European Journal of Cancer 155 (2021) 200–215

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Review

Gastrointestinal cancer classification and prognostication from histology using deep learning: Systematic review

Sara Kuntz [a], Eva Krieghoff-Henning [a], Jakob N. Kather [b], Tanja Jutzi [a], Julia Höhn [a], Lennard Kiehl [a], Achim Hekler [a], Elizabeth Alwers [c], Christof von Kalle [i], Stefan Fröhling [d], Jochen S. Utikal [e,f], Hermann Brenner [c,g,h], Michael Hoffmeister [c], Titus J. Brinker [a,*]

[a] *Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[b] *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*
[c] *Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[d] *Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[e] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[f] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[g] *Division of Preventive Oncology, German Cancer Research Center (DKFZ), National Center for Tumor Diseases (NCT), Heidelberg, Germany*
[h] *German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[i] *Department of Clinical-Translational Sciences, Charité University Medicine and Berlin Institute of Health (BIH), Berlin, Germany*

Received 16 June 2021; accepted 6 July 2021
Available online 11 August 2021

KEYWORDS
Gastrointestinal cancer;
Colorectal cancer;
Gastric cancer;
Esophageal cancer;
Pathology;
Digital biomarker;
Artificial intelligence;
Deep learning;

Abstract ***Background:*** Gastrointestinal cancers account for approximately 20% of all cancer diagnoses and are responsible for 22.5% of cancer deaths worldwide. Artificial intelligence–based diagnostic support systems, in particular convolutional neural network (CNN)–based image analysis tools, have shown great potential in medical computer vision. In this systematic review, we summarise recent studies reporting CNN-based approaches for digital biomarkers for characterization and prognostication of gastrointestinal cancer pathology.

Methods: Pubmed and Medline were screened for peer-reviewed papers dealing with CNN-based gastrointestinal cancer analyses from histological slides, published between 2015 and 2020.Seven hundred and ninety titles and abstracts were screened, and 58 full-text articles were assessed for eligibility.

[*] *Corresponding author*: Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.ejca.2021.07.012
0959-8049/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Convolutional neural network

Results: Sixteen publications fulfilled our inclusion criteria dealing with tumor or precursor lesion characterization or prognostic and predictive biomarkers: 14 studies on colorectal or rectal cancer, three studies on gastric cancer and none on esophageal cancer. These studies were categorised according to their end-points: polyp characterization, tumor characterization and patient outcome. Regarding the translation into clinical practice, we identified several studies demonstrating generalization of the classifier with external tests and comparisons with pathologists, but none presenting clinical implementation.

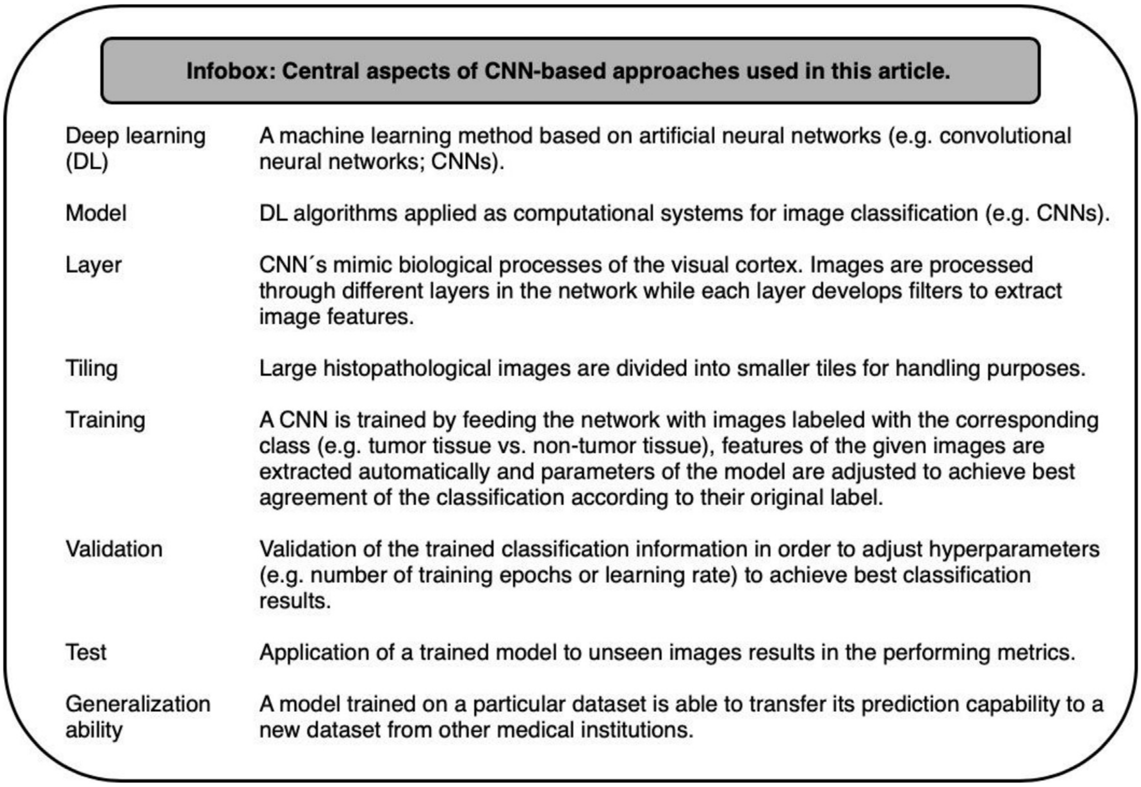
Conclusions: Results of recent studies on CNN-based image analysis in gastrointestinal cancer pathology are promising, but studies were conducted in observational and retrospective settings. Large-scale trials are needed to assess performance and predict clinical usefulness. Furthermore, large-scale trials are required for approval of CNN-based prediction models as medical devices.

© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Background

Gastrointestinal cancers comprise esophageal, gastric, colon and rectal tumors. As reported by the WHO, approximately 3.5 million new gastrointestinal cancer cases worldwide have been recorded in 2018. Whereas the incidence of esophageal cancer is comparatively low, gastric cancer (GC) is the fifth most frequent type of cancer and the third leading cause of cancer death. Colorectal cancer (CRC) is the third most common cancer worldwide after lung and breast cancer but the second leading cause of cancer death [1]. Although various predictive and prognostic biomarkers exist, high mortality rates for gastrointestinal cancer patients show that there is still potential to improve diagnostics to pave the way for more personalised therapy strategies leading to a better prognosis and/or fewer side effects.

Biomarkers in oncology are defined as indicators as well for the presence as for the absence of cancer or for tumor behavior such as responsiveness to therapy or recurrence risk [2]. A number of molecular biomarkers in oncology has been developed and enables more fine-tuned tumor treatment, at least for the more common cancer types [3,4]. Unfortunately, these additional molecular methods often require additional tissue material, this can be very time-consuming and cost-



Infobox: Central aspects of CNN-based approaches used in this article.

Deep learning (DL) A machine learning method based on artificial neural networks (e.g. convolutional neural networks; CNNs).

Model DL algorithms applied as computational systems for image classification (e.g. CNNs).

Layer CNN´s mimic biological processes of the visual cortex. Images are processed through different layers in the network while each layer develops filters to extract image features.

Tiling Large histopathological images are divided into smaller tiles for handling purposes.

Training A CNN is trained by feeding the network with images labeled with the corresponding class (e.g. tumor tissue vs. non-tumor tissue), features of the given images are extracted automatically and parameters of the model are adjusted to achieve best agreement of the classification according to their original label.

Validation Validation of the trained classification information in order to adjust hyperparameters (e.g. number of training epochs or learning rate) to achieve best classification results.

Test Application of a trained model to unseen images results in the performing metrics.

Generalization ability A model trained on a particular dataset is able to transfer its prediction capability to a new dataset from other medical institutions.

intensive leading to capacity difficulties in clinical workflows [5,6].

An enormous chance for precision oncology may originate from digital biomarkers based on whole-slide image (WSI) analysis. Altered signaling in cells, which may for instance be induced by gene mutations, can result in morphologic changes in affected cells [7–9]. In cancer diagnostics, pathology slides are routinely prepared from biopsies or resections and tumor tissue is stained with hematoxylin and eosin (HE). With a whole-slide scanner, histological slides can be converted into digital images containing complex visual information that can be extracted by artificial neural networks [10]. CNNs are state-of-the-art deep-learning (DL) algorithms that have been reported as very successful computational systems for image classification (Fig. 1) [10]. Many challenges remain to be addressed prior to clinical implementation of digital biomarkers, but it is conceivable that DL concepts may have tremendous benefits for diagnostic and therapeutic decisions [5,11,60–69]. Advances of digital biomarkers in DL-assisted tumor pathology for common cancer types have been reviewed recently [5,6].

