PRETRAINED VISION MODELS FOR PREDICTING HIGH-RISK BREAST CANCER STAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Cancer is increasingly a global health issue. Seconding cardiovascular diseases, cancers are the second biggest cause of death in the world with millions of people succumbing to the disease every year. According to the World Health Organization (WHO) report, by the end of 2020, more than 7.8 million women have been diagnosed with breast cancer, making it the world's most prevalent cancer. In this paper, using the Nightingale Open Science dataset of digital pathology (breast biopsy) images, we leverage the capabilities of pre-trained computer vision models for the breast cancer stage prediction task. While individual models achieve decent performances, we find out that the predictions of an ensemble model are more efficient. We also provide analyses of the results and explore pathways for better interpretability and generalization.

1 INTRODUCTION

Overall 2.3 million women have been diagnosed with breast cancer (BC) in 2020, and 685000 globally died from the disease. These statistics, coupled with the ones of the five last years, bring up to 7.8 million, the number of women alive who were diagnosed with BC by the end of 2020: more women have lost their lives due to BC than any other type of cancer, making it the most prevalent cancer in the world [1]. The treatment of BC can be efficient when the disease is detected at a very early stage, and there are mainly five stages of BC (Stage 0, Stage 1, Stage 2, Stage 3, and Stage 4). One of the most popular ways of detecting BC is mammography, which is a detailed X-ray scanning (or screening) of the breast with Magnetic Resonance Imaging (MRI). Another approach is breast biopsy: a biopsy is a medical process during which samples of cell tissues are collected to be examined in the laboratory with a microscope. A biopsy helps to locate the presence, cause, or extent of the disease.

Thanks to recent advances in Artificial Intelligence (AI), especially Deep Learning (DL), there has been a rise in research efforts to leverage the potential of DL-based systems to help in breast cancer detection. Gastounioti et al. (2022) provided a narrative review of AI in mammographic screening of breast cancer risk: the recent initiatives made use of pretrained Computer Vision (CV) models like ResNets He et al. (2016), DenseNets Huang et al. (2017), AlexNet Krizhevsky (2014), U-Net Ronneberger et al. (2015), cGAN Mirza & Osindero (2014), Szegedy et al. (2015), MobileNet Sandler et al. (2018), Inception Szegedy et al. (2016), and RetinaNet Lin et al. (2017). All of these methods were applied to Full-Field Digital Mammography (FFDM) in which X-rays mammograms are converted into electrical signals, in a binary classification (detecting BC or not BC) setting.

In this work, after briefly describing the Nightingale Open Science Dataset of Digital Pathology (NOSDDP) Bifulco et al. (2021); Mullainathan & Obermeyer (2022), we use slides from each biopsy digital image and pretrained CV models, to predict the stage of cancer of a patient (see Figure 1).

2 DATASETS

Existing works have demonstrated the ability of DL-based systems to predict the type of cancer based on X-ray scans and mammograms. DL-based algorithms also focus on features that some-

[1] Breast Cancer (WHO): `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`

times get overlooked or neglected by specialists (for instance the nature of the non-cancerous tissue surrounding the tumor) Beck et al. (2011). However, there are currently very few to no datasets, that link biopsy images to the cancer stage of the patients, and the respective outcomes: the NOSDDP is an attempt to bridge that gap. The NOSDDP contains 72400 biopsy slide images, from 4335 breast biopsies of 3425 unique patients, observed from 2014 to 2020. These slide images are linked to cancer stages and information about the mortality of patients (see Figure 1). Figure 2 presents the statics across the dataset, and its different splits (train and evaluation). The main takeaway is that most patients across the dataset are respectively in BC Stages 1, 2, and 0. Stages 1, and 0 are considered *early*, while Stage 2 can already be the characteristic of a *very advanced* BC. It is also important to notice that there are biopsies without stages (i.e. biopsies that have not been labeled with a specific cancer stage). In the next section, after describing our experimental setups which are: (1) the preprocessing of slide images, and (2) the pretrained CV models we use, we present the results and analyses from the different pretrained CV models.

