Integrative deep learning analysis improves colon adenocarcinoma patient stratification at risk for mortality

Jie Zhou,[a,b] Ali Foroughi pour,[a] Hany Deirawan,[c,d] Fayez Daaboul,[c] Thazin Nwe Aung,[e] Rafic Beydoun,[c] Fahad Shabbir Ahmed,[c,**] and Jeffrey H. Chuang[a,b,*]

[a]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[b]Department of Genetics and Genome Sciences, UCONN Health, Farmington, CT, USA
[c]Department of Pathology, Wayne State University, Detroit, MI, USA
[d]Department of Dermatology, Wayne State University, Detroit, MI, USA
[e]Department of Pathology, Yale University, New Haven, CT, USA

eBioMedicine
2023;94: 104726
Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104726

Summary

Background Colorectal cancers are the fourth most diagnosed cancer and the second leading cancer in number of deaths. Many clinical variables, pathological features, and genomic signatures are associated with patient risk, but reliable patient stratification in the clinic remains a challenging task. Here we assess how image, clinical, and genomic features can be combined to predict risk.

Methods We developed and evaluated integrative deep learning models combining formalin-fixed, paraffin-embedded (FFPE) whole slide images (WSIs), clinical variables, and mutation signatures to stratify colon adenocarcinoma (COAD) patients based on their risk of mortality. Our models were trained using a dataset of 108 patients from The Cancer Genome Atlas (TCGA), and were externally validated on newly generated dataset from Wayne State University (WSU) of 123 COAD patients and rectal adenocarcinoma (READ) patients in TCGA (N = 52).

Findings We first observe that deep learning models trained on FFPE WSIs of TCGA-COAD separate high-risk (OS < 3 years, N = 38) and low-risk (OS > 5 years, N = 25) patients (AUC = 0.81 ± 0.08, 5 year survival $p < 0.0001$, 5 year relative risk = 1.83 ± 0.04) though such models are less effective at predicting overall survival (OS) for moderate-risk (3 years < OS < 5 years, N = 45) patients (5 year survival p-value = 0.5, 5 year relative risk = 1.05 ± 0.09). We find that our integrative models combining WSIs, clinical variables, and mutation signatures can improve patient stratification for moderate-risk patients (5 year survival $p < 0.0001$, 5 year relative risk = 1.87 ± 0.07). Our integrative model combining image and clinical variables is also effective on an independent pathology dataset (WSU-COAD, N = 123) generated by our team (5 year survival $p < 0.0001$, 5 year relative risk = 1.52 ± 0.08), and the TCGA-READ data (5 year survival $p < 0.0001$, 5 year relative risk = 1.18 ± 0.17). Our multicenter integrative image and clinical model trained on combined TCGA-COAD and WSU-COAD is effective in predicting risk on TCGA-READ (5 year survival $p < 0.0001$, 5 year relative risk = 1.82 ± 0.13) data. Pathologist review of image-based heatmaps suggests that nuclear size pleomorphism, intense cellularity, and abnormal structures are associated with high-risk, while low-risk regions have more regular and small cells. Quantitative analysis shows high cellularity, high ratios of tumor cells, large tumor nuclei, and low immune infiltration are indicators of high-risk tiles.

Interpretation The improved stratification of colorectal cancer patients from our computational methods can be beneficial for treatment plans and enrollment of patients in clinical trials.

Funding This study was supported by the National Cancer Institutes (Grant No. R01CA230031 and P30CA034196). The funders had no roles in study design, data collection and analysis or preparation of the manuscript.

Keywords: Deep learning; Colorectal cancer; Overall survival; Digital pathology; Multimodal analysis

*Corresponding authors. 10 Discovery Dr, Farmington, CT 06032, USA.
**Corresponding author.
E-mail addresses: jeff.chuang@jax.org (J.H. Chuang), fahadahmed@wayne.edu (F.S. Ahmed).

Research in context

Evidence before this study

Prior to beginning this study in January 2021, we performed a literature review of published articles on PubMed using the keywords: "deep learning" AND "colorectal cancer" AND "survival" OR "prognosis" AND "integrative model" OR "multimodal", without restrictions on date or language. Our search did not reveal any research that had combined deep learning-extracted pathological image features, clinical data, and molecular features to stratify patients with colon adenocarcinoma (COAD) based on their risk of mortality, although individual data modality was reported to be associated with patient outcome. A few papers had examined only using pathological images to predict survival, without external validation beyond The Cancer Genome Atlas (TCGA). We hypothesized integrative models can improve patient stratification, and investigated this research question. Since then some articles have reported combining pathological and clinical features, but these have done so at the slide level, an approach that does not account for intratumoral heterogeneity. We hypothesized and investigated if integrative models that better account for tumor heterogeneity and patient's clinical context are more effective for risk prediction. We also hypothesized that an integrative deep learning model would be able to stratify borderline patients.

Added value of this study

In this study, we developed tile-based integrative models combining whole slide images (WSIs), clinical variables, and mutation signatures to stratify COAD patients based on their risk of mortality, and we externally validated the model on newly generated patient datasets. Our integrative model achieves patient stratification (5-year RR = 4.01 in TCGA) comparable to prior state-of-the-art large-scale studies trained on >5000 patients (Wulczyn, E. et al., 2020, RR = 3.35 and 2.7 on two validation sets), but we achieved this with <300 patients.

Our models are trained using a dataset of 108 COAD patients from TCGA, and are validated on an external cohort of 123 patients from Wayne State University (WSU-COAD) and 52 rectal adenocarcinoma patients from TCGA (TCGA-READ). In the TCGA-COAD cohort, we observe that deep learning models based only on WSI images accurately separate high-risk (OS < 3 years, N = 38) from low-risk (OS > 5 years, N = 25) patients (AUC = 0.81 ± 0.08, 5 year survival p < 0.0001, 5 year relative risk = 1.83 ± 0.04).
Within the clinically challenging class of moderate risk patients (3 years < OS < 5 years, N = 45), the image-only model is not effective at distinguishing patients with higher or lower risk (5 year survival p-value = 0.5, 5 year relative risk = 1.05 ± 0.09). However, our integrative models combining WSIs, clinical variables, and mutation signatures improve patient stratification for these moderate-risk patients (5 year survival p < 0.0001, 5 year relative risk = 1.87 ± 0.07). More importantly, our integrative model combining image and clinical variables is also effective at predicting patient mortality risk on the external validation cohort (WSU-COAD, 5 year survival p < 0.0001, 5 year relative risk = 1.52 ± 0.08) and TCGA-READ (5 year survival p < 0.0001, 5 year relative risk = 1.18 ± 0.17).

Implications of all the available evidence

Our results show architectural improvements to predictive models that enhance patient stratification. This can be useful to develop personalized treatment plans, such as closer follow-up for those with higher predicted risk. This is particularly beneficial for patients considered under current standards to be moderate risk, as their prognoses are often uncertain. Using our models to inform treatment strategies may improve survival rates and reduce mortality rates among COAD patients. Our integrative model is more efficient than other models at combining pathological and clinical features. Therefore, it can also guide design of accurate models for other cancer types while minimizing data requirements.

Introduction

Stratification of colon adenocarcinoma patients is based on standards established by the American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC)[1] and remains a challenging clinical decision. Colon adenocarcinoma (COAD) has an overall all-stages SEER 5-year survival of 63%,[2,3] and risk assessments impact decisions such as whether a patient receives additional chemotherapy or is inducted into a clinical trial. Tumor infiltrating lymphocytes (TIL) quantifications have been shown to be informative in recent years.[4–6] Nevertheless, improvements in biomarkers remain critical, either through incorporating additional biomarkers or better use of currently identified markers,[7–9] particularly for patients without a clear indication of high/low risk.[10,11] Clinical assessment of these patients can be difficult, hampering decisions about additional treatment, uptake of patients into clinical-trials, and proactive disease surveillance.[12,13] Therefore, automated computational models on patient data, including histopathology images, can address important needs in assessment and reproducibility of cancer management decisions.

Deep learning models have achieved high accuracy for detecting tumor regions[14] and identifying cancer subtypes[15] from hematoxylin and eosin (H&E)-stained whole slide images (WSIs). Such models have also been able to predict several clinically relevant genetic features, such as microsatellite instability (MSI)[16] and mutation status of key genes[17–19] with moderate accuracy. Deep learning models using WSIs have been studied to stratify patients based on survival risk.[20] However, these

models have room for improvement as they tend to only utilize a single data modality and rely on large datasets for model training, including in recent studies of colorectal cancer.[21,22]

We hypothesize that integrating H&E image data with other data modalities can improve risk stratification since clinical variables, mutation signatures, and gene expression profiles have individually been shown to be informative.[23] To address this question, we develop and evaluate integrative deep learning models that combine morphological features from H&E WSIs, clinical variables, MSI-status, and mutation status of key genes.[24–31] While prior studies have combined patient-level image features from WSIs with patient-level clinical variables, to the best of our knowledge, our work is the first to train at the tile level with patient-level information based on context-aware learning,[32] which we find improves performance.