In this review, we focused on CNN-based models that analyse histological images of gastrointestinal cancer and precursor lesions to determine tumor characteristics and behavior and their development towards clinical applicability. We concentrated on studies that characterise precursor lesions, perform molecular characterization of tumors or deal with prognostic or predictive biomarkers.

2. Methods

2.1. Search strategy

PubMed and Medline databases were searched for peer reviewed articles published in English. The search terms *gastrointestinal cancer*, *esophageal cancer, gastric cancer, stomach cancer, colorectal cancer, colon cancer, rectal cancer, deep learning, convolutional neural networks, histological images* and *whole-slide image* were combined. Search results were screened manually by two reviewers for relevance, and papers were chosen based on titles and abstracts.

2.2. Study selection

This review only includes studies that used CNNs to analyse HE images from gastrointestinal cancer patients. Only studies published between 2015 and 2020 were considered. To ensure statistical robustness of the presented classifiers, we excluded studies based on less than 50 WSIs or patients. Studies lacking a precise description of statistical data or information on the underlying dataset regarding the number of patients and/or WSIs were also not considered.

2.3. Study analysis

In this review, we designed a level system that defines evaluation categories according to clinical applicability

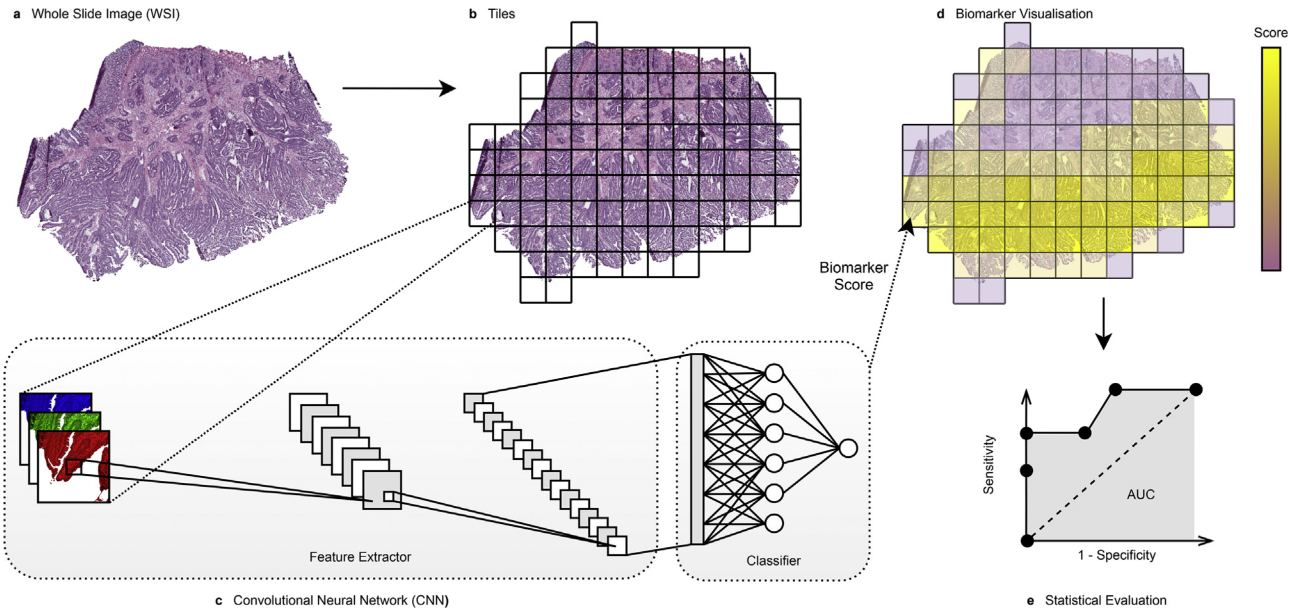


Fig. 1. Simplified pipeline of CNN-based image analysis in pathology. a Whole-slide image with HE staining of CRC (from TCGA). b Tessellation into smaller image tiles. c The CNN is composed of a feature extractor transforming tiles into feature vectors and a classifier operating on those vectors to discriminate between predefined classes. d Output scores of the CNN can be presented in a visualization map to enhance interpretability and e the statistical evaluation of the model is represented by plotting performance metrics like AUC. Abbreviations: AUC Area under the receiver operating curve, CNN Convolutional neural network, CRC colorectal cancer, HE hematoxylin and eosin, TCGA The Cancer Genome Atlas.

(Table 1). A classifier has to complete all levels to succeed in the aspired integration in clinical diagnostics.

2.4. Applied performance metrics

A CNN-based classification model assigns each image to a predefined class and can thereby distinguish between two (binary classification) or multiple classes (multiclass classification). The class assignment can be either correct or incorrect with respect to the pre-defined actual class or "label" of that image. Different statistical metrics exist to evaluate the quality of a binary classifier. Accuracy (acc) indicates the percentage of correct classifications. Sensitivity defines the correctly identified positives, whereas specificity regards the negative class. The area under the receiver operating characteristic curve (AUC or AUROC) indicates the performance of the classifier across all possible classification thresholds used to assign an item to the positive or negative class. For survival analyses, a commonly used metric is the hazard ratio (HR). For binary or dichotomised classifiers, the HR quantifies the ratio of the hazard rates of the two groups compared. The hazard ratio can also be used for ordinal classifiers or continuous variables to indicate the relative increase in hazard rates per one-category or one-unit increase in the classifier, respectively.

If an analysed study presented multiple DL models, only the approach with the best statistical values was referred to in this review and summarised in Table 2.

3. Results

In total, we identified 16 studies fulfilling our search criteria in the systematic literature search (Fig. 2; summarised in Table 2). Three papers presented classifiers for CRC polyp characterization, five studies described molecular characteristics influencing therapy selection and eight analyses were conducted to determine patient outcome from histological images (Fig. 3). We did not find any study analyzing tumor characteristics or predictions in esophageal cancer with respect to our search criteria. For GC, we identified only one study exclusively concentrating on the prediction of cancer-specific death in gastric tumor patients. Most studies were published for CRC or for a combined consideration of GC and CRC. We excluded three studies from this review because the datasets were too small with respect to our inclusion criteria [12–14] and another three studies because the number of WSIs was not reported and only numbers of tiles were indicated [15–17]. All excluded studies are briefly summarised in Supplementary Table 1.

Table 1
Structure to evaluate clinical applicability of a classifier.

Level	Characteristics of CNN-based classifiers in cancer diagnosis
1	Training and testing with only one usually comparatively small dataset
2	Testing with an external dataset to demonstrate its generalization independent from varying sample conditions
3	Comparison with the performance of pathologists to reveal its additional value
4	Testing in a clinical setting as a supportive system in combination with diagnosis of pathologists (At this point, the classifier is confronted with acute patient data for the first time and the classification result has influence on the patient's therapy. This application represents a change from a retrospective to a prospective analysis.)
5	Implementation in clinical routine to actively support cancer diagnosis

3.1. Polyp characterization

We found three papers dealing with CRC polyp characterization as a strategy to reduce CRC incidence by improved early diagnosis of adenomas. Colorectal polyps can be subdivided into various types with individual malignancy risks, while early detection endoscopy gives opportunity to surgically remove high-risk polyps before transformation into cancer occurs [18–20].

Korbar *et al.* [18] trained a CNN to characterise colorectal polyps in preventive CRC screening. Currently, the screening method that is considered the gold standard is colonoscopy. Histopathological characterization can supply additional information to differentiate between high-risk and low-risk polyps. Histopathology as confirmation is specifically required to distinguish hyperplastic from neoplastic polyps, only the latter being cancer precursors. The CNN model classified normal tissue and five types of polyps: hyperplastic polyp, sessile serrated polyp, traditional serrated adenoma, tubular adenoma, and tubulovillous/villous adenoma; designated according to the guidelines for colorectal cancer risk assessment and surveillance [21]. The classifier accurately differentiated between high-risk or low-risk polyps or normal cases (acc for all classes: 89.5–95.8%). There were two major sources of errors: low-risk polyps had the tendency to be classified as normal tissue, and the differentiation between hyperplastic and sessile serrated polyps was error-prone. This work was limited to a single dataset.

