Title

End-to-End diagnosis of breast biopsy images with transformers.

Authors

Mehta, Sachin
Lu, Ximing
Wu, Wenjun
et al.

Publication Date

2022-07-01

DOI

10.1016/j.media.2022.102466

Peer reviewed

HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2023 May 05.

Published in final edited form as:

Med Image Anal. 2022 July ; 79: 102466. doi:10.1016/j.media.2022.102466.

End-to-End diagnosis of breast biopsy images with transformers

Sachin Mehta[a,1], **Ximing Lu**[a,1], **Wenjun Wu**[a], **Donald Weaver**[b], **Hannaneh Hajishirzi**[a], **Joann G. Elmore**[c,2], **Linda G. Shapiro**[a,2,*]

[a]University of Washington, Seattle, USA

[b]Department of Pathology, The University of Vermont College of Medicine, USA

[c]David Geffen School of Medicine, University of California, Los Angeles, USA

Abstract

Diagnostic disagreements among pathologists occur throughout the spectrum of benign to malignant lesions. A computer-aided diagnostic system capable of reducing uncertainties would have important clinical impact. To develop a computer-aided diagnosis method for classifying breast biopsy images into a range of diagnostic categories (benign, atypia, ductal carcinoma in situ, and invasive breast cancer), we introduce a transformer-based hollistic attention network called HATNet. Unlike state-of-the-art histopatho-logical image classification systems that use a two pronged approach, i.e., they first learn local representations using a multi-instance learning framework and then combine these local representations to produce image-level decisions, HATNet streamlines the histopathological image classification pipeline and shows how to learn representations from gigapixel size images end-to-end. HATNet extends the bag-of-words approach and uses self-attention to encode global information, allowing it to learn representations from clinically relevant tissue structures without any explicit supervision. It outperforms the previous best network Y-Net, which uses supervision in the form of tissue-level segmentation masks, by 8%. Importantly, our analysis reveals that HATNet learns representations from clinically relevant structures, and it matches the classification accuracy of 87 U.S. pathologists for this challenging test set.

Keywords

Transformers; Histopathological images; Breast cancer; Image classification; Convolutional neural networks; Whole slide images

*Corresponding author at: Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, 98195, USA, shapiro@cs.washington.edu (L.G. Shapiro).

[1]S. Mehta and X. Lu are co-first authors.

[2]J.G. Elmore and L.G. Shapiro are co-senior authors.

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.

CRediT authorship contribution statement

Sachin Mehta: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing. **Ximing Lu:** Software, Validation, Writing – review & editing. **Wenjun Wu:** Validation. **Donald Weaver:** Writing – review & editing. **Hannaneh Hajishirzi:** Supervision. **Joann G. Elmore:** Funding acquisition, Supervision, Writing – review & editing. **Linda G. Shapiro:** Funding acquisition, Supervision, Writing – review & editing.

1. Introduction

Breast cancer is the most common non-skin cancer in women accounting for approximately 25% of all cancer instances world-wide (Makki, 2015; DeSantis et al., 2019). The "gold standard" for diagnosis of breast biopsy specimens relies on a pathologist's visual assessment of tissue sections and cognitive processing of learned cytologic and morphological criteria, including architectural and cellular changes in the tissue, alterations of the tumor microenvironment, and immune-mediated host response. Assessment of these morphological criteria is subjective and can be challenging for some cases, especially those in the middle of the breast diagnostic spectrum. Pathologists, even expert pathologists, cannot always reach consensus on diagnostically challenging cases; diagnostic disagreement occurs throughout the spectrum of benign to malignant lesions (Wells et al., 1998; Della Mea et al., 1997; Allison et al., 2014; Elmore et al., 2015). Diagnostic variability is a serious problem, as misclassifying breast cancer as benign may lead to delay and fatal progression of disease, while diagnosing a benign lesion as malignant may result in significant morbidity including overtreatment, unnecessary emotional strain, anxiety, and increased cost of care. Misdiagnosis of breast cancer has been a leading cause for malpractice claims for decades (Kern, 2001; Reisch et al., 2015). A computer-aided diagnostic system that would support pathologists by reducing classification uncertainties could have positive clinical impact.

This paper introduces a self-attention-based network called **H**olistic **AT**tention **Net** work (HATNet) for classifying breast biopsy images in an end-to-end manner. HATNet extends the self-attention network of Vaswani et al. (2017). The core principle is to factorize the input biopsy image into words (or patches) using a bag-of-words approach and then encode inter-word and inter-bag relationships in a hierarchical manner using self-attention. Self-attention enables interaction between inputs (bags or words), allowing the encoding of global information in an end-to-end fashion. This helps the network learn representations from clinically relevant tissue structures without any supervision, as shown in Fig. 1.

HATNet outperforms previous methods; it is 8% more accurate and about 2 × faster than the previous best network, Y-Net (Mehta et al., 2018b), and also matches the classification performance of participant pathologists on the test set. Our analysis further suggests that HATNet pays attention to important biomarkers (stromal tissue and ducts) in the diagnosis and classification of breast tissue, suggesting that there is clinical relevance to the method. To the best of our knowledge, this is the first work that (1) uses transformers to classify histopathological images in an end-to-end fashion and (2) correlates model decisions with clinically relevant structures. Our source code is available at https://github.com/sacmehta/HATNet.

2. Related work

Histopathological image classification

Convolutional neural networks (CNNs) are state-of-the-art networks for image classification (e.g., ResNet of He et al., 2016), including histopathological image analysis (Cireşan et al., 2013; Cruz-Roa et al., 2014; Xu et al., 2015; Hou et al., 2016; Gecer et al., 2018; Mehta et al., 2018b). Histopathological image classification methods often follow a bag-of-words

model for learning representations, wherein a whole slide image is treated as a bag, while image patches are treated as words (or instances).

Given the bag-of-words model, a first line of research focuses on extracting word-level representations using CNNs, which are then aggregated to produce image-level decisions. Feature selection-based aggregation methods allows identification of relevant features in these word representations (Cruz-Roa et al., 2014; Xu et al., 2015; Sun et al., 2019). However, such methods fail to capture the heterogeneity of diagnosis categories. To address the limitations of these methods, multi-instance learning (MIL) based methods have been proposed (Hou et al., 2016; Mercan et al., 2017; Gecer et al., 2018; Ilse et al., 2018; Wang et al., 2019b; Campanella et al., 2019; Lu et al., 2021a). In a MIL framework (Maron and Lozano-Pérez, 1998), a WSI is divided into words (or instances) and the same slide-level diagnostic label is assigned to all words within a particular slide. Because a slide-level label casts a weak label on all words in a given slide, these approaches are also categorized as weakly supervised. In general, these approaches are two pronged. They first generate word-level representations using a CNN and then combine these representations using different methods to produce a WSI-level decision. For instance, Hou et al. (2016) studies different approaches (e.g., thresholding, averaging, and majority voting) to combine word-level representations and produce a WSI-level diagnostic decision. Campanella et al. (2019) uses recurrent neural networks to combine word-level representations. Lu et al. (2021a) clusters word-level representations into positive and negative categories, and then weighs positive word-level representations by their relative scores to produce a WSI-level decision. Because some of these approaches also identify salient regions before fusion (e.g., Hou et al., 2016; Lu et al., 2021a), they are also known as saliency-based methods.

A second line of research considers tissue type, size, and distribution to produce image-level decisions (Lu and Mandal, 2015; Mehta et al., 2018a; Mercan et al., 2019). These approaches extend MIL-based approaches to tissue-level. These approaches produce word-level (or instance-level) segmentation masks, which are then combined to produce image-level segmentation masks. Tissue-level structural information (e.g., size and distribution) extracted from these masks is then used to produce diagnosis categories.

A third line of research integrates both saliency- and segmentation-based approaches (Mehta et al., 2018b; Thome et al., 2019; Heker and Greenspan, 2020; Hou et al., 2020; Wang et al., 2021). These approaches simultaneously produce saliency maps and segmentation masks, which are then combined to extract structural information about tissues and to produce image-level decisions.

Though these methods are effective in classifying histopatho-logical images, the context-capturing ability of saliency-based methods is limited to words and are not able to encode spatial relationships between words. Also, some of these methods require manual threshold selection to identify salient regions. The latter segmentation-based methods address these limitations; however, acquiring tissue-level segmentation labels at a large scale is difficult, because experts are required for annotating images. This work introduces a transformer-based method, HATNet, to address the limitations of existing methods. Similar to previous work, HATNet is based on the bag-of-words model. However, unlike existing methods, it

hierarchically aggregates information at different levels of the model using self-attention, which allows learning of spatial relationships between words and bags. HATNet outperforms existing methods (saliency-based or segmentation-based or their com- bination) by a significant margin. Moreover, this network learns representations from clinically relevant and variably-sized structures.

Spatial attention in vision models

The most widely studied attention mechanism in visual recognition tasks (image classification, segmentation and object detection) is the spatial attention mechanism (Zhou et al., 2016; Selvaraju et al., 2017), which weighs the activation maps (or spatial planes) to identify regions of interest. Initially introduced to provide explanations for CNN outputs, variants of this mechanism, including supervised (Yang et al., 2019; Yao and Gong, 2020) and unsupervised (Hu et al., 2018; Xu et al., 2018; Huang et al., 2019; Wang et al., 2020) methods, have been incorporated in CNNs to improve the performance across different visual recognition tasks (Howard et al., 2019; Woo et al., 2018), including medical imaging (Oktay et al., 2018; Abraham and Khan, 2019; Schlemper et al., 2019; Rundo et al., 2019; Tomita et al., 2019). In general, these networks introduce a spatial or channel-wise attention module within a CNN. For example, Attention U-Net (Oktay et al., 2018) incorporated an additive gating unit (similar to the squeeze-and-excitation unit of Hu et al. (2018)) between the encoder and the decoder blocks in the U-Net network to learn better representations. Identifying salient regions in histopathological images using spatial attention is difficult because of their large size (usually of the order of gigapixels). This paper introduces an end-to-end transformer-based network for classifying histopathological images.

