Journal of Pathology Informatics 15 (2024) 100377

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Short Communication

Evaluation of the precision and accuracy in the classification of breast histopathology images using the MobileNetV3 model

Kenneth DeVoe [a], Gary Takahashi [a,*], Ebrahim Tarshizi [a], Allan Sacker [b]

[a] Shiley-Marcos School of Engineering, Applied Artificial Intelligence MS Program, University of San Diego, 5998 Alcalá Park, San Diego, CA 92110, USA
[b] Department of Pathology, Providence St. Vincent Medical Center, 9205 SW Barnes Road, Portland, OR 97225, USA

ARTICLE INFO

Keywords:
MobileNetV3
Convolutional neural networks
Breast cancer
Clinical pathology
Histology
BreakHis

ABSTRACT

Accurate surgical pathological assessment of breast biopsies is essential to the proper management of breast lesions. Identifying histological features, such as nuclear pleomorphism, increased mitotic activity, cellular atypia, patterns of architectural disruption, as well as invasion through basement membranes into surrounding stroma and normal structures, including invasion of vascular and lymphatic spaces, help to classify lesions as malignant. This visual assessment is repeated on numerous slides taken at various sections through the resected tumor, each at different magnifications. Computer vision models have been proposed to assist human pathologists in classification tasks such as these. Using MobileNetV3, a convolutional architecture designed to achieve high accuracy with a compact parameter footprint, we attempted to classify breast cancer images in the BreakHis_v1 breast pathology dataset to determine the performance of this model out-of-the-box. Using transfer learning to take advantage of ImageNet embeddings without special feature extraction, we were able to correctly classify histopathology images broadly as benign or malignant with 0.98 precision, 0.97 recall, and an F1 score of 0.98. The ability to classify into histological subcategories was varied, with the greatest success being with classifying ductal carcinoma (accuracy 0.95), and the lowest success being with lobular carcinoma (accuracy 0.59). Multiclass ROC assessment of performance as a multiclass classifier yielded AUC values ≥ 0.97 in both benign and malignant subsets. In comparison with previous efforts, using older and larger convolutional network architectures with feature extraction pre-processing, our work highlights that modern, resource-efficient architectures can classify histopathological images with accuracy that at least matches that of previous efforts, without the need for labor-intensive feature extraction protocols. Suggestions to further refine the model are discussed.

Introduction

Breast cancer is the most common cancer afflicting females, second only to skin cancers, and now afflicts nearly 1 in 3 women each year.[1] This condition usually presents either as a palpable breast mass, or, as is more often the case, detected by screening mammography. Once a lesion is identified, a core needle biopsy is performed to assess the nature of the abnormality, and a surgical pathologist is tasked with precise identification of the lesion, as well as a panel of an associated standardized set of biomarker phenotypes.[2,3] Tumor behavior can reasonably be predicted by evaluating the histological features such as cellular morphology, the degree and nature of architectural distortion, as well as invasion through basement membranes into normal structures, including blood vessels and lymphatics.[4] It is pertinent to note that each case typically involves the evaluation of numerous slides involving various sections through the tumor, each at different magnifications, as features such as architectural patterns are more apparent with the lower power widefield objective, while cellular and nuclear detail are better evaluated using the higher power objectives. Making an accurate diagnosis in a timely manner is critical, as appropriate intervention when curative management is possible can lead to improved survival outcomes.[5]

Interpretation of mammographic imaging has benefited from computer-assisted identification of suspicious radiographic lesions.[6] Deep learning may also assist the pathologist in histological classification of surgical biopsies.[7] Previous efforts have utilized convolutional neural net (CNN) algorithms to broadly classify breast cancer pathological images into benign and malignant categories, using larger image processing architectures that have been used to successfully classify non-medical, more everyday images. Convolutional networks have formed the basis for virtually all computer vision architectures since the AlexNet model famously won the ImageNet competition in 2012, achieving an unprecedented top-5 error rate of 17.0%.[8] Since then, other groups have built upon this concept,[9] and subsequent years saw the appearance of newer architectures, such as deep residual networks[10] and inception networks,[11] among others. These

* Corresponding author at: 7575 SW Brentwood St, Portland, OR 97225, USA.
E-mail address: gtakahashi@sandiego.edu (G. Takahashi).

http://dx.doi.org/10.1016/j.jpi.2024.100377
Received 3 February 2024; Received in revised form 5 April 2024; Accepted 8 April 2024
Available online 10 April 2024

networks were tested on the ImageNet dataset, a large publicly available collection of 14.2 million labeled images across 21 841 synsets (categories),[12,13] and this resource has been a benchmark for evaluating new computer vision architectures.

Convolutional models excel by first learning to detect pertinent features of an image, such as edges, textures, and shapes. Deeper layers in the architecture then become trained to detect combinations of these features.[14] To enhance edges and contrast along the borders of features of interest, feature extraction techniques have been used. Semantic segmentation has been used to selectively apply coloration to specific features, such as the nuclei, or to the emphasize the boundaries of an infiltrating tumor mass.[15,16] By enhancing the contrast of the borders between 2 regions of concern, this could theoretically increase classification accuracy, but this comes at the cost of time spent pre-processing the data in this manner, and the time and effort needed to confirm the accuracy and precision of the segmentation. The ability to accurately classify pathology images without the need for special pre-processing, such as segmentation, would greatly enhance the ease and efficiency of machine-learning image classification.

Material and methods

Dataset selection

BreakHis_v1 (BreakHis) is a large dataset consisting of 9109 microscopic images collected from 82 patients. Of these, 7909 images were available for public downloading.[17] This dataset includes 2480 images of benign breast lesions, and 5429 images of malignant lesions. Regions of interest (ROIs) have been identified by a pathologist, and the images were obtained at the indicated magnifications (Fig. 1).

Benign breast lesions were divided into four categories: adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma. The malignant conditions selected for inclusion were ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. Each histological category was represented in four magnification levels: 40×, 100×, 200×, and 400×. These refer to the microscope objective magnification multiplied by the 10× ocular lens (Fig. 2).

We selected the BreakHis dataset as the most suitable for our objectives, as other breast cancer datasets that have been used in other publications either were no longer available, contained fewer or smaller images for classification, offered only binary classifications (i.e., benign vs. malignant) or were better suited for the detection of metastatic spread in the background of lymphoid tissue.

Exploratory data analysis

It is necessary to evaluate histopathological images at various levels of magnification. Images examined at low magnification may reveal glandular structures, stromal invasion, and architectural distortion by the neoplasm. Higher magnifications may reveal the presence of mitotic figures, nuclear atypia, degree of cellular differentiation, and abnormal growth patterns, as well as the presence or absence of invasion through basement membranes (signifying metastatic spread) as well as abnormal tumor invasion through lymphatic or vascular channels. Because each magnification contributes useful information that would be important to classification, we did not restrict the analysis to specific magnifications, but elected to train the model on all magnifications collectively.

