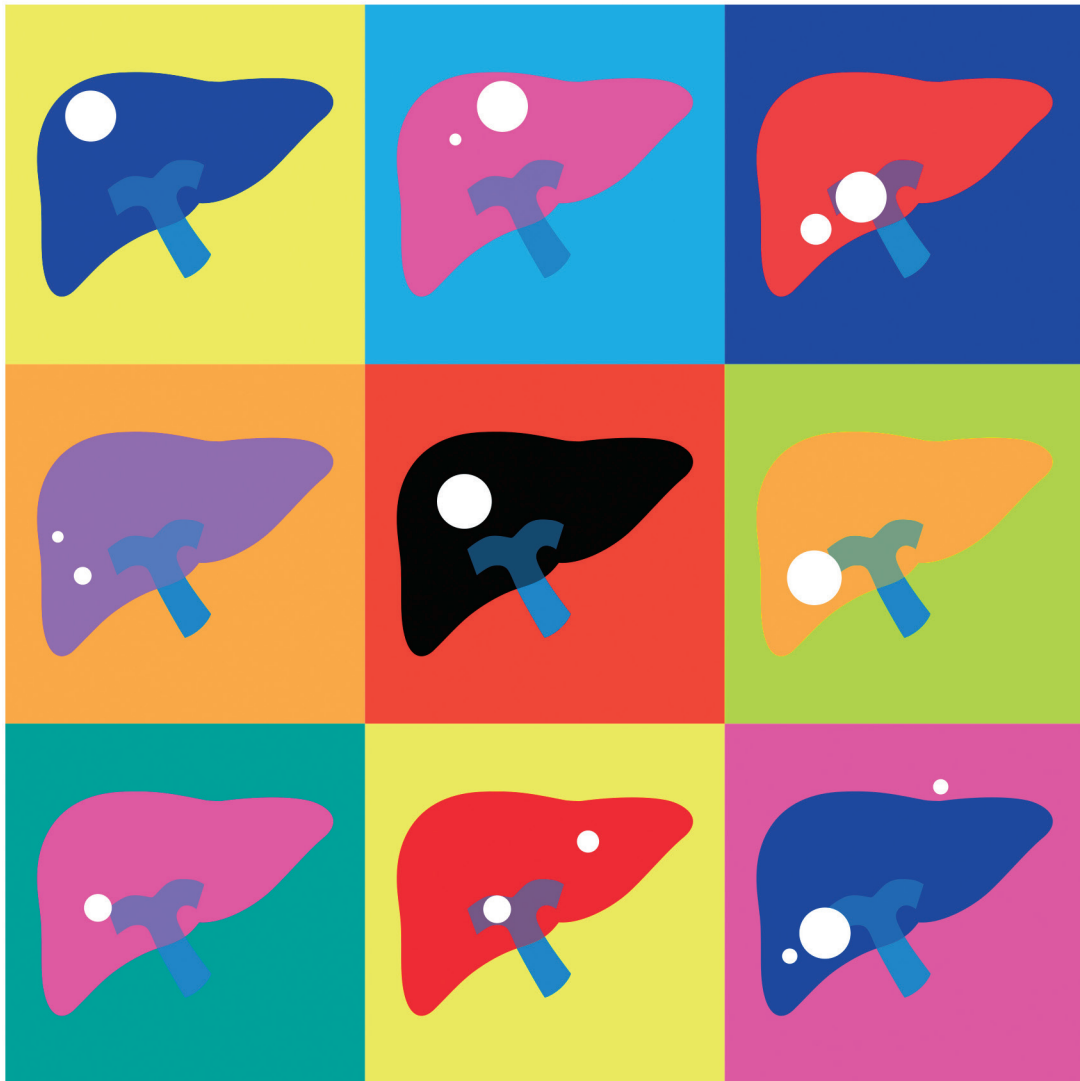# CLINICAL and MOLECULAR HEPATOLOGY

The forum for latest knowledge of hepatobiliary diseases

2022 KLCA-NCC HCC practice guideline

Insulin resistance index and NAFLD

Auranofin inhibits NAFLD

Depression and anxiety in HCC

KASL

Review

Deep learning-based prediction of molecular cancer biomarkers from tissue slides: A new tool for precision oncology

Sung Hak Lee[1] and Hyun-Jong Jang[2]

[1]Department of Hospital Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul; [2]Catholic Big Data Integration Center, Department of Physiology, College of Medicine, The Catholic University of Korea, Seoul, Korea

Molecular tests are necessary to stratify cancer patients for targeted therapy. However, high cost and technical barriers limit the application of these tests, hindering optimal treatment. Recently, deep learning (DL) has been applied to predict molecular test results from digitized images of tissue slides. Furthermore, treatment response and prognosis can be predicted from tissue slides using DL. In this review, we summarized DL-based studies regarding the prediction of genetic mutation, microsatellite instability, tumor mutational burden, molecular subtypes, gene expression, treatment response, and prognosis directly from Hematoxylin and Eosin-stained tissue slides. Although performance needs to be improved, these studies clearly demonstrated the feasibility of DL-based prediction of key molecular features in cancer tissues. With the accumulation of data and technical advances, the performance of the DL system could be improved in the near future. Therefore, we expect that DL could provide cost- and time-effective alternative tools for patient stratification in the era of precision oncology. **(Clin Mol Hepatol 2022;28:754-772)**

Keywords: Deep learning; Pathology; Molecular tests; Precision medicine; Precision oncology

INTRODUCTION

The introduction of targeted therapy for cancer treatment increased demand for identification of molecular targets, such as driver mutations in cancer cells.[1] However, many molecular tests, including next-generation sequencing, are not available to all cancer patients because of high cost and technical barriers. Recently, deep learning (DL) has been applied to predict molecular biomarkers directly from Hematoxylin and Eosin (H&E)-stained cancer tissue slides.[2] After introduction of high-speed digital slide scanners and the recent US

Food and Drug Administration approval of digitized whole slide images (WSIs) as primary specimens for diagnosis, digitization of tissue slides has been widely adopted in clinical practice.[3] Accumulation of digitized tissue images enables DL model training for prediction of many molecular pathology test results, such as mutational status of driver genes and microsatellite instability (MSI). Furthermore, the response to specific anticancer agents and the prognosis of cancer patients can be predicted by DL. Since H&E-stained tissue slides are prepared for almost all cancer patients, the DL-based method can be a cost- and time-effective alternative tool for

Corresponding author : Hyun-Jong Jang
Department of Physiology, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea
Tel: +82-2-2258-7274, Fax: +82-2-532-9575, E-mail: hjjang@catholic.ac.kr
https://orcid.org/0000-0003-4535-1560

clinical decision making.[4] In this review, we explain the current state of DL applications for the prediction of genetic mutation, MSI, tumor mutational burden (TMB), molecular subtypes, gene expression, treatment response, and prognosis directly from H&E-stained tissue slides.

Because this review is written for cancer researchers and clinicians who are not familiar with DL, we would like to start with an explanation of the basic procedure to process the WSIs of tissue slides with DL to classify tissues into different types. As an example, Figure 1 shows how normal and tumor tissue classifiers can be trained. The left image in panel A shows a WSI with regions of normal and tumor tissues labeled by pathologists. The labeled regions in the WSI are split into small image patches and collected for each label, as demonstrated in the middle part of panel A. The split is inevitable because a WSI is usually a very large image with a size of 100,000×100,000 pixels when scanned at 40× magnification. The current state of DL systems does not allow such large images to be processed at once. Supplementary Figure 1 presents examples of 250×250 pixels tissue image patches at different magnifications. Oftentimes hundreds to thousands of WSIs must be collected for accurate DL models, and the total number of collected image patches sometimes exceeds one million for each class. DL is a process to train a deep neural network (DNN) to yield correct classification results. When the task is image classification, a specific DNN called a convolutional neural network (CNN) is used. Therefore, image patches are presented as inputs to the CNN, and the outputs of the CNN are processed to yield results. Then, the classification results are compared with the true labels of the images, and the CNN is iteratively modified to yield correct results. It is common for the CNN initially to fail to cor-
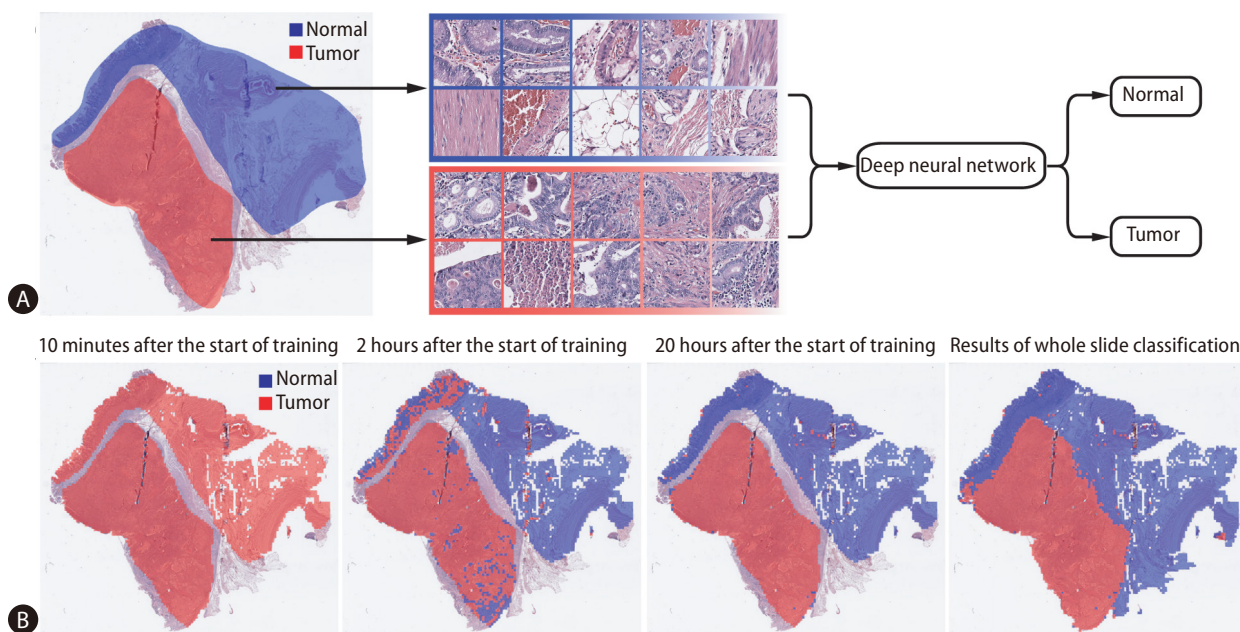

