VALIDATED MACHINE LEARNING MODELS FOR BREAST CANCER

PERFORMANCE OF EXTERNALLY VALIDATED MACHINE LEARNING MODELS BASED ON HISTOPATHOLOGY IMAGES FOR THE DIAGNOSIS, CLASSIFICATION, PROGNOSIS, OR TREATMENT OUTCOME PREDICTION IN FEMALE BREAST CANCER: A SYSTEMATIC REVIEW

By RICARDO GONZALEZ, MD, MPH

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for

the Degree Master of Science in eHealth

McMaster University MASTER OF SCIENCE (2022) Hamilton, Ontario (eHealth)

TITLE: Performance of externally validated machine learning models based on histopathology images for the diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer: A systematic review AUTHOR: Ricardo Gonzalez, MD, MPH (McMaster University) SUPERVISOR: Dr. Cynthia Lokker NUMBER OF PAGES: xi, 84

Lay Abstract

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer deaths in women. Microscopic analysis of tissues taken from the breast is the standard method for diagnosis. Using digital images of these tissues, researchers have been training computer software to identify and classify breast cancer and to predict future behavior and response to treatments. These computer algorithms are called “machine learning models.” It is important to test how well machine learning models perform with new images—ones that were not used during the development of the models and differ from the development data in some aspect such that they can be considered independent from the development data and process (external data). This systematic review looks at the performance of machine learning models that used microscopic pictures of breast cancer and were tested with external data.

Abstract

Background: Numerous machine learning (ML) models have been developed for breast cancer using various types of data (e.g., images, text). Successful external validation (EV) of ML models is considered as important evidence of their generalizability.

Objectives: Assess the performance of externally validated ML models based on histopathology images for diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer.

Methods: A systematic search of MEDLINE, EMBASE, CINAHL, IEEE, MICCAI, and SPIE conferences was performed for studies published between January 2010 and February 2022. The Prediction Model Risk of Bias Assessment Tool (PROBAST) was employed, and the results were narratively described.

Results: Of the 2339 retrieved citations, eight journal articles and two conference proceedings met inclusion criteria. Three studies externally validated ML models for diagnosis, four for classification, two for prognosis, and one for both classification and prognosis. Most studies used Convolutional Neural Networks and one used logistic regression algorithms. For diagnostic/classification models, the most common performance metrics reported in the EV were accuracy and area under the curve, which were above 87% and 90%, respectively, using pathologists' annotations/diagnoses as ground truth. The hazard ratios in the EV of prognostic ML models were between 1.7 (95% CI, 1.2–2.6) and 1.8 (95% CI, 1.3–2.7) to predict distant disease‑free survival; 1.91 (95% CI, 1.11-3.29) for recurrence, and between 0.09 (95% CI, 0.01–0.70) and 0.65 (95% CI, 0.43–0.98) for overall survival, using clinical data as ground truth.

Conclusion: Despite EV being an important step before the clinical application of a ML model, it hasn't been performed routinely. The large variability in the training/validation datasets, methods, performance metrics, and reported information limited the comparison of the models and the analysis of their results. Increasing the availability of validation datasets and implementing standardized methods and reporting protocols may facilitate future analyses.

Acknowledgements

I want to thank the following people, without whom I would not have been able to complete this systematic review and finish this master's degree program.

Dr. Cynthia Lokker, not only for providing guidance and permanent feedback throughout this project as a supervisor but also for her patience and unconditional support. Dr. Peyman Nejat, for his time and thoughtful contributions during the Title/Abstract and full/text screening processes and risk of bias/applicability assessments. Dr. Ashirbani Saha and Dr. Clinton Campbell, members of my thesis committee, for their time and for providing clear, constructive, and insightful commentaries to improve the original thesis proposal and the final manuscript. Dr. Andrew P. Norgan, for sharing with me his thoughts about some of the topics discussed in this thesis. Finally, my deepest gratitude to my family for all their support and for being my deepest source of motivation.

Contents

[1. Title 1](#_Toc130050908)

[2. Abstract 1](#_Toc130050909)

[3. Introduction 3](#_Toc130050910)

[3.1 Rationale 3](#_Toc130050911)

[3.2 Objective 8](#_Toc130050912)

[4. Methods 9](#_Toc130050913)

[4.1 Eligibility criteria 9](#_Toc130050914)

[4.2 Information sources 10](#_Toc130050915)

[4.3 Search strategy 10](#_Toc130050916)

[4.4 Selection process 10](#_Toc130050917)

[4.5 Data collection process 11](#_Toc130050918)

[4.6 Data collection 11](#_Toc130050919)

[4.7 Study risk of bias assessment 12](#_Toc130050920)

[4.8 Effect Measures 12](#_Toc130050921)

[4.9 Synthesis methods 12](#_Toc130050922)

[4.10 Reporting bias assessment 13](#_Toc130050923)

[4.11 Certainly assessment 13](#_Toc130050924)

[5. RESULTS 13](#_Toc130050925)

[5.1 Study selection 13](#_Toc130050926)

[5.2 Study characteristics 15](#_Toc130050927)

[5.2.1 ML models for diagnostic purposes 16](#_Toc130050928)

[5.2.2 ML models for classification purposes 18](#_Toc130050929)

[5.2.3 ML models for prognosis purposes 22](#_Toc130050930)

[5.2.4 ML model for classification and prognosis purposes 24](#_Toc130050931)

[5.3 Risk of bias in studies 26](#_Toc130050932)

[5.4 Results of individual studies 28](#_Toc130050933)

[5.4.1 ML for diagnostic purposes: 28](#_Toc130050934)

[5.4.2 ML for classification purposes: 29](#_Toc130050935)

[5.4.3 ML for prognostic purposes: 30](#_Toc130050936)

[5.4.4 ML for classification and prognosis purposes: 31](#_Toc130050937)

[6. DISCUSSION 32](#_Toc130050938)

[6.1 General interpretation of the results in the context of other evidence 32](#_Toc130050939)

[6.2 Limitations of the evidence included in the review 36](#_Toc130050940)

[6.3 Limitations of the current review 37](#_Toc130050941)

[6.4 Implications of the results for practice, policy, and future research 38](#_Toc130050942)

[7. CONCLUSION 45](#_Toc130050943)

[8. OTHER INFORMATION 45](#_Toc130050944)

[8.1 Registration and protocol 45](#_Toc130050945)

[8.2 Support 45](#_Toc130050946)

[8.3 Competing interests 46](#_Toc130050947)

[8.4 Availability of data, code, and other materials 46](#_Toc130050948)

[9. REFERENCES 47](#_Toc130050949)

[Appendix 1. Search strategies. 74](#_Toc130050950)

Lists of Figures and Tables

* Lists of Figures:

[Figure 1. ML model basic development steps 5](https://mctools-my.sharepoint.com/personal/gonzalez_ricardo_mayo_edu/Documents/RG/Ricardo%20Gonzalez%20-%20Thesis%20-%20Final%20version%20-%20Edited%20RG.docx#_Toc130050980)

[Figure 2. PRISMA flow diagram of the studies identification process for the systematic review 14](#_Toc130050981)

[Figure 3. Prediction model Risk Of Bias Assessment Tool (PROBAST) Graphical presentation – (1) Risk of Bias results and (2) Applicability 27](#_Toc130050982)

[Figure 4. ML models site-specific iterative development steps 40](https://mctools-my.sharepoint.com/personal/gonzalez_ricardo_mayo_edu/Documents/RG/Ricardo%20Gonzalez%20-%20Thesis%20-%20Final%20version%20-%20Edited%20RG.docx#_Toc130050983)

* Lists of Tables:

[Table 1. Glossary of terms. 6](#_Toc130051062)

[Table 2. Included studies: Purpose, authors (Publication year), country, and publication type 15](#_Toc130051063)

[Table 3 Datasets, sources, and preprocessing steps for development and validation of diagnostic models 17](#_Toc130051064)

[Table 4. Datasets, sources, and preprocessing steps for development and validation of classification models. 20](#_Toc130051065)

[Table 5. Datasets, sources, and preprocessing steps for development and validation of prognosis models for overall survival or distant disease‑free survival 23](#_Toc130051066)

[Table 6. Datasets, sources, and preprocessing steps for development and validation of DeepGrade 25](#_Toc130051067)

[Table 7. Risk of bias assessment (PROBAST) 26](#_Toc130051068)

[Table 8. Results of diagnostic models 28](#_Toc130051069)

[Table 9. Results of classification models 29](#_Toc130051070)

[Table 10. Results of prognostic models for predicting survival 30](#_Toc130051071)

* List of all Abbreviations and Symbols

|  |  |
| --- | --- |
| Abbreviation | Definition  |
| AUC | Area Under the Curve |
| CNN | Convolutional Neural Network |
| DCIS | Ductal Carcinoma in situ |
| DP | Digital Pathology  |
| EV | External Validation |
| FPR | False Positive Rate |
| H&E | Hematoxylin and Eosin  |
| HR | Hazard Ratios |
| ML | Machine Learning  |
| NHG | Nottingham HistologicalG.  |
| NPV | Negative Predictive Value |
| PPV | Positive Predictive Value |
| TIL | Tumor-Inﬁltrating Lymphocyte.  |
| TMA | Tissue Microarray |
| TNR | True Negative Rate |
| TPR | True Positive Rate  |
| WSIs | Whole Slide Images |

k) Declaration of Academic Achievement

I, Ricardo Gonzalez, declare this thesis to be my own work.

I am the sole author of this document. No part of this work has been published or submitted for publication or for a higher degree at another institution.

To the best of my knowledge, the content of this document does not infringe on anyone's copyright.

My supervisor, Dr. Cynthia Lokker, and the members of my supervisory committee, Dr. Ashirbani Saha and Dr. Clinton Campbell, have provided guidance and support at all stages of this project.

Dr. Peyman Nejat acted as an independent reviewer during the Title/Abstract and Full-text screening processes, verified the data extracted from the included studies, and contributed to the risk of bias/applicability assessment.

I completed the rest of the research work.

# Title

Performance of externally validated machine learning models based on histopathology images for the diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer: A systematic review

# Abstract

Background: Numerous machine learning (ML) models have been developed for breast cancer using various types of data (e.g., images, text). Successful external validation (EV) of ML models is considered as important evidence of their generalizability.

Objectives: Assess the performance of externally validated ML models based on histopathology images for diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer.

Methods: A systematic search of MEDLINE, EMBASE, CINAHL, IEEE, MICCAI, and SPIE conferences was performed for studies published between January 2010 and February 2022. The Prediction Model Risk of Bias Assessment Tool (PROBAST) was employed, and the results were narratively described.

Results: Of the 2339 retrieved citations, eight journal articles and two conference proceedings met inclusion criteria. Three studies externally validated ML models for diagnosis, four for classification, two for prognosis, and one for both classification and prognosis. Most studies used Convolutional Neural Networks and one used logistic regression algorithms. For diagnostic/classification models, the most common performance metrics reported in the EV were accuracy and area under the curve, which were above 87% and 90%, respectively, using pathologists' annotations/diagnoses as ground truth. The hazard ratios in the EV of prognostic ML models were between 1.7 (95% CI, 1.2–2.6) and 1.8 (95% CI, 1.3–2.7) to predict distant disease‑free survival; 1.91 (95% CI, 1.11-3.29) for recurrence, and between 0.09 (95% CI, 0.01–0.70) and 0.65 (95% CI, 0.43–0.98) for overall survival, using clinical data as ground truth.

Conclusion: Despite EV being an important step before the clinical application of a ML model, it hasn't been performed routinely. The large variability in the training/validation datasets, methods, performance metrics, and reported information limited the comparison of the models and the analysis of their results. Increasing the availability of validation datasets and implementing standardized methods and reporting protocols may facilitate future analyses.

# Introduction

## Rationale

Worldwide, cancer is a leading cause of death, and it is expected to limit increases in life expectancy in most countries in future decades (Bray et al., 2021). Female breast cancer is the most commonly diagnosed cancer and, in women, the most frequent cause of cancer mortality (Sung et al., 2021). Histopathological examination of breast tissue samples is the reference standard for cancer diagnosis and is used to determine the prognosis of a patient and risk factors to predict outcomes (Khened et al., 2021; Leong & Zhuang, 2011).

Pathologists have been using a microscope to make diagnoses for more than 100 years. Now, these slides can be scanned to be viewed on computer screens (Nam et al., 2020; Randell et al., 2015). The process of scanning glass slides to produce digital images (Whole-Slide Images or WSI) was initially called "Digital Pathology" (DP) (Nam et al., 2020). However, over the last decades, this term has evolved and is now used to describe many other related processes (Nam et al., 2020; Randell et al., 2015). According to the Digital Pathology Association, DP is "A blanket term that encompasses tools and systems to digitize pathology slides and associated meta-data, their storage, review, analysis, and enabling infrastructure" (Abels et al., 2019).

DP has been approved and implemented in laboratories for routine diagnosis in many countries and is expected to increase in the near future (Chong et al., 2020; Têtu & Evans, 2014). Beyond the potential benefits in improving workflow efficiency and enhancing diagnostic accuracy, DP facilitates the implementation of artificial intelligence (AI)-based tools for pathologists (Azam et al., 2021).

AI is a branch of computer science concerned with understanding and building intelligent entities (i.e., machines able to adapt to new situations) (Russell & Norvig, 2021; K. H. Yu et al., 2018). Machine Learning (ML) is a subﬁeld of AI that attempts to generate models that learn to make predictions on new data based on experience (Radakovich et al., 2020; Russell & Norvig, 2021). Three main types of ML are recognized: Supervised, unsupervised, and reinforcement learning. In supervised learning, models are trained with input data and output variables. They learn mapping functions (i.e., functions that map from input to output). They aim to predict the output variables when new input data is given (Russell & Norvig, 2021). In unsupervised learning, models are trained with input data and no corresponding output variables (i.e., no explicit feedback). They aim to learn patterns in the input data (Russell & Norvig, 2021). In reinforcement learning, models are trained by rewarding desired outputs and punishing the undesired ones. They "learn" by trying to maximize the number of future rewards (Russell & Norvig, 2021).

The followed when developing a ML model (Figure 1) include:

1. **Problem formulation: identifying and understanding a problem that needs a solution and specifying the part(s) of the problem suitable to be solved with a ML algorithm (Russell & Norvig, 2021).
2. ML algorithm selection: The type of algorithm is chosen based on the problem and data available (Radakovich et al., 2020).
3. Data preparation: Data needed to train the model is selected, gathered, and preprocessed (Radakovich et al., 2020). Preprocessing steps that might be used are de-identification,

Figure 1. ML model basic development steps

feature engineering and normalization (Norgeot et al., 2020).

