Original Article

Molecular Classification of Breast Cancer Using Weakly Supervised Learning

Wooyoung Jang¹, Jonghyun Lee²,³, Kyong Hwa Park⁴, Aeree Kim¹, Sung Hak Lee⁵, Sangjeong Ahn³,⁶,⁷

¹Department of Pathology, Korea University Guro Hospital, Korea University College of Medicine, Seoul, ²Department of Medical and Digital Engineering, Hanyang University College of Engineering, Seoul, ³Department of Pathology, ⁴Division of Oncology/Hematology, Department of Internal Medicine, Korea University Anam Hospital, Korea University College of Medicine, Seoul, ⁵Department of Hospital Pathology, Seoul St. Mary’s Hospital, College of Medicine, The Catholic University of Korea, Seoul, ⁶Artificial Intelligence Center, Korea University Anam Hospital, Korea University College of Medicine, Seoul, ⁷Department of Medical Informatics, Korea University College of Medicine, Seoul, Korea

Correspondence: Sangjeong Ahn
Department of Pathology, Korea University College of Medicine, Artificial Intelligence Center, Korea University College of Medicine, Department of Medical Informatics, Korea University Anam Hospital, Korea University College of Medicine, 73, Goryeodae-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: 82-2-920-5590 Fax: 82-2-920-6576 E-mail: vanitasahn@gmail.com

Co-correspondence: Sung Hak Lee
Department of Hospital Pathology, Seoul St. Mary’s Hospital, College of Medicine, The Catholic University, 222, Banpo-daero, Seocho-gu, Seoul 06591, Korea
Tel: 82-2-2258-1617 Fax: 82-2-2258-1627 E-mail: hajjjang@catholic.ac.kr

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi:10.4143/crt.2024.113
Abstract

Purpose

The molecular classification of breast cancer is crucial for effective treatment. The emergence of digital pathology has ushered in a new era in which weakly supervised learning leveraging whole-slide images has gained prominence in developing deep learning models because this approach alleviates the need for extensive manual annotation. Weakly supervised learning was employed to classify the molecular subtypes of breast cancer.

Methods

Our approach capitalizes on two whole-slide image datasets: one consisting of breast cancer cases from the Korea University Guro Hospital (KG) and the other originating from The Cancer Genomic Atlas dataset (TCGA). Furthermore, we visualized the inferred results using an attention-based heat map and reviewed the histomorphological features of the most attentive patches.

Results

The KG+TCGA-trained model achieved an area under the receiver operating characteristics value of 0.749. An inherent challenge lies in the imbalance among subtypes. Additionally, discrepancies between the two datasets resulted in different molecular subtype proportions. To mitigate this imbalance, we merged the two datasets, and the resulting model exhibited improved performance. The attentive patches correlated well with widely recognized histomorphologic features. The triple-negative subtype has a high incidence of high-grade nuclei, tumor necrosis, and intratumoral tumor-infiltrating lymphocytes. The luminal A subtype showed a high incidence of collagen fibers.

Conclusions

The artificial intelligence (AI) model based on weakly supervised learning showed promising performance. A review of the most attentive patches provided insights into the predictions of
the AI model. AI models can become invaluable screening tools that reduce costs and workloads in practice.

**Keywords**

Breast neoplasms, Weakly supervised learning, Molecular classification, Computational pathology
Introduction

Breast cancer emerged as the prevailing female malignancy globally, accounting for the fifth leading cause of cancer-related mortality [1]. Breast cancer is a heterogeneous group of diseases showing diverse response patterns to various treatment modalities. Establishing tailored treatment strategies is paramount, and multiple endeavors have been undertaken [2,3]. Currently, the most prevalent method employed in practice is a molecular classification-based approach using the gene expression profile: luminal A, luminal B, human epidermal growth factor receptor 2 (HER-2)-enriched, and basal-like, or triple-negative breast cancer (TNBC) [4]. Each subtype exhibits unique responses and prognoses to therapeutic interventions, necessitating a subtype-specific treatment approach [4].

