Database, 2022, 1–10
DOI: https://doi.org/10.1093/database/baac093
Database update

BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images

Nadia Brancati[1,*,], Anna Maria Anniciello[2,¶], Pushpak Pati[3,4,**], Daniel Riccio[1,5,‡‡], Giosuè Scognamiglio[2,††], Guillaume Jaume[3,6,**], Giuseppe De Pietro[1,†], Maurizio Di Bonito[2,§], Antonio Foncubierta[3,‡], Gerardo Botti[2,†], Maria Gabrani[3,†], Florinda Feroce[2,¶] and Maria Frucci[1,‡]**

[1]Institute for High Performance Computing and Networking of the Research Council of Italy, 111 Via Pietro Castellino, ICAR-CNR, Naples 80131, Italy
[2]National Cancer Institute – IRCCS – Fondazione Pascale, 53 Via Mariano Semmola, Naples 80131, Italy
[3]IBM Research – Säumerstrasse 4, 8803 Rüschlikon, Zurich, Switzerland
[4]ETH, Rämistrasse 101, 8092, Zurich, Switzerland
[5]Department of Electrical Engineering and Information Technologies, Via Claudio, University of Naples Federico II, 21, Naples 80125, Italy
[6]EPFL Rte Cantonale, Lausanne 1015, Switzerland

[*]Corresponding author: Tel: +00390816139228; Email: nadia.brancati@icar.cnr.it
[†]Were responsible for the conceptualization of this project.
[‡]Contributed to the conceptualization and were responsible for execution and management of this project.
[§]Was responsible for the institutional clinical databases and contributed to the annotation process.
[¶]Contributed to the clinical design of the dataset, data annotations and documentation for the project and organization of the validation of results.
[**]Contributed to the dataset organization and validation.
[††]Collected cases and extracted and scanned slides from their institutional clinical databases.
[‡‡]Contributed to the conceptualization of this project.

Citation details: Brancati, N., Anniciello, A.M., Pati, P. *et al.* BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. *Database* (2022) Vol. 2022: article ID baac093; DOI: https://doi.org/10.1093/database/baac093

Abstract

Breast cancer is the most commonly diagnosed cancer and registers the highest number of deaths for women. Advances in diagnostic activities combined with large-scale screening policies have significantly lowered the mortality rates for breast cancer patients. However, the manual inspection of tissue slides by pathologists is cumbersome, time-consuming and is subject to significant inter- and intra-observer variability. Recently, the advent of whole-slide scanning systems has empowered the rapid digitization of pathology slides and enabled the development of Artificial Intelligence (AI)-assisted digital workflows. However, AI techniques, especially Deep Learning, require a large amount of high-quality annotated data to learn from. Constructing such task-specific datasets poses several challenges, such as data-acquisition level constraints, time-consuming and expensive annotations and anonymization of patient information. In this paper, we introduce the BReAst Carcinoma Subtyping (BRACS) dataset, a large cohort of annotated Hematoxylin and Eosin (H&E)-stained images to advance AI development in the automatic characterization of breast lesions. BRACS contains 547 Whole-Slide Images (WSIs) and 4539 Regions Of Interest (ROIs) extracted from the WSIs. Each WSI and respective ROIs are annotated by the consensus of three board-certified pathologists into different lesion categories. Specifically, BRACS includes three lesion types, i.e., benign, malignant and atypical, which are further subtyped into seven categories. It is, to the best of our knowledge, the largest annotated dataset for breast cancer subtyping both at WSI and ROI levels. Furthermore, by including the understudied atypical lesions, BRACS offers a unique opportunity for leveraging AI to better understand their characteristics. We encourage AI practitioners to develop and evaluate novel algorithms on the BRACS dataset to further breast cancer diagnosis and patient care.

Database URL: https://www.bracs.icar.cnr.it/

Introduction

Histology images contain both complex and ambiguous information, thus challenging pathologists to perform a robust, reproducible and efficient analysis. Furthermore, histology images are very large, which makes their analysis cumbersome and time-consuming. With advances in Computer-Aided-Diagnosis (CAD), Artificial Intelligence (AI) techniques, especially Machine Learning (ML) and Deep Learning (DL), have the potential to address the aforementioned bottlenecks (1–4). Recent advancements in DL have demonstrated superior capabilities compared to classical ML approaches for CAD (5–11). The crucial advantage of DL approaches is their ability to learn task-specific salient features directly from the training data. These techniques can identify discriminative morphological patterns from large datasets to diagnose histology images in a standardized and objective manner. However, this superiority comes at the cost of acquiring large, high-quality, variable and unbiased annotated

Received 18 March 2022; Revised 16 September 2022; Accepted 1 October 2022

Key Points

- This dataset provides a large and heterogeneous set of realistic breast histology images both at WSI and ROI levels.
- The provided ROIs range over variable dimensions by entirely including the diagnostic lesion, thus avoiding the loss of diagnostically relevant information.
- The images are acquired from a large number of patients encompassing large variability.
- Seven different subtypes of lesions are included, two of them representing atypical lesions, also known as precancerous lesions.

training datasets. Indeed, there exist several challenges in adopting such techniques in digital pathology (12), such as (i) the requirement of large annotated datasets, (ii) the need for sufficiently variable data to set up cross-patient experiments, (iii) the inclusion of diagnostically challenging lesions, that are generally difficult and expensive to acquire, (iv) the utilization of sub-region annotations to delineate Region Of Interest (ROI) of the Whole-Slide Images (WSIs), (v) the coverage of diagnostic spectrum and (vi) coping with data leakage and noisy annotations. Although several datasets for diagnosing breast histology images exist (13–17), they do not meet all the aforementioned criteria. For instance, some datasets focus on specific diseases that include only binary classes (13, 14), as they aim to categorize lesions only into benign and malignant classes, which do not depict a sufficiently large spectrum of subtypes in breast cancer diagnosis. On the other hand, datasets handling multiple classes (15, 16) include only a small number of training samples (both at WSI and ROI levels) collected from a few patients, thus limiting the dataset variability. Many of these datasets contain standardized images without clinical artifacts, e.g., staining anomalies, ink marks, tissue folding, blurred regions, tears, etc. Consequently, these datasets do not comprehensively represent real-world breast cancer diagnosis, since many of the clinical artifacts do not prevent pathologists from making a diagnosis. Thus, it is necessary to collect a breast cancer dataset consisting of heterogeneous images across the diagnostic spectrum which is comparable to real-world diagnosis performed by the pathologists.