Examining multiple random samples from the dataset revealed that the images were of good quality and suitable for training. The size of each image in the BreakHis_v1 dataset is 700 × 460 pixels and is stored in PNG format. MobileNetV3 requires images of equal height and width, with three channels for RGB color representation, and the dataset images were resized accordingly. The crop_to_aspect_ratio parameter was left at the default setting, allowing for width compression so that the software would not crop images indiscriminately, potentially eliminating key histological features from being available for training.

In prior studies, feature extraction has been utilized in other efforts as a means of strengthening features, such as contrast enhancement to emphasize edges or semantic segmentation to highlight regions of interest (ROIs). Implementation of this step has required input from trained pathologists to confirm the correct identification of ROI, and in the setting of metastatic breast cancer.[18] We evaluated the performance of our model without implementing feature extraction, so as to classify breast pathology into four benign and four malignant categories, against the background of normal breast tissue, a potentially more challenging task.

Data augmentation

The BreakHis dataset contains approximately 110–140 images at each magnification in most of the neoplasm categories, but for the most common neoplasms such as fibroadenomas, there are 230–260 images at each magnification, and between 790 and 900 images for ductal carcinomas. Categorical representation in this dataset reflects the clinical prevalence of these histological entities, as ductal carcinomas and fibroadenomas are the most common malignant and benign histologies, respectively.[19,20] While a more evenly balanced dataset is usually preferred to decrease

	40X	100X	200X	400X	TOTAL
ADENOSIS	114	113	111	106	444
FIBROADENOMA	253	260	264	237	1014
PHYLLODES	109	121	108	115	453
TUBULAR ADENOMA	149	150	140	130	569
	40X	**100X**	**200X**	**400X**	**TOTAL**
DUCTAL CA	864	903	896	788	3451
LOBULAR CA	156	170	163	137	626
MUCINOUS CA	205	222	196	169	792
PAPILLARY CA	145	142	135	138	560

Fig. 1. BreaKHis dataset data distribution.

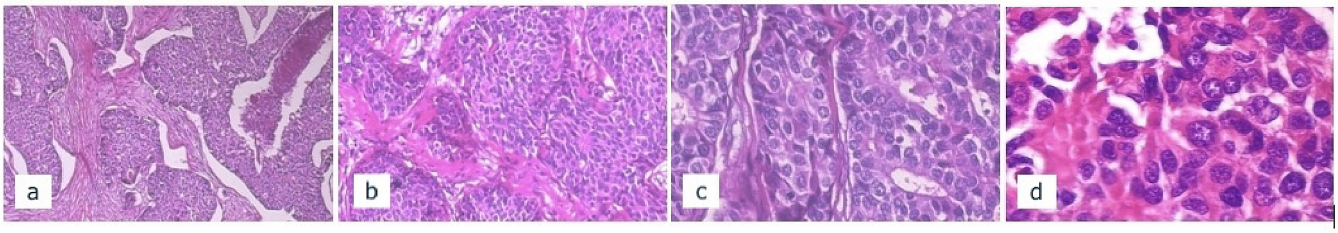

Fig. 2. Examples from the BreaKHis_v1 ductal carcinoma dataset, seen at magnifications of: (a) 40×, (b) 100×, (c) 200×, (d) 400×. Lower magnifications more clearly display architectural details of tumor penetration into normal breast tissue, while at higher magnifications, cellular and nuclear detail become more important. Presence of metastatic cells in the lymphatic channels and capillaries is better appreciated at the higher magnifications.

overfitting to one category, it may not be entirely disadvantageous to having a model be trained based on a weighted representation that reflects the actual prevalence. Nonetheless, data augmentation over all images is a common technique to manage unevenness in the dataset. In this endeavor, this was implemented by introducing a simple random horizontal image reflection (mirror images) to the model, effectively doubling our virtual sample size. We elected not to implement more distortive effects, such as shear. Other data augmentative effects, such as rotation would be meaningless with histology images, and the various magnifications would provide zoom information. We assessed the effect of data augmentation on the training and validation accuracy curves, which will be presented below.

Model selection

The MobileNet family is built on the concept of depthwise-separable convolutions, which has allowed it to achieve image classification accuracy comparable to that of larger models while being efficient and and incurring lower computational cost.[21] MobileNetV3 built on MobileNetV2's improvements by introducing the squeeze-and-excite algorithm in parallel to all components of the residual block; using Neural Architectural Search to improve accuracy[22]; and replacing some of the sigmoid activations with the hard swish activation function which is computationally less "expensive" than the sigmoid activation function (a major factor in mobile devices), to improve performance but without compromising on accuracy. This enabled some layer reduction in the last stages of the model without decreasing accuracy. MobileNetV3 is divided into small and large versions, with the small version to be run on reduced-resource devices (Fig. 3).

Model architecture

The MobileNetV3 large architecture offers improved accuracy over the companion MobileNetV3 small architecture. The ImageNet dataset embeddings are enabled by default,[23] however, it was not clear that these parameter weights obtained from training on the widely varied images obtained from the Internet would be helpful in classifying histology images. We decided to implement transfer learning, utilizing weights and biases of earlier layers, which detect shapes and edges in images, while training on the later layers, which register activations based on more complex image features. This will be discussed more in detail in the section on Layerwise Learning.

A key hyperparameter is the model learning rate, which controls how rapidly the parameters of the model converge to the local minimum of the categorical cross-entropy loss function,[24] used to assess the accuracy of classification after each training epoch. As the loss function local minimum is approached, optimal adjustments of the learning rate at each training epoch can help to prevent overshoot, and subsequent deterioration of accuracy. The optimal freezing layer as well as setting the magnitude of the learning rate are heuristically determined. We tested a range of epoch decay values, which were first entered into a spreadsheet to facilitate automation. In this way, the model could be programmed to autonomously run multiple training sessions using different hyperparameter values, while model accuracy was recorded.