RNA-based signature assays, including the Predictor Analysis of Microarray 50, are fundamental for breast cancer subtyping [5]. In clinical settings, immunohistochemistry-based modalities are widely used with four surrogate markers (estrogen (ER), progesterone receptor (PR), HER-2, and Ki-67) to determine the breast cancer subtype [6]. However, the adoption of genetic assays has been limited by their high cost, long turnaround time, and requirement for suitable tissue samples [7], whereas immunohistochemistry is not only labor-intensive and time-consuming but also has an inherent disadvantage of inter- and intra-observer variability [8]. Therefore, developing an artificial intelligence (AI) trained with routine and universal hematoxylin and eosin (H&E) whole-slide images (WSIs) as a screening tool could significantly alleviate the workload of pathologists.

Deep learning has proven to be particularly suitable for image analysis [10], and with progress in computing power, it has become a valuable tool in computational pathology [5]. Deep learning image analysis has been applied in various medical fields, encompassing basic applications such as tumor detection, subtyping, and grading and advanced applications such
Ordinarily, a WSI is an extremely high-resolution image that cannot be managed in the same manner as a natural image. Thus, a prevalent strategy for dealing with WSI involves partitioning a complete image into smaller units referred to as instances (or patches) [11]. The condition of the WSI is then determined by merging the predictions from individual instances. However, this strategy requires labor-intensive instance-level annotation, which is practically unfeasible. To overcome this limitation, multiple instance learning (MIL) [12], a weakly supervised learning strategy that does not require exhaustive local patch-level annotations, is often applied, thereby substantially reducing the amount of manual effort [9,13,14]. The MIL model requires only slide-level class labels, which can be readily obtained from digitalized pathology information systems [14], providing an advantage for large-scale datasets. Moreover, despite weak supervision, MIL models exhibit comparable and, in some cases, even superior performance compared with conventional supervised learning models [9,14]. Owing to these advantages, MIL is now widely used in image classification tasks in pathology [15].

Only a few studies have employed deep learning to predict the molecular subtypes of breast cancer based on H&E WSIs [5,16,17]. Couture et al. used deep learning to train a binary classifier (basal-like vs. nonbasal-like). Jaber et al. employed deep convolutional neural networks and traditional machine learning methods such as principal component analysis and support vector machines to train an AI model. Liu et al. applied MIL and implemented various procedures, including discriminative patch selection and local outlier factors, to address the noisy label problem. However, these studies have limitations, such as being restricted to binary classification, which has limited clinical significance [16], requiring significant manual effort for dataset preparation [5], or necessitating complex MIL frameworks consisting of multiple stages instead of end-to-end neural networks [17].
In this study, we applied the MIL method with weak slide-level annotation and minimal modifications to reduce workload. Specifically, we applied constrained-attention-multiple-instance learning (CLAM) [18], an enhanced MIL algorithm with attention-based learning, to automatically identify subregions with high diagnostic value. The AI model was trained with H&E WSIs to classify breast cancer into four molecular subtypes. We evaluated the performance and reviewed the histomorphological features of the most attentive patches to investigate any interpretable tendency to gain insight from the model.

Materials and Methods

1. Dataset Configurations

This study employed two datasets: The Cancer Genomic Atlas Breast Cancer (TCGA) dataset and the Korea University Guro Hospital breast cancer (KG) dataset. The TCGA dataset comprised 1,009 patients and 1,072 WSIs associated with breast cancer, whereas the KG dataset comprised 480 patients and 604 WSIs. Patients who underwent surgical resection for breast cancer between 2018-01-01 and 2021-12-31 were included in the KG dataset. Patients who received neoadjuvant chemotherapy or had recurrent breast cancer or ductal carcinoma in situ were excluded from the study. The inclusion and exclusion criteria and the number of WSIs for each step are shown in Fig. 1. The total number of WSIs was 1,676. WSIs were constructed from sections obtained from formalin-fixed paraffin-embedded blocks and stained with hematoxylin and eosin. The TCGA dataset was scanned at resolutions of 20× and 40×. In contrast, the KG dataset was scanned at a resolution of 40×. Ethical approval for this study was granted by the Institutional Review Board (IRB) of the Korea University Guro Hospital, and the need for informed consent was waived (IRB number: 2023GR0308).

To classify patients into distinct subtypes, a surrogate subtyping approach based on ER,
PR, HER-2, and Ki-67 status studied by immunohistochemistry was employed in both the KG and TCGA datasets.

Subtypes were determined based on specific criteria [6]:

1) Luminal A: When ER and PR were positive, and HER-2 was negative, the Ki-67 labeling index was lower than 14%.