Vision transformers

Recent work (e.g., Dosovitskiy et al., 2021; Touvron et al., 2021) has extended the transformers of Vaswani et al. (2017) (described in Section 3) for vision tasks. Though these approaches are effective in learning global representations, they exhibit sub-standard optimizability (i.e., they require a large amount of training data and heavy regularization). This is likely because vision transformers lack image-specific inductive biases (Xiao et al., 2021; Dai et al., 2021). Moreover, extending these approaches to histopathological images is challenging primarily because of their large size (e.g., images in our dataset are 2,000 × larger than the ImageNet dataset of Russakovsky et al. (2015)). This work extends transformers using bag-of-words model to classify breast biopsy images in an end-to-end fashion. Specifically, we introduce a bottom-up decoding method that allows us to hierarchically encode the information from words to bags to images and produce diagnostic categories. Because the spatial order of bags and words in each bag is preserved in HATNet's top-down and bottom-up approach, HATNet implicitly incorporates inductive biases, similar to CNNs. We believe that this property, along with HATNet's ability to encode global information, allows HATNet to learn representations from clinically relevant structures without any explicit supervision, and deliver better performance than CNN-based methods (Section 5.5). We note that our observation is consistent with recent parallel work on the ImageNet dataset (Russakovsky et al., 2015), which also shows that vision transformers benefit from spatial inductive biases (e.g., Xiao et al., 2021; Dai et al., 2021; Graham et al., 2021).

3. Background: Transformers

Transformers (Vaswani et al., 2017) allow inputs to interact with each other, so that the model can automatically find important inputs on which to focus. The transformer module consists of two parts: (1) multi-head attention (MHA) that models relationships between inputs, and (2) a feed forward network (FFN) that learns wider representations. For an input $\mathbf{X} \in \mathbb{R}^{N \times d}$ with N d-dimensional instances (words and bags in our case), transformers learn the representations as:

$$\mathbf{Y} = \text{Transformer}(\mathbf{X}) = \text{FFN}(\text{MHA}(\mathbf{X}_Q = \mathbf{X}, \mathbf{X}_K = \mathbf{X}, \mathbf{X}_V = \mathbf{X})) \quad (1)$$

where $\mathbf{X}_Q$, $\mathbf{X}_K$, and $\mathbf{X}_V$ are the inputs to the query, key, and value branches in the multi-head attention, respectively. For simplicity, residual connections are not shown in Eq. (1).

Because of large spatial dimensions of histopathological images (e.g., $11\text{k} \times 10\text{k}$ in our dataset), learning visual representations of WSIs with transformers is challenging. On an average, the vision transformer of Dosovitskiy et al. (2021) will have about $N = 430\text{k}$ words for a WSI in our dataset. Because of the quadratic computational cost of MHA (i.e., $\mathcal{O}\left(N^2 d\right)$), applying transfomers to WSIs is computationally intractable. This work extends the vision transformers using the bag-of-words model for learning global representations from very large images in an end-to-end fashion.

4. HATNet: Holistic attention network

State-of-the-art CNN-based classification networks stack convolutional layers and down-sampling layers to learn representations at multiple scales (Simonyan and Zisserman, 2014; He et al., 2016). These networks are difficult to apply to histopathological images, primarily because the resolution of these medical images (e.g., 11k × 10k) are much larger than images used in standard image classification tasks (e.g., 224 × 224 in the ImageNet dataset Russakovsky et al., 2015). To address this resolution challenge, a standard approach is to learn word-wise (or patch-wise) representations using a sliding window method (Hou et al., 2016; Mehta et al., 2018b; Gecer et al., 2018; Iizuka et al., 2020). Though these approaches are effective for histopathological image analysis, the context-capturing ability of such approaches is still limited to word-level, and it is difficult to train such systems in an end-to-end manner.

This paper unifies the separate components of histopathological image analysis (i.e., first learn the word-wise representations *independently* and then fuse these local representations to produce image-level decisions) into a single neural network. Our network, a Holistic ATtention Network (HATNet), uses representations from the entire image at once to produce the diagnostic decision. This means that HATNet reasons globally about the entire input image and all variably-sized structures in the image. The HATNet design enables end-to-end training and inference while delivering pathologist-level performance.

HATNet extends the transformer architecture using a bag-of-words approach and is shown in Fig. 2 (a). We call our model a Holistic ATtention Network (HATNet) because of its ability

to learn inter-word and inter-bag representations in an end-to-end fashion. With *attention*, we emphasize the progressive hierarchical refining from words to bags to image to produce the classification output. Briefly, HATNet first encodes inter-word representations using self-attention (Section 4.1). These representations are then combined to produce bag-level representations (Section 4.2). The inter-bag rep- resentations (Section 4.3) are encoded and then combined to pro- duce image-level representations (Section 4.4). These representations are classified to produce the diagnosis category (Section 4.5). Because of the bottom-up decoding (words → bags → image), representations learned using HATNet are expressive and allow the identification of important words and bags corresponding to clinically relevant structures in an image. We believe that this will help us build tools to annotate clinically important words and explain diagnosis decisions.

4.1. Word-to-word attention

The word-to-word attention module, shown in Fig. 2(b), is comprised of a transformer unit (Section 3) with multi-head attention and a feed-forward network, allowing us to model the interactions between words and identify important words in the whole slide image.

The input image $\mathbf{I} \in \mathbb{R}^{w \times h}$ with width w and height h is first divided into n non-overlapping bags $\mathbf{I} = \left(\mathbf{B}^1, \cdots, \mathbf{B}^n\right) \in \mathbb{R}^{\frac{w}{\sqrt{n}} \times \frac{h}{\sqrt{n}}}$, where $\mathbf{B}^i$ represents the ith bag. Each bag $\mathbf{B}^i$ is then divided into m non-overlapping words $\mathbf{B}^i = (\mathbf{W}^i_1, \cdots, \mathbf{W}^i_m) \in \mathbb{R}^{\frac{w}{\sqrt{nm}} \times \frac{h}{\sqrt{nm}}}$, where $\mathbf{W}^i_j$ represents the j th word in the ith bag. Following previous works (e.g., Hou et al., 2016; Mehta et al., 2018b; Lu et al., 2021a), the words $\mathbf{W}^i_j$ inside each bag $\mathbf{B}^i$ are fed to a CNN to produce word-level representations for each bag: $\mathbf{B}^i_{cnn} = (\widehat{\mathbf{W}}^i_1, \cdots, \widehat{\mathbf{W}}^i_m) \in \mathbb{R}^d$. The representations from the CNN does not encode inter-word relationships. Inter-word relationships in each bag $\mathbf{B}^i_{cnn}$ are encoded using the transformer unit (Section 3) to produce contextualized word encoded using the transformer unit (Section 3) to produce contextualized word embeddings (CWEs) $\mathbf{B}^i_{w2w} \in \mathbb{R}^{m \times d}$ as:

$$\mathbf{B}^{\mathbf{i}}_{w2w} = \mathrm{FFN}\left(\mathrm{MHA}(\mathbf{X}_Q = \mathbf{B}^i_{cnn}, \mathbf{X}_K = \mathbf{B}^i_{cnn}, \mathbf{X}_V = \mathbf{B}^i_{cnn})\right) \tag{2}$$

The multi-head attention (MHA) enables the encoding of inter-word relationships, and the feed forward network (FFN) allows the system to learn wider representations.

4.2. Word-to-bag attention

The word-to-word attention produces CWEs for each bag. These word-level representations are aggregated to produce bag-level representations (see Fig. 2(c)) by linearly combining the words inside each bag $\mathbf{B}^i_{w2w}$. Specifically, each word in $\mathbf{B}^i_{w2w}$ is mapped from $\mathbb{R}^d$ to $\mathbb{R}^1$ using a projection function Ψ. Since each bag has m words, this projection function Ψ produces a vector of length m. A linear transformation $\bar{\boldsymbol{\beta}}_{w2b} \in \mathbb{R}^{m \times m}$ and softmax functions are then applied to produce m coefficients, which are then used to linearly combine words in $\mathbf{B}^i_{w2w}$ to produce bag-level representations $\bar{\mathbf{B}}^i_{w2b} \in \mathbb{R}^d$ as:

$$\overline{\mathbf{B}}^{i}_{w2b} = \text{softmax}\Big(\mathbf{\Psi}(\mathbf{B}^{i}_{w2w})\overline{\boldsymbol{\beta}}_{w2b}\Big)\mathbf{B}^{i}_{w2w},\ 1 \leq i \leq n \tag{3}$$

Similarly, the word-level representations obtained from the CNN for each bag $\mathbf{B}^{i}_{cnn}$ are also combined using Ψ, linear transformation $\widehat{\boldsymbol{\beta}}_{w2b} \in \mathbb{R}^{m \times m}$, and the softmax function to produce bag-level representations, $\widehat{\mathbf{B}}^{i}_{w2b} \in \mathbb{R}^{d}$.

$$\widehat{\mathbf{B}}^{i}_{w2b} = \text{softmax}\Big(\mathbf{\Psi}(\mathbf{B}^{i}_{cnn})\ \widehat{\boldsymbol{\beta}}_{w2b}\Big)\mathbf{B}^{i}_{cnn},\ 1 \leq i \leq n \tag{4}$$

4.3. Bag-to-bag attention

The representation in $\overline{\mathbf{B}}_{w2b}$ encodes global information about all words in a bag using multi-headed self-attention, while the representation in $\widehat{\mathbf{B}}_{w2b}$ encodes local information (obtained using the CNN) about all words in a bag to produce bag-level representations. However, these bag-level representations do not encode information about surrounding bags. To encode inter-bag relationships, bag-to-bag attention (see Fig. 2(d)) is applied. The bag-to-bag attention module is similar to the word-to-word attention module (Section 4.1), except that $\widehat{\mathbf{B}}_{w2b}$ (Eq. (4)) is used as context to $\overline{\mathbf{B}}_{w2b}$ (Eq. (3)). With this attention, we are able to encode local and global information in the input effectively. We note that this attention also mimics the typical skip-connection mechanism in neural networks (He et al., 2016; Ronneberger et al., 2015) and helps improve the performance.

