Journal of Pathology Informatics 15 (2024) 100363

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Review Article

Publicly available datasets of breast histopathology H&E whole-slide images: A scoping review

Masoud Tafavvoghi [a,*], Lars Ailo Bongo [b], Nikita Shvetsov [b], Lill-Tove Rasmussen Busund [c], Kajsa Møllersen [a]

[a] Department of Community Medicine, Uit The Arctic University of Norway, Tromsø, Norway
[b] Department of Computer Science, Uit The Arctic University of Norway, Tromsø, Norway
[c] Department of Medical Biology, Uit The Arctic University of Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Breast cancer
Computational pathology
Deep learning
Whole-slide images
Publicly available datasets

ABSTRACT

Advancements in digital pathology and computing resources have made a significant impact in the field of computational pathology for breast cancer diagnosis and treatment. However, access to high-quality labeled histopathological images of breast cancer is a big challenge that limits the development of accurate and robust deep learning models. In this scoping review, we identified the publicly available datasets of breast H&E-stained whole-slide images (WSIs) that can be used to develop deep learning algorithms. We systematically searched 9 scientific literature databases and 9 research data repositories and found 17 publicly available datasets containing 10 385 H&E WSIs of breast cancer. Moreover, we reported image metadata and characteristics for each dataset to assist researchers in selecting proper datasets for specific tasks in breast cancer computational pathology. In addition, we compiled 2 lists of breast H&E patches and private datasets as supplementary resources for researchers. Notably, only 28% of the included articles utilized multiple datasets, and only 14% used an external validation set, suggesting that the performance of other developed models may be susceptible to overestimation. The TCGA-BRCA was used in 52% of the selected studies. This dataset has a considerable selection bias that can impact the robustness and generalizability of the trained algorithms. There is also a lack of consistent metadata reporting of breast WSI datasets that can be an issue in developing accurate deep learning models, indicating the necessity of establishing explicit guidelines for documenting breast WSI dataset characteristics and metadata.

Contents

Introduction

One important area of active research in pathology is the use of deep learning for analyzing H&E histopathology whole-slide images (WSIs)-the gold-standard for the clinical diagnosis of cancer.[1] Deep learning algorithms can identify complex patterns in billion-pixel microscope images that may not be readily apparent to human experts (an example is shown in Fig. 1). For instance, deep learning models have been used to predict

* Corresponding author.
E-mail address: masoud.tafavvoghi@uit.no (M. Tafavvoghi).

http://dx.doi.org/10.1016/j.jpi.2024.100363
Received 14 September 2023; Received in revised form 24 November 2023; Accepted 23 January 2024
Available online 01 February 2024

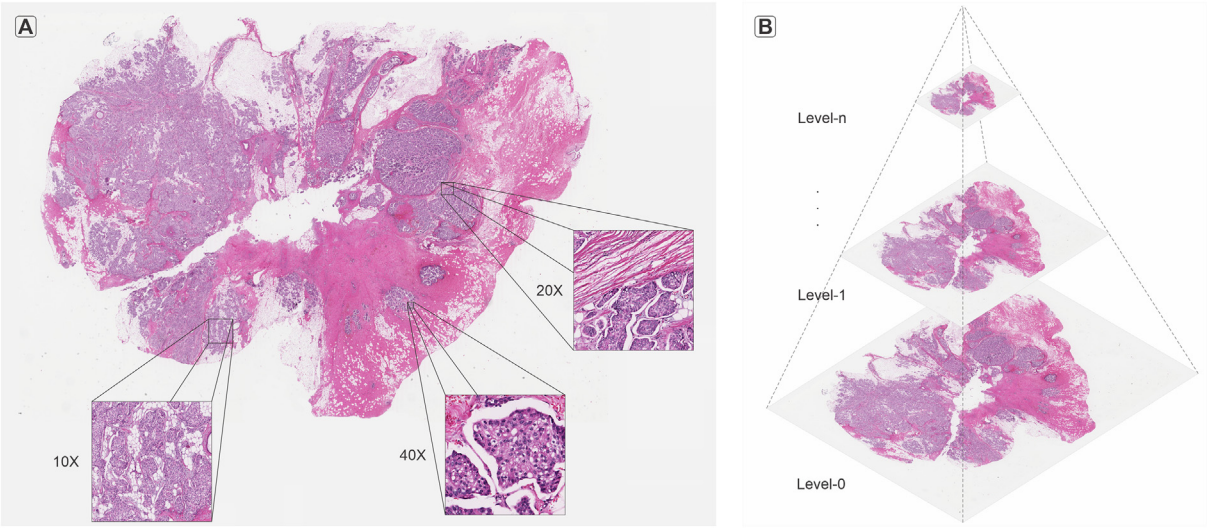

Fig. 1. An example of breast H&E WSI from TCGA-BRCA dataset: (A) Illustration of 3 tile images in 10×, 20×, and 40× magnifications in a WSI; (B) WSI's multi-resolution pyramid with the highest resolution in level-0.

breast cancer recurrence[2] and classify breast cancer subtypes[3] by using histopathological WSIs.

The lack of adequately labeled datasets is a significant limitation in computational pathology for breast cancer, as the performance of deep learning algorithms depends on the availability of sufficient high-quality training and validation data. The use of large and diverse training datasets allows algorithms to identify complex patterns and non-linear relationships more accurately. In addition, using large and independent validation datasets can increase the reliability of models and mitigate overfitting risk, which in turn improves the generalizability of the models.[4]

With advances in technology to produce and utilize data, there is a growing recognition of the benefits of data sharing, so greater openness in scientific research is being advocated for by scientists. The FAIR principles,[5] developed in response to this push for open-access data, provide a framework for making research data more Findable, Accessible, Interoperable, and Reusable. Based on the FAIR principles, medical data should be easily findable by both humans and machines by using a standardized and well-documented approach for metadata and data description and by making use of appropriate metadata standards, taxonomies, and ontologies. Data should be made accessible to all researchers authorized to access it per relevant ethical and legal frameworks. It should be feasible to integrate the data with other datasets and software tools in a seamless manner. Interoperability can be facilitated by using open data formats, data models, and data dictionaries. In addition, data should be designed with the intention of supporting data reuse, allowing other researchers to build upon the data and reproduce the results. To support data reuse, data should be accompanied by complete and accurate documentation, licensing information, and data citation. This approach enriches access to larger and more diverse datasets,[6] enhances faster development of deep learning models,[7] and improves their accuracy and performance,[8] which can consequently lead to better patient outcomes and improved quality of care.[9]

Like other medical data, histopathology images are protected by ethical and legal regulations related to privacy, security, and consent. To share medical data, data owners should adhere to relevant regulatory requirements, such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. After the EU GDPR revised the new regulations in May 2018,[10] medical data sharing has become more challenging due to concerns over maintaining the confidentiality of patient health information, especially in light of high-profile data breaches and incidents of data misuse. In addition to privacy and security concerns, intellectual property and regulatory issues can pose significant barriers to sharing medical data. Resource constraints can also limit the ability of researchers to share medical data. The costs associated with collecting, storing, and curating medical data can be substantial, and many researchers may lack the necessary resources or expertise to manage and share their data effectively.

As the field of big data analysis continues to expand, having access to summaries of existing data has become more advantageous for researchers. Such overviews can assist in identifying relevant datasets without having to begin a new data collection process. Additionally, having a comprehensive and standardized set of public datasets can facilitate the reproducibility and comparability of research findings across different studies. A systematic review of available public datasets can help to identify gaps and limitations in the existing datasets and opportunities for improving the quality and diversity of available datasets. To address this, Hulsen[11] has conducted a systematic review, providing an overview of publicly available patient-centered datasets of prostate cancer presented in imaging, clinical, and genomics categories. He identified 42 publicly available datasets that can efficiently support prostate cancer researchers in selecting appropriate data resources. He found that most datasets do not follow the FAIR principles, as some have legacy issues and need decoding work that might increase the possibility of human error. In Wen et al,[12] the authors have systematically reviewed characteristics of publicly available datasets of skin cancer images, which can be leveraged for the advancement of machine learning algorithms for skin cancer diagnosis. They have reported 21 open-access datasets and 17 open-access atlases available for data extraction. They came to the conclusion that there is inconsistency in reporting image metadata, and population representations are limited in open-access datasets of skin cancer. Leung et al[13] have reviewed the datasets available for machine learning in genomic medicine, including an overview of available omic datasets. They suggested using multiple data sources to rectify problems arising from the missing information from individual datasets.

This scoping review aims to identify and assess the characteristics of all publicly available datasets of breast H&E WSIs to reduce the demand for setting up new studies to collect data for the development of deep learning algorithms. This overview helps to identify potential data sources, the suitability of each dataset for specific tasks in computational pathology, and their quality and biases to ensure the generalizability of machine learning models. To the best of our knowledge, there has not been any study specifically targeting the available datasets of breast H&E WSIs. However, several studies[14–18] have mentioned a small number of such datasets, suggesting the necessity for a comprehensive overview of all available datasets in this field.

Methods

We conducted a scoping review based on the PRISMA-ScR guidelines[19] to identify all publicly available breast H&E WSI datasets appropriate for

deep learning (Supplement 1). Because this scoping review does not evaluate direct health outcomes, it was not eligible to be registered with PROSPERO.[20]

Our inclusion criteria were papers using, reviewing, or mentioning any publicly available dataset of human breast H&E WSIs. These may be introduced in machine learning challenges and contests or published for research purposes. A search was conducted in July 2023 using the following criteria: ("deep learning" OR "machine learning") AND ("whole slide images" OR WSI) AND (breast) AND (histology OR histopathology OR pathology) AND (data OR dataset OR "data set"). In total, nine scientific literature databases were queried: Pubmed, Medline, MDPI, Web of Science, Science Direct, Semantic Scholar, IEEE-Explore, Association for Computing Machinery (ACM) digital library, and the dbpl computer science bibliography. The search for our queries were not limited to only titles and abstracts but across all fields in the search engines, including full-text content and other relevant data fields. To ensure a manageable scope for this review, results were limited to full-text articles in English published between the years 2015 and 2023, and the following exclusion criteria were applied:

1. Not of human breast tissue, for instance, use of histology images of canine or mouse.
2. Using other modalities like CT or MRI instead of histology images.
3. Images of other organs (like lung, skin, etc.) rather than breast.
4. Only patch or image tile datasets instead of WSIs.
5. Non-image data like genomics or clinical data.
6. Tissues not stained with H&E, e.g. immunohistochemistry (IHC) images.
7. Not publicly available datasets.
8. Use of unoriginal or subset datasets derived entirely from public datasets.

Fig. 2 summarizes the workflow diagram of the data collection. Of the 2152 articles from the search results and cross-referencing, 636 articles were removed as duplicates. The remaining 1516 identified articles were then screened by title, abstract, and full-text, respectively, by two independent reviewers (MT and (KM or LAB or NS)) by using the Mendeley reference management software. Subsequently, studies meeting the inclusion criteria were meticulously chosen and annotations were added to indicate the datasets utilized in each selected paper. In the event of disagreement on the inclusion, a third co-author (KM or LB) checked the article to make the final decision on whether it should be included or not. Out of 1516 articles, 756, 198, and 386 were excluded by their title, abstract, and full-text assessment, respectively. This resulted in 176 articles that were included in this review.