Figure 1. Representative example of the training procedure for a deep learning-based classifier. (A) Normal and tumor tissue image patches are collected for the training of a deep neural network based on the labeling by pathologists. (B) Evolution of training performance during the training procedure. First three images: classification results for the image patches are overlaid on the labeled regions at different time points of the training. Last image: after 20 hours of training, the entire tissue was classified to reveal the overall distribution of normal and tumor regions.

Abbreviations:

AUROC, area under the receiver operating characteristics curve; CNN, convolutional neural network; CNV-H, copy-number variation high; CNV-L, copy-number variation low; CRC, colorectal cancer; DL, deep learning; DNN, deep neural network; DSS, disease-specific survival; ER, estrogen receptor; FFPE, formalin-fixed paraffin-embedded; FGFR, fibroblast growth factor receptor; GAN, Generative Adversarial Network; GC, gastric cancer; GI, gastrointestinal; H&E, Hematoxylin and Eosin; HCC, hepatocellular carcinoma; HR, hazard ratio; ICI, immune checkpoint inhibitor; INF-γ, interferon-gamma; MMR, mismatch repair; MSI, microsatellite instability; MSI-H, high level of MSI; MSK-IMPACT, model performed well with the external validation cohort; OS, overall survival; PAM50, Prediction Analysis of Microarray 50; PD-L1, programmed death-ligand 1; PR, progesterone receptor; TCGA, The Cancer Genome Atlas; TIL, tumor-infiltrating lymphocyte; TMB, tumor mutational burden; WSI, whole slide image

rectly classify images. As shown in the first image of Figure 1B, the CNN classifies all image patches as tumors at the beginning of training, which is incorrect. As training progresses, the network yields a larger proportion of correct results because the mathematical weights used to calculate prediction are adjusted based on incorrect results. The last image of panel B is the classification results for an entire WSI after 20 hours of training. However, different from tumor classification, many molecular biomarkers inside the WSI are not identified and labeled by pathologists because it is unclear which regions show the biomarker-specific features. Therefore, WSIs are usually labeled as entirely negative or positive for specific biomarkers, and all image patches from a WSI have the same label. An example of this scenario regarding prediction of MSI status is presented in Supplementary Figure 2.

Published papers on the prediction of genetic mutation, MSI, TMB, molecular subtype, gene expression, treatment response, and prognosis are summarized in the following sections. Training and external validation cohorts are important for evaluating the results of such studies. Most studies adopted molecular test results and tissue slides provided by The Cancer Genome Atlas (TCGA) for either training or validation. The type of CNN and the size of image patches are also important in the DL-based evaluation of tissue images. Finally, performance measures should be compared between studies; most studies defined the DL-based model performance as the area under the receiver operating characteristics curve (AUROC). These key features of the reviewed papers are summarized in tables.

GENETIC MUTATION

An increasing number of drugs targeting specific mutations in cancer cells has been developed. For example, gefitinib and erlotinib are effective against non-small cell lung cancers with EGFR mutation.[5] Vemurafenib is effective against BRAF-mutated metastatic colorectal cancer (CRC),[6] and Sotorasib selectively and irreversibly targets KRAS[G12C] in solid cancers.[7] Mutational status should be assessed by various sequencing or staining methods to select patients for these targeted drugs. Additional cost, time, and tissue samples are required for these tests. Therefore, not every patient can be properly tested for optimal treatment options. Many studies have tried to predict the mutational status of various

genes from H&E-stained diagnostic tissue slides of multiple cancer types to select patients for these targeted drugs. Table 1 summarizes the results of the studies.

One of the first studies demonstrating the feasibility of DL-based mutation prediction for H&E-stained tissue slides was published on bioRxiv in 2016. Schaumberg et al.[8] tried to predict the mutational status of the SPOP gene from prostate cancer tissue slides provided by TCGA. The model performed well with the external validation cohort (MSK-IMPACT), demonstrating the feasibility of DL-based mutation prediction.

In 2018, seminal work for DL-based prediction of mutational status of multiple genes directly from H&E-stained histopathologic images of lung cancer was published by Coudray et al.[9] The authors first tested whether DL could discriminate between adenocarcinoma and squamous cell carcinoma of the lung. Then, they focused on adenocarcinoma for determination of the 10 most frequently mutated genes. DL models could discriminate the mutational status of STK11, KRAS, SETBP1, EGFR, FAT1, and TP53 with the AUROCs of 0.845, 0.814, 0.785, 0.754, 0.739, and 0.674, respectively.

Because the mutational status of BRAF and NRAS is important for therapeutic decision-making regarding those with stage III/IV melanoma, Kim et al.[10] built prediction models for BRAF and NRAS mutations. Interestingly, predictive performance was affected by the thickness of the tumor and ulceration status. In skin cancer, these parameters reflect complex underlying molecular mechanisms that control the phenotype of mutated genes. Therefore, independent models considering different thicknesses and ulceration statuses can improve the overall performance of mutation prediction in melanoma.

Tsou and Wu tried to discriminate two mutually exclusive drivers of papillary thyroid carcinoma: BRAFV600E and mutated RAS.[11] The accuracies of the classifiers were 63.6% and 90% for BRAF and RAS mutations, respectively, with an overall AUROC of 0.951 for the test set.

Differentiation status and mutations were predicted by Chen et al.[12] from liver cancer tissue slides. The authors first tried to classify the liver cancer tissues into well, moderately, and poorly differentiated tumors and compared the results with pathologists. Then, they predicted the mutational status of the 10 most commonly mutated genes. Four of them, CTNNB1, FMN2, TP53, and ZFX4, were predicted with the AUROCs ranging from 0.71 to 0.89.

Because the mutational statuses of IDH genes (IDH1 and

Table 1. Summaries for studies on the prediction of genetic mutation

Study	Target (cancer type)	Training cohort	Validation cohort	Neural network	Patch size (magnification)	Performance measure
Schaumberg et al.[8]	SPOP (prostate cancer)	TCGA	MSK-IMPACT	ResNet-50	224×224 pixels (N/A)	AUROC: 0.74
Coudray et al.[9]	STK11, KRAS, SETBP1, EGFR, FAT1, and TP53 (lung cancer)	TCGA	NYU Langone Medical Center	Inception v3	512×512 pixels (20×)	AUROC: 0.674–0.845
Kim et al.[10]	BRAF and NRAS (melanoma)	NYU Langone Health	TCGA	Inception v3	229×229 pixels (20×)	AUROC: 0.83 (BRAF) and 0.92 (NRAS)
Tsou and Wu[11]	BRAF and RAS (thyroid cancer)	TCGA	None	Inception v3	512×512 pixels (5×)	AUROC: 0.951
Chen et al.[12]	CTNNB1, FMN2, TP53, and ZFX4 (liver cancer)	TCGA	SRRSH	Inception v3	256×256 pixels (20×)	AUROC: 0.71–0.89
Liu et al.[13]	IDH (glioma)	TCGA+YUH	None	ResNet-50	256×256 pixels (N/A)	AUROC: 0.927
Fu et al.[14]	151 gene:cancer pairs (various cancers)	TCGA	METABRIC	Inception v4	512×512 pixels (20×)	AUROC: 0.098–0.972
Kather et al.[15]	Mutations with a prevalence above 2% (various cancers)	TCGA	None	ShuffleNet	512×512 pixels (20×)	AUROC: 0.55–0.8
Noorbakhsh et al.[17]	TP53 (various cancers)	TCGA	None	Inception v3	512×512 pixels (20×)	AUROC: 0.65–0.80
Jang et al.[16]	APC, KRAS, PIK3CA, SMAD4, and TP53 (colorectal cancer)	TCGA	SSMH	Inception v3	360×360 pixels (20×)	AUROC: 0.645–0.809
Yang et al.[19]	DNMT3A, EGFR, PBRM1, STK11, and TP53 (lung cancer)	TCGA	None	ResNet	512×512 pixels (N/A)	AUROC: 0.71–0.87
Loeffler et al.[20]	FGFR3 (bladder cancer)	TCGA	Aachen University	ShuffleNet	512×512 pixels (20×)	AUROC: 0.701
Jang et al.[21]	CDH1, ERBB2, KRAS, PIK3CA, and TP53 (gastric cancer)	TCGA	SSMH	Inception v3	360×360 pixels (20×)	AUROC: 0.661–0.862

Because some studies resized or cropped the input image patches during preprocessing, the final patch size and magnification are presented in the table. In some cases, the magnification was not clearly specified and thus noted as (N/A). The area under the receiver operating characteristics curves (AUROCs) are for the held-out test sets from the training datasets. If it is not provided, AUROCs for external validation cohort are presented.