Multi-head attention is first applied to $\widehat{\mathbf{B}}_{w2b}$ to encode inter-bag representations and produce $\widehat{\mathbf{B}}_{b2b} \in \mathbb{R}^{n \times d}$ as:

$$\widehat{\mathbf{B}}_{b2b} = \text{MHA}(\mathbf{X}_Q = \widehat{\mathbf{B}}_{w2b}, \mathbf{X}_K = \widehat{\mathbf{B}}_{w2b}, \mathbf{X}_V = \widehat{\mathbf{B}}_{w2b}) \tag{5}$$

To allow every bag $\widehat{\mathbf{B}}_{b2b}$ obtained from a CNN to attend over every bag $\overline{\mathbf{B}}_{w2b}$ obtained after word-level self-attention, another multi-head attention in which $\widehat{\mathbf{B}}_{b2b}$ serves as a query and $\overline{\mathbf{B}}_{w2b}$ serves as keys and values is applied to produce contextualized bag embeddings (CBEs) $\mathbf{B}_{b2b} \in \mathbb{R}^{n \times d}$. Mathematically, the bag-to-bag attention operation is defined as:

$$\mathbf{B}_{b2b} = \text{FFN}\Big(\text{MHA}(\mathbf{X}_Q = \widehat{\mathbf{B}}_{b2b}, \mathbf{X}_K = \overline{\mathbf{B}}_{w2b}, \mathbf{X}_V = \overline{\mathbf{B}}_{w2b})\Big) \tag{6}$$

4.4. Bag-to-image attention

The inter-bag representations encoded in $\mathbf{B}_{b2b} \in \mathbb{R}^{n \times d}$ are aggregated to produce image-level representations. Similar to word-to-bag attention (Section 4.2), these bag-level representations are combined using a function Ψ and linear transformation $\boldsymbol{\beta}_{b2i} \in \mathbb{R}^{n \times n}$ to produce image-level representations $\mathbf{I}_{b2i} \in \mathbb{R}^{d}$.

$$\mathbf{I}_{b2i} = \text{softmax}(\mathbf{\Psi}(\mathbf{B}_{b2b})\boldsymbol{\beta}_{b2i})\mathbf{B}_{b2b} \tag{7}$$

Because of the bottom-up decoding (words to bags to image), these representations are expressive and allows the system to identify important words and bags in an image (Fig. 3).

4.5. Classification and loss

HATNet classifies $\mathbf{I}_{b2i} \in \mathbb{R}^d$ into C-diagnosis classes using a linear classifier with weights $\beta_{cls} \in \mathbb{R}^{d \times C}$ as:

$$\hat{y} = \text{softmax}(\mathbf{I}_{b2i}\, \boldsymbol{\beta}_{cls}) \tag{8}$$

To train HATNet, the cross-entropy loss $\mathscr{L}$ between the ground truth y and prediction $\hat{y}$ is minimized. During evaluation, the index that has the highest confidence score in $\hat{y}$ is chosen as the predicted class label.

5. Experimental results

5.1. Dataset and evaluation

Breast biopsy dataset and ground truth consensus reference—The breast biopsy dataset consists of 240 whole slide images with haematoxylin and eosin (H&E) staining (Elmore et al., 2015). The image dataset was designed to include a higher prevalence of cases from diagnostic categories that have lower prevalence in the general population, providing a robust and challenging image dataset. A group of three expert pathologists independently interpreted these cases and then met to discuss the cases using a modified Delphi method to provide a reference consensus label per slide (Custer et al., 1999). The pathologists' assessments were grouped into 4 diagnostic categories: (1) benign without atypia (including non-proliferative and proliferative without atypia), (2) atypia, (3) ductal carcinoma in situ (DCIS), and (4) invasive car-cinoma. The consensus labels are our ground truth diagnoses. The expert pathologists also marked 422 regions of interest (ROIs) that best supported the diagnoses. Following previous studies on this dataset that aims to build directed computer-aided second opinion systems (Mercan et al., 2017; Mehta et al., 2018a; 2018b; Mercan et al., 2019; Gecer et al., 2018), we use these ROIs to train and evaluate our method, randomly splitting the dataset into 164 training, 42 validation, and 216 test ROIs (see Table 1). Note that clinically, each slide can have multiple ROIs. Therefore, we ensured that all ROIs corresponding to a slide are in the same set (training + validation or test).

Outcome metrics—The following metrics were used to evaluate the performance of HATNet (Tharwat, 2018)

- Classification (or Top-1) accuracy counts the number of times the predicted label is the same as the ground truth label and is defined as:

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where TP, FP, TN, and FN denotes the true positive, false positive, true negative, and false negatives respectively.

- F1-score is a harmonic mean of precision P and recall R and is defined as:

$$\text{F1-score} = \frac{2PR}{P + R}$$

where $P = \frac{\text{TP}}{\text{TP} + \text{FP}}$ and $R = \frac{\text{TP}}{\text{TP} + \text{FN}}$.

- Sensitivity measures the proportion of ROIs from positive cases that are correctly classified and is defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Specificity measures the proportion of ROIs from negative cases that are correctly classified and is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- Area under receiver operating characteristics curve (ROC-AUC) is a graph that is obtained by varying the threshold for diagnostic decision, illustrating the discrimination ability of the classifier.

The values of these metrics range between zero and one, and higher values of these metrics mean better performance.

Accuracy data from U.S. pathologists—To compare the results from HATNet with the interpretations of practicing U.S. pathologists, we used data from a prior clinical study of 87 pathologists who interpreted these same cases (Allison et al., 2014; Elmore et al., 2015; Elmore et al., 2017). Each pathologist interpreted a random subset of 60 cases and their diagnoses were classified into the same four diagnostic categories. This resulted in 22 independent diagnostic labels (on average) per slide and gave us a way to compare human pathologist results to HATNet.

Structure-level annotations for saliency-annotation agreement—The bottom-up decoding (word to bag to image embedding) approach is expressive and allows HATNet to identify important words and bags in a ROI. We rank the CWEs (Section 4.1) and CBEs (Section 4.3) based on their self-attention score obtained using the transformer unit to identify the top-k words and bags respectively, where k is a variable used in our experiments. To determine if these top-k words and bags are clinically relevant, we study the agreement between these salient regions (bags and words) and clinical biomarkers (stromal tissue and ductal regions) for which we have annotations from expert pathologists at the ROI-level as ground truth (Mehta et al., 2018a; Li et al., 2020). The Dice score is used as a quantitative metric to assess the agreement rate between the ground truth and the salient regions. Mathematically, the Dice score is equal to twice the area of overlap between the ground truth and salient region divided by the total number of pixels in the ground truth and the salient region. The value of k is varied from 10% to 60%.

5.2. Architecture

The ROIs are split into non-overlapping bags with a spatial dimension of 1792 × 1792. Each bag is then split into non-overlapping words with a spatial dimension of 256 × 256, resulting in $m = 49$ non-overlapping words. These words are fed to off-the-shelf CNNs to extract word-level representations. In our experiments, three state-of-the-art light-weight CNNs pretrained on the ImageNet dataset (Russakovsky et al., 2015) were studied: (1) ESPNetv2 (Mehta et al., 2019), (2) MobileNetv2 (Sandler et al., 2018), and (3) MNASNet (Tan et al., 2019). ESPNetv2 follows an Inception-style design (Szegedy et al., 2015) and uses four simultaneous 3 × 3 depth-wise convolutions with different dilation rates, allowing to learn multi-scale representations. MobileNetv2 follows a ResNet-style design (He et al., 2016). To improve the computational efficiency, MobileNetv2 uses 3 × 3 depth-wise convolutions instead of 3 × 3 standard convolutions. MNASNet uses the same basic building block as MobileNetv2; however, it uses neural architecture search (Zoph et al., 2018) to identify the optimal model configuration, which provides best trade-off between different parameters. The proposed network is generic and any off-the-shelf heavy-weight CNNs can be used to extract word-level representations. Heavy-weight networks, such as VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016), were not explored because of resource constraints.

The dimensionality of word-level representations varies from CNN to CNN. Therefore, the output of a CNN is linearly projected to a 256-dimensional space ($d = 256$). To encode the inter-word and inter-bag representations, 4 heads were used in multi-head attention. The function Ψ was used to aggregate word-level representations into bag-level representations (Section 4.2) and bag-level representations into image-level representations (Section 4.4). In our experiments, three different functions were studied: (1) Euclidean distance (or L2 norm), (2) Manhattan distance (or L1 norm), and (3) mean of a vector.

5.3. Training

HATNet is trained end-to-end using the ADAM optimizer of Kingma and Ba (2014) with a learning rate warm-up strategy. The learning rate is first warmed up from 10^{-7} to 10^{-4} in 600 iterations, and then the model is trained for the next 50 epochs with a learning rate of 10^{-4}. After that, the learning rate is decayed by half, and the model is trained for another 50 epochs. Our model takes about 36 h for training on two NVIDIA GeForce GTX 1080 GPUs, each with a memory of 8 GB. Gradients are accumulated for 8 iterations before the weights are updated, yielding an effective batch size of 8 ROIs per update. Training data is augmented by randomly resizing (192 × 192, 224 × 224, 256 × 256, 288 × 288, 320 × 320), flipping, and rotating (angle: −10° to 10°) the words. For evaluation, a single model is obtained by averaging the best 5 validation checkpoints. Compared to the best model on the validation set, averaged models delivered 1 to 1.5 points higher accuracy.

5.4. Baseline networks

We compare our method with the following methods:

1. **Bag-of-words model with hand-crafted features** (Gecer et al., 201): This method follows a multi-instance learning (MIL) framework and splits an input

image (bag) into words. Following Basavanhally et al. (2013), LAB and LBP histogram features are extracted from these words. These word-level features are concatenated and then classified using logistic regression into diagnostic categories with and without saliency. Similar to a standard practice in MIL-based saliency approaches (Hou et al., 2016; Wang et al., 2018), the class with majority voting in saliency maps is selected as a diagnostic category. The results of these approaches are summarized in rows R2 and R3 of Table 2.

2. **Bag-of-words model with deep features** (Gecer et al., 2018): This method extends the MIL framework with CNNs. Specifically, word-level representations are obtained using a deep convolutional neural network, FCN (with VGG as a backbone) of Long et al. (2015). These representations are used to identify discriminative or salient regions. In addition to majority voting-based method, a learned fusion method of Hou et al. (2016) is also tried to model the relationships between words. The results of these approaches are summarized in rows R4-R7 of Table 2.

3. **Multi-resolution segmentation network (MRSegNet)** (Mehta et al., 2018a): MRSegNet has two stages: (1) tissue-level segmentation and (2) diagnostic classification. The first stage is a multi-resolution encoder-decoder network which combines the outputs of many words (or patches) at different resolutions to reduce segmentation errors. In the second stage, histogram and co-occurrence features are extracted from tissue-level segmentation masks, which are then classified using a multi-layer perceptron into different diagnostic classes. The results of this method are given in row R8 of Table 2.