In addition to the scientific literature databases, we searched nine online databases and repositories known to contain public datasets. We used search strings in the Supplement 2 to find breast histopathology datasets in The Cancer Imaging Archive, US National Institutes of Health (NIH), Google Data Research, Zenodo, Figshare, Github, Kaggle, Grand Challenge, and the Papers with Code platforms. All the relevant search results that featured breast histopathological images were reviewed by the first

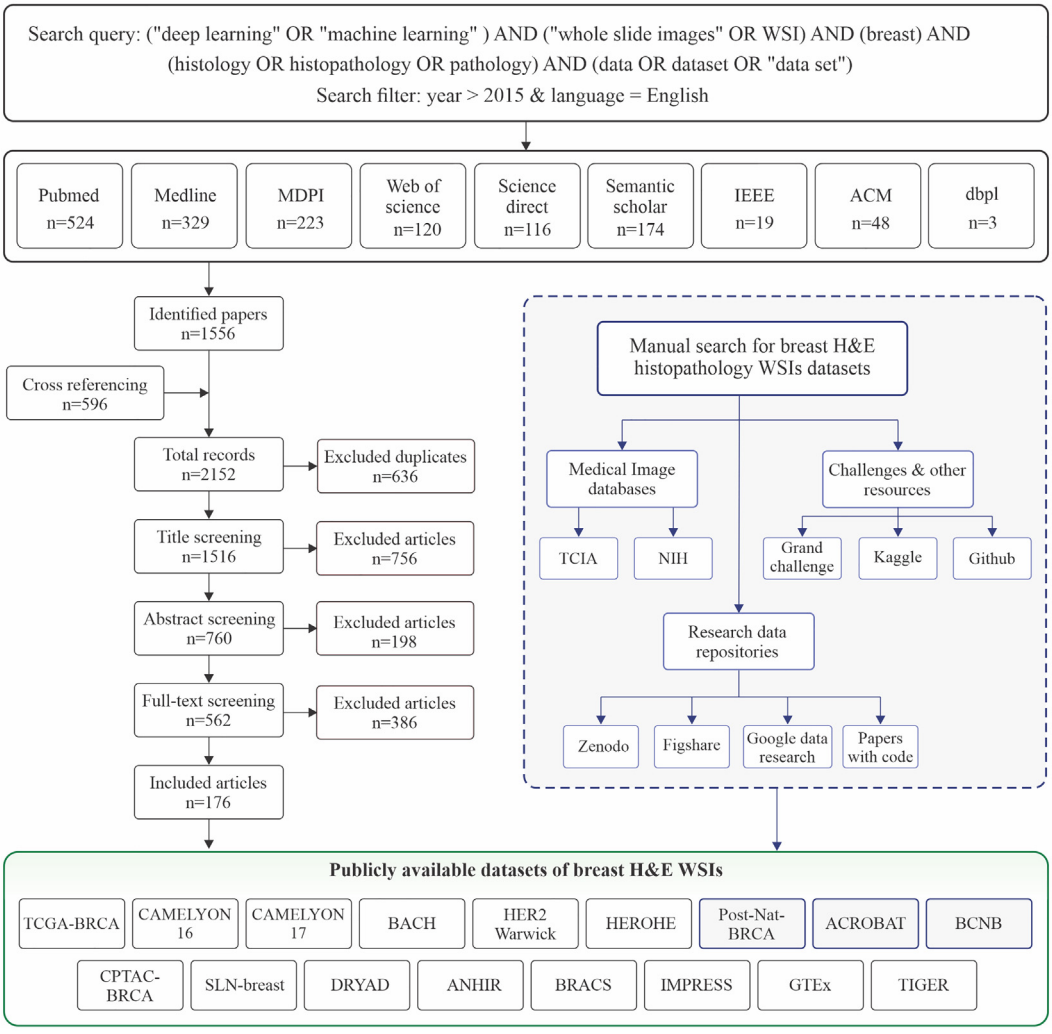

Fig. 2. Data collection workflow. The dashed line box represents the manual search for publicly available datasets of breast H&E WSIs.

Table 1
Publicly available datasets of breast histological H&E WSIs.

Dataset	AKA	Source	Pub. year	WSIs	Patients	Annotation/Labels	Clinical data	Scanner	Pixel size (μm/pixel)	Size (GB)	Image format
ACROBAT	–	Sweden	2023	1153	1153	Landmark pairs for the WSIs in the validation and test sets	–	NanoZoomer X360 and XR	0.91	1164	TIFF
ANHIR	–	Spain	2019	5	–	Coordinates of tumor area	–	Aperio AT2	0.25	0.2	JPG
BACH	ICIAR 2018	Portugal	2018	30	–	10 WSIs have coordinates of ROIs, labeled pixel-wise	–	Leica SCN400	0.50	7	SVS, TIFF
BCNB	–	China	2022	1058	1058	Coordinates of tumor regions	✓	Iscan Coreo	–	33	JPG
BRACS	–	Italy	2020	547	189	Labels for 6 different subtypes	–	Aperio AT2	0.25	1100	SVS
Camelyon16	–	Netherlands	2016	399	399	ROI polygons, pN-stage labels	–	Pannoramic 250, NanoZoomer-XR	0.24	1160	TIFF
Camelyon17	–	Netherlands	2017	1000	200	ROI polygons, pN-stage labels	–	Pannoramic 250, NanoZoomer-XR, Philips IntelliSite	0.24	2950	TIFF
CPTAC-BRCA	CPTAC	USA	2021	642	134	PAM50 molecular subtypes, tumor stage, etc. from the clinical data	✓	–	0.25, 0.50	113	SVS
DRYAD	–	USA	2018	584	–	ROI polygons, binary masks of invasive regions	–	–	–	4.6	PNG
GTEx-breast	–	USA	–	894	894	–	✓	–	0.50	80	SVS
HER2	Warwick	UK	2016	86	86	HER2 score, percentage of cells with complete membrane staining	–	NanoZoomer C9600	0.23	20	SVS
HEROHE	–	Portugal	2020	500	360	Binary labels of HER2 + /- status	–	Pannoramic 1000	0.24	820	MRXS
IMPRESS	–	USA	2023	126	126	Histological subtypes, tumor size, response to therapy, etc.	✓	Hamamatsu	0.50	27	SVS
Post-NAT-BRCA	–	Canada	2021	96	54	Cellularity and cell label, ER, PR, and HER2 scores	✓	Aperio	0.50	43	SVS
SLN-Breast	–	USA	2021	130	78	Binary labels of metastasis status	–	Aperio	0.50	53	SVS
TCGA-BRCA	TCGA	USA	–	3111	1098	Tumor histology and molecular subtypes, given treatment, etc. from the clinical data	✓	–	0.25	1640	SVS
TIGER	–	Netherlands	2022	370	370	ROI polygons, Lymphocyte and Plasma cells indicators, TIL values	–	–	0.50	169	TIFF

reviewer (MT), and their associated metadata and documentation were examined to find other public datasets of breast H&E WSIs that were not included in the selected papers.

Results

The 176 included papers used or reviewed one or more of 14 public datasets of breast H&E WSIs (Table 1). In addition, we manually identified three datasets: the Post-Nat-BRCA dataset in the Cancer Imaging Archive repository and the ACROBAT and BCNB datasets in the Grand Challenge website, none of which were used in any of the selected articles. These 17 publicly available datasets comprise 10,385 breast H&E WSIs appropriate for machine learning use. An additional 89 datasets of histological breast imagery were identified but were not included as they did not fall within the scope of this review. Of these, 32 datasets were tiles (Supplement 3) rather than WSIs, and 57 datasets were privately held or required data use agreements rather than being publicly available (Supplement 4). The corresponding clinical data have been published for only six datasets. Interested readers can find detailed information on the available clinical variables for each dataset in Supplement 5.

There are also three public datasets of breast histopathological WSIs that have acquired all or part of the data from other publicly available datasets: TUPAC16,[21] DRYAD, and TIGER with 821, 195, and 151 WSIs from the TCGA-BRCA, respectively (derived datasets of image tiles can be found in Supplement 3). Therefore, there is a risk of obtaining an overly optimistic performance estimate for trained models if derived datasets are used for the validation. This is because the model has already seen some of the data during training, and using derived datasets for validation may lead to an overestimation of the model's ability to generalize to new, unseen data. However, such derived datasets may be published with extra information not provided in the original datasets. For example, TUPAC16 has 500 WSIs in the training set, all derived from the TCGA-BRCA, with corresponding tumor proliferation and molecular proliferation scores as ground truth which were not included in the original TCGA-BRCA dataset. Including such extra information would be highly advantageous in developing models in breast computational pathology. Table 2 shows details of the derived datasets of breast histopathology WSIs.

Datasets description

ACROBAT dataset[22]: This dataset is part of the CHIME breast cancer study in Sweden, published in the AutomatiC Registration Of Breast cAncer Tissue (ACROBAT) challenge. ACROBAT entails 4212 WSIs from 1153 female primary breast cancer patients, where 1153 and 3059 images are H&E and IHC stained, respectively. The slides are digitized using Hamamatsu NanoZoomer S360 and NanoZoomer XR scanners with 0.23 μm/pixel resolution. However, the published images are in TIFF format with 10× and lower resolutions (pixel size of 0.91 μm) to reduce the dataset size. In addition to the WSIs, the dataset includes annotations of landmark pairs between H&E and IHC images for the validation ($n = 200$) and test ($n = 606$) sets.

ANHIR dataset[23]: This dataset is from the Automatic Non-rigid Histological Image Registration (ANHIR) challenge, which was part of the IEEE International Symposium on Biomedical Imaging (ISBI) 2019. ANHIR contains WSIs of different types of tissue, including breast. Breast WSIs are stained with H&E and IHC and scanned with Leica Biosystems Aperio AT2 with 40× magnitude and 0.253 Âμm/pixel resolution. Images are marked manually with landmarks with standard ImageJ structure and coordinate frame.

BACH[24]: The BreAst Cancer Histology images dataset is from the challenge held as part of the International Conference on Image Analysis and Recognition (ICIAR 2018). The dataset includes H&E-stained WSIs and patches. There are 400 patches with 2048 × 1536 resolution, image-wise labeled in four different classes, along with annotations produced by two medical experts. BACH consists of 30 WSIs, acquired by Leica SCN400 scanner in SVS format, out of which 10 WSIs have coordinates of benign,

Table 2
Derived datasets of breast H&E WSIs.

Dataset	Year	WSIs	Source	Added information	Comments
TUPAC16	2016	821	TCGA	Tumor and molecular proliferation scores	Ground truth is provided for 500 WSIs
DRYAD	2018	195	TCGA	Binary masks for annotated invasive regions in down-sized WSIs	CINJ, CWRU, and HUP WSIs in full-size are not publicly available
		+40	CINJ		
		+110	CWRU		
		+239	HUP		
TIGER- WSIROIS	2022	151	TCGA	Region annotations on WSIs	All 195 WSIs have ROI polygons of different tissue regions and annotations of plasma and lymphocyte cells
		+26	RUMC		
		+18	JB		

in situ carcinoma, and invasive carcinoma regions, labeled pixel-wise by two pathologists.

BCNB[25]: The Early Breast Cancer Core-Needle Biopsy WSI Dataset is the only publicly available dataset of breast histopathological WSIs from Asia. This dataset has 1058 WSIs from 1058 breast cancer patients in China. Images are scanned using an Iscan Coreo pathological scanner, and tumor regions of each image are annotated by two pathologists. Furthermore, the clinical data, including the patient's age, tumor size, histology and molecular subtypes, number of lymph node metastases, and their status of HER2, ER, and PR, is made publicly available alongside the WSIs.

BRACS dataset[26]: BReAst Carcinoma Subtyping dataset is collected at the Istituto Nazionale dei Tumori, Italy using an Aperio AT2 scanner at 0.25 μm/pixel for 40× resolution. BRACS contains 547 WSIs of 189 patients, labeled in 7 classes. Benign tumors are labeled normal, pathological benign, and usual ductal hyperplasia. Atypia tumors are labeled flat epithelial atypia and atypical ductal hyperplasia, and malignant tumors have ductal carcinoma in situ and invasive carcinoma labels. In addition, 4539 regions of interest acquired from 387 WSIs are labeled and provided in .png files.

The Camelyon 16 and 17 datasets[27,28]: The Cancer Metastases in Lymph Nodes Challenge 2016 consists of 399 WSIs of H&E-stained lymph node sections collected in two centers in the Netherlands. Images are annotated with a binary label, and the ground truth for images containing metastases is available in WSI binary masks and plain text files in .xml format, providing the contour vertices of the metastases area. The dataset has 269 images in normal and metastasis classes for training and 130 WSIs for testing. The Camelyon 17 is the extended version of Camelyon 16 comprising 1399 unique H&E-stained WSIs, with an additional 1000 images added to the previous dataset. These 1000 WSIs are collected equally at five medical centers in the Netherlands, each providing 200 images from 40 patients (five slides per patient). In Camelyon 17, images of 100 patients are provided for training, and images of 100 other patients for testing. This dataset has detailed contours of metastasis boundaries on a lesion level for 50 WSIs and pN-stage labels for the patients in training data.

CPTAC-BRCA dataset[29]: The Clinical Proteomic Tumor Analysis Consortium Breast Invasive Carcinoma Collection consists of 642 WSIs of 134 patients, scanned at 20× magnification. The published images have two different resolutions: 0.25 and 0.5 μm/pixel, which can be important when creating image tiles. (Fig. 3). In addition to the slides, clinical, proteomics, and genomic data are available for researchers.

DRYAD dataset[30]: This dataset consists of four different cohorts: The Cancer Genome Atlas (TCGA), Cancer Institute of New Jersey (CINJ), Case Western Reserve University (CWRU), and Hospital at the University of Pennsylvania (HUP) each allotting 195, 40, 110, and 239 WSIs of breast tissue from ER + patients. Slides are scanned by Aperio and Ventana whole-slide scanners at 40× magnification with 0.246 and 0.23 μm pixel width, respectively. Images in the published dataset are down-sized (32:1) WSIs with binary masks for annotated invasive regions. In Cruz-Roa et al, Celik et al, and Ektefaie et al,[31–33] the authors did not mention any use of the DRYAD dataset, but they used the CINJ and HUP datasets that are included in DRYAD. Therefore, papers using these two datasets are included in our study, supposing that they have used part of the DRYAD dataset; as to our knowledge, the CINJ and HUP datasets are not separately available to the public in any database.