We show that integrative analysis improves patient stratification and enables training of reliable models using smaller sample sizes, which we demonstrate using TCGA-COAD[33] and an independently generated dataset from Wayne State University (WSU-COAD). We further validate our COAD models on rectal adenocarcinoma patients (TCGA-READ). Our integrative model demonstrates superior performance to models using only one data type and is more robust to staining differences than a model using only WSIs. Our model outputs interpretable heatmaps, which are informative of morphologies of high risk. These results demonstrate how integrative computational analysis of colorectal adenocarcinomas can improve prediction of outcomes.

Methods

We have provided a simplified explanation of our methodological process in the Supplementary Methodological Appendix. This is intended to facilitate comprehension for readers with a clinical background. For those interested in a deeper understanding of the machine learning processes involved, the full technical details are available in this Methods section.

Data and study design

TCGA-COAD cohort

336 Formalin-Fixed Paraffin-Embedded (FFPE) hematoxylin and eosin (H&E) stained TCGA-COAD WSIs were downloaded from the GDC (Genomic Data Commons) data portal. The following clinical variables were downloaded from the cBioPortal webpage[34,35]: patient age at diagnosis, gender, tumor (T) stage, nodes (N) stage, and metastasis (M) stage. Mutation statuses of 207 genes were downloaded from the cBioPortal webpage (see Supplementary Figure S10 for the full gene list). Patients were grouped by their overall survival (OS): low-risk (LR, OS > 5 years, N = 25), moderate-risk (MR, 3<OS < 5 years, N = 45), high-risk (HR, OS < 3 years, N = 38), and loss to follow-up (time to last follow up < 3 years, patient status: alive, N = 228 censored patients).

WSIs from these 228 censored patients were also used for the training of the computational tumor detector. We used these patients for this task because their WSIs are disjoint from the WSIs used to train the survival predictions. This removes any potential spurious correlations that could arise if the tumor/non-tumor separator were trained on the same WSIs as used for the survival model.

Wayne State University validation cohort

123 patients' H&E stained FFPE samples and corresponding clinical data were collected from Wayne State University (WSU). The clinical data include patient age at diagnosis, gender, T stage, N stage and M stage. Patients were grouped as HR (N = 17), LR (N = 97) and MR (N = 9). There was no loss to follow-up in this cohort.

Combined multi-center cohort: A multi-center cohort (N = 115) was obtained by combining the WSU (N = 43) validation cohort and TCGA-COAD (N = 72). TCGA-COAD slides obtained from Indivumed (N = 36) were removed (see Model training and assessment section). Patients were grouped as HR (N = 39), LR (N = 42) and MR (N = 34).

TCGA-READ cohort

165 FFPE TCGA-READ WSIs corresponding to 164 patients were downloaded from the GDC data portal. The following clinical variables were downloaded from the cBioPortal webpage: patient age at diagnosis, gender, tumor (T) stage, nodes (N) stage, and metastasis (M) stage. Mutation statuses of 207 genes were downloaded from the cBioPortal webpage. Patients were grouped by their overall survival (OS): low-risk (LR, OS > 5 years, N = 13), moderate-risk (MR, 3<OS < 5 years, N = 22), high-risk (HR, OS < 3 years, N = 17), and loss to follow-up (time to last follow up < 3 years, patient status: alive, N = 112 censored patients).

Data preparation

Tumor annotation

Tumor regions of WSIs were annotated by 3 expert pathologists by multi-scoping the specimens to produce a consensus among the slides. Pathologists used the Aperio ImageScope software version 12.4.3[36] for annotation. Our pathologists only annotated highly pure tumor regions. Tumor areas were exported from the Aperio software in The Extensible Markup Language (XML) format, with X and Y coordinates corresponding to the annotated tumor regions. Tumor masks were generated for each slide image by connecting the coordinates, dilating, and eroding the areas using the OpenCV package in python.[37]

Image pre-processing

H&E stained WSIs were acquired in SVS format. All images were downsampled to 20 × magnification, corresponding to a resolution of 0.5 μm/pixel. Each WSI was manually reviewed and the tumor area was annotated by expert pathologists. Regions with excess background or containing no tissue were removed as previously described.[38] Image slides were tiled into non-overlapping patches of 512 × 512 pixels. Tiles with >50% overlap with tumor masks were labeled as tumor tiles. The remaining tiles were labeled non-tumor.

Clinical data pre-processing

Five clinical variables related to patient outcomes were selected by the pathology team: age at diagnosis, gender, and TNM (Tumor, Nodes, and Metastasis) staging of colonic adenocarcinomas: tumor (T) stage, nodes (N) stage, and metastases (M) stage based on the college of American pathology (CAP) protocol for Colon and Rectum, Resection 2021 (v4.2.0.0)). Age was encoded numerically, and other variables were encoded as integers.

Molecular data pre-processing: 207 genes from 11 canonical cancer pathways[24–31] and the top 11 most commonly mutated genes in TCGA-COAD were selected. A 10% threshold was used to filter out genes that are not frequently mutated in TCGA-COAD patients, resulting in a total of 26 genes (see Supplementary Figure S10). Microsatellite instability (MSI) status was also considered due to its impact in colon cancers.[39] Supplementary Table S1 describes patient characteristics.

Model training and assessment

Train-test splits and cross-validation

We used Monte Carlo cross-validation to assess the model performance. When training and testing on the same cohort, we randomly split our cohort into paired training (70%) and testing (30%) sets to generate 100 training/testing set pairs. The predictive accuracy was assessed in each split. The results (AUC values and survivorship values) were then averaged over the splits.

For the cross-validation analysis when testing on all patients, we randomly split the entire data set into a 70%/30% train/test split. We then selected only the high and low risk samples from the 70% subset for model training. The test set, constituting the remaining 30%, included all sample types: high risk, low risk, moderate risk, and those patients lost to follow-up within 3 years. This approach allows one to perform cross-validation analyses on uncensored test sets while training on binarized high/low sets.

For tests focused on how moderate risk samples impact survival prediction (for example, when adding the moderate risk patients to the high + low for testing, but training on high + low), we used the following procedure. We randomly split the high + low set into a 70%/30% train/test split. We then took the 30% test set and added the remaining moderate risk samples. This procedure was used in the analyses of "Images are informative of colon adenocarcinoma risk" and "Integrative analysis improves stratification of moderate risk patients" sections.

Network architecture

InceptionV3[40] features pre-trained on Image-Net[41] were fed to a two-layer multi-layer perceptron (MLP) following the parameters of[38]: The first layer has 1024 neurons followed by ReLU activation and drop-out. The second layer is the classification layer with softmax activation. Parameter initialization and batch size (=512) was set according to[38] L1-L2 regularization values and number training epoch were the two hyper-parameters optimized over subsets of data before the final cross validation step (number of epochs = 10, L1-L2 regularization, regularization of 10e-4 for both L1 and L2 penalties. The model only using WSIs was used for hyper-parameter optimization. For the computational tumor detector we used the same architecture from.[38]

Feature construction

Mutation status was encoded as a binary variable. Age was encoded as a continuous variable. Other clinical variables were encoded as integers as well as one-hot-encoded variables. The deep learning survival model used the one-hot-encoding outputs. Random forest models trained on clinical and mutation status considered both encodings. Integer encoding resulted in higher AUCs and was used throughout. Clinical and/or mutation status variables were concatenated with the tile level Inception V3 features (see Fig. 1).

Deep learning model training

Deep learning models were trained to predict risk either from WSIs only, or as integrative models that combine WSIs and other data modalities. A key difference of our method compared to others is how we use local information in the training of the integrative models. In previous approaches[21] tile-level image features are averaged within a patient to create patient-level image features. Patient-level image features are then used with patient level clinical variables to train the classification model. In our integrative models, however, each tile is concatenated with the patient clinical features, and training is done across all tiles. We did this because we found that using patient-level image features in the training yielded inferior performance (AUC = 0.68 ± 0.09 for deep learning Cox model and AUC = 0.81 ± 0.08 for image-only model).