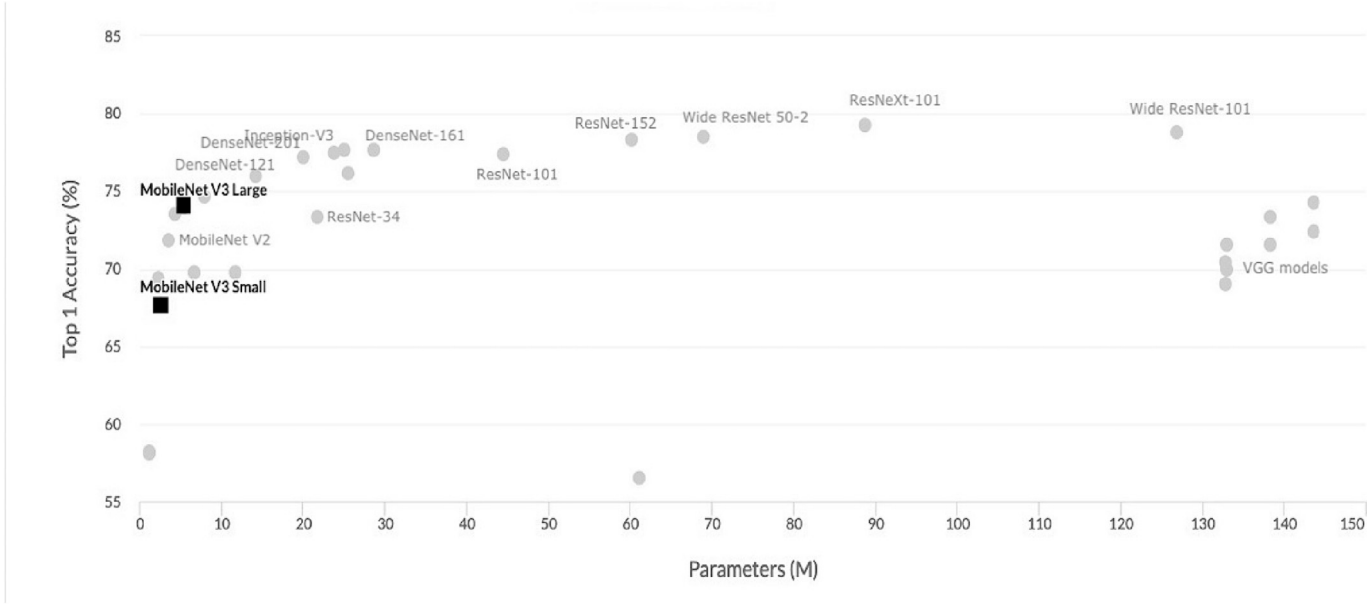

Fig. 3. Top accuracy of the MobileNetV3 models as compared with other computer vision architectures, stratified by the number of trainable parameters in the model (Adopted from Pandey[38]).

Layer type	Output Shape	Param #
InputLayer	[(None, 224, 224, 3)]	0
Sequential	(None, 224, 224, 3)	0
MobileNetV3large (Functional)	(None, 7, 7, 960)	2996352
GlobalAveragePooling2D	(None, 960)	0
Dropout	(None, 960)	0
Dense	(None, 64)	61504
BatchNormalization	(None, 64)	256
Dense	(None, 32)	2080
Dense	(None, 16)	528
Dense	(None, 8)	136

Total parameters: 3060856 (11.68 MB)

Fig. 4. Architecture of model used for image classification.

Training

The BreakHis dataset was split into training, validation, and test subsets at ratios of 0.75/0.15/0.10. Outputs from MobileNetV3 were flattened with a Global_Average_Pooling2D layer, then passed to four Dense layers with ReLU activation, followed by a Dense layer with softmax activation to eight outputs. Dropout regularization and BatchNormalization were used to reduce overfitting to the training set. As shown in Fig. 4, the final model consisted of approximately 3.0 million parameters, with 98% associated with MobileNetV3 Large. Hyperparameters of learning rate epoch decay, freezing layer setting were loaded into a spreadsheet, and used to autonomously train the model for 13 epochs. The workflow is illustrated in Fig. 5. Accuracy measurements on the validation set were recorded, and are displayed below. The full Python code is available at https://github.com/kdevoe/MobileNetV3_Breast_Cancer_Detection.

Results

Model training and hyper-parameter tuning were focused on three key areas; data augmentation, layer selection for training, and learning rate selection.

Fig. 6 depicts the accuracy on the training and validation sets, and the effect of data augmentation. In our model, accuracy in the training and validation sets exhibited minimal separation. After 10 epochs of training,

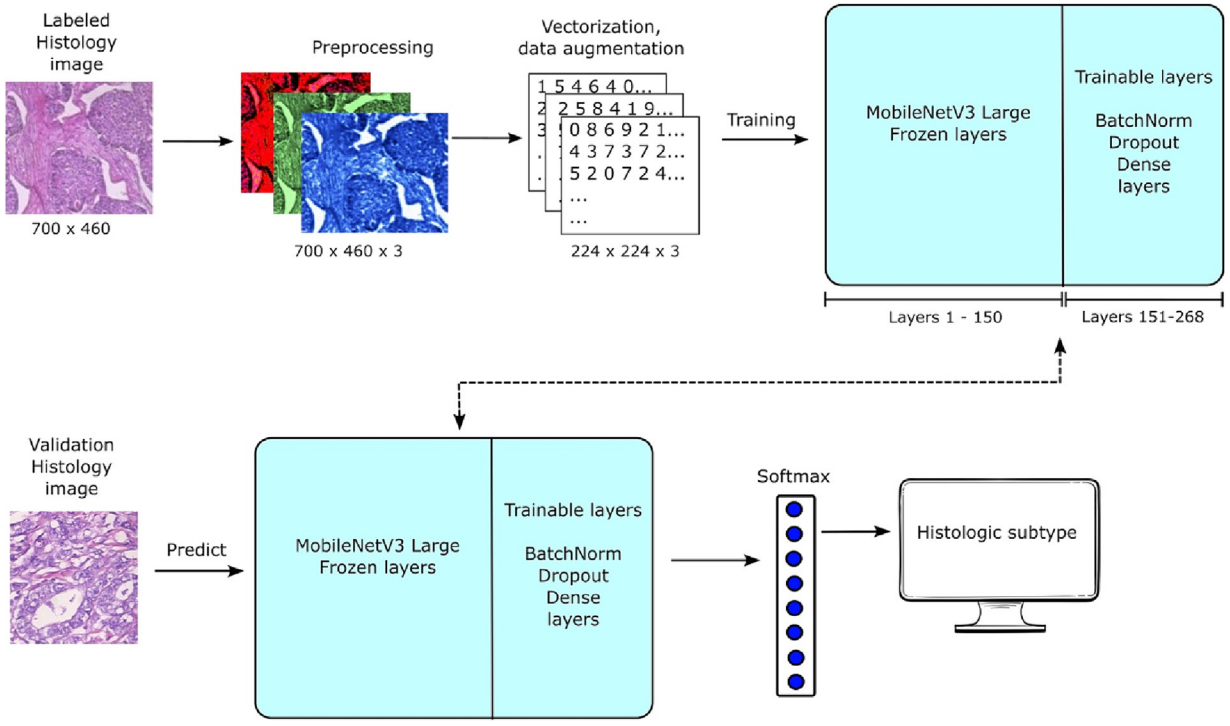

Fig. 5. Workflow diagram. Histology images are pre-processed, then vectorized. Data augmentation is applied and fed to the MobileNetV3 model, which is trained for 13 epochs. The trained model is fed an image from the validation set, on which the model makes a prediction, selecting from one of eight possible labels. The output is then displayed.