1. Model training: Using an iterative process to reduce the error in its predictions or in attainment of its objectives, ML model's parameters (i.e., variables whose values can be learned) are defined (Maleki et al., 2020).
2. Hyperparameter tuning: The optimal set of hyperparameters (e.g., number of hidden layers, number of neurons in each layer, learning rates) are selected (Maleki et al., 2020).
3. Model evaluation: The performance of the model is validated/tested (Maleki et al., 2020). As discussed below, this can be done internally or externally.
4. Model deployment/maintenance: The ﬁnal model is packaged appropriately, deployed, monitored, and maintained (Radakovich et al., 2020; Russell & Norvig, 2021).

The terms "validation" and "test" are most commonly used interchangeably in the ML literature when mentioning model evaluation (Maleki et al., 2020). However, the word "validation" is also frequently utilized to describe data used for hyperparameter tuning ("validation dataset") (Maleki et al., 2020) and/or to select the ML model (Homeyer et al., 2022). The lack of standardized terminology may confuse researchers in the ML community (Maleki et al., 2020). Here, the following terms are used:

Table 1. Glossary of terms.

|  |  |
| --- | --- |
| Terms | Definitions |
| Internal validation  | Model evaluation conducted with data extracted from the input dataset (Ho et al., 2020; Nagendran et al., 2020; Norgeot et al., 2020; Park et al., 2021) |
| EV | Model evaluation conducted with data extracted from independent datasets (Ho et al., 2020; Nagendran et al., 2020; Norgeot et al., 2020; Park et al., 2021) |
| Training/Tuning datasets | Used for model training, optimization, and/or selection (Norgeot et al., 2020) |
| Internal validation dataset | Created with data extracted from an input dataset (Ho et al., 2020; Nagendran et al., 2020; Park et al., 2021) that was set aside from the training/tuning dataset at the beginning of the study to evaluate the final version of a ML model a single time (Norgeot et al., 2020) |
| EV dataset | Created with data extracted from an independent dataset (Ho et al., 2020; Nagendran et al., 2020; Park et al., 2021) to evaluate the final version of a ML model a single time (Norgeot et al., 2020) |
| EV: External validation |

Cross-validation, bootstrapping, and split-sample (two-way splits or three-way splits) are commonly used for internal validation and can be considered an industry standard. The independent datasets used during EV should be ideally extracted from a different setting or source; for example, another clinic or hospital system (Ho et al., 2020; Nagendran et al., 2020; Norgeot et al., 2020; Park et al., 2021). Good performance in EV is considered proof of model generalizability (Ho et al., 2020; Norgeot et al., 2020). In clinical practice, predictive models are commonly used. Since safety is a key consideration, ML models implemented in this setting should provide estimates of the uncertainty associated with model results, be interpretable and explainable, and achieve good performance in EV (Nwanosike et al., 2021; Russell & Norvig, 2021).

Recent advances in ML methods, increasing computer power, and unprecedented amounts of data are reshaping cancer care. Numerous ML models have been developed for a range of tasks, such as cancer risk prediction, detection, diagnosis, classification, grading, staging, treatment selection, prognosis, treatment response-prediction, treatment discovery, and patient follow-up (Bhinder et al., 2021; Kann et al., 2021; Luchini et al., 2022). However, most of these models have not been externally validated, limiting their generalization and implementation in different clinical settings (Adeoye et al., 2021; Akazawa & Hashimoto, 2021; Bang et al., 2021; Nagendran et al., 2020; Nwanosike et al., 2021; Park et al., 2021; Shelmerdine et al., 2021; Shi et al., 2021; Twilt et al., 2021). It is important to note that DP is still in the early phase of adoption, and therefore a key bottleneck to EV is the lack of widely available and well annotated DP datasets (Mazo et al., 2022; K. H. Yu et al., 2018). Further, standardization of methods to acquire and annotate DP has not been achieved (Champion et al., 2014; Lindman et al., 2019).

To the best of our knowledge, a systematic review assessing the performance of externally validated ML models based on histopathology images for diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer have not been published. This study intends to fill this gap in the literature.

## Objective

The aim of this systematic review was to assess the performance of externally validated ML models based on histopathology images for diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer.

# Methods

## Eligibility criteria

Studies that externally validated the performance of ML models, using any ML methodology, for diagnosis, classification, prognosis, or treatment response prediction of female breast cancer using data extracted directly from histopathology images were included.

Studies were selected according to the criteria outlined below.

* Date of publication: From January 1, 2010, to February 28, 2022
* Language: English
* Species: Humans
* Sex: Female
* Ages: All
* Article types: All original research journal papers and conference proceedings from IEEE, SPIE conferences and MICCAI.
* ML models:
	+ For diagnosis, classification, prognosis, or treatment outcome prediction. Focused on female breast cancer (invasive tumors or carcinomas in situ).
	+ Using data extracted directly from images of histopathology slides (stained with hematoxylin/eosin or other histochemical stains).
	+ Externally validated with breast cancer images: performance was tested with breast cancer images extracted from independent datasets. This EV must be mentioned or implied in the Title/Abstract.
	+ At least one performance metric reported.

Exclusion criteria:

* + ML models used to predict biomarkers.

## Information sources

Searches were performed in MEDLINE via OVID, EMBASE via OVID, CINAHL via EBSCO, IEEE via IEEE Xplore, MICCAI via Springer link, and SPIE conferences. Search strategies incorporated relevant text words in all databases and database taxonomies, e.g., Medical Subject Headings (MeSH) in MEDLINE, Subject Headings in EMBASE, and Exact Subject Headings in CINAHL.

## Search strategy

No study type or sex limits were applied. The search strategies were developed by authors and reviewed by a health sciences librarian with expertise in systematic reviews. The search strategies are shown in Appendix 1. All the searches were conducted in March, 2022.

## Selection process

The authors developed, tested, and refined screening questions based on the eligibility criteria. Search results were imported into Rayyan to remove duplicates and for title/abstract screening. Two reviewers independently conducted the initial title and abstract screening. Selected full-text articles were downloaded for all titles that met the inclusion criteria or where there was any uncertainty. Two reviewers independently conducted the full-text screening using Mendeley. Additional information was requested from study authors when needed. A third person acted as an adjudicator to resolve any conflicts.

## Data collection process

One author independently extracted data from included studies with verification by a second. Articles were separated into two groups: 1) ML models related to diagnosis/classification and 2) ML models related to prognosis/treatment outcome prediction.

## Data collection

The following information was extracted: Authors, publication date, country of study, objectives, information related to the ML models, and main results, including all performance measures in the EV of the ML models. The following information related to the ML models was extracted from all eligible studies: Algorithms employed; source, number, and type of images used for algorithm development (training/tuning and internal validation datasets) and EV; histological type of lesions/tumors contained in the training and EV datasets; details on preprocessing of the images; hardware/software platform(s) noted for annotating/computational purposes and ground truth (which is captured with the gold standard). From studies for prognosis or that predicted treatment outcomes, the following data were also extracted: study design, outcomes, duration of follow-up, and treatment details (if applicable).

## Study risk of bias assessment

The risk of bias by using PROBAST (prediction model risk of bias assessment tool) for non-randomized studies was assessed by one author and confirmed by another.

## Effect Measures

The effect measures were all performance metrics, such as accuracy, Area Under the Curve (AUC), Precision, Recall, etc used to evaluate the ML model during EV.

## Synthesis methods

The main study characteristics and findings are presented in the text and summarized in tables. Results are narratively described. A meta-analysis was not conducted because of the likely heterogeneity of the studies in algorithms, type of images, and reported outcomes.

## Reporting bias assessment

The risk of bias by using PROBAST (prediction model risk of bias assessment tool) for non-randomized studies was assessed by one author and confirmed by another.

## Certainly assessment

 The authors did not assess the confidence in cumulative evidence.

## RESULTS

## Study selection

The search queries identified 2157 articles and 182 conference proceedings. After removing 328 duplicates and excluding 1961 publications during the title and abstract screening, 50 were assessed during full-text screening, and 10 were included in the review (Figure 1).

Figure 2. PRISMA flow diagram of the studies identification process for the systematic review



## Study characteristics

Three studies externally validated ML models used for diagnosis, four for classification, two for prognosis, and one for both classification and prognosis. Eight were published as journal articles, and two as conference proceedings. Two in 2017, two in 2018, one in 2019, one in 2020, three in 2021, and one in 2022.

Three studies were conducted in Colombia and the United States, two in the United States, two in China, one in Sweden, one in Sweden and the United States, and one in Finland (Table 1).

Table 2. Included studies: Purpose, authors (Publication year), country, and publication type

|  |  |  |  |
| --- | --- | --- | --- |
| Purpose | Authors (year) | Country of study  | Publication Type |
| Diagnosis | Cano et al. (2018) | Colombia & United States  | Conference proceeding |
| Cruz-Roa et al. (2017) | Colombia & United States  | Journal article |
| Cruz-Roa et al. (2018) | Colombia & United States  | Journal article |
| Classification | Colon-Cartagena et al. (2020)  | United States  | Conference proceeding |
| Mi et al. (2021)  | China | Journal article |
| Radiya-Dixit et al. (2017) | United States  | Journal article |
| Yang et al. (2019) | China | Journal article |
| Prognosis | Bai et al. (2021) | Sweden & United States  | Journal article |
| Bychkov et al. (2022) | Finland | Journal article |
| Classification & Prognosis | Wang et al. (2021) | Sweden | Journal article |

A detailed description of the studies included in this systematic review is found below. To facilitate their comparison, they are grouped based on their purpose (i.e., for diagnosis, classification, prognosis, or classification & prognosis purposes).

## ML models for diagnostic purposes

In one conference proceeding reported by Cano et al. (Cano et al., 2018) and two journal articles published by Cruz-Roa (Cruz-Roa et al., 2017, 2018), ML models for diagnostic purposes were externally validated. Cano et al. aimed to detect invasive ductal carcinomas (Cano et al., 2018) and Cruz-Roa et al. estrogen receptor-positive (ER+) invasive breast cancer (Cruz-Roa et al., 2017, 2018).

Cano et al. utilized cases from the Hospital of the University of Pennsylvania (HUP) and The Cancer Genome Atlas (TCGA) (Table 2). They conducted experiments using cases from one dataset for training/tuning purposes and from the other for internal validation and vice versa. The EV was performed with cases from New Jersey Cancer Institute (CINJ) (Cano et al., 2018). In contrast, Cruz-Roa et al. utilized cases from HUP and Case Western Reserve University (CWRU) (Cruz-Roa et al., 2017) or CWRU/ University Hospitals Case Medical Center (UHCMC) (Cruz-Roa et al., 2018) for training/ tuning purposes, from CINJ for internal validation and from TCGA/CWRU/UHCMC (Cruz-Roa et al., 2017) or TCGA only (Cruz-Roa et al., 2018) for EV. A detailed description of the datasets is shown in Table 3.

Table 3 Datasets, sources, and preprocessing steps for development and validation of diagnostic models

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cano et al. (2018) | Cruz-Roa et al. (2017) | Cruz-Roa et al. (2018) |
| Datasets: |
| Type of lesions / tumors | Invasive ductal carcinomas | ER+ invasive breast cancer  | ER+ invasive breast cancer  |
| Training / tuning datasets: source (n)  | HUP§ (239)TCGA§ (172) | HUP (239)CWRU/UHCMC (110) | HUP (239)CWRU (110) |
| IV datasets: source (n) | TCGA§ (172) HUP§ (239)  | CINJ (40)  | CINJ (40)  |
| EV datasets: source (n) | CINJ (40)  | TCGA (195)CWRU/ UHCMC (21) §§ | TCGA (195) |
| Preprocessing steps: |
| Color normalization | NS | Yes | Yes  |
| Data augmentation | NS | Yes | Yes |
| CINJ: New Jersey Cancer Institute. CWRU: Case Western Reserve University. ER+: Estrogen receptor-positive. EV: External validation. HUP: Hospital of the University of Pennsylvania. IV: Internal validation. NS = Not specified. TCGA: The Cancer Genome Atlas. UHCMC: University Hospitals Case Medical Center. § Internal validation with cases from the TCGA was conducted after the models were trained with cases from HUP, and internal validation with cases from HUP was conducted after the models were trained with cases from the TCGA. §§ The test data set used by Cruz-Roa et al. (2017) included positive and negative controls. Positive controls were extracted from the TCGA. Negative controls were extracted from normal breast tissue regions adjacent to invasive ductal carcinomas of patients diagnosed at UHCMC/CWRU. It is unknown if regions with invasive ductal carcinomas extracted from the same patients were included in the training dataset.  |

While Cruz-Roa et al. utilized patches of 200 x 200 μm (Cruz-Roa et al., 2017) or 101 x 101 pixels (Cruz-Roa et al., 2018) extracted from WSIs scanned with 40x magnification for training purposes, Cano et al. used patches of 50 x 50 pixels extracted with 20x magnification (Cano et al., 2018). The software platforms used for annotation purposes by Cruz-Roa et al. were ImageScope v11.2 (Aperio) and Image Viewer v3.1.4 (Ventana) (Cruz-Roa et al., 2017, 2018). Cano et al. did not specify those.

After comparing different CNNs, Cano et al. found that one with nine layers had the best performance. It was composed of three convolutional-pooling layers of sixteen neurons per layer, three convolutional-pooling layers of thirty-two neurons per layer, one fully connected layer of sixteen neurons, one fully connected layer of thirty-two neurons, and one sigmoid classiﬁcation layer (Cano et al., 2018). Cruz-Roa et al. used CNNs with one convolutional and pooling layer, one fully connected layer of 256 neurons, and a classification layer (Cruz-Roa et al., 2017, 2018).

## ML models for classification purposes

In one conference proceeding, Colon-Cartagena et al. externally validated a ML to differentiate several histologic variants of high-grade DCIS (i.e., Comedo-type, Cribriform, Micropapillary, and Solid) (Colon-Cartagena et al., 2020). In two journal articles, Mi et al. and Yang et al. validated ML models to differentiate benign from malignant lesions of the breast (Mi et al., 2021; Yang Z. et al., 2019), and in another journal article, Radiya-Dixit E. et al. validated a ML model to differentiate Ductal Carcinoma in situ (DCIS) from Usual Ductal Hyperplasia (UDH) (Radiya-Dixit E. et al., 2017).