TCGA, The Cancer Genome Atlas; MSK-IMPACT, model performed well with the external validation cohort; N/A, not applicable; AUROC, area under the receiver operating characteristics curve; NYU, New York University; SRRSH, Sir Run-Run Shaw Hospital; YUH, Yeditepe University Hospital; METABRIC, Molecular Taxonomy of Breast Cancer International Consortium; SSMH, Seoul St. Mary's Hospital.

IDH2) are important prognostic and therapeutic biomarkers for glioma, Liu et al.[13] tested the feasibility of IDH mutation prediction from histopathologic glioma images using DL. They tried to augment existing images using a Generative Adversarial Network (GAN). GANs can create virtual images of glioma with and without IDH mutation by iteratively improving generated images until they are indiscernible from the training set. Using a GAN, the AUROC was slightly improved from 0.920 to 0.927. Although there was some improvement, these results indicated that generative modeling could provide only a slight improvement of performance because GAN-created images are only extended virtual copies of existing images and cannot provide new information.

Fu et al.[14] tried to build a general-purpose classifier for multiple mutations in various cancers. They first constructed an Inception v4-based classifier for discrimination among 42 tissue types, including 28 cancers and 14 normal tissues. Then, the classifier was reused to predict the mutational status of multiple genes. In detail, regression methods were applied to the 1,536 features from the final layer of the classifier for cancer type-specific mutation prediction. The AUROCs ranged from 0.098 for the FBXW7 gene in head and neck cancer to 0.972 for the STAG2 gene in bladder cancer. The performance seemed suboptimal because an end-to-end model was not used for mutation prediction; however, the possibility of a pan-cancer classifier was well demonstrated.

Kather et al.[15] published another pan-cancer approach for detection of clinically actionable genetic alterations. They aimed to predict mutations affecting at least four patients, a prevalence above 2%. The results were suboptimal compared to those of other studies that investigated specific cancer types. For example, the AUROCs for APC, KRAS, SMAD4, and TP53 mutations in CRC were 0.63, 0.6, 0.61, and 0.64, respectively, which were much lower than those of our previous study, 0.771, 0.778, 0.693, and 0.809.[16]

Noorbakhsh et al.[17] trained classifiers to predict TP53 gene mutational status from breast, lung, stomach, colon, and bladder cancers from the TCGA datasets and obtained the AUROCs of 0.7447, 0.7969, 0.6532, 0.7825, and 0.7094, respectively. Importantly, they found that the classifiers could identify mutational status more accurately when tumor tissues were more homogenous. As we discuss later, tumor heterogeneity affects the performance of DL-based tissue classifiers.

In our previous study, mutations in APC, KRAS, PIK3CA, SMAD4, and TP53 genes were predicted from CRC tissue slides.[16] We trained classifiers for frozen and formalin-fixed paraffin-embedded (FFPE) tissues separately because they are morphologically different. We previously showed that DL-based classifiers are incompatible between frozen and FFPE tissues.[18] The AUROCs ranged from 0.693 to 0.809 and from 0.645 to 0.783 for frozen and FFPE tissues, respectively. When we combined the TCGA data and our own data for the training of the models, the performance was improved. These results indicate that collection of larger dataset is essential to improve the performance of DL-based classifiers.

Yang et al.[19] investigated the feasibility of mutational status prediction for DNMT3A, EGFR, PBRM1, STK11, and TP53 genes in lung cancer, which are thought to be candidate markers for immunotherapy response. The AUROCs were between 0.71 and 0.87.

Because fibroblast growth factor receptor (FGFR) inhibitor was approved as a targeted therapy in bladder cancer, Loeffler et al.[20] tried to predict the mutational status of the FGFR gene from tissues of bladder cancer. Furthermore, they compared the performance of the DL model with that of pathologists because there are some visually recognizable histologic features in FGFR-mutated tumors. The AUROC of the DL model was 0.701, and it outperformed pathologists. Interestingly, one tissue had available information on intratumor heterogeneity, which was investigated with multi-region sequencing. DL correctly delineated the heterogeneous regions in the tissue.

We also tried to predict mutations in gastric cancer (GC) tissue slides.[21] Similar to the study for the CRC, we separately trained the classifiers for frozen and FFPE tissues and predicted the mutational status of CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes. The AUROCs ranged from 0.727 to 0.862 and from 0.661 to 0.858 for frozen and FFPE tissues, respectively. The performance could also be enhanced by combining the TCGA data with our own data, highlighting the importance of large datasets for prediction of mutational status from H&E-stained tissue slides.

MSI

Accumulation of insertions or deletions in the repeated units of microsatellites by impaired DNA mismatch repair (MMR) causes MSI.[22] Deficient MMR also increases overall mutation rates and leads to expression of neoantigens, which

attract immune cells. To avoid immune surveillance, tumor cells express several immune checkpoint ligands.[23] Therefore, MSI could be a marker for patient response to immune checkpoint inhibitors (ICIs).[24] Furthermore, MSI is an important prognostic marker.[25] Many studies tested the feasibility of MSI status prediction from tissue slides using DL (Table 2).

The first comprehensive study to investigate the feasibility of DL-based MSI status prediction was published by Kather et al.[26] in 2019. They focused on gastrointestinal (GI) cancers, and GC and CRC datasets from the TCGA program were used to train the classifiers. The AUROCs were 0.81 and 0.77 for the FFPE tissue slides of GC and CRC, respectively, and 0.84 for the frozen tissue slides of CRC. When the DL models were tested on an Asian GC cohort and the TCGA endometrial cancer dataset, performance was much decreased. These results indicated that the classifiers did not extend beyond the ethnicity and cancer types presented in the training datasets. Therefore, current generalizability of DL-based classifiers for MSI status appears to be limited.

Cao et al.[27] also tested DL-based MSI status prediction in CRC. The AUROC was 0.8848 for frozen tissue slides from the TCGA. When the classifier was tested on an Asian CRC cohort, the AUROC was only 0.6497. The authors implemented the transfer learning scheme to improve the performance on the Asian-CRC cohort. When the parameters of the classifier trained on the TCGA dataset were transferred to retrain a new classifier for the Asian-CRC cohort, the AUROC was improved to 0.8504. These results confirmed that transfer learning is a plausible method to modify a classifier to improve performance on other datasets.

Echle et al.[28] formed the MSIDETECT consortium to collect a large set of CRC data and improve the performance of MSI prediction. They recruited more than 8,000 patients and improved the AUROC to 0.92, which is much better than their previous TCGA-based study with the AUROCs between 0.77 and 0.84.[26] Although they clearly demonstrated the importance of a large dataset for better performance, the importance of ethnicity on predictive performance was not tested because there was no Asian cohort involved.

Wang et al.[29] trained a prediction model for MSI status in uterine corpus endometrial carcinoma. The AUROC of their model was 0.73, not superior to that of Kather et al.[26]

Similar to the study by Liu et al.,[13] Krause et al.[30] trained a GAN to enlarge the tissue datasets for MSI research by generating virtual tissue images. The use of GAN-created images

Table 2. Summaries for studies on the prediction of microsatellite instability

Study	Target (cancer type)	Training cohort	Validation cohort	Neural network	Patch size (magnification)	Performance measure
Kather et al.[26]	Microsatellite instability (gastrointestinal cancer)	TCGA	DACHS (colorectal), KCCH (gastric)	ResNet-18	512×512 pixels (20×)	AUROC: 0.77–0.84
Cao et al.[27]	Microsatellite instability (colorectal cancer)	TCGA	Asian CRC cohort	ResNet-18	224×224 pixels (20×)	AUROC: 0.8848
Echle et al.[28]	Microsatellite instability (colorectal cancer)	TCGA+DACHS+QUASAR+NLCS	YCR-BCIP	ShuffleNet	512×512 pixels (20×)	AUROC: 0.92
Wang et al.[29]	Microsatellite instability (endometrial cancer)	TCGA	None	ResNet-18	512×512 pixels (20×)	AUROC: 0.73
Krause et al.[30]	Microsatellite instability (colorectal cancer)	TCGA+NLCS	None	ShuffleNet	512×512 pixels (20×)	AUROC: 0.742–0.777
Yamashita et al.[31]	Microsatellite instability (colorectal cancer)	SUMC	TCGA	MobileNet v2	224×224 pixels (about 10×)	AUROC: 0.931
Lee et al.[32]	Microsatellite instability (colorectal cancer)	TCGA	SSMH	Inception v3	360×360 pixels (20×)	AUROC: 0.861–0.942

TCGA, The Cancer Genome Atlas; DACHS, Darmkrebs Chancen der Verhütung durch Screening; KCCH, Kanagawa Cancer Center Hospital; AUROC, area under the receiver operating characteristics curve; CRC, colorectal cancer; QUASAR, Quick and Simple and Reliable trial; NLCS, Netherlands Cohort Study; YCR-BCIP, Yorkshire Cancer Research Bowel Cancer Improvement Programme; SUMC, Stanford University Medical Center; SSMH, Seoul St. Mary's Hospital.

improved the AUROC from 0.757 to 0.777. Compared to the study by Liu et al.,[13] this improvement was more prominent.

