)} + a_{(l)}]. \quad (2)$$

This implies that in worse cases when the network

fails to learn representative features, i.e. $w_{(l+2)} = 0$ and $b_{(l+2)} = 0$, the output still remains an identity mapping of the input a_l. In residual networks, a series of residual blocks along with intermediate normalization layers was used; thus improving the learning of the deep neural networks. In this work, the ResNet50 implementation from Keras, which has 152 layers and requires an RGB image as an input with size 224×224, was used.

The above explained three DCNN models were trained using two transfer learning strategies which are discussed in the next section.

3.4 Transfer Learning Strategies

The above-mentioned DCNNs were utilized for two transfer learning strategies. For the first strategy, a pre-trained DCNN model to extract off-the-shelf features followed by a linear classifier was used. In the second strategy, a multi-input pre-trained DCNN model followed by a softmax classifier was used. Both strategies were performed on a commercially available PC system intel® Core™ with NVIDIA GeForce GTX 1060, 6GB with python packages: Keras(Chollet et al., 2015), Tensorflow(Abadi et al., 2015), Scikit-learn (Pedregosa et al., 2011), Scipy (Jones et al., 2001) and Numpy (Oliphant, 2006).

3.4.1 DCNN as Feature Extractor

In the first strategy (figure 2a), features $\mathbf{z}_i \in \mathbb{R}^m, i = (1,2,3,4,5)$ were extracted for patches of each stain type i using the pre-trained VGG16, Inceptionv3 and ResNet50 networks. The patches were resized according to the model's input size requirement. For a patch of a single stain type, 25,088 features were extracted by the VGG16 (feature shape: 1, 7, 7, 512), 51,200 features were calculated by the Inceptionv3 (feature shape: 1, 5, 5, 2048) and 2048 features were obtained by the ResNet50 (feature shape: 1, 1, 1, 2048). For data fusion approach, the features from all five stain types were concatenated, $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5) \in \mathbb{R}^d \ (d \gg m)$ resulting in ∼0.12 million features by the VGG16 model, ∼0.25 million features by the Inceptionv3 model and 10,240 features by the ResNet50 model per patch. For histological imaging data, i.e. without the data fusion approach, the features extracted only from the H&E images were used. In both cases, the large feature dimension of each patch was reduced by principal component analysis (PCA) model, and classified as normal or tumour using linear discriminant analysis (LDA) model (Hastie et al., 2009). The PCA-LDA model was evaluated using internal and external cross-validation scheme explained elsewhere (Guo et al., 2017). Shortly, the internal cross-validation was used to optimize the number of PC's of the PCA-LDA model. The external cross-validation was used to predict an independent test dataset based on the PCA-LDA model. The external cross-validation used leave-one-patient-out cross-validation, such that the patches acquired from TMAs of 23 patients were used at least once as an independent test dataset. The internal cross-validation used 10 fold cross-validation. The predictions by the PCA-LDA model acquired for the patches from the external cross-validation step were voted to assign each TMA into a tumour or normal class. Based on the predicted TMA labels (obtained after majority voting of the patches) and true TMA labels, metrics like confusion matrix, mean sensitivity and mean F1 score were reported. The mean sensitivity and the mean F1 score were calculated using an average of the mean sensitivities and the mean F1 scores for the tumour and normal class, respectively. Lastly, the transfer learning strategy 1 was performed for all the three DCNNs and their classification performance based on TMAs was compared.

3.4.2 Fine-tuning of DCNN

In the second strategy (figure 2b), for histological imaging data, a single-input DCNN was used; whereas for the data fusion approach, a multi-input DCNN was used. The multi-input DCNN model $\mathcal{N}$ was constructed using five pre-trained models of the same architecture; for instance, five pre-trained VGG networks each using a stain type image as an input. The input to the multi-input DCNN model was the five stained images (H&E, ER, Her2, Ki-67 and PR). The last layer of the multi-input DCNN models was concatenated and followed by a dense layer with two outputs (corresponding to the normal and tumour class) with a softmax activation layer. The softmax activation layer mapped the non-normalized output of the model $\mathcal{N}$ to the distribution of K probabilities and is defined as