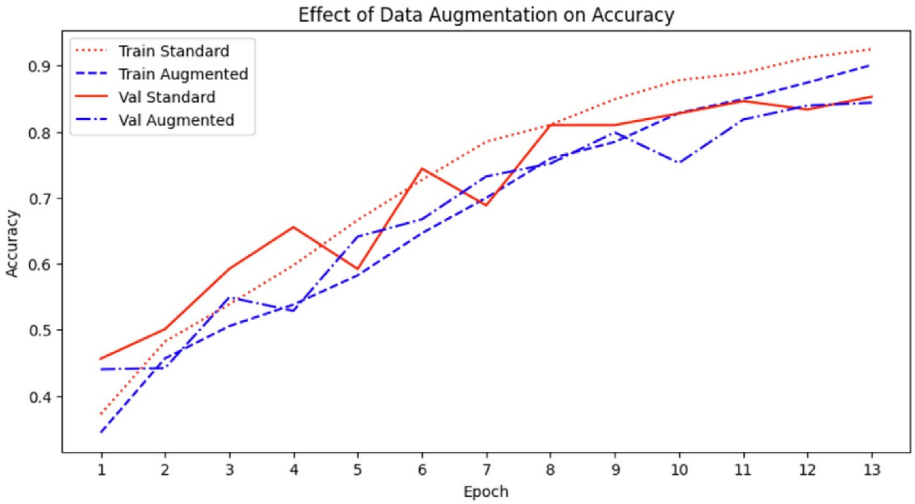

Fig. 6. Comparison of training vs. validation curves with and without data augmentation. Overall, data augmentation reduces overfitting, bringing the validation results closer to the training results. Training runs were performed with an initial learning rate of 0.001, epoch decay rate of 0.95, and fine-tuning start layer of 150.

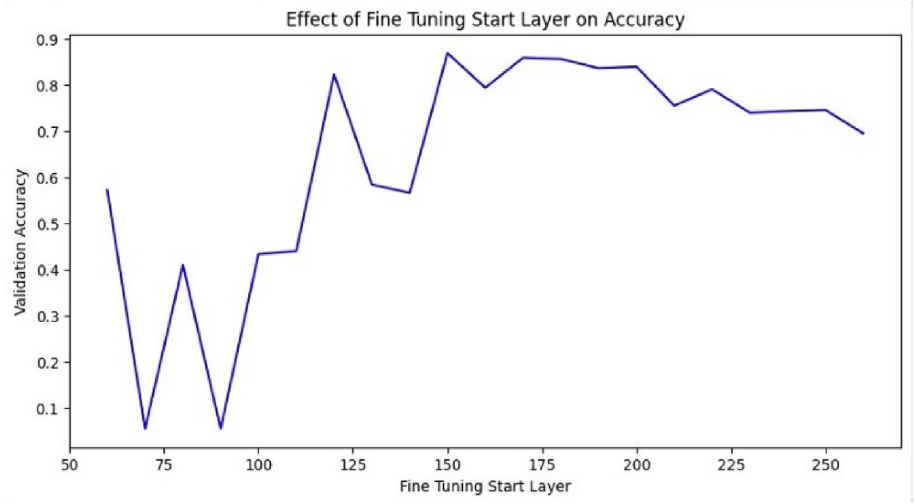

Fig. 7. Validation accuracy of the model after 13 epochs based on the initial fine-tuning layer. For reference, the model includes 268 layers total, with 263 layers from MobileNetV3. Each training session used an initial learning rate of 0.001 and epoch decay rate of 0.95. This graph represents 21 unique model training runs.

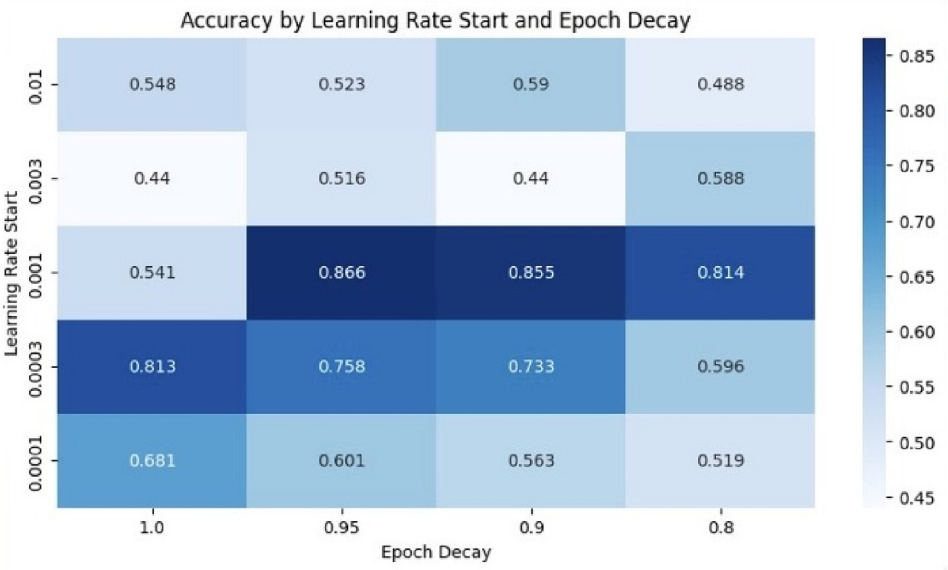

Fig. 8. Validation accuracy of the model based on the initial learning rate and epoch decay rate. Fine-tuning was started at layer 150 out of 268 total layers for all models, with training run for 13 epochs.

validation accuracy reached a plateau with no further improvement. Training and validation accuracy tracked closely, and the impact of data augmentation was minimal.

Layerwise learning

Implementation of transfer learning during model training involves setting the hyperparameter that determines the layer at which to freeze (preserve) pre-trained ImageNet embeddings, beyond which training will occur in the MobileNetV3 model. If the freezing layer is set too distally, then an insufficient number of layers of the model will be fine-tuned on the new dataset, and performance on tumor images will be inadequate. However, if the freezing layer is set too proximally, then too much of the pre-trained embeddings will be overwritten, and edge-and shape-detecting functionality garnered from pre-training on ImageNet will be lost (Fig. 7). Initiating the training prior to layer 130 led to poor, erratic performance of the overall model, as the ability to recognize basic image features learned from ImageNet had been overwritten. On the other hand, delaying the start of training until after layer 150 resulted in a slight decline in performance, likely due to inadequate fine-tuning with the tumor dataset. Based on these results, layer 150 was selected as the optimal threshold, beyond which the model parameters would be made available for training.