(1) Image-only model: for the image-based model, we utilized the Inception V3 architecture that was pre-trained on the ImageNet database as described previously.[38] The cached 2048 global average pooling layer features of InceptionV3 were extracted and written to

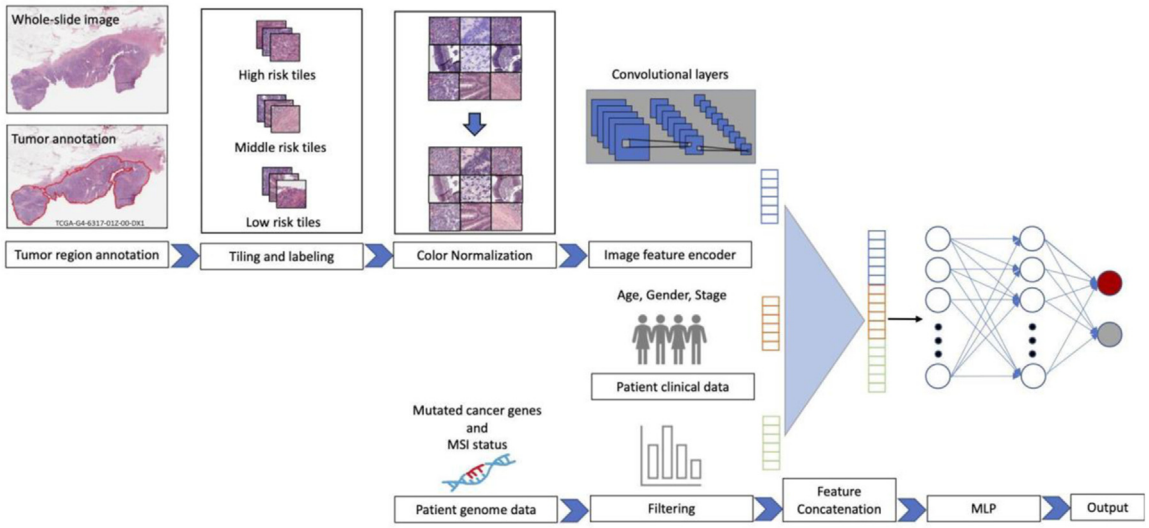

***Fig. 1:* The integrative CNN model.** Tumor regions of WSIs are annotated by expert pathologists. WSIs are tiled, and tiles overlapping with pathologist tumor annotations (>50% overlap) are used for survival analysis. Tiles are color normalized using the Macenko method and passed through an Inception V3 model pre-trained on Image-Net. Tile level CNN features are concatenated with patient level clinical variables and mutation status. These features are fed to a multi-layer perceptron to predict patient risk.

disk for downstream analysis. (2) Integrative models: we designed integrative prognostic models integrating WSIs with different combinations of data modalities. We concatenated tile-level InceptionV3 features with the feature vectors encoding clinical variables and/or mutation signature. The final feature vector was fed to the two-layer MLP. We under-sampled tumor tiles of the majority class to mitigate the effects of class imbalance. To address potential batch effects, we utilized the Macenko method[42] to normalize the stain color across training and independent test data sets. (3) Deep learning Cox model: we trained a Cox proportional

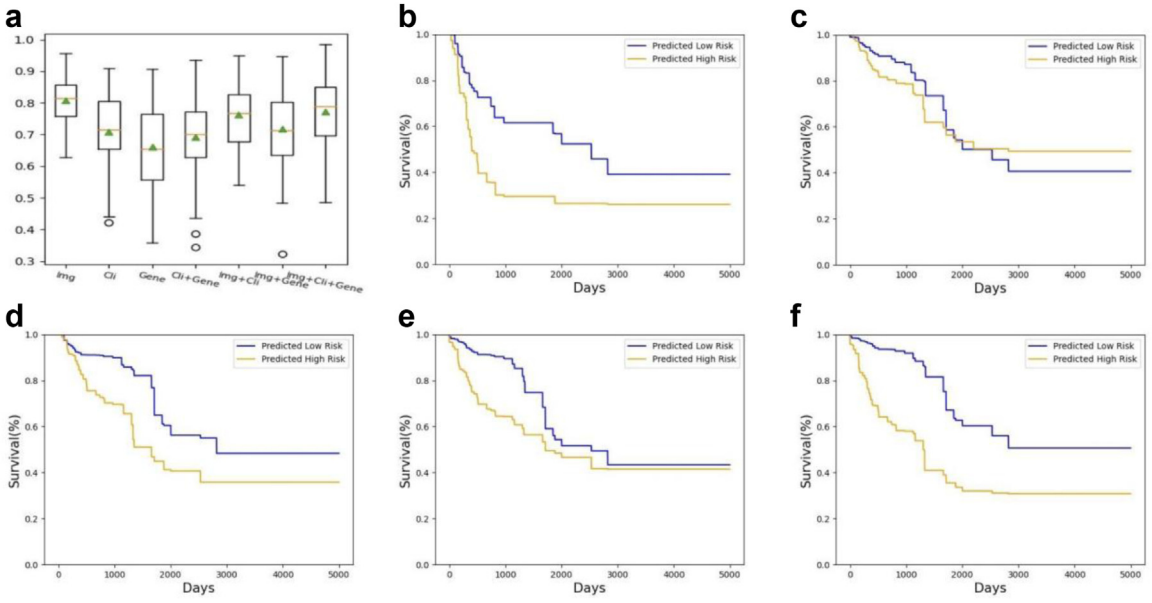

***Fig. 2:* Integrative analysis improves stratification of moderate risk patients.** (a) AUCs for prediction of High/Low risk class by various models. Kaplan–Meier curves of patients from the (b) High/Low and (c) High (N = 38)/Moderate (N = 45)/Low (N = 25) clinical groups, stratified by predicted risk class from the image-only model. Kaplan–Meier plots of High/Moderate/Low patients as stratified by the (d) image & clinical, (e) image & mutation, and (f) image & clinical & mutation models.

hazards model using patient-level image features extracted from Inception V3 transfer learning architecture. Specifically, slide-level image features were generated using the median value of all tumor tiles, and were fed into a Cox proportional hazards regression model implementation of the statsmodels (v0.13.2) package. We randomly split the entire data set into a 70%/30% train/test split, which includes loss-to-follow up and moderate risk patients in addition to HR and LR patients. The median hazard scores of patients in each class (HR or LR) of the train set were averaged, serving as the threshold on the hazard score for predicting class labels in the test set.

Random forest for survival analysis

We used random forests to train 3 separate models stratifying patients based on clinical variables, mutation status, and combined clinical and mutational signatures. Random forests were implemented using scikit-learn version 1.02[43] with default settings except the number of trees was set to 300. We tested several MLPs, but they performed inferior to the random forest model and were less stable. Therefore, the random forest model was used as the final classifier.

Decision tree for integrative survival analysis

We used the decision tree implementation of scikit-learn version 1.02 with default hyperparameters. We used clinical variables and risk scores (probability of being high risk) of image-only models trained on TCGA-COAD and WSU-COAD combined dataset to separate TCGA-READ patients.

Cox proportional model for integrative survival analysis

We used the Cox proportional hazard model implementation of lifelines version 0.27.4 with default hyperparameters. We used clinical variables and risk scores (probability of being high risk) of image-only models trained on TCGA-COAD and WSU-COAD combined dataset to separate TCGA-READ patients.

Model assessment

Model performance was evaluated on patients in the test set. Tile-level risk probabilities were averaged to construct patient level scores. A threshold of 0.5 was used to predict patients as High risk (HR) or Low risk (LR). No threshold optimization was performed. The Kaplan–Meier (KM) curves were plotted using the averaged survivorship at each time point in each cohort. 3 year and 5 year survivorships were used to assess model performance. In addition to KM plots, the mean and standard deviation of the area under the receiver operating characteristic (AUROC) on the test set was used to measure classifier performance in separating HR and LR patients. We statistically compared survivorship between high-risk and low-risk groups, as designated by our model, at both the 3-year and 5-year time points. The null hypothesis for this comparison is that there is no difference in survivorship between the high-risk and low-risk groups at either the 3-year or the 5-year mark. We used the Mann–Whitney U test to assess these groups, giving us a quantifiable measure of the difference in survivorship between these groups at the specified time points.

Relative risk score

The mean and standard deviation of relative-risk at 3 year, 5 year, and median survival points were calculated to compare KM curves. For each test set relative-risk was calculated as follows, where SL and SH denote the survivorship of predicted LR and predicted HR patients, respectively:

$$Relative\ Risk = \frac{1-SH}{1-SL}$$

For the comparison to study,[21] we used their reported SL and SH values to calculate relative risk scores.

C-index

The concordance index (C-index) is calculated using vital status, overall survival time and risk scores of deep learning models. Scikit-survival (version 0.20.0) is used to calculate the C-index.

Feature importance assessment

We used SHAP (SHapley Additive exPlanations)[44–46] to explain the predictions of our trained models. SHAP measures the impact of each feature value on the predictions of a machine learning model for a single input. The average SHAP impact across a dataset quantifies the overall variable importance for a fixed machine learning model. The KernelExplainer function of SHAP was used to measure importance of clinical variables and InceptionV3 features in the integrative deep

TCGA test set

High/Low	5-year LR	5-year HR	5-year p-value
Image-only	0.615	0.296	2.13E-25
Image & clinical	0.645	0.189	1.92E-23
Image & mutation	0.524	0.283	1.37E-21
Image, clinical & mutation	0.658	0.324	1.20E-19
Clinical-only	0.548	0.271	1.03E-24
Mutation-only	0.554	0.313	1.58E-20
Clinical & mutation	0.552	0.266	5.42E-25

WSU test set

High/Low	5-year LR	5-year HR	5-year p-value
Image-only	0.573	0.397	5.05E-13
Clinical-only	0.608	0.486	3.28E-10
Image & clinical	0.535	0.371	1.80E-22

Table 1: **Survivorship of High/Low risk patients in TCGA and WSU test set.**

TCGA test set						
High/Moderate/Low	3-year LR	3-year HR	5-year LR	5-year HR	3-year p-value	5-year p-value
Image-only	0.836	0.786	0.586	0.564	1.96E-05	5.36E-01
Image & clinical	0.899	0.697	0.65	0.449	1.25E-20	1.33E-14
Image & mutation	0.894	0.644	0.591	0.496	7.70E-20	2.21E-03
Image & clinical & mutation	0.919	0.581	0.671	0.355	2.76E-30	6.69E-30
Clinical-only	0.687	0.663	0.418	0.275	3.72E-02	1.54E-21
Mutation-only	0.437	0.459	0.437	0.459	1.00E + 00	1.00E + 00
Clinical & mutation	0.822	0.768	0.63	0.443	6.19E-02	3.75E-20
WSU test set						
High/Moderate/Low	3-year LR	3-year HR	5-year LR	5-year HR	3-year p-value	5-year p-value
Image-only	0.858	0.857	0.662	0.646	5.00E-01	1.84E-01
Clinical-only	0.605	0.555	0.507	0.453	3.25E-03	2.14E-03
Image & clinical	0.747	0.605	0.682	0.545	1.14E-09	2.15E-05

Table 2: **Survivorship of H/M/L patients of TCGA and WSU test set.**

learning model. 50 randomly selected tiles were used to estimate variable importance of the integrative model. The clinical-only model, being a random forest, uses the TreeExplainer function of SHAP to measure the importance of each clinical variable. Beeswarm plots depict the impact of top variables on each patient, and bar plots depict the average SHAP value magnitudes of top variables for each class. For each variable group total importance is defined as the sum of the importance of all variables in the group (e.g. all clinical variables or all InceptionV3 image features).