$$P(\mathbf{r})_i = \frac{\exp(r_i)}{\sum_{j=1}^{K} \exp(r_j)}, \tag{3}$$

where $\mathbf{r} = (r_1, \cdots, r_K)$ and $K = 2$ for a binary classification task. During the training process, the last two layers were fine-tuned using Adam optimizer (Kingma and Ba, 2014) with a learning rate 0.001 and mini-batch size of 5 patches. To allocate higher class weight for the minority class (here, the normal class), the weighted binary cross-entropy loss function

$$\mathcal{L} = -\sum_{i}^{K} \alpha_i y_i log(P(\mathbf{r})_i) \tag{4}$$

was used, where $\alpha_i = \frac{1}{\#K_i}$, y_i, $P(\mathbf{r})_i$ are the weight, ground truth and the probability from the softmax activation layer of the i^{th} class in K, respectively. The model was evaluated using the mean sensitivity and the mean F1 score similar to transfer learning strategy 1.

For the evaluation of the single and multi-input DCNN, the dataset was divided into three parts: training, validation and testing. In every iteration, patches of one patient were used as an independent test dataset and the patches of remaining patients were used as training and validation dataset. To avoid any training bias, the training and validation datasets were randomly split patient-wise such that patches from 30% patients were used as validation dataset and the rest as the training dataset. In other words, during each iteration, patches of one patient were used as the test dataset, patches of 16 patients formed the training dataset and patches of remaining 6 patients belonged to the validation dataset. The combination of 16 and 6 patients in training and validation datasets were chosen randomly. The iterations were repeated until all 23 patients were used as an independent test dataset. Further, every iteration was executed for ten epochs, and validation sensitivity was monitored for early stopping of the model training. The model with best validation sensitivity was used for predicting the independent test dataset in that iteration. In this way, the patches of all 23 patients were used individually as an independent test dataset, and majority voting of the patches similar to transfer learning strategy 1 was performed. The confusion matrices and average of the mean sensitivities for the normal and tumour classes were evaluated using the independent test dataset. Subsequently, transfer learning strategy 2 was performed for all the three pre-trained DCNN models with the same hyper-parameter setting.

3.4.3 ROC Curve Analysis for TMAs

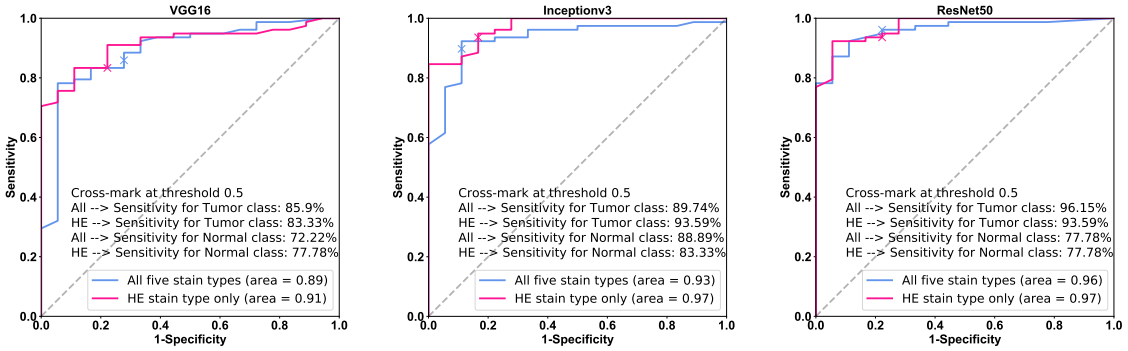
The results of the two transfer learning strategies were obtained as ROC curves showing the true and the false positive rate for the tumour class. The ROC curves were evaluated for TMAs based on the majority voting of the selected patches. To achieve ROC curves for TMAs, the model output in the form of probabilities of each patch for the tumour class was thresholded using 100 different values in the range [0, 1]. This led to predictions for patches with different threshold values. Subsequently, the predictions for patches obtained for each threshold value were majority voted to obtain a prediction for a TMA. The predictions for TMAs were used to calculate the true positive rate, the false positive rate and the ROC curve, as shown in figure 3 and 4. The predictions for the TMAs obtained with 0.5 threshold were used to obtain the confusion matrix, mean sensitivities and mean F1 scores as reported in table 1, 2 and 3.