GTEx-Breast dataset[34]: The Genotype-Tissue Expression (GTEx) project hosts gene expression levels of 44 human tissues. This project has published 894 breast tissue histology images, consisting of 306 and 588 WSIs of female and male breast tissues dissected from the central breast subareolar region of the right breast. The images are collected from different centers in the USA and have short pathology notes. Additionally, GTEx provides an annotation file with detailed information about the samples.

HER2-Warwick dataset[35]: The data are part of the HER2 scoring contest organized by the University of Warwick, the University of Nottingham, and the Academic–Industrial Collaboration for Digital Pathology consortium. The dataset comprises 86 H&E-stained WSIs of invasive breast carcinomas acquired from 86 patients. IHC-stained images, the ground-truth data in the form of HER2 scores, and the percentage of cells with complete membrane staining are also provided in this dataset.

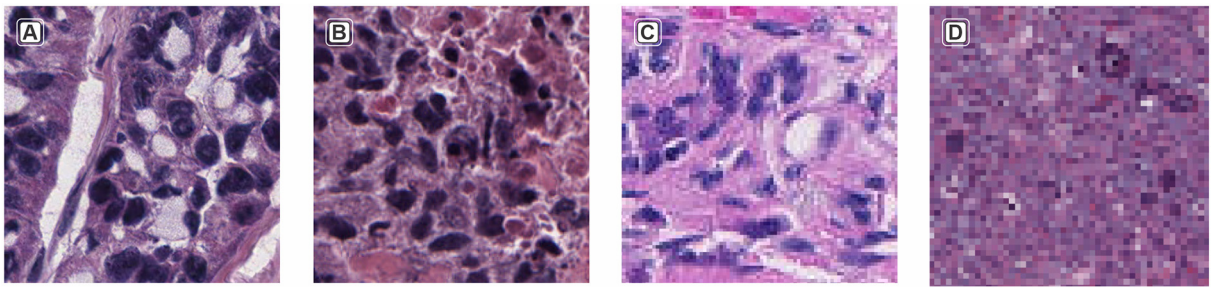

Fig. 3. Comparative resolutions of four image tiles: (A) and (B) are two image tiles cropped from WSIs in the CPTAC-BRCA dataset with pixel widths of 0.25 μm and 0.5 μm, respectively. (C) shows an image tile from the ACROBAT dataset with a larger pixel width of 0.91 μm, resulting in reduced image resolution. (D) displays a tile from a down-sized WSI in the DRYAD dataset with an unknown pixel width, where the image quality is insufficient for cell detection purposes.

HEROHE dataset[36]: This dataset is presented in the HER2 on hematoxylin and eosin (HEROHE) challenge, aimed at predicting HER2 status in breast cancer by using only H&E-stained WSIs. This dataset entails 360 invasive breast cancer cases (144 HER2+ and 216 HER2-) for training and 150 cases (60 HER2+ and 90 HER2-) for testing. The WSIs in training and test sets are from different patients to maintain the independence between the two datasets. WSIs are scanned by 3D Histech Pannoramic 1000 in .mrxs format. Only a binary classification indicating positive or negative HER2 status is available for the HEROHE dataset, and the location of the invasive carcinoma is not annotated.

IMPRESS dataset[37]: This dataset comprised 126 breast H&E WSIs from 62 female patients with HER2-positive breast cancer and 64 female patients diagnosed with triple-negative breast cancer. All participants underwent neoadjuvant chemotherapy followed by surgical excision. In addition to the H&E images, the dataset has IHC stained WSIs of the same slides and their corresponding scores. All the slides are scanned using a Hamamatsu scanner with 20× magnification. The IMPRESS dataset is published with clinical data (cohort metadata) for both patient groups, including patients' age and tumor size, as well as annotations for biomarkers such as PD-L1, CD-8, and CD-163.

Post-NAT-BRCA[38]: The Post-neoadjuvant therapy (NAT) breast cancer dataset is from a cohort with residual invasive breast cancer following NAT. The dataset is composed of 96 WSIs from 54 patients. The slides were scanned by an Aperio scanner at 20× magnification at Sunnybrook Health Sciences Centre in Canada. Clinical data, including patients' age, ER, PR, and HER2 status, is also available with tumor cellularity and cell label annotations.

SLN-Breast[39]: The Breast Metastases to Axillary Lymph Nodes dataset consists of 130 H&E WSIs of axillary lymph nodes from 78 patients, among them 36 WSIs have metastatic breast carcinoma. Slides were scanned with a Lecia Aperio scanner at 20× magnification at Memorial Sloan Cancer Center in the USA. Images are labeled in two classes, positive or negative breast cancer metastases.

TCGA-BRCA[40]: The Cancer Genome Atlas (TCGA) Breast Cancer study is an inclusive, experimental study of breast invasive carcinoma, coordinated and updated regularly by the US National Cancer Institute for research purposes. This dataset entails 3111 H&E-stained WSIs of breast cancer from 1086 female and 12 male patients and is the largest publicly available dataset of breast histopathological WSIs. The TCGA-BRCA includes matched H&E WSIs, gene expression data, and clinical information. We could not find any published region annotations for the WSIs, but there are external sources like cBioPortal that host comprehensive well-organized details of the patients in this dataset.

TIGER[41]: This dataset is released in three formats as the training set for the Tumor InfiltratinG lymphocytes in breast cancER challenge. WSIROIS dataset has 195 WSIs from 195 patients with HER2+ and TNBC breast cancer. Images are collected from three sources: 151 WSIs of TNBC cases from the TCGA-BRCA, 26 WSIs from Radboud University Medical Center (RUMC) with both HER2+ and TNBC cases, and 18 WSIs of HER2+ and TNBC breast cancer cases from Jules Bordet Institut in Belgium. All the published images have 0.5 μm per pixel width and have annotated ROIs indicating seven different tissue regions, as well as 8 × 8 μm^2 bounding boxes indicating lymphocytes and plasma cells. The second dataset, called WSIBULK, consists of 93 WSIs from RUMC and JB with annotation of regions containing invasive tumor cells, and the third dataset, WSITILS, has only TIL values annotation of 82 WSIs without any manual region annotations. Images within WSIROIS, WSIBULK, and WSITILS are unique within each subset, with no duplications across the three subsets.

Datasets descriptive statistics

149[14,31–33,42–186] out of the 176 included papers used public datasets of breast H&E WSIs actively for different algorithm development purposes such as segmentation, classification, prognostic predictions, and color normalization of histology images. The remaining 27 articles[15–18,187–209] have reviewed or mentioned these datasets. These review papers are not included in the subsequent statistical analysis of datasets utilization in this article. Fig. 4A shows the frequency of active use of breast histopathology WSI datasets. Almost half of the studies (52%) have used the TCGA-BRCA

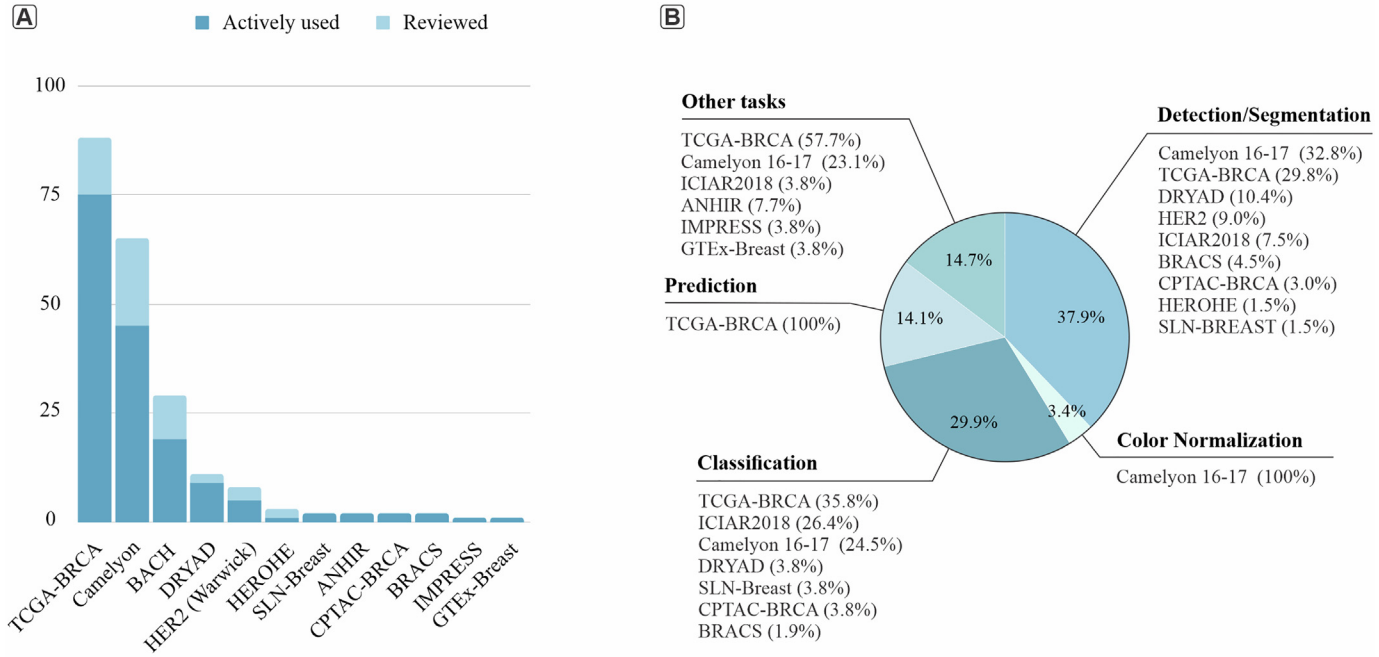

Fig. 4. Publicly available breast H&E WSI datasets usage across selected studies. (A) Usage frequency of breast public datasets in the included articles. The Post-Nat-BRCA, ACROBAT, and BCNB datasets were not used or mentioned in any reviewed articles. Note: In several included papers, multiple datasets are employed. (B) Share of active use of breast histopathology datasets for different tasks. The *Other tasks* category comprises noise elimination, exploration of the tumor immune microenvironment, editing WSIs, feature engineering, investigation of bias in data, staining evaluation, histological grading, gene expression localization, crowdsourcing, obtaining tumor purity maps, scoring nucleoli, making image search and registration tools, and TIL assessment tools.

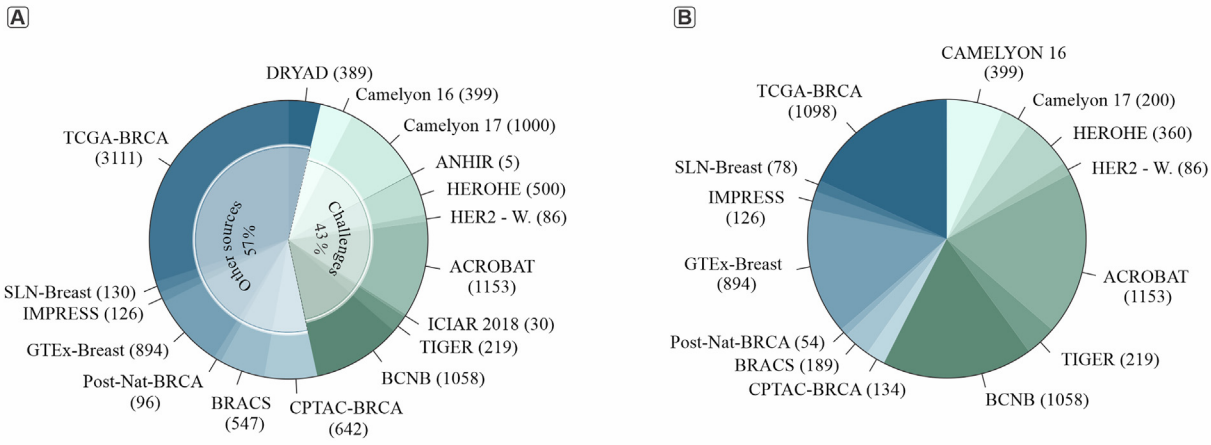

Fig. 5. Distribution of images and patients in publicly available datasets of breast H&E WSIs. (A) Number of WSIs in each dataset and proportion of image sources. The duplicate images in the DRYAD and TIGER datasets derived from the TCGA-BRCA are excluded. (B) Number of patients in each dataset. The number of duplicate patients in the TIGER dataset derived from the TCGA-BRCA ($n = 151$) is not counted, and datasets that do not report the number of breast cancer cases are excluded.

actively that highlights the significance of this dataset in breast computational pathology and its value as a resource for future studies.

The TCGA-BRCA is the only dataset used for developing prediction models using breast WSIs that might be explained by the fact that it has one of the largest cohorts among the publicly available datasets of breast WSIs, and it comes with clinical and genomics data. TCGA-BRCA has the largest contribution in the classification and *Other tasks* categories and is the second most utilized dataset for detection/segmentation tasks, following closely behind the Camelyon dataset. Of note, the Camelyon dataset stands out in the color normalization category as the sole dataset chosen for this particular task (Fig. 4B).