To train and internally validate their models, Colon-Cartagena et al. utilized cases from the Departmental archives of Virginia Commonwealth University (Colon-Cartagena et al., 2020), Mi et al. from the Peking Union Medical College Hospital (Mi et al., 2021), Radiya-Dixit et al. from the Massachusetts General Hospital and Beth Israel Deaconess Medical Center (Radiya-Dixit E. et al., 2017), and Yang et al. from the BACH dataset (Yang Z. et al., 2019). To externally validate them, Mi et al. used cases from BreakHis dataset and BACH (Mi et al., 2021), Radiya-Dixit et al. from Beth Israel Deaconess Medical Center (Radiya-Dixit E. et al., 2017), and Yang et al. from BreakHis (Yang Z. et al., 2019). Colon-Cartagena et al. did not specify the source of the EV dataset (Colon-Cartagena et al., 2020). The size of the datasets is shown in Table 4.

Table 4. Datasets, sources, and preprocessing steps for development and validation of classification models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Colon-Cartagena et al. (2020) | Mi et al. (2021) | Radiya-Dixit et al. (2017) | Yang et al. (2019) |
| Datasets: |
| Type of lesions / tumors | High-grade DCIS | Training/IV: Normal, benign lesions, DCIS, and invasive carcinomas  | DCIS and UDH | Training/IV: Benign and malignant lesions from BACH\*\* dataset  |
| EV: benign and malignant lesions from BreakHis\* and BACH\*\* datasets | EV: Benign and malignant lesions from BreakHis\* dataset |
| Training / tuning datasets: source (n) | VCU (334) | PUMCH (540) | MGH (116)§ | BACH (400) |
| IV datasets: source (n) | VCU (80)  | PUMCH (540) | MGH (116)BIDMC (51)§ | BACH (100) |
| EV datasets: source (n) | NS (31) | BreakHis (7909) BACH (430) | BIDMC (51)§ | BreakHis (1995) |
| Preprocessing steps: |
| Color normalization | NS | NS | NS | NS |
| Data augmentation | NS | Yes | NS | Yes |
| Other | NS | - Image resizing- Random color perturbations  | NS | NS |
| BIDMC: Beth Israel Deaconess Medical Center. DCIS: Ductal Carcinoma in situ. EV: External validation. IV: Internal validation. MGH: Massachusetts General Hospital. NS: Not specified. PUMCH: Peking Union Medical College Hospital. UDH: Usual Ductal Hyperplasia. VCU: Virginia Commonwealth University. \* BreakHis dataset: (1) Benign lesions: Adenosis, Fibroadenomas, Phyllodes tumors, and Tubular adenoma. (2) Malignant lesions: Ductal Carcinoma, Lobular carcinoma, Mucinous carcinoma, and Papillary carcinoma. \*\* BACH dataset: Normal, benign, carcinomas in situ, and invasive carcinomas.§ External validation performed with cases from the BIDMC after training the model with cases from the MGH. All cases from both institutions were combined to train and test the model when conducting the internal validation (using cross-validation). |

For training and internal validation purposes, Mi et al. used patches of 1024 x 1024 pixels extracted from WSIs scanned with 40x magnification, and for EV, patches of 2048 x 1536 pixels extracted with 4x, 10x, 20x, and 40x magnifications (Mi et al., 2021). Yang et al. utilized images of 2048 x 1536 pixels scanned with 20x magnification for training/internal validation and of 700 x 460 pixels scanned with 40x magnification for EV (Yang Z. et al., 2019). The preprocessing steps reported by them are shown in Table 4. Colon-Cartagena et al. and Radiya-Dixit et al. did not provide this information.

Mi et al. used ASAP for annotation purposes (Mi et al., 2021), and Radiya-Dixit et al. used Fiji (ImageJ, National Institutes of Health) for nuclei segmentation. Colon-Cartagena et al. or Radiya-Dixit et al. did not mention the software platforms used for annotation or computational purposes (Radiya-Dixit E. et al., 2017).

Regarding the algorithms, Colon-Cartagena et al. and Yang et al. used ResNet models. In addition to ResNet-50, Colon-Cartagena et al. utilized a coding-free image classifier developed with the IBM Watson Visual Recognition platform (Colon-Cartagena et al., 2020). The latter combined three fine-tuned CNNs that employed ResNet-101, ResNet-152, and DenseNet-161 as backbone networks (Yang Z. et al., 2019). Mi et al. added a fully connected layer (with 1024 neurons) and a Softmax layer to the basic model of Inception V3 and used it as a patch-level classifier. Radiya-Dixit et al. employed a “Combined model with Active Feature Extraction – CAFE,” based on two logistic regression algorithms.

## ML models for prognosis purposes

In two journal articles, prognostic models were externally validated. One model aimed to predict the overall survival in patients (Bai et al., 2021), and the other the distant disease‑free survival (Bychkov et al., 2022).

Bai et al. trained a model with cases from the Pathology Department of Yale School of Medicine to predict the overall survival of patients with triple-negative breast cancer with five model-derived tumor-infiltrating lymphocytes (TILs) variables. The overall survival was deﬁned as the time between primary diagnosis of the tumor to death (any cause) or date of last censoring if still alive. To externally validate their model, they retrospectively collected information from other patients in the Pathology Department of Yale School of Medicine (one cohort diagnosed between 1962 and 2006 with a median follow-up of 63.8 months and another cohort diagnosed between 1981 and 2012 with a median follow-up of 64.8 months), the TCGA (who were operated between 1996 and 2013 and with 13 months of median follow-up), and the Swedish National Breast Cancer Quality Registry - WTS Sweden (from patients enrolled in the "SCAN-B study" between 2010 and 2015 with a median follow-up of 49.7 months) (Bai et al., 2021).

Bychkov et al. trained and internally validated a model to predict distant disease‑free survival in patients with breast cancer (histological subtypes not specified) using tissue microarrays from the FinProg series. Distant disease‑free survival was time from randomization to detection of distant metastasis. They externally validated their model with data extracted from the FinHer trial (multicenter randomized clinical trial) conducted between 2000 and 2003 with women (up to 65 years old) who had axillary lymph node‑positive breast cancer or a high‑risk (node‑negative) cancer and were treated with docetaxel or vinorelbine followed by fluorouracil, epirubicin, and cyclophosphamide. HER2+ patients were also randomly assigned to receive concomitant trastuzumab or not (Bychkov et al., 2022). Additional details of the datasets are found in Table 5.

Table 5. Datasets, sources, and preprocessing steps for development and validation of prognosis models for overall survival or distant disease‑free survival

|  |  |  |
| --- | --- | --- |
|  | Bai et al. (2021) | Bychkov et al. (2022) |
| Datasets: |
| Type of lesions / tumors | Triple-negative breast cancer | Breast cancer |
| Training / tuning datasets: source (n)  | Yale School of Medicine (95)§ | FinProg (693) |
| IV datasets: source (n) | Yale School of Medicine (171)§ | FinProg (354) |
| EV datasets: source (n) | Yale School of Medicine (417)§WTS Sweden (216)TCGA (116) | FinHer trial (712) |
| Preprocessing steps: |
| Color normalization | Yes | Yes |
| Data augmentation | NS | Yes |
| EV: External Validation. IV: Internal Validation. NS: Not Specified. TCGA: The Cancer Genome Atlas. § Training, IV and EV datasets contained cases from different cohorts of patients of the Yale School of Medicine.  |

For training and internal validation, Bai et al. utilized WSIs scanned at 20x (with a pixel size of 0.4986 mm x 0.4986 mm for images of the Yale cohorts and 0.4537 x 0.4537 mm for those from the WTS Sweden). For EV, they used WSIs from the Yale cohorts with the same scanning characteristics and from the TCGA (other details not specified) (Bai et al., 2021). Bychkov et al. used center crops (of 2100 x 2100 pixels) of Tissue Microarrays (TMA) during the training and internal validation processes and extracted patches (of 950 x 950 pixels) from WSIs for testing purposes (scanning magnifications not specified) (Bychkov et al., 2022). The preprocessing steps are shown in Table 6.

QuPath (version 0.1.2) was used by Bai et al. to build the automated TIL scoring algorithm (Bai et al., 2021). Bychkov et al. used WebMicroscope (Aiforia Technologies Oy, Helsinki, Finland) to segment individual TMA images from WSIs. They also used PyTorch (Facebook's AI Research lab‑FAIR) to implement deep learning architectures and Adam as an optimization algorithm for model training (Bychkov et al., 2022).

Regarding the algorithms, Bai et al. trained a CNN with eight hidden layers for cell classification (Bai et al., 2021). Bychkov et al. trained CNNs using ResNet as a backbone network (with a dropout layer introduced before the fully connected blocks) to create two types of prognosis models: Single task (i.e., with the distant disease‑free survival data only. Named by the authors as "Solo" models) and models trained in a multitask fashion (i.e., predicting ER and HER2 status together with the distant disease‑free survival data) (Bychkov et al., 2022).

## ML model for classification and prognosis purposes

In a journal article, Wang et al. externally validated a model to assign Nottingham Histological Grades to primary invasive breast cancers (i.e., classification purposes) and predict their recurrence-free survival (i.e., prognosis) (Wang Y. et al., 2022).

They trained and internally validated a model called "DeepGrade" with cases extracted from the Stockholm South General Hospital, Karolinska University Hospital, and TCGA. Its EV was conducted with cases from patients diagnosed with breast cancer in Lund between 2010 and 2019 that were included in the Sweden Cancerome Analysis Network - Breast initiative [SCAN-B] study. None of them underwent neoadjuvant chemotherapy (Wang Y. et al., 2022) (Table 6).

Table 6. Datasets, sources, and preprocessing steps for development and validation of DeepGrade

|  |  |
| --- | --- |
|  | Wang et al. (2022) |
| Dataset |
| Type of lesions / tumors | Primary invasive breast cancer (subtypes not specified) |
| Training / tuning datasets: source (n)  | SSGH, KUH, and TCGA (844) |
| IV datasets: source (n) | SSGH, KUH, and TCGA (351) |
| EV dataset: source (n) | SCAN-B (1262) |
| Preprocessing step |
| Color normalization | Yes |
| Data augmentation | Yes |
| EV: External Validation. IV: Internal Validation. KUH: Karolinska University Hospital. SCAN-B: Sweden Cancerome Analysis Network-Breast project. SSGH: Stockholm South General Hospital. TCGA: The Cancer Genome Atlas. |

From WSIs scanned at 40x, Wang et al. extracted patches of 598 x 598 pixels. Images were preprocessed with color normalization and data augmentation using standard Python packages. Deep learning was performed using the Keras (2.2.4) framework with TensorFlow (1.12) backend. The trained model was an Inception V3 (Wang Y. et al., 2022), initialized with weights from the model pretrained on the "ImageNet" (Russakovsky et al., 2015). They used ADAM for optimization (Wang Y. et al., 2022).

## Risk of bias in studies

According to the PROBAST (Moons et al., 2019), two studies were at a high risk of bias due to the sample size of the validation dataset (Table 7 and Figure 3). Therefore, when used in practice, their predictive performance will likely be lower than that reported. For other studies, the risk of bias was unclear due to the lack of more detailed information about the datasets (all but one of them) and the statistical analyses (studies validating prognostic ML models). There was “low concern” regarding applicability for all studies (Table 7 and Figure 3).

Table 7. Risk of bias assessment (PROBAST)

|  |  |  |  |
| --- | --- | --- | --- |
| Authors | ROB | Applicability | Overall |
| PAR | PRE | OC | ANA | PAR | PRE | OC | ROB | Applicability |
| Cano et al. | ? | + | + | + | + | + | + | ? | + |
| Cruz-Roa et al.  | ? | + | + | + | + | + | + | ? | + |
| Cruz-Roa et al.  | ? | + | + | + | + | + | + | ? | + |
| Colon-Cartagena et al. | ? | ? | + | - | + | + | + | - | + |
| Mi et al. | ? | + | + | + | + | + | + | ? | + |
| Radiya-Dixit, E. et al. | ? | ? | + | - | + | + | + | - | + |
| Yang et al.  | ? | + | + | + | + | + | + | ? | + |
| Bai et al. | ? | + | + | ? | + | + | + | ? | + |
| Bychkov et al. | + | + | + | ? | + | + | + | ? | + |
| Wang et al. | ? | + | + | ? | + | + | + | ? | + |
| ANA: Analysis. OC: Outcome. PAR: Participants. PRE: Predictors. "PROBAST" = Prediction model Risk Of Bias Assessment Tool. "ROB" = Risk of bias. "−"= high ROB or high concern regarding applicability. “+” = Low ROB or low concern regarding applicability. “?” = Unclear ROB or unclear concern regarding applicability. |

Figure 3. Prediction model Risk Of Bias Assessment Tool (PROBAST) Graphical presentation – (1) Risk of Bias results and (2) Applicability

## Results of individual studies

The ground truth used to assess the performance of models developed for diagnostic or classification purposes were pathologists' annotations/opinions and, for prognostic models, clinical data. The results are summarized below (Table 8)

## ML for diagnostic purposes:

While Cano et al. selected accuracy as the only performance metric for the EV of their model (Cano et al., 2018), Cruz Roa et al. reported several others (Cruz-Roa et al., 2017, 2018) (Table 8).

Table 8. Results of diagnostic models

|  |  |  |  |
| --- | --- | --- | --- |
| Study | Cano et al. (2018) | Cruz Roa et al. (2017) | Cruz-Roa et al. (2018) |
| Dataset | TCGA | CWRU/UHCMC\* |
| AC (%) | 88.78 | - | - | - |
| Dice (%) | - | 75.86 | - | 76 |
| PPV (%) | - | 71.62 | - | 72 |
| NPV (%) | - | 96.77 | 100 | 97 |
| TRP (%) | - | 86.91 | - | 87 |
| TNR (%) | - | 92.18 | 99.64 | 92 |
| FPR (%) | - | 7.82 | 0.36 | 8 |
| FNR (%) | - | 13.09 | - | 13 |
| HM | - | - | - | √ |
| AC: Accuracy. HM: Heatmaps. CWRU: Case Western Reserve University. Dice: Dice coefficient. FNR: False Negative Rate. FPR: False Positive Rate. NPV: Negative Predictive Value. PPV: Positive Predictive Value. TCGA: The Cancer Genome Atlas. TNR: True Negative Rate. TPR: True Positive Rate. UHCMC: University Hospitals Case Medical Center. √: Good concordance between predictions of HASHI and pathologists' annotations.\* Not all performance metrics were calculated because the "normal" dataset did not have cancer annotations. |

## ML for classification purposes:

The most common performance metric used in their EV was accuracy (Colon-Cartagena et al., 2020; Mi et al., 2021; Yang Z. et al., 2019), followed by Area Under the Curve (AUC) (Radiya-Dixit E. et al., 2017; Yang Z. et al., 2019). Mi et al. reported the accuracies of their model in the BreakHis dataset for four different magnifications (Mi et al., 2021), and Yang et al. included the performance metrics for the entire dataset (overall) and the benign and malignant categories (Yang Z. et al., 2019). In summary, accuracy ranged from 87.2% (Mi et al., 2021) to 99.75% (Yang Z. et al., 2019), AUC from 91.8% (Radiya-Dixit E. et al., 2017) to 99.99% (Yang Z. et al., 2019), and precision and recall from 99.20% to 100% (Yang Z. et al., 2019).