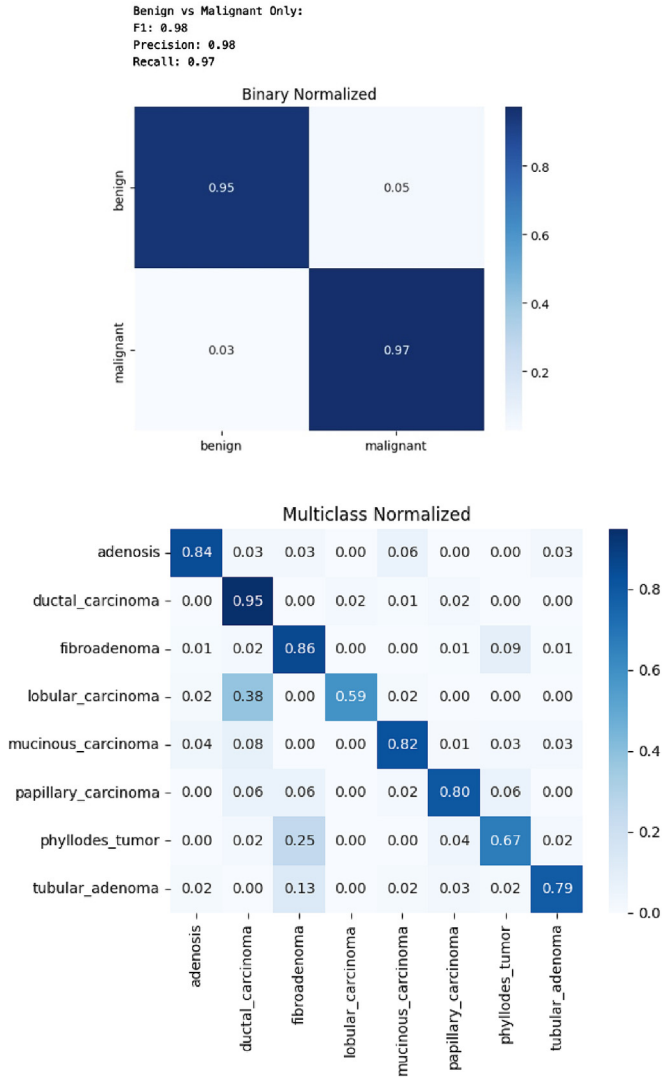

Fig. 9. (a) Confusion matrix for accuracy generated for the binary classification of benign vs. malignant. (b) Confusion matrix for accuracy in classifying histological subtypes.

Learning rate selection (epoch decay rate)

The heuristically determined initial learning rate for training is the coefficient modulating the adjustments made to the weights and biases of the model parameters after each epoch of training. As the training proceeds and the loss function converges to the local minimum, it may be advantageous to reduce this value progressively, so as not to overshoot the target weightings.[24] There is no standard learning rate schedule, and various implementations have been developed empirically to maximize test accuracy for particular benchmarks.[25] We chose to reduce the learning rate by a constant factor at each epoch of training, which we refer to as the epoch decay rate. Fig. 8 shows the model validation accuracy as a function of initial learning rate and epoch decay rate. As indicated previously, fine-tuning began at layer 150 out of the total of 268 layers. We tested several learning rate settings without epoch decay and found that the highest accuracy was achieved with a learning rate of 0.0003. However, when epoch decay was incorporated, the highest accuracy was achieved with an initial learning rate of 0.001 and an epoch decay rate of 0.95. More aggressive epoch decay rates of 0.9 and 0.8 resulted in further reduction of model accuracy, suggesting excessive attenuation of training efforts at later epochs. Therefore, for the final analysis, a start learning rate of 0.001 and epoch decay of 0.95 was selected.

Classification accuracy

The trained model was then used to predict on images in the test set, and was able to classify the histology slides as benign vs. malignant with a recall of 0.97, a precision of 0.98, and an F1 score of 0.98 (Fig. 9a). 95% confidence intervals were calculated according to the method outlined by Newcombe.[26] Amongst the subtypes of each major category, the model had the greatest success in identifying ductal carcinoma, adenosis, and tubular adenoma, with accuracies of ≥ 0.9. Mucinous carcinoma and phyllodes tumors were accurately identified in 0.85 of cases. Accurate classification for papillary carcinoma was 0.74, and lobular carcinoma was 0.56 (Figs. 9b, 10).

The F1 score, precision, and recall scores were calculated for the classification into the four benign and four malignant subtypes. The performance of the model was highest with classifying ductal carcinoma, likely because of the larger number of samples on which to train. Performance was lowest with phyllodes tumor, of which the fewest samples were provided in the dataset.

The tabular data in Fig. 10 can also be visualized as receiver operating curves (ROC), as depicted in Fig. 11. The ROC curves confirm the relative classification performance.

Incorrectly classified images

To gain insight into how the model erred in classifying some of the histopathology images, we identified the misclassified images, comparing their predicted labels with the ground-truth label. A few of these images are presented in Fig. 8.

In Fig. 12a, the image that was misclassified as lobular carcinoma contained mainly tissue stroma or possibly tissue necrosis, with very few identifiable cells. In Fig. 12b, the image misclassified as phyllodes tumor was reasonable in quality, however, in this selected image, it may be difficult for even a human pathologist to make the distinction between fibroadenoma and phyllodes tumor. The image in Fig. 12c, also misclassified as phyllodes tumor, is of such high magnification that the image consists of a sheet of nuclei with dispersed chromatin and indistinct cellular borders, and the possibilities of its provenance are numerous. The final image in Fig. 12d, misclassified as ductal carcinoma, consists mainly of cellular outlines, and may possibly represent a region of adipose tissue, with little material that can be identifiable as carcinoma. It is, therefore,

	F1 score	Precision score	Recall score	Count
Adenosis	0.831 ± 0.028	0.818 ± 0.041	0.844 ± 0.038	345
Ductal carcinoma	0.922 ± 0.007	0.898 ± 0.012	0.948 ± 0.009	2576
Fibroadenoma	0.808 ± 0.020	0.765 ± 0.030	0.857 ± 0.025	753
Lobular carcinoma	0.692 ± 0.030	0.841 ± 0.033	0.587 ± 0.045	468
Mucinous carcinoma	0.855 ± 0.020	0.89 ± 0.025	0.823 ± 0.031	594
Papillary carcinoma	0.769 ± 0.028	0.741 ± 0.041	0.8 ± 0.038	437
Phyllodes tumor	0.681 ± 0.036	0.696 ± 0.050	0.667 ± 0.051	331
Tubular adenoma	0.842 ± 0.024	0.906 ± 0.028	0.787 ± 0.039	428

Fig. 10. F1 score, precision score, and recall scores (±95% confidence intervals) for classification into the histological subtypes.

not unreasonable for images to have been misclassified, as they would pose a challenge for even expert human evaluation (Bayes optimal error rate).