The same group published an updated version of their colorectal polyp classifier with an extended test set originating from 24 medical institutions [19]. The performance of the model was also compared with the performance of pathologists. Classification was reduced to four types: tubular adenoma, tubulovillous or villous adenoma, hyperplastic polyp and sessile serrated adenoma. The analysis of the internal test set revealed accuracies of 92.4–95.5% which were on par with the performance of five pathologists (acc 89.8–94.3% depending on the polyp class). The external dataset from multiple institutions was classified with accuracies ranging from 84.5 to 89.5%, also with no statistical differences to the performance by pathologists. The error analysis showed that the CNN made similar divergent classifications as the pathologists, mostly subclassifications within the adenomatous lesions or the serrated lesions.

Song *et al.* [20] presented a CNN model which was trained and validated on a cohort from a Chinese Hospital and performed a binary classification to detect the presence of adenoma. The classification system achieved an accuracy of 90.4% and performed on par with pathologists. To test the model's generalization ability, external datasets from two other medical institutions were analysed and the model maintained its high accuracy (>90%). Before starting clinical trials, the authors plan to increase the training set with more adenoma subtypes.

3.2. Molecular characterization

We identified five studies aiming to predict molecular characteristics from CRC images. No studies for esophageal cancer or GC with respect to molecular characterization have been found. The molecular signature of tumors is important to select a targeted and individual treatment against the underlying molecular key components of the tumor and to enhance the responsiveness of a selected therapy. Genetic alterations often result in cell morphological changes, so that CNN-based image analysis has the potential to reveal the genetic profile from tumor tissue [22–26].

An early approach for image-based genetic subtyping of CRC was presented in 2017 [22]. The analysis was based on CRC subtypes proposed by Budinska *et al.* [27] (subtypes A-E) which is very similar to the consensus molecular subtypes (CMS) proposed by Guinney *et al.* [28]. The dataset was obtained from one clinical trial. The multiclass classifier achieved an overall accuracy of 0.84, with high sensitivity of subtype A (1.0; serrated or papillary morphology) but poor results for subtype E (0.57; group with mixed phenotypes). Next, relapse-free survival of the different subtypes (lower risk [subtype A and B] versus higher risk [subtypes C, D, E]) was analysed and the image-based prediction (HR = 1.56) provided similar results as the molecular-based analysis (HR 1.75), showing that the predictions from the DL model could be used as a prognosticator for relapse-free survival. The study was limited by the dataset from one medical institute.

Another model for predicting CRC subtypes from histological images was developed on the basis of the CMS system [26,28]. The authors showed that the subtype can be gleaned from HE images as distinct image phenotypes designated image-based CMS (imCMS). The imCMS classifier was trained on resection samples from a multicenter cohort and tested on two external cohorts. Using domain adversarial training [29] improved the classification accuracy so that the imCMS classifier achieved AUCs of 90, 84 and 85% in the internal testset, in the publicly available Cancer Genome Atlas (TCGA [30]) resection samples and in the GRAMPIAN cohort biopsy samples, respectively. It is noteworthy that the classifier could be successfully applied on resection samples and preoperative biopsy samples. The visual revision of tiles with the highest prediction confidence for each imCMS revealed that histological patterns correlate well with the biological characteristics of CMS and a high level of agreement was found between both classification approaches. Moreover, with regard to the transcriptional tumor heterogeneity there was a high level of agreement between the computational and molecular classification schemes. In a prognostic analysis, the imCMS status produced comparable values as the established CMS system in all three cohorts. Furthermore, the imCMS classifier was able to predict the status of samples previously unclassifiable by RNA expression profiling for technical reasons.

The responsiveness of gastrointestinal cancer to immunotherapy strongly correlates with microsatellite instability (MSI) [31]. In clinical diagnostics, MSI is diagnosed by immunochemistry or genetic analyses by polymerase chain reaction. However, MSI can also be detected on HE slides with CNN-based image analysis [23]. On the basis of automated tumor detection (AUC > 0.99), MSI and microsatellite stability were classified in GC and CRC samples embedded in paraffin as well as in CRC snap-frozen samples. The approach successfully identified the microsatellite status in paraffin (GC samples AUC = 0.81; CRC samples AUC = 0.77) and snap-frozen samples (CRC samples AUC = 0.84). An external multicenter cohort was used to demonstrate the robustness for CRC samples (AUC = 0.84). For GC, external testing was performed with data from a Japanese cohort which resulted in a lower AUC of 0.69 which may indicate that generalization across different ethnic populations is challenging. Because MSI is a pan-cancer biomarker, an additional external dataset of endometrial cancer tissue was analysed and still the DL algorithm classified MSI with an AUC of 0.75.

A follow-up study was presented by the same group with the focus on gene status prediction of MSI and mismatch-repair deficiency (dMMR) as its cause in large external cohorts [24]. In clinical diagnostics, both genetic alterations are used for appropriate treatment selection in CRC patients. Microsatellite instability and dMMR are associated with distinct morphological patterns [32] revealing the enormous potential benefit of image-based genetic predictions. In contrast to molecular analysis, computational analysis might be used for high-throughput pre-screening and low-cost evaluation of GC and CRC tissues [24]. The classifier was trained

Table 2

The table summarises all studies analysed in this review with respect to the classification task, the aspect of clinical implementation, the datasets and the applied metrics. For multiclass analyses, the authors use multiple binary classification strategies, such as 'one versus rest' (OvR) or 'one versus one' (OvO). Binary classification metrics such as AUROC, sensitivity and specificity are provided using OvR with the lowest and highest values given, if not otherwise specified. Hazard ratio values are based on dichotomised data, if not indicated otherwise. Additional information on the utilised datasets regarding data sources, samples sizes and cancer types is summarised in Supplementary Table 2. Abbreviations used in the table are explained below.[a]

Study	Cancer type task	Classification	Clinical applicability	Datasets (#WSIs/ #patients)	# Samples (#WSIs/#patients)			Results					CNN
					Training	Internal test	External test	Acc [%]	Sens [%]	Spec [%]	AUROC [%]	HR (95% CI)	
Adenoma/polyp characterization													
[18]	CRC	6 classes	level 1	697/-	458/-	239/-	–/–	89.5-95.8[b]	81.1-95.8				ResNet-D
[19]	CRC	4 classes	level 1–3	1) 508/508 2) 238/179	326/- (1)	157 (1)	238/- (2)	92.4-95.5[b] 84.5-89.5[2]	78.9-97.1 60.3-97.6	94.6-97.5 87.8-97.5			ResNet
[20]	GC	binary classification: adenoma yes/no	level 1–3	1) 411/- 2) 63/- 3) 105/-	177/- (1)	194/- (1)	168/- (2 + 3)	90.4 >90%	89.3 90 (2) 93.4 (3)	79 92.3 93.2	92 –		DeepLabv2 with ResNet-34
Tumor subtyping													
[22]	CRC	5 classes: CRC subtypes a) low-risk versus high-risk group	level 1	300/-	100/-	200/-	–/–	84	57-100			RFS 1.56 (1.10–2.21)	VGG-F
[23]	GC, CRC	a) 3 classes tumor detection b) binary classification: MSI versus MSS	level 1+2	1) 94/81 2) 1062/1026 3) 378/- 4) 185/185 5) 492/492	a) 94/81 (1) b) -/644 (2)	b) -/418 (2)	b) -/378 (3) -/185 (4) -/492 (5)				a) >99[c] b) 77[d]; 84 b) 84 (3) 69 (4) 75 (5)		ResNet18
[24]	CRC	a) binary tumor detection b) multiclass: MSI and dMMR status	level 1+2	1) -/426 2) -/2013 3) -/1770 4) -/2197 5a) -/771, resection 5b) -/1531, biopsy	-/5500 (1–4)	-/906 (1–4)	-/771 (5a) -/1531/5b)		98	52	92 92-96 (5a) 88-89 (5b)		Shufflenet (taken from Kather *et al.*, 2020 [33])
[25]	CRC	binary classification: MSI versus MSS	level 1+2	1) -/429 2) -/785	-/300 (1)	-/129 (1)	-/785 (2)		91	77	88 85		ResNet18
[26]	CRC	a) 5 classes: CRC subtypes b) prediction of patient outcome imCMS1 versus imCMS2 imCMS3 versus imCMS2 imCMS4 versus imCMS2	level 1+2	1) 666/362, resection 2) 468/463, resection 3) 406/223, biopsy sample	510/278 (1)	N/A	431/430 (2) 265/144 (3)				a) 90 (1)[c] a) 84 (2)[c] 85 (3)[c]	b) RFS (3) 1.43 (0.13, 15.75) 3.18 (0.58, 17.35) 5.75 (1.05, 31.41)	Inception-V3