4. **Structural features** (Mercan et al., 2019): This method extracts structural features from tissue-level segmentation masks produced using MRSegNet. These features allows capturing structural changes in ductal regions, an important biomarker for cancer diagnosis in the breast (Kinne et al., 1989; Page and Jensen, 1996; Zhang et al., 2012; Shah et al., 2016). The results are summarized in row R9 of Table 2.

5. **Y-Net** (Mehta et al., 2018b): CNNs with multi-scale field of view yield better performance across different vision tasks (e.g., He et al., 2015; Chen et al., 2017; Zhao et al., 2017; Mehta et al., 2019; Wang et al., 2019a). These methods re-sample either the feature maps at different spatial resolutions (e.g., SPPNet He et al., 2015 and PSPNet Zhao et al., 2017) or the weights of a convolutional kernel using dilated convolutions (e.g., DeepLabv3 Chen et al., 2017 and ESPNetv2 Mehta et al., 2019) to learn multi-scale representations. Y-Net uses these multi-scale view approaches to learn better representations. It also generalizes the U-Net architecture of Ronneberger et al. (2015) and adds a classification branch, which allows it to jointly predict the tissue-level segmentation mask and the saliecy map. The saliency map is then combined with the tissue-level segmentation mask to produce a discriminative segmentation mask. Similar to MRSegNet, Y-Net extracts histogram and co-occurrence features, which are then

used for classifying diagnostic classes. The results are summarized in row R10 of Table 2.

5.5. Main results

Comparison with existing methods—Table 2 shows that HATNet outperforms state-of-the-art methods significantly. For example, HATNet (R13) improves the performance of the best saliency- (or MIL-) based models (R5, R7) by about 15%. When compared to approaches that use tissue-level segmentation masks (R8-R10) to capture the structural changes in biopsies, HATNet delivers better performance. In particular, HATNet improves the F1-score of the previous best segmentation-based approach (R10) by 8%. Overall, these results shows that HATNet is effective. We note that ensembling the three HATNet models (R14) further improves the accuracy and senstivity by 1%.

Furthermore, Table 3 shows that HATNet is fast. HATNet with MNASNet is about 1.8 × faster and 8% more accurate than the previously best reported network, i.e., Y-Net. The two-tailed p-value between HATNet and Y-Net is less than 0.0001, which indicates that the accuracy improvement of 8% is statistically significant. Also, HATNet is a stable network because run-to-run variation with three different random seeds (0, 100, and 1000) is low (about 0.2%).

Comparison with pathologists—HATNet achieves similar performance to practicing U.S. pathologists who interpreted these same cases in all quantitative metrics (HATNet (R13) vs. practicing pathologists (R1): 0.70 vs. 0.70 (accuracy), 0.70 vs. 0.71 (F1-score), 0.70 vs. 0.70 (sensitivity) and 0.90 vs. 0.90 (specificity)). We further analyze the misclassifications of HATNet and pathologists. For each case i, we obtain a pathologist score p_i, where p_i is the percentage of pathologists who misdiagnosed the case. The average over all the p_i's for all these cases was 0.61. So, we can say that on an average if HATNet misdiagnoses a case, 61% of the pathologists who diagnosed the same case also got it wrong. Therefore, we believe that HATNet can act as a directed second opinion system. For more details, see our work (Lu et al., 2021b) that uses HATNet to understand the visual similarities between diagnosed and misdiagnosed cases.

Saliency-annotation agreement analysis—Several clinical studies have shown that ductal regions and stromal tissue are important bio-markers for diagnosing breast cancer (Kinne et al., 1989; Page and Jensen, 1996; Arendt et al., 2010; Zhang et al., 2012; Conklin and Keely, 2012; Mao et al., 2013; Shah et al., 2016; Plava et al., 2019; DeSantis et al., 2019). Briefly, ducts are thin tubes in the breast and are responsible for carrying milk from lobules (milk glands) to the nipples. These regions are useful in identifying cancers that began in milk ducts, for example, DCIS (Kinne et al., 1989; Page and Jensen, 1996; Shah et al., 2016). On the other hand, the stroma is part of the breast tissue with a structural and developmental role and may be involved in tumor promotion and progression. Many clinical studies have underlined the importance of stroma in tumor progression along with its contribution to risk factors that determines tumor formation (Arendt et al., 2010; Conklin and Keely, 201).

Table 2 shows that HATNet learns better representations, resulting in significant performance gains compared to existing methods. A closer analysis in Figs. 4 and 5 reveals that our model pays attention to these important bio-markers, which helps it to achieve these gains.

- **Ductal regions:** To evaluate if our model pays attention to ductal regions or not, we compute the overlap between ductal regions (marked by experts) and top-k bag predictions of our model using dice score.[3] We use bags for saliency-annotation agreement with ductal regions because these variably-sized regions are large in size. Fig. 4 g shows the results. When considering top-50% bag predictions, HATNet achieves a dice score of 0.68. Furthermore, Fig. 4 a–f shows that HATNet is able to differentiate between ducts of variable size and texture. This shows that HATNet identifies ductal regions as an important structure.
- **Stromal tissue:** We compute the overlap between pixel-level annotations of stromal tissue and top-k words predicted by our model to determine whether our model pays attention to stromal tissue. We use the dice score to measure the overlap and vary k from 10% to 60%. Figure 5 g shows that HATNet achieves a dice score of about 0.75 when the top-50% word predictions are considered. This indicates that HATNet also identifies stroma as an important tissue. This is further strengthened by visualizations in Fig. 5 b–f, which shows that the majority of the top-50% words lie in stromal tissue.

The ability of HATNet to learn representations at different granularities (bags and words) allowed us to correlate model decisions with structurally different clinical entities, demonstrating that HATNet is effective in modeling inter-bag and inter-word relationships. Though HATNet's model decisions correlate with clinically-relevant structures, we want to note that it learns content-aware representations and pays attention to regions other than ducts and stromal tissue. For example, in Fig. 3 a, the most important identified bags and words do not belong to ducts or stromal tissue.

5.6. Model ablations

Effect of function Ψ—Table 4 compares the performance of three different Ψ functions with ESPNetv2 as a base feature extractor. Euclidean distance delivers the best performance. The model has 1% higher accuracy, sensitivity, and specificity values compared to the other two functions. In the rest of the experiments, we use Euclidean distance as a Ψ function.

HATNet with words only—Standard vision transformers (e.g., Dosovitskiy et al., 2021; Touvron et al., 2021) uses only words for learning visual representations. To understand the effectiveness of our hierarchical approach, we trained HATNet with words only. Results in Table 5 shows that hierarchical approach improves the performance over word-only model significantly. These observations are consistent with concurrent works that also shows that

[3]We are interested in evaluating if our model pays attention to ductal regions or stroma region. Therefore, we only use the top-k bags or words inside these regions while computing the dice score.

hierarchical approaches help vision transformers learn better representations (e.g., Liu et al., 2021; Mehta and Rastegari, 2022). In the rest of our experiments, we use both bags and words.

Effect of positional encoding—Positional embeddings are used in transformer-based models to incorporate positional information (Vaswani et al., 2017; Dosovitskiy et al., 2021). We found that the HATNet is insensitive to positional embeddings (see Table 5). This is likely because the top-down (image to bags to words) and bottom-up (words to bags to image) approach in the HATNet implicitly encodes the position of words and bags. As a result, it does not require any explicit positional information. Therefore, we do not use positional embeddings in the rest of our experiments.

Effect of bag and word sizes—Table 6 compares the performance of our model with three different bag-word size configurations using ESPNetv2 as a base feature extractor. The bag size of 1792 and word size of 256 delivered slightly better performance than the others. In the rest of the experiments, we use this bag-word size configuration.

Effect of $\widehat{\mathbf{B}}_{w2b}$—We noted in Section 4.3 that $\widehat{\mathbf{B}}_{w2b}$ aggregates local information and mimics skip-connections. To study its importance, we replaced $\mathbf{X}_Q = \widehat{\mathbf{B}}_{b2b}$ (which was derived from $\widehat{\mathbf{B}}_{w2b}$ in Eq. (5)) with $\mathbf{X}_Q = \overline{\mathbf{B}}_{w2b}$ in Eq. (6). As a result, the accuracy of HATNet dropped by 2% (0.67 to 0.65). This shows that information encoded via this skip-connection helps learn better representations. Our findings are consistent with recent (and parallel) work on vision transformers on the ImageNet dataset (Russakovsky et al., 2015), which also shows that vision transformer-based networks deliver better performance when both local and global information are encoded in contrast to global information only (e.g., Xiao et al., 2021; Dai et al., 2021). Therefore, we leave Eq. (6) as is and use $\mathbf{X}_Q = \widehat{\mathbf{B}}_{w2b}$ in the rest of our experiments.

Effect of different base feature extractors—Figure 6 compares the class-wise performance of HATNet with three different base feature extractors. HATNet with MNASNet delivers similar or better class-wise F1-score, sensitivity, and specificity values, except for the invasive case where MobileNetv2 has a higher sensitivity value.

Figure 7 plots the overall and class-wise receiver operating characteristics of HATNet with different base feature extractors. We observe that HATNet with MNASNet delivers the best performance (higher ROC-AUC) compared to the other two networks. Similarly, in Table 2 (R11-R13), HATNet delivers the best overall performance with MNASNet across different evaluation metrics. HATNet with MNASNet has 6% and 5% higher F1-score than with ESPNetv2 and MobileNetv2, respectively.

6. Discussion

This paper introduces a novel deep learning approach, HATNet, for classifying regions of interest (ROIs) of breast biopsy whole slide H&E images. Our experimental results showed that HATNet is able to achieve a pathologists-level performance on a challenging dataset that includes the full spectrum of diagnostic cases. Importantly, HATNet pays attention

to ductal regions and stromal tissue, two important clinical biomarkers in breast cancer diagnosis.

