Received: 22 September 2022 | Revised: 6 June 2023 | Accepted: 10 June 2023

DOI: 10.1002/mp.16574

RESEARCH ARTICLE

MEDICAL PHYSICS

# A hierarchical self-attention-guided deep learning framework to predict breast cancer response to chemotherapy using pre-treatment tumor biopsies

**Khadijeh Saednia**[1,2] | **William T. Tran**[2,3,4] | **Ali Sadeghi-Naini**[1,2,4,5]

[1]Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, Ontario, Canada

[2]Department of Radiation Oncology, Sunnybrook Health Sciences Center, Toronto, Ontario, Canada

[3]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada

[4]Temerity Centre for AI Research and Education in Medicine, University of Toronto, Toronto, Ontario, Canada

[5]Physical Sciences Platform, Sunnybrook Research Institute, Toronto, Ontario, Canada

**Correspondence**
Dr. Ali Sadeghi-Naini, Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada.
Email: asn@yorku.ca

## Abstract

**Background:** Pathological complete response (pCR) to neoadjuvant chemotherapy (NAC) has demonstrated a strong correlation to improved survival in breast cancer (BC) patients. However, pCR rates to NAC are less than 30%, depending on the BC subtype. Early prediction of NAC response would facilitate therapeutic modifications for individual patients, potentially improving overall treatment outcomes and patient survival.

**Purpose:** This study, for the first time, proposes a hierarchical self-attention-guided deep learning framework to predict NAC response in breast cancer patients using digital histopathological images of pre-treatment biopsy specimens.

**Methods:** Digitized hematoxylin and eosin-stained slides of BC core needle biopsies were obtained from 207 patients treated with NAC, followed by surgery. The response to NAC for each patient was determined using the standard clinical and pathological criteria after surgery. The digital pathology images were processed through the proposed hierarchical framework consisting of patch-level and tumor-level processing modules followed by a patient-level response prediction component. A combination of convolutional layers and transformer self-attention blocks were utilized in the patch-level processing architecture to generate optimized feature maps. The feature maps were analyzed through two vision transformer architectures adapted for the tumor-level processing and the patient-level response prediction components. The feature map sequences for these transformer architectures were defined based on the patch positions within the tumor beds and the bed positions within the biopsy slide, respectively. A five-fold cross-validation at the patient level was applied on the training set (144 patients with 9430 annotated tumor beds and 1,559,784 patches) to train the models and optimize the hyperparameters. An unseen independent test set (63 patients with 3574 annotated tumor beds and 173,637 patches) was used to evaluate the framework.

**Results:** The obtained results on the test set showed an AUC of 0.89 and an F1-score of 90% for predicting pCR to NAC a priori by the proposed hierarchical framework. Similar frameworks with the patch-level, patch-level + tumor-level, and patch-level + patient-level processing components resulted in AUCs of 0.79, 0.81, and 0.84 and F1-scores of 86%, 87%, and 89%, respectively.

**Conclusions:** The results demonstrate a high potential of the proposed hierarchical deep-learning methodology for analyzing digital pathology images of pre-treatment tumor biopsies to predict the pathological response of breast cancer to NAC.

**KEYWORDS**
attention mechanism, breast cancer, deep learning, digital pathology (DP), neoadjuvant chemotherapy, transformers

## 1 | INTRODUCTION

Breast cancer (BC) is the most common cancer in women and the second leading cause of cancer-related death worldwide.[1] About 20% of breast cancer patients are diagnosed with locally advanced breast cancer (LABC).[2,3] LABC includes stage III and a subset of stage IIB BC and may extend to the skin or chest wall or involve the axillary lymph nodes.[4–6] LABC patients typically have a poorer prognosis than early-stage BC due to the high risk of cancer progression, local recurrence and metastasis.[7–9] The standard of care to treat LABC involves neoadjuvant chemotherapy (NAC) followed by surgery and in select cases, adjuvant therapies, such as radiotherapy, endocrine therapy and targeted drugs.[10,11] Despite multimodal treatment plans, LABC patients exhibit low overall survival, and outcomes are highly dependent on tumor response to NAC[12], thus pathological response is correlated to survival.[13] However, only about 30% of LABC patients achieve a pathological complete response (pCR) to NAC.[14,15] The definitive method for pathological assessment of NAC response is histopathology on the surgical excisions. This limits the opportunity to modify NAC based on tumor response. Therefore, predicting chemotherapy response either before or during early intervals of NAC could improve therapy outcomes by facilitating response-guided drug treatments.[16,17]

Prior research has explored various quantitative imaging approaches for assessing chemotherapy response in breast cancer patients at the time of diagnosis or early after starting treatments.[17–20] Quantitative biomarkers derived using different medical imaging modalities, including ultrasound, diffuse optical imaging, magnetic resonance imaging (MRI), and positron emission tomography (PET), have shown promise in characterizing breast cancer in terms of responsiveness to chemotherapy, particularly when coupled with machine learning (ML) models.[21–24] However, the histopathological assessment remains the gold standard to report a cancer diagnosis and characterize tumors to steer treatment decisions. In a recent study, digitized diagnostic tumor biopsies were applied to predict response to NAC in BC patients.[25] Several pathomic features were extracted from segmented nuclei in digital pathology images. A gradient boosting machine (GBM) was adapted for NAC response prediction. The results show the potential of quantitative information derived from the pre-treatment digital pathology images for predicting NAC outcomes.

Recent research has explored the efficacy of deep learning (DL) methods in various medical image analysis applications.[26,27] Studies have investigated the performance of several DL frameworks in analyzing histopathological images for tumor grading and subtyping and predicting patient survival.[28–31] These studies have demonstrated the potential of the DL models and, specifically, the deep convolutional networks in digital pathology image analysis. However, the size of whole slide images (WSIs) remains a constraint in training the DL models due to memory limitations. Specifically, developing adequate DL models for analyzing the WSIs at high magnification is challenging as the high-resolution WSIs can be as large as 150000 × 150000 pixels.[32,33] This issue can be addressed by resizing (downscaling) the WSIs or extracting small patches for analysis by the DL models. Downscaling reduces the image resolution and potentially the efficacy of the information derived by the model. The patching approach requires the development of efficient strategies for fusing the patch-level information while considering the global dependencies to make relevant conclusions on the WSI level.

Several recent studies have focused on developing efficient DL models for image analysis and classification in computer vision applications. The current state-of-the-art models are categorized into three general approaches. The first category includes the convolutional neural network (CNN) models. The Xception model,[34] a well-known model in this group, is an extended version of the Inception-V3 architecture[35] in which the Inception modules have been replaced by depth-wise separable convolutions. This model could outperform the Inception-V3 model in classifying the ImageNet dataset.[36] The second category comprises the transformers[37] that utilize an encoder-decoder architecture with a self-attention mechanism.[38] The self-attention mechanism in the transformers differentially weights the significance of each part of the input data for the target analysis. The recently introduced vision transformer (ViT) architecture[39] has demonstrated high performance in extracting the global relations in the input images.[40] However, compared to

the CNN-based models, this architecture requires additional training data to yield adequate generalizability.[41] The last category includes the networks that combine the convolutional layers with the attention mechanisms. The convolutional block attention module (CBAM),[42] a recently proposed attention mechanism, can infer independent channel and spatial attention maps. The CoAtNet[43] is the most recent state-of-the-art architecture that stacks depth-wise convolutional layers and the self-attention mechanism to improve the generalizability of the model.[43]

This study proposes a hierarchical attention-guided deep learning framework to predict pCR to NAC in breast cancer patients using digital histopathological images of pre-treatment biopsy specimens. The proposed model consists of three modules, including patch-level processing, tumor-level processing, and patient-level response prediction. The hierarchical flow presented in this study could overcome the difficulty of deriving the global relations between different tumor areas in high-resolution WSIs while utilizing the local information within the tumor regions in the analysis.