2) Luminal B:
   a. When HER-2 was negative, and ER was positive, at least one of the following criteria was satisfied.
      i. PR was negative.
      ii. Ki-67 labeling index was either equal to or higher than 14%.
   b. When HER-2 was positive, ER also had to be positive.

3) HER-2 enriched: When HER-2 was positive and both ER and PR were negative.

4) TNBC: When ER, PR, and HER-2 were all negative.

To ensure the robustness and reliability of the results, the datasets were randomly shuffled at the WSI level. A partition ratio of 6:2:2 was used to allocate the shuffled data to the training, validation, and test sets, respectively. The entire analysis was repeated five times to account for variations and uncertainties in the dataset. This repetition was conducted using a fivefold cross-validation scheme, which allowed for a comprehensive evaluation and validation of the classification performance.

2. Instance Generation of WSIs

Owing to hardware limitations, WSIs typically require subdivisions into smaller instances [18]. In this study, we used the method proposed by the CLAM framework to split the WSI into instances and generate features for each instance. This process can be described
by the following three steps:

1) Background Exclusion: To isolate the regions of interest, a preliminary step involves excluding the background from the WSI.

2) Instantiation: Individual instances were generated in an isolated region of interest. The size of each instance was $256 \times 256$ pixels.

3) Feature Extraction: Once instances have been identified and generated, feature extraction is performed for each instance. The pre-generated instances were resized to $224 \times 224$ pixels and fed into the feature extractor. For this purpose, this study employed a pretrained ResNet 50 model, originally trained on the ImageNet dataset (19). The feature extractor transformed the instances into meaningful representations by capturing relevant characteristics. Consequently, the resulting WSI representation comprised the number of significant instances multiplied by the feature dimension, which, in this case, was set to 1,024.

3. Molecular Subtype Classification via MIL

The features obtained using the CLAM framework were trained using the CLAM-SB model provided within the framework. Subsequently, multi-label predictions were made. An additional experiment was conducted to enhance the efficacy of the framework. We integrated TCGA and KG datasets. This integration aimed to address label imbalance in the dataset and augment the training data, thereby potentially improving the overall learning performance.

The CLAM-SB model operates by taking the WSI-compressed data as input, where each instance is represented by a feature dimension. The model maps the input data into four classes. In contrast, the CLAM framework is a multi-task learning model incorporating an attention mechanism, which assigns weights to important instances for effective learning, and an instance clustering strategy. The instance clustering method uses attention weights to assign
pseudo-labels and employs them as clusters. This approach assists the model in distinguishing between various instances and improving its overall performance.

The performances of the trained models were demonstrated by the area under the receiver operating characteristics (AUROC). A pairwise comparison of the AUROCs of the trained model was conducted using the Mann-Whitney U test to investigate the benefit of merging the datasets.

4. Histomorphologic Reviews of the Attentive Patches

We selected the top attentive patches based on their attention weights. The attentive patches were 420, 420, 150, and 411 for luminal A, luminal B, HER-2, and TNBC, respectively. They were reviewed by a pathologist (WJ), and the histomorphological features of the patches were recorded. In the case of tumor patches, features including nuclear grade, tumor necrosis, intratumoral tumor-infiltrating lymphocytes (TILs), and stromal TILs were analyzed, as shown in Fig. 2A. Nuclear grade was recorded using a two-tier system, categorizing conventional low and intermediate grades as low and conventional high grades as high. Intratumoral TILs were defined as lymphocytes within a tumor cluster directly contacting cancer cells without stromal cell intervention [20]. Stromal TILs were defined as lymphocytes found in the stroma between carcinoma cells within the boundary of a tumor without directly infiltrating tumor cell nests [21]. For quantitative analysis, we aimed to score stromal TILs as a continuous parameter using conventional scoring methods (i.e., calculating the percentage of the stromal area occupied by mononuclear inflammatory cells over the total stromal area within the tumor). However, the patches were too small to reliably represent the total stromal area within the tumor. As a result, stromal TILs were recorded solely based on their unequivocal presence within the patch. In the non-tumor patches, the recorded histomorphological features included lymphoid aggregates,
collagen fibers, red blood cells, neutrophils, and skin, which are illustrated in Fig. 2B. Chi-squared tests assessed the association between these features and the four subtypes. For the features that were statistically significant across the four subtypes, a pairwise comparison between the two subtypes was performed using the $p$-value adjusted by the Bonferroni correction.