The available ground truth is a limiting factor in using public datasets of breast H&E WSIs for computational pathology, especially when employing supervised algorithms for training the models. Only 43.3% of WSIs have labels for breast cancer subtypes, 33.5% of the images have annotations of regions of interest as ground truth, 11.4% have binary labels of breast cancer metastasis, 5.1% of images are provided by HER2 status labels or scores, and 6.7% of the images do not have any annotations, which restrains the use of these datasets for specific tasks like classification of breast cancer subtypes. Nonetheless, the available WSIs can be utilized for training self-supervised and semi-supervised models.

Camelyon 16 and 17, HEROHE, ICIAR 2018, HER2-Warwick, ANHIR, ACROBAT, BCNB, and TIGER datasets are published in challenges and contests and comprise 43% of the publicly available breast H&E WSIs. The other eight datasets: TCGA-BRCA, SLN-Breast, Post-Nat-BRCA, BRACS, DRYAD, GTEx-Breast, IMPRESS, and CPTAC-BRCA are collected and made available for research purposes (Fig. 5A). One intriguing aspect of the identified datasets is the number of patients included, which varies considerably between studies and may have important implications for the generalizability and reliability of trained models. The TCGA-BRCA is the largest open-access dataset of breast H&E WSIs with 3311 images from 1098 patients (Fig. 5B), that is extended regularly by adding new slides to the dataset.

Discussion

The present study aimed to investigate the availability and suitability of open-access histopathology datasets for the development of deep learning algorithms in breast tissue analysis. In pursuit of this, we identified 17 publicly available datasets of breast H&E WSIs that may appear to be a substantial amount for developing deep learning algorithms. However, it is important to note that the publicly available datasets of breast H&E often lack detailed metadata descriptions (Fig. 6). For instance, the number of patients is not reported in three datasets. Furthermore, the clinical data necessary for the development of prognostic tools are only available for six datasets: BCNB, Post-Nat-BRCA, TCGA-BRCA, GTEx-Breast, CPTAC-BRCA, and

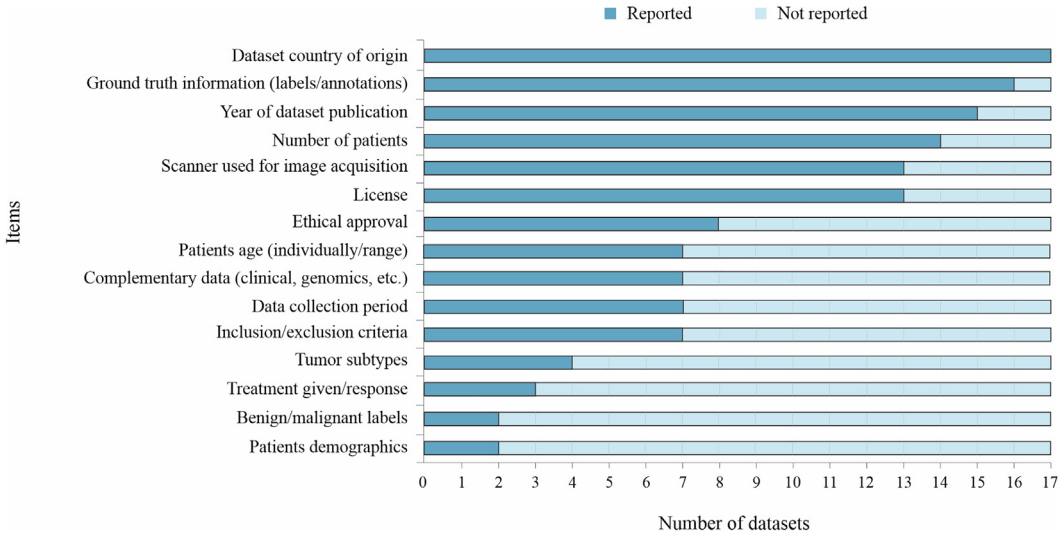

Fig. 6. Report of public breast H&E WSI datasets' characteristics and metadata.

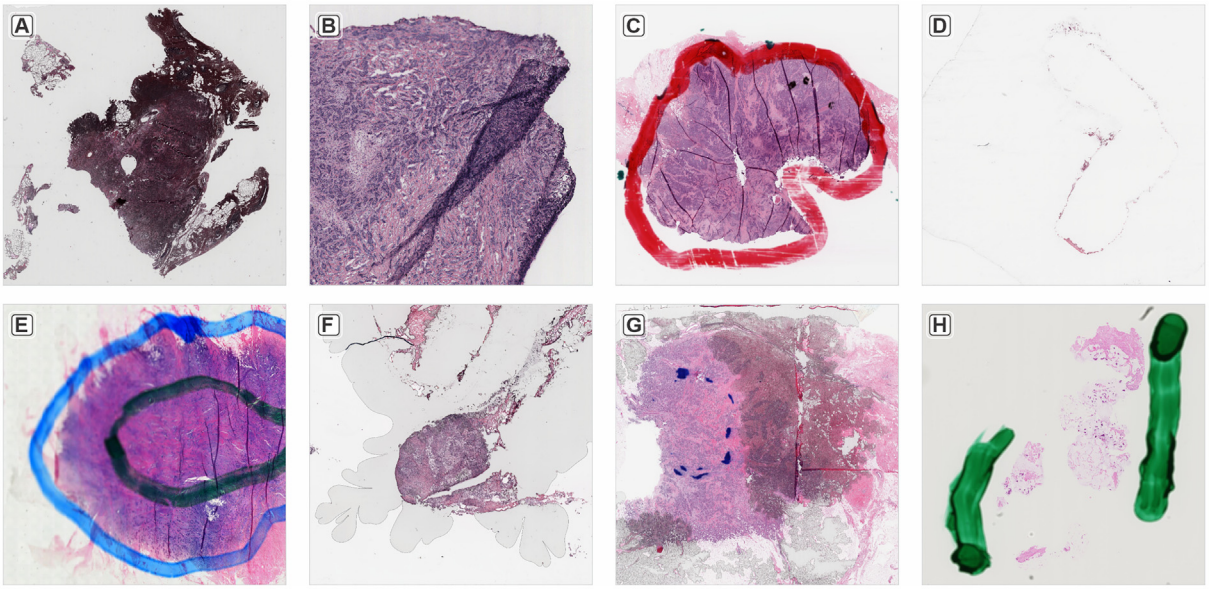

Fig. 7. Examples of artifacts and quality issues in breast histopathological WSIs. (A) A WSI from the TCGA-BRCA with inconsistent staining. (B) Tissue folding in a WSI in the CPTAC-BRCA. (C) Pathologist's marker sign on a WSI from the TCGA-BRCA with several tissue foldings. (D) A slide from the CPTAC-BRCA exhibiting predominantly vacant areas. (E) A blurred WSI with marker signs from a private dataset. (F) Presence of air bubbles in a WSI of CPTAC-BRCA leading to focal image disruptions. (G) An image from the TCGA-BRCA with Marker signs, air bubbles, and inconsistent colors in some regions. (H) A WSI from the HER2-Warwick dataset showing marker signs on the slide.

IMPRESS, with the latter four being collected in the USA. TCGA-BRCA is the sole dataset published before 2021, which could explain its exclusive utilization for developing predictive models in the articles included in this study. None of the included papers or web pages hosting the breast H&E WSI datasets provided an explicit statement of adherence to the FAIR principles, and the level of metadata and documentation provided by the dataset publishers varied. This variability in metadata and documentation could potentially affect the findability and reusability of the datasets, highlighting the need for improved adherence to the FAIR principles to enhance their accessibility and usability. Furthermore, inconsistencies in the data format and structure were found across the available datasets, which could limit the interoperability of the datasets.

The development of deep learning algorithms for breast computational pathology could be confounded by the quality of available WSIs. Variations in staining, tissue preparation, and image scanning processes can result in artifacts WSIs, affecting the integrity of the data. Additionally, marker signs or annotations on the slides may inadvertently introduce noise or bias into the data. Therefore, accurate recognition and mitigation of such artifacts are vital to maintaining image-based analyses' fidelity. Figure 7 shows examples of such artifacts and marker signs on the images.

The WSIs of the TCGA-BRCA and the Camelyon datasets are widely employed in breast computational pathology. This widespread usage has implications for the generalizability of machine learning algorithms. As the TCGA dataset is collected in the USA, it may not be representative of the breast cancer population in other regions or countries. In addition, TCGA-BRCA has a high proportion of white women compared to American Indian and Hispanic patients (Fig. 8A) and a high proportion of patients with infiltrating duct carcinoma (70%). Additionally, this dataset includes a large percentage of samples from younger women, which may

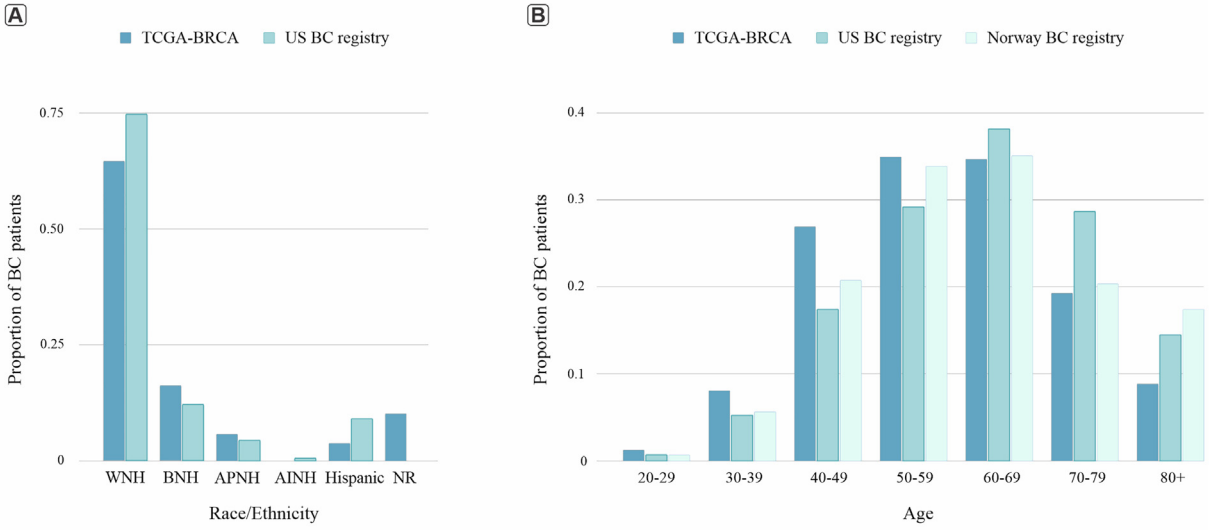

Fig. 8. Distribution of age and ethnicity among patients in the TCGA dataset. (A) Distribution of breast cancer cases in different ethnicities. WNH = White Non-Hispanic. BNH = Black Non-Hispanic. APNH = Asian and Pacific Islander Non-Hispanic. AINH = American Indian and Alaska Native Non-Hispanic. NR = Not Reported. (B) Comparison of breast cancer cases in the TCGA-BRCA dataset, the US breast cancer registry[212] and the breast cancer registry of Norway.[213]

not represent the entire breast cancer population as the disease is more common in older women (Fig. 8B). The composition of patient demographics is not reported in the Camelyon dataset, which limits the ability to analyze the potential impact of demographic diversity on the developed algorithms and findings.

The issue of biases is not exclusive to the TCGA-BRCA dataset; it has also been observed in other studies. A study on the representativeness of the TCGA bladder cancer cohort[210] revealed biases in this dataset. The authors found that patients captured in the TCGA-BCa cohort demonstrate a higher risk disease profile compared to the reference cystectomy series, and consequently, their rates of overall survival and disease-specific survival are lower. Another study[211] found that Black Americans are not adequately represented in the majority of cancer cases within the TCGA datasets compared to clinical and mortality datasets. They also stated that Asian Americans are overrepresented in the TCGA dataset for most cancers. These biases are significant factors that should be acknowledged during the validation of computer-assisted tools, playing a vital role in maintaining the models' robustness, applicability, and transferability across different cohorts of breast cancer patients.

Deep learning models can benefit from external datasets to improve their ability to generalize to new data and enhance their performance on a specific task. Using external data to validate trained algorithms can help ensure their generalizability, identifying overfitting, and assessing their performance across different datasets. Therefore, the real measure of a model's predictive ability lies in its performance on an independent dataset that was not employed in its initial development,[214] as the performance of these models often diminishes when applied to a new cohort beyond the original development population.[215] Typically, there is a lack of external validation in the algorithms developed for breast computational pathology. Among the 149 non-review papers, only 41 incorporated multiple datasets; of those, 21 integrated private datasets alongside public ones for model development or validation. Notably, only 21 studies used an external validation/test set (Table 3) implying that the performance of other developed models could be subject to overestimation.

The potential benefits of using private datasets can make it well worth the effort to get access to such datasets. Incorporating private datasets in both the training and validation processes can improve machine learning models' diversity, representativeness, and overall performance. Private datasets can also contain unique or hard-to-obtain data that might not be available through public sources. For example, WSIs acquired from patients who have taken specific treatments like immunotherapy and the response to this treatment does not exist in any publicly available datasets of breast WSIs.