## 2 | MATERIALS AND METHODS

### 2.1 | Dataset

This retrospective study was conducted following institutional ethics review board (IRB) approval at Sunnybrook Health Sciences Centre, Toronto, Canada. Patients were included in the study based on the following criteria: biopsy-confirmed diagnosis of invasive breast cancer, age (18+), and treatment with Anthracycline- and Taxane-based neoadjuvant chemotherapy followed by surgery (any type). There were 207 patients identified and included in the study. The mean age of patients was $51.1 \pm 10.4$ years (range: $28 - 79$ years). The mean tumor size was $5.01 \pm 2.9$ *cm*. The clinical nodal (N) stage was N0 (no positive lymph nodes) for 24.6% ($n = 51$); N1 ($1 - 3$ positive lymph nodes) for 66.2% ($n = 137$); N2 ($4 - 9$ positive lymph nodes) for 8.2% ($n = 17$); and N3 (≥ten positive lymph nodes) for 1% ($n = 2$) of patients. Among 207 patients, 62.3%, 54.1%, 62.3% and 41.5% had tumors with an ER+, PR+, and HER2+ receptor status, respectively. Most of the patients ($n = 192$; 93%) had invasive ductal carcinoma (IDC), while those with invasive lobular carcinoma (ILC) constituted a smaller cohort ($n = 15$; 7%).

All patients had a breast core needle biopsy before NAC with a pathological review as part of the standard of care. The formalin-fixed paraffin-embedded (FFPE) blocks containing core biopsy specimens were microtomed into 4 $\mu m$ sections. The specimen sections were prepared onto histopathology slides and stained with hematoxylin and eosin (H&E). Digital histopathological images of the H&E-stained slides were acquired using a digital pathology imaging system (Huron Digital Pathology, St. Jacob's, Canada). The images were acquired at 40X (pixel size: 0.2 $\mu m$) for all patients. The digital WSIs were manually reviewed for artifacts or occlusions within the specimen before analysis, and any slides associated with a distorted or blurry image were re-imaged.

The treatment response was assessed for each patient after surgery and categorized into pathological complete response (pCR) versus pathological non-complete response (non-pCR; i.e., exhibiting residual disease) as ground truth labels for evaluating the developed models. A standard assessment method using the residual cancer burden index (RCBI) was applied to assess treatment response. An RCBI score of 0 (i.e., pCR) was defined as the absence of residual invasive and nodal disease.[44] Patients who demonstrated residual disease were classified as non-pCR (i.e., RCBI $> 0$).[44] All pathology reviews (pre-treatment and post-surgery histopathology) were performed by board-certified breast pathologists as part of the patient's standard of care. The pathological evaluations after surgery demonstrated 25.2% ($n = 52$) of the patients with a pCR and 74.8% ($n = 155$) with a non-pCR.

### 2.2 | Preprocessing and dataset splitting

Using the Sedeen software package,[45] an expert pathologist annotated the tumor bed areas on the WSIs. The tumor bed annotations were pre-processed on the three-channel RGB images for patch extraction. Tumor margins were included, when required, to obtain non-overlapping patches with a size of $512 \times 512$ pixels (pixel size: 0.2 $\mu m$) from the annotated tumor regions. However, only the patches with more than 10% tumor tissue and less than 10% white background were retained for the study (Figure 1). In total, 1,733,421 patches were included in this study.

Out of the 207 patients in this study, 155 and 52 patients were labeled as non-pCR and pCR, respectively. The number of patches for each patient varied from 30 to 14418 (*median* $= 1755$). The non-pCR and pCR classes included 1,423,210 and 310,211 patches, respectively. As such, the ratio of the pCR to non-pCR class at the patient and patch level was 25.2% and 17.9%, respectively. Figure 2a shows the distribution and quartiles of the number of patches per patient in the pCR and non-pCR classes of the dataset. A stratified random splitting approach was applied to split the data at the patient level into the training (70%; $n = 144$ with 9,430 annotated tumor beds and 1,559,784 patches) and test sets (30%; $n = 63$ with 3,574 annotated tumor beds and 173,637 patches). The stratified random splitting was performed considering the response label and the number of extracted patches for each patient. The first quartile, median and third quartile of the number of patches in each response class (pCR and non-pCR)

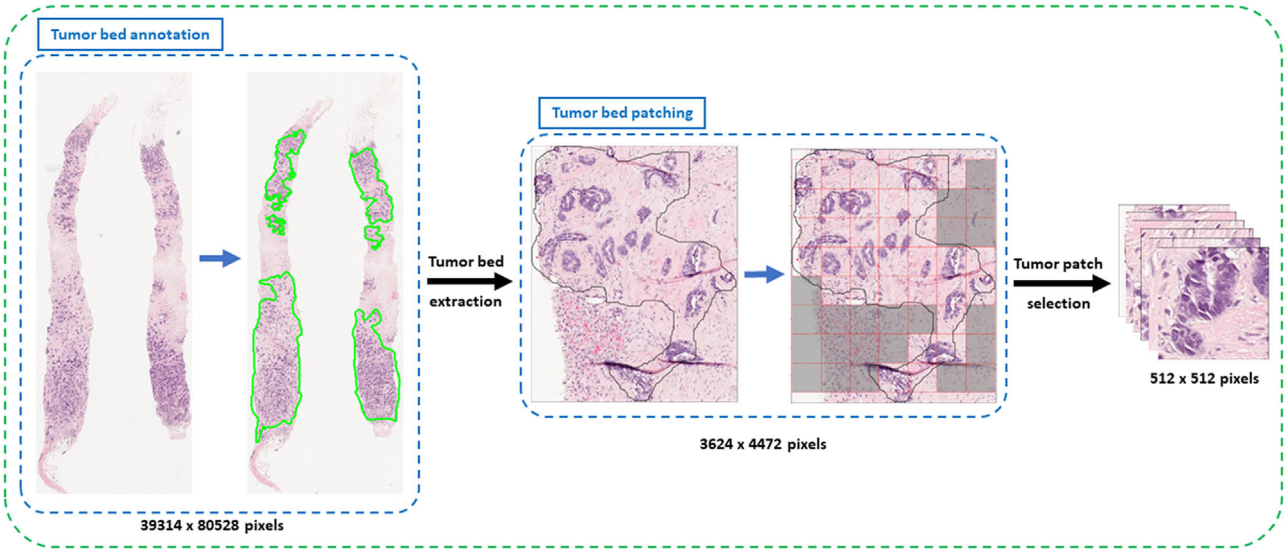
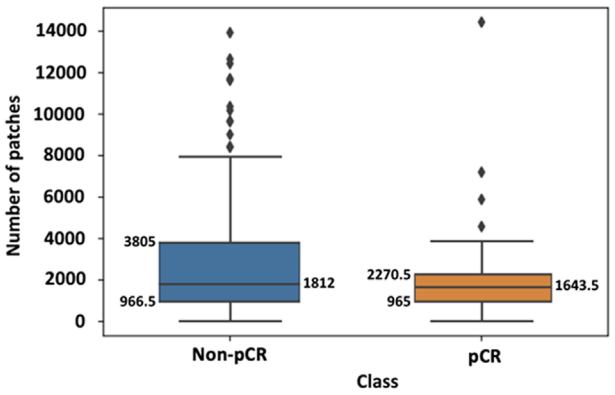



**FIGURE 1** Overview of preprocessing steps. The tumor beds were annotated (green contours) by an expert pathologist. Patches (size = 512 × 512 pixels) were extracted from the tumor beds. Patches with more than 10% tumor tissues and fewer than 10% white background were retained for the study.




| Class (pCR = 0, non-pCR = 1) | Criterion Discretized number of patches | Stratifying bin number: 4×Class + criterion |
|---|---|---|
| 0 | 0 (< quartile1) | 0 |
| 0 | 1 (>quartile1, <median) | 1 |
| 0 | 2 (>median, <quartile3) | 2 |
| 0 | 3 (>quartile3) | 3 |
| 1 | 0 (< quartile1) | 4 |
| 1 | 1 (>quartile1, <median) | 5 |
| 1 | 2 (>median, <quartile3) | 6 |
| 1 | 3 (>quartile3) | 7 |

**FIGURE 2** Stratified random data splitting at the patient level based on the number of patches in each dataset class. (a) The box plot presents the distribution and quartiles of the number of patches per patient in each class. (b) The criterion for stratification.

were used to stratify the patients into eight different bins for random sampling, as shown in Figure 2b. A similar procedure was utilized during five-fold cross-validation on the training set to optimize the framework's hyperparameters (described further in Section 2.4). The training and test sets obtained using the stratified random splitting approach were assessed, using statistical tests, for possible inhomogeneities in terms of clinical feature distribution between the two sets. A Pearson's chi-squared homogeneity test was used for categorical variables. The continuous variables were assessed using a *t*-test. Table 1 presents the cohort's clinical information in the training and test sets, and the results of associated statistical analysis.