4 RESULTS

The main aim of this work was to confirm that the data fusion approach can achieve better breast cancer diagnosis than histological imaging data based on performance metrics. This was confirmed by one of the two transfer learning strategies. The results are divided in three parts as shown in table 1, 2 and 3. Table 1 and 2 report performance metrics obtained for transfer learning strategy 1 and transfer learning strategy 2, with and without data fusion approach, respectively. Table 3 shows a comparison of the two transfer learning strategies using only the data fusion approach. In table 1, 2 and 3 report values for the VGG16, the Inceptionv3 and the ResNet50 models. These values were evaluated for 96 TMAs acquired by majority voting of the five patches extracted from each TMA.

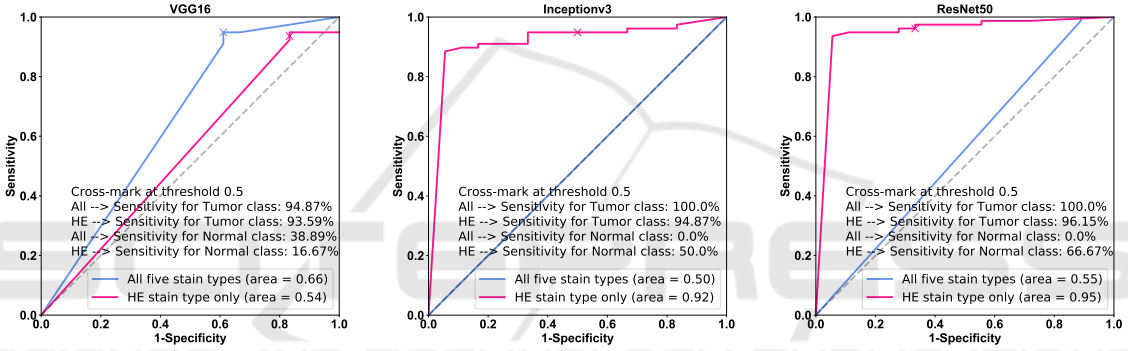
The results in table 1 show that the pre-trained features acquired from the data fusion approach yield slightly higher mean sensitivities and mean F1 scores in comparison to the pre-trained features extracted from the histological imaging data. Higher mean sensitivities using the data fusion approach were seen for at least two of the three DCNNs. Higher mean F1 score using the data fusion approach was seen only for the ResNet50 model. Specifically, the pre-trained features obtained from the data fusion approach using the Inceptionv3 and the ResNet50 models showed mean sensitivities 89.32% and 86.97%, respectively. Similarly, the mean F1 scores for the two models were 85.47% and 87.80%, respectively. In comparison, the pre-trained features from the histological data using the same DCNN model showed mean sensitivities 88.46% and 85.68%, respectively. Thus, there was approximately 2% increase in the model performance by data fusion approach based on the mean sensitivity, which is significant from a clinical perspective. However, the VGG16 model showed higher mean sensitivity (80.56%) using histological imaging data compared to the mean sensitivity calculated for the data fusion approach (79.06%). Overall, it can be seen that transfer learning using pre-trained DCNN features and a linear classification model (PCA-LDA) based on data fusion approach show a slight improvement in breast cancer detection in some cases for a small dataset as in our study.

Contrarily, table 2 obtained by the transfer learning strategy 2 shows lower mean sensitivities for



(a) (b) (c)

Figure 3: (a-c) show ROC curves for the VGG16, the Inceptionv3 and the ResNet50 networks using the transfer learning strategy 1 based on TMAs. The blue line shows ROC curve for the PCA-LDA model trained using the pre-trained DCNN features obtained from the histological and IHC imaging data, whereas the pink line shows ROC curve for the PCA-LDA model trained using pre-trained DCNN features extracted from the histological imaging data only. The cross-mark shows the true and the false positive rate on the ROC curve with 0.5 threshold.



(a) (b) (c)

Figure 4: (a-c) show ROC curves for the VGG16, the Inceptionv3 and the ResNet50 networks using the transfer learning strategy 2 based on TMAs. The blue line shows ROC curve for the multi-input DCNN model fine-tuned using the histological and IHC imaging data, whereas the pink line shows the ROC curve for the single-input DCNN model fine-tuned using only the histological imaging data. The cross-mark shows the true and the false positive rate at 0.5 threshold.