One potential limitation in this scoping review is the possibility of missing relevant datasets due to the search strategy or selection criteria used. To

Table 3
Utilization of multiple breast H&E WSI datasets in the included articles.

Study	Aim	Development dataset	External validation set
33	BC Classification	TCGA, PD	DRYAD (HUP, CINJ)
42	Region detection and classification	ICIAR, DRYAD	TCGA
44	Detection and overexpression of HER2	HER2-W.	TCGA
46	Metastasis detection	Cam17	PD
53	Treatment prediction	TCGA	PD
55	Segmentation and classification	TCGA, Cam16, Cam17, TUPAC, SLN	-
62	Scoring nucleoli in invasive BC	TCGA	PD
64	Editing WSIs with GANs	Cam16, Cam17	-
67	Metastasis detection	Cam16, Cam17	-
73	Staining evaluation	TCGA, ICIAR, TUPAC	Cam17
77	Classification of invasive ductal carcinoma	PD	TCGA
78	Segmentation of WSIs	Cam16, Cam17	-
90	Detection of fiber orientation disorder	TCGA	PD
96	Detection and classification	TCGA, CPTAC, Cam16, Cam17	PD
104	Detection of estrogen receptor status	TCGA, PD	-
106	Detection of HER2 status	HER2	TCGA
109	Making an image search tool	TCGA, PD	-
112	Color normalization and classification	Cam17	Cam16
121	Segmentation	Cam16, ICIAR	-
126	Prediction of patient staging and node status	TCGA, TUPAC	-
128	Detection of BC	Cam16 and 17	-
130	Making a TILS assessment tool	PD	TCGA
131	Color normalization	Cam16, TUPAC	-
132	Making a prognostic tool	TCGA, PD	-
135	Prediction of DNA repair deficiency	TCGA	PD
139	Metastasis detection	Cam16, Cam17	-
140	Histological grading	TCGA	PD
141	Prediction of molecular phenotypes	TCGA	PD
153	Color normalization	Cam16, Cam17	-
154	Segmentation	TCGA, ICIAR, DRYAD	-
156	Segmentation	BRACS, PD	-
157	Metastasis detection	TCGA, BRACS	-
158	Color normalization	Cam16, Cam17	-
159	Spatial characterization of TILs	TCGA, PD	-
161	Prediction of response to NAC	TCGA	IMPRESS
163	Testing the generalizability of a model	Cam16, Cam17, PD	-
167	Prediction of DNA-repair deficiency	PD	TCGA
168	Prediction of BC recurrence	PD	TCGA
169	Detection of HER2 status	TCGA	HER2-W., PD
175	BC classification	ICIAR, Cam16	-
186	HER2 detection and response prediction	PD	TCGA

Studies utilizing non-breast datasets, TMAs, and image tiles along with the breast WSIs for training, validation, or as an external test set are excluded from this table. BC = Breast Cancer. PD = Private Data. Cam = Camelyon. NAC = Neoadjuvant Chemotherapy. TMAs = Tissue Microarrays.

mitigate this limitation, a comprehensive and well-defined search strategy was developed to ensure that all relevant datasets were captured. Nevertheless, the Post-Nat-BRCA, ACROBAT, and BCNB datasets were not found in any of the papers identified in our selected literature databases, and we found them during the manual screening of research data repositories. Another limitation is the use of only English language search terms and inclusion criteria. This approach may have resulted in the exclusion of relevant datasets that were unavailable in English or primarily in other languages. Future systematic or scoping reviews of publicly available datasets of breast H&E WSIs could benefit from broader search strategies that include searches in multiple languages. This would increase the likelihood of identifying relevant datasets that are not primarily in English and thus reduce potential language-related bias. However, the feasibility of such an approach will depend on the availability of resources, expertise in multiple languages, and the research question being addressed. Additionally, changes in the availability of datasets over time may further restrict the relevance and applicability of such reviews of publicly available datasets. The review should be conducted within a well-defined time frame to address this limitation, and the publication date of included studies should be clearly reported. In summary, our study examined the availability and suitability of publicly available datasets of H&E-stained histopathology WSIs that can be used in breast computational pathology. This data overview can save significant time and effort by providing a starting point without the need for setting up a data collection study.

Despite the significant number of WSIs, we found limitations in metadata descriptions, inadequate clinical data, and inconsistencies in the format and structure of datasets. Additionally, the presence of biases within widely used datasets, such as TCGA-BRCA, raises concerns regarding the generalizability of models. Therefore, it is crucial to improve adherence to FAIR principles, enhance metadata descriptions, and address biases. Moreover, incorporating diverse datasets, including private and external sources, promises to improve model performance and generalizability.

Declaration of competing interest

All authors declare they have no conflicts of interest.

Acknowledgements

The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2024.100363.

References

1. Liu L, Feng W, Chen C, Liu M, Qu Y, Yang J. Classification of breast cancer histology images using MSMV-PFENet. Scient Rep 2022;12(1):17447.
2. Yang J, Ju J, Guo L, et al. Prediction of her2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. Computat Struct Biotechnol J 2022;20:333–342.
3. Srikantamurthy MM, Rallabandi VP, Dudekula DB, Natarajan S, Park J. Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid cnn-lstm based transfer learning. BMC Medical Imaging 2023;23(1):1-15.
4. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv: 1207.0580; 2012.
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The fair guiding principles for scientific data management and stewardship. Sci Data 2016;3, 160018.
6. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124(7):962–969.
7. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. . J R Soc Interface 2018;15(141). 20170387.
8. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New Engl J Med 2019;380(14):1347–1358.
9. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Informatics 2017;22(5):1589–1604.
10. BA Simell, OM Törnwall, I Hämäläinen, H-E Wichmann, G Anton, P Brennan, L Bouvard, N Slimani, A Moskal, M Gunter, et al. Transnational access to large prospective cohorts in Europe: Current trends and unmet needs. New Biotechnol, 49:98–103, 2019.
11. Hulsen T. An overview of publicly available patient-centered prostate cancer datasets. Translat Androl Urol 2019;8(suppl 1):S64.
12. Wen D, Khan SM, Xu AJ, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digital Health 2022;4(1):e64–e74.
13. Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a reviewof computational problems and data sets. Proc IEEE 2015;104(1):176–197.
14. Brancati N, Anniciello AM, Pati P, et al. Bracs: a dataset for breast carcinoma subtyping in H&E histology images. Database 2022;2022:baac093.
15. Zeiser FA, da Costa CA, Roehe AV, da Rosa Righi R, Marques NMC. Breast cancer intelligent analysis of histopathological data: a systematic review. Appl Soft Comput 2021;113, 107886.
16. Duggento A, Conti A, Mauriello A, Guerrisi M, Toschi N. Deep computational pathology in breast cancer. Seminars in Cancer Biology. Elsevier; 2021. p. 226–237.
17. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology, an overview and future trends. Med Image Anal 2018;47:45–67.
18. Liew XY, Hameed N, Clos J. A review of computer-aided expert systems for breast cancer diagnosis. Cancers 2021;13(11):2764.
19. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467–473.
20. Booth A, Clarke M, Dooley G, et al. The nuts and bolts of prospero: an international prospective register of systematic reviews. Syst Rev 2012;1(1):1–9.
21. Tumor Proliferation Assessment Challenge. Grand Challenge. https://tupac.grand-challenge.org/ 2016. Accessed 07 Jul. 2023.
22. AutomatiC Registration Of Breast cAncer Tissue (ACROBAT). Grand Challenge. https://acrobat.grand-challenge.org/ 2023. Accessed 07 Jul. 2023.
23. Automatic Non-rigid Histological Image Registration (ANHIR). Grand Challenge. https://anhir.grand-challenge.org/ 2019. Accessed 07 Jul. 2023.
24. ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. Grand Challenge. https://iciar2018-challenge.grand-challenge.org/ 2018. Accessed 07 Jul. 2023.
25. Early Breast Cancer Core-Needle Biopsy WSI (BCNB). Grand Challenge. https://bcnb.grand-challenge.org/ 2022. Accessed 07 Jul. 2023.
26. BRACS: BReAst Carcinoma Subtyping. Institute of High-Performance Computing and Networking. 2020 https://www.bracs.icar.cnr.it/. Accessed 07 Jul. 2023.
27. CAMELYON16. Grand Challenge. https://camelyon16.grand-challenge.org/ 2016. Accessed 07 Jul. 2023.
28. CAMELYON17. Grand Challenge. https://camelyon17.grand-challenge.org/ 2017. Accessed 07 Jul. 2023.
29. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium. The Clinical Proteomic Tumor Analysis Consortium Breast Invasive Carcinoma Collection (CPTAC-BRCA). The Cancer Imaging Archive; 2020. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70227748 2020. Accessed 07 Jul. 2023.
30. Cruz-Roa A, Gilmore H, Basavanhally A, et al. High-throughput adaptive sampling for . whole-slide histopathology image analysis (hashi) via convolutional neural networks: application to invasive breast cancer detection. PloS One 2018;13(5):e0196828. https://doi.org/10.5061/dryad.1g2nt41. Accessed 07 Jul. 2023 via.
31. Cruz-Roa A, Basavanhally A, González F, et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Imaging 2014: Digital Pathology. SPIE; 2014. p. 904103.
32. Celik Y, Talo M, Yildirim O, Karabatak M, Rajendra Acharya U. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. Pattern Recognit Lett 2020;133:232–239.
33. Ektefaie Y, Yuan W, Dillon DA, et al. Integrative multiomics-histopathology analysis for breast cancer classification. NPJ Breast Cancer 2021;7(1):147.
34. The Genotype-Tissue Expression (GTEx). gtex portal. https://gtexportal.org/home/histologyPage. Accessed 07 Jul. 2023.
35. Her2 Scoring Contest. Tissue Image Analytics (TIA) Centre. https://warwick.ac.uk/fac/cross_fac/tia/data/her2contest/ 2016. Accessed 07 Jul. 2023.
36. HEROHE. Grand Challenge. https://ecdp2020.grand-challenge.org/ 2002. Accessed 07 Jul. 2023.
37. Huang Z, Shao W, Han Z, et al. Artificial intelligen .ce reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. NPJ Precision Oncol 2023;7(1):14. Accessed 07 Jul. 2023 via: https://drive.google.com/drive/folders/1fNf-F_aplm6ACJTWO1vGqbb-DdaP4K_r.
38. Martel AL, Nofech-Mozes S, Salama S, Akbar S, Peikari M. Assessment of Residual Breast Cancer Cellularity after Neoadjuvant Chemotherapy using Digital Pathology [Data set]. Cancer Imaging Arch 2019. https://doi.org/10.7937/TCIA.2019.4YIBTJNO.
39. Campanella G, Hanna MG, Brogi E, Fuchs TJ. Breast metastases to axillary lymph nodes. Cancer Imaging Archive; 2019.
40. The Cancer Genome Atlas (TCGA). Genomic Data Commons Data Portal (GDC). https://portal.gdc.cancer.gov/projects/TCGA-BRCA. Accessed 07 Jul. 2023.
41. Tumor InfiltratinG lymphocytes in breast cancER. Grand Challenge. 2022. https://tiger.grand-challenge.org/Home/. Accessed 07 Jul. 2023.
42. Ahmed S, Tariq M, Naveed H. Pmnet: a probability map based scaled network for breast cancer diagnosis. Comput Med Imaging Graphics 2021;89, 101863.
43. Amgad M, Elfandy H, Hussein H, et al. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics 2019;35(18):3461–3467.