Yamashita et al.[31] compared the performance of a DL-based MSI status prediction model with that of five GI pathologists. The pathologists discriminated the MSI tissues based on 10 known tissue features of MSI CRC tumors, such as Crohn's-like reaction and signet ring-cell differentiation. Regarding the reader experiment (40 cases), the AUROC of the DL model was 0.865. The mean AUROC for the five pathologists was 0.605. Therefore, the DL model outperformed the pathologists.

We also tested the feasibility of DL-based MSI status prediction and obtained the AUROCs of 0.942 and 0.861 for frozen and FFPE tissues, respectively, from the TCGA CRC datasets.[32] The classification performance was not satisfactory on our own Asian dataset, with an AUROC of 0.787. This result reiterated the lack of compatibility of TCGA-based MSI prediction models on Asian cohort. Then, we trained a new classifier with both TCGA and our own dataset. The new classifier performed well on both datasets, with the AUROCs of 0.892 and 0.972 for the TCGA and our dataset, respectively. These results demonstrated the importance of large multi-national datasets for better performance. We also showed that the application of a classifier trained on the CRC tissues from original sites could not be extended to metastasized CRC tissues in the liver and lung (AUROC, 0.484). These results indicated that the morphologic features of MSI in tumor tissues were different between primary and metastasized tumors. We also showed that the DL model discriminated MSI tissue images based on previously known features, such as cribriform pattern, high tumor-infiltrating lymphocyte (TIL) density, and mucinous differentiation. Because these features were more prominent in primary tumors than metastasized tumors, the DL algorithm could discriminate MSI more clearly in primary tumors.

TMB

TMB is defined as the total number of somatic mutations in the coding area of a tumor genome.[33] Immune cells infiltrate because of high neoantigen expression in tumors with high TMB; thus, tumor cells express several immune checkpoint ligands to avoid immune surveillance. Therefore, like MSI, TMB can be a clinical biomarker for response to ICIs.[34] TMB has

been predicted in many studies (Table 3).

One of the first studies to predict TMB status was published by Xu et al.[35] in 2019. They trained a DL model to discriminate high vs. low TMB in bladder cancer and obtained an AUROC of 0.75. They also showed that the probability of survival was higher for patients with tissues predicted to have high TMB by the classifier.

Jain and Massoud[36] predicted TMB status by integrating three DL models with lung adenocarcinoma images at different magnifications (5×, 10×, and 20×). Although the AUROCs were only 0.72, 0.80, and 0.81 for 5×, 10×, and 20× models, respectively, the combined model achieved an AUROC of 0.92.

Another study by Xu et al.[35] extended their previous study and predicted TMB status from bladder and lung cancers, with the AUROCs of 0.752 and 0.742, respectively.[37] Most importantly, they analyzed intratumoral heterogeneity with DL and discovered better prognoses for tumors with high TMB and low intratumoral heterogeneity than for highly heterogenous high-TMB tumors. Only DL-based approaches can obtain information on intratumoral heterogeneity with affordable cost.

Shimada et al.[38] first tried to discriminate high TMB from CRC tissue slides based on TIL counts and achieved an AUROC of 0.910. Then, a DL-based model was trained and yielded an AUROC of 0.934. Although the improvement was not large, DL can achieve better results without the laborious process of TIL counting.

Sadhwani et al.[39] tried a histologic subtype-based approach to develop an interpretable model for TMB prediction. They trained a DL model to discriminate nine histologic features from lung adenocarcinoma tissues. Then, the information on tissue features was combined with clinical data to predict TMB status. Although this approach can help to improve the interpretability of DL models, the overall performance was inferior to that of a previous study.[36]

MOLECULAR SUBTYPES

Cancers can be subclassified into molecular subtypes based on molecular pattern.[40-43] Molecular subtypes provide important information for clinical decision-making because they show different prognoses and therapeutic responses. However, molecular subtyping is costly and technically diffi-

Table 3. Summaries for studies on the prediction of tumor mutational burden

Study	Target (cancer type)	Training cohort	Validation cohort	Neural network	Patch size (magnification)	Performance measure
Xu et al.[35]	Tumor mutational burden (bladder cancer)	TCGA	None	Xception	1,024×1,024 pixels (20×)	AUROC: 0.75
Jain and Massoud[36]	Tumor mutational burden (lung cancer)	TCGA	None	Inception v3	512×512 pixels (5×, 10×, 20×)	AUROC: 0.92
Xu et al.[37]	Tumor mutational burden (bladder and lung cancers)	TCGA	None	Xception	512×512 pixels (20×)	AUROC: 0.742–0.752
Shimada et al.[38]	Tumor mutational burden (colorectal cancer)	TCGA+Japanese CRC cohort	None	Inception v3	300×300 pixels (N/A)	AUROC: 0.934
Sadhwani et al[39]	Tumor mutational burden (lung cancer)	TCGA	None	Inception v3	512×512 pixels (10×)	AUROC: 0.71

TCGA, The Cancer Genome Atlas; AUROC, area under the receiver operating characteristics curve; CRC, colorectal cancer.

cult. Therefore, DL-based prediction of molecular subtypes can be a cost-effective tool for patient stratification. Recently, the molecular subtypes of various cancers have been predicted by DL from tissue images (Table 4).

Couture et al.[44] tried to discriminate the intrinsic molecular subtypes of breast cancer. Based on the subtypes discriminated by the Prediction Analysis of Microarray 50 (PAM50), they trained a classifier to discriminate basal-like vs. non-basal-like (including luminal A, luminal B, and HER2-enriched subtypes) tumors from tissue slides. The accuracy of the molecular subtype classifier was 77%. Because they only tried to discriminate one subtype from three others, the overall performance of the DL system on classification of the four molecular subtypes in breast cancer could not be estimated.

Based on tissue type classifiers, Kather et al.[15] tried to discriminate the intrinsic molecular subtypes for breast, colorectal, gastric, and lung cancers. Basal-like, Her2-enriched, luminal A, and luminal B subtypes of breast cancer were detectable with the AUROCs between 0.61 and 0.86. For CRC and GC, the AUROCs for the pan-GI subtypes, including GI-hypermutated indel, GI genome stable, GI-chromosomally unstable, GI-hypermutated-single-nucleotide variant predominant, and GI Epstein-Barr-virus-positive, ranged from 0.23 to 0.78. TCGA molecular subtypes LUAD1 to 6 in lung adenocarcinoma yielded an AUROC of up to 0.74. These results indicate that tissue morphology reflects the characteristics of the intrinsic molecular subtypes.

Another study regarding PAM50-based molecular subtype discrimination in breast cancer was published by Jaber et al.[45] The classification accuracy for the four classes (basal-like, Her2-enriched, luminal A, and luminal B) was 67.27%. Because there was an imbalance between the number of classes, they tried to distinguish two classes (basal-like and non-basal-like), and the accuracy was 87.21%. Therefore, it is possible for the performance of DL models to be improved if large amounts of data on Her2-enriched, luminal A, and luminal B cases are collected. They also showed that patients with heterogeneous cancer tissues with mixed basal-like and luminal A characteristics had intermediate survival compared to homogenous basal-like and luminal A groups, suggesting the importance of tumor heterogeneity for prognosis.

Hong et al.[46] first trained a classifier to discriminate the histologic types of endometrial cancer into endometrioid or serous subtypes and achieved an AUROC of 0.969. Then, they tried to discriminate the four molecular subtypes: POLE ultra-

Table 4. Summaries for studies on the prediction of molecular subtypes

Study	Target (cancer type)	Training cohort	Validation cohort	Neural network	Patch size (magnification)	Performance measure
Couture et al.[44]	Molecular subtypes (breast cancer)	CBCS3	None	VGG16	800×800 pixels (20×)	Accuracy: 77%
Kather et al.[15]	Molecular subtypes (breast, colorectal, gastric, and lung cancers)	TCGA	None	ShuffleNet	512×512 pixels (20×)	AUROC: 0.24–0.86
Jaber et al.[45]	Molecular subtypes (breast cancer)	TCGA	None	Inception v3	400×400 pixels (5×, 10×, 20×)	Accuracy: 67.27%
Hong et al.[46]	Molecular subtypes (endometrial cancer)	TCGA+TCIA	None	InceptionResNet	299×299 pixels (2.5×, 5×, 10×)	AUROC: 0.827–0.934
Sirinukunwattana et al.[47]	Molecular subtypes (colorectal cancer)	FOCUS	TCGA+GRAMPIAN	Inception v3	299×299 pixels (3×, 12×)	AUROC: 0.86–0.92
Yu et al.[48]	Molecular subtypes (lung cancer)	TCGA	ICGC	VGGNet	224×224 pixels (about 5×)	AUROC: 0.7–0.892
Woerl et al.[49]	Molecular subtypes (bladder cancer)	TCGA	CCC-EMN	ResNet-50	512×512 pixels (40×)	AUROC: 0.76–0.89

CBCS3, Carolina Breast Cancer Study Phase 3; TCGA, The Cancer Genome Atlas; AUROC, area under the receiver operating characteristics curve; TCIA, The Cancer Imaging Archive; ICGC, International Cancer Genome Consortium; CCC-EMN, Comprehensive Cancer Center Erlangen Metropol Region Nuremberg.

mutated, high level of MSI (MSI-H) hypermutated, copy-number variation low (CNV-L), and copy-number variation high (CNV-H). The AUROCs for CNV-H and MSI-H were 0.934 and 0.827, respectively.