the data fusion approach in comparison to the performance achieved by using histological imaging data alone. Except for the multi-input VGG16 network, the multi-input Inceptionv3 and the multi-input ResNet50 network trained with a combination of histological and IHC imaging data predicted all normal patches as tumour patches. Thus, the multi-input Inceptionv3 and the multi-input ResNet50 model achieved mean sensitivity of 50% and mean F1 score of 44.83%; whereas, the multi-input VGG16 network showed mean sensitivity of 66.88% and mean F1 score of 70.86% for the data fusion approach. The mean sensitivity of the single-input VGG16 network declined to 55.13% when only histological imaging data was used. On the other hand, the single-input Inceptionv3 and the single-input ResNet50 models using histological imaging data showed an opposite trend with comparatively higher mean sensitivities of 72.44% and 81.41%, and higher mean F1 scores of 75.66% and 83.78%, respectively. Overall, it was observed that transfer learning performed by fine-tuning the last layer of the pre-trained multi-input DCNNs result in lower mean sensitivities for the data fusion approach. This behaviour can be a consequence of the small sample size. It is clear from the results that fine-tuning the last layer of DCNNs is not the best approach for our small breast cancer dataset. Thus, it is suspected that the fine-tuning of all the layers of a DCNN will decrease the model performance further. However, fine-tuning of all layers for large breast cancer dataset should be investigated in the future.

Lastly, the performance of the two transfer learning strategies for the data fusion approach is summarized in table 3, where higher mean sensitivities are reported for strategy 1, i.e. using pre-trained features from the VGG16, the Inceptionv3 and the ResNet50

model. The training of the PCA-LDA model based on pre-trained features of the Inceptionv3 and the ResNet50 network yield promising results. The results from the VGG16 network are lower in comparison to the other two models for transfer learning strategy 1, but higher for transfer learning strategy 2.

The performance of the two transfer learning strategies based on TMAs is summarized in the form of ROC curves in figure 3 and 4. The ROC curve calculated for the data fusion approach and histological imaging data at various thresholds is depicted in blue and pink, respectively. The AUC values given in the figure legend show lower values for the data fusion approach in comparison to the AUC values calculated using histological imaging data. This trend is observed for both the transfer learning strategies. From figure 3 and 4, it can be inferred that the overall performance of DCNN models trained using an H&E image is better for both transfer learning strategies. However, the final performance of the models in terms of mean sensitivities evaluated at 0.50 threshold is better for the data fusion approach in some cases. The mean sensitivities cross-marked in each subplot of figure 3 and 4 are calculated at 0.50 threshold coincide with the values reported in table 1, 2 and 3. These values are evaluated for TMA's by performing majority voting of the five patches in each TMA. The ROC curves at threshold 0.50 which is mostly used to evaluate the model performance, show higher mean sensitivities for data fusion approach than using histological data, at least for the Inceptionv3 and the ResNet50 model in transfer learning strategy 1 (figure 3). Nevertheless, the AUC derived from the ROC curves for transfer learning strategy 2 (figure 4) show low mean sensitivities for all the DCNN networks. The inconsistency in the results of two transfer learning strategies can be due to various reasons discussed below.

5 DISCUSSION

Based on the results, three critical findings can be discussed.

5.1 Data Fusion vs. Histological Imaging

The results showed that the data fusion approach, i.e. combining histological and IHC imaging data, increases the model performance by $\sim$2%. However, the increase in model performance was achieved only for transfer learning strategy 1, where features were extracted from a pre-trained DCNN followed by binary classification using the PCA-LDA model. It is important to mention that the analysis was performed on a limited number of TMAs and it is suspected that the results can improve with an increasing number of TMAs, at least for the transfer learning strategy 1. Furthermore, the data fusion approach can largely increase the feature dimension of the data, thus increasing computational complexity. Nevertheless, these limitations are the cost of performing reliable and early breast cancer diagnosis. In future studies, feature dimension can be reduced by extracting features from the last layers and a comparative study can be performed.