## 2.3 | Response prediction framework

Figure 3 demonstrates the scheme of the proposed hierarchical deep-learning framework for therapy response prediction. The patient-wise sampler in the framework addresses the imbalance issue of the training data at patient and patch levels. The sampler applies a weighted sampling strategy for under-sampling the majority class in the training set based on the total number of patches and the number of patients in each class:

$$w_p = \frac{S_{pt}}{S_{pp} \times S_{pc}}$$

**TABLE 1** Demographic and clinical characteristics of the patients. The distribution of each variable was compared between the training and test sets using a Pearson's chi-squared homogeneity test for categorical variables and a *t*-test for continuous variables; the p-values are reported in the last column.

| Patient demographics and clinicopathological characteristics | Count (%) | | | | *p*-value |
|---|---|---|---|---|---|
| | Training ($n = 144$) | | Test ($n = 63$) | | |
| | pCR ($n = 36$) | non-pCR ($n = 108$) | pCR ($n = 16$) | non-pCR ($n = 47$) | $p = 0.91$ |
| Median age (Years) | 47.5 | 51 | 49.5 | 51 | $p = 0.61$ |
| *Menopausal status* | | | | | |
| Pre/Peri-menopausal | 15 (42%) | 56 (52%) | 9 (56%) | 24 (51%) | $p = 0.69$ |
| Post-menopausal | 21 (58%) | 52 (48%) | 7 (44%) | 23 (49%) | |
| *Tumor Size* | | | | | |
| Mean tumor size (mm; ± SD) | 39.8 ± 21.1 | 55.8 ± 29.5 | 37.1 ± 23.9 | 49.6 ± 30.7 | $p = 0.22$ |
| *Nodal status (N Stage)* | | | | | |
| No positive lymph nodes (N0) | 14 (39%) | 22 (21%) | 7 (44%) | 8 (17%) | $p = 0.59$ |
| 1–3 Positive lymph nodes (N1) | 21 (58%) | 75 (69%) | 8 (50%) | 33 (70%) | |
| 4–9 Positive lymph nodes (N2) | 1 (3%) | 10 (9%) | 1 (6%) | 5 (11%) | |
| ≥10 Positive lymph nodes (N3) | 0 (0%) | 1 (1%) | 0 (0%) | 1 (2%) | |
| *Receptor status* | | | | | |
| ER positive | 13 (36%) | 79 (73%) | 3 (19%) | 34 (72%) | $p = 0.81$ |
| PR positive | 12 (33%) | 68 (63%) | 2 (13%) | 30 (64%) | $p = 0.27$ |
| HER2 positive | 24 (67%) | 38 (35%) | 10 (63%) | 14 (30%) | $p = 0.51$ |
| *Histology* | | | | | |
| Invasive ductal carcinoma | 36 (100%) | 96 (89%) | 16 (100%) | 44 (94%) | $p = 0.53$ |
| Invasive lobular carcinoma | 0(0%) | 12 (11%) | 0 (0%) | 3 (6%) | |
| *Nottingham grade* | | | | | |
| 1 | 1 (3%) | 4 (4%) | 0 (0%) | 1 (2%) | $p = 0.47$ |
| 2 | 9 (25%) | 39 (36%) | 2 (13%) | 22 (47%) | |
| 3 | 26 (72%) | 65 (60%) | 14 (87%) | 24 (51%) | |
| *Other clinical information* | | | | | |
| Inflammatory breast cancer | 2 (6%) | 11 (10%) | 2 (13%) | 5 (11%) | $p = 0.64$ |

where, $w_p$ is the calculated weight for each patient, $S_{pt}$ is the total number of patches in the training set, $S_{pp}$ is the total number of patches for the patient, and $S_{pc}$ is the total number of patches in the associated class of the patient (pCR or non-pCR). After calculating the weights for all patients in the training set, they are normalized such that the sum of all weights adds up to one. All patches of each patient are assigned equal weights for selection by the sampler. The sampler size (number of patches in each epoch) is tuned as a hyperparameter during the system's training process.

The framework includes a hierarchical flow of patch-level processing, tumor-level processing, and patient-level response prediction. It utilizes a self-attention-guided convolutional network architecture and two customized ViT network architectures for hierarchical deep-learning-based analysis of the digital pathology images for NAC response prediction. In the patch-level processing module, a modified CoAtNet model with two convolutional components and two self-attention modules (Figure 3b) extracts the descriptive features from the pathology image patches. The input patches are downsampled to 256 × 256 pixels using a two-layer convolutional block. A feature map of size 768 associated with each patch is collected from the last layer (global pooling layer) before the fully connected layer in the network. The sequence of feature maps for each tumor bed is generated by stacking the sorted patch-level feature maps of the corresponding tumor bed annotation in the associated digital pathology image. Specifically, the feature maps are sorted based on the position of their associated patch in the tumor bed from top left to bottom right. The generated feature map sequence is fed as input to the first ViT model that includes sixteen encoder blocks (Figure 3c) to explore the relations in aggregated features of each tumor bed. The initial positional encoding vector for the tumor-level processing module is defined based on the generated sequence of the sorted patches. To adapt the sequence of feature maps for the ViT architecture, the input size is defined as the (sequence length × size of feature map vector). Since the number of patches varies for different tumor beds, a