Image comparison across centers

The WSU validation cohort and TCGA-COAD cohort were compared to assess relative image quality and compatibility. We observed stronger differences between the WSU images and TCGA Indivumed slides than between the WSU images and TCGA-COAD slides from other TCGA centers (see Supplementary Figure S6). This difference was observable despite Macenko normalization. Removing Indivumed slides reduced stain differences and improved generalizability of our TCGA-models to WSU data. For this reason we removed Indivumed slides from the multicenter analysis as well. The outlier behavior of the TCGA Indivumed slides has been reported in prior studies of TCGA WSIs[47] as well.

Quantitative analysis of predictive regions

probability of being high risk of the image-only model using WSU-COAD and TCGA-COAD data was averaged over all splits to obtain slide-level probabilities. For HR patients, tiles with HR probability >0.9 were identified as predictive tiles. For patients with more than 500 predictive tiles, the top 3% tiles with the highest probabilities of being HR were selected to limit the number of predictive tiles of each patient. A similar procedure was carried out for LR patients where LR probabilities were used to determine predictive tiles. HoverNet[48] pretrained on the PanNuke dataset[49] was used to segment and annotate cells (labels: non-label, tumor, inflammatory, connective, necrosis, and no-neo, class probability >0.9) of predictive tiles. Total cell count and ratio of tumor and inflammatory cells for each tile was saved. Area of identified tumor nuclei was computed using scikit-image version 0.20. Average area of tumor nuclei within each tile was saved.

HR and LR patients whose average probability for correct label across training splits was above 0.75 were considered as "easy-to-identify". HR and LR patients who were misclassified even when considered as training data in a split comprised "misclassified" patients. Kolmogorov–Smirnov test (scipy version 1.10.1) of differences between informative tiles of HR and LR patients in easy-to-identify and misclassified groups are reported.

Ethics

The study did not require new ethical approval, as it is encompassed by prior IRB-20-05-2248.

Role of the funding source

This study was supported by the National Cancer Institutes (Grant No. R01CA230031 and P30CA034196). The funders had no roles in study design, data collection and analysis or preparation of the manuscript.

Results

Images are informative of colon adenocarcinoma risk

We first investigated to what extent WSIs alone are predictive of patient risk in TCGA-COAD (Fig. 1). We binned patients as high-risk (HR, OS < 3 years, N = 38), moderate risk (MR, 3 years < OS < 5 years, N = 45), and low risk (LR, OS > 5 years, N = 25) based on overall survival (see Methods). We trained a convolutional

neural network (CNN) to predict risk from WSIs using the HR and LR patients as binary training sets. This yields what we refer to as the image-only model (see Methods). In cross-validation tests, the image-only model is able to distinguish HR and LR patients (AUC = 0.81 ± 0.08, see Fig. 2a), and patients predicted to be HR vs. LR have well-separated survival curves (see Fig. 2b and Table 1, p-value = 2.13e-25, 5 year relative risk = 1.83 ± 0.04). However, the separation between survival curves decreases when MR (3<OS < 5) patients are added to the test set (see Fig. 2c). The image-only model is unable to stratify patients (Supplementary Figure S1a, C-index = 0.49 ± 0.16, 5 year RR = 1.06 ± 0.22, RR-CI = [0.89, 1.19], p-value>0.5) when all patients are included in the test set (i.e. HR, LR, MR, and survivors lost to follow-up within 3 years, see Methods).

Our approach trains on HR and LR patients only, and we observed that this binarization of the training data was important to the predictive success. As a comparison, we trained a Cox proportional hazard model, which is based on survival times from all patients (HR, LR, MR, and lost-to-follow-up, see Methods). Despite this consideration of all patients, the Cox model did poorly in separating survival outcomes in cross-validation tests (Supplementary Figure S1b, 5 year relative risk = 1.04 ± 0.09, p-value = 2.34e-3, C-index = 0.51 ± 0.10). Therefore we used the binarized approach for training in subsequent analyses.

Next, we compared the image-only model to models based on clinical variables and/or mutation statuses (see Methods). The image-only (AUC = 0.81 ± 0.08) model performance was superior to models using only clinical variables (clinical-only model, see Fig. 2a and Supplementary Figure S1c, AUC = 0.71 ± 0.12) or only mutation status (mutation-only model, see Fig. 2a and Supplementary Figure S1d, AUC = 0.66 ± 0.12), as well as to an integrative model combining clinical and mutation information (clinical & mutation model, see Fig. 2a and Supplementary Figure S1e, AUC = 0.69 ± 0.11). These results indicate WSIs are a rich source of information for separating HR and LR patients. Similarly, the clinical-only, mutation-only, and clinical & mutation models were less effective than the image-only model in separating the survival curves when MR patients were included in the test set (see Supplementary Figure S1i and j and Table 2).

Integrative analysis improves stratification of moderate risk patients

We next tested whether an integrative model combining WSIs, clinical variables, and mutation status, hereafter called the image & clinical & mutation model, would improve patient stratification. We found that the fully integrative model performs similarly to the image-only model in separating HR and LR patients (see Fig. 2a), but performs superiorly when MR patients are also included in the test set (compare Fig. 2f and Supplementary Figure S1h). Moreover, even when all patients, including those lost early to follow-up, were included in the test set (see Methods), the integrative model was still successful in separating patients (Supplementary Figure S1k, C-index = 0.69 ± 0.19, 5 year RR = 2.74 ± 0.63, RR-CI = [2.01, 3.12], p-value = 6.29e-19). This finding is similar to the results of[50] on skin cancers, which reported that an integrative model has comparable performance to single-data-type models for distinguishing patients with strong survivorship differences, but provides additional benefit for low confidence cases.

We also investigated integrative models utilizing two data modalities (image & clinical and image & mutation models, see Fig. 2d and e and Supplementary Figure S1e–g) for stratifying patients. The integrative models using only two data types were inferior to the image & clinical & mutation model, though the image & clinical model was superior to the image & mutation model. Both the image & clinical and image & mutation models outperform the clinical & mutation model (see Supplementary Figure S1f, 1g and 1e). As shown in Table 2, the image & clinical & mutation model (RR = 4.01 ± 0.07, p-value = 2.76e-30) provided stronger separation of survival curves than the image-only (3 year RR = 1.29 ± 0.11, p-value = 1.96e-5, 5 year RR = 0.95 ± 0.09, p-value = 5.36e-1) and the clinical-only (3 year RR = 1.02 ± 0.10, p-value = 3.72e-2, 5 year RR = 1.23 ± 0.08, p-value = 1.54e-21) models at the 3-year and 5-year time points.