We introduce BReAst Carcinoma Subtyping (BRACS), a large cohort of Hematoxylin and Eosin (H&E)-stained images to advance CAD of breast lesions. BRACS presents the following advantages over the extant breast cancer image datasets: (i) it includes a large and heterogeneous set of realistic breast histology images (both at WSI and ROI levels), (ii) ROIs range over variable dimensions by entirely including the diagnostic lesion, thus avoiding the loss of diagnostically relevant information, (iii) the images are acquired from a large number of patients encompassing large variability and (iv) two atypical lesion categories, also known as precancerous lesions, are included along with other categories. In particular, we consider the following lesion types, Normal (N), Pathological Benign (PB), Usual Ductal Hyperplasia (UDH), Flat Epithelial Atypia (FEA), Atypical Ductal Hyperplasia (ADH), Ductal Carcinoma in Situ (DCIS) and Invasive Carcinoma (IC). Thus, BRACS represents a more realistic benchmark for breast cancer automatic diagnosis by including several types of typical and atypical tissue samples over a wide variety of WSIs and ROIs extracted from a large number of patients.

Methods

The BRACS dataset is created to support the development of breast cancer diagnostic methods through the automatic analysis of histology images. The dataset was built through the collaboration of the National Cancer Institute—Scientific Institute for Research, Hospitalization and Healthcare (IRCCS) 'Fondazione G. Pascale', the Institute for High Performance Computing and Networking (ICAR) of National Research Council (CNR) and International Business Machines (IBM) Research—Zurich. The dataset was acquired from patients between 2019 and 2020, by board-certified clinicians of the Department of Pathology at the National Cancer Institute—IRCCS 'Fondazione G. Pascale' in Naples (Italy). The samples were generated from H&E-stained breast tissue biopsy slides and were selected based on the diagnostic reports of the patients. The age of the patients range from 16 to 86 years old, with about 61% of patients in the range of 40–60 and only a few patients of aged <20 or >80.

As the introduction of AI in medical applications has opened up a wide debate on the ethical problems that this practice can raise, it is mandatory to find the right trade-off between the use of medical data and the protection of patient privacy. In the specific case of BRACS, the purpose is not to study the disease over time, but to improve the accuracy of still images analysis algorithms. Thus, all confidential data as well as the label from the slide images and the patient code in the corresponding files have been completely eliminated.

Dealing with rich and comprehensively annotated histology images is a complex and time-consuming task and involves a large number of experts in very different fields. In the specific case, the experts involved are the biologists who select and scan the slides, the pathologists who make the annotations of the slides and the computer scientists who guide and support the annotation process (18). Indeed, the interaction between the various experts brings out critical issues to be addressed with motivated decisions to make the image acquisition and annotation process feasible and at the same time useful for possible experimentation (19).

The WSIs have been obtained by scanning slides that were selected by a biologist of the pathological anatomy department. This selection was performed on the basis of the pathology reports obtained from clinical routine. Each report is associated with a set of slides containing small tissue parts extracted from the primary sample and processed with the H&E staining method. A pathology report also includes the final diagnosis performed by pathologists on the most significant lesion subtype detected through the analysis of associated slides. Clearly the subtype associated with the report can appear in one or more slides, but not necessarily in all of them. In order to take into account ethical issues, all WSIs were de-identified before making them available to the pathologists. Image quality was also examined, and WSIs for which pathologists were not able to make a decision due to dramatic aberrations (i.e. out-of-focus and too-high staining irregularity) were eliminated from the image dataset. Conversely, images with low quality but of a sufficient standard to guarantee the AI downstream have been kept in the dataset (20).

A WSI typically includes several lesions of different subtypes. Moreover, the inclusion of atypical breast lesions at

both WSI and ROI levels significantly increases the complexity of the annotation task and makes it prone to observer variability. In order to ensure a high reliability of the annotations, three expert pathologists have been involved to annotate both WSIs and ROIs. To support pathologists in the image annotation, the QuPath software (21) was provided, as it allows zooming, scrolling and marking. In a first step, WSIs have been inspected by each pathologist, who assigned the corresponding label according to the most aggressive tumor subtype she/he has detected in the image. In a second step, all annotations were collectively checked by three pathologists, and those with disagreement were further discussed and re-annotated when consensus was reached or discarded otherwise. The number of disagreements was relatively low, and the consensus was found by observing together the most significant regions.

A subset of the annotated WSIs was split into three disjoint subsets, each of which was assigned to one of the pathologists. Each pathologist extracted a set of ROIs from her/his subset, being careful that at least one ROI with the same subtype of the WSI was selected and that the set of the extracted ROIs might keep the dataset balanced with respect to the classes. Each extracted ROI corresponds to a unique category and can include single or multiple glandular structures (22). Consequently, ROIs have variable size. By using the QuPath software, the ROI is marked with an appropriate color encoding the corresponding label. Figure 1 presents the annotation procedure for a sample WSI and its corresponding ROIs.

The number of extracted ROIs per WSI ranges from 0 to 119, with an average of 11 ROIs per WSI, selected across all WSIs to collectively encompass the lesion heterogeneity. This aspect is crucial for representing pathogenesis and disease progression and consequently allowing for the inclusion of sufficiently variable data for DL model training. In particular, for the annotation of ROIs, the interaction between the pathology and Information Technology (IT) teams was very high to guarantee the inclusion of information-rich examples, inter-class balancing and noise-free annotations. Also for