Discussion

We have shown that a fine-tuned convolutional neural network, optimized for image classification and designed for increased accuracy, decreased latency, and decreased resource utilization, was able to properly categorize breast cancer histopathology images as to being benign vs. malignant with ≥97% precision and recall. Classification of dataset images into the given histological subtypes was achieved with moderate success. Accuracy was best with the classification of ductal carcinoma, which is important since it is the most common subtype of breast cancer, representing 70%–80% of all breast cancers.[27] Images representing this histological subtype were most represented in the BreaKHis dataset, which could account for the relative accuracy in classifying ductal carcinoma, as there were more examples for the model to train on. Because the model trains to decrease loss, the increased prevalence of this category might have led to more rapid reduction in loss function values than for the other categories. Nevertheless, confident identification of infiltrating ductal carcinoma would be of value, as it could potentially assist in the confirmation of this histology in over half of breast cancer cases.

To put our findings into context, we will review some of the work that inspired our own efforts. Spanhol et al[17] used six feature extractors and four different unsupervised classifiers on the BreaKHis dataset at each of the magnifications available. Their model best classified malignant histology correctly, with accuracies of around 0.94, but the accuracy of classifying benign tissue ranged from 0.38 to 0.75 at various magnifications. Error was most noted in classifying fibroadenomas, which constituted around 30% of errors at every magnification. The ROC AUC was around 0.8 at each of the four magnification levels, however, the AUC with all magnifications collectively was not reported. As stated above, pathologists extract architectural information at low magnification and nuclear and cellular detail and tissue invasiveness at high magnification, so constraining the model to classify at one magnification level is an unnecessary restriction.

Zhu et al[28] trained a custom CNN model with ideas taken from the Inception architecture (residual network connections), but with the "squeeze-excite" features of MobileNetV3. Their model was trained on the BreakHis and BACH breast cancer datasets. They also reported accuracy based on each magnification class, and achieved greater than 0.9 AUC on ROC analysis, but this was based on channel pruning of the model, ostensibly to decrease computing burden, although it raises some concern about how this model would generalized to other datasets.

Araújo et al[29] studied an unspecified CNN (but which resembles AlexNet) together with a support vector machine classifier, and trained on the Bioimaging 2015 challenge dataset, consisting normal as well as benign, and malignant breast cancer, some invasive and some in situ. "Patchwise accuracy" was 0.667–0.776. Image-wise accuracy was 0.778–0.833, using a voting system, the process of which is not well-described.

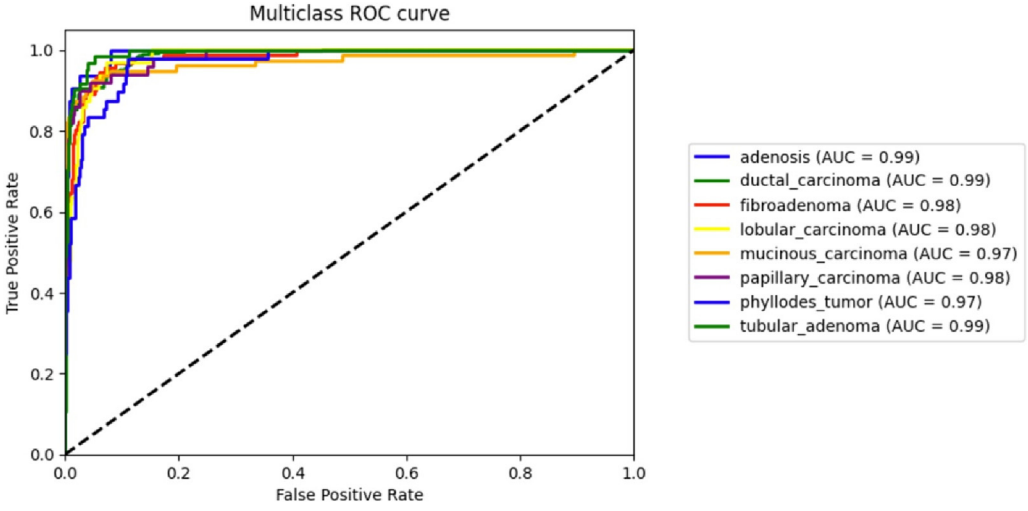

Fig. 11. ROC AUC curves depicting accuracy of the classifier across the histological subtypes.