Prediction heatmaps reveal the morphology associated with risk

We analyzed the prediction heatmaps of several representative TCGA-COAD slides to gain insight about the underlying morphologies that CNNs associate with risk (see Fig. 3, Supplementary Figure S2, Supplementary Figure S3). These heatmaps were generated using the image & clinical & mutation model and show the risk probability for each tile as predicted by the CNN. Pathologist review suggests that nuclear shape, nuclear size pleomorphism, intense cellularity, and abnormal structures are indicative of high risk. Low risk tiles tend to have more regular and small cells.[51–54]

We then quantified the differences between informative tiles of easy-to-identify HR and LR patients (see methods, Supplementary Figure S4). HR tiles were more cellular (HR: 291 ± 125, LR: 269 ± 120, raw p-value = 1.5e-50) and contained more tumor cells (HR: 134 ± 99, LR: 112 ± 86, raw p-value = 6.5e-37). Their tumor cells were also larger in size (HR: 216 ± 70, LR: 201 ± 72, p-value<1e-100). We saw little immune activity across informative tiles of TCGA-COAD patients (immune cell ratio = 0.014 ± 0.03). We further quantified the differences between misclassified HR and LR patients. Misclassified LR patients had informative tiles

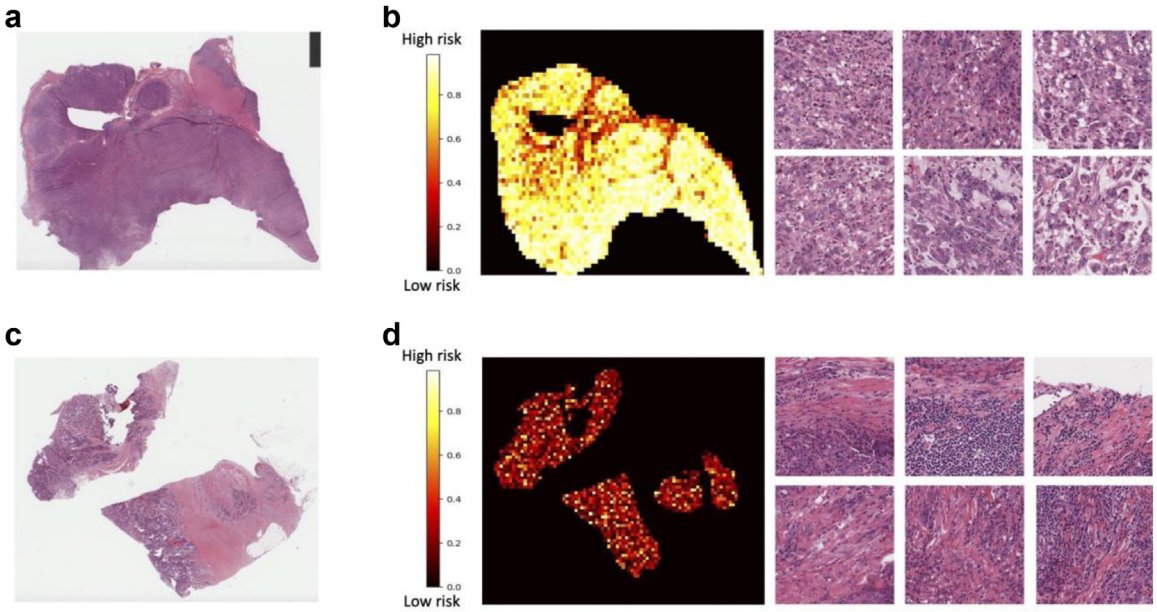

Fig. 3: **Representative H&E slides from TCGA test set and their predicted heatmaps.** WSIs of (a) a high risk patient and (c) a low risk patient. The prediction heatmaps of (b-left) a high risk patient and (d-left) a low risk patient. Example tiles predicted as (b-right) high risk and (d-right) low risk from (a) high risk patient and (c) low risk patient, respectively.

with higher cellularity (HR: 284 ± 88, LR: 313 ± 102, raw p-value = 4.8e-6) and higher number of tumor cells (HR: 110 ± 70, LR: 128 ± 91, p-value = 1.1e-6) than misclassified HR patients.

Pure tumor regions are more informative of risk

Accurate identification of tumor regions within a WSI is a key preliminary step affecting risk classification. To test whether pathologist annotation of tumor regions can be replaced with a computational method, we used pathologist annotations of 228 independent WSIs (see Methods) to build a computational tumor detector. This detector showed high accuracy (Fig. 4a, AUC >92%). Some other works have reported higher AUCs for computationally identifying tumor regions,[55] though this is likely due to variations in pathologist annotation methods. For example, some of our "false positives" are due to the fact that only a subset of tumor regions in a

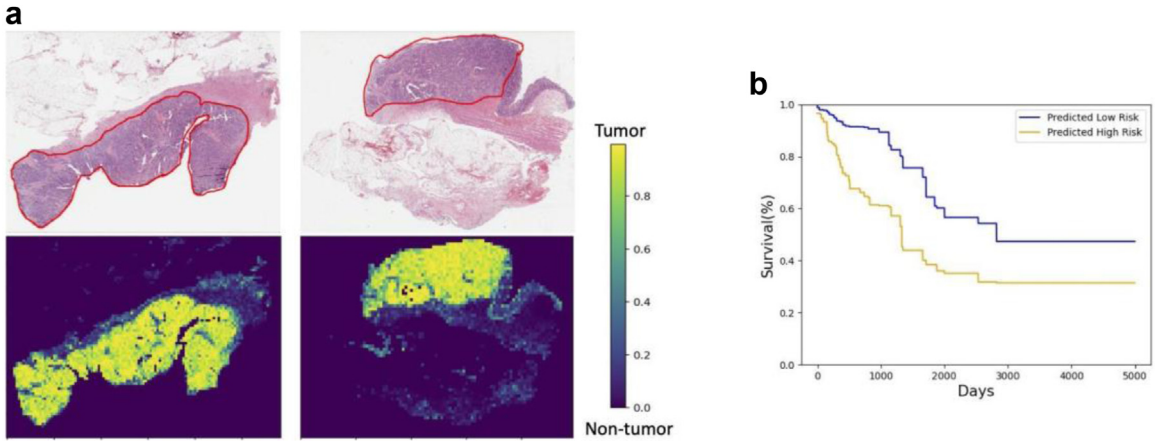

Fig. 4: **Accurate tumor detection improves survival prediction.** (a) Ground truth annotations of tumor regions from pathologists, circled with redlines (top) and tumor prediction heatmaps (bottom). (b) Kaplan–Meier curve of predicted high and low risk patients on the full set of High (N = 38)/Moderate (N = 45)/Low (N = 25) risk patients, determined by applying the image, clinical & mutation model to predicted tumor regions.

slide were selected for annotation (see Supplementary Figure S5). Other discrepancies between our computational predictions and pathologist annotations appear to be related to pathologists' implicit thresholds for tumor annotation. Manual inspection of several "false positive" computational predictions indicate they do contain tumor cells but at lower purity than pathologist annotated regions.

The Kaplan–Meier curve of the image & clinical & mutation model using the computational tumor detector as input is shown in Fig. 4b. As in Fig. 2f, there is a clear separation between the high and low risk curves. However, the separation is lower using the computational tumor detector than using pathologist annotations. The 5 year relative risks when using pathologist annotated and deep learning-predicted tumor regions are 1.83 and 1.65, respectively (p-value <0.05). These results suggest that the integrative model is more effective using only pure tumor regions as input, while computational tumor predictions tend to include low-purity regions that reduce performance.

Validation of TCGA models on Wayne State hospital data

We validated our TCGA-trained models on an independent COAD dataset from Wayne State University (WSU). We collected and annotated tumor regions (N = 123, see Methods), and stratified patients as HR (N = 17), LR (N = 97), or MR (N = 9) similar to the TCGA-COAD cohort. For analyses involving the WSU-COAD cohort, we did not include mutation data as it was not available.

We first considered a test set that included all HR, LR and MR cases together. The image-only model was unable to stratify high and low risk patients for this test set (3 year survival, image-only p-value = 5.0e-01, Fig. 5a). The clinical-only model provided a statistically significant but modest stratification (3 year RR = 1.12 ± 0.33, RR-CI = [0.65, 1.67], p-value = 3.3e-03, Fig. 5b). However, the image & clinical model provided superior separation of the patient cohort (3 year RR = 1.53 ± 0.27, RR-CI = [1.03, 2.21], p-value = 1.14e-09, see Fig. 5c and Table 2), consistent with expectations from the intra-TCGA analysis.

We next considered the simpler problem in which only the HR and LR patients were in the test set. As expected we found better stratification than in the HR/MR/LR case for all models: image-only (3 year RR = 1.41 ± 0.19, RR-CI = [0.78, 2.01], p-value = 5.05e-13), clinical-only (3 year RR = 1.3 ± 0.23, RR-CI = [0.87, 1.74], p-value = 3.28e-10), and image & clinical models (3 year RR = 1.46 ± 0.09, RR-CI = [1.12, 1.82], p-value = 1.80e-22). Notably, the image & clinical model has performance superior to the image-only and clinical-only models (Fig. 5d–f). Interestingly, we observe significant p-values for the image & clinical model in the TCGA data and WSU in Table 2 at both time points. However, for the image-only model, the p-values are significant in the TCGA data at the 3-year time point (2e-5) but not the 5-year time point, and they are not significant at either time point in the WSU data. This suggests that the integrative model is more robust to stain differences, as might be expected from its addition of clinical features.

Our pathologists further evaluated the heatmaps of the image & clinical model in the WSU cohort (see Supplementary Figure S2). These confirmed similar findings to the TCGA test set, i.e. that nuclear shape, nuclear size pleomorphism, intense cellularity, and abnormal structures are associated with high risk.

We then quantified the differences between informative tiles of easy-to-identify HR and LR patients (see Methods, Supplementary Figure S4). HR tiles contained more tumor cells (HR: 126 ± 91, LR: 113 ± 108, raw p-value = 1.6e-98), had higher tumor cell ratios (HR: 0.47 ± 0.23, LR: 0.33 ± 0.28, raw p-value <1e-100), had larger tumor nuclei (HR: 252 ± 74, LR: 183 ± 97, raw p-value <1e-100), contained fewer immune cells (HR: 4 ± 7, LR: 26 ± 55, raw p-value <1e-100), and had lower ratios of immune cells (HR: 0.018 ± 0.03, LR:0.084 ± 0.012, raw p-value<1e-100). We further quantified the differences between misclassified HR and LR patients. Misclassified LR patients had informative tiles with higher cellularity (HR: 230 ± 106, LR: 273 ± 163, raw p-value = 8.4e-28).