(continued on next page)

Table 2 (*continued*)

Study	Cancer type task	Classification	Clinical applicability	Datasets (#WSIs/ #patients)	# Samples (#WSIs/#patients)			Results					CNN
					Training	Internal test	External test	Acc [%]	Sens [%]	Spec [%]	AUROC [%]	HR (95% CI)	
Patient outcome													
[34]	CRC	Prediction: low-risk patient versus high-risk patient	level 1, 3	-/420	-/220	-/140	−/−				69	b) DSS 2.3 (1.79−3.03)	VGG16
[35]	RC	a) 9 tissue classes b) prediction: stroma-high versus stroma-low group	level 1, 3	129/-	74/-	55/-	−/−	a) 94.6	a) 90.8-98.4	a) 97.7 −99.9		b) DSS 2.48 (1.29−4.78) DFS: 2.05 (1.11−3.78)	ConvNets
[36]	CRC	a) 9 tissue classes b) prediction of OS, DSS, RFS: high versus low 'deep stroma score'	level 1−3	1) 86/- 2) 25/- 3) 862/500 4) 409/409	86/- (1)	25/- (2)	-/909 (3 + 4)	a) 94.3	a) 59-100	a) 99-100	a) 98-100	b) OS 1.63 (1.14 −2.33) DSS 2.29 (1.5 −3.48) RFS 1.92 (1.34−2.76)	VGG19
[37]	CRC	a) 9 tissue classes b) prediction of OS: stroma-high versus stroma-low group	level 1−3	1) 85/- 2) 86/- 3) 158/499 4) 25/- 5) 72/315	277/- (1 −3)	42/- (3)	25/- (4) 48/- (5)	a) 95.72 (4) 97.46 (5)				b) OS (3) 1.72 (1.24 −2.37) b) OS (5) 2.08 (1.26 −3.42)	VGG19
[38]	CRC	Prediction of CSS a) good versus uncertain and poor prognosis b) good and uncertain versus poor c) good versus poor	level 1+2	1) -/160 2) -/576 3) -/970 4) -/767 5) -/1122	-/2473 (1 −4)		-/1122 (5)	a) 76 b) 67	a) 52 b) 69	a) 78 b) 66	c) 71.3	CSS poor versus good: 3.04 (2.07−4.47) uncertain versus good: 1.89 (1.14 −3.15)	DoMore v1
[39]	CC, stage III	a) 9 tissue classes b) prediction: high- versus low-recurrence group and poor- versus good-prognosis group	level 1+2	1) 86/- 2) 168/168 3) 47/47	187/- (1 + 2)	67/- (2)		75.5				b) recurrence groups 8.98 (2.82 −28.53) prognosis groups 10.69 (2.91 −39.27)	InceptionResNetV2
							47/- (3)	77				b) recurrence group 10.27	

									(2.18−48.47) prognosis group 5.03 (1.79 −14.13)	
[40]	CRC	a) 4 tissue classes b) prediction of distant metastasis[e]	level 1	1) -/72 2) -/30	70%	30%	−/−	a) 90.4-99.9[f] b) 58.6-63.8	b)[g] 2.11 (0.79 −5.60) 0.39 (0.19 −0.80)	Spatially Constrained CNN
[41]	GC	prediction of CSD	level 1	-/248	N/A	N/A	−/−		Ki67/CD20 1.36 (1.07 −1.74) CD20/CD68 1.34 (1.07 −1.67) Ki67/CD68 1.47 (1.15 −1.89)	GoogLeNet

[a] acc, accuracy; AUROC, area under the receiver operating characteristic; CC, colon cancer; CRC, colorectal cancer; CSD, cancer-specific death; CSS, cancer-specific survival; DFS, disease-free survival; DSS, disease-specific survival; FFPE, formalin-fixed paraffin-embedded; GC, gastric cancer; HR, hazard ratio; MSI, microsatellite instability; MSS, microsatellite stability; N/A, not available; RC, rectal cancer; RFS, recurrence-free survival; sens, sensitivity; spec, specificity.

[b] Range of OvR accuracies.

[c] Macro average of OvR.

[d] AUC of 77 refers to CRC samples, AUC of 84 refers to GC samples.

[e] Distant metastasis-free survival analysis is based on features ‘connection frequency of smooth muscle ratio’ and ‘appearance based on inflammatory tissue’.

[f] OvO.

[g] HR is based on interquartiles.

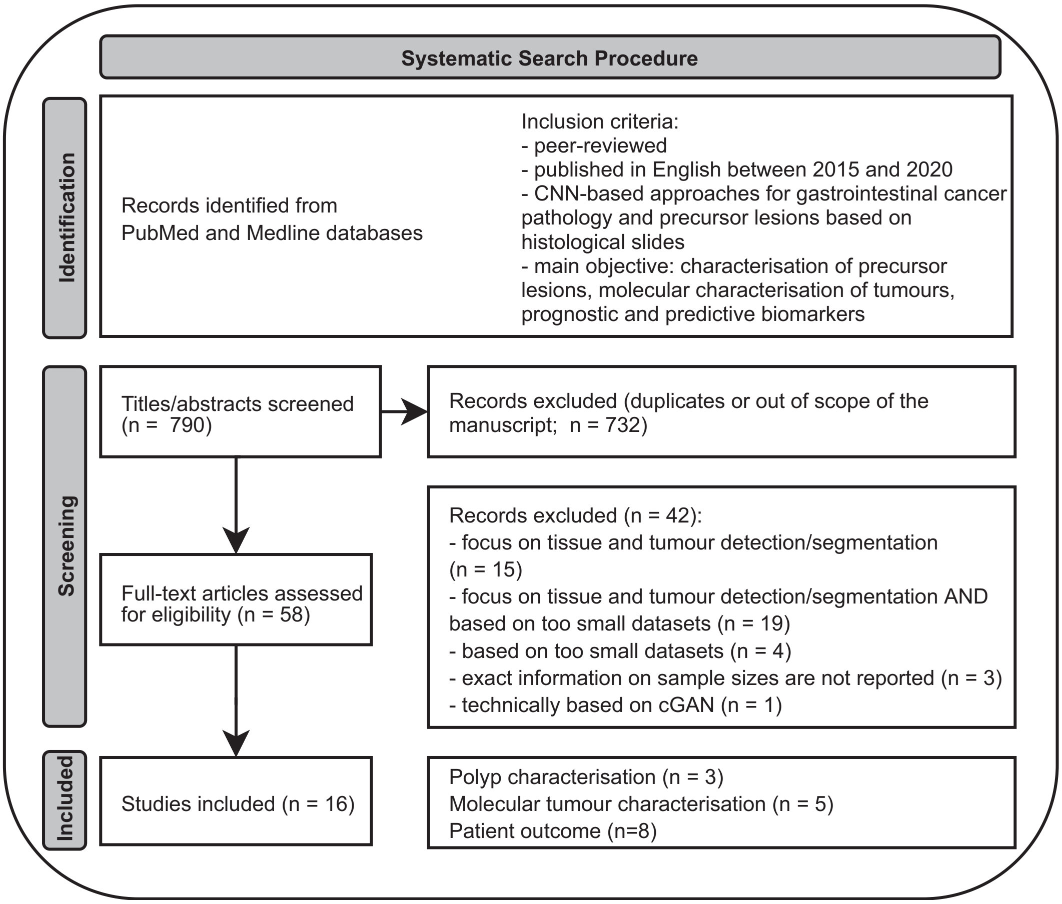

Systematic Search Procedure
Identification
Records identified from PubMed and Medline databases
Inclusion criteria:
- peer-reviewed
- published in English between 2015 and 2020
- CNN-based approaches for gastrointestinal cancer pathology and precursor lesions based on histological slides
- main objective: characterisation of precursor lesions, molecular characterisation of tumours, prognostic and predictive biomarkers
Screening
Titles/abstracts screened (n = 790)
Records excluded (duplicates or out of scope of the manuscript; n = 732)
Full-text articles assessed for eligibility (n = 58)
Records excluded (n = 42):
- focus on tissue and tumour detection/segmentation (n = 15)
- focus on tissue and tumour detection/segmentation AND based on too small datasets (n = 19)
- based on too small datasets (n = 4)
- exact information on sample sizes are not reported (n = 3)
- technically based on cGAN (n = 1)
Included
Studies included (n = 16)
Polyp characterisation (n = 3)
Molecular tumour characterisation (n = 5)
Patient outcome (n=8)