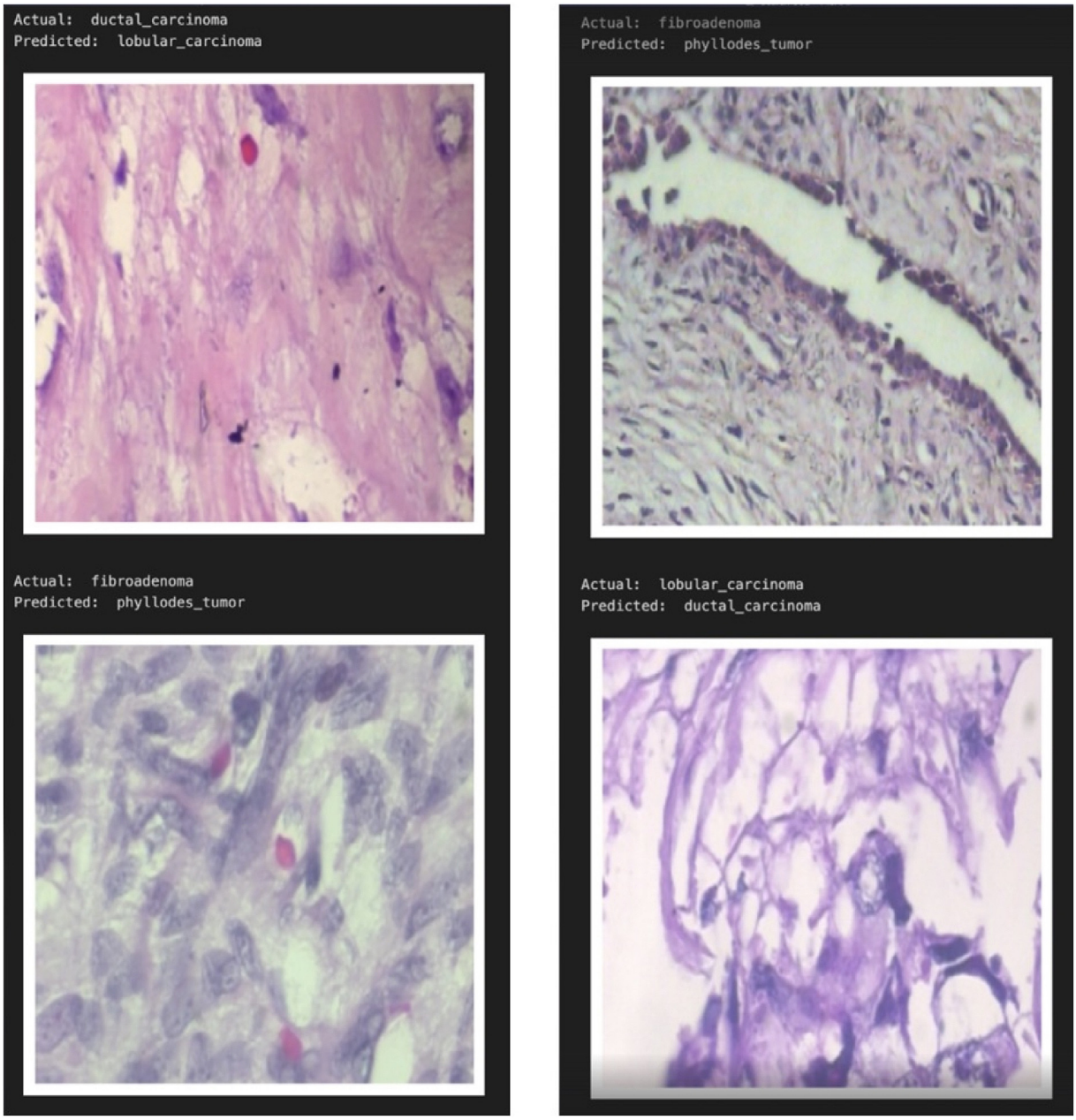

Fig. 12. A sample of misclassified images. (a) Ductal carcinoma misclassified as lobular carcinoma; (b) fibroadenoma misclassified as phyllodes tumor; (c) fibroadenoma misclassified as phyllodes tumor; (d) lobular carcinoma misclassified as ductal carcinoma.

Joshi et al[30] trained an Xception model on the BreaKHis and IDC breast cancer datasets. The objective was the classify the material into benign and malignant categories. ROC AUC was 0.921 on the BreakHis dataset and 0.881 on the IDC dataset. Accuracy data regarding further subclassification into histological subsets was not presented.

Amerikanos et al[31] used feature augmentation with semantic segmentation to train Facebook AI Research's Detectron2 architecture on the TUPAC16 challenge dataset. The investigators sought to avoid having a human pathologist determine the segmentation, and used an AlexNet trained with parameters published previously.[32] The objective of this study, however, was to automate the identification of nuclear to facilitate feature selection, and not histological classification.

We recognize that there are more histological entities than are represented in this limited dataset, such as triple-negative breast cancers and mixed-histologies. Indeed, we coded a predictor that performed well on selected random images from the dataset but performed less accurately on selected histopathology slides obtained from the Internet. These images were resized to the 700×468 pixel and converted to a dataset as in the BreakHis dataset. The images obtained from the Internet were chosen so as to match as closely as possible the images in the BreakHis dataset. However, despite this effort, the model performed rather poorly on the small internet dataset, with an accuracy never exceeding 0.6. We feel that the primary reason for this is the difference in image quality in the BreakHis dataset and ones available on the Internet (see Supplemental Fig. 2). Differences in the intensity of hematoxylin and eosin staining can result in differences in contrast of features such as cellular architecture, nuclear structure, and stromal detail, all of which can influence training of convolutional blocks that detect edges, corners and gradients. Tafavvoghi *et al* highlights variability in image quality amongst several available public breast cancer datasets,[33] and therefore classification accuracy between image repositories may more a reflection of differences in intrinsic source characteristics than model performance.

Another factor contributing to difficulty with model classification when attempting to predict on new images is the variability attributable to differences between low- and high-grade malignancies. Details of this kind were not provided with the BreakHis dataset, and all examples of one tumor subtype were placed into the same diagnosis bin. Accuracy may also be improved if our model were able to train using labels informing about tumor grade. Morphological pleomorphism in poorly differentiated malignancies can challenge even human pathologists, and therefore immunohistochemical data helps to confirm visual assessment of hematoxylin-eosin stained slides.

One of the known limitations of deep learning in convolutional architectures is that the trained weight embeddings are notoriously in a black box. Image data that is passed through CNNs undergo drastic transformations through convolution, pooling, flattening, such that examination of layers deep within the model reveal no discernible relationship to the original input (Fig 3). It has been difficult to elucidate precisely how a model is progressively trained to classify images so as to provide step-by-step documentation as to the process of generating predictions. A surrogate means of validation has been to report performance indicators on standardized datasets and infer that the accuracy and precision are replicated. At this time, the inability of convolutional models to justify and document the basis for classification supports the argument that these deep learning models will not replace a trained human pathologist, but instead may play an assistive role as long the pathologist does not second-guess his/her own reading, and place undue reliance on an erroneous reading of the trained model.[34]

Conclusions

Building upon a modern CNN architecture, designed for accuracy and efficiency, and to be compact enough to run on mobile devices, we were able to develop and train a model on a dataset of breast histopathological images, such that it was able to predict the classification of the images as being benign vs. malignant in nature, with a high degree of precision and accuracy. It was also moderately successful in classifying the pathology into one of eight subcategories.

The fine-tuning process was greatly aided by automating the training process, to heuristically identify optimal hyperparameter settings, such as freezing layer determination, as well as initial learning rate and epoch decay rate. Overfitting was addressed with data augmentation as well as using Dropout and BatchNormalization. Although the process we described did not require special feature extraction pre-processing, such as semantic segmentation, it likely that there would still be benefit to the selection of features that emphasize the best characteristics of each histological entity.

The latest member of the MobileNet family of CNNs was used in this project, but there is no consensus as to the "best" convolutional network for image classification. Abid et al[35] reported high accuracy (98.61%) in the classification of multi-modal medical images as to whether they were images of pathology slides, hand or chest radiographs, endoscopic and tomographic imaging using the ResNet50 model. In the Natural Language Processing field, it has been hypothesized that larger models with a greater number of trainable parameters are more "sample-efficient" and more performant on a wider range of sample data.[36] This paradigm was famously countered by the "Chinchilla" paper,[37] which demonstrated that a smaller model trained with an optimal number of tokens showed better performance compared to models with a larger number of parameters, and a similar situation might also hold for image classification models. Nonetheless, larger CNNs such as VGGNet, GoogLeNet/Inception, ResNet, and DenseNet would be expected to be similarly successful in achieving good accuracy in pathology image classifications.