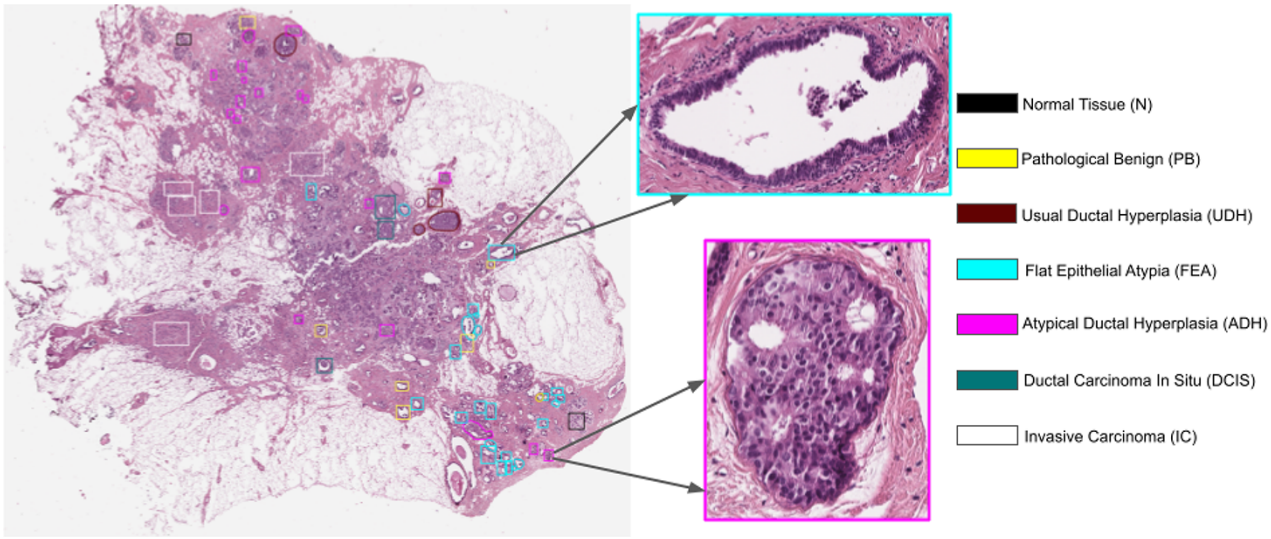

Figure 1. Example of a WSI and its associated ROIs. A fixed palette is used to mark the tumor subtype of the lesion.

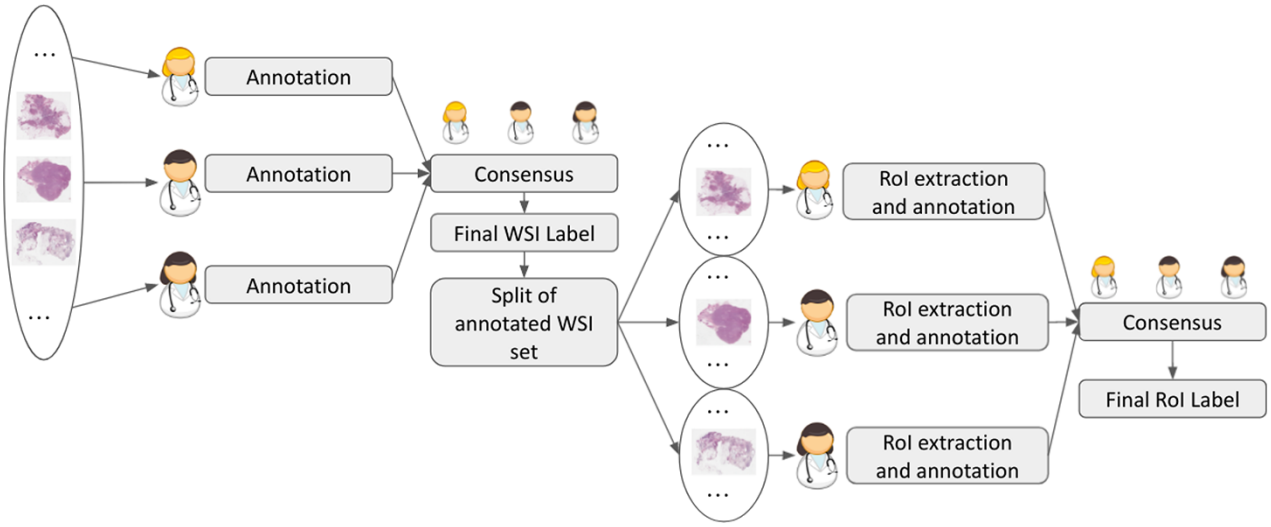

Figure 2. Annotation process of WSIs and ROIs.

ROIs, a second step to reach the consensus on the final annotation has been performed. Indeed, for some ROIs, pathologists were in difficulty on deciding among different lesion classes, as different subtypes of lesions have similar morphological features (e.g. ADH and DCIS) and cannot be easily classified disregarding the WSI-level context. Thus, the pathologists examined together the corresponding WSI, to jointly assign a final label to the ROI. Figure 2 summarizes the annotation process of both WSIs and ROIs.

An important characteristic of BRACS is the inclusion of atypical lesions, ADH and FEA. While ignored in other public datasets, these categories remain important as they might be indicators of either (i) the presence of abnormalities in the neighboring breast tissue that could go undetected, e.g., due to the extraction of small tissue samples, or (ii) a high risk of onset of future carcinoma, i.e., development of DCIS and IC. In addition, these lesions cannot be detected by mammography or other breast imaging techniques nor can they be felt during a clinical breast examination. When detected in a core biopsy, more frequent imaging follow-up and often surgical excision are recommended (23). BRACS also includes lesion subtypes belonging to benign and malignant types. In particular, benign lesions are subtyped as either noncancerous lesions (PB) or inflammatory responses (UDH). Malignant lesions are categorized as either DCIS or IC. Finally, histology images representing normal tissue sample are classified in the N category.

In order to clarify the description of the different tissue subtypes, a brief description of the mammary gland should be considered. The breast is a modified apocrine sweat gland, made up of 15–25 independent glandular units called lobes, each of which is formed by a compound tubulo-acinar gland. The lobes are composed of adipose tissue and divided by connective tissue septa. Inside each lobe, the main ducts branch into terminal ducts, each of which leads to a lobule that is made up of many berries to form the ductulo-lobular terminal unit. Detailed information on lesions included in BRACS can be found in (24). The specific features of the different sample tissue subtypes are briefly summarized in the following, and a representative example for each of them is shown in Figure 3.

Normal Tissue

In normal mammary glandular tissue, there are two types of epithelial cells (the luminal layer and the basal myoepithelial layer) and two types of stromal cells (interlobular stroma and intralobular stroma). Differently from PB, the ratio between epithelial component and stroma is preserved.

Pathological Benign

Benign breast lesions can be grouped according to the risk of developing invasive carcinoma and include several groups of histological entities classified in relation to morphology. In our study, because of differential diagnosis, in the PB category we included both non-proliferative lesions and proliferative lesions with the exception of UDH, FEA and ADH, which were considered as three independent subtypes. Therefore, PB includes cyst, apocrine metaplasia, ductal ectasia, squamous metaplasia, atrophy, stromal fibrosis, mastitis, sclerosing adenosis, papilloma, radial scar and simple and complex fibroadenoma.