5.2 Strategy 1 vs. Strategy 2

From the results shown in table 3 it is clear that transfer learning strategy 1 outperforms the transfer learning strategy 2 for our breast cancer dataset. For transfer learning strategy 2, the misclassification of the under-represented normal class as tumour class is higher. This means that transfer learning strategy 2 performed by merging and fine-tuning the last layer of the pre-trained multi-input model causes 'negative transfer learning' showing lower binary classification performance. Although the past studies (Kensert et al., 2018; Mormont et al., 2018) have shown that transfer learning strategy 2 for medical imaging data can provide good classification performances, these studies used a single-input DCNN for fine-tuning; whereas, in this study a multi-input DCNN was used. Thus, training a large multi-input network on a small dataset can cause the model to overfit and degrade its performance. Degradation in model performance can also be a consequence of transferring features of top layers from two different domains (Yosinski et al., 2014). Specifically, the transferability of features can be negatively affected when the source task (e.g. classification of the ImageNet dataset) is different from the target task (e.g. breast cancer detection). Thus, transfer learning of features for different domains should be performed cautiously (Yosinski et al., 2014). Further, merging and fine-tuning only the last layer and initializing the weights of the whole network based on the ImageNet dataset transferred the specific features (learned in top layers) of the non-medical domain to the medical domain, thus decreasing the classification performance in the strategy 2. To improve the performance of a DCNN model by the transfer learning strategy 2, initializing and fine-tuning weights of the top and intermediate layers of the multi-input DCNN model should be investigated in future studies.

So far, limitations of the transfer learning strategy 2 were discussed, now it is important to discuss few

limitations of the transfer learning strategy 1. One of the limitations is the need for an aggressive downsampling of the pathological images according to the input size of the pre-trained DCNN, ignoring the essential information. Although it is also possible to use a desired input image size by removing the fully connected layers of a pre-trained DCNN, downsampling our patches of size 1024×1024 to the model's input size facilitated the best classification performance. Extracting smaller size patches to increase the number of patches were also evaluated during the analysis. However, it was observed that small size patches increased the dataset size but decreased the biologically significant tissue features in each patch. Irrespective of our acceptable results using the pre-trained DCNNs as feature extractors, the interpretability of the transferred features is questionable. It is difficult to obtain an intuitive understanding of the transferability of non-medical features obtained from the ImageNet dataset to the medical domain. Thus, it is important to investigate transferring features from the medical domain to improve the breast cancer classification rate in future.

5.3 Effect of DCNN Architecture

It was clear from the results that acquiring a good classification rate using data fusion approach is dependent on the DCNN model. For transfer learning strategy 1, the Inceptionv3 and the ResNet50 network achieved better classification performances. While for transfer learning strategy 2, the multi-input VGG16 network achieved good classification performance. Furthermore, for transfer learning strategy 1, the Inceptionv3 and the VGG16 provided a large number of features (as they were combined from multiple modalities) in comparison to the ResNet50 network. Large feature dimension not only increased the dataset size but also increased the memory requirement. However, large feature dimension obtained by large DCNNs like the Inceptionv3 and the ResNet50 proved to be beneficial for training the PCA-LDA model in transfer learning strategy 1. While for transfer learning strategy 2, it was seen that large DCNN like the multi-input Inceptionv3 and the multi-input ResNet50 networks easily overfit and degrade model performance. It is suspected that large networks in multi-input fashion like the Inceptionv3 and the ResNet50 network generates a large number of trainable parameters which degrades model performance during fine-tuning. Furthermore, the time required to fine-tune the last layers of networks increases with network size.

6 CONCLUSION

The results show that combining histological imaging data along with IHC imaging data (estrogen receptor, progesterone receptor, human epidermal growth factor-2 and Ki-67) can improve breast cancer classification rate as compared to histological imaging data alone. The improvement in the classification performance was approximately 2% when deep convolutional neural networks (DCNN) were used as feature extractors (i.e. transfer learning strategy 1). However, the classification performance degraded when fine-tuning of the last layer of the multi-input DCNN (i.e. transfer learning strategy 2) was performed. Out of all three pre-trained networks, the pre-trained residual network and inception network as feature extractor outperformed the binary classification task (tumour vs normal), while the pre-trained VGG network as feature extractor obtained reasonable results. On the other hand, the VGG network showed better performances than the residual network and the inception network when fine-tuning of last layers was performed. The increase in performance by 2% for diagnosing breast cancer is explainable, because this task is normally performed using H&E, so the advancement is limited. Nevertheless, the data fusion approach can substantially improve differential diagnosis, which is important from a clinical perspective. Therefore, combining histology and IHC staining technique should be encouraged in future for more complicated tasks like a differential diagnosis or the prognosis of breast cancer patients. Overall, this comparative study showed that transfer learning could be utilized to diagnose breast cancer based on the combined histological and IHC imaging data with acceptable results. However, it is important to perform this study on a larger dataset in future. On large dataset, transfer learning strategy 3 i.e. training a DCNN from scratch can also be investigated. Furthermore, the data fusion approach can be performed to characterize stages of breast cancer in future.