5. Clinicopathologic Features According to the Subtype Predicted by Deep Learning Model

To assess the performance of the model in terms of its association with clinicopathological variables, statistical analyses were conducted using the KG test dataset. After selecting the tumor characteristics of breast cancer that correlate with molecular subtypes [22], the associations between the true and predicted labels were investigated. The chi-square test was used for binary variables, while the Kruskal-Wallis test was used for continuous variables to compare the proportions of characteristics among the molecular subtypes.

6. Hardware and Software

Single Nvidia A100 GPUs with 80GB of VRAM were used to train the model. Python 3.7 and Pytorch framework 1.13 were employed to build the model. The CLAM source code (https://github.com/mahmoodlab/CLAM) was modified slightly for the datasets. Overall hyperparameters were followed by the original CLAM framework, except for the learning rate ($1e^{-4}$) and instance loss (focal loss, $\gamma = 2.0$).

Results

1. Patients and Datasets

The label distributions are summarized in S1 Table. A notable class imbalance was
observed across all datasets, with particular emphasis on the significant underrepresentation of the HER-2 class compared with the other classes. Specifically, when considering the luminal A and luminal B subtypes, TCGA dataset exhibited a higher proportion of luminal A cases. However, this trend was reversed in the KG dataset, in which a higher proportion of luminal B cases was observed. Additionally, in relative terms, TNBC appeared more prevalent than the HER-2 class but less common than the other classes.

2. Molecular Subtype Classification

A procedural schematic of the investigation is shown in Fig. 3. The gigapixel WSIs underwent fragmentation and compression using a pretrained image encoder and subsequently underwent weak label classification via MIL.

The classification performance of the AI models was evaluated using the AUROC. The AUROCs of the models based on the training and test datasets are presented in Table 1 and visualized in Fig. 4. Significantly, the model's AUROCs demonstrated an upward trend when trained with the merged datasets (KG+TCGA) compared with training with KG or TCGA only. The model trained with KG+TCGA and tested with KG demonstrated the highest AUROC (0.749) among the models. These findings underscore the advantages of merging datasets to alleviate label imbalances. The confusion matrix of the KG+TCGA-trained model is shown in Fig. 5. The AI models demonstrated promising performance, particularly in TNBC, achieving 67.14% accuracy despite their relatively low proportion in the label distribution.

3. Attention Visualization of Each Subtype

During the training phase, the process of procuring a weak label for a WSI necessitates the assignment of attention weights to individual instances, signifying their relative significance
toward the overall label. These attention weights are then combined via a weighted sum, which enables the determination of the class for the entire WSI. Using this method, it becomes feasible to discern the class membership of each instance and indirectly comprehend the knowledge absorbed by the model from the distinct features of the instances that receive heightened attention. An example of attention visualization is graphically shown in Fig. 6.

4. Histomorphologic Reviews of the Attentive Patches

Regardless of the subtype, the attentive patches contained more non-tumor patches than tumor patches, as shown in S2 Table. Following this review, the histomorphological features of these patches were recorded, as shown in Table 2, Fig. 7, and S3 Fig.

The chi-square test was used to evaluate the correlation between histomorphological features and molecular subtypes. The results showed that nuclear grade, tumor necrosis, intratumoral TILs, lymphoid aggregates, and collagen fibers were statistically significant variables (p < 0.05). Further pairwise chi-square tests were performed between the two subtypes for the aforementioned features to identify statistically significant (p < 0.008, adjusted using the Bonferroni correction) pairs. The results are shown in S4 Table. Molecular subtypes were arranged in order of the frequency of specific histomorphological features, as illustrated in Table 3.

TNBC and HER-2 displayed the highest number of tumor patches with high-grade nuclei. HER-2 exhibited the highest number of tumor necrosis patches, followed by TNBC. In the case of intratumoral TIL patches, both HER-2 and TNBC exhibited a greater number than luminal A. Lastly, luminal A had the highest quantity of collagen fiber patches among all subtypes.
5. AI Model Can Identify Clinically Distinct Subtypes, Concordant with Molecular Classification

In both the true and predicted labels, the frequencies of high-grade nuclei differed significantly among the subtypes, as shown in the S5 Table. The results suggest a similar association between the predicted labels and clinicopathological variables, as observed with true labels. Additionally, while the predicted labels exhibited differences in mean tumor size, with a p-value of 0.062, the true labels did not exhibit significant differences, with a p-value of 0.249, indicating that the predicted labels demonstrated an even stronger association. The findings imply that the model performs well in terms of its association with tumor characteristics.