Fig. 2. Flow diagram of systematic literature search inspired by PRISMA guidelines [59].

on a very large and diverse cohort to establish a classifier fulfilling the requirements for a future clinical implementation. To predict MSI and dMMR status, a retraining of the previously described algorithm [33] with four different international cohorts with over 8000 WSIs from CRC was performed. Training on each cohort alone revealed performances with AUROCs of 0.74–0.92 and 0.67–0.82 for intracohort or intercohort testing, respectively. A 3-fold cross-validation on all four international cohorts yielded an AUROC of 0.92. Incremental training sample size up to 5500 images showed best results with approximately 5000 training images. The main objective of the paper was to validate the MSI and dMMR classifier on an external patient cohort with 771 patients. The classification performance evaluation yielded an AUROC of 0.96. Prediction performance was stable over different CRC features such as right- versus left-sided tumors or colon versus rectal cancer. According to BRAF and KRAS status, slightly better performances were obtained for wild-type tumors. Microsatellite instability prediction was stable over CRC stages I–III, but a little less reliable in stage IV. No difference in performance was observed for histological tumor grading. When the final classifier that was trained on resection samples, was tested on histological images of biopsies, the AUROC was reduced to 0.78. However, the AUROC could be restored to 0.89 when the classifier was trained on biopsy data before. This study was limited by variations in genetic testing methods and tumor heterogeneity. Furthermore, the aspect of hereditary or sporadic MSI/dMMR, as well as ethnic differences, was not investigated due to the lack of appropriate samples in the datasets.

Another recent study also addressed the prediction of MSI and presented a combined approach of a CNN with two independent multi-instance learning (MIL) pipelines resulting in the Ensemble Patch Likelihood Aggregation (EPLA) model [25]. In contrast to the method of Kather *et al.* [23], the EPLA classification is based on a few key patches that contribute most to the prediction of MS status and need to be profoundly weighted in the calculation of MS scores. When trained and tested on the CRC data from the TCGA cohort, EPLA achieved an AUC of 0.88. On external data from a Chinese CRC cohort, the model performance was initially reduced (AUC 0.65) but could be augmented to an AUC of 0.85 by transfer learning with 10% of the Asian cohort data. Transfer learning enables EPLA to overcome data diversity regarding tissue preparation (flash-frozen versus paraffin-embedded samples) and

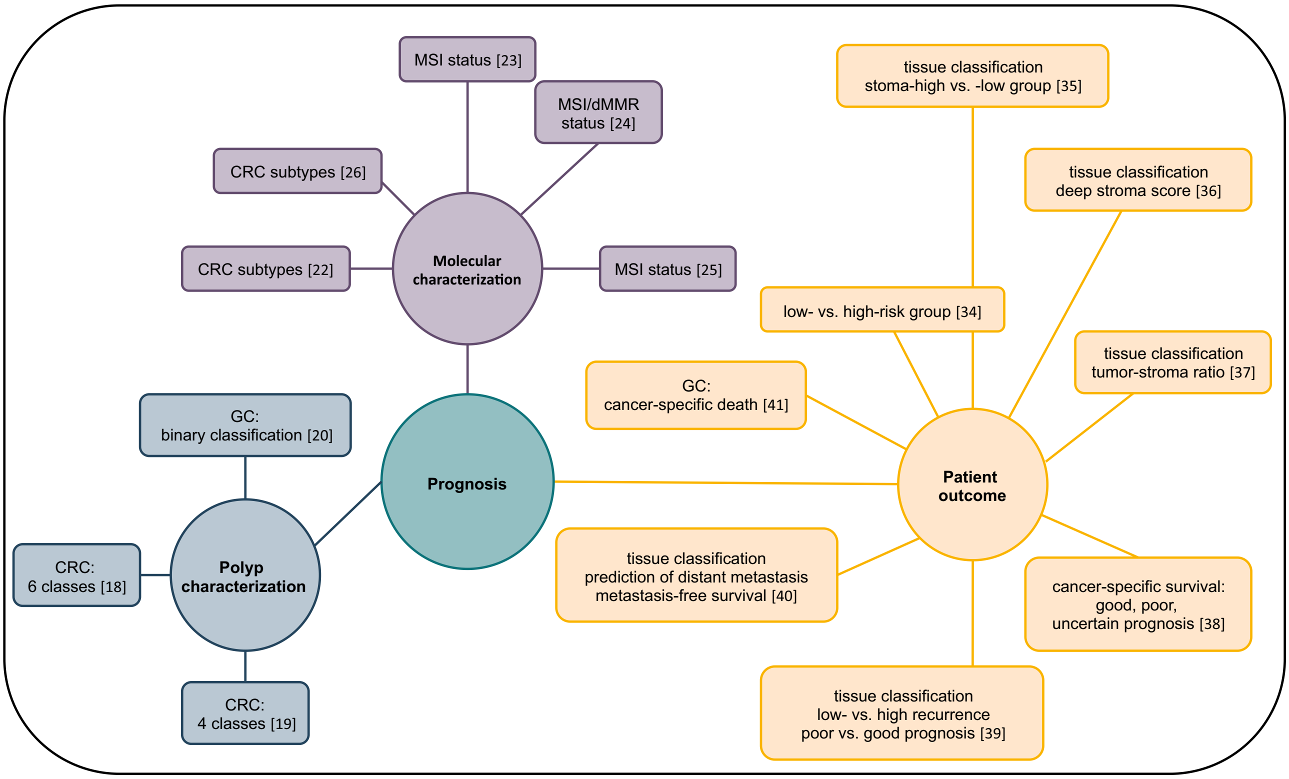

MSI status [23]
MSI/dMMR status [24]
CRC subtypes [26]
CRC subtypes [22]
Molecular characterization
MSI status [25]
GC: binary classification [20]
Prognosis
CRC: 6 classes [18]
Polyp characterization
CRC: 4 classes [19]
tissue classification stoma-high vs. -low group [35]
tissue classification deep stroma score [36]
low- vs. high-risk group [34]
tissue classification tumor-stroma ratio [37]
GC: cancer-specific death [41]
Patient outcome
tissue classification prediction of distant metastasis metastasis-free survival [40]
cancer-specific survival: good, poor, uncertain prognosis [38]
tissue classification low- vs. high recurrence poor vs. good prognosis [39]

Fig. 3. Overview of CNN applications in the field of gastrointestinal cancer pathology and prognosis. The reported studies are subdivided into three groups based on their endpoints: polyp characterization, molecular characterization and prediction of patient outcome. Studies are indicated by the superscript reference number. Abbreviations: CRC colorectal cancer, GC gastric cancer, MSI microsatellite status, dMMR mismatch-repair deficiency.

differences in ethnicity (Western population in TCGA cohort versus Asian cohort). The model was able to classify MS status equally successful in cases with CRC stage I–III or IV. To increase explainability, the authors compared the predicted image signatures with transcriptomic profiles and DNA mutation clusters profiled by whole-exome sequencing. Known genomic biomarkers or pathways correlate well with the prediction supporting the interpretability of the EPLA model.

3.3. Patient outcome

Eight studies presented digital biomarkers to determine tumor prognosis and performed analyses of patient outcome by assigning patients into low-risk or high-risk groups [34–41]. Because studies indicate that only some CRC patients benefit from adjuvant treatments, prognostic biomarkers are needed for the selection of appropriate cancer therapies [42].

Bychkov *et al.* [34] presented a model to predict five-year disease-specific survival of CRC patients directly from tumor tissue microarray samples without tissue classification as an intermediate step. They combined a CNN and a recurrent neural network for feature extraction and classification. The study was based on images from 420 CRC patients, and an individual risk score was calculated to classify patients into a low- or high-risk group (HR 2.3). However, the performance of the classification model (AUC of 0.69) compared favorably with the visual assessment of pathologists (AUC of 0.58) or the histological grade assessed at the time of primary diagnosis (AUC of 0.57). The classifier was not tested on an external dataset.