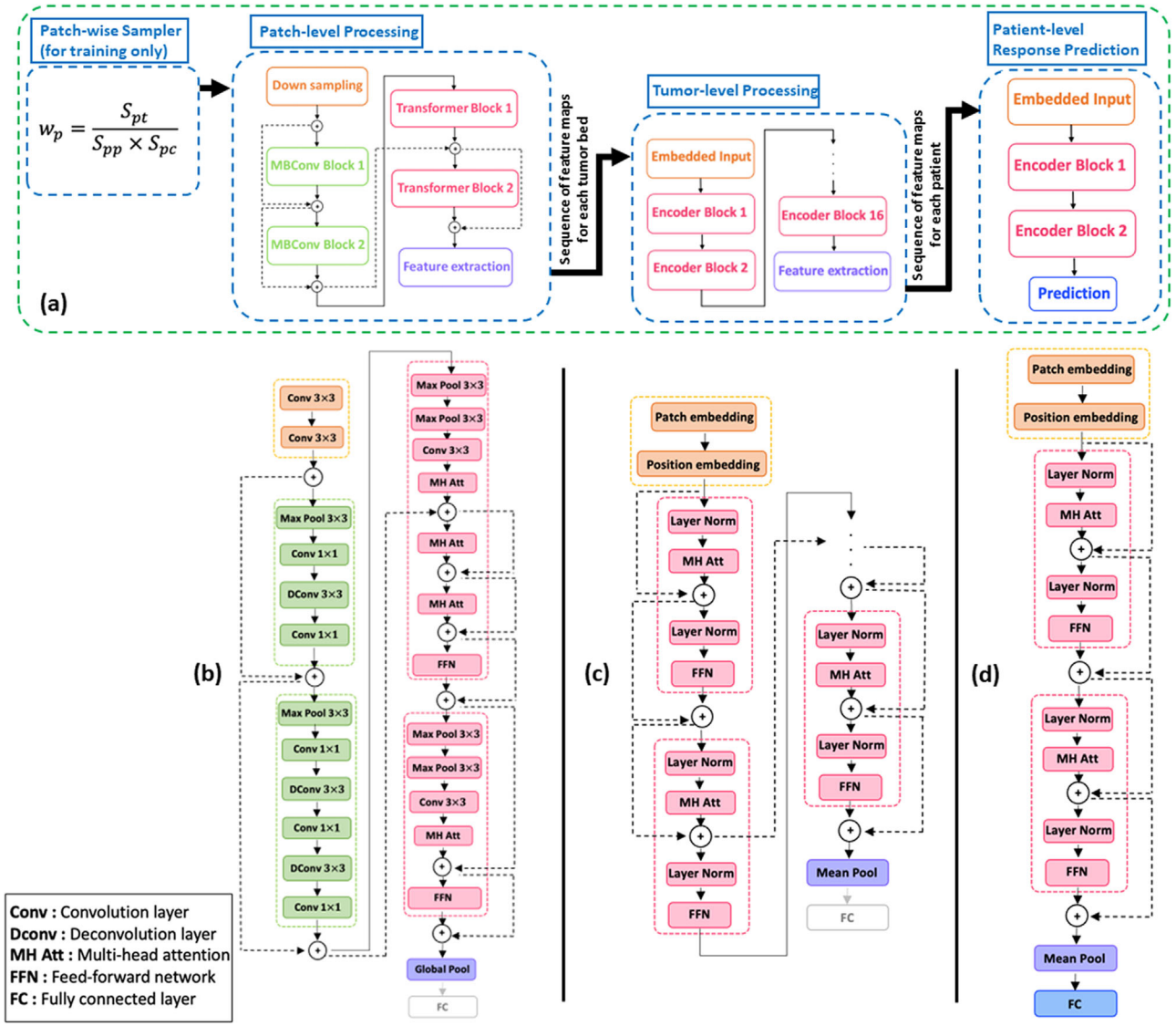



**FIGURE 3** Schematic representation of the proposed hierarchical deep-learning framework for NAC outcome prediction. The framework (a) consists of a patch-level sampler for training and three levels of processing. Detailed architectures of the processing modules are shown for the patch-level processing module (b), tumor-level processing module (c), and patient-level response prediction module (d). The component colors in (b), (c) and (d) show the associated block in (a).

threshold is set for the maximum length of the sequence (tuned as a hyperparameter during the system's training process). For the tumor beds with a smaller number of patches than the threshold, a zero-masking approach is applied, while for the beds with a larger number of patches, the starting patch of the sequence is randomly selected. A similar method is used in the patient-level response prediction network (Figure 3d) to aggregate the information obtained at the tumor-level processing while exploiting the global dependencies between the tumor areas for therapy response prediction. A feature map of size 1024 associated with each tumor bed is extracted from the last layer (mean pooling layer) of the tumor-level processing ViT before the fully connected layer. A sequence of the feature maps is generated for each patient to feed into the patient-level response prediction ViT network that includes two encoder blocks. Similar feature map sorting and sequence length thresholding strategies applied for the tumor-level processing ViT are used to create the input sequence of this ViT network.

## 2.4 | System training

The framework's hyperparameters were optimized using a grid search approach with five-fold cross-validation on the training set. The sampler size was tuned to 10000 patches for each training epoch. The number of patches per tumor bed and tumor bed regions

per patient varied between 1 to 3879 (*median* = 6) and 1 to 285 (*median* = 44), respectively. Accordingly, the threshold for the maximum length of the sequence in the tumor-level and patient-level processing modules was tuned to 10 and 50, respectively. These thresholds resulted in an input size of 10 × 768 and 50 × 1024 for the associated networks, respectively. The batch size was tuned to 200 for all three networks. A maximum number of 100 epochs was used for training each network. The learning rate was tuned to 0.05 for the patch-level and tumor-level processing networks and 0.01 for the patient-level response prediction network. After the hyperparameter tuning, 20% of the training set patients ($n$ = 29) were selected as the validation set using the stratified random splitting approach, and the framework with optimized hyperparameters was trained using the remaining patients in the training set (80% of the training set, $n$ = 115). Early stopping was utilized for all networks based on the validation loss during training to prevent overfitting.

## 2.5 | Evaluation

The performance of the proposed framework was assessed on the independent unseen test set using the accuracy, sensitivity, specificity, loss, F1-score, and area under the receiver operating characteristics (ROC) curve (AUC). A threshold value of 0.5 was used as the cut-off to calculate the sensitivity and specificity. To evaluate the performance and effectiveness of the proposed hierarchical framework, ablation experiments were performed with models that only incorporated the patch-level processing, patch-level + tumor-level, and patch-level + patient-level processing modules. In another set of experiments, maximum voting was applied on the output of the patch-level and patch-level + tumor-level models to obtain the tumor-level and patient-level response prediction results for comparison with those of the hierarchical models. Specifically, the maximum voting was applied over the responses (pCR/non-pCR) predicted by the patch-level or patch-level + tumor-level model for all the patches/tumor regions associated with a tumor region/patient. Separate experiments were conducted using two other frameworks with different network architectures for patch-level processing. The first framework utilized an Xception model coupled with CBAM attention as a state-of-the-art CNN-based model, while the second framework applied a pure self-attended architecture with a ViT model for processing the patches.

A gradient class activation map (Grad-CAM) approach[46] was applied to visualize the trained patch-level processing attention mechanisms. The Grad-CAM approach provides information on salient regions in an image for a specific class to permit interpreting the network decisions based on the model attention. The attention heatmaps were generated for the patches presented to the trained framework based on their predicted label. The generated heatmaps were stitched together using their position information to create complete attention maps for individual tumor areas. The visualization heatmaps were reviewed to assess and compare the efficacy of attention mechanisms in different networks.

## 3 | RESULTS

Table 2 shows the performance of the proposed framework compared to other similar models. The results demonstrate that the CoAtNet architecture as the patch-level processing module outperformed the Xception model with CBAM attention and the ViT model, with an accuracy of 81% on the test set, compared to 79% and 78%, respectively. Results of the ablation experiments demonstrate that the three-level hierarchical frameworks could outperform the patch-level only and the two-level (patch-level + tumor-level and patch-level + patient-level) processing frameworks. The patch-level + tumor-level and patch-level + patient-level processing frameworks consisting of cascaded ViT models resulted in an AUC of 0.78 and 0.77, respectively, on the test set. In contrast, a similar architecture with a three-level hierarchy resulted in an AUC of 0.80. Also, the two-level hierarchical models (patch-level + tumor-level and patch-level + patient-level) with the Xception + CBAM architecture followed by the ViT module could achieve an AUC of 0.80 and 0.82, respectively. In contrast, a similar model with three levels of processing could achieve an AUC of 0.86. The tumor-level and patient-level results obtained through maximum voting on the outputs of the patch-level and patch-level + tumor-level models demonstrate a better performance of the corresponding hierarchical models in response prediction. The best performance was achieved by the proposed framework with a three-level hierarchy consisting of the CoAtNet architecture as a patch-level processing module and two ViT architectures for the tumor-level processing and patient-level response prediction. This model resulted in an accuracy of 86% on the test set and a sensitivity, specificity, F1-score and AUC of 87%, 83%, 90% and 0.89, respectively.

Figure 4 compares the AUC of the two-level and three-level hierarchical architectures with different patch-level processing modules (ViT, Xception+CBAM, and CoAtNet). The AUCs range between 0.77 and 0.89, with the best results associated with the three-level hierarchical framework using the CoAtNet as the patch-level processing component.