Discussion

In this study, AI models for predicting breast cancer molecular subtypes were developed using weakly supervised learning based on H&E WSIs. The performance of the models, assessed using AUROCs, exhibited an upward trend when merged with the training dataset (KG+TCGA), reaching the highest AUROC of 0.749. Attentive patches were visualized as a heatmap based on the attention weights of the individual patches. A subsequent histomorphological review of the top attentive patches revealed interpretable tendencies that were biologically plausible and aligned with the well-known histomorphological features of certain molecular subtypes.

Among previous studies investigating AI models for breast cancer molecular subtyping, the model by Liu et al. [17] demonstrated promising performance. However, the model was trained only on an in-house dataset and required complex modifications to a MIL framework consisting of multiple stages. In contrast, we used a publicly available TCGA dataset
in addition to the in-house dataset and a minimally modified CLAM framework to achieve end-to-end deep learning, making the model more versatile and accessible.

Both the KG and TCGA datasets exhibited severe label imbalances. Class-imbalanced data can negatively impact the performance of classification models and is a common challenge in developing machine learning models [18]. Recently, several studies have been conducted to overcome this imbalanced label distribution problem [19]. Although technical adjustments may be a viable solution, merging two datasets with different label distributions to increase positive bags in underrepresented classes is also reliable for enhancing performance. Our model demonstrated enhanced performance, especially in classes with sparse positive bags such as HER-2 and TNBC, when trained with the merged dataset (KG+TCGA).

While the model demonstrated promising performance in evaluation metrics such as AUROC, we also sought to assess its clinical significance. This assessment is essential to determine the model's potential for future application in actual clinical settings. To this end, we investigated the association between clinicopathological variables of breast cancer and both the true and predicted labels. The analysis revealed that the frequencies of high-grade nuclei showed a similar association between predicted labels and clinicopathological variables as observed with true labels. Additionally, mean tumor size demonstrated an even stronger association in the predicted labels, compared to the true labels. These findings suggest that the model not only performs well in evaluation metrics but also effectively correlates with clinically significant variables.

We generated a heatmap based on the attention weights assigned to the patches. Histomorphological review of the most attentive patches revealed the inclusion of tumor patches and non-tumor patches. The frequency of non-tumor patches consistently surpassed that of tumor patches for all subtypes. Compared with tumor patches, non-tumor patches could
provide unexpected but valuable insights, which may contribute to novel hypotheses of the underlying biological mechanisms. For instance, Brockmoeller et al. [23] suggested that ‘inflamed fat’ outside the primary tumor is a risk factor for lymph node metastasis in early-stage colorectal cancer. Based on these findings, we conducted a histomorphological analysis of the tumor and non-tumor patches to gain insight into explainable AI.

Interpretable tendencies in the histomorphological features extracted from the attentive patches were observed, which aligned with the widely reported histomorphological characteristics of certain breast cancer subtypes. For tumor patches, the findings included: (1) a higher occurrence of high-grade tumor patches in TNBC than in all other subtypes [24], (2) more tumor necrosis patches in TNBC than in luminal B [24], and (3) a greater presence of intratumoral TILs in TNBC than in luminal A [25,26]. Regarding non-tumor patches, more collagen fibers were observed in luminal A than in all the other subtypes [27]. However, certain features, such as lymphoid aggregates, pose challenges in their interpretation. In addition, certain tendencies contradicted widely known observations, such as HER-2 exhibiting a higher frequency of tumor necrosis than TNBC. It is important to note that because the AI model arbitrarily selects patches with an attention mechanism, the frequency of a specific histomorphological feature among the top attentive patches may not accurately depict the actual frequency of the same feature in that specific subtype. This discrepancy may explain the contradictions observed.

Our study had certain limitations that warrant consideration. First, two notable distinctions exist between the TCGA and KG datasets: (1) the multiplicity of institutions as the data source and (2) ethnic composition. The TCGA dataset was collected from multiple hospitals, whereas the KG dataset was collected from a single hospital. Regarding ethnic composition, TCGA dataset had a relatively low representation of the Asian population,
constituting only 6.1% [28], whereas most individuals in the KG dataset were of Asian descent. Since the proportion of molecular subtypes of breast cancer varies across different racial groups [29], this difference could influence the outcome. Therefore, these disparities between the two datasets may have contributed to the diminished performance of the AI model trained on TCGA dataset when applied to the KG test set and vice versa. Second, MIL-based methods have some inherent limitations. They treated patches from different locations on a slide as independent entities, disregarding the potential contextual relationships between patches. This limitation became apparent when evaluating features that require understanding the stroma surrounding the tumor, such as stromal TILs. To address this issue, future approaches may benefit from more context-aware techniques, such as graph-convolutional networks [30].