Earlier studies on diagnosing breast cancer using machine learning have focused on binary classification tasks, i.e., benign vs. malignant or DCIS vs. non-DCIS (Spanhol et al., 2015; Cruz-Roa et al., 2017; Bolhasani et al., 2020). For example, Spanhol et al. (2015) introduced BreakHis dataset and studied the binary classification (benign vs. malignant) using CNNs. Each histopathological sample in the dataset has a spatial dimension of 700 × 460 pixels. Because the full spectrum of breast cancer diagnosis is more complex then the binary classification tasks and the spatial resolution of samples in clinical settings is an order of magnitudes larger than the ones in the BreakHis dataset, Aresta et al. (2019) introduced the BACH dataset that provides multi-class diagnostic annotations (benign, DCIS, and invasive) for 40 variably-sized whole slide images (training set: 30; validation set: 10). The dataset also provides the performance of two pathologists as a reference. Similar to the BACH dataset, we introduced a breast biopsy dataset in our previous studies that provide multi-class diagnostic annotations (benign, atypia, DCIS, and invasive) for 240 variably-sized whole slide images, including an *independent* test set of 119 whole slide images. Unlike the BACH dataset, the images in our dataset were interpreted by 87 U.S. pathologists in an independent study; allowing us to compare the performance of HATNet with pathologists while accounting for the variability in diagnostic decisions among pathologists.

Most of the histopathological image classification networks for different tissue types (e.g., lung Hou et al., 2016 and colon Raczkowski et al., 2019 cancer) are multi-stage. Similar to these networks, the baseline networks in our study also have multiple stages. For example, Y-Net of Mehta et al. (2018b) has two stages. Such approaches are hindered in learning global representations. The HATNet brings these different stages under one umbrella and enables learning local (word-wise) and global (across words and bags) representations in a hierarchical and end-to-end fashion. This ability of aggregating information hierarchically at different levels (image, bags, and words) allows HATNet to learn representations from clinically relevant areas.

Previous work on this dataset used features extracted from tissue-level segmentation masks for diagnostic decisions. The ROI-level classification system of Mercan et al. (2019) used structural features extracted from tissue-level segmentation masks to predict the diagnosis. Using the same 4-classes as the current study, their system achieved an overall accuracy of 0.56 (R9 in Table 2). Y-Net of Mehta et al. (2018b) allowed for simultaneous classification and segmentation, and achieved a 4-class accuracy of 0.62 (R10 in Table 2). HATNet achieves a classification accuracy of 0.70 and outperforms these prior methods by a significant margin. Besides performance improvement, HATNet is 1. 8 × faster than the previous best model, Y-Net.

Unlike previous work, HATNet identifies important duct and stromal image areas of each ROI. HATNet uses self-attention at different levels (bags and words) to identify the salient areas. We studied the agreement between these salient areas and the annotations of clinical biomarkers (ducts and stroma) from expert pathologists. The fact that many of the bags that HATNet found important belonged to ductal regions also correlates with clinical studies

(Kinne et al., 1989; Page and Jensen, 1996; Shah et al., 2016) as well as our previous analysis of the importance of ducts in breast cancer diagnosis (Mercan et al., 2016; Li et al., 2020). The fact that many of the words that HATNet found important belonged to stromal tissue also correlates with clinical studies (Arendt et al., 2010; Conklin and Keely, 2012) and our previous analyses of stromal tissue in diagnostic classification (Mehta et al., 2018a; 2018b; Mercan et al., 2019). We emphasize that unlike our previous studies that supervised CNNs with the information about ducts and stromal tissue, HATNet figured out the importance of these regions (ductal architecture and stromal organization) in the diagnostic classification *without* explicit supervision about these biomarkers. How stroma and ducts may be architecturally important for classification is a topic for further study.

Strengths and limitations

This study introduces a novel diagnostic system using self-attention that allows the learning of representations in an end-to-end manner. One strength of this work is its study of the full clinical range of breast pathology (benign, atypia, DCIS, and Invasive) on an *independent* test set, not just a binary classification of tissue (e.g., invasive vs. non-invasive Spanhol et al., 2015, DCIS or non-DCIS Bolhasani et al., 2020). Another distinctive feature of our study is the ability to compare the classification decisions of our system with the data from multiple practicing U.S. pathologists who independently interpreted the same cases.

Despite the great promise of deep learning methods in pathology, we recognize the limitations of our study. HATNet was trained and validated on 204 ROIs (or 121 cases) and tested on 216 ROIs (or 119 cases). HATNet should be tested on a different independent set of breast biopsy cases to study its unbiased effectiveness. Also, this work only studies breast tissue. It should be tested on different tissue types to study its generalizability. Additionally, similar to previous work on this dataset and other datasets (e.g., BreakHis Spanhol et al., 2015 and BACH Aresta et al., 2019), HATNet was designed as a *directed second opinion* system wherein pathologists mark a region that they want to study carefully for final diagnosis. However, HATNet is generic and we believe that it can be extended to entire whole slide images either directly using a transformer-based bag-of-words approach (Fig. 2) or using a ROI detection system (Mercan et al., 2016; Gecer et al., 2018).

7. Conclusion

A diagnosis of cancer and pre-invasive risk lesions relies on human pathologists, and yet these diagnoses can be challenging, with marked intra- and inter-observer variability reported Elmore et al. (2015). With whole slide digital imaging now approved by the FDA, we will see growth in available data to develop and validate machine learning tools to help support pathologists in difficult cases. We introduced an end-to-end attention-based network, HATNet, for classifying breast biopsy images. HATNet extends bag-of-words models using Transformers to learn global representations. Our approach effectively aggregates inter-word and inter-bag representations, allowing HATNet to learn representations from clinically relevant areas and helping us explain its predictions. We believe that this ability to point out areas important to its diagnosis will facilitate improved interactions between computer-aided diagnostic tools and clinicians, helping to reduce classification uncertainties.

Acknowledgements

This work was supported in part by the National Cancer Institute awards (R01 CA172343, R01 CA140560, and RO1 CA200690).

References

Abraham N, Khan NM, 2019. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 683–687.

Allison KH, Reisch LM, Carney PA, Weaver DL, Schnitt SJ, O'Malley FP, Geller BM, Elmore JG, 2014. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. Histopathology 65 (2), 240–251. [PubMed: 24511905]

Arendt LM, Rudnick JA, Keller PJ, Kuperwasser C, 2010. Stroma in breast development and disease. In: Seminars in Cell & Developmental Biology, vol. 21. Elsevier, pp. 11–18. [PubMed: 19857593]

Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, Marami B, Prastawa M, Chan M, Donovan M, et al. , 2019. Bach: grand challenge on breast cancer histology images. Med. Image Anal. 56, 122–139. [PubMed: 31226662]

Basavanhally A, Ganesan S, Feldman M, Shih N, Mies C, Tomaszewski J, Madabhushi A, 2013. Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides. IEEE Trans. Biomed. Eng. 60 (8), 2089–2099. [PubMed: 23392336]

Bolhasani H, Amjadi E, Tabatabaeian M, Jassbi SJ, 2020. A histopathological image dataset for grading breast invasive ductal carcinomas. Inf. Med. Unlocked 19, 100341.

Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ, 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309. [PubMed: 31308507]

Chen L-C, Papandreou G, Schroff F, Adam H, 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J, 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 411–418.

Conklin MW, Keely PJ, 2012. Why the stroma matters in breast cancer: insights into breast cancer patient outcomes through the examination of stromal biomarkers. Cell Adhes. Migr. 6 (3), 249–260.

Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, Shih N, Tomaszewski J, Madabhushi A, 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Medical Imaging 2014: Digital Pathology, vol. 9041. International Society for Optics and Photonics, p. 904103.

Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NN, Tomaszewski J, González FA, Madabhushi A, 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Sci. Rep. 7 (1), 1–14. [PubMed: 28127051]

Custer RL, Scarcella JA, Stewart BR, 1999. The modified Delphi technique-a rotational modification.

Dai Z, Liu H, Le QV, Tan M, 2021. CoAtNet: Marrying Convolution and Attention for All Data Sizes. In: Beygelzimer A, Dauphin Y, Liang P, Wortman Vaughan J (Eds.). Advances in Neural Information Processing Systems. https://openreview.net/forum?id=dUk5Foj5CLf.

Della Mea V, Puglisi F, Bonzanini M, Forti S, Amoroso V, Visentin R, Dalla Palma P, Beltrami CA, 1997. Fine-needle aspiration cytology of the breast: a preliminary report on telepathology through internet multimedia electronic mail. Mod. Pathol. 10 (6), 636–641. [PubMed: 9195583]

DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, Jemal A, Siegel RL, 2019. Breast cancer statistics, 2019. CA Cancer J. Clin. 69 (6), 438–451. [PubMed: 31577379]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N, 2021. An image is worth 16 × 16

words: transformers for image recognition at scale. In: International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy

Elmore JG, Longton GM, Pepe MS, Carney PA, Nelson HD, Allison KH, Geller BM, Onega T, Tosteson AN, Mercan E, et al. , 2017. A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. J. Pathol. Inform. 8.

Elmore, et al. , 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA.

Gecer B, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG, 2018. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. Pattern Recognit. 84, 345–356. [PubMed: 30679879]

Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jegou H, Douze M, 2021. LeViT: A vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12259–12269.

He K, Zhang X, Ren S, Sun J, 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916. [PubMed: 26353135]

He K, Zhang X, Ren S, Sun J, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heker M, Greenspan H, 2020. Joint liver lesion segmentation and classification via transfer learning. Medical Imaging with Deep Learning. https://openreview.net/forum?id=qGWHiEgYAs

Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH, 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433.

Hou R, Grimm LJ, Mazurowski MA, Marks JR, King LM, Maley CC, Hwang S, Lo JY, 2020. A multi-task deep learning method in simultaneously predicting occult invasive disease in ductal carcinoma in situ and segmenting microcalcifications in mammography (Conference presentation). In: Medical Imaging 2020: Computer-Aided Diagnosis, vol. 11314. International Society for Optics and Photonics, p. 1131405.

Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, et al., 2019. Searching for mobilenetv3. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1314–1324.

Hu J, Shen L, Sun G, 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W, 2019. CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 603–612.

Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M, 2020. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Sci. Rep. 10 (1), 1–11. [PubMed: 31913322]

Ilse M, Tomczak J, Welling M, 2018. Attention-based deep multiple instance learning. In: International Conference on Machine Learning. PMLR, pp. 2127–2136.