Figure 5 presents the attention heatmaps obtained for two representative tumor regions: one with a pCR and the other with a non-pCR outcome after NAC. The heatmaps were generated for each patch in the

**TABLE 2** Performance of different architectures in predicting NAC response using pre-treatment digitized pathology slides of core needle biopsies. The best value in each column is in bold. Acc: accuracy, Sens: sensitivity, Spec: specificity, AUC: area under the ROC curve.

| | | Training set | | | | Validation set | | | | Independent test set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | | **Loss** | **Acc** | **Sens** | **Spec** | **Loss** | **Acc** | **Sens** | **Spec** | **Loss** | **Acc** | **Sens** | **Spec** | **AUC** | **F1-score** |
| **ViT + (None/ViT) + (None/ViT)** | **Patch-level** | 0.61 | 0.76 | 0.76 | 0.75 | 0.61 | 0.77 | 0.78 | 0.74 | 0.59 | 0.78 | 0.79 | 0.75 | 0.78 | 0.84 |
| | **Max-voting on Patch-level (Tumor-level Results)** | - | 0.73 | 0.73 | 0.73 | - | 0.74 | 0.75 | 0.72 | - | 0.74 | 0.75 | 0.73 | - | 0.82 |
| | **Max-voting on Patch-level (Patient-level Results)** | - | 0.72 | 0.73 | 0.72 | - | 0.72 | 0.73 | 0.71 | - | 0.73 | 0.74 | 0.72 | - | 0.81 |
| | **Patch-level + Tumor-level** | 0.60 | 0.77 | 0.78 | 0.75 | 0.60 | 0.78 | 0.79 | 0.75 | 0.59 | 0.79 | 0.80 | 0.75 | 0.78 | 0.85 |
| | **Max-voting on Patch-level + Tumor-level (Patient-level Results)** | - | 0.76 | 0.77 | 0.75 | - | 0.76 | 0.77 | 0.74 | - | 0.78 | 0.79 | 0.75 | - | 0.84 |
| | **Patch-level + Patient-level** | 0.59 | 0.79 | 0.79 | 0.78 | 0.60 | 0.78 | 0.79 | 0.76 | 0.57 | 0.80 | 0.81 | 0.78 | 0.77 | 0.86 |
| | **Patch-level + Tumor-level + Patient-level** | 0.57 | 0.80 | 0.80 | 0.79 | 0.58 | 0.79 | 0.80 | 0.76 | 0.55 | 0.82 | 0.83 | 0.79 | 0.80 | 0.87 |
| **Xception (+CBAM) + (None/ViT) + (None/ViT)** | **Patch-level** | 0.57 | 0.77 | 0.80 | 0.74 | 0.58 | 0.77 | 0.79 | 0.73 | 0.56 | 0.79 | 0.80 | 0.76 | 0.79 | 0.85 |
| | **Max-voting on Patch-level (Tumor-level Results)** | - | 0.76 | 0.78 | 0.75 | - | 0.77 | 0.78 | 0.74 | - | 0.78 | 0.79 | 0.76 | - | 0.84 |
| | **Max-voting on Patch-level (Patient-level Results)** | - | 0.76 | 0.77 | 0.76 | - | 0.76 | 0.77 | 0.74 | - | 0.76 | 0.77 | 0.75 | - | 0.83 |
| | **Patch-level + Tumor-level** | 0.55 | 0.80 | 0.81 | 0.77 | 0.56 | 0.80 | 0.81 | 0.75 | 0.55 | 0.81 | 0.82 | 0.77 | 0.80 | 0.86 |
| | **Max-voting on Patch-level + Tumor-level (Patient-level Results)** | - | 0.78 | 0.79 | 0.77 | - | 0.79 | 0.80 | 0.75 | - | 0.79 | 0.80 | 0.76 | - | 0.85 |
| | **Patch-level + Patient-level** | 0.54 | 0.82 | 0.83 | 0.79 | 0.53 | 0.82 | 0.84 | 0.79 | 0.53 | 0.83 | 0.84 | 0.81 | 0.82 | 0.88 |
| | **Patch-level + Tumor-level + Patient-level** | 0.52 | 0.84 | 0.84 | **0.84** | **0.51** | **0.84** | **0.85** | **0.82** | 0.50 | 0.85 | 0.86 | **0.83** | 0.86 | **0.90** |
| **CoAtNet + (None/ViT) + (None/ViT)** | **Patch-level** | 0.56 | 0.78 | 0.80 | 0.76 | 0.58 | 0.78 | 0.80 | 0.74 | 0.54 | 0.81 | 0.82 | 0.78 | 0.79 | 0.86 |
| | **Max-voting on Patch-level (Tumor-level Results)** | - | 0.76 | 0.77 | 0.76 | - | 0.77 | 0.78 | 0.74 | - | 0.78 | 0.80 | 0.77 | - | 0.85 |
| | **Max-voting on Patch-level (Patient-level Results)** | - | 0.76 | 0.77 | 0.75 | - | 0.76 | 0.77 | 0.74 | - | 0.77 | 0.79 | 0.76 | - | 0.84 |
| | **Patch-level + Tumor-level** | 0.54 | 0.80 | 0.83 | 0.78 | 0.55 | 0.80 | 0.81 | 0.76 | 0.53 | 0.82 | 0.83 | 0.78 | 0.81 | 0.87 |
| | **Max-voting on Patch-level + Tumor-level (Patient-level Results)** | - | 0.78 | 0.80 | 0.75 | - | 0.79 | 0.80 | 0.76 | - | 0.80 | 0.81 | 0.78 | - | 0.86 |
| | **Patch-level + Patient-level** | 0.52 | 0.83 | 0.85 | 0.80 | 0.56 | 0.81 | 0.83 | 0.78 | 0.51 | 0.85 | 0.86 | 0.80 | 0.84 | 0.89 |
| | **Patch-level + Tumor-level + Patient-level** | **0.48** | **0.85** | **0.86** | **0.84** | 0.52 | **0.84** | **0.85** | 0.81 | **0.48** | **0.86** | **0.87** | **0.83** | **0.89** | **0.90** |

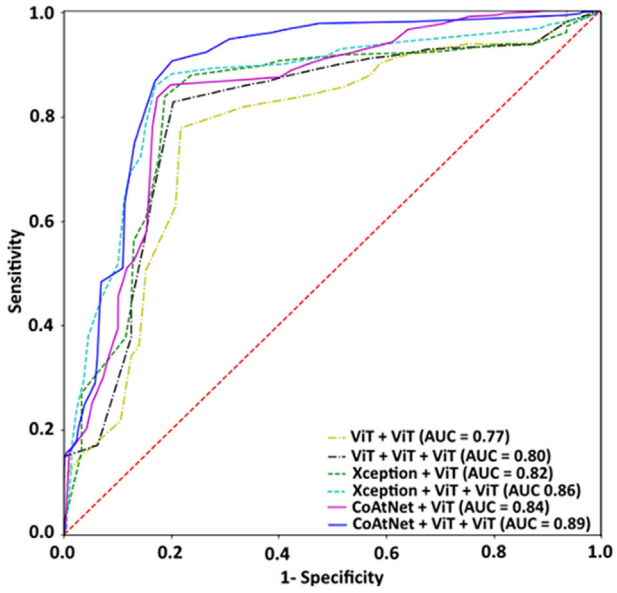



**FIGURE 4** Receiver operating characteristic (ROC) curves on the independent test set for the predictive models developed with two-level (patch-level + patient-level: ViT + ViT, Xception + ViT, and CoAtNet + ViT) and three-level (patch-level + tumor-level + patient-level: ViT + ViT + ViT, Xception + ViT + ViT, and CoAtNet + ViT + ViT) hierarchical framework. Different patch-level processing modules were utilized for comparison.

tumor bed region (contoured by an expert pathologist) using the Grad-CAM approach and overlaid on the original pathology image for visualization. Comparing the heatmaps associated with different patch-level processing modules shows that the CoAtNet architecture has focused more on the tumor regions than the other two networks.