The AI model presented here was trained on routine H&E WSI using weakly supervised learning and showed good performance. The interpretable tendencies observed in the review of the top attentive patches suggest the possibility of explainable AI. With further refinement, AI models have the potential to become invaluable screening tools, reducing costs and workload in clinical practice.

**Ethical Statement**

Ethical approval for this study was granted by the Institutional Review Board (IRB) of the Korea University Guro Hospital, and the need for informed consent was waived (IRB number: 2023GR0308).
Author Contributions

Conceived and designed the analysis: Park KH, Kim A, Lee SH, Ahn S.
Collected the data: Jang W, Lee SH, Ahn S.
Contributed data or analysis tools: Jang W, Lee J, Lee SH, Ahn S.
Performed the analysis: Jang W, Lee J, Lee SH, Ahn S.
Wrote the paper: Jang W, Lee J, Lee SH, Ahn S.

ORCID iDs

Wooyoung Jang: https://orcid.org/0000-0001-6611-5378
Sung Hak Lee: https://orcid.org/0000-0003-1020-5838
Sangjeong Ahn: https://orcid.org/0000-0003-4338-3235

Conflicts of Interest

Conflict of interest relevant to this article was not reported.

Acknowledgments

This work was supported by research grants as follows; from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2021-KH113146); from a Korea University Grant (grant number: K2319661); and from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR22C1302).
References


### Table 1. The classification performances (AUROC) of AI models based on the training and test datasets

<table>
<thead>
<tr>
<th>Training dataset</th>
<th>Test dataset</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>KG</td>
<td>0.674 (0.06)</td>
</tr>
<tr>
<td></td>
<td>TCGA</td>
<td>0.611 (0.05)</td>
</tr>
<tr>
<td>KG</td>
<td>KG+TCGA</td>
<td>0.628 (0.03)</td>
</tr>
<tr>
<td>TCGA</td>
<td>KG</td>
<td>0.471 (0.08)</td>
</tr>
<tr>
<td></td>
<td>TCGA</td>
<td>0.622 (0.04)</td>
</tr>
<tr>
<td>KG+TCGA</td>
<td>KG</td>
<td>0.605 (0.04)</td>
</tr>
<tr>
<td>KG</td>
<td>KG+TCGA</td>
<td>0.749 (0.03)</td>
</tr>
<tr>
<td></td>
<td>TCGA</td>
<td>0.741 (0.04)</td>
</tr>
<tr>
<td>KG+TCGA</td>
<td>KG+TCGA</td>
<td>0.725 (0.03)</td>
</tr>
</tbody>
</table>

Note. Each metric is the average value of five cross-validation runs, and the values in parentheses correspond to the standard deviation.

AUROC, area under the receiver operating characteristics; KG, Korea University Guro Hospital; TCGA, The Cancer Genomic Atlas.
Table 2. Histomorphological features of top attentive patches in each subtype

<table>
<thead>
<tr>
<th></th>
<th>Luminal A</th>
<th>Luminal B</th>
<th>HER-2</th>
<th>TNBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-grade nucleus</td>
<td>6.6%</td>
<td>2.4%</td>
<td>51.1%</td>
<td>95.9%</td>
</tr>
<tr>
<td>Tumor necrosis</td>
<td>18.4%</td>
<td>0.0%</td>
<td>63.8%</td>
<td>28.6%</td>
</tr>
<tr>
<td>Intratumoral TILs</td>
<td>2.6%</td>
<td>10.7%</td>
<td>19.1%</td>
<td>20.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Luminal A</th>
<th>Luminal B</th>
<th>HER-2</th>
<th>TNBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymphoid aggregates</td>
<td>52.3%</td>
<td>95.2%</td>
<td>89.3%</td>
<td>90.9%</td>
</tr>
<tr>
<td>Collagen fibers</td>
<td>45.1%</td>
<td>4.5%</td>
<td>17.5%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

HER-2, human epidermal growth factor receptor 2; TILs, tumor-infiltrating lymphocytes; TNBC, triple-negative breast cancer.
Table 3. Statistically significant differences between the number of patches with specific histomorphological features of each subtype.