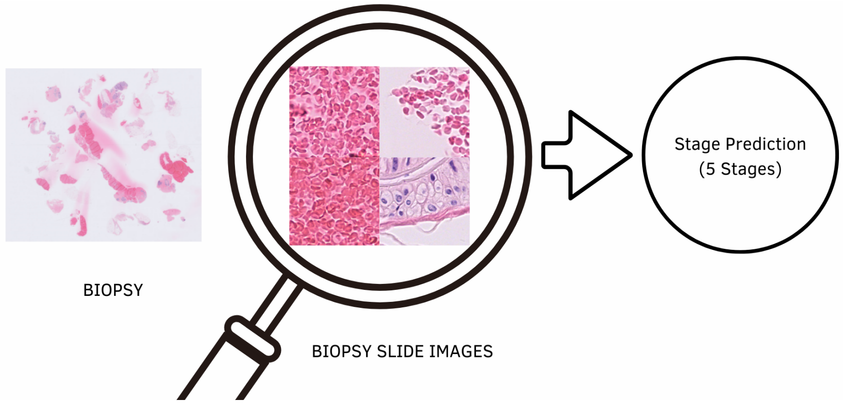

Figure 1: Description of the task: a biopsy generates slide images, that can be considered as different patches (parts) of the original biopsy. The cancer stage for the entire biopsy is the average of the individual stage prediction of each slide image.

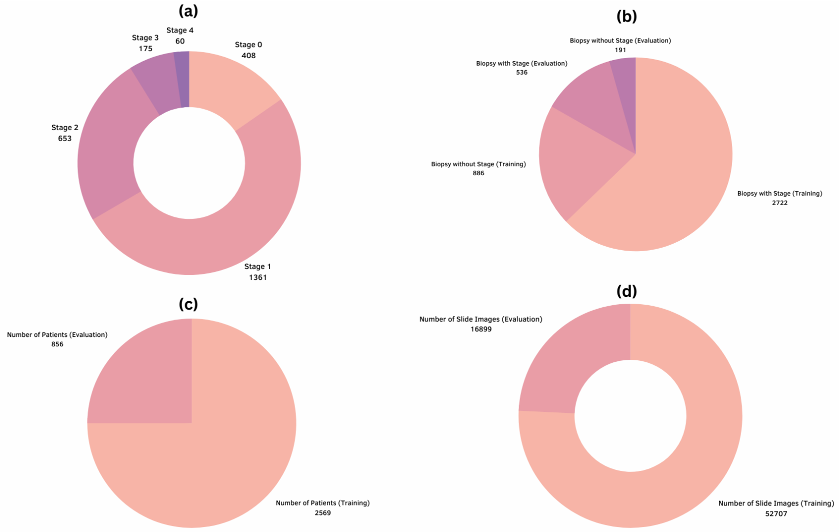

Figure 2: Statistics of the dataset: (a): Distribution of biopsies per cancer stage, (b): Distribution of labeled biopsies and non-labeled biopsies across the different dataset splits, (c): Number of patients used for each dataset split, and (d): Number of slide images corresponding to biopsies collected across the different dataset splits.

3 Experiments and Results

We make complete use of available data by assigning to the unlabeled biopsies, Cancer Stage 1 (the most frequent cancer stage). We then downsampled the high-resolution slide images to (224, 224), as supported by many existing pretrained CV models. The original training set has been split into two subsets (with 80:20 ratio): D_{Train} and D_{Eval}, while the initial evaluation set is used as a test dataset, we denote it D_{Test}. Image samples from D_{Train} have been augmented using randomized cropping and horizontal flipping, then normalized. The image samples from D_{Eval} and D_{Test} have only been normalized.