## 4 | DISCUSSION

In this study, a hierarchical deep learning framework was developed to predict breast cancer response to NAC using digital histopathological images of pre-treatment biopsy specimens. The proposed model consists of a patch-level processing module followed by a tumor-level processing module and a patient-level response prediction module. A self-attention-guided convolutional network based on the CoAtNet architecture was adapted for the patch-level processing step, with two ViT models for aggregating the sequence of feature maps at the tumor level and predicting the therapy response, respectively. The patch-level processing module was applied to capture the local correlations within the tumor microenvironment and extract the feature maps carrying relevant information of the pathology image patches. The tumor-level processing step was used to aggregate the local information for each tumor region. The patient-level prediction module utilized the sequence of information for various tumor regions to derive the global relations and predict the patient response to NAC. The proposed model could predict NAC response of patients in an independent test set with a sensitivity, specificity, and F1-score of 87%, 83%, and 90%, respectively. Comparing the attention heatmaps generated by various patch-level processing architectures demonstrates that the CoAtNet architecture paid more attention to the tumor areas than the surrounding healthy tissues. The proposed hierarchical model's performance shows that the combination of convolutional blocks with self-attention modules in the patch-level processing component can effectively extract local information within the tumor patches. Further, the relations between these features in each tumor area and at the patient level for NAC response prediction can be successfully modeled using the vision transformer modules.

The proposed hierarchical framework provides an effective approach for analyzing the WSIs at high resolution. Results of the ablation experiments in this study demonstrate that the three-level hierarchical model could outperform the patch-level only and the two-level hierarchical models in predicting the NAC response. The performance of a purely convolutional architecture, a completely self-attention-based model, and a combination of convolutional and self-attention components in extracting informative features from the tumor patches were compared. The results demonstrate that coupling the convolutional and self-attention modules leads to a more effective architecture for extracting local features. The positional embedding approach proposed in this study enables the framework to extract global relations between tumor areas and predict the pCR/non-pCR outcome for each patient using multi-head attention modules.

The obtained results in this study show that the local features extracted at the patch and tumor levels carry meaningful information for NAC response prediction. However, the relations within the aggregated patch-level information and the global dependencies between the tumor areas should also be considered for a more accurate response prediction at the tumor and patient levels. This is supported by the results of the comparative evaluations performed with maximum voting for predicting NAC response at the tumor and patient levels. Specifically, while the models with maximum voting receive the information from all patches or tumor regions in a WSI, they have demonstrated inferior performance compared to their hierarchical-model counterparts. Further, aggregating the patch- or tumor-level information using the maximum voting strategy has resulted in slightly lower response prediction accuracy compared to the individual patch/tumor-level results. These observations highlight the importance of a systematized strategy to fuse the patch/tumor-level information on the tumor/patient level for NAC response prediction. The size and number of the tumor beds vary among the biopsy samples, leading to large variations

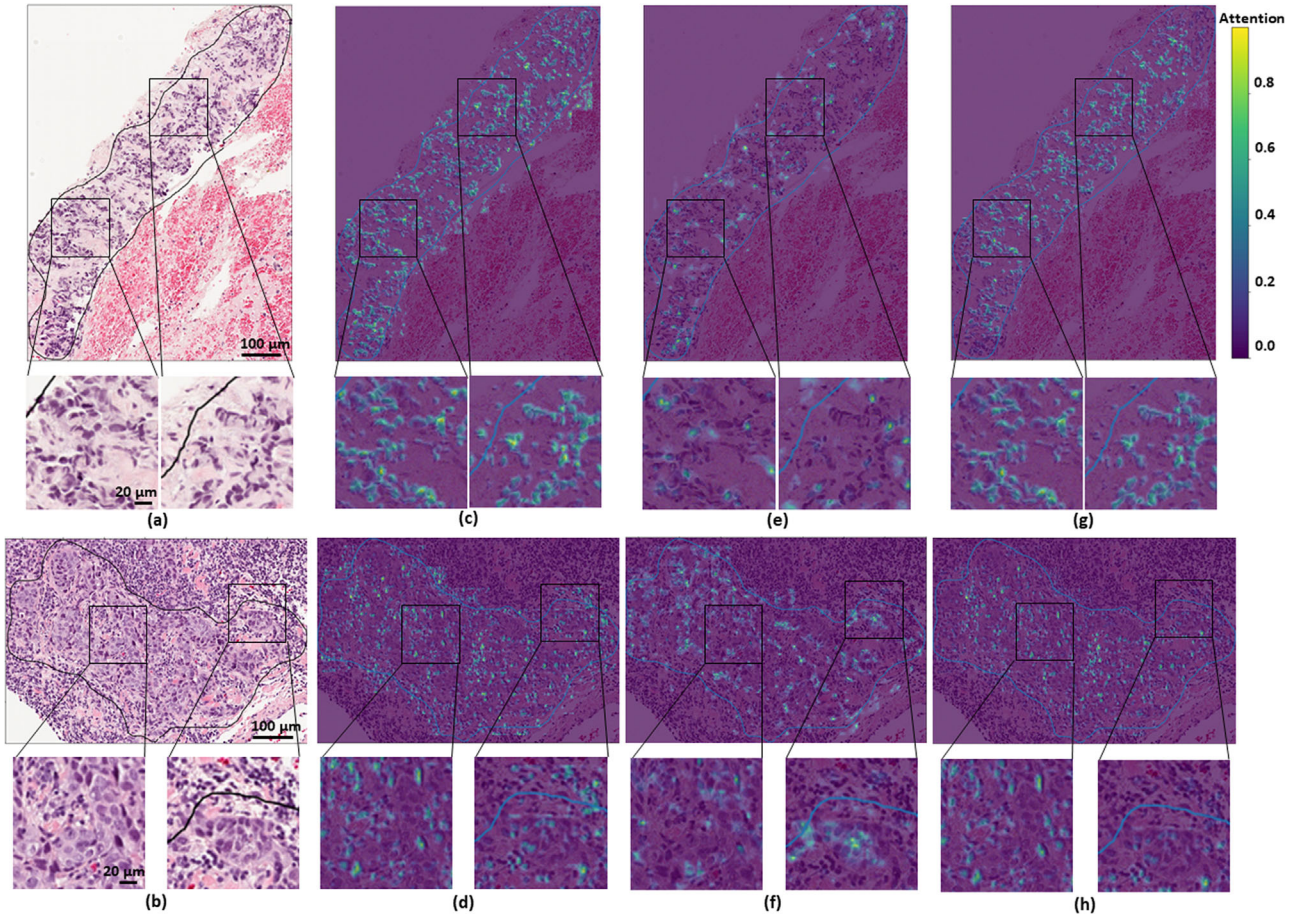



**FIGURE 5** Comparison of Grad-CAM attention heatmaps associated with different patch-level processing modules for two representative tumor regions with a pCR (top row) and non-pCR (bottom row) outcome after NAC, respectively: (a, b) tumor areas extracted from the H&E-stained WSIs, (c, d) attention heatmaps of ViT, (e, f) attention heatmaps of Xception + CBAM, (g, h) attention heatmaps of CoAtNet. The contours show the tumor bed annotations drawn by an expert pathologist. The CoAtNet architecture focuses more on the tumor area than the other networks.

in the number of patches per tumor bed and WSI. Such variations can potentially lead to inferior aggregated results on the tumor level and patient level based on maximum voting compared to the corresponding patch-level and patch-level + tumor-level models.

Information on cellular interactions and activities within the tumor microenvironment can potentially be captured from imaging. Enhancing diagnostic and response-guided therapy approaches may be possible by identifying biomarker signatures obtained through mapping tumor subcomponents and assessing biological heterogeneity in digital pathology images.[47] Previous studies have investigated the use of radiomic features from different medical imaging data collected at early stages of diagnosis for predicting the NAC response in BC.[19,48,49] The results obtained in those investigations demonstrate that intra-tumor heterogeneity quantified on imaging at pre-treatment can be associated with response to NAC. However, these images may not be acquired routinely for BC diagnosis. Integrating this information into a therapy response prediction framework may necessitate additional imaging data collection and processing that would not always be possible.