 Table 9. Results of classification models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Authors | Colon-Cartagena et al. (2020) | Mi et al. (2021) | Radiya-Dixit et al. (2017) | Yang et al.(2019) |
| Dataset/category | BreakHis (Magnification) | BACH, part A | BACH, part B | Benign | Malignant |
| AC (%) | 90 | 96.7 (4x)97.6 (10x)95.0 (20x)93.3 (40x) | 87.2 | - | - | 99.75 | 99.75 |
| AUC (%) | - | - | - | - | 91.8 | 99.99 | 99.99 |
| Precision (%) | - | - | - | - | - | 100 | 99.2 |
| Recall (%) | - | - | - | - | - | 99.64 | 100 |
| HM | - | - | - | √ | - | - | - |
| AC: Accuracy. AUC: Area Under the Curve. HM: Heatmaps. √: Good concordance between predictions of HASHI and pathologists' annotations. |

## ML for prognostic purposes:

Bai et al. reported that all the TILs variables had signiﬁcant prognostic association with overall survival (P ≤ 0.01 for all comparisons). However, as shown in Table 10, there was cell-speciﬁc variation in validation sets. In addition, the derived easTILs variable score had a good correlation with the pathologist-read sTILs in the WTS Sweden cohort (Bai et al., 2021). Bychkov et al. found that both "Solo" models (i.e., to predict the distant disease‑free survival data only) and those trained in a multitask fashion (i.e., predicting Estrogen Receptor and HER2 status together with the distant disease‑free survival data) significantly predicted distant disease‑free survival (Table 10). (Bychkov et al., 2022).

Table 10. Results of prognostic models for predicting survival

|  |  |  |
| --- | --- | --- |
| Authors | Bai et al. (2021) | Bychkov et al. (2022) |
| Dataset/model | TMA Yale1 | TMA Yale2 | WTS TCGA | WTS Sweden | “Solo” model | Multitask model  |
| HR (95% CI) p-value\* | High eTILs% | 0.64(0.43-0.94)p = 0.025 | 0.43(0.26-0.69)p = 0.0005 | 0.09(0.01-0.70)p = 0.02 | NS | - | - |
| HighetTILS% | 0.51 (0.32-0.81) p = 0.004 | 0.47(0.28-0.77)p = 0.003 | 0.1(0.01-0.80)p = 0.03 | NS | - | - |
| HighesTILs | 0.48 (0.25-0.89) p = 0.02 | 0.42(0.24-0.76)p = 0.004 | NS | NS | - | - |
| HigheaTILs (mm2) | 0.48 (0.31-0.74) p = 0.0009 | 0.62(0.37-1.01)p = 0.06 | 0.1(0.01-0.76)p = 0.03 | NS | - | - |
| High easTILs | 0.65 (0.43-0.98) p = 0.04 | 0.78(0.48-1.26)p = 0.31 | NS | 0.54(0.31-0.92)p = 0.02 | - | - |
| Predicted as “High risk” | - | - | - | - | 1.8(1.3-2.7)0.002 | 1.7(1.2-2.6)0.003 |
| Spearman r coefﬁcient (p-value) | High easTILs | 0.61§(p<0.0001) | 0.61§(p<0.0001) | NS | 0.63§(p<0.0001) | - | - |
| c-index | - | - | - | - | - | 0.57 | 0.57 |
| 95% CI: 95% confidence interval. c‑index: concordance between the CNN predicted risk score and the actual time‑to‑event data. eaTILs (mm2): Density of TILs over tumor region. easTILs: Density of TILs over stroma area that mimics the international TIL working group variable as read by pathologists. eTILs%: Proportion of TILs over tumor cells. esTILs%: Proportion of TILs over stromal cells. etTILs%: Proportion of TILs over all detected cells. sTIL: Stromal TILs. HR: Hazard ratios. TILs: Tumor-infiltrating lymphocytes\*Outcomes predicted: Distant disease‑free survival in patients with higher TILs scores (Bai et al. (2021)) and distant disease‑free survival in patients predicted as “High risk” by ML models (by Bychkov et al. (2022))§ Good correlation found when the CNN11-derived easTLs variable score was compared with the pathologist-read sTILs assessment. |

## ML for classification and prognosis purposes:

Wang et al. achieved an AUC of 0.907 when separating Nottingham Histological Grade 1 and Nottingham Histological Grade 3 invasive breast carcinomas and predicted recurrence-free survival rates between DeepGrade-classiﬁed Nottingham Histological Grade 1 and 3 patients that were similar to that of clinically assigned Nottingham Histological Grade 1 and 3 (defining recurrence as "locoregional or distant relapses, contralateral tumours or death"). It also provided a signiﬁcant prognostic value for stratiﬁcation of Nottingham Histological Grade 2 (P =0.0045), and those predicted as high grade showed an increased risk for recurrence, with a HR of 1.91 (95% CI, 1.11-3.29, P = 0.019). Pathologist opinion Histological grading assigned by pathologists in the clinical setting was used as the ground truth (Wang Y. et al., 2022).

## DISCUSSION

## General interpretation of the results in the context of other evidence

To our knowledge, this is the first systematic review specifically focused on assessing externally validated ML models based on histopathology images for diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer.

Validating ML models is essential to ensure they will perform the task they were developed for (Homeyer et al., 2022). In an internal validation, an input dataset is split into parts; one is used to train and potentially fine-tune a ML model, and the other to test it. An EV uses independently derived datasets (i.e., external) to train and test ML models (Ho et al., 2020; Nagendran et al., 2020; Norgeot et al., 2020; Park et al., 2021). As speciﬁc patterns learned from biased-training datasets are not expected to improve the performance of ML models when tested with independent datasets, only EV is considered important evidence of generalizability (Ho et al., 2020).

Although numerous journal articles and conference proceedings were identified with our search queries, the majority of them did not perform an EV of their ML models. This limitation has also been described for other ML models developed for medical purposes (Adeoye et al., 2021; Akazawa & Hashimoto, 2021; Bang et al., 2021; D. W. Kim et al., 2019; Nagendran et al., 2020; Nguyen et al., 2018; Nwanosike et al., 2021; Park et al., 2021; Shelmerdine et al., 2021; Shi et al., 2021; Twilt et al., 2021; Yao et al., 2020) and for diagnostic or predictive purposes on breast cancer patients specifically (Corti et al., 2022; Mazo et al., 2022; A. C. Yu et al., 2022). As previously stated, this may be related to the difficulty in finding appropriate external datasets (Mazo et al., 2022; A. C. Yu et al., 2022), nonadherence to guidelines that promote EV (Collins et al., 2015; Corti et al., 2022; Luo et al., 2016; Norgeot et al., 2020; Park & Han, 2018; A. C. Yu et al., 2022) and lack of awareness of their importance (A. C. Yu et al., 2022).

The reason why most of the included studies validated ML models for diagnosis or classification purposes is unknown. However, it may be partly explained by the fact that ML models used to predict biomarkers were excluded during the selection process, limiting the number of eligible studies used for prognosis or treatment-response prediction.

The accessibility to large data sets to train ML models may explain why most studies were conducted in the United States/Colombia, the United States, or China (Savage, 2020). Although Colombia has not been traditionally considered a leader in AI-related research output, in three studies, the first authors were affiliated to Colombian institutions and used datasets extracted from the United States.

There was significant heterogeneity in the studies regarding the amount and type of information reported. In addition, the performance metrics, datasets, ML models, image preprocessing, patch sizes, magnifications and platforms used for annotating/computational purposes were highly variable. This finding corroborates the lack of standardization on the methodology and reported information found in previous systematic reviews, such as those reported by Mazo et al. with studies using artiﬁcial intelligence tools to predict breast cancer recurrence (Mazo et al., 2022), by Gao et al. with ML-based breast cancer risk prediction models (Gao et al., 2021), by Cori et al. with artificial intelligence algorithms for prediction of treatment outcomes in breast cancer (Corti et al., 2022), by Nagendran et al. with deep learning algorithms for medical imaging (Nagendran et al., 2020) and by Yu et al. with deep learning algorithms with EV for radiologic diagnosis (A. C. Yu et al., 2022). As explained by other authors, this is an important limitation of these systematic reviews (A. C. Yu et al., 2022) that impedes from making rigorous comparisons (Gao et al., 2021), better understand findings (Dhiman et al., 2021) and limits models' generalizability and their clinical impact (Corti et al., 2022). Increasing adherence to existing and upcoming reporting guidelines (Collins et al., 2015; Luo et al., 2016; Norgeot et al., 2020; Shelmerdine et al., 2021; Sounderajah et al., 2020; Vasey et al., 2022) could be potentially improved by training authors on their practical use, enhancing the understanding or their content, encouraging and checking the adherence to them and involving experts on methodology and reporting on AI research groups (Blanco et al., 2019).

Using large and diverse datasets to address the variability that a ML model may find in real-life settings and to allow meaningful statistical analyses is recommended (Homeyer et al., 2022). Nevertheless, it was impossible to determine if the included studies were aligned with this recommendation. That is because detailed descriptions of the variability of the datasets were not included in the reports. In addition, as stated above, important differences among the studies in terms of the source, size, and histological type of lesions/tumors included in the training/tuning and validating datasets were noted.

The majority of experiments were performed using WSIs. As for many other tools developed for computational pathology (Zormpas-Petridis et al., 2020), most authors utilized images scanned with 20x and 40x magnifications. Data augmentation and color normalization were the most common preprocessing methods. Both have been widely used in computational pathology to help improve the generalizability of ML models. The first aims to increase the diversity of the training data by adding artificially generated variations of them (e.g., by rotating or mirroring the images) (Tellez et al., 2019). The latter tries to reduce the effect of color variations in the images (usually a consequence of different staining or scanning processes) (Boschman et al., 2022; Tellez et al., 2019). Except for some similarities found in the two studies written by the same author (Cruz-Roa et al., 2017, 2018), the hardware/software platforms used for annotation/computational purposes and the information published about them were different on each study.

All the included studies, but one, used CNNs, a category of Deep Neural Networks (DNNs). Unlike traditional computer vision approaches that require designing hand-engineered features (usually time-consuming and expensive), DNNs can learn to extract features automatically and classify data samples simultaneously (Khan et al., 2018). Besides this advantage, CNNs use two- (or high-) dimensional filters (AKA convolutional kernels) that have shown to be very powerful in learning patterns from high-dimensional input data (e.g., images and videos) (Khan et al., 2018). Consequently, CNNs became dominant during the last years (Iglesias et al., 2021; Yamashita et al., 2018). Even though transformers have outperformed CNNs in some computer vision tasks, they started to be used more recently and were not found in any of the included studies (Shmatko et al., 2022).

Considering that the performance of ML models when validated on external datasets usually diminishes, the results of most included studies can be regarded as encouraging. For example, all the accuracies and AUC achieved by the diagnostic or classification models were above 87% and 90%, respectively; Yang et al. obtained perfect or almost perfect precision and recall values, the prognostic models developed by Bychkov et al. and Wang et al., the Hazard Ratios (HR) were between 1.7 and 1.9 (with statistical significance), and all the machine TIL variables developed by Bai et al. were significantly associated with outcomes. However, as discussed below, the limited number of classes (AKA histological subtypes of lesions/tumours) in which the models were applied could restrict their usability in real-life clinical practices.

## Limitations of the evidence included in the review

As explained by several authors before (Corti et al., 2022; Gao et al., 2021; Mazo et al., 2022; A. C. Yu et al., 2022), the lack of consistency in the methods and performing metrics and the heterogeneity of the external validation datasets limited the comparison of the studies included in this systematic review and a deeper understanding of their results.

It is also relevant to mention that all the studies that externally validated ML models for diagnosis or classification purposes, trained and validated the models with broad categories of diseases (aka groups of tumors or "classes") or only with a specific subset of them (i.e., only a few examples of histological subtypes of lesions or "subclasses"). In clinical practices, pathologists always need to be able to recognize all the specific subtypes of lesions listed in the classifications regarded as standard. The World Health Organization (WHO) classifications are the most commonly used classifications for human tumors (*WHO Blue Books Web Site Launched – IARC*, n.d.). And as an example, the category "invasive breast carcinoma" contains more than twenty specific histological subtypes of tumors in its most recent edition (and some of them can be further subclassified, i.e., contain “sub-subclasses”) (Cserni, 2020). Although training/testing a model to recognize all the specific subtypes of lesions /tumours could be out of the scope of all or almost all ML models, this limitation will also undoubtedly restrict their use in clinical settings.

## Limitations of the current review

Although the search strategies were designed to be very comprehensive (See Appendix 1), the studies that did not mention or implicitly suggest in their Title/Abstract that their ML models underwent external validation were excluded. The same happened with those published before January 1, 2010, or after February 28, 2022. Consequently, some ML models that might have otherwise met the inclusion criteria were not included in this systematic review.

## Implications of the results for practice, policy, and future research

Recent advances in ML methods, increasing computer power, and unprecedented amounts of data are reshaping cancer care (Kann et al., 2021; Luchini et al., 2022). Numerous ML models for cancer have been developed; however, their application in real-life clinical settings is still limited (Adeoye et al., 2021; Akazawa & Hashimoto, 2021; Bang et al., 2021; Nagendran et al., 2020; Nwanosike et al., 2021; Park et al., 2021; Shi et al., 2021; Twilt et al., 2021). So far, most ML models have only been internally validated (Adeoye et al., 2021; Akazawa & Hashimoto, 2021; Bang et al., 2021; Nagendran et al., 2020; Nwanosike et al., 2021; Park et al., 2021; Shi et al., 2021; Twilt et al., 2021). Although obtaining good results in one or more EV datasets is considered one important step to ensure its generalizability (Ho et al., 2020), it is essential to note that this doesn't guarantee that the model will perform well in all other settings (Homeyer et al., 2022). As ML models’ prediction/recognition capabilities are restricted by the amount and diversity of examples used to train them, patterns not properly represented in training datasets are not expected to be adequately predicted/recognized in validation datasets (Tang et al., 2021; Tellez et al., 2019; Vali-Betts et al., 2021). The ability to understand and apply concepts as pathologists (Chow et al., 2021; Molavi, 2017) (instead of using a limited set of recurrent patterns learned from training datasets) could be needed if widely generalizable ML models are expected to be developed (Lake et al., 2017) And even if, for practical reasons, we opt to assume that future events (such as those expected to be predicted by ML models) will always resemble past events (e.g., those used to create training datasets), it is essential to be mindful that "universally" generalizable models may be unachievable if the generalization problem is approached from a philosophical perspective and the problems of induction, as those discussed by Lauc (Lauc, 2020), are contemplated.