<table>
<thead>
<tr>
<th>Patch</th>
<th>Most frequent subtypes</th>
<th>1&lt;sup&gt;st&lt;/sup&gt;</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt;</th>
<th>3&lt;sup&gt;rd&lt;/sup&gt;</th>
<th>4&lt;sup&gt;th&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-grade nucleus</td>
<td>TNBC</td>
<td>HER-2</td>
<td>Luminal A</td>
<td>Luminal B</td>
<td></td>
</tr>
<tr>
<td>Tumor necrosis</td>
<td>HER-2</td>
<td>TNBC</td>
<td>Luminal A</td>
<td>Luminal B</td>
<td></td>
</tr>
<tr>
<td>Intratumoral TILs</td>
<td>TNBC and &lt;sup&gt;a&lt;/sup&gt;) HER-2</td>
<td>Luminal A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-tumor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lymphoid aggregates</td>
<td>HER-2</td>
<td>Luminal A</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Luminal B</td>
<td>TNBC</td>
<td>Luminal A</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Collagen fibers</td>
<td>Luminal A</td>
<td>HER-2</td>
<td>Luminal B</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Luminal A</td>
<td>TNBC</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>)The conjunction “and” does not indicate any significant statistical difference between the subtypes. HER-2, human epidermal growth factor receptor 2; TILs, tumor-infiltrating lymphocytes; TNBC, triple-negative breast cancer.
Fig. 1. Patient inclusion, exclusion criteria, and dataset configurations. To optimize model performance and mitigate class imbalance, the TCGA dataset and the KG dataset (an in-house dataset) were consolidated. TCGA: Breast cancer cases from The Cancer Genomic Atlas, KG: Breast cancer cases from Korea University Guro Hospital.
Fig. 2. Representative patches of the recorded histomorphologic features (A. tumor patches, B. non-tumor patches). Tumor patches encompassed high-grade nucleus (1), tumor necrosis (2), intratumoral tumor-infiltrating lymphocytes (TILs) (3), and stromal TILs (4). Non-tumor patches included lymphoid aggregates (5), collagen fibers (6), red blood cells (7), neutrophils (8), and skin (9).
**Fig. 3.** Diagrammatic representation of the study design. The input whole-slide images (WSIs) were segmented into multiple instances, each encoded using a pretrained encoder (ResNet encoder). These instances were then aggregated and applied to classify breast cancer subtypes. The attention score, indicative of the degree of focus, identified the most significant instances pertinent to subtype classification.
**Fig. 4.** The area under the receiver operating characteristics of artificial intelligence models based on the training and test datasets. An upward trend with the merged dataset (KG + TCGA) is noted.
Fig. 5. Confusion matrix of prediction outputs. The diagonal axis represents accurate predictions, and each confusion matrix has been normalized.
Fig. 6. Visualization of attention within a given whole-slide image (WSI). Panel A provides a thumbnail of the WSI. Panel B presents the attention weights attributed to each instance. In Panel C, the magnified instances that are most significant for model predictions are depicted.
Fig. 7. Representative top attentive patches of human epidermal growth factor receptor 2 (HER-2) and triple-negative breast cancer (TNBC) subtypes. Both subtypes showed a high frequency of high-grade nucleus, tumor necrosis, and intratumoral tumor-infiltrating lymphocytes (TILs) patches.
### S1 Table. Label distribution of molecular subtypes in each dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Luminal A</th>
<th>Luminal B</th>
<th>HER-2</th>
<th>TNBC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA</td>
<td>616 (57.46)</td>
<td>243 (22.67)</td>
<td>39 (3.64)</td>
<td>174 (16.23)</td>
<td>1,072</td>
</tr>
<tr>
<td>KG</td>
<td>149 (24.67)</td>
<td>380 (62.91)</td>
<td>21 (3.48)</td>
<td>54 (8.94)</td>
<td>604</td>
</tr>
<tr>
<td>Total</td>
<td><strong>765 (45.64)</strong></td>
<td><strong>623 (37.18)</strong></td>
<td><strong>60 (3.58)</strong></td>
<td><strong>228 (13.60)</strong></td>
<td><strong>1,676</strong></td>
</tr>
</tbody>
</table>