Usual Ductal Hyperplasia

UDH has a rate of occurrence of 20% (25) and is characterized by an increase in the epithelial layers. It is a cohesive proliferation of disorderly distributed but oriented cells. It can have different architectural aspects (solid pattern, fenestrated pattern and micropapillary pattern). Even if UDH shares some architectural features with ADH and DCIS, it does not show atypia.

Flat Epithelial Atypia

FEA represents the 3.8–10% of core needle biopsy samples (25) and is a proliferative lesion characterized by low-grade cytological atypia, cell monomorphism, loss of polarity and orientation with respect to the basement membrane, presence of apical snout, endoluminal secretion and frequent calcifications.

Atypical Ductal Hyperplasia

ADH is a proliferation of monomorphic cells, which only partially fill the ductal spaces. Architectural aspects include a solid pattern, a cribriform pattern and a papillary pattern. The cytologic atypia is similar to that of low-grade DCIS, but the lesion spans no more than 2 mm or has an insufficient architectural atypia involving only partially ducts and/or lobules. Studies suggest that 5–20% of core needle biopsies are ADH and 10–20% of them generally upgrade to DCIS or IC (25).

Ductal Carcinoma in Situ

In situ carcinoma is a malignant proliferation of epithelial cells that fills the entire duct, without evidence of stroma invasion. Typically it involves multiple adjacent ductal spaces. It can have cribriform, solid, papillary and micropapillary patterns.

Invasive Carcinoma

IC is characterized by the invasion of tumor cells infiltrating the breast stroma with loss of peripheral myoepithelial cells. The presence of the myoepithelial cell layer is an important distinction of DCIS from IC.

Detecting certain subtypes is particularly challenging as some morphological patterns can be shared by several classes. For instance, ADH shares morphological similarities with DCIS. In certain cases, it even includes all the features of DCIS, but is simply limited in size. Also, UDH, ADH and DCIS are all characterized by an intraductal growth pattern, which makes these classes difficult to classify and differentiate in H&E-stained sections.

BRACS dataset characteristics

The BRACS dataset contains 547 WSIs related to 189 different patients. It also includes 4539 ROIs extracted from 387 WSIs collected on 151 patients. All slides were scanned with an Aperio AT2 scanner at 0.25 $\mu m/pixel$ using a magnification factor of 40×. Table 1 and Table 2 report the number of WSIs (with and without ROIs) and ROIs according to the lesion type and subtypes, respectively.

In recent years, several datasets have been proposed, with more and more samples and classes (26–31). Table 3 details existing public datasets of histology images for breast lesion classification. This table also includes information about BRACS dataset for comparison. Datasets that are either

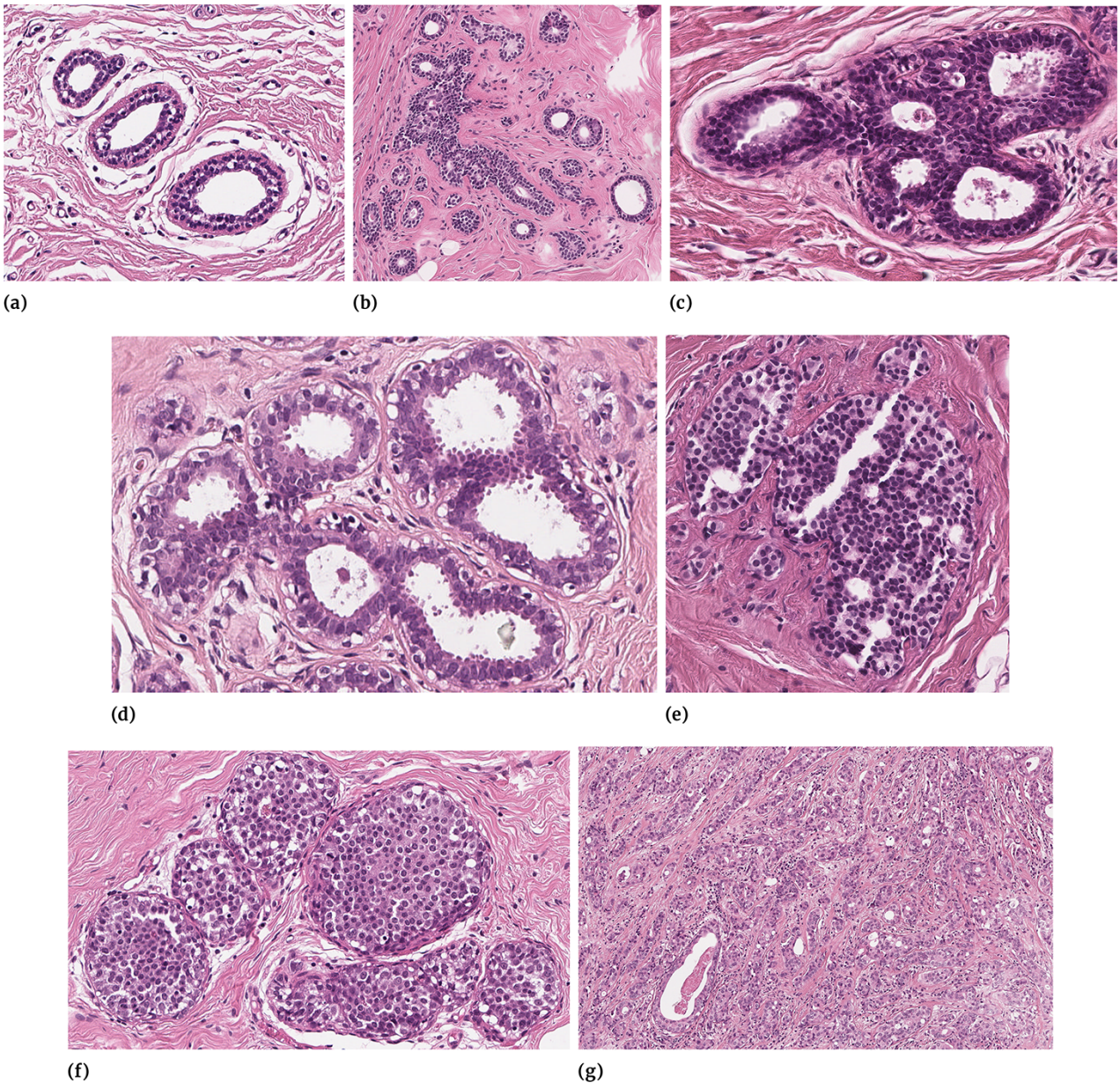

Figure 3. Examples of different tissue samples: (a) N, (b) PB, (c) UDH, (d) FEA, (e) ADH, (f) DCIS and (g) IC.