ACKNOWLEDGEMENTS

Financial support of the German Science Foundation (BO 4700/1-1, PO 563/30-1 and STA 295/11-19) and funding by the BMBF for the project Uro-MDD (FKZ 03ZZ0444J) are highly acknowledged.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Antony, J., McGuinness, K., O'Connor, N. E., and Moran, K. (2016). Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE.

Bayramoglu, N. and Heikkilä, J. (2016). Transfer learning for cell nuclei classification in histopathology images. In *European Conference on Computer Vision*, pages 532–539. Springer.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P. S., Parker, J. S., et al. (2009). Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, 101(10):736–750.

Chen, T. and Chefd'Hotel, C. (2014). Deep learning based automatic immune cell detection for immunohistochemistry images. In *International workshop on machine learning in medical imaging*, pages 17–24. Springer.

Chollet, F. et al. (2015). Keras. https://github.com/fchollet/keras.

Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567.

Damodaran, S. and Olson, E. M. (2012). Targeting the human epidermal growth factor receptor 2 pathway in breast cancer. *Hospital Practice*, 40(4):7–15.

Dobson, L., Conway, C., Hanley, A., Johnson, A., Costello, S., O'Grady, A., Connolly, Y., Magee, H., O'Shea, D., Jeffers, M., et al. (2010). Image analysis as an adjunct to manual her-2 immunohistochemical review: a diagnostic tool to standardize interpretation. *Histopathology*, 57(1):27–38.

Elledge, R. M., Green, S., Pugh, R., Allred, D. C., Clark, G. M., Hill, J., Ravdin, P., Martino, S., and Osborne, C. K. (2000). Estrogen receptor (er) and progesterone receptor (pgr), by ligand-binding assay compared with er, pgr and ps2, by immuno-histochemistry in predicting response to tamoxifen in metastatic breast cancer: A southwest oncology group study. *International journal of cancer*, 89(2):111–117.

Guo, S., Bocklitz, T., Neugebauer, U., and Popp, J. (2017). Common mistakes in cross-validating classification models. *Analytical Methods*, 9(30):4410–4417.

Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 582–591.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, Germany.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huynh, B. Q., Li, H., and Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501.

Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.

Kensert, A., Harrison, P. J., and Spjuth, O. (2018). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS DISCOVERY: Advancing Life Sciences R&D*, page 2472555218818756.

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., and Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine*, 27:317–328.

Kieffer, B., Babaie, M., Kalra, S., and Tizhoosh, H. R. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural

networks. In *Advances in neural information processing systems*, pages 1097–1105.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26.

Mormont, R., Geurts, P., and Marée, R. (2018). Comparison of deep transfer learning strategies for digital pathology. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2262–2271.

Oliphant, T. (2006). NumPy: A guide to NumPy. USA: Trelgol Publishing.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, 406(6797):747–752.

Pham, N.-A., Morrison, A., Schwock, J., Aviel-Ronen, S., Iakovlev, V., Tsao, M.-S., Ho, J., and Hedley, D. W. (2007). Quantitative image analysis of immunohistochemical stains using a cmyk color model. *Diagnostic pathology*, 2(1):1–10.

Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., and Vaidya, V. (2016). Understanding the mechanisms of deep transfer learning for medical images. In *Deep Learning and Data Labeling for Medical Applications*, pages 188–196. Springer.

Ribeiro, E., Uhl, A., Wimmer, G., and Häfner, M. (2016). Exploring deep learning and transfer learning for colonic polyp classification. *Computational and mathematical methods in medicine*, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Sheikhzadeh, F., Ward, R. K., van Niekerk, D., and Guillaud, M. (2018). Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks. *PloS one*, 13(1):e0190783.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global.

Van Ginneken, B., Setio, A. A., Jacobs, C., and Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*, pages 286–289. IEEE.

Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Veta, M., Pluim, J. P., Van Diest, P. J., and Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Webster, L., Bilous, A., Willis, L., Byth, K., Burgemeister, F., Salisbury, E., Clarke, C., and Balleine, R. (2005). Histopathologic indicators of breast cancer biology: insights from population mammographic screening. *British journal of cancer*, 92(8):1366–1371.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.