Kern KA, 2001. The delayed diagnosis of breast cancer: medicolegal implications and risk prevention for surgeons. Breast Dis. 12 (1), 145–158. [PubMed: 15687615]

Kingma DP, Ba J, 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kinne DW, Petrek JA, Osborne MP, Fracchia AA, DePalo AA, Rosen PP, 1989. Breast carcinoma in situ. Arch. Surg. 124 (1), 33–36. [PubMed: 2535929]

Li B, Mercan E, Mehta S, Knezevich S, Arnold CW, Weaver DL, Elmore JG, Shapiro LG, 2020. Classifying breast histopathology images with a ductal instance-oriented pipeline. In: 25th International Conference on Pattern Recognition.

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B, 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Long J, Shelhamer E, Darrell T, 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lu C, Mandal M, 2015. Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. Pattern Recognit. 48 (8), 2738–2750.

Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F, 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570. [PubMed: 33649564]

Lu X, Mehta S, Bruny¸ E, T.T., Weaver DL, Elmore JG, Shapiro LG, 2021. Analysis of regions of interest and distractor regions in breast biopsy images. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–4. Doi:10.1109/BHI50953.2021.9508513.

Makki J, 2015. Diversity of breast carcinoma: histological subtypes and clinical relevance. Clin. Med. Insights Pathol. 8, CPath–S31563.

Mao Y, Keller ET, Garfield DH, Shen K, Wang J, 2013. Stromal cells in tumor microenvironment and breast cancer. Cancer Metastasis Rev. 32 (1–2), 303–315. [PubMed: 23114846]

Maron O, Lozano-Pérez T, 1998. A framework for multiple-instance learning. Adv. Neural Inf. Process. Syst. 570–576.

Mehta S, Mercan E, Bartlett J, Weaver D, Elmore J, Shapiro L, 2018. Learning to segment breast biopsy whole slide images. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 663–672.

Mehta S, Mercan E, Bartlett J, Weaver D, Elmore JG, Shapiro L, 2018. Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 893–901.

Mehta S, Rastegari M, 2022. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations. https://openreview.net/forum?id=vh-0sUt8HlG

Mehta S, Rastegari M, Shapiro L, Hajishirzi H, 2019. ESPNetv2: a light-weight, power efficient, and general purpose convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9190–9200.

Mercan C, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG, 2017. Multi--instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans. Med. Imaging 37 (1), 316–325. [PubMed: 28981408]

Mercan E, Aksoy S, Shapiro LG, Weaver DL, Brunyé TT, Elmore JG, 2016. Localization of diagnostically relevant regions of interest in whole slide images: a comparative study. J. Digit. Imaging 29 (4), 496–506. [PubMed: 26961982]

Mercan E, Mehta S, Bartlett J, Shapiro LG, Weaver DL, Elmore JG, 2019. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. JAMA Netw. Open 2 (8). e198777–e198777 [PubMed: 31397859]

Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, Mc-Donagh S, Hammerla NY, Kainz B, et al. , 2018. Attention U-Net: learning where to look for the pancreas. Medical Imaging with Deep Learning.

Page DL, Jensen RA, 1996. Ductal carcinoma in situ of the breast: understanding the misunderstood stepchild. JAMA 275 (12).

Plava J, Cihova M, Burikova M, Matuskova M, Kucerova L, Miklikova S, 2019. Recent advances in understanding tumor stroma-mediated chemoresistance in breast cancer. Mol. Cancer 18 (1), 67. [PubMed: 30927930]

Razkowski Ł, Mo zejko˙ M, Zambonelli J, Szczurek E, 2019. ARA: accurate, reliable and active histopathological image classification framework with bayesian deep learning. Sci. Rep. 9 (1), 1–12. [PubMed: 30626917]

Reisch LM, Carney PA, Oster NV, Weaver DL, Nelson HD, Frederick PD, Elmore JG, 2015. Medical malpractice concerns and defensive medicine: a nationwide survey of breast pathologists. Am. J. Clin. Pathol. 144 (6), 916–922. [PubMed: 26572999]

Ronneberger O, Fischer P, Brox T, 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rundo L, Han C, Nagano Y, Zhang J, Hataya R, Militello C, Tangherloni A, Nobile MS, Ferretti C, Besozzi D, et al. , 2019. Use-Net: incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. Neurocomputing 365, 31–43.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. , 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C, 2018. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D, 2019. Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207. [PubMed: 30802813]

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Shah C, Wobb J, Manyam B, Kundu N, Arthur D, Wazer D, Fernandez E, Vicini F, 2016. Management of ductal carcinoma in situ of the breast: a review. JAMA Oncol. 2 (8), 1083–1088. [PubMed: 27253401]

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

Spanhol FA, Oliveira LS, Petitjean C, Heutte L, 2015. A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. 63 (7), 1455–1462. [PubMed: 26540668]

Sun C, Xu A, Liu D, Xiong Z, Zhao F, Ding W, 2019. Deep learning-based classification of liver cancer histopathology images using only global labels. IEEE J. Biomed. Health Inform..

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A, 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Tan M, Chen B, Pang R, Vasudevan V, Sandler M, Howard A, Le QV, 2019. MnasNet: platform-aware neural architecture search for mobile. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2820–2828.

Tharwat A, 2018. Classification assessment methods. Appl. Comput. Inf..

Thome N, Bernard S, Bismuth V, Patoureaux F, et al., 2019. Multitask classification and segmentation for cancer diagnosis in mammography. In: International Conference on Medical Imaging with Deep Learning–Extended Abstract Track.

Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S, 2019. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw. Open 2 (11). e1914645–e1914645 [PubMed: 31693124]

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H, 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008.

Wang H, Kembhavi A, Farhadi A, Yuille AL, Rastegari M, 2019. Elastic: improving CNNs with dynamic scaling policies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2258–2267.

Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C, 2020. Axial-DeepLab: stand-alone axial-attention for panoptic segmentation. arXiv preprint arXiv: 2003.07853.

Wang X, Chen H, Gan C, Lin H, Dou Q, Huang Q, Cai M, Heng P-A, 2018. Weakly supervised learning for whole slide lung cancer image classification. Med. Imaging Deep Learn..

Wang X, Chen H, Gan C, Lin H, Dou Q, Tsougenis E, Huang Q, Cai M, Heng P-A, 2019. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern..

Wang X, Fang Y, Yang S, Zhu D, Wang M, Zhang J, Tong K. y., Han X, 2021. A hybrid network for automatic hepatocellular carcinoma segmentation in H&E-stained whole slide images. Med. Image Anal. 68, 101914. [PubMed: 33285479]

Wells WA, Carney PA, Eliassen MS, Tosteson AN, Greenberg ER, 1998. Statewide study of diagnostic agreement in breast pathology. JNCI J. Natl. Cancer Inst. 90 (2), 142–145. [PubMed: 9450574]

Woo S, Park J, Lee J-Y, So Kweon I, 2018. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.

Xiao T, Singh M, Mintun E, Darrell T, Dollár P, Girshick R, 2021. Early convolutions help transformers see better. arXiv preprint arXiv:2106.14881.

Xu D, Wang W, Tang H, Liu H, Sebe N, Ricci E, 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3917–3925.

Xu Y, Jia Z, Ai Y, Zhang F, Lai M, Eric I, Chang C, 2015. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 947–951.

Yang H, Kim J-Y, Kim H, Adhikari SP, 2019. Guided soft attention network for classification of breast cancer histopathology images. IEEE Trans. Med. Imaging 39 (5), 1306–1315. [PubMed: 31634125]

Yao Q, Gong X, 2020. Saliency guided self-attention network for weakly and semi–supervised semantic segmentation. IEEE Access 8, 14413–14423.

Zhang B-N, Cao X-C, Chen J-Y, Chen J, Fu L, Hu X-C, Jiang Z-F, Li H-Y, Liao N, Liu D-G, et al. , 2012. Guidelines on the diagnosis and treatment of breast cancer (2011 edition). Gland Surg. 1 (1), 39. [PubMed: 25083426]

Zhao H, Shi J, Qi X, Wang X, Jia J, 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

Zoph B, Vasudevan V, Shlens J, Le QV, 2018. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710.

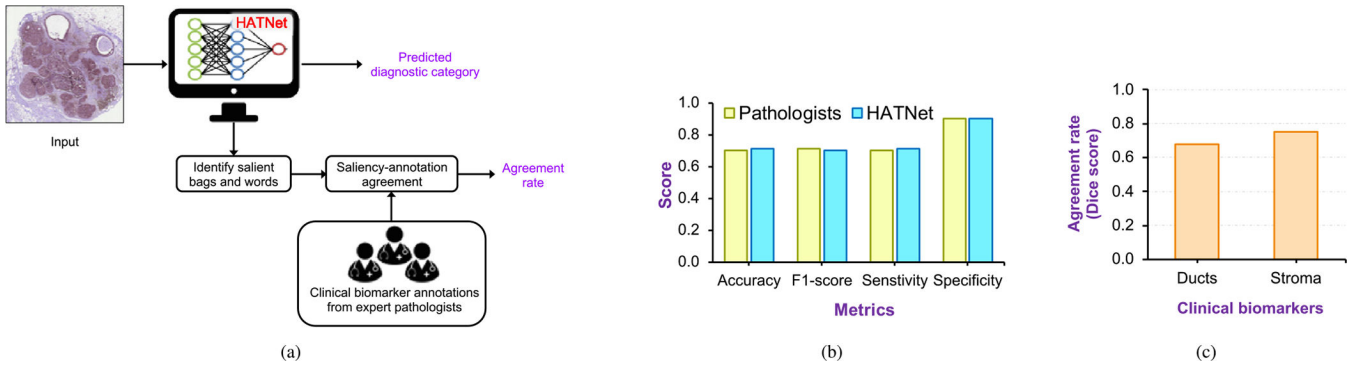


Fig. 1.

HATNet learns representations from clinically relevant biomarkers, allowing it to deliver similar performance to that of practicing pathologists. **(a)** HATNet for cancer diagnosis and interpretability, **(b)** performance comparison with HATNet and 87 U.S. pathologists, and **(c)** agreement rate of salient regions with clinical biomarker annotations from experts.