We finetuned 10 pretrained CV models: Resnet (18, 50, 152), EfficientNet Tan & Le (2021), ConvNext Liu et al. (2022) (M), WideResNet Zagoruyko & Komodakis (2016) (101), VGG Simonyan & Zisserman (2014), ResNext Xie et al. (2017) (101), RegNet Radosavovic et al. (2020) (X32GF), Swin Transformer Liu et al. (2021) (B), and MaxVit Tu et al. (2022). These models are all high-performing computer vision models, and their respective versions (or size) were chosen to cope with the memory available. For each model, we try various learning rates, with the *AdamW* optimizer. Each model was trained with a batch size of 32, and for 50 epochs: those were the efficient values for those hyperparameters. The predicted cancer stage PCS for a biopsy image B is the average of the predicted cancer stage for each subsequent slide image S:

$$PCS(B) = \frac{1}{|B|}\sum_{i} PCS(S_i)$$

where $|B|$ is the number of slides for the biopsy B, and $PCS(S_i) \in \{0, 1, 2, 3, 4\}$: this implies that the prediction for a biopsy image is a continuous value $\in [0, 4]$. Therefore, to measure the closeness to the true label which $\in \{0, 1, 2, 3, 4\}$, we use the Mean Square Error (MSE) metric:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(predictedstage_j - actualstage_j)^2$$

where the $actualstage \in \{0, 1, 2, 3, 4\}$. The different performances are summarized in Table 1.

Learning Rate	Resnet 18	Resnet 50	Resnet 152	EfficientNet (M)	ConvNext (Base)	Wide Resnet 101	VGG	RegNet	SwinT (B)	MaxVit
1e-4	1.009611	**0.970504**	1.005862	0.925831	1.009943	**0.931212**	**0.939045**	0.992109	0.898370	**0.855656**
1e-5	**1.001620**	1.008508	**1.001902**	0.987911	**0.957443**	1.012101	0.994687	**0.988756**	**0.897878**	0.893357
4e-4	1.020936	0.986704	1.007281	**0.829745**	1.110100	1.033937	1.107765	1.000450	1.231371	1.098370

Table 1: MSE of each pretrained CV model. Arranged per column (model), **the bold numbers represent the best performance of each model, with respect to learning rates**.

Looking at individual performances, we can see that EfficientNet performs better than other models. We believe this is due to their architecture, which combines families (like an ensemble) of individual baseline Neural Networks (NNs). Each NN uses a mobile inverted bottleneck convolution architecture, to optimize their respective accuracy and efficiency via neural architecture search Tan & Le (2021)[2]. Following the intuition of EfficientNet, we also tried to create a Deep Ensemble (E) of individual pretrained CV models (see Figure 3). Therefore, the predicted cancer stage of the ensemble model E, for a biopsy B is

$$PCS(B)_E = \frac{1}{|E|}\sum_{k=1}^{|E|} PCS(B)_k$$

where $PCS(B)_k$ is the $PCS(B)$ of the k-th model, and $|E|$ is the size of E i.e. the number of model composing E: the predicted cancer stage of the ensemble model E, for a biopsy B is hence the average of the predictions of each model of the ensemble E for that biopsy B.

We explore two strategies: (a) deep ensemble with all models, and (b) deep ensemble solely with the models which have an MSE lower than 1. We find out that in setup (a), the MSE is 0.632767 while

[2] EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling (GoogleAI Blog): `https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html`

in setup (b) the MSE is **0.5543481**. This shows that: (*) Deep Ensembles are better than individual models. This makes sense because in real-life, obtaining and aggregating the predictions of many experts (doctors) is better than relying on the opinion of a sole doctor. Moreover, each model learns different representations and features, which reduces the bias toward specific classes; and (**) the lower the losses of individual models, the lower the MSE of the Deep Ensemble.