Table 1. BRACS data distribution according to lesion type.

Data	Benign	Atypical	Malignant	Total
WSIs with ROIs	149	75	163	387
WSIs without ROIs	116	14	30	160
WSIs	265	89	193	547
ROIs	1837	1263	1439	4539

Table 2. BRACS data distribution according to lesion subtype.

Data	N	PB	UDH	FEA	ADH	DCIS	IC
WSIs with ROIs	17	77	55	34	41	51	112
WSIs without ROIs	27	70	19	7	7	10	20
WSIs	44	147	74	41	48	61	132
ROIs	484	836	517	756	507	790	649

(i) no longer accessible (27), (ii) subsets of already mentioned datasets (28) or (iii) targeting specific tasks, e.g., lesion proliferation scores prediction (30) and pN-stage prediction (31), are not mentioned in Table 3.

The IDC (13) and Camelyon16 (29) datasets focus on the detection of the presence of a given lesion. In particular, IDC provides ROIs at small spatial resolution (50×50 pixels) extracted from large areas of Invasive Ductal Carcinoma. We emphasize that even if the number of ROIs in BRACS is lower than in IDC, BRACS ROIs are on average much larger allowing the inclusion of whole glandular areas. A subset of Camelyon16 WSIs is also provided with annotations of metastases. However, BRACS includes a larger number of WSIs than Camelyon16 and more subtypes. BreakHis (32) and Breast Cancer Histology (BACH) (26) datasets are devoted to multi-classification tasks, but remain significantly smaller

Table 3. Popular publicly available breast histopathology image datasets and BRACS dataset.

	Lesion classes					
Dataset, Year	**Benign**	**Atypical**	**Malignant**	**Data type size (Magnification)**	**n. Pat.**	**Resolution in pixels**
IDC (13), 2014	IDC negative	–	IDC positive	*ROI* 277.524 (40×)	162	50 × 50
BreakHis (15), 2015	Adenoid fibroadenoma Phyllodes tumor Tubular adenoma	–	Carcinoma Lobular carcinoma Mucinous carcinoma Papillary carcinoma	*ROI* 1.995 (40×) 2.081 (100×) 2.013 (200×) 1.820 (400×)	82	700×460
Camelyon16 (29), 2016	Lymph nodes metastatic negative	–	Lymph nodes metastatic positive	*WSI* 400 (20× and 40×)	400	Variable size
BACH (26), 2018	Normal Benign	–	In situ carcinoma Invasive carcinoma	*ROI* 400 (200×) *WSI* 10 (20×)	3910	2048×1536 Variable size
TCGA-BRCA (33), 2016	Normal Benign	–	Invasive ductal carcinoma Invasive lobular carcinoma Special histologies Mixed histologies	*WSI* 1978 (20× and 40×)	1093	Variable size
CPTAC-BRCA (34), 2020	Normal Benign	–	Invasive ductal carcinoma Invasive lobular carcinoma Special histologies Mixed histologies	*WSI* 642 (20× and 40×)	134	Variable size
BRACS (35), 2021	Normal Benign UDH	FEA ADH	In situ carcinoma Invasive carcinoma	*ROI* 4537 (40×) *WSI* 547 (40×)	151 189	Variable size Variable size

Table 4. WSI-level split according to the lesion type.

	Benign	**Atypical**	**Malignant**	**Total WSIs**	**Total patients**
Train	203	52	140	395	133
Validation	30	14	21	67	25
Test	32	23	32	85	31

Table 5. WSI-level split according to the lesion subtype.

	N	**PB**	**UDH**	**FEA**	**ADH**	**DCIS**	**IC**
Train	27	120	56	24	28	40	100
Validation	10	11	9	6	8	9	12
Test	7	16	9	11	12	12	20

Table 6. ROI-level split according to the lesion type.

	Benign	**Atypical**	**Malignant**	**Total ROIs**	**Total patients**
Train	1460	1011	1186	3657	106
Validation	135	90	87	312	15
Test	242	162	166	570	30

Table 7. ROI-level split according to the lesion subtype.

	N	**PB**	**UDH**	**FEA**	**ADH**	**DCIS**	**IC**
Train	357	714	389	624	387	665	521
Validation	46	43	46	49	41	40	47
Test	81	79	82	83	79	85	81

than BRACS, both in terms of image size and number of samples, and include less subtypes than BRACS.

Moreover, BreakHis and BACH include fixed-size ROIs, while BRACS images are of arbitrary size. Assuming fixed-size tumor regions is a strong assumption that does not apply in real-life scenarios. Limiting the size of ROIs requires either (i) partially cutting the lesion, hence producing a loss of information that could be pivotal for a correct diagnosis or (ii) manually curating ROIs such that they all have similar sizes, which does not encompass tumor heterogeneity. By proposing samples of varying sizes, BRACS promotes the development of DL algorithms that need to be able to operate on inputs of different dimensionality. BRACS and BACH share the same malignant lesion subtypes, while BreakHis refines this class by partitioning it into four specific subtypes. All the benign lesion subtypes defined in BACH and BreakHis are also included in BRACS (normal and benign). In addition, BRACS includes the UDH lesion subtype that is not considered in BACH and BreakHis.

The Cancer Genome Atlas (TCGA) -Breast cancer (BRCA) dataset (33) includes a very large number of WSI images. However, it was not acquired specifically for image-based tumor classification, as it also includes clinical data and was mainly designed to facilitate studies on tumor classification algorithms based on multiple data sources. Indeed, in literature, TCGA-BRCA is mainly used to validate such kinds of algorithms. As regards its use for tumor classification only based on WSIs, the main drawback is that a partitioning of images in training/validation/testing sets is not provided, so limiting reproducibility of the experiments and comparison of the results with the state of the art. Secondly, many WSIs are labeled as mixed pathology, so researchers who use TCGA-BRCA for image classification either consider only a partial subset or submit the WSIs to their expert pathologists