As stated by the FDA, validation datasets must contain sufficient cases representative of those the product will likely encounter during its intended use (U.S. Food & Drug Administration., n.d.). Therefore, for real-life practice, comprehensive validation datasets would need to be created for each clinical setting where models are planned to be used (Homeyer et al., 2022). In addition, to improve their site-specific performances, datasets to retrain and fine-tune their hyperparameters would also need to be constructed (Homeyer et al., 2022). As shown in Figure 4, this can become an iterative process for each institution, considering that the ML model's performance would need to be monitored as new cases with previously unseen relevant characteristics would permanently arrive to be assessed (Rojas et al., 2022; Shankar et al., 2022; Symeonidis et al., 2022). Although the iterative nature of this process may not make ML models “universally” generalizable (Futoma et al., 2020; Lauc, 2020), it would certainly boost their learning capabilities by leveraging their ability to falsify prediction rules that lack empirical adequacy, as explained by Buchholz & Raidl (Buchholz & Raidl, 2022). Once these steps can be done automatically (John et al., 2021; Symeonidis et al., 2022) and some technical challenges are overcome (Maleki et al., 2020; Parisi et al., 2018), a site-specific autonomous endless self-learning process could be developed.



Figure 4. ML models site-specific iterative development steps

§ Validating the model with (external) cases extracted from different settings would be more relevant at the beginning of the process (i.e., before a ML model is deployed in an institution for the first time)

Another consequence of the above process is that privately owned ML models can become more generalizable when retrained with data extracted from institutions paying for them (G. Liu et al., 2022). Therefore, the implications of this improvement could be discussed while negotiating with vendors.

In addition, some considerations related to the applicability of ML models in pathology must be mentioned. WSIs carry large amounts of valuable information (Rosai, 2001). Some of this information has been visually recognizable for years and used for diagnostic/classification purposes and, secondarily or indirectly, to predict prognosis and treatment outcomes (Rosai, 2001). Other information, previously hidden from human eyes, can now be extracted using some ML models (Echle et al., 2021; Shmatko et al., 2022). This has created many opportunities with the potential to augment human skills (Harrison et al., 2021). For example, using clinical follow-up data as the gold standard (i.e., to capture the ground truth), hidden information has been found to be very powerful in predicting prognosis, treatment responses, and biomarkers (Echle et al., 2021; Shmatko et al., 2022). Nevertheless, some concerns have arisen and will still be subject to debate, such as the need for models' explainability (recently discussed by McCoy et al. (McCoy et al., 2022)), and some limitations will need to be overcome.

Beyond those related to pathologists’ intra- and inter-observer variability that could be considered inherent to their evaluative activities (Crowley et al., 2003; Fleming, 1996; Hamilton et al., 2009; Peters, 1996) (although the open debate between “naturalists” and “normativists” may be relevant to recognize and accept different standpoints) (Amoretti & Lalumera, 2022; Conley & Glackin, 2021; Kingma, 2010), and others whose solution might rely on innovative approaches or future technological capabilities; such as the scarcity of large, diverse, and granular datasets (Gildenblat & Klaiman, 2020; Haeyeh et al., 2022), and the constrains of training models with selected patches and not with complete WSIs (Ciga et al., 2021; Wu et al., 2022) or with labels extracted from pathology reports (which may not contain all the specific diagnoses found on each WSI or include unstandardized or out-of-date terminology) (Al-Sukhni et al., 2016; Ellis & Srigley, 2016; Garcia-Roig et al., 2013; Hung et al., 2019); there is an important one, which limits the scope of many ML models based on morphology for diagnostic/classification purposes, that needs to be discussed. As explained below, It relates to how pathologists make morphological diagnoses and the difference between hidden and diagnostically relevant information.

With the proper clinical/surgical data and relevant auxiliary tests’ results (e.g., obtained with immunohistochemical or molecular studies), pathologists make diagnoses by comparing the visual information they extract from patients’ tissues/samples against sets of diagnostic criteria (i.e., only after confirming that a tissue/sample meets some diagnostic criteria, pathologists assign a disease name to it) (Funkhouser, 2018; Hamilton et al., 2009). These sets of diagnostic criteria are listed in histopathology classifications (e.g., WHO Classification of Tumours) (Uttley et al., 2020) and generally use concepts (Chow et al., 2021; Molavi, 2017) to describe the presence/absence and the spatial distribution of some normal/abnormal cells and tissue components (Hamilton et al., 2009; Reddy et al., 2021; Tambasco et al., 2009). As histopathology classifications are typically created by pathologists and are intended to be used by pathologists (Uttley et al., 2020), only those cells/tissue components expected to be recognizable by them are included in the diagnostic criteria (J. M. Kim et al., 2019; L. H. Lee et al., 2017). Therefore, information hidden from pathologists' eyes that is not useful to assess if diagnostic criteria are met (even if it is valuable to predict prognosis, treatment outcomes, and molecular biomarkers (Shmatko et al., 2022)), could be considered irrelevant for morphologic diagnostic/classification purposes. This limitation may persist as long as morphologic classifications made by pathologists are needed to guide clinical decisions, and consequently, only pathologists' opinions are regarded as the gold standard (Salto-Tellez & Cree, 2019). On the other hand, developing ML models for morphologic diagnostic/classification purposes that recognize diagnostically relevant cells/tissue components may have some practical advantages. For example, it could be a reasonable approach to make ML models more applicable in the long term, considering that histopathology classifications usually change over time (i.e., are "moving targets") (Cserni, 2020; Pearson et al., 2008), but most diagnostically relevant cells/tissue components do not. In addition, due to their limited scope, the need to develop "explainable" algorithms could be less relevant, and ML models could gain regulatory agencies' approval easier if designed to improve pathologists' workflows (and not to make diagnoses directly) (Tosun et al., 2020). Lastly, better performances might be achievable with less effort. That is because cells/tissue components have some physical attributes (such as colors, shapes, and textures) that ML models could easily identify (e.g., using convolutional kernels) (Caicedo et al., 2009). In contrast, to recognize specific diseases, as stated above, ML models would need to assess if some diagnostic criteria are met; and the ability to understand some concepts included in them is beyond current ML models’ capabilities.

Finally, ML models that continuously integrate and assign specific weights (i.e., a relative importance) to personal (e.g., clinical, radiological, histopathological, laboratory medicine, multi-omics, self-reported and collected with wearable devices) and population-based empirical data (e.g., related to “social determinants of health”) may be developed (Amal et al., 2022; Chen et al., 2022; Kline et al., 2022) to predict health outcomes dynamically (C. Lee et al., 2022; Pickett et al., 2021). For some of these models (e.g., those created to predict treatment responses), both diagnostically-relevant and hidden information obtained from histopathology slides may play a new essential role. Although many challenges related to data governance/management (Baumfeld Andre et al., 2022; Chomutare et al., 2022; Fisher & Rosella, 2022; Haendel et al., 2018; J. Liu, 2022; Padron-Monedero et al., 2022; Peng et al., 2020), ethical/legal (Gerke et al., 2020; Kostick-Quenet et al., 2022; Mehta et al., 2020) and environmental (Dhar, 2020; Wolf et al., 2022) considerations would need to be addressed, these “dynamic multimodal ML models” may one day become cost-effective in different populations if conceived as the integrative tools needed to support “precision health” (Beckmann & Lew, 2016; Bove et al., 2022; Corti et al., 2023; Gambhir et al., 2018; Kline et al., 2022; Schüssler-Fiorenza Rose et al., 2019) and “learning health systems” (Friedman et al., 2017; Kasperbauer, 2021; Kohn et al., 2022).

## CONCLUSION

Despite EV being an important step before the clinical application of a ML model, it hasn't been performed routinely. The large variability in the training/validation datasets, methods, performance metrics, and information included in the reports limited the comparison of the models and the analysis of their results. Increasing the availability of validation datasets with further implementation of DP platforms and developing and enhancing the adherence to standardized methods and reporting protocols may facilitate future comprehensive analysis. The importance of EV during model development, deployment, and maintenance is discussed, and some considerations on the potential role of pathology in future data-driven healthcare systems are presented.

## OTHER INFORMATION

## Registration and protocol

This systematic review was not registered.

## Support

The search strategy was developed in conjunction with a McMaster Health Sciences Library librarian with expertise in systematic reviews.

## Competing interests

This systematic review was not externally funded. There are no financial conflicts of interest to disclose.

## Availability of data, code, and other materials

Data available upon request.

## REFERENCES

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto‐Arroyo, E., Beck, A. H., & Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: A white paper from the Digital Pathology Association. *The Journal of Pathology*, *249*(3), 286–294. https://doi.org/10.1002/path.5331

Adeoye, J., Tan, J. Y., Choi, S.-W., & Thomson, P. (2021). Prediction models applying machine learning to oral cavity cancer outcomes: A systematic review. *International Journal of Medical Informatics*, *154*, 104557. https://doi.org/10.1016/j.ijmedinf.2021.104557

Akazawa, M., & Hashimoto, K. (2021). Artificial intelligence in gynecologic cancers: Current status and future challenges - A systematic review. *Artificial Intelligence in Medicine*, *120*, 102164. https://doi.org/10.1016/j.artmed.2021.102164

Al-Sukhni, E., LeVea, C., Skitzki, J., Kane, J., & Francescutti, V. (2016). Key gaps in pathologic reporting for appendiceal mucinous neoplasms: Time for universal synoptic reporting? *Annals of Diagnostic Pathology*, *24*, 52–54. https://doi.org/10.1016/j.anndiagpath.2016.06.004

Amal, S., Safarnejad, L., Omiye, J. A., Ghanzouri, I., Cabot, J. H., & Ross, E. G. (2022). Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine*, *9*, 840262. https://doi.org/10.3389/fcvm.2022.840262

Amoretti, M. C., & Lalumera, E. (2022). Wherein is the concept of disease normative? From weak normativity to value-conscious naturalism. *Medicine, Health Care, and Philosophy*, *25*(1), 47–60. https://doi.org/10.1007/s11019-021-10048-x

Azam, A. S., Miligy, I. M., Kimani, P. K.-U., Maqbool, H., Hewitt, K., Rajpoot, N. M., & Snead, D. R. J. (2021). Diagnostic concordance and discordance in digital pathology: A systematic review and meta-analysis. *Journal of Clinical Pathology*, *74*(7), 448–455. https://doi.org/10.1136/jclinpath-2020-206764

Bai, Y., Cole, K., Martinez-Morilla, S., Ahmed, F. S., Zugazagoitia, J., Staaf, J., Bosch, A., Ehinger, A., Nimeus, E., Hartman, J., Acs, B., & Rimm, D. L. (2021). An open-source, automated tumor-infiltrating lymphocyte algorithm for prognosis in triple-negative breast cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *27*(20), 5557–5565. https://doi.org/10.1158/1078-0432.CCR-21-0325

Bang, C. S., Lee, J. J., & Baik, G. H. (2021). Computer-aided diagnosis of esophageal cancer and neoplasms in endoscopic images: A systematic review and meta-analysis of diagnostic test accuracy. *Gastrointestinal Endoscopy*, *93*(5), 1006-1015.e13. https://doi.org/10.1016/j.gie.2020.11.025

Baumfeld Andre, E., Carrington, N., Siami, F. S., Hiatt, J. C., McWilliams, C., Hiller, C., Surinach, A., Zamorano, A., Pashos, C. L., & Schulz, W. L. (2022). The current landscape and emerging applications for real-world data in diagnostics and clinical decision support and its impact on regulatory decision making. *Clinical Pharmacology & Therapeutics*, *112*(6), 1172–1182. https://doi.org/10.1002/cpt.2565

Beckmann, J. S., & Lew, D. (2016). Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Medicine*, *8*(1), 134. https://doi.org/10.1186/s13073-016-0388-7

Bhinder, B., Gilvary, C., Madhukar, N. S., & Elemento, O. (2021). Artificial intelligence in cancer research and precision medicine. *Cancer Discovery*, *11*(4), 900–915. https://doi.org/10.1158/2159-8290.CD-21-0090

Blanco, D., Altman, D., Moher, D., Boutron, I., Kirkham, J. J., & Cobo, E. (2019). Scoping review on interventions to improve adherence to reporting guidelines in health research. *BMJ Open*, *9*(5), e026589. https://doi.org/10.1136/bmjopen-2018-026589

Boschman, J., Farahani, H., Darbandsari, A., Ahmadvand, P., Van Spankeren, A., Farnell, D., Levine, A. B., Naso, J. R., Churg, A., Jones, S. J., Yip, S., Köbel, M., Huntsman, D. G., Gilks, C. B., & Bashashati, A. (2022). The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images. *The Journal of Pathology*, *256*(1), 15–24. https://doi.org/10.1002/path.5797

Bove, R., Schleimer, E., Sukhanov, P., Gilson, M., Law, S. M., Barnecut, A., Miller, B. L., Hauser, S. L., Sanders, S. J., & Rankin, K. P. (2022). Building a precision medicine delivery platform for clinics: The University of California, San Francisco, BRIDGE experience. *Journal of Medical Internet Research*, *24*(2), e34560. https://doi.org/10.2196/34560

Bray, F., Laversanne, M., Weiderpass, E., & Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16), 3029–3030. https://doi.org/10.1002/cncr.33587

Buchholz, O., & Raidl, E. (2022). A falsificationist account of artificial neural networks. *The British Journal for the Philosophy of Science*, 721797. https://doi.org/10.1086/721797

Bychkov, D., Joensuu, H., Nordling, S., Tiulpin, A., Kucukel, H., Lundin, M., Sihto, H., Isola, J., Lehtimaki, T., Kellokumpu-Lehtinen, P.-L., von Smitten, K., Lundin, J., & Linder, N. (2022). Outcome and biomarker supervised deep learning for survival prediction in two multicenter breast cancer series. *Journal of Pathology Informatics*, *13*(101528849), 9. https://doi.org/10.4103/jpi.jpi\_29\_21

Caicedo, J. C., Cruz, A., & Gonzalez, F. A. (2009). Histopathology image classification using bag of features and kernel functions. In C. Combi, Y. Shahar, & A. Abu-Hanna (Eds.), *Artificial Intelligence in Medicine* (pp. 126–135). Springer. https://doi.org/10.1007/978-3-642-02976-9\_17

Cano, F., Madabhushi, A., & Cruz-Roa, A. (2018). A comparative analysis of sensitivity of convolutional neural networks for histopathology image classification in breast cancer. *14th International Symposium on Medical Information Processing and Analysis*, *10975*, 277–284. https://doi.org/10.1117/12.2511647