Geessink *et al.* [35] published a DL prognosticator for disease-specific risk determination based on semi-automatically quantification of the tumor-stroma ratio (TSR) in rectal cancer. The ratio of tumor and intra-tumoral stroma is a strong predictor for survival in cancer patients with solid tumors [43–46]. The DL model was trained to segment relevant tissue types and was applied to image regions with manually pre-annotated stroma hot spots. Based on the tissue segmentation, the TSR was calculated, and patients were classified either into a stroma-low or stroma-high group. As expected, high intra-tumoral stroma content was associated with a poor prognosis. TSR was a prognostic factor for disease-specific survival (HR = 2.48) and for disease-free survival (HR = 2.05). Geessink *et al.* [35] pointed out that in contrast to a visually assessed TSR by human experts, the DL-based TSR can serve as a prognostic factor independent of clinicopathological variables. The algorithm was not tested on an external dataset. The proposed semi-automated TSR calculation can only be applied to therapy-naive patients because adjuvant treatment modifies tissue characteristics. One future goal of the authors is to develop a fully automated process to eliminate the requirement of manually annotated hot-spot regions for TSR quantification.

Kather *et al.* also determined the prognosis of CRC patients from histological slides but included an intermediate step of tissue classification [36]. The DL model was based on a large dataset from the TCGA cohort [30] and two German medical centers. The tissue classifier achieved an overall nine-class accuracy close to 99% and demonstrated its robustness on an external dataset (acc, 94.3%). The CNN classifier also generated comparable results to standard molecular analyses in assigning the CRC samples to one of the consensus molecular subtypes (CMS [28]). The abundance of defined tissue components could be aggregated in a prognostic score. The 'deep stroma score' was calculated from the non-tumor components of the tissue samples extracted by the classifier and served as an independent prognostic factor for overall survival at all tumor stages (HR = 1.99), with high prognostic power especially at advanced tumor stages. In contrast, pathologist-annotated stromal tissue estimation was surprisingly not prognostic at any tumor stage. A comparison with the CAF score (based on fibroblasts only and gene expression signatures) and the UICC TNM stage indicated that the 'deep stroma score' may further improve survival predictions of currently applied prognostic scores. The 'deep stroma score' was also a prognosticator for disease-specific survival (HR = 2.29), for relapse-free survival (HR = 1.92) and overall survival (HR = 1.63) in the external test cohort (TCGA cohort).

Zhao *et al.* [37] presented a fully automated approach for CRC tissue classification and demonstrated the prognostic value of TSR for overall survival of CRC patients. Training of the CNN model was conducted with images from various origins (TCGA cohort and Chinese cohort). Tissue analysis was performed according to the nine-tissue classification approach proposed by Kather *et al.* [36]. Patch-level segmentation on WSIs was performed with classification accuracies in all tissue classes ranging from 0.96 to 0.97, also when applied on two independent patient cohorts. The TSR was calculated from CNN-based tissue classification and resulted in a prognostic value between the stroma-low versus the stroma-high group in an independent test set (HR 2.08). In addition, a comparison between the CNN and pathologists' estimation of TSR revealed high agreement (intra-class correlation coefficient: 0.937). Finally, a combination of the predicted TSR with the independent risk factors stage and age in a prediction model provided a better discrimination performance than the reference model based on stage and age alone. The fully automated approach differed from the DL models developed by Kather *et al.* [36] or Geessink *et al.* [35], which required manual annotations.

Another DL approach for determining CRC outcome was published by Skrede *et al.* [38]. The study aimed to develop an automatic prognostic biomarker of patient outcome after primary colorectal cancer resection. The model was based on diverse data from four different cohorts. A first CNN was employed to segment tumor tissue, and two additional CNN ensembles classified patients into prognostic categories. Patients were categorised into either a good or poor prognosis group. In case the two CNN ensembles predict different outcomes, the patients were assigned as uncertain. The classifier was a robust predictor of cancer-specific survival in an external test cohort (poor versus good prognosis HR 3.84; uncertain versus good prognosis HR 1.89), and the prognostic power was consistent across tumor and nodal stages. The output from the two CNN ensembles resulted in a prognostic score which strongly associated with patient outcome (AUC of 0.71). The classifier provided similar HRs in an external test cohort from multiple medical centers indicating the generalization of this approach.

Jiang *et al.* [39] focused on developing a digital prognosticator for stage III CRC in order to enable better personalised therapy for these patients. A combination of the CNN network with machine classifier was trained for tissue classification and survival prediction. A Chinese cohort was used for an internal test set, and external testing was performed on the TCGA cohort (CRC stage III), with a relatively small sample size. The classifier predicted disease-free survival in stage III CRC in the independent TCGA cohort (high- versus low-risk recurrence HR 10.27). The prediction of poor versus good prognosis showed an HR value of 10.69 in the Chinese cohort and 5.03 in the TCGA cohort. A pilot analysis which morphological parameters correlated with the predictive recurrent risk did not reveal a clear candidate.

Sirinukunwattana *et al.* [40] proposed a DL algorithm as a new digital biomarker for estimating the risk of distant metastasis by analyzing the tumor microenvironment. Distant metastases are the major cause of death in CRC [47,48]. Two CNNs were trained: one for cell detection and one for cell classification [40]. Results from cell detection and classification were used to build up a cell network. A distribution of cell–cell connections was calculated resulting in a tissue phenotype for each image patch. Ultimately, the ratio of the area of these tissue phenotypes to the total tissue area yielded a tissue phenotype signature for each tumor sample. The analysis was performed on a group of patients with stage II CRC based on UICC [49] staging. A logistic regression analysis revealed that the markers 'connection frequency of smooth muscle ratio', the 'connection frequency of inflammation ratio', as well as the 'appearance based on inflammatory tissue' have the

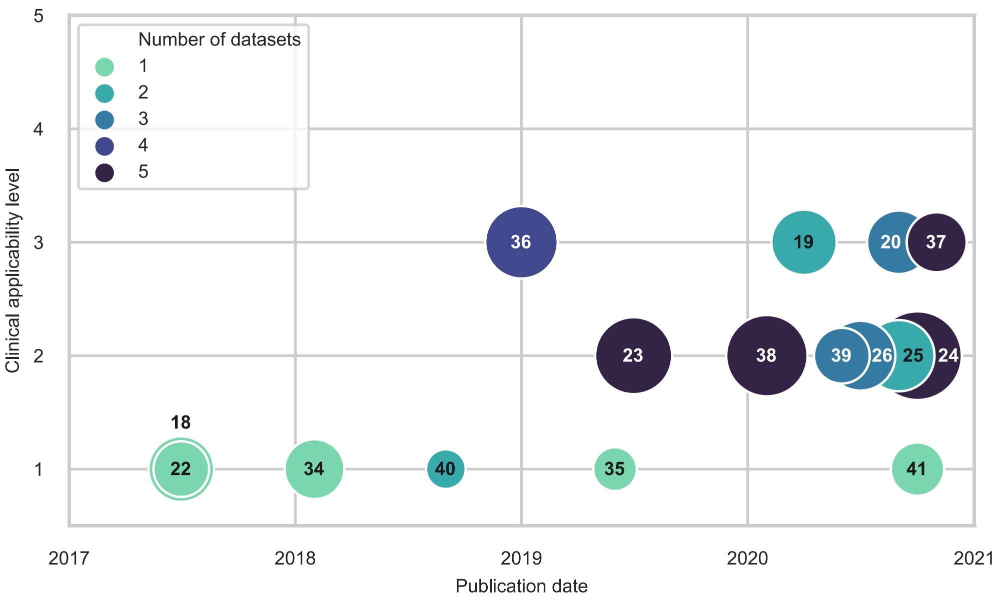


Fig. 4. Analysis of reported studies in this review: Recent publications from 2017 to 2020 show a trend towards larger samples sizes, more diverse datasets and clinical applicability. The level of clinical applicability according to Table 1 is plotted against the publication date, while the number inside the bubble indicates the reference number, the bubble size depends on the sample size, and the bubble color represents the number of applied datasets in the study. As reference numbers 34 and 35 comply with level 1 and 3 of clinical applicability, but lack the important step of testing the proposed CNN model on external data (Level 2), these the studies are depicted on level 1. Reference number 26 was sorted according to its online publication date. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

potential to predict the development of distant metastasis independently. A distant metastasis-free survival analysis showed that the features ‘connection frequency of smooth muscle ratio’ (HR 2.11; AUC 0.59) and ‘appearance based on inflammatory tissue’ (HR = 0.39; AUC 0.64) were potential prognosticators in a 5-year period analysis. Regarding the AUC values for distant metastasis prediction, it is obvious that the classifier performance is not satisfying yet. Cell types identified with HE stainings are limited and thus might have restrictive effects on the presented analysis. The authors argued that specific immunohistochemical stainings could augment the informative value of histological slides and thereby could improve the prediction of distant metastasis [40]. Furthermore, the interpretability of the classifier was reduced by the limited number of 27 metastatic cases in this study.