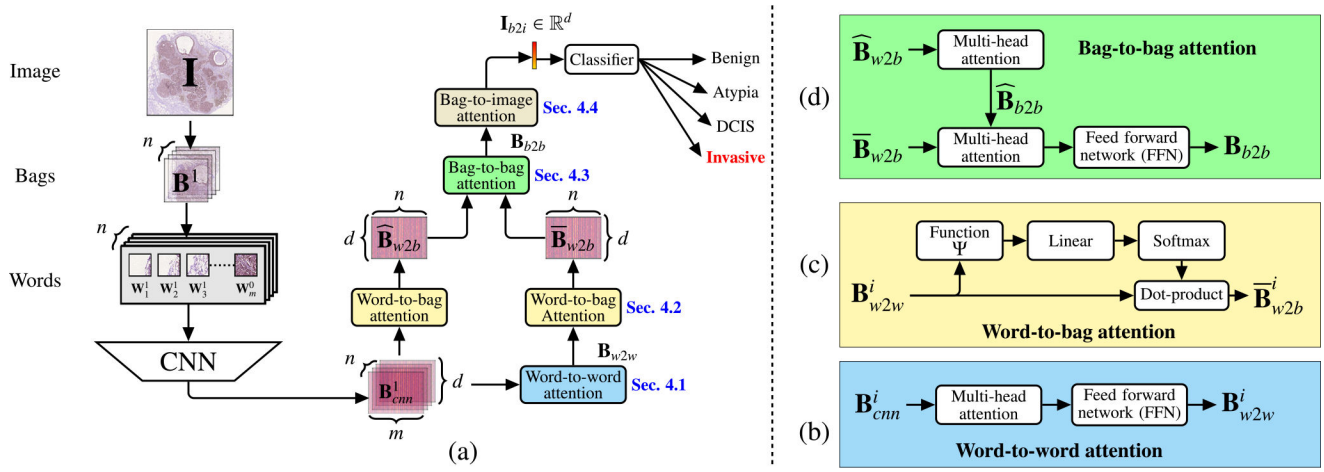


Fig. 2.

(a) HATNet: Our end-to-end holistic attention network for classifying breast biopsy images models the relationships between bags and words in a hierarchical manner using self attention. (b-d) Word-to-word, word-to-bag, and bag-to-bag attention modules are visualized; they allow the learning of relationships between bags and words using a bottom-up method. Note that the word-to-bag attention module for processing $\mathbf{B}_{cnn}$ and the bag-to-image attention module for processing $\mathbf{B}_{b2b}$ are similar to (c) and therefore, we do not visualize them here.

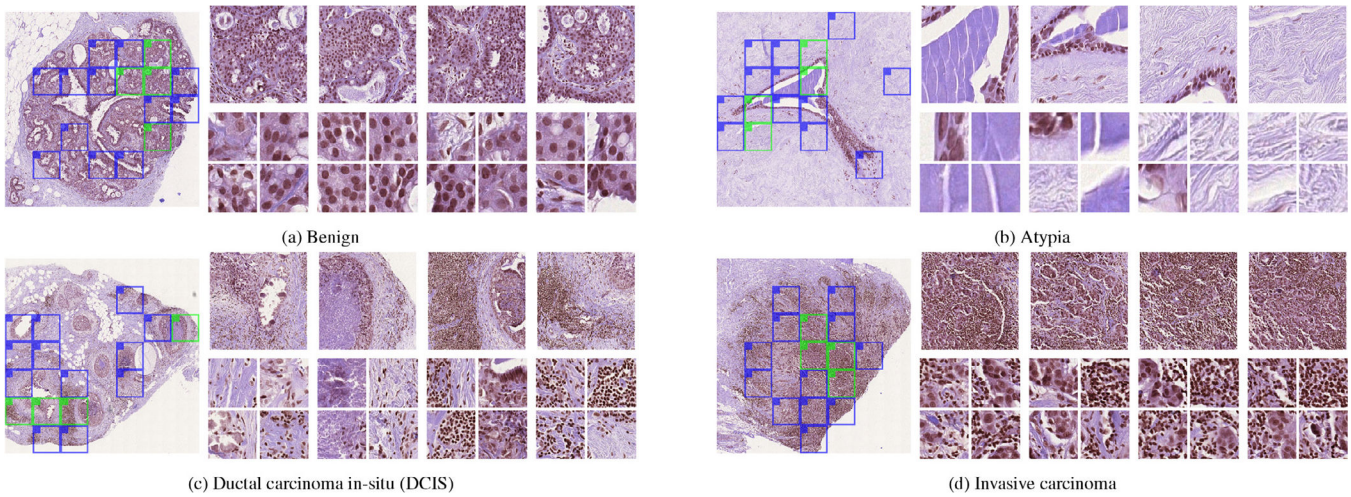


Fig. 3.

Example results of bags and words identified using HATNet across different diagnostic categories. HATNet aggregates information from different parts of the image and different textures. Here, each sub-figure of the breast biopsy image is shown on the left of each panel with the top-30% bags (top-4 in **green**, the rest in **blue**) identified using HATNet overlayed on the image. The upper right in each panel shows the top-4 bags, and the bottom right in each panel shows the top-4 words in each bag.

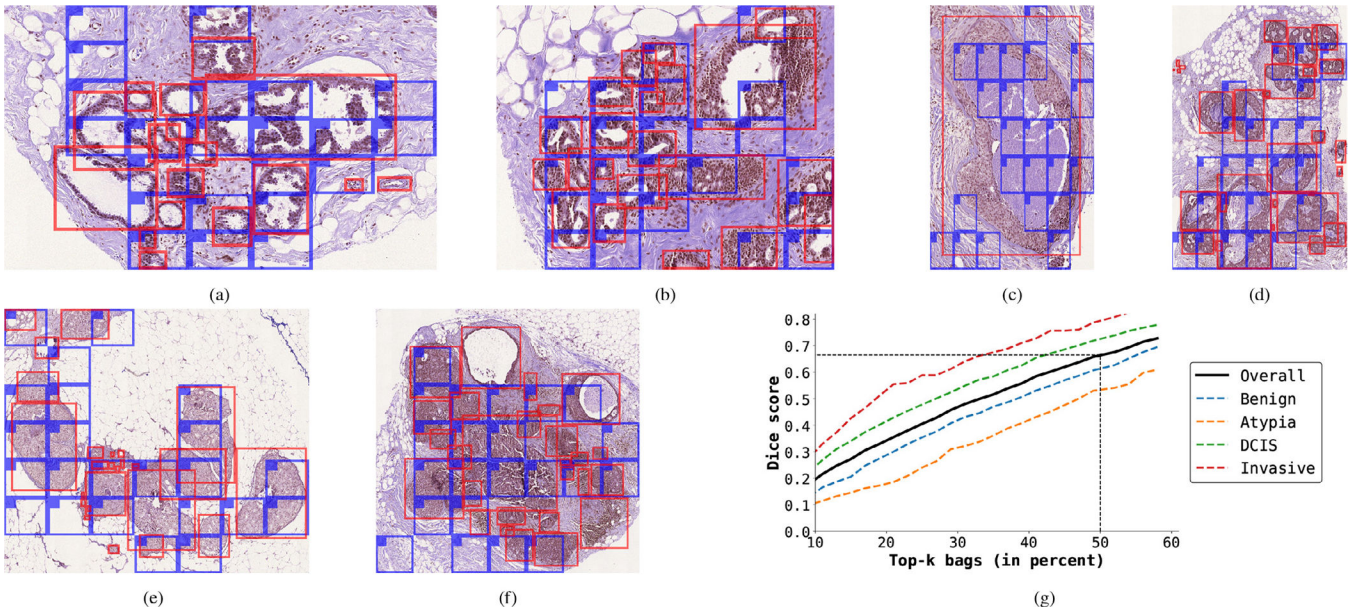


Fig. 4.

HATNet identifies ducts of variable size and texture as an important structure. In (a–f), ductal regions (marked by pathologists) are shown in red, while the top-50% bags predicted by HATNet are shown in blue. In (g), the dice score is plotted between ductal regions and top-k bag predictions (k varies from 10% to 60%) for different diagnostic classes.