Champion, A., Lu, G., Walker, G., Kothari, S., Osunkoya, A. O., & Wang, M. D. (2014). Semantic interpretation of robust imaging features for Fuhrman grading of renal carcinoma. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6446–6449. https://doi.org/10.1109/EMBC.2014.6945104

Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., & Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, *40*(8), 865-878.e6. https://doi.org/10.1016/j.ccell.2022.07.004

Chomutare, T., Tejedor, M., Svenning, T. O., Marco-Ruiz, L., Tayefi, M., Lind, K., Godtliebsen, F., Moen, A., Ismail, L., Makhlysheva, A., & Ngo, P. D. (2022). Artificial intelligence implementation in healthcare: A theory-based scoping review of barriers and facilitators. *International Journal of Environmental Research and Public Health*, *19*(23). https://doi.org/10.3390/ijerph192316359

Chong, Y., Kim, D. C., Jung, C. K., Kim, D.-C., Song, S. Y., Joo, H. J., Yi, S.-Y., & Medical informatics study group of the Korean Society of Pathologists. (2020). Recommendations for pathologic practice using digital pathology: Consensus report of the Korean Society of Pathologists. *Journal of Pathology and Translational Medicine*, *54*(6), 437–452. https://doi.org/10.4132/jptm.2020.08.27

Chow, Z. L., Indave, B. I., Lokuhetty, M. D. S., Ochiai, A., Cree, I. A., & White, V. A. (2021). Misleading terminology in pathology: Lack of definitions hampers communication. *Virchows Archiv*, *479*(2), 425–430. https://doi.org/10.1007/s00428-021-03069-7

Ciga, O., Xu, T., Nofech-Mozes, S., Noy, S., Lu, F.-I., & Martel, A. L. (2021). Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-88494-z

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55–63. https://doi.org/10.7326/M14-0697

Colon-Cartagena, L., Gayhart, M., & Robila, V. (2020). Coding-free platform for development of image analysis algorithm produces comparable results to established convoluted neural network architecture for sub-classifications of Ductal Carcinoma In Situ of the Breast. *Modern Pathology*, *33*, Article 2. https://doi.org/10.1038/s41379-020-0464-9

Conley, B. A., & Glackin, S. N. (2021). How to be a naturalist and a social constructivist about diseases. *Philosophy of Medicine*, *2*(1), Article 1. https://doi.org/10.5195/pom.2021.18

Corti, C., Cobanaj, M., Dee, E. C., Criscitiello, C., Tolaney, S. M., Celi, L. A., & Curigliano, G. (2023). Artificial intelligence in cancer research and precision medicine: Applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care. *Cancer Treatment Reviews*, *112*. https://doi.org/10.1016/j.ctrv.2022.102498

Corti, C., Cobanaj, M., Marian, F., Dee, E. C., Lloyd, M. R., Marcu, S., Dombrovschi, A., Biondetti, G. P., Batalini, F., Celi, L. A., & Curigliano, G. (2022). Artificial intelligence for prediction of treatment outcomes in breast cancer: Systematic review of design, reporting standards, and bias. *Cancer Treatment Reviews*, *108*. https://doi.org/10.1016/j.ctrv.2022.102410

Crowley, R. S., Naus, G. J., Stewart, J., III, & Friedman, C. P. (2003). Development of visual diagnostic expertise in pathology: An information-processing study. *Journal of the American Medical Informatics Association*, *10*(1), 39–51. https://doi.org/10.1197/jamia.M1123

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., González, F., & Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific Reports*, *7*, 46450. https://doi.org/10.1038/srep46450

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., & González, F. (2018). High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PloS One*, *13*(5). https://doi.org/10.1371/journal.pone.0196828

Cserni, G. (2020). Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica*, *112*(1), 25–41. https://doi.org/10.32074/1591-951X-1-20

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, *2*(8), Article 8. https://doi.org/10.1038/s42256-020-0219-9

Dhiman, P., Ma, J., Navarro, C. A., Speich, B., Bullock, G., Damen, J. A., Kirtley, S., Hooft, L., Riley, R. D., Calster, B. V., Moons, K. G. M., & Collins, G. S. (2021). Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *Journal of Clinical Epidemiology*, *138*, 60. https://doi.org/10.1016/j.jclinepi.2021.06.024

Echle, A., Rindtorff, N. T., Brinker, T. J., Luedde, T., Pearson, A. T., & Kather, J. N. (2021). Deep learning in cancer pathology: A new generation of clinical biomarkers. *British Journal of Cancer*, *124*(4), 686–696. https://doi.org/10.1038/s41416-020-01122-x

Ellis, D. W., & Srigley, J. (2016). Does standardised structured reporting contribute to quality in diagnostic pathology? The importance of evidence-based datasets. *Virchows Archiv: An International Journal of Pathology*, *468*(1), 51–59. https://doi.org/10.1007/s00428-015-1834-4

Fisher, S., & Rosella, L. C. (2022). Priorities for successful use of artificial intelligence by public health organizations: A literature review. *BMC Public Health*, *22*(1), 2146. https://doi.org/10.1186/s12889-022-14422-z

Fleming, K. A. (1996). Evidence-based pathology. *The Journal of Pathology*, *179*(2), 127–128. https://doi.org/10.1002/(SICI)1096-9896(199606)179:2<127::AID-PATH519>3.0.CO;2-Z

Friedman, C. P., Allee, N. J., Delaney, B. C., Flynn, A. J., Silverstein, J. C., Sullivan, K., & Young, K. A. (2017). The science of Learning Health Systems: Foundations for a new journal. *Learning Health Systems*, *1*(1), e10020. https://doi.org/10.1002/lrh2.10020

Funkhouser, W. K. (2018). Chapter 11 - Pathology: The clinical description of human disease. In W. B. Coleman & G. J. Tsongalis (Eds.), *Molecular Pathology (Second Edition)* (pp. 217–229). Academic Press. https://doi.org/10.1016/B978-0-12-802761-5.00011-0

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, *2*(9), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2

Gambhir, S. S., Ge, T. J., Vermesh, O., & Spitler, R. (2018). Toward achieving precision health. *Science Translational Medicine*, *10*(430), eaao3612. https://doi.org/10.1126/scitranslmed.aao3612

Gao, Y., Li, S., Jin, Y., Zhou, L., Sun, S., Xu, X., Li, S., Yang, H., Zhang, Q., & Wang, Y. (2021). *Assessment of performance of the machine learning-based breast cancer risk prediction model: A systematic review and meta-analysis (Preprint)*. JMIR Public Health and Surveillance. https://doi.org/10.2196/preprints.35750

Garcia-Roig, M., Gorin, M., Parra-Herran, C., Garcia-Buitrago, M., Kava, B., M, J., Ms, S., M, M., & G, C. (2013). Pathologic evaluation of non-neoplastic renal parenchyma in partial nephrectomy specimens. *World Journal of Urology*, *31*(4). https://doi.org/10.1007/s00345-011-0720-1

Gerke, S., Minssen, T., & Cohen, G. (2020). Chapter 12—Ethical and legal challenges of artificial intelligence-driven healthcare. In A. Bohr & K. Memarzadeh (Eds.), *Artificial Intelligence in Healthcare* (pp. 295–336). Academic Press. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Gildenblat, J., & Klaiman, E. (2020). *Self-supervised similarity learning for digital pathology* (arXiv:1905.08139). arXiv. https://doi.org/10.48550/arXiv.1905.08139

Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, ontology, and precision medicine. *The New England Journal of Medicine*, *379*(15), 1452–1462. https://doi.org/10.1056/NEJMra1615014

Haeyeh, Y. A., Ghazal, M., El-Baz, A., & Talaat, I. M. (2022). Development and evaluation of a novel deep-learning-based framework for the classification of renal histopathology images. *Bioengineering*, *9*(9). https://doi.org/10.3390/bioengineering9090423

Hamilton, P. W., van Diest, P. J., Williams, R., & Gallagher, A. G. (2009). Do we see what we think we see? The complexities of morphological assessment. *The Journal of Pathology*, *218*(3), 285–291. https://doi.org/10.1002/path.2527

Harrison, J. H., Jr, Gilbertson, J. R., Hanna, M. G., Olson, N. H., Seheult, J. N., Sorace, J. M., & Stram, M. N. (2021). Introduction to artificial intelligence and machine learning for pathology. *Archives of Pathology & Laboratory Medicine*, *145*(10), 1228–1254. https://doi.org/10.5858/arpa.2020-0541-CP

Ho, S. Y., Phua, K., Wong, L., & Bin Goh, W. W. (2020). Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, *1*(8), 100129. https://doi.org/10.1016/j.patter.2020.100129

Homeyer, A., Geißler, C., Schwen, L. O., Zakrzewski, F., Evans, T., Strohmenger, K., Westphal, M., Bülow, R. D., Kargl, M., Karjauv, A., Munné-Bertran, I., Retzlaff, C. O., Romero-López, A., Sołtysiński, T., Plass, M., Carvalho, R., Steinbach, P., Lan, Y.-C., Bouteldja, N., … Zerbe, N. (2022). Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. *Modern Pathology*, 1–11. https://doi.org/10.1038/s41379-022-01147-y

Hung, Y. P., Hunninghake, G. M., Miller, E. R., Putman, R., Nishino, M., Araki, T., Hatabu, H., Sholl, L. M., & Vivero, M. (2019). Incidental non-neoplastic parenchymal findings in patients undergoing lung resection. *Human Pathology*, *86*, 93. https://doi.org/10.1016/j.humpath.2019.01.002

Iglesias, L. L., Bellón, P. S., del Barrio, A. P., Fernández-Miranda, P. M., González, D. R., Vega, J. A., Mandly, A. A. G., & Blanco, J. A. P. (2021). A primer on deep learning and convolutional neural networks for clinicians. *Insights into Imaging*, *12*(1), 117. https://doi.org/10.1186/s13244-021-01052-z

John, M. M., Olsson, H. H., & Bosch, J. (2021). Towards MLOps: A framework and maturity model. *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 1–8. https://doi.org/10.1109/SEAA53835.2021.00050

Kann, B. H., Hosny, A., & Aerts, H. J. W. L. (2021). Artificial intelligence for clinical oncology. *Cancer Cell*, *39*(7), 916–927. https://doi.org/10.1016/j.ccell.2021.04.002

Kasperbauer, T. J. (2021). Conflicting roles for humans in learning health systems and AI-enabled healthcare. *Journal of Evaluation in Clinical Practice*, *27*(3), 537–542. https://doi.org/10.1111/jep.13510

Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). Convolutional Neural Network. In S. Khan, H. Rahmani, S. A. A. Shah, & M. Bennamoun (Eds.), *A Guide to Convolutional Neural Networks for Computer Vision* (pp. 43–68). Springer International Publishing. https://doi.org/10.1007/978-3-031-01821-3\_4

Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., & Srinivasan, B. (2021). A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific Reports*, *11*(1), 11579. https://doi.org/10.1038/s41598-021-90444-8

Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2019). Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. *Korean Journal of Radiology*, *20*(3), 405. https://doi.org/10.3348/kjr.2019.0025

Kim, J. M., Sohn, J. H., Cho, M.-Y., Kim, W. H., Chang, H. K., Jung, E. S., Kook, M.-C., Jin, S.-Y., Chae, Y. S., Park, Y. S., Kang, M. S., Kim, H., Lee, J. H., Park, D. Y., Kim, K. M., Kim, H., Suh, Y. J., Seol, S. Y., Jung, H.-Y., … You, J. H. (2019). Inter-observer reproducibility in the pathologic diagnosis of gastric intraepithelial neoplasia and early carcinoma in endoscopic submucosal dissection specimens: A multi-center study. *Cancer Research and Treatment*, *51*(4), 1568–1577. https://doi.org/10.4143/crt.2019.019

Kingma, E. (2010). Paracetamol, Poison, and Polio: Why Boorse’s account of function fails to distinguish health and disease. *The British Journal for the Philosophy of Science*, *61*(2), 241–264.

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., & Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00712-8

Kohn, M. S., Topaloglu, U., Kirkendall, E. S., Dharod, A., Wells, B. J., & Gurcan, M. (2022). Creating learning health systems and the emerging role of biomedical informatics. *Learning Health Systems*, *6*(1), e10259. https://doi.org/10.1002/lrh2.10259

Kostick-Quenet, K. M., Cohen, I. G., Gerke, S., Lo, B., Antaki, J., Movahedi, F., Njah, H., Schoen, L., Estep, J. E., & Blumenthal-Barby, J. S. (2022). Mitigating racial bias in machine learning. *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics*, *50*(1), 92–100. https://doi.org/10.1017/jme.2022.13

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lauc, D. (2020). Machine learning and the philosophical problems of induction. In S. Skansi (Ed.), *Guide to Deep Learning Basics: Logical, Historical and Philosophical Perspectives* (pp. 93–106). Springer International Publishing. https://doi.org/10.1007/978-3-030-37591-1\_9

Lee, C., Light, A., Saveliev, E. S., van der Schaar, M., & Gnanapragasam, V. J. (2022). Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate cancer. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00659-w

Lee, L. H., Yantiss, R. K., Sadot, E., Ren, B., Calvacanti, M. S., Hechtman, J. F., Ivelja, S., Huynh, B., Xue, Y., Shitilbans, T., Guend, H., Stadler, Z. K., Weiser, M. R., Vakiani, E., Gӧnen, M., Klimstra, D. S., & Shia, J. (2017). Diagnosing colorectal medullary carcinoma: Interobserver variability and clinicopathological implications. *Human Pathology*, *62*, 74–82. https://doi.org/10.1016/j.humpath.2016.12.013

Leong, A. S.-Y., & Zhuang, Z. (2011). The changing role of pathology in breast cancer diagnosis and treatment. *Pathobiology: Journal of Immunopathology, Molecular and Cellular Biology*, *78*(2), 99–114. https://doi.org/10.1159/000292644

Lindman, K., Rose, J. F., Lindvall, M., Lundström, C., & Treanor, D. (2019). Annotations, ontologies, and whole slide images – development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue. *Journal of Pathology Informatics*, *10*. https://doi.org/10.4103/jpi.jpi\_81\_18

Liu, G., Xu, T., Ma, X., & Wang, C. (2022). Your model trains on my data? Protecting intellectual property of training data via membership fingerprint authentication. *IEEE Transactions on Information Forensics and Security*, *17*, 1024–1037. https://doi.org/10.1109/TIFS.2022.3155921

Liu, J. (2022). Social data governance: Towards a definition and model. *Big Data & Society*, *9*(2), 20539517221111350. https://doi.org/10.1177/20539517221111352

Luchini, C., Pea, A., & Scarpa, A. (2022). Artificial intelligence in oncology: Current applications and future perspectives. *British Journal of Cancer*, *126*(1), 4–9. https://doi.org/10.1038/s41416-021-01633-1

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T. B., Venkatesh, S., & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research*, *18*(12). https://doi.org/10.2196/jmir.5870

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine learning algorithm validation: From essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics of North America*, *30*(4), 433–445. https://doi.org/10.1016/j.nic.2020.08.004

Mazo, C., Aura, C., Rahman, A., Gallagher, W. M., & Mooney, C. (2022). Application of artificial intelligence techniques to predict risk of recurrence of breast cancer: A systematic review. *Journal of Personalized Medicine*, *12*(9). https://doi.org/10.3390/jpm12091496

McCoy, L. G., Brenna, C. T. A., Chen, S. S., Vold, K., & Das, S. (2022). Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, *142*, 252–257. https://doi.org/10.1016/j.jclinepi.2021.11.001

Mehta, M. C., Katz, I. T., & Jha, A. K. (2020). Transforming global health with AI. *New England Journal of Medicine*, *382*(9), 791–793. https://doi.org/10.1056/NEJMp1912079

Mi, W., Li, J., Guo, Y., Ren, X., Liang, Z., Zhang, T., & Zou, H. (2021). Deep learning-based multi-class classification of breast digital pathology images. *Cancer Management and Research*, *13*(101512700), 4605–4617. https://doi.org/10.2147/CMAR.S312608

Molavi, D. W. (2017). Chapter 2. Descriptive terms in anatomic pathology. In *The practice of surgical pathology: A beginner’s guide to the diagnostic process* (Second Edition, pp. 6–15). Springer.