44. Anand D, Kurian NC, Dhage S, et al. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. J Pathol Inform 2020;11(1):19.
45. Aresta G, Araújo T, Kwok S, et al. Bach: grand challenge on breast cancer histology images. Med Image Anal 2019;56:122–139.
46. Bandi P, Geessink O, Manson Q, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Trans Med Imaging 2018;38(2):550–560.
47. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 2017;318(22):2199–2210.
48. Bokor P, Hudec L, Fabian O, Benesova W. Weighted multi-level deep learning analysis and framework for processing breast cancer WSIs. arXiv preprint arXiv:2106.14708; 2021.
49. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8): 1301–1309.
50. Çelik G, Talu MF. Resizing and cleaning of histopathological images using generative adversarial networks. Phys AStat Mech Its Appl 2020;554, 122652.
51. Chaudhury S, Shelke N, Sau K, Prasanalakshmi B, Shabaz M. A novel approach to classifying breast cancer histopathology biopsy images using bilateral knowledge distillation and label smoothing regularization. Computat Math Methods Med 2021;2021: 1-11.
52. Chen J, Jiao J, He S, Han G, Qin J. Few-shot breast cancer metastases classification via unsupervised cell ranking. IEEE/ACM Trans Computat Biol Bioinform 2019;18(5): 1914–1923.
53. Cho SY, Lee JH, Ryu JM, et al. Deep learning from HE slides predicts the clinical benefit from adjuvant chemotherapy in hormone receptor-positive breast cancer patients. Scient Rep 2021;11(1):17363.
54. Ciga O, Martel AL. Learning to segment images with classification labels. Med Image Anal 2021;68, 101912.
55. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. Mach Learn Appl 2022;7, 100198.
56. Cruz-Roa A, Gilmore H, Basavanhally A, et al. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: application to invasive breast cancer detection. PLoS One 2018;13(5), e0196828.
57. Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Scient Rep 2017;7(1):46450.
58. De Matos J, Ataky STM, de Souza Britto Jr A, de Oliveira LES, Koerich AL. Machine learning methods for histopathological image analysis: a review. Electronics 2021;10 (5):562.
59. Dhillon A, Singh A. eBreCaP: extreme learning-based model for breast cancer survival prediction. IET Syst Biol 2020;14(3):160–169.
60. Diao JA, Wang JK, Chui WF, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nat Commun 2021;12(1):1613.
61. Eddy JA, Thorsson V, Lamb AE, et al. Cri iatlas: an interactive portal for immuno-oncology research. F1000Research 2020;9.
62. Elsharawy KA, Toss MS, Raafat S, et al. Prognostic significance of nucleolar assessment in invasive breast cancer. Histopathology 2020;76(5):671–684.
63. Elsharawy KA, Gerds TA, Rakha EA, Dalton LW. Artificial intelligence grading of breast cancer: a promising method to refine prognostic classification for management precision. Histopathology 2021;79(2):187–199.
64. Blanco RF, Rosado P, Vegas E, Reverter F. Medical image editing in the latent space of generative adversarial networks. Intel-Based Med 2021;5, 100040.
65. Milagro Fernández-Carrobles M, Serrano I, Bueno G, Déniz O. Bagging tree classifier and texture features for tumor identification in histological images. Proc Comput Sci 2016;90:99-106.
66. Fischer W, Moudgalya SS, Cohn JD, Nguyen NTT, Kenyon GT. Sparse coding of pathology slides compared to transfer learning with deep neural networks. BMC Bioinform 2018;19:9-17.
67. Zanjani FG, Zinger S, Piepers B, Mahmoudpour S, Schelkens P, de With PHN. Impact of jpeg 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images. J Med Imaging 2019;6(2):027501.
68. Graham S, Vu QD, Raza SEA, et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal 2019;58, 101563.
69. Guo Z, Liu H, Ni H, et al. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. Scient Rep 2019;9(1):882.
70. He B, Bergenstråhle L, Stenbeck L, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng 2020;4(8):827–834.
71. Hegde N, Hipp JD, Liu Y, et al. Similar image search for histopathology: Smily. NPJ Digit Med 2019;2(1):56.
72. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun 2021;12(1):4423.
73. Choudhary A, Wu H, Tong L, Wang MD. Learning to evaluate color similarity for histopathology images using triplet networks. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2019. p. 466–474.
74. Jaber MI, Song B, Taylor C, et al. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. Breast Cancer Res 2020;22:1-10.
75. Jiao Y, Yuan J, Qiang Y, Fei S. Deep embeddings and logistic regression for rapid active learning in histopathological images. Comput Methods Prog Biomed 2021;212, 106464.
76. Kalra S, Tizhoosh HR, Shah S, et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. NPJ Digit Med 2020;3(1):31.
77. Kanavati F, Tsuneki M. Breast invasive ductal carcinoma classification on whole slide images with weakly-supervised and transfer learning. Cancers 2021;13(21):5368.
78. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B. A generalized deep learning framework for whole-slide image segmentation and analysis. Scient Rep 2021;11(1):11579.
79. Kim Y-G, Kim S, Cho CE, et al. Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections. Scient Rep 2020;10(1):21899.
80. Krithiga R, Geetha P. Deep learning based breast cancer detection and classification using fuzzy merging techniques. Mach Vision Appl 2020;31:1-18.
81. Kumar A, Prateek M. Localization of nuclei in breast cancer using whole slide imaging system supported by morphological features and shape formulas. Cancer Manage Res 2020:4573–4583.
82. Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans Med Imaging 2017;36(7):1550–1560.
83. La Barbera D, Polónia A, Roitero K, Conde-Sousa E, Mea VD. Detection of her2 from haematoxylineosin slides through a cascade of deep learning classifiers via multi-instance learning. J Imaging 2020;6(9):82.
84. Le H, Gupta R, Hou L, et al. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. Am J Pathol 2020;190(7):1491–1504.
85. Lee S, Amgad M, Mobadersany P, et al. Interactive classification of whole-slide imaging data for cancer researchers. Cancer Res 2021;81(4):1171–1177.
86. Lei G, Xia Y, Zhai D-H, et al. Neurocomputing 2020;406:267–273.
87. Levy-Jurgenson A, Tekpli X, Kristensen VN, Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Scient Rep 2020;10(1):18802.
88. Li C, Lu X. Computer-aided detection breast cancer in whole slide image. 2021 International Conference on Computer, Control and Robotics (ICCCR). IEEE; 2021. p. 193–198.
89. Li H, Qiu L, Wang M. Informed attentive predictors: a generalisable architecture for prior knowledge-based assisted diagnosis of cancers. Sensors 2021;21(19):6484.
90. Li H, Bera K, Toro P, et al. Collagen fiber orientation disorder from h&e images is prognostic for early stage breast cancer: clinical trial validation. NPJ Breast Cancer 2021;7 (1):104.
91. Lin H, Chen H, Graham S, Dou Q, Rajpoot N, Heng P-A. Fast scannet: fast and dense analysis of multigigapixel whole-slide images for cancer metastasis detection. IEEE Trans Med Imaging 2019;38(8):1948–1958.
92. Litjens G, Bandi P, Bejnordi BE, et al. 1399 H&E-stained sentinel lymph node sections . of breast cancer patients: the CAMELYON dataset. GigaScience 05, 2018;7(6):giy065. https://doi.org/10.1093/gigascience/giy065. ISSN 2047-217X.
93. Liu Y, Kohlberger T, Norouzi M, et al. Artificial intelligence–based breast cancer nodal metast .asis detection: insights into the black box for pathologists. Arch Pathol Lab Med 10, 2018;143(7):859–868. https://doi.org/10.5858/arpa.2018-0147-OA. ISSN 0003-9985.
94. López-Pérez M, Amgad M, Morales-Álvarez P, et al. Learning from crowds in digital pathology using scalable variational gaussian processes. Scient Rep 06, 2021;11(1): 11612. https://doi.org/10.1038/s41598-021-90821-3. ISSN 2045-2322.
95. Lu MY, Chen RJ, Kong D, et al. Federated learning for computational pathology on gigapixel whole slide images. Med Image Anal 2022;76:102298. https://doi.org/10.1016/j.media.2021.102298. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003431.
96. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervis .ed computational pathology on whole-slide images. Nat Biomed Eng 06, 2021;5(6):555–570. https://doi.org/10.1038/s41551-020-00682-w. ISSN 2157-846X.
97. Lu Z, Zhan X, Wu Y, et al. Brcaseg: a deep learning approach for tissue quantification and genomic correlations of histopathological images. Genom Proteom Bioinform 2021;19(6):1032–1042. https://doi.org/10.1016/j.gpb.2020.06.026. ISSN 1672-0229. URL: https://www.sciencedirect.com/science/article/pii/S1672022921001522.
98. Mi W, Li J, Guo Y, et al. Deep learning-based multi-class classification of breast digital pathology images. Cancer Manage Res 2021;13:4605–4617. https://doi.org/10.2147/CMAR.S312608. PMID: 34140807.
99. Monjo T, Koido M, Nagasawa S, Suzuki Y, Kamatani Y. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. Scient Rep 03, 2022;12(1):4133. https://doi.org/10.1038/s41598-022-07685-4. ISSN 2045-2322.
100. Mukundan R. Analysis of image feature characteristics for automated scoring of her2 in histology slides. J Imaging 2019;5(3). https://doi.org/10.3390/jimaging5030035. ISSN 2313-433X. URL: https://www.mdpi.com/2313-433X/5/3/35.
101. Mukundan R. Image features based on characteristic curves and local binary patterns for automated her2 scoring. J Imaging 2018;4(2). https://doi.org/10.3390/jimaging4020035. ISSN 2313-433X. URL: https://www.mdpi.com/2313-433X/4/2/35.
102. Munien C, Viriri S, Rakhshan V. Classification of hematoxylin and eosin-stained breast cancer histology microscopy images using transfer learning with efficientnets. Computat Intel Neurosci 2021:5580914. https://doi.org/10.1155/2021/5580914. ISSN 1687-5265.
103. Muñoz-Aguirre M, Ntasis VF, Rojas S, Guigó R. Pyhist: A histol:ogical image segmentation tool. PLoS Computat Biol 2020;16(10):e1008349. https://doi.org/10.1371/journal.pcbi.1008349.
104. Naik N, Madani A, Esteva A, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat Commun 11, 2020;11(1): 5727. https://doi.org/10.1038/s41467-020-19334-3. ISSN 2041-1723.