Robustness of separating moderate risk patients into high/low risk groups

We combined the Wayne State and TCGA data to more exhaustively investigate how MR patients can be computationally stratified into high and low risk groups. We used the image & clinical model and trained on HR and LR patients, analogous to Fig. 2d. Given the small number of MR patients in the WSU cohort (N = 9), we tested this in two ways: (1) training on WSU and testing on TCGA (N = 45 in MR group), and (2) forming a combined multicenter dataset (TCGA + WSU) and testing/training on subsets.

First, we considered the model trained on WSU patients. We confirmed that the model trained on WSU HR and LR patients is able to effectively stratify a test set made of TCGA HR and LR patients (Supplementary Figure S7). We then tested how the model can stratify TCGA MR patients by risk. The model is able to stratify MR patients into higher risk and lower risk sets (5-year p-value = 0.03), though as expected stratification is not as distinct as for the HR/LR test sets. Second, we trained a model from the combined WSU + TCGA set. As expected, this model was able to stratify a reserved set of HR and LR patients by risk (Supplementary Figure S7). It also was able to separate MR patients into higher risk and lower risk (5-year p-value = 5.63e-14), with a highly

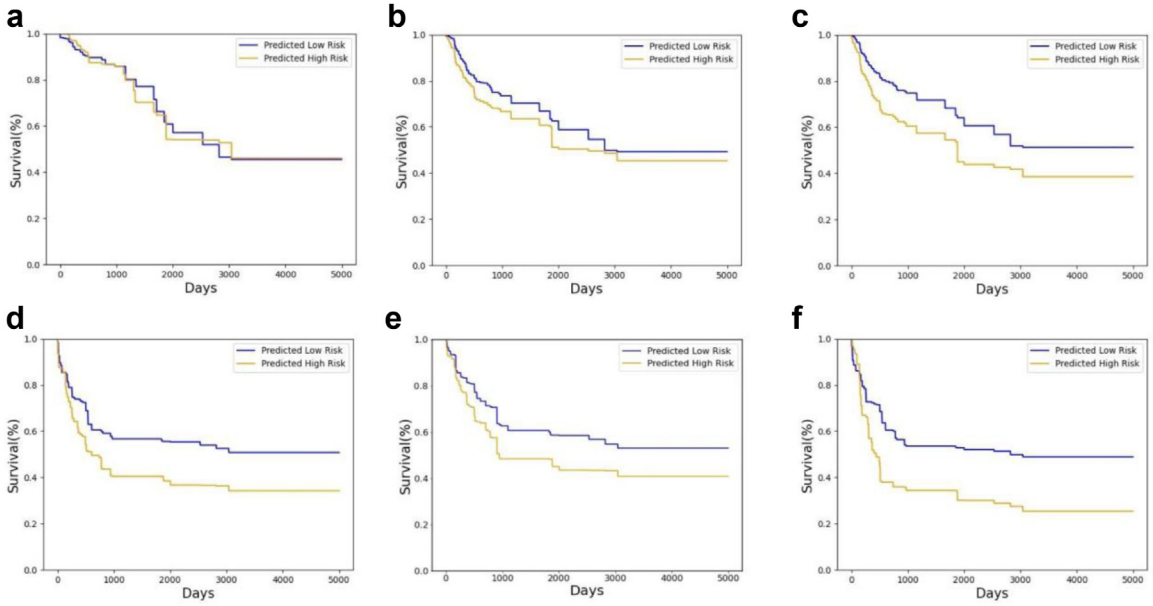

Fig. 5: **External validation using Wayne State data.** a–c. The Kaplan–Meier curves of High/Moderate/Low patients using TCGA-trained classifiers (N = 72) for (a) image-only, (b) clinical-only, and (c) image & clinical models tested on external data (N = 123). d–f. The Kaplan–Meier curves of High/Low patients using TCGA-trained classifiers for (d) image-only, (e) clinical-only, and (f) image & clinical models tested on external data.

significant p-value. Interestingly, both of the MR stratification tests yielded long term survival ratio differences in the two predicted groups. Our results indicate that MR patients share enough similarities with HR and LR patients to improve survival stratification.

Validation of colon adenocarcinoma models on rectal adenocarcinoma

We tested if our COAD models generalize to READ. Our integrative models trained on TCGA-COAD and tested on TCGA-READ separate patients (Fig. 6 and

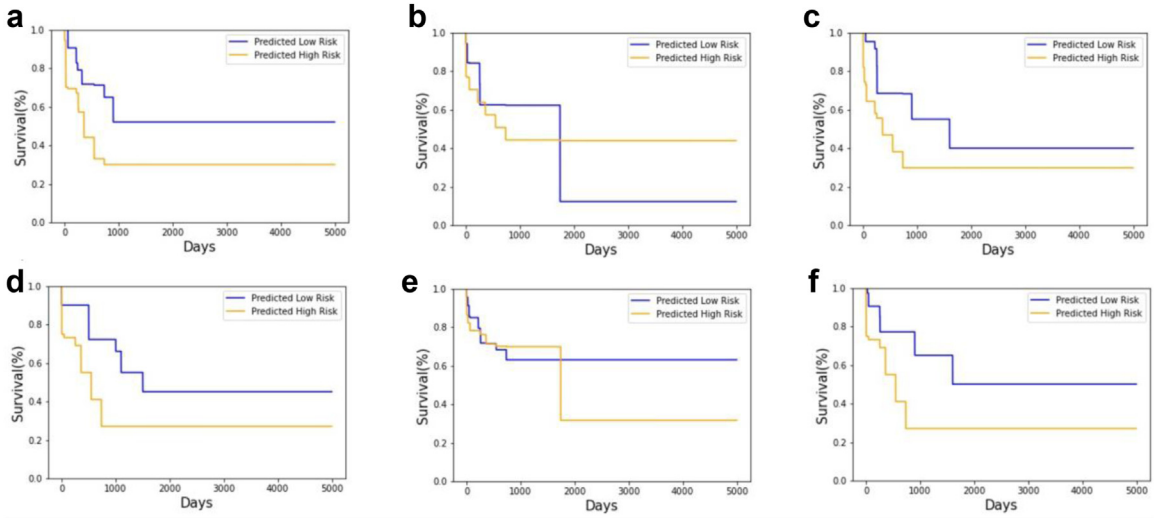

Fig. 6: **Models trained on COAD patients separate READ patients:** Kaplan–Meier plots of (a) image-only model trained on TCGA-COAD separating (a) HR and LR TCGA-READ (N = 30) and (b) HR, MR, and LR TCGA-READ patients (N = 52). Kaplan–Meier plots of (c) image & clinical, and (d) image & clinical & genomic models trained on TCGA-COAD separating (a) HR, MR, and LR TCGA-READ patients (N = 52). Kaplan–Meier plots of (e) image-only and (f) image & clinical models trained on TCGA-COAD and WSU-COAD combined dataset separating HR, MR, and LR TCGA-READ patients (N = 52).

Supplementary Figure S8). Consistent with previous results on COAD patients (sections 3.1 and 3.5), the image-only model was successful in separating HR and LR patients (5 year RR = 1.52 ± 0.31, RR-CI = [1.22, 1.89], p-value = 2.67e-20), but could not stratify patients when MR patients were included in the test set (5 year RR = 0.67 ± 0.22, RR-CI = [0.42, 0.91], p-value>0.5, Fig. 6). Additionally, our image & clinical (5 year RR = 1.18 ± 0.17, RR-CI = [0.86, 1.39], p-value = 1.33e-09) and image & clinical & genomic (5 year RR = 1.24 ± 0.22, RR-CI = [0.96, 1.44], p-value = 1.69e-15) models separate patients (HR, MR, and LR patients combined, Fig. 6).

Our integrative image & clinical model trained on TCGA-COAD and WSU-COAD combined data stratifies READ patients (5 year RR = 1.82 ± 0.13, RR-CI = [1.26, 2.34], p-value = 8.21e-25, Fig. 6). Our image-only model trained on the combined dataset and tested on TCGA-READ separated HR and LR patients (3 year RR = 1.48 ± 0.27, RR-CI = [0.98, 1.81], p-value = 1.41e-23, Supplementary Figure S8), but did not separate patients when MR patients were included (3 year RR = 0.82 ± 0.39, RR-CI = [0.51, 1.14], p-value>0.5, Fig. 6).

Our image only model was able to separate HR and LR patients when restricted to stage 2 (3 year RR = 1.9 ± 0.67, RR-CI = [0.93, 3.47], p-value = 8.2e-39), non metastatic (3 year RR = 1.56 ± 0.59, RR-CI = [0.91, 4.18], p-value = 2.8e-22), lymph node negative (3 year RR = 1.58 ± 0.85, RR-CI = [0.51, 3.2], p-value = 8.2e-5), and metastatic-or-stage 3 (3 year RR = 1.36 ± 0.14, RR-CI = [1.25, 1.5], p-value = 4.3e-34) patients (Supplementary Figure S8). While recent studies suggest deep learning models may be sensitive to the differences between COAD and READ,[56] our COAD models were successful in separating READ patients.