Four consensus molecular subtypes (CMS1 to 4) of CRC were classified by Sirinukunwattana et al.[47] from tissue images. They adopted the domain adversarial training scheme, which enforces the DL model not to learn specific features confined to a specific cohort. Therefore, the network could learn more general features. The AUROCs for CMS1 to 4 were 0.86, 0.91, 0.92, and 0.89, respectively, yielding a macro-average AUROC of 0.9.

Yu et al.[48] trained classifiers to discriminate normal and tumor tissues from non-small cell lung cancer tissues and further classified the tumor tissues into adenocarcinoma and squamous cell carcinoma. Based on the classification results, the three transcriptomic subtypes of lung adenocarcinoma, terminal respiratory unit, proximal inflammatory, and proximal proliferative subtypes, and the four transcriptomic subtypes of lung squamous cell carcinoma, classical, basal, secretory, and primitive subtypes, were classified. The AUROCs for the adenocarcinoma subtypes ranged from 0.771 to 0.892, and the AUROCs for the squamous cell carcinoma subtypes were around 0.7.

Based on a custom panel of 21 genes, Woerl et al.[49] predicted four molecular subtypes of muscle invasive bladder cancer. The AUROCs for discrimination of double negative, basal, luminal, and luminal p53-like subtypes were 0.76, 0.89, 0.88, and 0.89, respectively. They found that DL-based prediction outperformed morphology-based prediction by four pathologists.

EXPRESSION OF GENES AND PROTEINS

Expression of specific genes and proteins has a significant impact on prognostication and therapeutic decision-making.[50-53] DL has been successfully applied for prediction of expression status, particularly for breast and lung cancers (Table 5).

Estrogen receptor (ER) expression has significant implications in the prognosis and treatment of breast cancer.[54] Couture et al.[44] trained a DL model to test the ER status of breast cancer tissues. The discrimination accuracy of ER-positive and ER-negative tissues was 84%.

Programmed death-ligand 1 (PD-L1) expression status is

Table 5. Summaries for studies on the prediction of expression of genes and proteins

Study	Target (cancer type)	Training cohort	Validation cohort	Neural network	Patch size (magnification)	Performance measure
Couture et al.[44]	Estrogen receptor (breast cancer)	CBCS3	None	VGG-16	800×800 pixels (20×)	Accuracy: 84%
Sha et al.[56]	PD-L1 status (lung cancer)	Own cohort (Tempus Labs Chicago)	None	ResNet-18	446×446 and 32×32 pixels (10×)	AUROC: 0.80
Rawat et al.[57]	Estrogen receptor, progesterone receptor, HER2 (breast cancer)	TCGA	ABCTB	ResNet-34	224×224 pixels (20×)	AUROC: 0.71–0.88
Naik et al.[58]	Estrogen receptor, progesterone receptor, HER2 (breast cancer)	TCGA+ABCTB	Multiple centers	ResNet-50	256×256 pixels (20×)	AUROC: 0.778–0.92
He et al.[59]	Expression of multiple genes (breast cancer)	Own cohort	TCGA	DenseNet-121	224×224 pixels (20×)	Correlation coefficient: 0.52
Schmauch et al.[60]	Expression of multiple genes (various cancers)	TCGA	None	ResNet-50	224×224 pixels (20×)	Correlation coefficient: 0.47
Levy-Jurgenson et al.[61]	Expression of multiple genes (breast and lung cancer)	TCGA	None	Inception v3	512×512 pixels (20×)	AUROC: 0.44–0.85

CBCS3, Carolina Breast Cancer Study Phase 3; PD-L1, programmed death-ligand 1; AUROC, area under the receiver operating characteristics curve; TCGA, The Cancer Genome Atlas; ABCTB, Australian Breast Cancer Tissue Bank.

one of the biomarkers for clinical response to cancer immunotherapy.[55] PD-L1 expression status is usually evaluated by visual assessment of immunostained tissue slides and suffers from inter-observer variability. Sha et al.[56] tried to predict PD-L1 status from tissue slides of non-small cell lung cancer and achieved an AUROC of 0.80. The prediction performance was much better for adenocarcinoma than for squamous cell carcinoma.

In addition to ER status, progesterone receptor (PR) and HER2 receptor statuses are also important biomarkers for breast cancer. Rawat et al.[57] tried to predict the statuses of ER, PR, and HER2, which are usually assessed by immunohistochemistry. The AUROCs for ER, PR, and HER2 statuses were 0.88, 0.78, and 0.71, respectively.

Naik et al.[58] enhanced the performance of prediction models for ER, PR, and HER2 statuses in breast cancer by applying the attention mapping method. They obtained the AUROCs of 0.92, 0.81, and 0.778 for ER, PR, and HER2 statuses, respectively.

Spatial transcriptomics technology allowed the expression of multiple genes to be measured from multiple locations within each tissue sample. He et al.[59] trained classifiers based on spatial transcriptomics data and predicted spatial variation in the expression of multiple genes. In 102 of 250 genes, the predictions were positively correlated with the experimentally measured expression. The expression of GNAS, ACTG1, FASN, DDX5, and XBP1 genes was especially well predicted. This study was unique in that the correlation of spatial expression patterns between actual experimental data and DL-based prediction could be examined because the data were obtained through spatial transcriptomics.

Schmauch et al.[60] tried to predict gene expression profiles from multiple tumor types presented by the TCGA program. The study included 8,725 patients with 28 cancer types. An average of 3,627 genes per cancer type was predicted to have altered expression with statistically significant correlation. They also tested the spatial correlation of their prediction results with tissues immunostained for CD3 and CD20. The staining levels were well correlated with the prediction results, confirming the excellent spatial prediction of CD3 and CD20 expression.

Levy-Jurgenson et al.[61] trained DL-based prediction models for the expression level of various RNAs and miRNAs from tissue slides of breast and lung cancers. Five genes for breast cancer and three genes for lung adenocarcinoma were pre-

dicted with the AUROCs higher than 0.6. Furthermore, the authors demonstrated that heterogeneity in expression level is a reliable negative predictor of survival. For many cases, highly heterogeneous expression patterns were correlated with poor prognosis.

TREATMENT RESPONSE AND PROGNOSIS

The aforementioned molecular biomarkers offer information for prediction of treatment response and prognosis. Because prediction of these biomarkers was feasible with DL, direct prediction of treatment response and prognosis could be possible. Therefore, many studies have tested DL-based prediction of treatment response and prognosis from H&E-stained tissue slides (Table 6).

Hu et al.[62] tried to predict the anti-PD-1 response in melanoma and lung cancer. Because the TCGA dataset did not offer information on immunotherapy responses, they adopted interferon-gamma (INF-γ) scores as a surrogate for the training of anti-PD-1 response prediction in a DL model. Then, they used their own dataset with anti-PD-1 therapy response data to validate the model. The AUROCs for melanoma and lung cancer were 0.778 and 0.645, respectively. Interestingly, they compared the performance of their INF-γ-based model with that of a TIL-based model because TIL has been thought as a biomarker for immunotherapy response. The AUROC of the TIL-based model was only 0.58 for melanoma. Therefore, their results indicated that INF-γ scores could be a good biomarker to predict anti-PD-1 response.

Johannet et al.[63] directly trained immunotherapy response prediction models on metastasized melanoma tissues of patients treated with anti-CTLA-4, anti-PD-1, or a combination of anti-CTLA-4 and anti-PD-1 therapy. The prediction results were better for lymph nodes than soft tissue, with the AUROCs of 0.857 and 0.583, respectively, and an overall AUROC of 0.691. Furthermore, the prediction results were improved by incorporating clinical data into the model. The AUROC of the combined model was 0.793.

Bychkov et al.[64] trained a model to discriminate low- and high-risk groups of CRC patients directly from H&E-stained tissue microarray images. The DL model yielded an AUROC of 0.69 with a hazard ratio (HR) of 2.3. A visual risk score was assessed by visual examination of tissue slides by three expert pathologists. The AUROC of the visual risk score was only

0.58, with an HR of 1.67. Therefore, the DL model for prognosis prediction outperformed visual assessment by pathologists.

Mobadersany et al.[65] trained a time-to-event model to predict patient outcomes from glioma tissues. Because it was a time-to-event model, Harrell's C index was measured to assess concordance between the model and actual survival rather than AUROC. The model achieved a C index of 0.741. When genomic information including IDH mutation status and 1p/19q co-deletion status were integrated into the model, the C index was increased to 0.781. These results indicate that tissue morphology and genetic alteration status contain complementary information for prediction of survival.

Kather et al.[66] tried to predict the overall survival (OS) of CRC patients. They trained a DL model to classify tissue images into adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and tumor tissues. Then, the weighted sum of the activation of the neural network on non-tumor tissue, including adipose tissue, debris, lymphocytes, smooth muscle, and cancer-associated stroma, was calculated as a deep stromal score. The deep stromal score was confirmed to be an independent prognostic factor for OS, with an HR of 1.99. This result demonstrates that non-tumor stromal tissues can provide prognostic information for survival.