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, *170*(1), W1–W33. https://doi.org/10.7326/M18-1377

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ (Clinical Research Ed.)*, *368*, m689. https://doi.org/10.1136/bmj.m689

Nam, S., Chong, Y., Jung, C. K., Kwak, T.-Y., Lee, J. Y., Park, J., Rho, M. J., & Go, H. (2020). Introduction to digital pathology and computer-aided pathology. *Journal of Pathology and Translational Medicine*, *54*(2), 125–134. https://doi.org/10.4132/jptm.2019.12.31

Nguyen, A. V., Blears, E. E., Ross, E., Lall, R. R., & Ortega-Barnett, J. (2018). Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: A systematic review and meta-analysis. *Neurosurgical Focus*, *45*(5), E5. https://doi.org/10.3171/2018.8.FOCUS18325

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., & Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nature Medicine*, *26*(9), 1320–1324. https://doi.org/10.1038/s41591-020-1041-y

Nwanosike, E. M., Conway, B. R., Merchant, H. A., & Hasan, S. S. (2021). Potential applications and performance of machine learning techniques and algorithms in clinical practice: A systematic review. *International Journal of Medical Informatics*, *159*, 104679. https://doi.org/10.1016/j.ijmedinf.2021.104679

Padron-Monedero, A., Suárez, R. S., Bogaert, P., Abboud, L., Van Oyen, H., Tolonen, H., Tijhuis, M. J., Palmieri, L., Haneef, R., Gallay, A., Lapao, L., Nogueira, P. J., Ziese, T., Seeling, S., Vukovic, J., & Noguer-Zambrano, I. (2022). Integrating technical and political views for a sustainable European Distributed Infrastructure on Population Health. *Archives of Public Health*, *80*(1), 29. https://doi.org/10.1186/s13690-022-00790-w

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2018). *Continual lifelong learning with neural networks: A review*. https://doi.org/10.1016/j.neunet.2019.01.012

Park, S. H., Choi, J., & Byeon, J. S. (2021). Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean Journal of Radiology*, *22*(3), 442–453. https://doi.org/10.3348/kjr.2021.0048

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*(3), 800–809. https://doi.org/10.1148/radiol.2017171920

Pearson, B. E., Markert, J. M., Fisher, W. S., Guthrie, B. L., Fiveash, J. B., Palmer, C. A., & Riley, K. (2008). Hitting a moving target: Evolution of a treatment paradigm for atypical meningiomas amid changing diagnostic criteria. *Neurosurgical Focus*, *24*(5), E3. https://doi.org/10.3171/FOC/2008/24/5/E3

Peng, C., Goswami, P., & Bai, G. (2020). A literature review of current technologies on health data integration for patient-centered health management. *Health Informatics Journal*, *26*(3), 1926–1951. https://doi.org/10.1177/1460458219892387

Peters, T. S. (1996). Pathology peer review—A concept for consideration. *Toxicologic Pathology*, *24*(5), 654–656. https://doi.org/10.1177/019262339602400523

Pickett, K. L., Suresh, K., Campbell, K. R., Davis, S., & Juarez-Colunga, E. (2021). Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Medical Research Methodology*, *21*(1), 216. https://doi.org/10.1186/s12874-021-01375-x

Radakovich, N., Nagy, M., & Nazha, A. (2020). Machine learning in haematological malignancies. *The Lancet Haematology*, *7*(7), e541–e550. https://doi.org/10.1016/S2352-3026(20)30121-6

Radiya-Dixit E., Zhu D., & Beck A.H. (2017). Automated Classification of benign and malignant proliferative breast lesions. *Scientific Reports*, *7*(1), 9900. https://doi.org/10.1038/s41598-017-10324-y

Randell, R., Ruddle, R. A., & Treanor, D. (2015). Barriers and facilitators to the introduction of digital pathology for diagnostic work. *Studies in Health Technology and Informatics*, *216*, 443–447.

Reddy, V. B., David, O., Spitz, D. J., & Haber, M. H. (2021). *Gattuso’s differential diagnosis in surgical pathology*. Elsevier Health Sciences.

Rojas, J. C., Fahrenbach, J., Makhni, S., Cook, S. C., Williams, J. S., Umscheid, C. A., & Chin, M. H. (2022). Framework for integrating equity into machine learning models: A case study. *Chest*, *161*(6). https://doi.org/10.1016/j.chest.2022.02.001

Rosai, J. (2001). The continuing role of morphology in the molecular age. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *14*(3), 258–260. https://doi.org/10.1038/modpathol.3880295

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach. Global edition* (4th ed.). Pearson Education Limited.

Salto-Tellez, M., & Cree, I. (2019). Cancer taxonomy: Pathology beyond pathology. *European Journal of Cancer (Oxford, England : 1990)*, *115*. https://doi.org/10.1016/j.ejca.2019.03.026

Savage, N. (2020). The race to the top among the world’s leaders in artificial intelligence. *Nature*, *588*(7837), S102–S104. https://doi.org/10.1038/d41586-020-03409-8

Schüssler-Fiorenza Rose, S. M., Contrepois, K., Moneghetti, K. J., Zhou, W., Mishra, T., Mataraso, S., Dagan-Rosenfeld, O., Ganz, A. B., Dunn, J., Hornburg, D., Rego, S., Perelman, D., Ahadi, S., Sailani, M. R., Zhou, Y., Leopold, S. R., Chen, J., Ashland, M., Christle, J. W., … Snyder, M. P. (2019). A longitudinal big data approach for precision health. *Nature Medicine*, *25*(5), 792–804. https://doi.org/10.1038/s41591-019-0414-6

Shankar, S., Herman, B., & Parameswaran, A. G. (2022). *Rethinking streaming machine learning evaluation* (arXiv:2205.11473). arXiv. https://doi.org/10.48550/arXiv.2205.11473

Shelmerdine, S. C., Arthurs, O. J., Denniston, A., & Sebire, N. J. (2021). Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Health & Care Informatics*, *28*(1), e100385. https://doi.org/10.1136/bmjhci-2021-100385

Shi, Z., Zhang, Z., Liu, Z., Zhao, L., Ye, Z., Dekker, A., & Wee, L. (2021). Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal cancer: A systematic review of clinical outcome prediction after concurrent chemoradiotherapy. *European Journal of Nuclear Medicine and Molecular Imaging*. https://doi.org/10.1007/s00259-021-05658-9

Shmatko, A., Ghaffari Laleh, N., Gerstung, M., & Kather, J. (2022). Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology. *Nature Cancer*, *3*(9). https://doi.org/10.1038/s43018-022-00436-4

Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A. K., Greaves, F., Karthikesalingam, A., King, D., Liu, X., Markar, S. R., McInnes, M. D. F., Panch, T., Pearson-Stuttard, J., Ting, D. S. W., Golub, R. M., Moher, D., Bossuyt, P. M., & Darzi, A. (2020). Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nature Medicine*, *26*(6), Article 6. https://doi.org/10.1038/s41591-020-0941-1

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Symeonidis, G., Nerantzis, E., Kazakis, A., & Papakostas, G. A. (2022). MLOps—Definitions, tools and challenges. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 0453–0460. https://doi.org/10.1109/CCWC54503.2022.9720902

Tambasco, M., Costello, B. M., Kouznetsov, A., Yau, A., & Magliocco, A. M. (2009). Quantifying the architectural complexity of microscopic images of histology specimens. *Micron*, *40*(4), 486–494. https://doi.org/10.1016/j.micron.2008.12.004

Tang, H., Sun, N., & Shen, S. (2021). Improving generalization of deep learning models for diagnostic pathology by increasing variability in training data: Experiments on osteosarcoma subtypes. *Journal of Pathology Informatics*, *12*, 30. https://doi.org/10.4103/jpi.jpi\_78\_20

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., & van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, *58*, 101544. https://doi.org/10.1016/j.media.2019.101544

Têtu, B., & Evans, A. (2014). Canadian licensure for the use of digital pathology for routine diagnoses: One more step toward a new era of pathology practice without borders. *Archives of Pathology & Laboratory Medicine*, *138*(3), 302–304. https://doi.org/10.5858/arpa.2013-0289-ED

Tosun, A. B., Pullara, F., Becich, M. J., Taylor, D. L., Fine, J. L., & Chennubhotla, S. C. (2020). Explainable AI (xAI) for anatomic pathology. *Advances in Anatomic Pathology*, *27*(4), 241–250. https://doi.org/10.1097/PAP.0000000000000264

Twilt, J. J., van Leeuwen, K. G., Huisman, H. J., Fütterer, J. J., & de Rooij, M. (2021). Artificial intelligence based algorithms for prostate cancer classification and detection on magnetic resonance imaging: A narrative review. *Diagnostics (Basel, Switzerland)*, *11*(6), 959. https://doi.org/10.3390/diagnostics11060959

U.S. Food & Drug Administration., P. (n.d.). *DEN200080.Letter.DENG*. 6.

Uttley, L., Indave, B. I., Hyde, C., White, V., Lokuhetty, D., & Cree, I. (2020). Invited commentary—WHO Classification of Tumours: How should tumors be classified? Expert consensus, systematic reviews or both? *International Journal of Cancer*, *146*(12), 3516–3521. https://doi.org/10.1002/ijc.32975

Vali-Betts, E., Krause, K. J., Dubrovsky, A., Olson, K., Graff, J. P., Mitra, A., Datta‑Mitra, A., Beck, K., Tsirigos, A., Loomis, C., Galvao Neto, A., Adler, E., & Rashidi, H. H. (2021). Effects of image quantity and image source variation on machine learning histology differential diagnosis models. *Journal of Pathology Informatics*, *12*. https://doi.org/10.4103/jpi.jpi\_69\_20

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, *28*(5), Article 5. https://doi.org/10.1038/s41591-022-01772-9

Wang Y., Acs B., Robertson S., Liu B., Solorzano L., Wahlby C., Hartman J., & Rantalainen M. (2022). Improved breast cancer histological grading using deep learning. *Annals of Oncology*, *33*(1), 89–98. https://doi.org/10.1016/j.annonc.2021.09.007

*WHO Blue Books web site launched – IARC*. (n.d.). IARC News. Retrieved February 13, 2023, from https://www.iarc.who.int/news-events/who-blue-books-web-site-launched/

Wolf, R. M., Abramoff, M. D., Channa, R., Tava, C., Clarida, W., & Lehmann, H. P. (2022). Potential reduction in healthcare carbon footprint by autonomous artificial intelligence. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00605-w

Wu, Y., Cheng, M., Huang, S., Pei, Z., Zuo, Y., Liu, J., Yang, K., Zhu, Q., Zhang, J., Hong, H., Zhang, D., Huang, K., Cheng, L., & Shao, W. (2022). Recent advances of deep learning for computational histopathology: Principles and applications. *Cancers*, *14*(5), 1199. https://doi.org/10.3390/cancers14051199

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, *9*(4), Article 4. https://doi.org/10.1007/s13244-018-0639-9

Yang Z., Ran L., Zhang S., Xia Y., & Zhang Y. (2019). EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. *Neurocomputing*, *366*((Yang, Xia) Research&Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China), 46–53. https://doi.org/10.1016/j.neucom.2019.07.080

Yao, A. D., Cheng, D. L., Pan, I., & Kitamura, F. (2020). Deep learning in neuroradiology: A systematic review of current algorithms and approaches for the new wave of imaging technology. *Radiology: Artificial Intelligence*, *2*(2), e190026. https://doi.org/10.1148/ryai.2020190026

Yu, A. C., Mohajer, B., & Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*, *4*(3), e210064. https://doi.org/10.1148/ryai.210064

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zormpas-Petridis, K., Noguera, R., Ivankovic, D. K., Roxanis, I., Jamin, Y., & Yuan, Y. (2020). SuperHistopath: A deep learning pipeline for mapping tumor heterogeneity on low-resolution whole-slide digital histopathology images. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.586292

# Appendix 1. Search strategies.

1. OVID Medline Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

1. "neoplasms, glandular and epithelial"/ or carcinoma/ or adenocarcinoma, mucinous/ or carcinoma, adenoid cystic/ or carcinoma, ductal/ or carcinoma, ductal, breast/ or carcinoma, lobular/ or carcinoma, mucoepidermoid/ or carcinoma, neuroendocrine/ or cystadenocarcinoma, mucinous/ or carcinoma, adenosquamous/ or carcinoma, papillary/ or carcinoma, squamous cell/ or "neoplasms, cystic, mucinous, and serous"/ or cystadenocarcinoma/ or "neoplasms, ductal, lobular, and medullary"/ or carcinoma, medullary/ or breast neoplasms/ or "hereditary breast and ovarian cancer syndrome"/ or inflammatory breast neoplasms/ or triple negative breast neoplasms/ or unilateral breast neoplasms/

2. (cancer or carcinoma$ or malignant).tw.

3. breast/ or mammary glands, human/

4. (breast or mammar$).tw.