There was only one study dealing with the prediction of cancer-specific death in GC. Meier *et al.* [41] analysed HE and immunohistochemically stained tissue to identify survival-related features and predict the risk of cancer-specific death. As indicators for immunogenicity and aggressive growth, they looked at immune cell markers (CD8, CD20, CD68) and a proliferation marker (Ki-67). Risk heat maps were plotted to visualise the structures that the CNN associated with specific risks enabling an evaluation by pathologists and uncovering biological patterns underlying the risk prediction. B-cell (CD20+)–dominated clusters and subregions of Ki67+ cells were related to a low risk of cancer-specific death. For cancer-specific death, combined risk scores for Ki67, CD20 and/or CD68 revealed prognostic value (Ki67 & CD20: HR 1.36; CD20 & CD68: HR 1.34; Ki67 & CD68: HR 1.47; according to the main text and table). The same survival analysis was performed on the corresponding HE images but did not result in any significant values. Furthermore, a 5-year survival classification analysis and the average densities of marker cells did not yield any significant results. As future approaches, the authors plan to expand the study with the use of WSIs from resections, additional makers such as CD4, FOXP3 and CD163, a larger cohort and an external validation set.

3.4. Comparison of reported studies

We analysed all 16 studies with respect to their clinical applicability, sample size of data used and number of different datasets (data summarised in Table 2 and Fig. 4). The number of published studies increased continuously since 2017, considerably since 2020 which represents the progress and expectations achieved by medical computer vision. A trend towards larger datasets and especially generalization of the CNN-models with external datasets can be observed. Development towards clinical trials and applicability remains to be investigated and should be the focus of future research.

4. Discussion

In this review, we analysed 16 studies focusing on digital biomarkers for characterization and prognostication of gastrointestinal cancer using CNN-based classifiers. Data types and amounts and setups of the analysed studies are very heterogeneous, making a direct and quantitative comparison impossible. To assess the status with respect to clinical applicability, we designed a system adapted from Echle *et al.* [5], which defines different levels that a DL-based algorithm has to complete to succeed in the aspired clinical implementation (Tables 1 and 2).

Six studies performed the CNN-based image analysis with only one dataset (Level 1; Tables 1 and 2). These were mainly early studies published in 2017 or 2018 [18,22,34]. Meanwhile, improvements in data sharing and public dataset distribution have facilitated the acquisition of big datasets for research. Consequently, later studies could refer to additional external datasets (Fig. 4). Nevertheless, the problem of proprietary datasets still exists, impairing reproducibility of the presented methods and slowing down progress in this field.

Ten studies employed big datasets and conducted external tests of the classifier (Level 2), demonstrating that the classifiers were able to interpret images from another source correctly, despite possible differences in color, slide preparation, tissue properties or WSI scanner modalities. This suggests that the classification is based on true biological differences. As any system to be applied in a clinical routine, this absolutely requires robustness and generalization ability towards data from different institutions; recent developments reveal a trend to train the classifier on big and diverse datasets to achieve this generalization ability early on [24,36–38].

Six studies compared the performance between artificial intelligence (AI) systems and pathologists (Level 3). The performance of the classifier was on par with pathologists or even better in all cases, although publication bias may also contribute to this result. Because of these results, it is tempting to overestimate the power of AI-based systems in a real-life situation. However, as CNN systems are susceptible to artefacts, AI-based image classification systems should rather be employed as powerful assistance systems for pathologists' decisions, thus ensuring a plausibility check by human experts [50,51]. Other important aspects to consider when comparing the performance of these tools are explainability and transparency. To improve the trustworthiness of computational techniques, visualization of the regions in the underlying histological images on which the decision is based, for instance in the form of saliency maps [52] highlighting the areas that the CNN's decision was based on, would be highly desirable as to increase the acceptance of machine learning–based diagnostic tools by pathologists and patients alike. Several studies show a positive attitude of most patients towards AI-based applications for cancer diagnostics [53–55].

Ethnic differences of patient cohorts may be of remarkable relevance for the correct classification of gastrointestinal cancer samples [23,25,37]. Differences in frequencies and dietary habits are reported, but more investigation is required with respect to biological differences [1,56,57]. To encompass these apparent differences across ethnic populations, international collaborations and data exchange are required [11].

We found no study that already applied a CNN-model as a supportive system in a clinical setting (Table 2 and Fig. 4). This step includes an important change of perspective: from the retrospective analysis of past events to the prospective way of analyzing current cancer samples in a real-life scenario without knowing the results in advance. One aspect that limits clinical applicability is that in many cases AI-systems do not match the accuracy of currently used molecular methods, especially in real-life settings. This is especially true for image-based genetic testing [23–25]. Studies should clarify whether image-based genetic classification could be used as a first prescreening tool to decide which tumor samples need to be examined molecularly. A very high accuracy has been achieved with respect to MSI status [23,25]. In this case, genotype–phenotype correlations even seem to be robust enough to use the developed algorithms on images of many cancer entities to predict MSI status [33] A study published in 2021, and therefore not included in our analysis, also presented a DL-based MSI testing with good performance [58].

A general and fundamental limitation of DL algorithms is that they can only be as good as the labelling of the training set. Besides correct labelling, the training set should be as diverse as possible because features that are not included in the training set cannot be learned by the CNN. For a CNN to perform better than pathologists, it must be labelled as objectively as possible and not solely based on histopathology.

Besides the fact that CNN-based classifiers are costsaving and timesaving and provide objectivity in diagnostics, the power of AI-based systems may lie in establishing new biomarkers delivering new benefits for cancer patients. So far, most research is directed towards trying to detect known molecular biomarkers in histological slides. The future challenge is to find new digital biomarkers to generate additional prognostic or predictive value in addition to already established pathological methods/markers. Still, more barriers have to be overcome before such applications can be introduced into clinical routine. As depicted in Table 1, several levels of approval have to be passed through to define a safe and clinically useful classifier. As yet, none of the reported studies have shown a satisfying performance in a real-life setting that would enable the initiation of clinical trials. In addition, heterogeneity with respect to samples sizes, tile sizes, available cohorts (im) balanced classes, specimen used (paraffin-embedded or frozen samples) and applied CNN techniques so far precludes firm conclusions on the best techniques and strategies for a given task. Therefore, systematic evaluation of the impact of model architecture, training strategies and data sets in further studies is another essential step on the road towards a future successful clinical implementation.

In summary, enormous effort was made to develop classification models with high generalization ability in the past years. However, to approach implementation in routine clinical practice, studies in a more realistic setting leading up to clinical trials are required.

CRediT author statement

Sara Kuntz: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Roles/Writing - original draft; Writing – review & editing. **Eva Krieghoff-Henning:** Conceptualization; Funding acquisition; Methodology; Supervision; Validation; Writing – review & editing. **Jakob Nikolas Kather:** Writing – review & editing. **Tanja Jutzi:** Conceptualization; Funding acquisition; Supervision; Writing – review & editing. **Julia Höhn:** Data curation; Formal analysis; Writing – review & editing. **Lennard Kiehl:** Data curation; Formal analysis; Visualization. **Achim Hekler:** Writing – review & editing. **Elizabeth Alwers:** Writing – review & editing. **Christof von Kalle:** Writing – review & editing. **Stefan Fröhling:** Writing – review & editing. **Jochen Sven Utikal:** Writing – review & editing. **Hermann Brenner:** Writing – review & editing. **Michael Hoffmeister:** Writing – review & editing. **Titus Josef Brinker:** Conceptualization; Funding acquisition; Resources; Writing – review & editing.

Ethics approval and consent to participate

As this study is purely based on the analysis of published data, ethics approval was waived by the ethics committee of the University of Heidelberg.

Consent for publication

Not applicable.

Data availability

Not applicable.

Funding

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Tumorverhalten-Prädiktions-Initiative: Smarte Daten für die patientenzentrierte Präzisionsonkologie bei Melanom, Brust- & Prostatakrebs; funding code: 2519DAT712; grant holder: Titus J. Brinker, German Cancer Research Center, Heidelberg, Germany).