Courtiol et al.[67] trained a model to predict the survival time of patients with malignant mesothelioma. The model achieved a C index of 0.643. They also showed that the most relevant image patches for prediction of survival were located in stromal regions rather than tumor regions. The results also indicated that non-tumor stromal tissues are important for prognostic prediction.

Skrede et al.[68] tried to predict the survival of CRC patients with cancer-specific survival as a primary endpoint. They collected huge datasets of four training cohorts and one validation cohort. Ensemble results of 10 total models, five models trained with 10× tissue images and five models trained with 40× tissue images, were used to improve the performance of prediction model for the discrimination of good or poor prognostic groups. The ensemble model could discriminate good or poor prognosis with an HR of 3.84.

Kulkarni et al.[69] proposed a DL method to predict visceral recurrence and disease-specific survival (DSS) from the H&E-stained tissues of primary melanoma. They implemented a complex model that incorporated information from DL-

Table 6. Summaries for studies on the prediction of treatment response and prognosis

Study	Target (cancer type)	Training cohort	Validation cohort	Neural network	Patch size (magnification)	Performance measure
Hu et al.[62]	Anti-PD1 response (melanoma and lung cancer)	TCGA	PUCH	Xception	256×256 pixels (20×)	AUROC: 0.645–0.778
Johannet et al.[63]	Immune checkpoint inhibitors response (melanoma)	NYU	Vanderbilt University	Inception v3	299×299 pixels (10× or 20×)	AUROC: 0.691–0.793
Bychkov et al.[64]	Prognosis (colorectal cancer)	HUCH	None	VGG-16	224×224 pixels (40×)	AUROC: 0.69
Mobadersany et al.[65]	Prognosis (glioma)	TCGA	None	VGG-19	256×256 pixels (20×)	Harrell's C index: 0.741
Kather et al.[66]	Prognosis (colorectal cancer)	TCGA	DACHS	VGG-19	224×224 pixels (20×)	HR: 1.99
Courtiol et al.[67]	Prognosis (mesothelioma)	French MESOBANK	TCGA	ResNet-50	224×224 pixels (20×)	C index: 0.643
Skrede et al.[68]	Prognosis (colorectal cancer)	AkUH, AUH, GCCS, VICTOR trial	QUSAR2 trial	MobileNet v2	448×448 pixels (10×, 40×)	HR: 3.84
Kulkarni et al.[69]	Recurrence and prognosis (melanoma)	CUIMC, NYUMC, GHS, ISMMS	YSM	Simple 5-layers CNN with RNN	100×100 pixels (8×)	AUROC: 0.905 HR: 58.7
Wulczyn et al.[70]	Prognosis (various cancers)	TCGA	None	Similar to MobileNet	256×256 pixels (N/A)	HR: 1.58
Fu at al.[14]	Prognosis (various cancers)	TCGA	METABRIC	Inception v4	512×512 pixels (20×)	C index: 0.53–0.67
Saillard et al.[71]	Prognosis (liver cancer)	HMUH	TCGA	ResNet	224×224 pixels (20×)	C index: 0.78
Wang et al.[72]	Prognosis (stomach cancer)	CHH and JXCH	None	ResNet-50	768×768 pixels (20×)	HR: 2.05
Wulczyn et al.[73]	Prognosis (colorectal cancer)	MUG	None	Similar to MobileNet	256×256 pixels (5×)	AUROC: 0.69–0.70
Shim et al.[74]	Recurrence (lung cancer)	3 hospitals from CUK	2 hospitals from CUK	ResNet-50	224×224 pixels (10×, 40×)	AUROC: 0.76–0.77

PD1, programmed death 1; TCGA, The Cancer Genome Atlas; PUCH, Peking University Cancer Hospital; AUROC, area under the receiver operating characteristics curve; NYU, New York University; HUCH, Helsinki University Central Hospital; DACHS, Darmkrebs Chancen der Verhütung durch Screening; HR, hazard ratio; MESOBANK, mesothelioma biobank; AkUH, Akershus University Hospital; AUH, Aker University Hospital; GCCS, Gloucester Colorectal Cancer Study; VICTOR, venetoclax with low dose cytarabine for acute myeloid leukaemia; QUSAR2, QUick And Simple And Reliable2; CUIMC, Columbia University Irving Medical Center; NYUMC, New York University Medical Center; GHS, Geisinger Health Systems; ISMMS, Icahn School of Medicine at Mount Sinai; YSM, Yale School of Medicine; CNN, convolutional neural network; RNN, recurrent neural network; N/A, not applicable; METABRIC, Molecular Taxonomy of Breast Cancer International Consortium; HMUH, Henri Mondor University Hospital; CHH, Changhai Hospital; JXCH, Jiangxi Provincial Cancer Hospital; MUG, Medical University of Graz; CUK, Catholic University of Korea.

based tissue analysis and morphology-based features. The model predicted distant recurrence with an AUROC of 0.905 and DSS with an HR of 58.7.

From the combined datasets of 10 cancers, including bladder, breast, colon, head and neck, kidney, liver, lung, ovary, and stomach cancers, Wulczyn et al.[70] tried to predict prognosis. They mixed tissue image patches from patients from all 10 cancer groups with good and poor prognoses and then trained a unified model. However, the threshold to determine the low- and high-risk groups of each cancer was set individually for better performance. The overall HR for the 10 cancer types was 1.58. The model could sub-stratify low- and high-risk groups across stage II and III cancers but not stage I and IV cancers. When separately analyzed by cancer type, DSS was predicted for breast, colon, head and neck, kidney, and liver cancers with statistical significance.

Fu et al.[14] also tried prognosis prediction for multiple cancers. They reused features from the Inception v4-based classifier for discrimination of tissue types to train a predictive model for OS. Compared to the prediction based on tumor grade and subtype, the DL model yielded significantly better results for 15 of 18 cancer types, with concordance ranging from 0.53 to 0.67.

Saillard et al.[71] trained predictive models for survival of hepatocellular carcinoma (HCC) patients. Their models could stratify patients with different prognoses even after stratification for other clinicopathologic features, such as disease stage, satellite nodules, alpha-fetoprotein serum level, and vascular invasion. The results indicated that the model could capture unique information from tissue images that is nonredundant with other variables known to affect survival.

Wang et al.[72] adopted a unique approach to extract prognostic information from the dissected lymph nodes of GC patients. By applying a segmentation network, they first extracted lymph node regions and removed other tissues, such as fat and muscle. Then metastasized tumor and normal lymph node tissues were separated with another classifier network. Finally, the area ratio of metastasized tumor vs. lymph node was calculated as a candidate marker for prediction of prognosis. The ratio could discriminate between good and poor prognosis groups with an HR of 2.05.

Wulczyn et al.[73] tried to predict 5-year DSS for patients with stage II and III CRC. They first separated tumor tissue with the Inception-v3-based classifier. Then, tumor tissues were used to train a MobileNet-based prognosis prediction model. The model predicted 5-year DSS with the AUROCs of 0.70 and 0.69 for stage II and III cancer, respectively. When they investigated the correlation of clinicopathologic features and high-risk scores from the DL model, higher T and N categories showed higher risk scores.

Shim et al.[74] tried to predict the recurrence of early-stage lung adenocarcinoma. They adopted a multi-scale approach by combining the feature layers from two independent neural networks analyzing 10× and 40× tissue images. Their model predicted recurrence with the AUROCs of 0.77 and 0.76 for external validation cohorts I and II, respectively. They also found a high probability of recurrence to be associated with tumor necrosis, discohesive tumor cells, and atypical nuclei.

OTHER APPLICATIONS OF DL FOR TISSUE ANALYSIS

Because many pathologic evaluation processes suffer from inter- and intra-observer variability, DL can help to improve the reliability of quantitative evaluation of tissue slides.[75] Therefore, in addition to the prediction of molecular biomarkers, DL can be applied to assist many other pathologic tasks. In this section, these other application fields will be briefly introduced with liver tissue assessment as an example. First of all, DL can assist basic diagnosis of HCC by discriminating tumor tissues from normal liver tissues. Pathologists should also discriminate between HCC and cholangiocarcinoma during the assessment of primary liver cancers. A DL-based assistant tool improved the accuracy of the pathologists' discrimination between HCC and cholangiocarcinoma.[76] Furthermore, tumor tissues can be subclassified by DL depending on differentiation status.[12,77] As these examples showed, DL can provide important tools to assist liver cancer assessment from tissue slides. In addition, liver tissues should be analyzed to evaluate other liver diseases. For the evaluation of nonalcoholic fatty liver disease, histologic changes such as ballooned hepatocytes or fat accumulation can be assessed with DL.[78] DL-based automatic assessment of liver fibrosis can improve liver fibrosis quantification and scoring.[79] During liver transplantation, pathologists should evaluate donor liver biopsies for accepting or discarding the donor liver. DL can help to quantify the percentage of steatosis to evaluate frozen biopsies of donor liver.[80] Therefore, DL can be

used to assist almost every aspect of pathologic workflow to improve the objective evaluation of tissue slides.