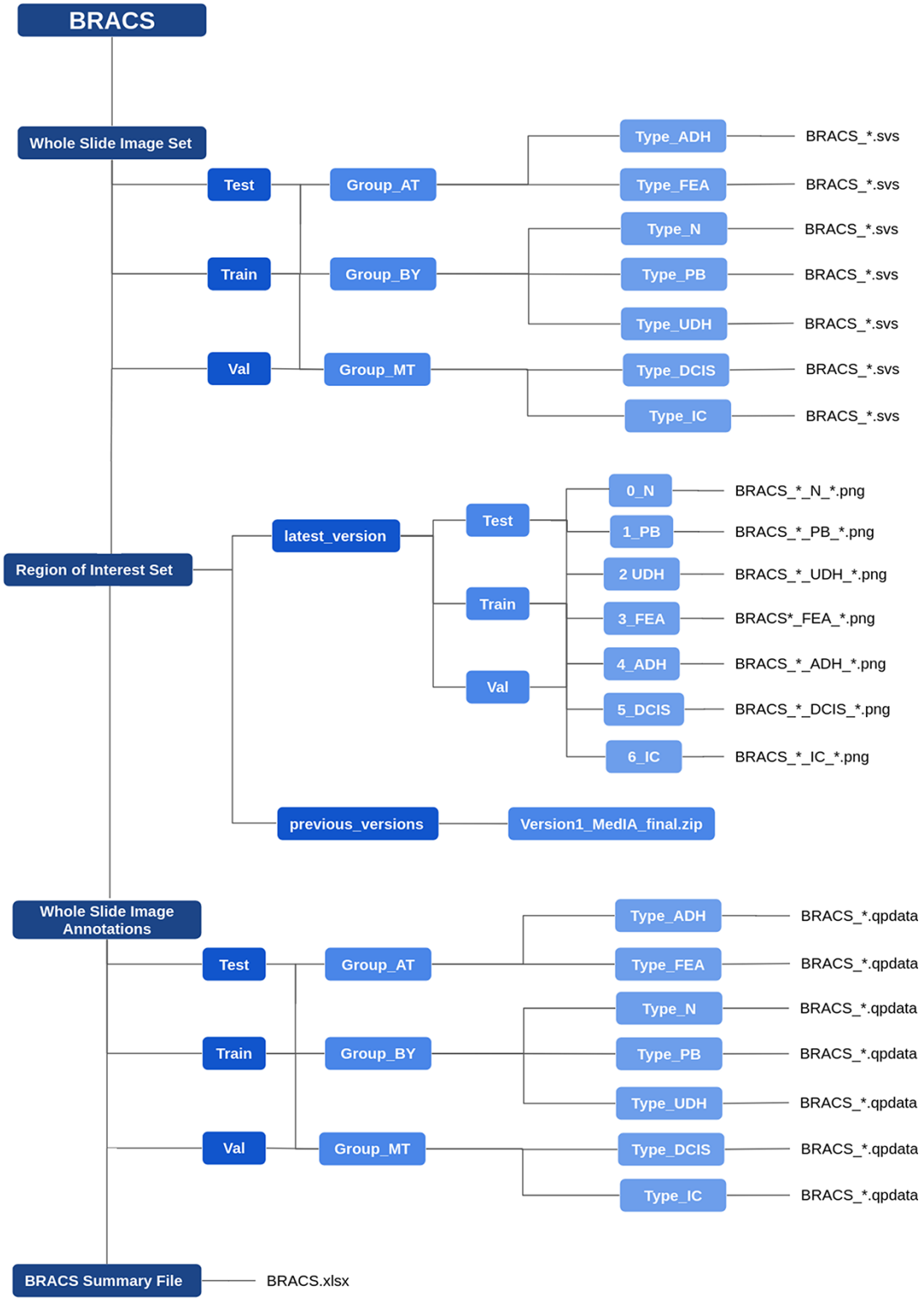

Figure 4. The organization of BRACS dataset folders.

to get a more refined set of tumor subtypes. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) -BRCA (34) also contains both WSIs and clinical data. In particular, it includes a subset of images from TCGA-BRCA, for which it considers a much higher number of protein features. However, some of the WSIs are not provided with the corresponding label. Like TCGA-BRCA, CPTAC-BRCA is designed to evaluate tumor classification algorithms based on multiple data sources and shows the same limits in its applicability to evaluate image-based tumor classification methods. Moreover, neither TCGA-BRCA nor CPTAC-BRCA include annotations for ROIs, unlike BRACS, which instead

provides a large number of annotated ROIs for each tumor subtype.

In summary, BRACS characteristics are unique, as they allow for multi-class classification task of breast cancer lesions, including challenging atypical lesions. In terms of the number of patients, WSIs and ROIs is also the largest dataset of histology images providing a standardized benchmark to evaluate tumor classification algorithms.

Dataset organization

To foster reproducibility and following ML best practices, we provide pre-defined WSI- and ROI-level splits in training, validation and test sets. Data split was generated such that all the WSIs extracted from a patient belong to the same set. Similarly, all the ROIs extracted in a given WSI are assigned to the same split. By following this approach, we avoid that different sets being including in correlated patient- and slide-level information, which could lead to overly optimistic prediction results (36). Table 4 and Table 5 present the number of WSIs included in the train, validation and test splits. Information about the number of patients in each set is provided in Table 4. Equivalent information for ROIs is shown in Table 6 and Table 7. At ROI level, the ratio between the most common subtype (PB with 714 samples) and the least common one (N with 357) is around two. At WSI level, the most common subtype is PB with 120 samples and the least common one is FEA with 24 samples, hence the ratio is approximately 4. Considering the constraint of patient- and WSI-level split and the fact that atypical lesions are more rare than malignant ones, BRACS offers a rather balanced set that can directly be used for training DL systems. The class imbalance across sets is due to the following aspects. First of all, the extraction of an equal number of WSIs (ROIs) for each lesion subtype was quite difficult to plan a priori. Secondly, not all WSIs of a patient have the same label, and the number of WSIs for each patient could not be defined a priori. Moreover, atypical lesions barely appear as the most severe lesion in the WSIs. Similar limitations hold for the ROIs, but in this case, the variability of the number of ROIs into a WSI make the set-balancing task even more difficult. Finally, to ensure having a wide variability of data for the training DL models, the number of samples selected for the training set is higher than that considered for the validation and test sets. The trade-off between balancing data and the absence of contamination in the reference sets has been extensively evaluated. We opted to avoid the sharing of indirect information between the reference sets instead of obtaining more balanced sets, by considering that the balancing problem can be mitigated by employing data augmentation (e.g. by affine transformations).

The BRACS dataset can be publicly accessed and downloaded via the BRACS website (35). Anyone registering and agreeing with the terms of use (Creative Commons CC0 license) can freely download it. Once registered, the user can access via File Transfer Protocol (FTP) to the server containing all the data.