Conflict of interest statement

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

TJB would like to disclose that he is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69,120 Heidelberg, Germany, https://smarthealth.de) which developed the teledermatology services AppDoc (https://online-hautarzt.net) and Intimarzt (https://Intimarzt.de), outside the scope of the submitted work.

JNK reports a consulting role at Owkin, France.

All remaining authors have declared no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2021.07.012.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.

[2] Duffy MJ. Tumor markers in clinical practice: a review focusing on common solid cancers. Med Princ Pract 2013;22:4–11.

[3] Harbeck N, Gnant M. Breast cancer. Lancet 2017;389:1134–50.

[4] Villalobos P, Wistuba II. Lung cancer biomarkers. Hematol Oncol Clin North Am 2017;31:13–29.

[5] Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer 2020;124:686–96.

[6] Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun 2020;40:154–66.

[7] Fischer EG. Nuclear morphology and the biology of cancer cells. Acta Cytol 2020;64:511–9.

[8] Shia J, Schultz N, Kuk D, Vakiani E, Middha S, Segal NH, et al. Morphological characterization of colorectal cancers in the Cancer Genome Atlas reveals distinct morphology-molecular associations: clinical and biological implications. Mod Pathol 2017;30:599–609.

[9] Fontana E, Eason K, Cervantes A, Salazar R, Sadanandam A. Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. Ann Oncol 2019;30:520–7.

[10] Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019;16:703–15.

[11] Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. Nat Rev Gastroenterol Hepatol 2020;17:591–2.

[12] Awan R, Sirinukunwattana K, Epstein D, Jefferyes S, Qidwai U, Aftab Z, et al. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. Sci Rep 2017;7: 16852.

[13] Shapcott M, Hewitt KJ, Rajpoot N. Deep learning with sampling in colon cancer histology. Front Bioeng Biotechnol 2019;7:52.

[14] Liu Y, Li X, Zheng A, Zhu X, Liu S, Hu M, et al. Predict ki-67 positive cells in H&E-Stained images using deep learning independently from IHC-stained images. Front Mol Biosci 2020;7: 183.

[15] Lichtblau D, Stoean C. Cancer diagnosis through a tandem of classifiers for digitized histopathological slides. PloS One 2019;14: e0209274.

[16] Mori H, Miwa H. A histopathologic feature of the behavior of gastric signet-ring cell carcinoma; an image analysis study with deep learning. Pathol Int 2019;69:437–9.

[17] Xu Y, Jia Z, Wang L-B, Ai Y, Zhang F, Lai M, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinf 2017;18:281.

[18] Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inform 2017;8:30.

[19] Wei JW, Suriawinata AA, Vaickus LJ, Ren B, Liu X, Lisovsky M, et al. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw Open 2020;3:e203398.

[20] Song Z, Yu C, Zou S, Wang W, Huang Y, Ding X, et al. Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists. BMJ Open 2020;10:e036423.

[21] Lieberman DA, Rex DK, Winawer SJ, Giardiello FM, Johnson DA, Levin TR. Guidelines for colonoscopy surveillance after screening and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2012;143:844–57.

[22] Popovici V, Budinská E, Dušek L, Kozubek M, Bosman F. Image-based surrogate biomarkers for molecular subtypes of colorectal cancer. Bioinformatics 2017;33:2002–9.

[23] Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25:1054–6.

[24] Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. Gastroenterology 2020;159:1406–16.

[25] Cao R, Yang F, Ma S-C, Liu L, Zhao Y, Li Y, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. Theranostics 2020;10:11080–91.

[26] Sirinukunwattana K, Domingo E, Richman SD, Redmond KL, Blake A, Verrill C, et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. Gut 2021;70:544–54.

[27] Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol 2013;231: 63–76.

[28] Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350–6.

[29] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. In: Csurka G, editor. Domain adaptation in computer vision applications. Cham: Springer; 2017. p. 189–209.

[30] The Cancer Genome Atlas Research Network, Fearon ER, Bass AJ, Sjoblom T, Wood LD, Umar A, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.

[31] André T, Shiu KK, Kim TW, Jensen BV, Jensen LH, Punt C, et al. Pembrolizumab in microsatellite-instability-high advanced colorectal cancer. N Engl J Med 2020;383:2207–18.

[32] De Smedt L, Lemahieu J, Palmans S, Govaere O, Tousseyn T, Van Cutsem E, et al. Microsatellite instable vs stable colon carcinomas: analysis of tumour heterogeneity, inflammation and angiogenesis. Br J Cancer 2015;113:500.

[33] Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 2020;1:789–99.

[34] Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 2018;8:3395.

[35] Geessink OGF, Baidoshvili A, Klaase JM, Bejnordi BE, Litjens GJS, Van Pelt GW, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. Cell Oncol 2019;42:331–41.

[36] Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med 2019;16:e1002730.

[37] Zhao K, Li Z, Yao S, Wang Y, Wu X, Xu Z, et al. Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. EBioMedicine 2020;61:103054.

[38] Skrede O-J, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 2020; 395:350–60.

[39] Jiang D, Liao J, Duan H, Wu Q, Owen G, Shu C, et al. A machine learning-based prognostic predictor for stage III colon cancer. Sci Rep 2020;10:10333.

[40] Sirinukunwattana K, Snead D, Epstein D, Aftab Z, Mujeeb I, Tsang YW, et al. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Sci Rep 2018;8: 13692.

[41] Meier A, Nekolla K, Hewitt LC, Earle S, Yoshikawa T, Oshima T, et al. Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. Hip Int 2020;6: 273–82.

[42] Danielsen HE, Hveem TS, Domingo E, Pradhan M, Kleppe A, Syvertsen RA, et al. Prognostic markers for colorectal cancer: estimating ploidy and stroma. Ann Oncol 2018; 29:616–23.

[43] Wu J, Liang C, Chen M, Su W. Association between tumor-stroma ratio and prognosis in solid tumor patients: a systematic review and meta-analysis. Oncotarget 2016;7:68954–65.

[44] Huijbers A, Tollenaar RA, v Pelt GW, Zeestraten EC, Dutton S, McConkey CC, et al. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. Ann Oncol 2013;24:179–85.

[45] Aurello P, Berardi G, Giulitti D, Palumbo A, Tierno SM, Nigri G, et al. Tumor-Stroma Ratio is an independent predictor for overall survival and disease free survival in gastric cancer patients. Surgeon 2017;15:329–35.

[46] Wang K, Ma W, Wang J, Yu L, Zhang X, Wang Z, et al. Tumor-stroma ratio is an independent predictor for survival in esophageal squamous cell carcinoma. J Thorac Oncol 2012;7:1457–61.

[47] Kim HJ, Choi G-S. Clinical implications of lymph node metastasis in colorectal cancer: current status and future perspectives. Ann Coloproctol 2019;35:109–17.

[48] Nanduri LK, Hissa B, Weitz J, Schölch S, Bork U. The prognostic role of circulating tumor cells in colorectal cancer. Expert Rev Anticancer Ther 2019:1077–88.

[49] Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 2010;17:1471–4.

[50] Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020;31:137–43.

[51] Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019;118:91–6.

[52] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. 2014. arXiv:1312.6034.

[53] Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. JAMA Dermatol 2020;156:501–12.

[54] Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, Utikal JS, Hauschild A, Schadendorf D, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med 2020;7:233.

[55] Jonmarker O, Strand F, Brandberg Y, Lindholm P. The future of breast cancer screening: what do participants in a breast cancer screening program think about automation using artificial intelligence? Acta radiologica open 2019;8. 2058460119880315.

[56] Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet 2019;394:1467–80.

[57] Gonçalves WGE, Dos Santos MH de P, Lobato FMF, Ribeiro-Dos-Santos Â, de Araújo GS. Deep learning in gastric tissue diseases: a systematic review. BMJ Open Gastroenterol 2020;7:e000371.

[58] Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol 2021;22:132–41.

[59] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021;18:e1003583.

[60] Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019;111:30–7.

[61] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 2019;111:148–54.

[62] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019;113:47–54.

[63] Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019;118:91–6.

[64] Hekler A, Utikal SJ, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019;115:79–83.

[65] Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 2019;119:11–7.

[66] Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 2019.

[67] Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 2019;120:114–21.

[68] Höhn J, Krieghoff-Henning E, Jutzi TB, von Kalle C, Utikal JS, Meier F, et al. Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification. Eur J Cancer 2021;149:94–101.

[69] Brinker TJ, Kiehl L, Schmitt M, Jutzi TB, Krieghoff-Henning E, Krahl D, et al. Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours. Eur J Cancer 2021;154:227–34.