DISCUSSION

Although these studies clearly demonstrated the potential of DL for prediction of molecular biomarkers, the performance is generally far behind the threshold for clinical applicability. However, we expect that performance will be improved soon because more data for the training of the DL system will be available within a few years. Many studies demonstrated that performance can be enhanced when multi-national and multi-institutional data are collected to be used in training. It has been less than 5 years since digital slide scanners have been widely adopted in hospitals. Therefore, we are still in an early phase of digital pathology, and much more data will likely be accumulated in the near future. It will be interesting to see the improvement in performance with ongoing research.

Small image patch-based classification provides one of the strongest advantages of the DL-based approach. As demonstrated in Figure 2, tumor heterogeneity can be automatically revealed by this approach. When the classification result of each small image patch is overlaid on a WSI, a detailed distribution of tissue regions with different molecular profiles can be easily analyzed. Although spatial sequencing techniques, such as single-cell sequencing and spatial transcriptomics technology, can provide information on tumor heterogeneity,[81,82] technical difficulty and high cost limit their application in the clinic. The innate ability of the DL method to reveal spatial tumor heterogeneity can be an alternative tool to these costly methods. Many studies utilized the heterogeneity information provided by DL-based classifiers to investigate the impact of intratumoral heterogeneity.[17,20,35,45,61] Generally, highly heterogeneous tumors impart poorer prognoses than homogenous tumors. Tumor heterogeneity is also a source of difficulty for the correct prediction of molecular biomarkers. When tissue has very heterogeneous classification results for a biomarker, it is difficult to determine whether the tissue is negative or positive for the biomarker. Studies aimed at the correct prediction of biomarkers from highly heterogeneous classification results are necessary to improve the reliability of DL-based molecular biomarker prediction.
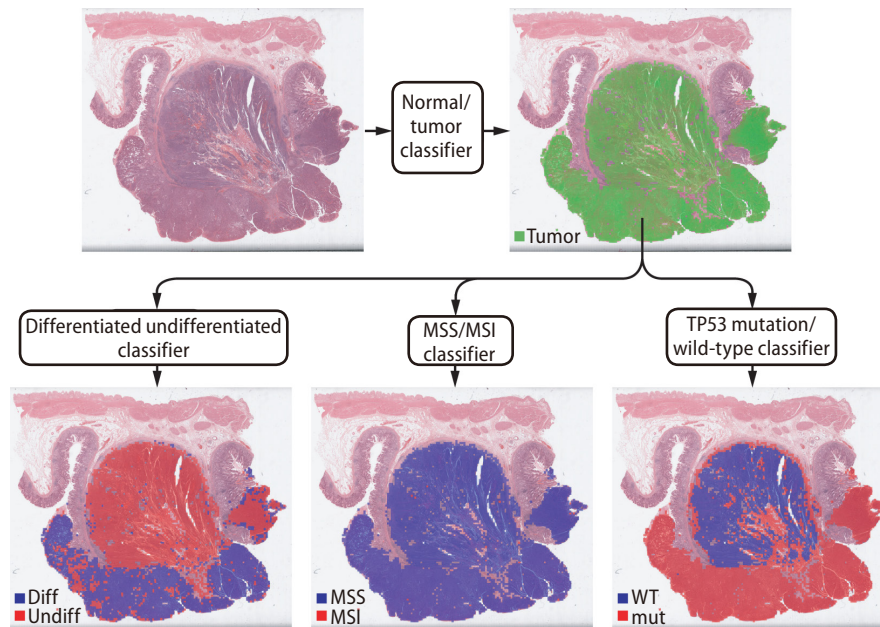

Figure 2. Representative example of a gastric tissue slide showing different levels of heterogeneity for different entities. Upper part: only tumor tissue patches are selected for the next classification tasks. Lower part: selected tumor patches are classified for tumor differentiation status, microsatellite instability (MSI) status, and TP53 mutational status. The tissue is heterogeneous for tumor differentiation status, homogeneous for MSI status, and heterogeneous for TP53 mutational status. diff, differentiated; undiff, undifferentiated; MSS, microsatellite stable; WT, wild-type; mut, mutated.

The ability of DL to perform detailed analysis of tumor heterogeneity from multiple tissue regions is specifically advantageous for liver cancer. Multiple tumor nodules in the liver can have different grades and heterogenous molecular profiles because they can arise from genetically independent clones.[83] Because tumor grades and molecular profiles are important for prognosis of cancer patients, it is inappropriate to make treatment decisions based on a small portion of heterogeneous tumor tissues. Therefore, detailed analysis of multiple tissue regions in multiple tumor nodules is imperative for accurate assessment of liver cancer. DL can help to quantitatively analyze tumor tissue grades and molecular properties in whole tissue specimens in a cost-effective manner. Improved prognostication and better treatment decision using information from the DL system can enhance treatment response for liver cancer patients.

Another strength of the DL-based approach on the H&E-stained tissue slides is that it can be widely adopted for retrospective analysis of cancer treatment response and prognosis. As described, H&E-stained slides are available for almost all cancer patients and can be used to study the correlation of specific biomarkers with treatment response and prognosis. No additional specimens are necessary for DL-based analysis. Since the cost of DL-based molecular biomarkers prediction is also negligible, retrospective studies will likely be encouraged. Therefore, clinicians could utilize the data from their previous treatment experience with DL models to study the impact of specific molecular biomarkers. Furthermore, the costs of prospective clinical trials can be reduced. Because drug responses significantly differ between patients with different mutational profiles and gene expression profiles, many clinical trials have started to adopt the umbrella platform strategy, which assigns treatment arms based on the specific molecular traits of cancer patients.[84,85] It is very costly to determine the molecular profiles of patients in a trial. If the DL-based approach can be made applicable, the costs of molecular tests for patient assignment will be greatly reduced.

Although the DL approach is very promising, there are some hurdles to clinical adoption. First, the black-box nature of DL limits the interpretability of DL-based prediction results. Because a DNN has millions to billions of numerical parameters to be modified during the training process, it is very hard to understand how a DNN classifies a tissue image patch into a specific class. Without an understanding of the basis of the decision, it will be difficult to trust a DL system. Efforts for explainable artificial intelligence will help to enhance the interpretability of DL systems.[86] Another barrier is the need for an individual DL model for every biomarker for each specific cancer. The generalizability of DL-based prediction models for molecular biomarkers is not very high. Therefore, prediction models for each biomarker should be separately developed for each cancer type. It will take years to develop an entire set of clinically actionable prediction systems for molecular biomarkers in major cancers.

The feasibility of DL-based prediction of molecular cancer biomarkers has been extensively studied thus far. Based on the promising results from these studies, we expect that DL models will be widely adopted to support clinical decisions for management of cancer patients in the near future. A DL system can be applied either as a pre-screening tool or as a post-test quality management tool. With pre-screening, definitive cases will not be further tested to save cost. For quality management, diagnosis and molecular test results can be reassured by DL to enhance the reliability of the pathologic reports. Therefore, DL will be an essential tool in the era of precision oncology.

Authors' contribution

Acknowledgements

Conflicts of Interest

The authors have no conflicts to disclose.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Clinical and Molecular Hepatology website (http://www.e-cmh.org).

REFERENCES

1. Chae YK, Pan AP, Davis AA, Patel SP, Carneiro BA, Kurzrock R, et al. Path toward precision oncology: review of targeted therapy studies and tools to aid in defining "actionability" of a molecular lesion and patient management support. Mol Cancer Ther 2017;16:2645-2655.

2. Murchan P, Ó'Brien C, O'Connell S, McNevin CS, Baird AM, Sheils O, et al. Deep learning of histopathological features for the prediction of tumour molecular genetics. Diagnostics (Basel) 2021;11:1406.

3. 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;16:703-715.

4. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.

5. Costa DB, Nguyen KS, Cho BC, Sequist LV, Jackman DM, Riely GJ, et al. Effects of erlotinib in EGFR mutated non-small cell lung cancers with resistance to gefitinib. Clin Cancer Res 2008;14: 7060-7067.

6. Kopetz S, Guthrie KA, Morris VK, Lenz HJ, Magliocco AM, Maru D, et al. Randomized trial of irinotecan and cetuximab with or without vemurafenib in BRAF-mutant metastatic colorectal cancer (SWOG S1406). J Clin Oncol 2021;39:285-294.

7. Hong DS, Fakih MG, Strickler JH, Desai J, Durm GA, Shapiro GI, et al. KRASG12C Inhibition with sotorasib in advanced solid tumors. N Engl J Med 2020;383:1207-1217.

8. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. BioRxiv. 2018 Oct 1. doi: 10.1101/064279.

9. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-1567.

10. Kim RH, Nomikou S, Coudray N, Jour G, Dawood Z, Hong R, et al. A deep learning approach for rapid mutational screening in melanoma. BioRxiv. 2020 Aug 19. doi: 10.1101/610311.

11. Tsou P, Wu CJ. Mapping driver mutations to histopathological subtypes in papillary thyroid carcinoma: applying a deep convolutional neural network. J Clin Med 2019;8:1675.

12. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis Oncol 2020;4:14.

13. Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020;10:7733.

14. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer 2020;1:800-810.

15. 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;1:789-799.

16. Jang HJ, Lee A, Kang J, Song IH, Lee SH. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. World J Gastroenterol 2020;26:6207-6223.

17. Noorbakhsh J, Farahmand S, Foroughi Pour A, Namburi S, Caruana D, Rimm D, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nat Commun 2020;11:6367.

18. Jang HJ, Song IH, Lee SH. Generalizability of deep learning system for the pathologic diagnosis of various cancers. Appl Sci 2021;11:808.