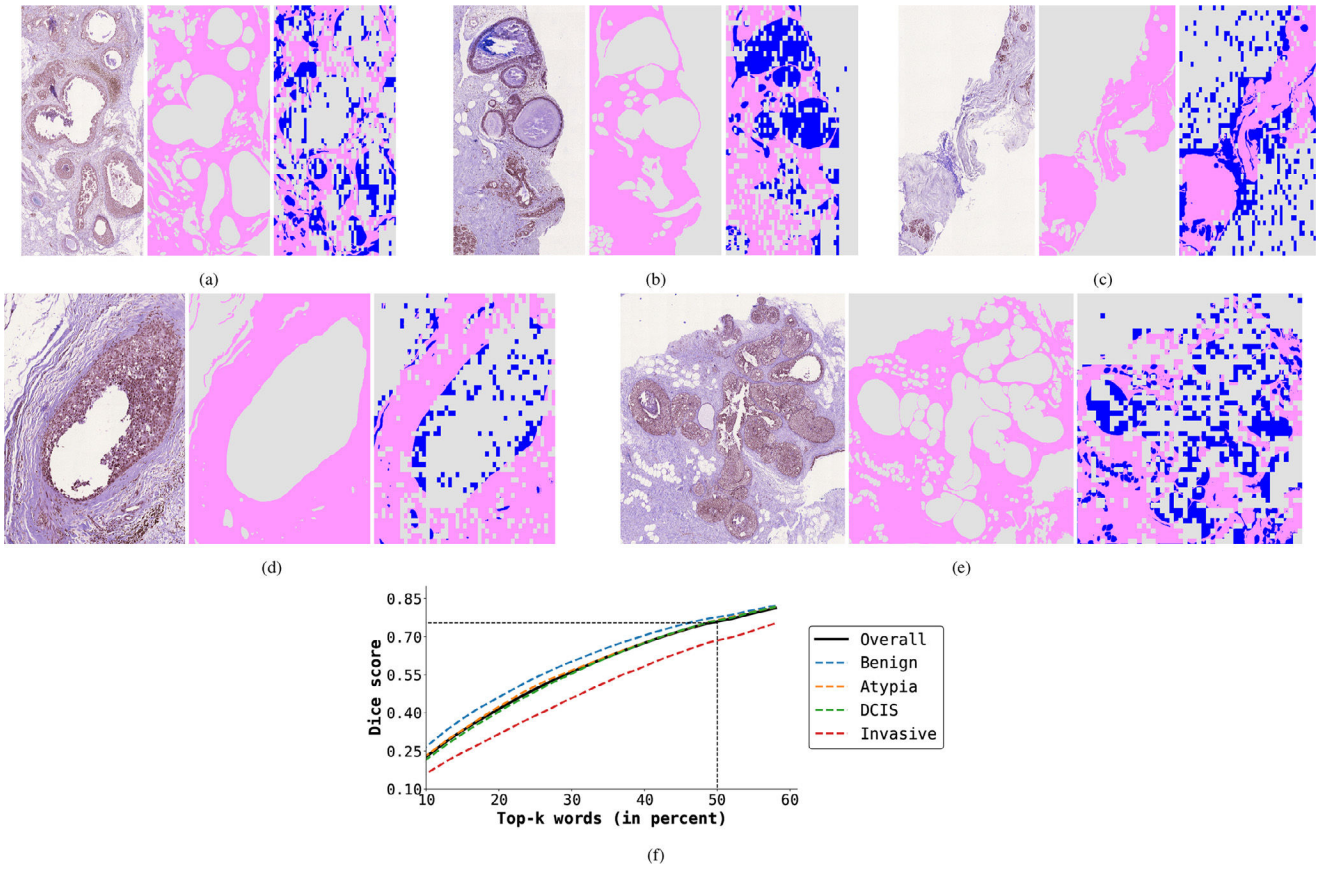
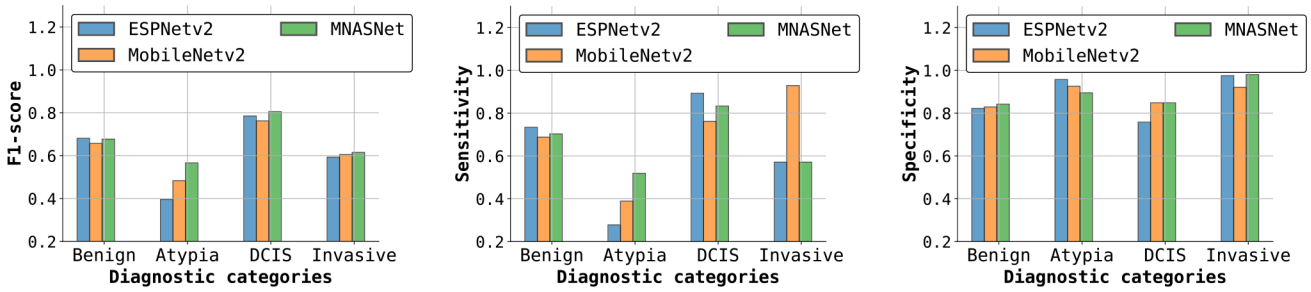


Fig. 5.

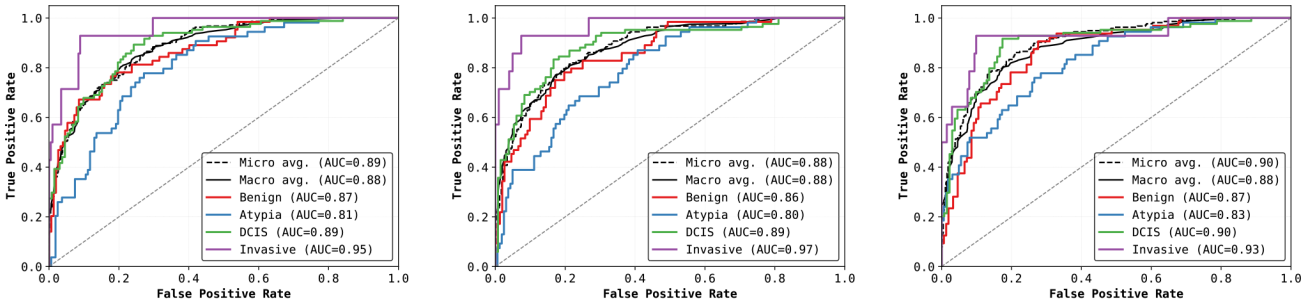
HATNet identifies stroma as an important tissue. In (a-e), each sub-figure is organized from left to right as: breast biopsy image, **stroma tissue** labeled by pathologists, and the top-50% words (words that belong to stroma tissue are shown in **pink** while the remaining words are shown in **blue**) identified using our model. The remaining 50% words are shown in white. In (f), we plot the dice score between stromal tissue and top-k word predictions (k varies from 10% to 60%) for different diagnostic classes.



(a) F1-score (b) Sensitivity (c) Specificity

Fig. 6.

Class-wise performance comparison of HATNet with different CNN architectures. Overall, the models with MNASNet as a base feature extractor performs a little better than the other two networks across different metrics. However, MNASNet has a low sensitivity score for Invasive Cancer, while MobileNetv2 does much better in this regard.



(a) ESPNetv2 (b) MobileNetv2 (c) MNASNet

Fig. 7.

Receiver operating characteristic (ROC) curves of HATNet with different CNN architectures. The models with MNASNet as a base feature extractor has slightly higher area under curve (AUC) than the other two.

Table 1.

Statistics of breast biopsy whole slide image dataset. (**a**) shows the distribution of ROIs for training, validation, and test set while (**b**) shows the slide-wise distribution for training + validation and test set. The 206 ROIs corresponding to 121 slides in the training + validation set are split randomly in 80:20 ratio to obtain training (164) and validation (42) ROIs.

Diagnostic Category	**Number of ROIs**				**Average ROI size (in pixels)**
	Training	**Validation**	**Test**	**Total**	
Benign	48	13	64	125	6,731 × 5,839
Atypia	40	8	54	102	10,668 × 8,967
DCIS	60	17	84	161	10,778 × 9,547
Invasive	16	4	14	34	23,866 × 21,402
Total	164	42	216	422	10,880 × 9,558

Diagnostic	**Number of whole slide images**		
Category	**Training + Validation**	**Test**	**Total**
Benign	39	39	79
Atypia	32	30	62
DCIS	39	39	79
Invasive	11	11	22
Total	121	119	240

Table 2.

Comparison with state-of-the-art networks.

HATNet outperforms existing methods by a significant margin. Network parameters are reported for single models only. We use majority voting for ensembling the models. † These works split the dataset (240 slides) into training (180 slides) and validation (60 slides) sets, and reports the performance on validation set. For completeness, we only report the accuracy of these methods. Note that the performance of networks in R8-R14 is on the same independent test set of 119 slides (training+validation/test slides: 121/119; Table 1).

Row No.	Model	Parameters			Evaluation metrics				
		CNN	Attn.		Accuracy	F1-score	Sensitivity	Specificity	ROC-AUC
R1	Pathologists (avg. of 87 practicing pathologists)			0.70	0.71	0.70	0.90		
R2†	LAB & LBP hand-crafted features (w/o saliency)			0.28					
R3†	LAB & LBP hand-crafted features (w/ saliency)			0.45					
R4†	Bag-of-word (majority voting w/o saliency)			0.23					
R5†	Bag-of-word (majority voting w/ saliency)			0.55					
R6†	Bag-of-word (learned fusion w/o saliency)			0.38					
R7†	Bag-of-word (learned fusion w/ saliency)			0.55					
R8	MRSegNet with histogram and co-occurence features	26.03 M	NA		0.55	0.56	0.55	0.85	
R9	MRSegNet with structural features	26.03 M	NA		0.56	0.57	0.56	0.85	
R10	Y-Net	3.91 M	NA		0.62	0.62	0.62	0.87	
R11	HATNet (w/ ESPNetv2)	2.21 M	2.37 M		0.67	0.64	0.67	0.89	0.89
R12	HATNet (w/ MobileNetv2)	2.22 M	2.37 M		0.66	0.65	0.66	0.89	0.88
R13	HATNet (w/ MNASNet)	3.10 M	2.37 M		**0.70**	**0.70**	**0.70**	**0.90**	**0.90**
R14	HATNet (Ensemble)	NA	NA		**0.71**	**0.70**	**0.71**	**0.90**	**0.90**

Table 3

Comparison with Y-Net in terms of accuracy and inference time.

HATNet is fast and accurate compared to previous best model (Y-Net). The two-tailed p-value between HATNet and Y-Net is less than 0.0001. Inference time is measured on a machine with a single NVIDIA GTX 1080 Ti GPU, and is an average across 100 trails on the validation set. The accuracy is an average of three models trained with different random seeds (0, 100, and 1000). The training time for HATNet and Y-Net is about 1.5 days. The machine used for measuring inference time has four NVIDIA GTX 1080 Ti GPU, 64 GB RAM, and 64 core Intel®Xeon®CPU. For inference time, we used only one GPU and disabled the other three GPUs by using *CUDA_VISIBLE_DEVICES = 0* command.

Model	Accuracy	Inference time
Y-Net	0.62 ± 0.0074	3. 93 s ± 20 ms
HATNet (w/ ESPNetv2)	0.67 ± 0.0021	2. 63 s ± 19 ms
HATNet (w/ MobileNetv2)	0.66 ± 0.0032	2.17 s ± 10 ms
HATNet (w/ MNASNet)	**0.70** ± 0.0024	**2.13** s ± **12** ms

Table 4

Effect of different Ψ functions.

Function Ψ	Accuracy	F1-score	Senstivity	Specificity	ROC-AUC
Euclidean distance	0.67	0.64	0.67	0.89	0.89
Manhattan distance	0.66	0.64	0.66	0.88	0.89
Mean	0.66	0.64	0.66	0.88	0.89

Table 5

Effect of hierarchical learning and positional embeddings (PE).

Model	Accuracy	F1-score	Senstivity	Specificity	ROC-AUC
HATNet (words)	0.50	0.50	0.50	0.83	0.79
HATNet (words + bags)	**0.67**	**0.64**	**0.67**	**0.89**	**0.89**
HATNet (words + bags + PE)	0.65	0.65	0.65	0.87	0.88

Table 6

Effect of different bag and word sizes. Note the number of words in each configuration are the same (i.e. 49).

Bag size	Word size	Accuracy	F1-score	Senstivity	Specificity	ROC-AUC
1792 × 1792	256 × 256	0.67	0.64	0.67	0.89	0.89
2016 × 2016	288 × 288	0.67	0.64	0.67	0.89	0.88
2240 × 2240	320 × 320	0.66	0.64	0.66	0.88	0.89