A recent study has demonstrated the potential of hand-crafted pathomic features coupled with conventional ML models in predicting BC response to NAC.[25] The ML models were developed using the training data acquired from 111 patients, where the best model achieved an accuracy of 84% (sensitivity: 85%; specificity: 82%) on the test set (38 patients). The observations of that research agree with the results obtained in this study. However, extracting hand-crafted features could be affected by feature extraction protocols that influence their reproducibility. Also, while the hand-crafted feature-based conventional ML models have a decent potential in analyzing imaging data, they are bounded by the information provided by a set of features defined by closed-form mathematical equations. The data-driven deep-learning models such as the one implemented in this study have shown higher

performance in analyzing large-scale data and better scalability with growing datasets. A very recent study has applied deep learning methods to predict BC response to NAC using three parallel pathology images as the input.[50] The digital pathology images were obtained from the histopathology slides with H&E staining, and Ki67 and phosphohistone H3 (PHH3) immunohistochemistry. The models were developed using a training set of 43 patients and evaluated on a test set with 30 patients. Using maximum voting on the patch-level results, their best model achieved a test accuracy of 93% on the patient level for detecting pCR to NAC. Those results, albeit obtained on a relatively small dataset, support the utility of deep learning models in conjunction with digital pathology images of tumor biopsies for NAC response prediction. However, the models developed in that study require multimodal pathology images with immunohistochemistry that are not routinely performed on pre-treatment tumor biopsies in the clinic.

The data in this study was acquired from a single institution retrospectively. Future multi-institutional studies with external validation are required for a more rigorous evaluation of the developed framework for NAC response prediction. The current framework requires manual annotation of the tumor regions by pathologists, which is tedious and time-consuming in the clinical setting. Automating the tumor annotation process can streamline the pathology workflow considerably. Future works can address this issue by developing automated tumor annotation methods and investigating their efficacy when integrated with the NAC response prediction framework. The framework proposed in this study directly analyses the pathology image patches with no explicit pre-processing step for nucleus detection and tumor cell classification. Future studies may investigate the potential impact of incorporating such pre-processing steps in the framework to extract tumor microenvironment features of the cancer cells. Nevertheless, the methods proposed in this study can potentially be adapted for analyzing digital pathology WSIs in other applications. This includes diagnostic and prognostic applications for various cancer types such as breast,[51] prostate,[52] liver,[53] and lung carcinomas.[54]

## 5 | CONCLUSIONS

This study presented an automated hierarchical framework for analyzing digital pathology images of biopsy specimens and predicting NAC response at pre-treatment with promising results. The effectiveness of combining convolutional blocks with self-attention modules for hierarchical analysis of high-resolution histopathological images was demonstrated. The results of this study pave the way toward a response-guided therapy paradigm for individual breast cancer patients and motivate future studies on larger multi-institutional datasets for further investigation of the proposed methodologies.

## ACKNOWLEDGMENTS

The authors would like to thank Andrew Lagree, Marie A. Alera, Lauren Fleshner, Audrey Shiner, Ethan Law, and Brianna Law for assistance with data acquisition. This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (Grant #: RGPIN-2016-06472), Tri-Council New Frontiers in Research Fund (NFRF; Grant # NFRFE-2019–00193), Lotte and John Hecht Memorial Foundation, and the Terry Fox Foundation (Grant #: 1083). A.S.N. holds the York Research Chair in Quantitative Imaging and Smart Biomarkers, and an Early Researcher Award from the Ontario Ministry of Colleges and Universities.

## CONFLICT OF INTEREST STATEMENT

K.S., W.T.T. and A.S.N. are inventors of a patent (Application No. 63/495390, pending) on deep learning of digital pathology images of pre-treatment tumor biopsies to predict breast cancer response to chemotherapy.

## REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33. https://doi.org/10.3322/caac.21708
2. Tryfonidis K, Senkus E, Cardoso MJ, Cardoso F. Management of locally advanced breast cancer—perspectives and future directions. *Nat Rev Clin Oncol*. 2015;12(3):147-162. https://doi.org/10.1038/nrclinonc.2015.13
3. Giordano SH. Update on locally advanced breast cancer. *Oncologist*. 2003;8(6):521-530. https://doi.org/10.1634/theoncologist.8-6-521
4. Dhanushkodi M, Sridevi V, Shanta V, et al. Locally advanced breast cancer (LABC): real-world outcome of patients from Cancer Institute, Chennai. *JCO Glob Oncol*. 2021;(7):767-781. https://doi.org/10.1200/GO.21.00001
5. Sadeghi-Naini A, Papanicolau N, Falou O, et al. Quantitative ultrasound evaluation of tumor cell death response in locally advanced breast cancer patients receiving chemotherapy. *Clin Cancer Res*. 2013;19(8):2163-2174. https://doi.org/10.1158/1078-0432.CCR-12-2965
6. Moghadas-Dastjerdi H, Sha-E-Tallat HR, Sannachi L, Sadeghi-Naini A, Czarnota GJ. A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative CT and machine learning. *Sci Rep*. 2020;10(1):10936. https://doi.org/10.1038/s41598-020-67823-8
7. Falou O, Sadeghi-Naini A, Prematilake S, et al. Evaluation of neoadjuvant chemotherapy response in women with locally advanced breast cancer using ultrasound elastography. *Transl Oncol*. 2013;6(1):17-24. https://doi.org/10.1593/tlo.12412
8. Sannachi L, Gangeh M, Tadayyon H, et al. Breast cancer treatment response monitoring using quantitative ultrasound and texture analysis: comparative analysis of analytical models. *Transl Oncol*. 2019;12(10):1271-1281. https://doi.org/10.1016/j.tranon.2019.06.004
9. Pfob A, Sidey-Gibbons C, Lee HB, et al. Identification of breast cancer patients with pathologic complete response in the breast after neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. *Eur J Cancer*. 2021;143:134-146. https://doi.org/10.1016/j.ejca.2020.11.006