We compared our integrative image & clinical model with integrative models using a decision tree or the Cox proportional hazards model, in which patient-level image-based risk is combined with clinical variables (see Methods, Supplementary Figure S8). Our image & clinical model (3 year RR = 1.82 ± 0.13, RR-CI = [1.43, 2.14], p-value = 8.21e-25) performed superior to the decision tree (3 year survival RR = 1.06 ± 0.44, RR-CI = [0.53, 2.2], p-value = 0.37) and the Cox model (3 year survival RR = 1.37 ± 0.6, RR-CI = [0.42, 1.6], p-value = 1.7e-11).

We then combined WSU-COAD, TCGA-COAD, and TCGA-READ data to further investigate if MR patients can be computationally stratified into high and low risk groups via the multicenter approach of section 3.6. We observed (Supplementary Figure S9) our image & clinical model (3 year survival RR = 2.26 ± 0.56, RR-CI = [1.78, 2.92], p-value = 4.4e-28) is superior to image-only (3 year survival RR = 1.99 ± 0.39, RR-CI = [1.53, 2.10], p-value = 7.2e-08) and clinical-only (3 year survival RR = 1.67 ± 0.55, RR-CI = [1.13, 2.14], p-value = 5.9e-06) models.

Feature importance for colon adenocarcinoma risk

To improve interpretability of our deep learning models, we used SHAP[44–46] to measure the contribution of each clinical or Inception v3 image feature to the model output (see Methods). We describe results for the model trained on TCGA and tested on Wayne State. We found that T stage, M stage, and age are the most impactful features in the integrative model (see Methods, Fig. 7). Although only two InceptionV3 features have comparable importance to these clinical variables, the total importance of InceptionV3 features (11.84) is higher than clinical variables (6.63). This may be explained by the fact that image contributions are spread across 2048 InceptionV3 features, while there are only 6 clinical variables for each patient. Interestingly, although individual clinical variables have high importance, the clinical-only model does not separate patients, suggesting the importance of cross-talk between clinical and image features.

Discussion

While the utility of individual data modalities, such as clinical variables, mutation signatures, and WSIs, for patient stratification has been established,[7–11,21,22,57,58] our study demonstrates that integrative analysis improves patient risk stratification even for the challenging case of patients with intermediate survival times. Our image & clinical model showed more robustness to stain differences than the image-only model (section 3.5). While our image & clinical model successfully separated patients when MR patients were included in the test set, image-only and clinical-only models performed poorly (see sections 3.1 and 3.2). Of potential importance is that cross-talk, i.e. variable–variable interactions, between image features and clinical variables, is informative of patient risk (see Tables 1 and 2; see Figs. 2 and 5). Quantifying the crosstalk between each image feature and each clinical variable is an open research question for non-parametric deep learning models.

Our approach showed comparable performance even though we used a much smaller dataset size (231 COAD patients, 52 READ patients) than other recent studies (>5000 patients,[21] >2800 patients,[22] >1000 patients[59]). For example, in this study,[21] they used more than 5000 cases to predict 5-year disease-specific survival for colorectal cancer. Their survival rates for high and low-risk groups were 53% vs. 86% (validation set 1, 5 year RR = 3.35) and 46% vs. 80% (validation set 2, 5 year RR = 2.7). For our method trained on <300 patients, although our image model did slightly worse (image-only: 5 year RR = 1.83), the integrative model was superior (image & clinical & mutation: 5 year RR = 4.01). Furthermore, in[21] the AUC for predicting high/low risk was 0.70, but our

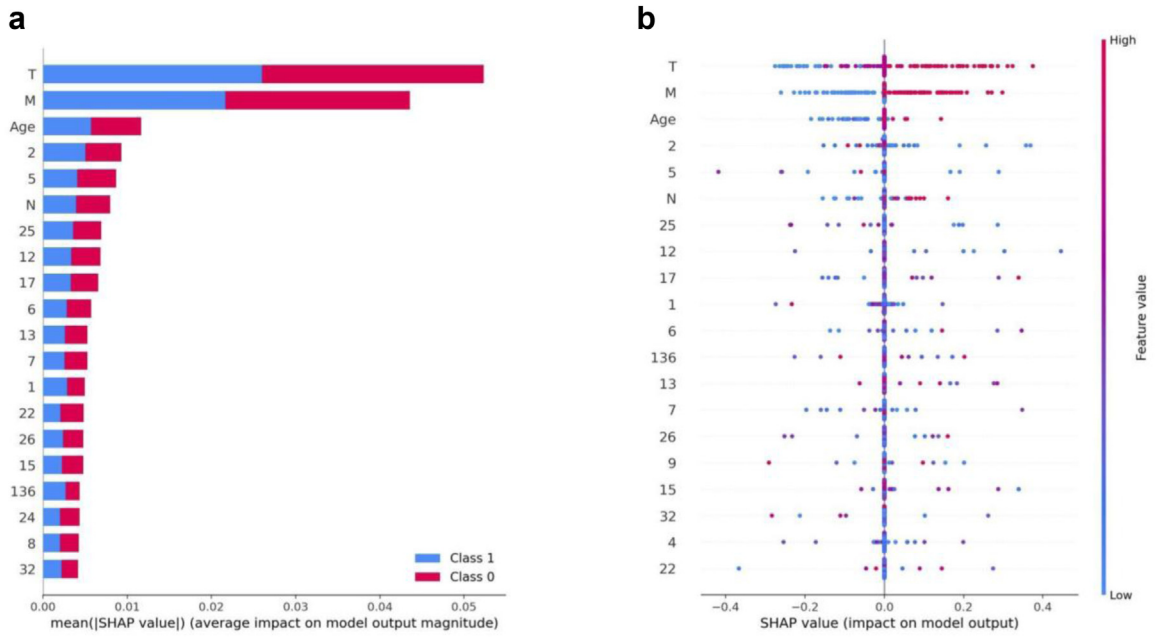

***Fig. 7*: SHAP values of individual features of image & clinical model applied to Wayne State data.** (a) Bar plot of the average SHAP values for top predicted features to illustrate global feature importance in class 1(High risk) and class 0 (Low risk). (b) SHAP values of top features across the Wayne State dataset. The plot sorts features by the sum of SHAP value magnitudes over all samples. The color represents the feature value (red high, blue low).

integrative model achieved an AUC of 0.80. This is of particular interest for translational research as large samples are difficult to obtain for many cancer types. The lower sample size requirements of our models can be attributed to a number of strengths. First, restricting image analysis to regions with high tumor content is crucial for reliable risk assessment (Results and Fig. 4). Second, our training process uses classification-based losses and combines local tile-level with global patient-level information to improve model training. Although overall survival is widely used as the dependent variable in survival analysis, it is not a direct measure of patient underlying risk,[60] as outcomes can be influenced by uncertain factors such as each patient's access to effective treatments. Prior theoretical results have suggested that simplifying continuous data into discrete categories reduces noise in related estimation problems.[61,62] We observed the quantization strategy improves patient stratification (see section 3.1), which is consistent with recent survival analysis studies using gene expression data.[60] Third, restricting to patient subsets with strong survival differences produces a more reliable training dataset. While previous works have assigned a continuous risk score to all patients, e.g. Cox hazard ratio, and identified MR patients as a post-process,[21] our approach of binary classification of MR patients yields clearer results.

The combining of local tile-level image features with patient level information has theoretical advantages as it is a form of context-aware learning.[32] In our analysis (section 3.7), this approach performed superior to those which first combine image features into an initial risk score, and combine patient information and the image-based score as a secondary step.[21,63] Our approach is superior to,[21] where tile-level image features are first combined to a patient level image features, then patient level image and clinical variables are combined afterwards. Particularly, only our model was able to detect the importance of cross-talk between local image features and clinical variables. In,[21] almost all of the signal was due to the image features (73–80%) with a lesser contribution from clinical features (T, N, and grade total: 18%) and no apparent cross-talk.

Prediction heatmaps from our computational model enable identifying regions that are informative of risk, improving model interpretability and discovery of novel prognostic markers. Specifically, our pathologist evaluations of the model predictions resulted in the findings that nuclear shape, nuclear size pleomorphism, intense cellularity, and abnormal structures are associated with higher risk (see Results and Fig. 3). Our predictions also comport with known histopathological risk features. Histopathological tumor grading is used in the College of American Pathologists (CAP) protocol for colon cancer reporting as part of the diagnostic standard template, and has been shown to correlate with patient survival.[51–54] We observed a similar trend in both the TCGA-COAD and the WSU sets during annotation and clinical data collection. Such stratification based on

histology is used in pathology reports as either a 2-tier, 3-tier, or 4-tier classification of tumors from well differentiated to poorly differentiated.[64] Comparable patterns of high and low risk morphology were detected by the deep learning model as shown in Fig. 3 and Supplementary Figure S2.