The data are organized as follows. The WSIs are stored in the 'Whole Slide Image Set' folder, that includes the train, validation and test data. Each data split folder is further partitioned in Benign (BT), Atypical (AT) and Malignant (MT) folders, each of which includes folders corresponding to lesion subtypes. The WSIs are stored as svs files. All the files follow the same naming convention. For instance, the file 'BRACS_1238.svs' refers to the slide ID 1238, whose label is defined by the name of the folder that contains it. The ROIs are stored in the 'Region of Interest Set' folder, which follows the same structure as the WSI set. The files are stored in png format, where the file 'BRACS_1238_PB_32.png' refers to the ROI number 32, extracted from the WSI named 'BRACS_1238.svs' and labeled as Pathological Benign. The folder also includes a 'previous_versions' archive that contains a zip file with data that have been used in a series of publications during the dataset collection process, e.g., (37–40). The WSI annotations are stored in the 'The Whole Slide Image Annotations' folder, which follows the same structure as the WSI set. It includes annotation files in qpdata format (based on QuPath (21)) for visualizing the ROIs inside their corresponding WSI. Finally, a summary file is provided as an xlsx file, which reports for each WSI, its label, reference set (training/validation/test), corresponding patient ID and the number of associated ROIs, if any. Figure 4 highlights the folder organization of BRACS.

Table 8. Results at WSI level for 3-class classification task.

	Benign	Atypical	Malignant	Total
F-measure	74.4	57.2	78.0	69.8
Precision	75.6	51.5	86.5	71.2
Recall	72.5	65.2	71.9	69.9
Accuracy	–	–	–	70.3

Table 9. Results at ROI level for 7-class classification task.

	N	PB	UDH	FEA	ADH	DCIS	IC	TOT
F-measure	73.5	45.0	34.3	64.0	24.0	58.3	81.0	54.3
Precision	71.8	43.0	41.4	58.6	32.6	50.9	80.5	54.1
Recall	75.3	46.8	29.3	69.9	19.0	68.2	81.5	56.0
Accuracy	–	–	–	–	–	–	–	55.9

Results and discussion

In order to show the potential utility of this dataset, we performed two DL-based multi-classification experiments on the BRACS dataset that can be considered as a baseline for researchers. As the BRACS dataset is particularly unbalanced, especially considering the atypical lesions, the performance was evaluated in terms of the F-measure.

Already during the construction of the BRACS dataset, its utility for DL was evaluated on the 7-class (i.e. N, PB, UDH, ADH, FEA, DCIS and Invasive) classification of ROIs using the data available in version 1 of the folder structure described in Figure 4. This approach exploits graph representations of tissue and Graph Neural Network (GNN) to predict the class at ROI level (37). The obtained weighted F-measure is 61.5% with a standard deviation of 0.9.

WSI-level classification was performed on 3-class (i.e. Benign, Atypical and Malignant) classification of WSIs exploiting the network architecture presented in (41) on the latest version of the dataset. The Convolutional Neural Network (CNN) structure consists of a compressing path and a learning path. In the compressing path, the gigapixel image is packed into a grid-based feature map by using a residual network devoted to the feature

extraction of each patch into which the image has been divided. In the learning path, attention modules are applied to the grid-based feature map, taking into account spatial correlations of neighboring patch features to find ROIs, which are then used for the final whole-slide analysis. Due to computational limits, 10 × magnification WSIs have been considered. WSIs were normalized by using the method proposed in (42). For the augmentation, in addition to affine transformations, the normalized WSIs at 5 × and 2.5 × magnification were also used. The obtained results in terms of F-measure, precision, recall and total accuracy are shown in Table 8.

The code for replicating the results at WSI level can be found at http://github.com/nadiabrancati/ABNN-WSI-Classification.

The CNN architecture implemented for the WSIs was also tested on the ROIs, but with a 7-class classification protocol. The results of this experiment are shown in Table 9.

Conclusion

In this paper we have presented a new dataset, namely BRACS, of histological images of breast cancer, which present a subdivision of the images in seven subtypes. It includes two classes of atypical subtypes, namely FEA and ADH. Moreover, it provides both WSIs and ROIs in far greater numbers than other existing databases. Particular attention was paid to the distribution of WSIs and ROIs with respect to the different subtypes, so as to make the dataset as balanced as possible. This makes the dataset particularly useful for automatic benchmarking in breast cancer diagnosis. We also tested a CNN architecture (41), which can be considered as a baseline for future comparisons. A further extension of this dataset is currently underway, in order to make it suitable for carrying out large-scale experiments and international challenges.

Acknowledgements

We would like to acknowledge Engr. Alessandro Manzoni from the Scientific Institute for Research, Hospitalization and Healthcare Fondazione Pascale for his support for management of technical instruments. We would like to acknowledge Mario Sicuranza Ph.D. from the Institute for High Performance Computing and Networking of National Research Council for the development and technical implementation of the BReAst Carcinoma Subtyping site.