10. Rueth NM, Lin HY, Bedrosian I, et al. Underuse of trimodality treatment affects survival for patients with inflammatory breast cancer: an analysis of treatment and survival trends from the national cancer database. *J Clin Oncol*. 2014;32(19):2018-2024. https://doi.org/10.1200/JCO.2014.55.1978
11. Scholl SM, Fourquet A, Asselain B, et al. Neoadjuvant versus adjuvant chemotherapy in premenopausal patients with tumours considered too large for breast conserving surgery: preliminary results of a randomised trial: S6. *Eur J Cancer*. 1994;30(5):645-652. https://doi.org/10.1016/0959-8049(94)90537-1
12. Meng X, Chang X, Wang X, Guo Y. Development and validation a survival prediction model and a risk stratification for elderly locally advanced breast cancer. *Clin Breast Cancer*. Published online 2022. https://doi.org/10.1016/j.clbc.2022.06.002
13. Sethi D, Sen R, Parshad S, Sen J, Khetarpal S, Garg M. Histopathologic changes following neoadjuvant chemotherapy in locally advanced breast cancer. *Ind J Cancer*. 2013;50(1):58. https://doi.org/10.4103/0019-509X.112301
14. Chuthapisith S, Eremin JM, El-Sheemy M, Eremin O. Neoadjuvant chemotherapy in women with large and locally advanced breast cancer: chemoresistance and prediction of response to drug therapy. *Surgeon*. 2006;4(4):211-219. https://doi.org/10.1016/S1479-666X(06)80062-4
15. Sethi D, Sen R, Parshad S, Sen J, Khetarpal S, Garg M. Histopathologic changes following neoadjuvant chemotherapy in locally advanced breast cancer. *Indian J Cancer*. 2013;50(1):58. https://doi.org/10.4103/0019-509X.112301
16. Haque W, Verma V, Hatch S, Suzanne Klimberg V, Brian Butler E, Teh BS. Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy. *Breast Cancer Res Treat*. 2018;170(3):559-567. https://doi.org/10.1007/s10549-018-4801-3
17. Tudorica A, Oh KY, Chui SYC, et al. Early prediction and evaluation of breast cancer response to neoadjuvant chemotherapy using quantitative DCE-MRI. *Transl Oncol*. 2016;9(1):8-17. https://doi.org/10.1016/j.tranon.2015.11.016
18. Tahmassebi A, Wengert GJ, Helbich TH, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Invest Radiol*. 2019;54(2):110-117. https://doi.org/10.1097/RLI.0000000000000518
19. Moghadas-Dastjerdi H, Sha-E-Tallat HR, Sannachi L, Sadeghi-Naini A, Czarnota GJ. A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative CT and machine learning. *Sci Rep*. 2020;10(10936). https://doi.org/10.1038/s41598-020-67823-8
20. Taleghamar H, Jalalifar SA, Czarnota GJ, Sadeghi-Naini A. Deep learning of quantitative ultrasound multi-parametric images at pre-treatment to predict breast cancer response to chemotherapy. *Sci Rep*. 2022;12(1):2244. https://doi.org/10.1038/s41598-022-06100-2
21. Tran WT, Gangeh MJ, Sannachi L, et al. Predicting breast cancer response to neoadjuvant chemotherapy using pretreatment diffuse optical spectroscopic texture analysis. *Br J Cancer*. 2017;116(10):1329-1339. https://doi.org/10.1038/bjc.2017.97
22. Taleghamar H, Moghadas-Dastjerdi H, Czarnota GJ, Sadeghi-Naini A. Characterizing intra-tumor regions on quantitative ultrasound parametric images to predict breast cancer response to chemotherapy at pre-treatment. *Sci Rep*. 2021;11(1):14865. https://doi.org/10.1038/s41598-021-94004-y
23. Hayashi M, Yamamoto Y, Iwase H. Clinical imaging for the prediction of neoadjuvant chemotherapy response in breast cancer. *Chin Clin Oncol*. 2020;9(3):31. doi:10.21037/cco-20-15
24. Romeo V, Accardo G, Perillo T, et al. Assessment and prediction of response to neoadjuvant chemotherapy in breast cancer: a comparison of imaging modalities and future perspectives. *Cancers (Basel)*. 2021;13(14). https://doi.org/10.3390/cancers13143521
25. Saednia K, Lagree A, Alera MA, et al. Quantitative digital histopathology and machine learning to predict pathological complete response to chemotherapy in breast cancer patients using pre-treatment tumor biopsies. *Sci Rep*. 2022;12(1):9690. https://doi.org/10.1038/s41598-022-13917-4
26. Rehman A, Ahmed Butt M, Zaman M. A Survey of Medical Image Analysis Using Deep Learning Approaches. In: *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE; 2021:1334-1342. https://doi.org/10.1109/ICCMC51019.2021.9418385
27. Kim KG. Book Review: Deep Learning. In: *Healthcare Informatics Research*. Vol 22. 2016:351-354. https://doi.org/10.4258/hir.2016.22.4.351
28. Zadeh Shirazi A, Fornaciari E, Bagherian NS, Ebert LM, Koszyca B, Gomez GA. DeepSurvNet: deep survival convolutional network for brain cancer survival rate classification based on histopathological images. *Med Biol Eng Comput*. 2020;58(5):1031-1045. https://doi.org/10.1007/s11517-020-02147-3
29. Saednia K, Tran WT, Sadeghi-Naini A. Automatic characterization of breast lesions using multi-scale attention-guided deep learning of digital histology images. *Comput Methods Biomech Biomed Eng Imaging Vis*. Published online 2022:1-9. https://doi.org/10.1080/21681163.2022.2058415
30. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559-1567. https://doi.org/10.1038/s41591-018-0177-5
31. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. 2021;124(4):686-696. https://doi.org/10.1038/s41416-020-01122-x
32. Dimitriou N, Arandjelović O, Caie PD. Deep learning for whole slide image analysis: an overview. *Front Med (Lausanne)*. 2019;6(264). https://doi.org/10.3389/fmed.2019.00264
33. Clunie D, Hosseinzadeh D, Wintell M, et al. Digital imaging and communications in medicine whole slide imaging connectathon at digital pathology association pathology visions 2017. *J Pathol Inform*. 2018;9(1):6. https://doi.org/10.4103/jpi.jpi_1_18
34. Chollet F. Xception: Deep Learning With Depthwise Separable Convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
35. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
36. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009:248-255.
37. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In Guyon I, Luxburg U von, Bengio S, et al. (Eds.), *Advances in neural information processing systems*. Vol 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
38. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In: *Third International Conference on Learning Representations (ICLR2015)*. 2015.
39. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. Published online 2020. http://arxiv.org/abs/2010.11929
40. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. *ACM Comput Surv*. 2022;54(10s):1-41. https://doi.org/10.1145/3505244

41. Paul S, Chen PY. Vision transformers are robust learners. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022;36(2):2071-2081. https://doi.org/10.1609/aaai.v36i2.20103
42. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*. Published online 2018:3-19.
43. Dai Z, Liu H, Le Q v, Tan M. CoAtNet: marrying convolution and attention for all data sizes. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, eds. *Advances in Neural Information Processing Systems*. Vol 34. Curran Associates, Inc.; 2021:3965-3977. https://proceedings.neurips.cc/paper/2021/file/20568692db622456cc42a2e853ca21f8-Paper.pdf
44. Symmans WF, Peintinger F, Hatzis C, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol*. 2007;25(28):4414-4422. https://doi.org/10.1200/JCO.2007.10.6823
45. Martel AL, Hosseinzadeh D, Senaras C, et al. An image analysis resource for cancer research: PIIP—pathology image informatics platform for visualization, analysis, and management. *Cancer Res*. 2017;77(21). https://doi.org/10.1158/0008-5472.CAN-17-0323
46. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE; 2017:618-626. https://doi.org/10.1109/ICCV.2017.74
47. Heindl A, Nawaz S, Yuan Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Lab Invest*. 2015;95(4):377-384. https://doi.org/10.1038/labinvest.2014.155
48. Ha S, Park S, Bang JI, Kim EK, Lee HY. Metabolic Radiomics for Pretreatment 18F-FDG PET/CT to Characterize Locally Advanced Breast Cancer: Histopathologic Characteristics, Response to Neoadjuvant Chemotherapy, and Prognosis. *Sci Rep*. 2017;7(1):1556. https://doi.org/10.1038/s41598-017-01524-7
49. Kolios C, Sannachi L, Dasgupta A, et al. MRI texture features from tumor core and margin in the prediction of response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *Oncotarget*. 2021;12(14):1354-1365. doi:10.18632/oncotarget.28002
50. Duanmu H, Bhattarai S, Li H, et al. A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images. *Bioinformatics*. 2022;38(19):4605-4612. https://doi.org/10.1093/bioinformatics/btac558
51. Lagree A, Shiner A, Alera MA, et al. Assessment of digital pathology imaging biomarkers associated with breast cancer histologic grade. *Curr Oncol*. 2021;28(6):4298-4316. https://doi.org/10.3390/curroncol28060366
52. Bulten W, Kartasalo K, Chen PHC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28(1):154-163. https://doi.org/10.1038/s41591-021-01620-2
53. Taylor-Weiner A, Pokkalla H, Han L, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. *Hepatology*. 2021;74(1):133-147. https://doi.org/10.1002/hep.31750
54. Shim WS, Yim K, Kim TJ, et al. DeepRePath: identifying the prognostic features of early-stage lung adenocarcinoma using multi-scale pathology images and deep convolutional neural networks. *Cancers (Basel)*. 2021;13(13). https://doi.org/10.3390/cancers13133308


**How to cite this article:** Saednia K, Tran WT, Sadeghi-Naini A. A hierarchical self-attention-guided deep learning framework to predict breast cancer response to chemotherapy using pre-treatment tumor biopsies. *Med Phys*. 2023;50:7852–7864. https://doi.org/10.1002/mp.16574