19. Yang Y, Yang J, Liang Y, Liao B, Zhu W, Mo X, et al. Identification and validation of efficacy of immunological therapy for lung cancer from histopathological images based on deep learning. Front Genet 2021;12:642981.

20. Loeffler CML, Ortiz Bruechle N, Jung M, Seillier L, Rose M, Laleh NG, et al. Artificial intelligence–based detection of FGFR3 mutational status directly from routine histology in bladder cancer: a possible preselection for molecular testing? Eur Urol Focus 2022;8:472-479.

21. Jang HJ, Lee A, Kang J, Song IH, Lee SH. Prediction of genetic alterations from gastric cancer histopathology images using a fully automated deep learning approach. World J Gastroenterol 2021;27:7687-7704.

22. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite instability: a review of what the oncologist should know. Cancer Cell Int 2020;20:16.

23. Llosa NJ, Cruise M, Tam A, Wicks EC, Hechenbleikner EM, Taube JM, et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. Cancer Discov 2015;5:43-51.

24. Eso Y, Shimizu T, Takeda H, Takai A, Marusawa H. Microsatellite instability and immune checkpoint inhibitors: toward precision medicine against gastrointestinal and hepatobiliary cancers. J Gastroenterol 2020;55:15-26.

25. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 2005;23:609-618.

26. 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; 25:1054-1056.

27. Cao R, Yang F, Ma SC, 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;10:11080-11091.

28. 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;159:1406-1416.e11.

29. Wang T, Lu W, Yang F, Liu L, Dong Z, Tang W, et al. Microsatellite instability prediction of uterine corpus endometrial carcinoma based on H&E histology whole-slide imaging; 2020 Apr 3-7; Iowa City, IA, USA. Manhattan, New York: IEEE; 2020 May. 1289-1292 p.

30. Krause J, Grabsch HI, Kloor M, Jendrusch M, Echle A, Buelow RD, et al. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. J Pathol 2021;254:70-79.

31. 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;22:132-141.

32. Lee SH, Song IH, Jang HJ. Feasibility of deep learning-based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. Int J Cancer 2021;149:728-740.

33. Galuppini F, Dal Pozzo CA, Deckert J, Loupakis F, Fassan M, Baffa R. Tumor mutation burden: from comprehensive mutational screening to the clinic. Cancer Cell Int 2019;19:209.

34. Klempner SJ, Fabrizio D, Bane S, Reinhart M, Peoples T, Ali SM, et al. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: a review of current evidence. Oncologist 2020;25:e147-e159.

35. Xu H, Park S, Lee SH, Hwang TH. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. BioRxiv. 2020 Dec 20. doi: 10.1101/554527.

36. Jain MS, Massoud TF. Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nat Mach Intell 2020;2:356-362.

37. Xu H, Park S, Clemenceau JR, Radakovich N, Lee SH, Hwang TH. Deep transfer learning approach to predict tumor mutation burden (TMB) and delineate spatial heterogeneity of TMB within tumors from whole slide images. Cold Spring Harbor Lab 2020;1:554527.

38. Shimada Y, Okuda S, Watanabe Y, Tajima Y, Nagahashi M, Ichikawa H, et al. Histopathological characteristics and artificial intelligence for predicting tumor mutational burden-high colorectal cancer. J Gastroenterol 2021;56:547-559.

39. Sadhwani A, Chang HW, Behrooz A, Brown T, Auvigne-Flament I, Patel H, et al. Comparative analysis of machine learning approaches to classify tumor mutation burden in lung adenocarcinoma using histopathology images. Sci Rep 2021;11:16605.

40. Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer 2017;17:268.

41. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015;21:449-456.

42. Hu F, Zhou Y, Wang Q, Yang Z, Shi Y, Chi Q. Gene expression classification of lung adenocarcinoma into molecular subtypes. IEEE/ACM Trans Comput Biol Bioinform 2020;17:1187-1197.

43. Chia SK, Bramwell VH, Tu D, Shepherd LE, Jiang S, Vickery T, et al. A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. Clin Cancer Res 2012;18:4465-4472.

44. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 2018;4:30.

45. Jaber MI, Song B, Taylor C, Vaske CJ, Benz SC, Rabizadeh S, et al. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. Breast Cancer Res 2020;22:12.

46. Hong R, Liu W, DeLair D, Razavian N, Fenyö D. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. Cell Rep Med 2021;2:100400.

47. 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;70:544-554.

48. Yu KH, Wang F, Berry GJ, Ré C, Altman RB, Snyder M, et al. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. J Am Med Inform

Assoc 2020;27:757-769.

49. Woerl AC, Eckstein M, Geiger J, Wagner DC, Daher T, Stenzel P, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. Eur Urol 2020;78:256-264.

50. Reis-Filho JS, Westbury C, Pierga JY. The impact of expression profiling on prognostic and predictive testing in breast cancer. J Clin Pathol 2006;59:225-231.

51. Nagy Á, Munkácsy G, Győrffy B. Pancancer survival analysis of cancer hallmark genes. Sci Rep 2021;11:6047.

52. El Sayed R, El Jamal L, El Iskandarani S, Kort J, Abdel Salam M, Assi H. Endocrine and targeted therapy for hormone-receptor-positive, HER2-negative advanced breast cancer: insights to sequencing treatment and overcoming resistance based on clinical trials. Front Oncol 2019;9:510.

53. 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;21:1350-1356.

54. Lumachi F, Brunello A, Maruzzo M, Basso U, Basso SM. Treatment of estrogen receptor-positive breast cancer. Curr Med Chem 2013;20:596-604.

55. Nishino M, Ramaiya NH, Hatabu H, Hodi FS. Monitoring immune-checkpoint blockade: response evaluation and biomarker development. Nat Rev Clin Oncol 2017;14:655-668.

56. Sha L, Osinski BL, Ho IY, Tan TL, Willis C, Weiss H, et al. Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. J Pathol Inform 2019;10:24.

57. Rawat RR, Ortega I, Roy P, Sha F, Shibata D, Ruderman D, et al. Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. Sci Rep 2020;10:7275.

58. Naik N, Madani A, Esteva A, Keskar NS, Press MF, Ruderman D, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat Commun 2020;11:5727.

59. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng 2020;4:827-834.

60. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun 2020;11:3877.

61. Levy-Jurgenson A, Tekpli X, Kristensen VN, Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Sci Rep 2020;10:18802.

62. Hu J, Cui C, Yang W, Huang L, Yu R, Liu S, et al. Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images. Transl Oncol 2021;14:100921.

63. Johannet P, Coudray N, Donnelly DM, Jour G, Illa-Bochaca I, Xia Y, et al. Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma. Clin Cancer Res 2021;27:131-140.

64. 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.

65. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018;115:E2970-E2979.

66. 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.

67. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med 2019;25:1519-1525.

68. Skrede OJ, 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;395:350-360.

69. Kulkarni PM, Robinson EJ, Sarin Pradhan J, Gartrell-Corrado RD, Rohr BR, Trager MH, et al. Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. Clin Cancer Res 2020;26:1126-1134.

70. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS One 2020;15:e0233678.

71. Saillard C, Schmauch B, Laifa O, Moarii M, Toldo S, Zaslavskiy M, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology 2020;72:2000-2013.

72. Wang X, Chen Y, Gao Y, Zhang H, Guan Z, Dong Z, et al. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nat Commun

2021;12:1637.

73. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med 2021;4:71.

74. Shim WS, Yim K, Kim TJ, Sung YE, Lee G, Hong JH, et al. Deep-RePath: identifying the prognostic features of early-stage lung adenocarcinoma using multi-scale pathology images and deep convolutional neural networks. Cancers (Basel) 2021;13:3308.

75. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform 2018; 9:38.

76. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med 2020;3:23.

77. Aatresh AA, Alabhya K, Lal S, Kini J, Saxena PUP. LiverNet: efficient and robust deep learning model for automatic diagnosis of sub-types of liver hepatocellular carcinoma cancer from H&E stained liver histopathology images. Int J Comput Assist Radiol Surg 2021;16:1549-1563.

78. Arjmand A, Angelis CT, Christou V, Tzallas AT, Tsipouras MG, Glavas E, et al. Training of deep convolutional neural networks to identify critical liver alterations in histopathology image samples. Appl Sci 2020;10:42.

79. Yu Y, Wang J, Ng CW, Ma Y, Mo S, Fong ELS, et al. Deep learning enables automated scoring of liver fibrosis stages. Sci Rep 2018;8:16016.

80. Sun L, Marsh JN, Matlock MK, Chen L, Gaut JP, Brunt EM, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. EBioMedicine 2020;60:103029.

81. Zhang Y, Wang D, Peng M, Tang L, Ouyang J, Xiong F, et al. Single-cell RNA sequencing in cancer research. J Exp Clin Cancer Res 2021;40:81.

82. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature 2021;596:211-220.

83. Chan LK, Tsui YM, Ho DW, Ng IO. Cellular heterogeneity and plasticity in liver cancer. Semin Cancer Biol 2022;82:134-149.

84. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol 2014;11:81-90.

85. Lee J, Kim ST, Kim K, Lee H, Kozarewa I, Mortimer PGS, et al. Tumor genomic profiling guides patients with metastatic gastric cancer to targeted treatment: the VIKTORY umbrella trial. Cancer Discov 2019;9:1388-1405.

86. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-215.