References

1. Madabhushi,A. and Lee,G. (2016) Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.*, **33**, 170–175. 10.1016/j.media.2016.06.037.
2. Tizhoosh,H.R. and Pantanowitz,L. (2018) Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inf.* **38**, 9.
3. Srinidhi,C.L., Ciga,O. and Martel,A.L. (2020) Deep neural network models for computational histopathology: a survey. *Med. Image Anal.* **67**, 101813.
4. de Matos,J., Ataky,S.T.M., de Souza Britto,A. *et al.* (2021) Machine Learning Methods for Histopathological Image Analysis: a Review. *Electronics*, **10**, 562. 10.3390/electronics10050562.
5. Araújo,T., Aresta,G., Castro,E. *et al.* (2017) Classification of breast cancer histology images using convolutional neural networks. *PloS One*, **12**, e0177544. 10.1371/journal.pone.0177544.
6. Bardou,D., Zhang,K. and Ahmad,S.M. (2018) Classification of breast cancer based on histology images using convolutional neural networks. *IEEE Access*, **6**, 24680–24693. 10.1109/ACCESS.2018.2831280.
7. Sudharshan,P., Petitjean,C., Spanhol,F. *et al.* (2019) Multiple instance learning for histopathological breast cancer image classification. *Expert Syst. Appl.*, **117**, 103–111. 10.1016/j.eswa.2018.09.049.
8. Duggento,A., Conti,A., Mauriello,A. *et al.* (2020) Deep computational pathology in breast cancer. *Semin. Cancer Biol.*, 226–237
9. Benhammou,Y., Achchab,B., Herrera,F. *et al.* (2020) BreakHis based breast cancer automatic diagnosis using deep learning: taxonomy, survey and insights. *Neurocomputing*, **375**, 9–24. 10.1016/j.neucom.2019.09.044.
10. Sharma,S. and Mehra,R. (2020) Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *J. Digit. Imaging*, **33**, 632–654. 10.1007/s10278-019-00307-y.
11. Chugh,G., Kumar,S. and Singh,N. (2021) Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis. *Cogn. Comput.* **13**, 1451–1470.
12. Asif,A., Rajpoot,K., Snead,D. *et al.* (2021) Towards Launching AI Algorithms for Cellular Pathology into Clinical & Pharmaceutical Orbits, preprint, arXiv:211209496.
13. Janowczyk A. *et al* IDC: Invasive Ductal Carcinoma (2014) http://www.andrewjanowczyk.com/use-case-6-invasive-ductal-carcinoma-idc-segmentation/ (16 September 2022, date last accessed).
14. Bejnordi,B.E., Veta,M., Van Diest,P.J. *et al.* (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, **318**, 2199–2210. 10.1001/jama.2017.14585.
15. Spanhol,F.A., Oliveira,L.S., Petitjean,C. *et al.* (2015) A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.*, **63**, 1455–1462. 10.1109/TBME.2015.2496264.
16. Aresta,G., Araújo,T., Kwok,S. *et al.* (2019) Bach: grand challenge on breast cancer histology images. *Med. Image Anal.*, **56**, 122–139. 10.1016/j.media.2019.05.010.
17. Veta,M., Heng,Y.J., Stathonikos,N. *et al.* (2019) Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med. Image Anal.*, **54**, 111–121. 10.1016/j.media.2019.02.012.
18. Cui,M. and Zhang,D.Y. (2021) Artificial intelligence and computational pathology. *Lab. Invest.*, **101**, 412–422. 10.1038/s41374-020-00514-0.
19. Wahab,N., Miligy,I.M., Dodd,K. *et al.* (2022) Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. *J. Pathol.: Clin. Res.*, **8**, 116–128.
20. Haghighat,M., Browning,L., Sirinukunwattana,K. *et al.* (2021) PathProfiler: automated Quality Assessment of Retrospective Histopathology Whole-Slide Image Cohorts by Artificial Intelligence, A Case Study for Prostate Cancer Research., medRxiv.
21. Bankhead,P., Loughrey,M.B., Fernández,J.A. *et al.* (2017) QuPath: open source software for digital pathology image analysis. *Sci. Rep.*, **7**, 1–7. 10.1038/s41598-017-17204-5.
22. Dudgeon,S.N., Wen,S., Hanna,M.G. *et al.* (2021) A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. *J. Pathol. Inf.* **12**, 45.
23. Frattaruolo,A., Pallavera,L., Martella,E. *et al.* (2010) Flat epithelial atypia and atypical ductal hyperplasia: carcinoma underestimation rate. *The Breast J.*, **16**, 55–59. 10.1111/j.1524-4741.2009.00850.x.
24. Tan,P.H., Ellis,I. and Allison,K. Breast Tumors. (2010) WHO classification of tumours series. In: (5th edn.) Vol. 2 *WHO Classification of Tumors Editorial Board Lyon (France): international Agency for Research on Cancer 2019*.

25. Gonzalez R.S. *et al* Pathology Outlines. (2022) https://www.pathologyoutlines.com/breast.html (25 May 2022, date last accessed).
26. Araújo T. *et al* (2018) BACH: the Grand Challenge on BreAst Cancer Histology Images. https://iciar2018-challenge.grand-challenge.org/Dataset/ (16 September 2022, date last accessed).
27. Ludovic,R., Daniel,R., Nicolas,L. *et al*. (2013) Mitosis detection in breast cancer histological images An ICPR 2012 contest. *J. Pathol. Inf.* **4**, 8.
28. Polonia A. and Eloy C. Bioimaging. (2015), http://www.bioimaging2015.ineb.up.pt/dataset.html (16 September 2022, date last accessed).
29. Bejnordi B.E. *et al* Camelyon. (2016). https://camelyon16.grand-challenge.org/Data/ (16 September 2022, date last accessed).
30. Veta M. *et al* TUPAC16: Tumor proliferation assessment challenge (2016). https://tupac.grand-challenge.org/ (03 October 2022, date last accessed).
31. Geessink O. *et al* Camelyon. (2017). https://camelyon17.grand-challenge.org/ (16 September 2022, date last accessed).
32. Spanhol F. *et al* (2015). https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ (16 September 2022, date last accessed).
33. Lingle,W., Erickson,B., Zuley,M. *et al*. (2016) Radiology data from the cancer genome atlas breast invasive carcinoma [TCGA-BRCA] collection. *J. Pathol. Inf.*, **10**, K9.
34. Krug,K., Jaehnig,E.J., Satpathy,S. *et al*. (2020) Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell*, **183**, 1436–1456. 10.1016/j.cell.2020.10.036.
35. Brancati N. BRACS: BReAst Carcinoma Subtyping (2021), https://bracs.icar.cnr.it (16 September 2022, date last accessed).
36. Bussola,N., Marcolini,A., Maggio,V. *et al*. (2021) AI slipping on tiles: data leakage in digital pathology. In: *International Conference on Pattern Recognition Springer*. pp. 167–182.
37. Pati,P., Jaume,G., Alisha Fernandes,L. *et al*. (2020) HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop on GRaphs in biomedicAl Image anaLysis*. Springer Nature Switzerland, Basel.
38. Pati,P., Jaume,G., Foncubierta,A. *et al*. (2021) Hierarchical Graph Representations in Digital Pathology. *Med. Image Anal.* **75**, 102264.
39. Jaume,G., Pati,P., Foncubierta-Rodriguez,A. *et al* . (2020) Towards explainable graph representations in digital pathology. In: ICML Workshop on Computational Biology, 1–5. arXiv preprint arXiv:2007.00311
40. Jaume,G., Pati,P., Bozorgtabar,B. *et al*. (2021) Quantifying Explainers of Graph Neural Networks in Computational Pathology. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.
41. Brancati,N., De Pietro,G., Riccio,D. *et al*. (2021) Gigapixel Histopathological Image Analysis Using Attention-Based Neural Networks. *IEEE Access*, **9**, 87552–87562. 10.1109/ACCESS.2021.3086892.
42. Macenko,M., Niethammer,M., Marron,J.S. *et al*. (2009) A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: from Nano to Macro IEEE*. IEEE Computer Society, pp. 1107–1110.