As these models improve further, becoming more efficient and even less resource-demanding, and as pathologists develop trust in their accuracy, trained CNNs could become a useful addition to the pathologist's diagnostic toolkit.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used NO artificial intelligence software to generate text for the manuscript. The only artificial intelligence used was the convolution neural network as described in the article itself. We take full responsibility for the content of the publication.

The work in this manuscript has not been published or submitted for publication elsewhere.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2024.100377.

References

1. American Cancer Society. Key Statistics for Breast Cancer. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html; 2023.
2. College of American Pathologists. Protocol for the Examination of Resection Specimens from Patients with Invasive Carcinoma of the Breast, Version 4.9.0.0. Posted June 2023. https://documents.cap.org/protocols/Breast.Invasive_4.9.0.0.REL_CAPCP.pdf.
3. College of American Pathologists. Template for Reporting Results of Biomarker Testing of Specimens from Patients with Carcinoma of the Breast. Version 1.5.0.1. Posted March 2023. https://documents.cap.org/documents/Breast.Bmk_1.5.0.1.REL_CAPCP.pdf.
4. Weigelt B, Geyer FC, Reis-Filho JS. Histological types of breast cancer: how special are they? Mol Oncol 2010;4(3):192–208. https://doi.org/10.1016/j.molonc.2010.04.004.
5. Howlader N, Noone AM, Krapcho M, et al, eds. SEER Cancer Statistics Review, 1975-2016. Bethesda, MD: National Cancer Institute; April 2019.based on November 2018 SEER data submission, posted to the SEER web site. https://seer.cancer.gov/archive/csr/1975_2016/.
6. Loizidou K, Elia R, Pitris C. Computer-aided breast cancer detection and classification in mammography: a comprehensive review. Comp Biol Med 2023;153:106554. https://doi.org/10.1016/j.compbiomed.2023.106554.
7. Wang S, Yang DM, Rong R, Zhan X. Pathology image analysis using segmentation deep learning algorithms. Am J Pathol 2019;189(9):1686–1698. https://doi.org/10.1016/j.ajpath.2019.05.007.
8. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf; 2012.
9. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8:53. https://doi.org/10.1186/s40537-021-00444-8.
10. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Xplore; 2016. https://doi.org/10.1109/CVPR.2016.90.
11. Alom MD, Hasan M, Yakopcic C, Taha TM. Inception Recurrent Convolutional Neural Network for Object Recognition. arXiv:170407709.2017. https://doi.org/10.48550/arXiv.1704.07709.
12. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Xplore; 2009. https://doi.org/10.1109/CVPR.2009.5206848.
13. ImageNet. https://www.image-net.org/.
14. Lakshmanan V, Görner M, Gillard R. *Practical Machine Learning for Computer Vision. Chapter 3. Image Vision.* O'Reilly Media. 2021:67-123.
15. Ding K, Zhou M, Wang H, Gevart O, Metaxas D, Zhang S. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. Sci Data 2023;10:231. https://doi.org/10.1038/s41597-023-02125-y.
16. Janowczyk A, Madabushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29. https://doi.org/10.4103/2153-3539.186902.
17. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng (TBME) 2016;63(7):1455–1462. https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/.
18. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide image. Nat Med 2019;25(8):1301–1309. https://doi.org/10.1038/s41591-019-0508-1.
19. McDivitt RW, Stevens JA, Lee NC, Wingo PA, Rubin GL, Gersell D. Histologic types of benign breast disease and the risk for breast cancer. Cancer 1992;69(6):1408–1414. https://doi.org/10.1002/1097-0142(19920315)69:6/1408::aid-cncr2820690617%3E3.0.co;2-c.

20. Tan BY, Tan PH. A diagnostic approach to fibroepithelial breast lesions. Surg Pathol Clin 2017;11(1):17–42. https://doi.org/10.1016/j.path.2017.09.003.
21. MobileNetV3. Accessed 14 August 2023. https://paperswithcode.com/lib/torchvision/mobilenet-v3.
22. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H. Searching for MobileNetV3. Conference: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00140
23. MobileNet, MobileNetV2, and MobileNetV3. https://keras.io/api/applications/mobilenet/.
24. Russell S, Norvig P. Computation graphs for deep learning. Artificial Intelligence A Modern Approach. 4th ed. Hoboken, NJ: Pearson; 2020. p. 670.758.
25. Wolfe C.R. The best learning rate schedules. Deep (Learning) Focus. https://cameronwolfe.substack.com/p/the-best-learning-rate-schedules
26. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17:857–872. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e.
27. Li CI, Anderson BO, Daling JR. Trends in incidence rates of invasive lobular and ductal breast carcinoma. JAMA 2003;289(11):1421–1424. https://doi.org/10.1001/jama.289.11.1421.
28. Zhu C, Song F, Wang Y, Dong H, Guo Y, Liu J. Breast cancer histopathology image classification through assembling multiple compact CNNs. BMC Med Inform Decis Mak 2019;19:198. https://doi.org/10.1186/s12911-019-0913-x.
29. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. PLoS ONE 2017;12(6);e0177544. https://doi.org/10.1371/journal.pone.0177544.
30. Joshi SA, Bongale AM, Olsson PO, Urolagin S, Dharro D, Bongale A. Enhanced pre-trained xception model transfer learned for breast cancer detection. Computation 2023;11(3):59. https://doi.org/10.3390/computation11030059.
31. Amerikanos P, Maglogiannis I. Image analysis in digital pathology utilizing machine learning and deep neural networks. J Pers Med 2022;12(9):1444. https://doi.org/10.3390/jpm12091444.
32. Janowcyzk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Informatics 2016;7(1):29. https://doi.org/10.4103/2153-3539.186902.
33. Tafavvoghi M, Bongo LA, Shvetsov N, Busund LR, Møllersen K. Publicly available datasets of breast histopathology H&E whole-slide images. a scoping review. arXiv: 2306.01546v2 2023. https://arxiv.org/pdf/2306.01546.pdf.
34. Zhang Y, Liao QV, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. New York, NY, USA: ACM; 2020;11 pages. https://doi.org/10.1145/3351095.3372852.
35. Abid MH, Ashraf R, Mahmood T, Faisal CMN. Multi-modal medical image classification using deep residual network and genetic algorithm. PLoS ONE 2023;18(6), e0287786. https://doi.org/10.1371/journal.pone.0287786.
36. Kaplan J, McCandish S, Henighan T, et al. *Scaling laws for neural language models.* 2020. arXiv:2001.08361v1 [csLG].
37. Hoffman J, Borgeaud S, Mensch A, et al. *Training compute-optimal large language models.* 2022. arXiv:2203.15556v1 [cs.CL].
38. Pandey A. Depth-wise convolution and depth-wise separable convolution. Medium 2018. https://medium.com/@zurister/depth-wise-convolution-and-depth-wise-separable-convolution-37346565d4ec.