Note. The value in the bracket indicates the proportion of each class. HER-2, human epidermal growth factor receptor 2; KG, Korea University Guro Hospital; TCGA, The Cancer Genomic Atlas; TNBC, triple-negative breast cancer.
### S2 Table. Number of top attentive patches and non-tumors in each subtype

<table>
<thead>
<tr>
<th></th>
<th>Luminal A</th>
<th>Luminal B</th>
<th>HER-2</th>
<th>TNBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top attentive patches</td>
<td>420</td>
<td>420</td>
<td>150</td>
<td>411</td>
</tr>
<tr>
<td>Tumor</td>
<td>76 (18.1%)</td>
<td>84 (20%)</td>
<td>47 (31.3%)</td>
<td>49 (11.9%)</td>
</tr>
<tr>
<td>Non-tumor</td>
<td>344 (81.9%)</td>
<td>336 (80.0%)</td>
<td>103 (68.7%)</td>
<td>362 (88.1%)</td>
</tr>
</tbody>
</table>

HER-2, human epidermal growth factor receptor 2; TNBC, triple-negative breast cancer.
S3 Fig. Representative top attentive patches of the luminal A and B subtypes. Both subtypes showed a high frequency of collagen fibers and lymphoid aggregates. Luminal B did not exhibit patches of tumor necrosis.
**S4 Table.** Chi-square test results of the association between histomorphological features and molecular subtypes

<table>
<thead>
<tr>
<th>Feature</th>
<th>Tumor Between all four subtypes (Statistically significant: p &lt; 0.05)</th>
<th>Tumor Between two subtypes (statistically significant: p &lt; 0.008&lt;sup&gt;a&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LA, LB</td>
<td>LA, HER-2</td>
</tr>
<tr>
<td>High-grade nucleus</td>
<td>&lt; 0.001</td>
<td>0.258</td>
</tr>
<tr>
<td>Tumor necrosis</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Intratumoral TILs</td>
<td>0.007</td>
<td>0.044</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Non-tumor Between all four subtypes (Statistically significant: p &lt; 0.05)</th>
<th>Non-tumor Between two subtypes (statistically significant: p &lt; 0.008&lt;sup&gt;a&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LA, LB</td>
<td>LA, HER-2</td>
</tr>
<tr>
<td>Lymphoid aggregates</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Collagen fibers</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Statistically significant $p$ values are highlighted in bold. HER-2, human epidermal growth factor receptor 2; LA, luminal A; LB, luminal B; TILs, tumor-infiltrating lymphocytes; TNBC, triple-negative breast cancer. <sup>a</sup> For pairwise comparisons between the two subtypes, the $p$ value was adjusted using Bonferroni correction.
**S5 Table.** Clinicopathologic correlation between ground truth and prediction

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>True labels</th>
<th>Predicted labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Luminal A</td>
<td>Luminal B</td>
</tr>
<tr>
<td>Number (% of total)</td>
<td>24 (26%)</td>
<td>56 (61%)</td>
</tr>
<tr>
<td>Mean age</td>
<td>59.1</td>
<td>55.1</td>
</tr>
<tr>
<td>Mean tumor size (cm)</td>
<td>1.8</td>
<td>2.0</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ductal</td>
<td>14 (58%)</td>
<td>46 (82%)</td>
</tr>
<tr>
<td>Nuclear grade</td>
<td>0 (0%)</td>
<td>21 (38%)</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodal involvement</td>
<td>5 (21%)</td>
<td>15 (27%)</td>
</tr>
<tr>
<td>Multifocality</td>
<td>6 (25%)</td>
<td>9 (16%)</td>
</tr>
<tr>
<td>LVI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 (13%)</td>
<td>10 (18%)</td>
</tr>
<tr>
<td>EIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>7 (29%)</td>
<td>15 (27%)</td>
</tr>
</tbody>
</table>

Note. Statistically significant *p*-values are highlighted in bold. *P*-values that are not statistically significant but lower than 0.1 is highlighted in italic. The chi-square test was employed for binary variables, while the Kruskal-Wallis test was utilized for continuous variables to compare the proportion of characteristics among the molecular subtypes. The value in the bracket corresponds to the percentage within each label. AI, artificial intelligence; HER-2, human epidermal growth factor receptor 2; TNBC, triple-negative breast cancer; LVI, lymphovascular invasion; EIC, extensive intraductal component.