105. Noorbakhsh J, Farahmand S, pour AF, et al. Deep learning-based cross-classifications r . eveal conserved spatial behaviors within tumor histological images. Nat Commun 12, 2020;11(1):6367. https://doi.org/10.1038/s41467-020-20030-5. ISSN 2041-1723.
106. Oliveira SP, Pinto JR, Gonçalves T, et al. Weakly-supervised classification of her2 expression in breast cancer haematoxylin and eosin stained slides. Appl Sci 2020;10 (14). https://doi.org/10.3390/app10144728. ISSN 2076-3417. URL: https://www.mdpi.com/2076-3417/10/14/4728.
107. Oner MU, Chen J, Revkov E, et al. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. Patterns (New York, NY) 12, 2021;3(2):100399. https://doi.org/10.1016/j.patter.2021.100399.
108. Öztürk Ş, Akdemir B. Hic-net: a deep convolutional neural network model for classification of histopathological breast images. Comput Elect Eng 2019;76:299–310. https://doi.org/10.1016/j.compeleceng.2019.04.012. ISSN 0045-7906. URL: https://www.sciencedirect.com/science/article/pii/S0045790618320007.
109. Pantanowitz L, Michelow P, Hazelhurst S, et al. A digital pathology solution to resolve the tissue floater conundrum. Arch Pathol Lab Med 07, 2020;145(3):359–364. https://doi.org/10.5858/arpa.2020-0034-OA. ISSN 0003-9985.
110. Park J, Chung YR, Kong ST, et al. Aggregation of cohorts for histopathological diagnosis with deep morphological analysis. Scient Rep 02, 2021;11(1):2876. https://doi.org/10.1038/s41598-021-82642-1. ISSN 2045-2322.
111. Patil SM, Tong L, Wang MD. Generating region of interests for invasive breast cancer in histopathological wholeslide-image. 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC); 2020. p. 723–728. https://doi.org/10.1109/COMPSAC48688.2020.0-174.
112. Pérez-Bueno F, Serra JG, Vega M, Mateos J, Molina R, Katsaggelos AK. Bayesian K-SVD fo:r H and E blind color deconvolution. applications to stain normalizationdata augmentation and cancer classification. Comput Med Imaging Graphics 2022;97:102048. https://doi.org/10.1016/j.compmedimag.2022.102048. ISSN 0895-6111. URL: https://www.sciencedirect.com/science/article/pii/S0895611122000210.
113. Pérez-Bueno F, Vega M, Sales MA, et al. Blind color deconvolution, normalization, and classification of histological images using general super Gaussian priors and Bayesian inference. Comput Meth Prog Biomedi 2021;211:106453. https://doi.org/10.1016/j.cmpb.2021.106453. ISSN 0169-2607. URL: https://www.sciencedirect.com/science/article/pii/S0169260721005277.
114. Phan NN, Huang C-C, Tseng L-M, Chuang EY. Predicting breast cancer gene expression signature by applying deep convolutional neural networks from unannotated pathological images. Front Oncol 2021;11. https://doi.org/10.3389/fonc.2021.769447. ISSN 2234-943X.
115. Qu H, Zhou M, Yan Z, et al. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. npj Precis Oncol 09, 2021;5(1):87. https://doi.org/10.1038/s41698-021-00225-9. ISSN 2397-768X.
116. Riasatian A, Babaie M, Maleki D, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. Med Image Anal 2021;70:102032. https://doi.org/10.1016/j.media.2021.102032. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000785.
117. Ruan J, Zhu Z, Wu C, Ye G, Zhou J, Yue J. A f:ast and effective detection framework for whole-slide histopathology image analysis. PLoS ONE 2021;16(5):e0251521. https://doi.org/10.1371/journal.pone.0251521.
118. Runz M, Rusche D, Schmidt S, Weihrauch MR, Hesser J, Weis C-A. Normalization of he-stained histological images using cycle consistent generative adversarial networks. Diagnos Pathol 08, 2021;16(1):71. https://doi.org/10.1186/s13000-021-01126-y. ISSN 1746-1596.
119. J Saltz, R Gupta, L Hou, T Kurc, P Singh, V Nguyen, D Samaras, KR Shroyer, T Zhao, R Batiste, J Van Arnam, Cancer Genome Atlas Research Network, I Shmulevich, AUK Rao, AJ Lazar, A Sharma, and V Thorsson. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep, 23 (1):181–193.e7, 04 2018. https://doi.org/10.1016/j.celrep.2018.03.086.
120. Schmauch B, Romagnoni A, Pronier E, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. Nat Commun 08, 2020;11(1):3877. https://doi.org/10.1038/s41467-020-17678-4.
121. Schmitz R, Madesta F, Nielsen M, et al. Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture. Med Image Anal 2021;70:101996. https://doi.org/10.1016/j.media.2021.101996. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000426.
122. Shao W, Wang T, Sun L, et al. Multi-task multimodal learning for joint diagnosis and prognosis of human cancers. Med Image Anal 2020;65:101795. https://doi.org/10.1016/j.media.2020.101795. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301596.
123. Sheikh TS, Lee Y, Cho M. Histopathological classification of breast cancer images using a multi-scale input and multi-feature network. Cancers 2020;12(8). https://doi.org/10.3390/cancers12082031. ISSN 2072-6694. URL: https://www.mdpi.com/2072-6694/12/8/2031.
124. Shi X, Su H, Xing F, Liang Y, Qu G, Yang L. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. Med Image Anal 2020;60:101624. https://doi.org/10.1016/j.media.2019.101624. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841519301604.
125. Srinidhi CL, Kim SW, Chen F-D, Martel AL. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Med Image Anal 2022;75:102256. https://doi.org/10.1016/j.media.2021.102256. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003017.
126. Srivastava A, Kulkarni C, Huang K, Parwani A, Mallick P, Machiraju R. Imitating pathologist based assessment with interpretable and context based neural network modeling of histology images. Biomed Inform Insights 2018;10:1178222618807481. https://doi.org/10.1177/1178222618807481. PMID: 30450002.
127. Sui D, Guo M, Zhang Y, Zhang L. Pyramid deconvolution net: Breast cancer detection using tissue and cell encoding information. Proceedings of the 4th International Conference on Big Data Research. ICBDR '20. New York, NY, USA: Association for Computing Machinery; 2021. p. 84–88. https://doi.org/10.1145/3445945.3445960. ISBN 9781450387750.
128. Zhao Q, Sui D, Liu W, et al. A pyramid architecturebased deep learning framework for breast cancer detection. BioMed Res Int 10, 2021;2021:2567202. https://doi.org/10.1155/2021/2567202. ISSN 2314-6133.
129. Sun D, Li A, Tang B, Wang M. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Comput Meth Prog Biomed 2018;161:45–53. https://doi.org/10.1016/j.cmpb.2018.04.008. ISSN 0169-2607. URL: https://www.sciencedirect.com/science/article/pii/S016926071830018X.
130. Sun P, He J, Chao X, et al. A computational tumor-infiltrating lymphocyte assessment method comparable with visual reporting guidelines for triple-negative breast cancer. EBioMedicine 08, 2021;70:103492. https://doi.org/10.1016/j.ebiom.2021.103492.
131. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal 2019;58:101544. https://doi.org/10.1016/j.media.2019.101544. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841519300799.
132. Thagaard J, Stovgaard ES, Vognsen LG, et al. Automated quantification of stil density with H&E-based digital image analysis has prognostic potential in triple-negative breast cancers. Cancers (Basel) 06, 2021;13(12):3050. https://doi.org/10.3390/cancers13123050.
133. Uchida S, Kojima T, Sugino T. Clinicopathological features, tumor mutational burden, and tumour-infiltrating lymphocyte interplay in erbb2-mutated breast cancer: in silico analysis. Pathol Oncol Res 2021;27. https://doi.org/10.3389/pore.2021.633243. ISSN 1532-2807.
134. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Scient Rep 06, 2021;11(1):13505. https://doi.org/10.1038/s41598-021-92799-4. ISSN 2045-2322.
135. Valieris R, Amaro L, de Toledo Osório CAB, et al. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. Cancers 2020;12(12). https://doi.org/10.3390/cancers12123687. ISSN 2072-6694. URL: https://www.mdpi.com/2072-6694/12/12/3687.
136. Valkonen M, Kartasalo K, Liimatainen K, Nykter M, Latonen L, Ruusuvuori P. Metastasis detection from whole slide images using local features and random forests. Cytometry Part A 2017;91(6):555–565. https://doi.org/10.1002/cyto.a.23089.
137. Venet L, Pati S, Feldman MD, Nasrallah MP, Yushkevich P, Bakas S. Accurate and robust alignment of differently stained histologic images based on greedy diffeomorphic registration. Appl Sci 2021;11(4). https://doi.org/10.3390/app11041892. ISSN 2076-3417. URL: https://www.mdpi.com/2076-3417/11/4/1892.
138. Vizcarra J, Place R, Tong L, Gutman D, Wang MD. Fusion in breast cancer histology classification. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. BCB '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 485–493. https://doi.org/10.1145/3307339.3342166. ISBN 9781450366663.
139. Wang L, Sun L, Zhang M, et al. Exploring pathologist knowledge for automatic assessment of breast cancer metastases in whole-slide image. . Proceedings of the 29th ACM International Conference on Multimedia. MM '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 255–263. https://doi.org/10.1145/3474085.3475489. ISBN 9781450386517.
140. Wang Y, Acs B, Robertson S, et al. Improved breast cancer histological grading using deep learning. Ann Oncol 2022;33(1):89–98. https://doi.org/10.1016/j.annonc.2021.09.007. ISSN 0923-7534. URL: https://www.sciencedirect.com/science/article/pii/S0923753421044860.
141. Wang Y, Kartasalo K, Weitz P, et al. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression–morphology analysis in breast cancer. Cancer Res 10, 2021;81(19):5115–5126. https://doi.org/10.1158/0008-5472.CAN-21-0482. ISSN 0008-5472.
142. Wodzinski M, Skalski A. Multistep, automatic and nonrigid image registration method for histology samples acquired using multiple stains. Phys Med Biol Jan, 2021;66(2), 025006. https://doi.org/10.1088/1361-6560/abcad7.
143. Thomas W, Eijkman CS, Rohr K. Adversarial domain adaptation to improve automatic breast cancer grading in lymph nodes. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018. p. 582–585. URL: https://api.semanticscholar.org/CorpusID:44160118.
144. Wu C, Ruan J, Ye G, et al. Identifying tumor in whole-slide images of breast cancer using transfer learning and adaptive sampling. 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI); 2019. p. 167–172. https://doi.org/10.1109/ICACI.2019.8778616.
145. Wulczyn E, Steiner DF, Xu Z, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS ONE 2020;15(6), e0233678. https://doi.org/10.1371/journal.pone. 0233678.
146. Xing F, Xie Y, Shi X, Chen P, Zhang Z, Yang L. Towards pixel-to-pixel deep nucleus detection in microscopy images. BMC Bioinform 09, 2019;20(1):472. https://doi.org/10.1186/s12859-019-3037-5. ISSN 1471-2105.
147. Xu S, Lu Z, Shao W, et al. Integrative analysis of histopathological images and chromatin accessibility data for estrogen receptor-positive breast cancer. BMC Med Genomics 12, 2020;13(11):195. https://doi.org/10.1186/s12920-020-00828-4. ISSN 1755-8794.
148. Xu Z, Verma A, Naveed U, Bakhoum SF, Khosravi P, Elemento O. Deep learning predicts chromosomal instability from histopathology images. iScience 04, 2021;24(5):102394. https://doi.org/10.1016/j.isci.2021.102394. ISSN 2589-0042.
149. Yang J, Ju J, Guo L, et al. Prediction of her2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal

deep learning. Computat Struct Biotechnol J 2022;20:333–342. https://doi.org/10.1016/j.csbj.2021.12.028. ISSN 2001-0370. URL: https://www.sciencedirect.com/science/article/pii/S2001037021005377.
150. Ye J, Luo Y, Zhu C, Liu F, Zhang Y. Breast cancer image classification on WSI with spatial correlations. . ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. p. 1219–1223. URL: https://api.semanticscholar.org/CorpusID:145921037.
151. Zhang H, Liu J, Yu Z, Wang P. Masg-gan: a multi-view attention superpixel-guided generative adversarial network for efficient and simultaneous histopathology image segmentation and classification. Neurocomputing 2021;463:275–291. https://doi.org/10.1016/j.neucom.2021.08.039. ISSN 0925-2312. URL: https://www.sciencedirect.com/science/article/pii/S0925231221012236.
152. Zhang W, Zhu C, Liu J, Wang Y, Jin M. Cancer metastasis detection through multiple spatial context network. Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition. ICCPR '19. New York, NY, USA: Association for Computing Machinery; 2020. p. 221–225. https://doi.org/10.1145/3373509.3373567. ISBN 9781450376570.
153. Zheng Y, Jiang Z, Zhang H, Xie F, Shi J, Xue C. Adaptive color deconvolution for histological wsi normalization. Comput Meth Prog Biomed 2019;170:107–120. https://doi.org/10.1016/j.cmpb.2019.01.008. ISSN 0169-2607. URL: https://www.sciencedirect.com/science/article/pii/S0169260718312161.
154. Zeiser FA, da Costa CA, de Oliveira Ramos G, Bohn HC, Santos I, Roehe AV. Deepbatch: a hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images. Exp Syst Appl 2021;185:115586. https://doi.org/10.1016/j.eswa.2021.115586. ISSN 0957-4174. URL: https://www.sciencedirect.com/science/article/pii/S095741742100988X.
155. Bagchi A, Pramanik P, Sarkar R. A multi-stage approach to breast cancer classification using histopathology images. Diagnostics (Basel) Dec 2022;13(1):126. https://doi.org/10.3390/diagnostics13010126.
156. Chen Y, Zhou Y, Chen G, et al. Segmentation of breast tubules in H&E images based on a dks-doubleu-net model. Biomed Res Int Sep 2022;2022:2961610. https://doi.org/10.1155/2022/2961610.
157. Chen S, Xiang J, Wang X, et al. Deep learning-based pathology signature could reveal lymph node status and act as a novel prognostic marker across multiple cancer types. Brit J Cancer Jul 2023;129(1):46–53. https://doi.org/10.1038/s41416-023-02262-6. ISSN 1532-1827.
158. Cong C, Liu S, Di Ieva A, Pagnucco M, Berkovsky S, Song Y. Colour adaptive generative networks for stain normalisation of histopathology images. Med Image Anal Nov 2022;82:102580. https://doi.org/10.1016/j.media.2022.102580. ISSN 1361-8415.
159. Fassler DJ, Torre-Healy LA, Gupta R, et al. Spatial characterization of tumor-infiltrating lymphocytes and breast cancer progression. Cancers (Basel) Apr 2022;14(9):2148. https://doi.org/10.3390/cancers14092148. ISSN 2072-6694.
160. Pour AF, White BS, Park J, Sheridan TB, Chuang JH. Deep learning features encode interpretable morphologies within histological images. Sci Rep Jun 2022;12(1):9428. https://doi.org/10.1038/s41598-022-13541-2. ISSN 2045-2322.
161. Huang Z, Shao W, Han Z, et al. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. npj Precis Oncol Jan 2023;7(1):14. https://doi.org/10.1038/s41698-023-00352-5. ISSN 2397-768X.
162. Huang J, Mei L, Long M, et al. Bm-net: Cnn-based mobilenet-v3 and bilinear structure for breast cancer detection in whole slide images. Bioengineering (Basel) Jun 2022;9(6):261. https://doi.org/10.3390/bioengineering9060261. ISSN 2306-5354.
163. Jarkman S, Karlberg M, Pocevičiūtė M, et al. Generalization of deep learning in digital pathology: experience in breast cancer metastasis detection. Cancers (Basel) Nov 2022;14(21):5424. https://doi.org/10.3390/cancers14215424. ISSN 2072-6694.
164. Jia F, Tan L, Wang G, Jia C, Chen Z. Asuper-resolution network using channel attention retention for pathology images. PeerJ Comput Sci 2023;9:e1196. https://doi.org/10.7717/peerj-cs.1196.
165. Jiang S, Suriawinata AA, Hassanpour S. Mhattnsurv: multi-head attention for survival prediction using whole-slide pathology images. Comput Biol Med 2023;158:106883. https://doi.org/10.1016/j.compbiomed.2023.106883. ISSN 0010-4825. URL: https://www.sciencedirect.com/science/article/pii/S0010482523003487.
166. Jin X, Huang T, Wen K, Chi M, An H. Histossl: self-supervised representation learning for classifying histopathology images. Mathematics 2023;11(1). https://doi.org/10.3390/math11010110. ISSN 2227-7390. URL: https://www.mdpi.com/2227-7390/11/1/110.
167. Lazard T, Bataillon G, Naylor P, et al. Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. Cell Rep Med 2022;3(12):100872. https://doi.org/10.1016/j.xcrm.2022.100872.
168. Liu X, Yuan P, Li R, et al. Predicting breast cancer recurrence and metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. Comput Biol Med 2022;146:105569. https://doi.org/10.1016/j.compbiomed.2022.105569. ISSN 0010-4825. URL: https://www.sciencedirect.com/science/article/pii/S0010482522003614.
169. Lu W, Toss M, Dawood M, Rakha E, Rajpoot N, Minhas F. Slidegraph(+): Whole slide image level graphs to predict her2 status in breast cancer. Med Image Anal August 2022;80:102486. https://doi.org/10.1016/j.media.2022.102486. ISSN 1361-8423.
170. Mondol RK, Millar EKA, Graham PH, Browne L, Sowmya A, Meijering E. hist2rna: an efficient deep learning architecture to predict gene expression from breast cancer histopathology images. Cancers (Basel) April 2023;15(9):2569. https://doi.org/10.3390/cancers15092569. ISSN 2072-6694.
171. Mou T, Liang J, Vu TN, Tian M, Gao Y. A comprehensive landscape of imaging feature-associated RNA expression profiles in human breast tissue. Sensors (Basel) January 2023;23(3):1432. https://doi.org/10.3390/s23031432. ISSN 1424-8220.
172. Sandarenu P, Millar EKA, Song Y, et al. Survival prediction in triple negative breast cancer using multiple instance learning of histopathological images. Scient Rep August 2022;12(1):14527. https://doi.org/10.1038/s41598-022-18647-1. ISSN 2045-2322.
173. Sheikh TS, Kim J-Y, Shim J, Cho M. Unsupervised learning based on multiple descriptors for wsis diagnosis. Diagnostics June 2022;12(6):1480. https://doi.org/10.3390/diagnostics12061480. ISSN 2075-4418.
174. Sun K, Chen Y, Bai B, Gao Y, Xiao J, Yu G. Automatic classification of histopathology images across multiple cancers based on heterogeneous transfer learning. Diagnostics March 2023;13(7):1277. https://doi.org/10.3390/diagnostics13071277. ISSN 2075-4418.
175. Tian J, Wang Y, Chen Z, Luo X, Xu X. Diagnose like doctors: weakly supervised fine-grained classification of breast cancer. ACM Trans Intell Syst Technol Feb 2023;14(2). https://doi.org/10.1145/3572033. ISSN 2157-6904.
176. Wang Z, Saoud C, Wangsiricharoen S, James AW, Popel AS, Sulam J. Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. IEEE Trans Med Imaging December 2022;41(12):3952–3968. https://doi.org/10.1109/TMI.2022.3202759. ISSN 1558-254X.
177. Wang R, Gu Y, Yang J. Cancer metastasis fast location based on coarse-to-fine network. 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML); 2022. p. 223–227. https://doi.org/10.1109/CACML55074.2022.00044.
178. Wu F, Liu P, Fu B, Ye F. Deepgcnmil: Multi-head attention guided multi-instance learning approach for whole-slide images survival analysis using graph convolutional networks. 2022 14th International Conference on Machine Learning and Computing (ICMLC). ICMLC 2022. New York, NY, USA: Association for Computing Machinery; 2022. p. 67–73. https://doi.org/10.1145/3529836.3529942. ISBN 9781450395700.
179. Wu X, Shi Y, Liu H, Li A, Wang M. Learning comprehensive multimodal representation for cancer survival prediction. Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. MLNLP '22. New York, NY, USA: Association for Computing Machinery; 2023. p. 332–336. https://doi.org/10.1145/3578741.3578806. ISBN 9781450399067.
180. Zheng H, Zhou Y, Huang X. Spatiality sensitive learning for cancer metastasis detection in whole-slide images. Mathematics 2022;10(15). https://doi.org/10.3390/math10152657. ISSN 2227-7390.
181. Zheng H, Zhou Y, Huang X. Improving cancer metastasis detection via effective contrastive learning. Mathematics 2022;10(14). https://doi.org/10.3390/math10142404. ISSN 2227-7390.
182. Schirris Y, Gavves E, Nederlof I, Horlings HM, Teuwen J. Deepsmile: contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. Med Image Anal 2022;79:102464. https://doi.org/10.1016/j.media.2022.102464. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001116.
183. Shen Y, Shen D, Ke J. Identify representative samples by conditional random field of cancer histology images. IEEE Trans Med Imaging 2022;41(12):3835–3848. https://doi.org/10.1109/TMI.2022.3198526.
184. Verdicchio M, Brancato V, Cavaliere C, Isgrò F, Salvatore M, Aiello M. A pathomic approach for tumor-infiltrating lymphocytes classification on breast cancer digital pathology images. Heliyon March 2023;9(3):e14371. https://doi.org/10.1016/j.heliyon.2023.e14371. ISSN 2405-8440.
185. Dehkharghanian T, Bidgoli AA, Riasatian A, et al. Biased data, biased AI: deep networks predict the acquisition site of tcga images. Diag Pathol May 2023;18(1):67. https://doi.org/10.1186/s13000-023-01355-3. ISSN 1746-1596.
186. Farahmand S, Fernandez AI, Ahmed FS, et al. Deep learning trained on hematoxylin and eosin tumor region of interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. Mod Pathol Jan 2022;35(1):44–51. https://doi.org/10.1038/s41379-021-00911-w. ISSN 1530-0285.
187. Cooper LAD, Demicco EG, Saltz JH, Powell RT, Rao A, Lazar AJ. Pancancer insights from the cancer genome atlas: the pathologist's perspective. J Pathol 2018;244(5):512–524. https://doi.org/10.1002/path.5028.
188. Dai B, Wu K, Wu T, et al. Faster-ppn: towards real-time semantic segmentation with dual mutual learning for ultra-high resolution images. . Proceedings of the 29th ACM International Conference on Multimedia. MM '21. NewYork, NY, USA: Association for Computing Machinery; 2021. p. 1957–1965. https://doi.org/10.1145/3474085.3475352. ISBN 9781450386517.
189. Gu H, Huang J, Hung L, "Anthony" Chen X. Lessons learned from designing an ai-enabled diagnosis tool for pathologists. Proc ACM Hum Comput Interact Apr 2021;5(CSCW1). https://doi.org/10.1145/3449084.
190. Hägele Miriam, Seegerer Philipp, Lapuschkin Sebastian, Bockmayr Michael, Samek Wojciech, Klauschen Frederick, Müller Klaus-Robert, Binder Alexander. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci Rep 2020;10(1):6423. https://doi.org/10.1038/s41598-020-62724-2. ISSN2045-2322. URL https://doi.org/10.1038/ s41598-020-62724-2.
191. Jansen C, Annuscheit J, Schilling B, et al. Curious containers: a framework for computational reproducibility in life sciences with support for deep learning applications. Future Gen Comput Syst 2020;112:209–227. https://doi.org/10.1016/j.future.2020.05.007. ISSN 0167-739X. URL: https://www.sciencedirect.com/science/article/pii/S0167739X19318096.
192. Lee K, Lockhart JH, Xie M, et al. Deep learning of histopathology images at the single cell level. Front Artif Intel 2021;4:754641. https://doi.org/10.3389/frai.2021.754641.
193. Li X, Li C, Rahaman MM, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. Artif Intel Rev 2022;55(6):4809–4878. https://doi.org/10.1007/s10462-021-10121-0. ISSN 1573-7462.
194. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135.
195. de Maturana EL, Pineda S, Brand A, Van Steen K, Malats N. Toward the integration of omics data in epidemiological studies: still a "long and winding road". Genet Epidemiol 2016;40(7):558–569. https://doi.org/10.1002/gepi.21992.

196. Graziani M, Andrearczyk V, Marchand-Maillet S, Müller H. Concept attribution: Explaining cnn decisions to physicians. Comput Biol Med 2020;123:103865. https://doi.org/10.1016/j.compbiomed.2020.103865. ISSN 0010-4825. URL: https://www.sciencedirect.com/science/article/pii/S0010482520302225.
197. Qaiser T, Mukherjee A, Chaitanya Reddy PB, et al. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 2018;72(2):227–238. https://doi.org/10.1111/his.13333.
198. Salvi M, Acharya UR, Molinari F, Meiburger KM. The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. Comput Biol Med January 2021;128:104129. https://doi.org/10.1016/j.compbiomed.2020.104129. ISSN 0010-4825.
199. Schneider L, Laiouar-Pedari S, Kuntz S, et al. Integration of deep learning-based image analysis and genomic data in cancer pathology: a systematic review. Eur J Cancer 2022;160:80–91. https://doi.org/10.1016/j.ejca.2021.10.007. ISSN 0959-8049. URL: https://www.sciencedirect.com/science/article/pii/S0959804921011606.
200. Shahid AH, Singh MP. Computational intelligence techniques for medical diagnosis and prognosis: problems and current developments. Biocybernet Biomed Eng 2019;39(3):638–672. https://doi.org/10.1016/j.bbe.2019.05.010. ISSN 0208-5216. URL: https://www.sciencedirect.com/science/article/pii/S0208521619300452.
201. Sobhani F, Robinson R, Hamidinekoo A, Roxanis I, Somaiah N, Yuan Y. Artificial intelligence and digital pathology: opportunities and implications for immuno-oncology. Biochim Biophys Acta (BBA) Rev Cancer 2021;1875(2):188520. https://doi.org/10.1016/j.bbcan.2021.188520. ISSN 0304-419X. URL: https://www.sciencedirect.com/science/article/pii/S0304419X21000196.
202. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal 2021;67:101813. https://doi.org/10.1016/j.media.2020.101813. ISSN 1361-8415. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301778.
203. Steiner DF, Chen P-HC, Mermel CH. Closing the translation gap: Ai applications in digital pathology. Biochim Biophys Acta (BBA) Rev Cancer 2021;1875(1):188452. https://doi.org/10.1016/j.bbcan.2020.188452. ISSN 0304-419X. URL: https://www.sciencedirect.com/science/article/pii/S0304419X20301712.
204. Tripathi S, Singh SK, Lee HK. An end-to-end breast tumour classification model using context-based patch modelling – a bilstm approach for image classification. Comput Med Imaging Graphics 2021;87:101838. https://doi.org/10.1016/j.compmedimag.2020.101838. ISSN 0895-6111. URL: https://www.sciencedirect.com/science/article/pii/S0895611120301336.
205. Caldonazzi N, Rizzo PC, Eccher A, et al. Value of artificial intelligence in evaluating lymph node metastases. Cancers (Basel) Apr 2023;15(9):2491. https://doi.org/10.3390/cancers15092491.
206. Couture HD. Deep learning-based prediction of molecular tumor biomarkers from H&E: a practical review. J Personal Med Dec 2022;12(12):2022. https://doi.org/10.3390/jpm12122022. ISSN 2075-4426.
207. Kim I, Kang K, Song Y, Kim T-J. Application of artificial intelligence in pathology: Trends and challenges. Diagnostics (Basel) 2022;12(11):2794. https://doi.org/10.3390/diagnostics12112794.
208. Wu Y, Cheng M, Huang S, et al. Recent advances of deep learning for computational histopathology: Principles and applications. Cancers February 2022;14(5):1199. https://doi.org/10.3390/cancers14051199. ISSN 2072-6694.
209. Zhao Y, Zhang J, Hu D, Qu H, Tian Y, Cui X. Application of deep learning in histopathology images of breast cancer: a review. Micromachines December 2022;13(12):2197. https://doi.org/10.3390/mi13122197. ISSN 2072-666X.
210. Seiler R, Black PC, Thalmann G, Stenzl A, Todenhöfer T. Is the cancer genome atlas (TCGA) bladder cancer cohort representative of invasive bladder cancer? Urol Oncol Semin Orig Investig 2017;35(7):458.e1–458.e7. https://doi.org/10.1016/j.urolonc.2017.01.024. ISSN 1078-1439. URL: https://www.sciencedirect.com/science/article/pii/S1078143917300595.
211. Kim Jr I, Sarkar I. Racial representation disparity of population-level genomic sequencing efforts. Studies in Health Technology and Informatics, 264. ; 08, 2019. p. 974–978. https://doi.org/10.3233/SHTI190369.
212. U.S. Cancer Statistics Breast Cancer Stat Bite. United States Cancer Statistics (USCS). 2020. URL: https://www.cdc.gov/cancer/uscs/about/stat-bites/stat-bite-breast.htm. Accessed 25 July 2023. [Internet].
213. Cancer Registry of Norway. The Registries: Cancer Statistics. URL: https://www.kreftregisteret.no/en/The-Registries/Cancer-Statistics/ 2020. Accessed 25 July 2023. [Internet].
214. Ivanescu AE, Li P, George B, et al. The importance of prediction model validation and assessment in obesity and nutrition research. Int J Obesity 2016;40(6):887–894. https://doi.org/10.1038/ijo.2015.214. ISSN 1476-5497.
215. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. https://doi.org/10.1136/bmj.g7594. ISSN 0959-8146.