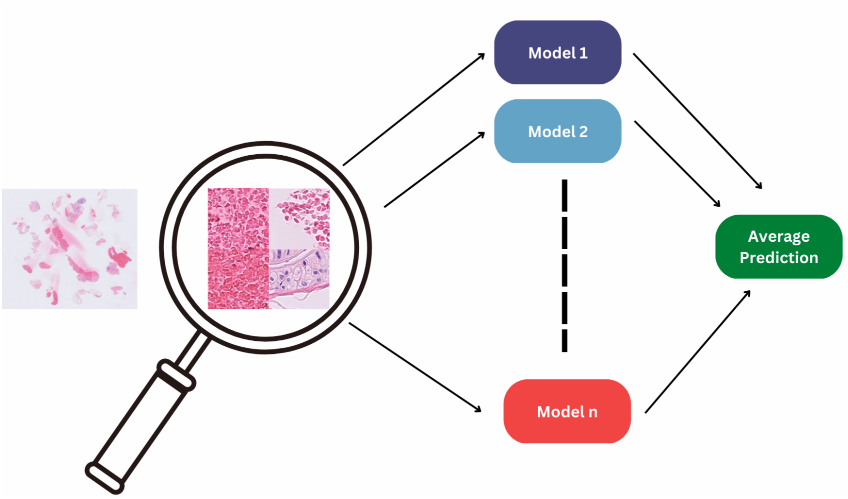

Figure 3: Deep Ensemble of each pretrained CV model: The average prediction for a biopsy is the average prediction of each model.

4 Looking Beyond Predictions: Exploring Causal Inference for a better Interpretability, Performance, and Generalization

It is fairly known that Neural Networks are black boxes. Many experimental high-performing models have failed when deployed in real-world settings, due to high distributional variation. Therefore, many efforts are being made to interpretability, and to enhance robustness. One of the subfields of ML research fields that could benefit medical imaging (the BC can be considered as a specific type of Medical Imaging problem) is Causal Inference. Causal Inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system. Causal Inference is very few to not explored in the Medical Imaging field, while its application could be very beneficial to the field of Medical Imaging. This is particularly important in the Medical Imaging context as it will allow answering the question ”*What if action x was performed?*”. This fits well the scope of precision healthcare whose goal is to target the right treatments to the right patients at the right time. Causal Inference would extremely benefit healthcare and precision medicine Sanchez et al. (2022); Vlontzos et al. (2022); Kaddour et al. (2022); Pölsterl & Wachinger (2021); Jesson et al. (2022) as it will enable to estimate of the Average Treatments Effects (ATE) while reducing structural bias and enhancing fairness across distribution and populations used in different studies. Causal Inference will also, therefore, increase robustness while allowing medical practitioners (e.g. doctors) to understand the predictions of the DL-based models, allowing more *”trust”*.

5 Conclusion

In this work, we explored the applicability of pretrained computer vision models, in the task of predicting high-risk breast cancer. We observed that Deep Ensemble models offer better performances, than single models. As future work, we opened discussions to causality in general and in the medical imaging context particularly; and hypothesized how it could help to build more robust models.

REFERENCES

Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113–108ra113, 2011. doi: 10.1126/scitranslmed.3002564. URL `https://www.science.org/doi/abs/10.1126/scitranslmed.3002564`.

Carlo Bifulco, Brian Piening, Tucker Bower, Ari Robicsek, Roshanthi Weerasinghe, Soohee Lee, Nick Foster, Nathan Juergens, Josh Risley, Katy Haynes, and Ziad Obermeyer. Identifying high-risk breast cancer using digital pathology images, 2021. URL `https://doi.org/10.48815/N5159B`.

Aimilia Gastounioti, Shyam Desai, Vinayak Ahluwalia, Emily Conant, and Despina Kontos. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. *Breast Cancer Research*, 24, 02 2022. doi: 10.1186/s13058-022-01509-z.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*, 2022.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Sendhil Mullainathan and Ziad Obermeyer. Solving medicine's data bottleneck: Nightingale open science. *Nature Medicine*, 28(5):897–899, May 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01804-4. URL `https://doi.org/10.1038/s41591-022-01804-4`.

Sebastian Pölsterl and Christian Wachinger. Estimation of causal effects in the presence of unobserved confounding in the alzheimer's continuum. In *International Conference on Information Processing in Medical Imaging*, pp. 45–57. Springer, 2021.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

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

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pp. 10096–10106. PMLR, 2021.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.

Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. A review of causality for learning algorithms in medical image analysis. *arXiv preprint arXiv:2206.05498*, 2022.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.