5. 3 or 4

6. 2 and 5

7. 1 or 6

8. algorithms/ or artificial intelligence/ or machine learning/ or deep learning/ or supervised machine learning/ or support vector machine/ or unsupervised machine learning/ or decision theory/ or decision trees/ or neural networks, computer/

9. machine learning.tw. or exp algorithm/ or algorithm.tw. or Automatic Data Processing/ or Automatic Data Processing.tw. or computer aided detection.tw. or artificial intelligence.tw.

10. 8 or 9

11. pathology/ or pathology, surgical/ or histology/ or histocytochemistry/ or immunohistochemistry/

12. (histopatholog$ or patholog$ or histolog$).tw.

13. ("whole slide imag$" or WSI or (digit$ and slide$)).tw.

14. 11 or 12 or 13

15. valid\*.tw.

16. diagnosis/ or clinical decision-making/ or clinical reasoning/ or diagnosis, computer-assisted/ or image interpretation, computer-assisted/ or diagnosis, differential/ or specimen handling/ or biopsy/ or biopsy, needle/ or biopsy, large-core needle/ or dissection/ or microscopy/ or photomicrography/ or diagnostic tests, routine/ or early diagnosis/ or "early detection of cancer"/ or prognosis/ or neoplasm staging/ or treatment outcome/ or disease-free survival/ or progression-free survival/ or response evaluation criteria in solid tumors/ or treatment failure/

17. (diagnos$ or classif$ or prognos$ or predict$).tw.

18. 16 or 17

1. Embase 1974 to 2022 February 28

1. neoplasm/ or malignant neoplasm/ or solid malignant neoplasm/ or carcinoma/ or breast tumor/ or breast cancer/ or breast carcinoma/ or breast adenocarcinoma/ or metastatic breast cancer/ or ductal carcinoma/ or breast ductal carcinoma/ or lobular carcinoma/ or inflammatory breast cancer/ or metaplastic carcinoma/ or adenosquamous carcinoma/ or papillary carcinoma/ or medullary carcinoma/ or neuroendocrine carcinoma/ or adenoid cystic carcinoma/ or breast cancer molecular subtype/ or luminal A breast cancer/ or luminal B breast cancer/ or basal like breast cancer/ or triple negative breast cancer/ or estrogen receptor positive breast cancer/ or estrogen receptor negative breast cancer/ or human epidermal growth factor receptor 2 positive breast cancer/ or human epidermal growth factor receptor 2 negative breast cancer/ or "hereditary breast and ovarian cancer syndrome"/

2. (cancer or carcinoma$ or malignant).tw.

3. breast/ or mammary gland/

4. (breast or mammar$).tw.

5. 3 or 4

6. 2 and 5

7. 1 or 6

8. machine learning/ or artificial neural network/ or computer vision/ or automated pattern recognition/ or back propagation/ or bayesian learning/ or classification algorithm/ or classifier/ or clustering algorithm/ or computer heuristics/ or convolution algorithm/ or cross validation/ or data mining/ or deconvolution algorithm/ or detection algorithm/ or Dijkstra's algorithm/ or dimensionality reduction/ or dynamic time warping/ or empirical mode decomposition/ or feature detection/ or feature extraction/ or feature extraction algorithm/ or "feature learning (machine learning)"/ or feature ranking/ or feature selection/ or feature selection algorithm/ or fuzzy system/ or generalized method of moments/ or greedy algorithm/ or hidden markov model/ or imaging algorithm/ or iterative closest point/ or k nearest neighbor/ or kernel method/ or knowledge discovery/ or learning algorithm/ or Levenberg Marquardt algorithm/ or Markov jump/ or markov state model/ or maximum entropy model/ or maximum likelihood method/ or Metropolis Hastings algorithm/ or model predictive control/ or multicriteria decision analysis/ or multifactor dimensionality reduction/ or Needleman Wunsch algorithm/ or network learning/ or online analytical processing/ or perceptron/ or radial basis function/ or radial basis function/ or random forest/ or recursive feature elimination/ or recursive partitioning/ or relevance vector machine/ or risk algorithm/ or rough set/ or successive projections algorithm/ or semi supervised machine learning/ or superposition algorithm/ or supervised machine learning/ or support vector machine/ or unsupervised machine learning/

9. machine learning.mp. or machine learning/ or algorithm.mp. or algorithm/ or information processing.mp. or information processing/ or artificial intelligence.mp. or artificial intelligence/

10. 8 or 9

11. pathology/ or general pathology/ or histopathology/ or molecular pathology/ or pathological anatomy/ or neuropathology/ or histology/ or histometry/ or histophotometry/ or brain histology/ or liver histology/ or immunohistochemistry.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword heading word, floating subheading word, candidate term word]

12. (histopatholog$ or patholog$ or histolog$).tw.

13. ("whole slide imag$" or WSI or (digit$ and slide$)).tw.

14. 11 or 12 or 13

15. valid$.tw.

16. diagnosis/ or early diagnosis/ or tumor diagnosis/ or cancer diagnosis/ or early cancer diagnosis/ or differential diagnosis/ or diagnosis related group/ or tumor biopsy/ or diagnostic test/ or cancer test/ or laboratory diagnosis/ or molecular diagnosis/ or computer assisted diagnosis/ or quantitative diagnosis/ or diagnostic accuracy/ or diagnostic test accuracy study/ or diagnostic error/ or missed diagnosis/ or diagnostic reasoning/ or cancer grading/ or cancer staging/ or prognosis/ or cancer prognosis/ or disease course/ or survival/ or survival prediction/ or cancer survival/ or cancer specific survival/ or survival rate/ or mean survival time/ or median survival time/ or overall survival/ or survival analysis/ or long term survival/ or cancer free survival/ or disease free survival/ or disease specific survival/ or disease free interval/ or event free survival/ or progression free survival/ or recurrence free survival/ or local progression free survival/ or distant progression free survival/ or local recurrence free survival/ or distant recurrence free survival/ or local disease free survival/ or distant disease free survival/ or metastasis free survival/ or distant metastasis free survival/ or failure free survival/ or local failure free survival/ or regional failure free survival/ or distant failure free survival/ or treatment free survival/ or post treatment survival/ or cancer recurrence/ or cancer regression/ or tumor recurrence/ or tumor regression/ or recurrent disease/ or relapse/ or remission/ or prediction/ or computer prediction/ or forecasting/ or predictive validity/ or predictive value/ or outcome assessment/ or response evaluation criteria in solid tumors/ or treatment response/ or treatment response time/ or treatment failure/ or treatment outcome/ or disease worsening with drug treatment/ or outcomes research/

17. (diagnos$ or classif$ or prognos$ or predict$).tw.

18. 16 or 17

19. 7 and 10 and 14 and 15 and 18

1. CINAHL. Interface: EBSCOhost Research Databases

( ((MH "Neoplasms, Glandular and Epithelial") OR (MH "Carcinoma") OR (MH "Carcinoma, Adenoid Cystic") OR (MH "Carcinoma, Ductal") OR (MH "Carcinoma, Ductal, Breast") OR (MH "Carcinoma, Lobular") OR (MH "Carcinoma, Neuroendocrine") OR (MH "Carcinoma, Papillary") OR (MH "Carcinoma, Squamous Cell") OR (MH "Neoplasms, Cystic, Mucinous, and Serous") OR (MH "Neoplasms, Ductal, Lobular, and Medullary") OR (MH "Breast Neoplasms") OR (MH "Hereditary Breast and Ovarian Cancer Syndrome") OR (((MH "Breast") OR (TI breast OR AB breast) OR (TI mammar\* OR AB mammar\*)) AND ((TI cancer OR AB cancer) OR (TI carcinoma\* OR AB carcinoma\*) OR (TI malignant OR AB malignant)))) ) AND ( ((MH "Algorithms") OR (MH "Artificial Intelligence") OR (MH "Machine Learning") OR (MH "Deep Learning") OR (MH "Support Vector Machine") OR (MH "Decision Trees") OR (MH "Neural Networks (Computer)") OR (TI “machine learning” OR AB “machine learning”) OR (TI algorithm OR AB algorithm) OR (TI “Automatic Data Processing” OR AB “Automatic Data Processing”) OR (TI “computer aided detection” OR AB “computer aided detection”) OR (TI “artificial intelligence“ OR AB “artificial intelligence”)) ) AND ( ((MH "Pathology") OR (MH "Histology") OR (MH "Histocytochemistry") OR (MH "Immunohistochemistry") OR (TI histopathology\* OR AB histopathology\*) OR (TI patholog\* OR AB patholog\*) OR (TI histolog\* OR AB histolog\*) or ("whole slide imag\*" OR WSI OR (digit\* AND slide\*))) ) AND ( (TI valid\* OR AB valid\*) ) OR ( ((MH "Diagnosis") OR (MH "Decision Making, Clinical") OR (MH "Clinical Reasoning") OR (MH "Diagnosis, Computer Assisted") OR (MH "Image Interpretation, Computer Assisted") OR (MH "Diagnosis, Differential") OR (MH "Specimen Handling") OR (MH "Biopsy") OR (MH "Biopsy, Needle") OR (MH "Dissection") OR (MH "Microscopy") OR (MH "Diagnostic Tests, Routine") OR (MH "Early Diagnosis") OR (MH "Early Detection of Cancer") OR (MH "Prognosis") OR (MH "Neoplasm Staging") OR (MH "Treatment Outcomes") OR (MH "Treatment Failure") OR (TI diagnos\* OR AB diagnos\*) OR (TI diagnos\* OR AB diagnos\*) OR (TI classif\* OR AB classif\*) OR (TI prognos\* OR AB prognos\*) or (TI predict\* OR AB predict\*)) ).

1. IEEE via IEEE *Xplore*

(("Mesh\_Terms":"neoplasms, glandular and epithelial" OR "Mesh\_Terms":carcinoma OR "Mesh\_Terms":adenocarcinoma, mucinous OR "Mesh\_Terms":carcinoma, adenoid cystic OR "Mesh\_Terms":carcinoma, ductal OR "Mesh\_Terms":carcinoma, ductal, breast OR "Mesh\_Terms":carcinoma, lobular OR "Mesh\_Terms":carcinoma, mucoepidermoid OR "Mesh\_Terms":carcinoma, neuroendocrine OR "Mesh\_Terms":cystadenocarcinoma, mucinous OR "Mesh\_Terms":carcinoma, adenosquamous OR "Mesh\_Terms":carcinoma, papillary OR "Mesh\_Terms":carcinoma, squamous cell OR "Mesh\_Terms":"neoplasms, cystic, mucinous, and serous" OR "Mesh\_Terms":cystadenocarcinoma OR "Mesh\_Terms":"neoplasms, ductal, lobular, and medullary" OR "Mesh\_Terms":carcinoma, medullary OR "Mesh\_Terms":breast neoplasms OR "Mesh\_Terms":"hereditary breast and ovarian cancer syndrome" OR "Mesh\_Terms":inflammatory breast neoplasms OR "Mesh\_Terms":triple negative breast neoplasms OR "Mesh\_Terms":unilateral breast neoplasms) OR ((("Index Terms":cancer OR "Index Terms":carcinoma OR "Index Terms":malignant) OR ("Abstract":cancer OR "Abstract":carcinoma OR "Abstract":malignant)) AND (("Mesh\_Terms":breast OR "Mesh\_Terms":mammary glands, human) OR ("Index Terms":breast OR "Index Terms":mammary OR "Index Terms":mammarian) OR ("Abstract":breast OR "Abstract":mammary OR "Abstract":mammarian)))) AND (("Mesh\_Terms":algorithms OR "Mesh\_Terms":artificial intelligence OR "Mesh\_Terms":machine learning OR "Mesh\_Terms":deep learning OR "Mesh\_Terms":supervised machine learning OR "Mesh\_Terms":support vector machine OR "Mesh\_Terms":unsupervised machine learning OR "Mesh\_Terms":decision theory OR "Mesh\_Terms":decision trees OR "Mesh\_Terms":neural networks, computer) OR ("IEEE Terms":Machine learning algorithms OR "IEEE Terms":Artificial intelligence OR "IEEE Terms":Machine learning OR "IEEE Terms":Deep learning OR "IEEE Terms":Supervised learning OR "IEEE Terms":Unsupervised learning OR "IEEE Terms":Semisupervised learning OR "IEEE Terms":Support vector machines OR "IEEE Terms":Artificial neural networks) OR ("Index Terms":machine learning OR "Index Terms":artificial intelligence) OR ("Abstract":machine learning)) AND (("Mesh\_Terms":pathology OR "Mesh\_Terms":pathology, surgical OR "Mesh\_Terms":histology OR "Mesh\_Terms":histocytochemistry OR "Mesh\_Terms":immunohistochemistry) OR ("Index Terms":histopathology OR "Index Terms":histopathological OR "Index Terms":pathology OR "Index Terms":pathological OR "Index Terms":histology OR "Index Terms":histological) OR ("Abstract":histopathology OR "Abstract":histopathological OR "Abstract":pathology OR "Abstract":pathological OR "Abstract":histology OR "Abstract":histological) OR ("Index Terms":"whole slide images" OR "Index Terms":"whole slide imaging" OR "Index Terms":WSIs) OR ("Abstract":"whole slide images" OR "Abstract":"whole slide imaging" OR "Abstract":WSIs) OR (("Index Terms":digital OR "Index Terms":digitizing OR "Index Terms":digitized) AND ("Index Terms":slides)) OR (("Abstract":digital OR "Abstract":digitizing OR "Abstract":digitized) AND ("Abstract":slides))) AND (("Index Terms":valid\*) OR ("Abstract":valid\*)).

1. MICCAI via Springer link

(Breast OR mammary) AND (Cancer OR carcinoma OR tumor OR tumors OR tumour OR tumours OR neoplasm OR neoplasms OR malignant) AND ("artificial intelligence" OR "machine Learning" OR "Deep learning" OR "neural network" OR "computer vision" OR Algorithm OR Algorithms) AND (pathology OR histology OR histopathology OR histological OR histopathological OR "whole slide images" OR "whole slide imaging" OR WSI OR WSIs OR "digital slides") AND (valid\*). Filters: within miccai & Conference Paper & 2010 - 2022.

1. SPIE conferences

ABSTRACT:(Breast OR mammary) AND ABSTRACT:(Cancer OR carcinoma OR tumor OR tumors OR tumour OR tumours OR neoplasm OR neoplasms OR malignant) AND ABSTRACT:("artificial intelligence" OR "machine Learning" OR "Deep learning" OR "neural network" OR "computer vision" OR Algorithm OR Algorithms) AND ABSTRACT:(pathology OR histology OR histopathology OR histological OR histopathological OR "whole slide images" OR "whole slide imaging" OR WSI OR WSIs OR "digital slides") AND ABSTRACT:(validated OR validation OR valid OR validating OR validate). Filters: Proceedings & 2010 – 2022.