Our integrative deep learning model has potential to improve clinical decision making. For example, patients with a higher predicted risk of mortality may receive personalized treatment plans with closer follow ups. Patients considered under current standards to be moderate risk may especially benefit, as their outcomes are difficult to predict[65,66] and better distinguishing their risk will be clinically useful.[67,68] Computational risk prediction may also present evidence for improved approval of expensive scans and improved patient counseling. To realize these translational goals, an important future direction will be to further explore the cross-talk between morphological features and clinical variables. Recent studies suggest that individual image deep learning features encode interpretable morphologies[69] and that small clusters of deep learning features encode distinct markers of risk.[21] Cross-talk can be further studied by identifying the distinct morphologies encoded by deep learning features in regions of interest, and then evaluating correlations between deep learning features and clinical variables in each risk group. While the current study establishes the utility of integrative models in stratifying moderate risk patients and reducing sample size requirements, larger multicenter datasets will be valuable to improve risk predictions and especially robustness to stain variations. Future research efforts should be devoted to confirming the applicability and efficacy of this neural network approach using more extensive patient cohorts for which comprehensive data, including matched clinical, imaging, and genomic mutation data, are available. These should be coupled with tests of the empirical robustness of predictions in spite of variations in data acquisition. Such broader studies will ensure the validity of our model across diverse population groups and cancer types.

Contributors

J.H.C. and F.S.A. conceptualized the idea. J.Z. and A.F. developed the pipeline and analyzed the data. J.Z. performed quality control and analyzed the WSU and TCGA data. A.F. analyzed the TCGA data. H.D., F.D., T.A., R.B. and F.S.A. provided pathological evaluations and tumor annotations. F.S.A. provided the WSU data and evaluated model predictions. J.Z., A.F., F.S.A, and J.H.C. prepared the manuscript. J.H.C. and F.S.A. led the project and were responsible for the decision to submit the manuscript. All authors approved the final paper. J.Z., A.F., H.D., F.D., R.B. and F.S.A have verified the underlying data.

Data sharing statement

WSU data is publicly available and can be downloaded from Zenodo (https://zenodo.org/10.5281/zenodo.8163751, https://zenodo.org/10.5281/zenodo.8170095). The code is available at (https://github.com/JieZhou8221/Colorectal-Cancer-Survival-Analysis/). TCGA data is publicly available and can be downloaded from the GDC portal (https://portal.gdc.cancer.gov/).

Declaration of interests

Fahad Shabbir Ahmed is a co-founder of ALGORISMUS, LLC and CDS Dental Material, LLC. The other authors declare no competing interests.

Acknowledgements

J.H.C. acknowledges support from NCI grant R01CA230031 and P30CA034196. A.F. acknowledges support from a JAX Scholar award, Farmington, CT, USA.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104726.

References

1 Weiser MR. AJCC 8th edition: colorectal cancer. *Ann Surg Oncol*. 2018;25(6):1454–1455.
2 Society AC. *Cancer facts & Figures 2021*. Atlanta: Ga: American Cancer Society; 2021.
3 Petrelli F, Tomasello G, Borgonovo K, et al. Prognostic survival associated with left-sided vs right-sided colon cancer: a systematic review and meta-analysis. *JAMA Oncol*. 2017;3(2):211–219.
4 Fuchs TL, Sioson L, Sheen A, et al. Assessment of tumor-infiltrating lymphocytes using International TILs Working Group (ITWG) system is a strong predictor of overall survival in colorectal carcinoma: a study of 1034 patients. *Am J Surg Pathol*. 2020;44(4):536–544.
5 Idos GE, Kwok J, Bonthala N, Kysh L, Gruber SB, Qu C. The prognostic implications of tumor infiltrating lymphocytes in colorectal cancer: a systematic review and meta-analysis. *Sci Rep*. 2020;10(1):1–14.
6 Zhao Y, Ge X, He J, et al. The prognostic value of tumor-infiltrating lymphocytes in colorectal cancer differs by anatomical subsite: a systematic review and meta-analysis. *World J Surg Oncol*. 2019;17(1):1–11.
7 Ahluwalia P, Kolhe R, Gahlay GK. The clinical relevance of gene expression based prognostic signatures in colorectal cancer. *Biochim Biophys Acta Rev Cancer*. 2021;1875(2):188513.
8 Chand M, Keller DS, Mirnezami R, et al. Novel biomarkers for patient stratification in colorectal cancer: a review of definitions, emerging concepts, and data. *World J Gastrointest Oncol*. 2018;10(7):145.
9 Koncina E, Haan S, Rauh S, Letellier E. Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges. *Cancers*. 2020;12(2):319.
10 Hull MA, Rees CJ, Sharp L, Koo S. A risk-stratified approach to colorectal cancer prevention and diagnosis. *Nat Rev Gastroenterol Hepatol*. 2020;17(12):773–780.
11 Thomas C, Mandrik O, Saunders CL, et al. The costs and benefits of risk stratification for colorectal cancer screening based on phenotypic and genetic risk: a health economic analysis. *Cancer Prev Res*. 2021;14(8):811–822.
12 Mármol I, Sánchez-de-Diego C, Pradilla Dieste A, Cerrada E, Rodriguez Yoldi MJ. Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int J Mol Sci*. 2017;18(1):197.
13 Biller LH, Schrag D. Diagnosis and treatment of metastatic colorectal cancer: a review. *JAMA*. 2021;325(7):669–685.
14 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*. 2021;124(4):686–696.
15 Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–1567.
16 Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054–1056.
17 Farahmand S, Fernandez AI, Ahmed FS, et al. Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. *Mod Pathol*. 2022;35(1):44–51.
18 Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health*. 2021;3(12):e763–e772.

19 Naik N, Madani A, Esteva A, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nat Commun*. 2020;11(1):1–8.
20 Wulczyn E, Steiner DF, Xu Z, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One*. 2020;15(6):e0233678.
21 Wulczyn E, Steiner DF, Moran M, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit Med*. 2021;4(1):1–13.
22 Skrede O-J, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet*. 2020;395(10221):350–360.
23 Alsinglawi B, Alshari O, Alorjani M, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci Rep*. 2022;12(1):1–10.
24 Serebriiskii IG, Connelly C, Frampton G, et al. Comprehensive characterization of RAS mutations in colon and rectal cancers in old and young patients. *Nat Commun*. 2019;10(1):1–12.
25 Gonzalez-Donquiles C, Alonso-Molero J, Fernandez-Villa T, Vilorio-Marqués L, Molina A, Martín V. The NRF2 transcription factor plays a dual role in colorectal cancer: a systematic review. *PLoS One*. 2017;12(5):e0177549.
26 Li X-L, Zhou J, Chen Z-R, Chng W-J. P53 mutations in colorectal cancer-molecular pathogenesis and pharmacological reactivation. *World J Gastroenterol*. 2015;21(1):84.
27 Schatoff EM, Leach BI, Dow LE. Wnt signaling and colorectal cancer. *Curr Colorectal Cancer Rep*. 2017;13(2):101–110.
28 He W-L, Weng X-T, Wang J-L, et al. Association between c-Myc and colorectal cancer prognosis: a meta-analysis. *Front Physiol*. 2018;9:1549.
29 Liang K, Zhou G, Zhang Q, Li J, Zhang C. Expression of hippo pathway in colorectal cancer. *Saudi J Gastroenterol*. 2014;20(3):188.
30 Tyagi A, Sharma AK, Damodaran C. A review on notch signaling and colorectal cancer. *Cells*. 2020;9(6):1549.
31 Slattery ML, Mullany LE, Sakoda LC, Wolff RK, Samowitz WS, Herrick JS. The MAPK-signaling pathway in colorectal cancer: dysregulated genes and their association with micrornas. *Cancer Inform*. 2018;17:1176935118766522.
32 Shaban M, Awan R, Fraz MM, et al. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Trans Med Imaging*. 2020;39(7):2395–2405.
33 Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. 2015;19(1A):A68.
34 Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2:401–404.
35 Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6(269):pl1–pl.
36 Biosystems L. *Aperio ImageScope-Pathology slide viewing software*. Leica Biosystems; 2015. https://www.leicabiosystems.com/sites/default/files/media_document-file/2021-02/MAN-0001-Rev-P_12.3.pdf.
37 Bradski G. The openCV library. *Dr Dobb's J Softw Tools Prof Program*. 2000;25(11):120–123.
38 Noorbakhsh J, Farahmand S, Namburi S, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun*. 2020;11(1):1–14.
39 Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology*. 2010;138(6):2073–2087.e3.
40 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
41 Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE; 2009.
42 Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE; 2009.
43 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
44 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.
45 Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
46 Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749–760.
47 Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12(1):1–13.
48 Graham S, Vu QD, Raza SEA, et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal*. 2019;58:101563.
49 Gamper J, Alemi Koohbanani N, Benet K, Khuram A, Rajpoot N. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: *Digital pathology: 15th European congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*. Springer International Publishing; 2019:11–19.
50 Höhn J, Krieghoff-Henning E, Jutzi TB, 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.
51 Chandler I, Houlston R. Interobserver agreement in grading of colorectal cancers—findings from a nationwide web-based survey of histopathologists. *Histopathology*. 2008;52(4):494–499.
52 Cho YB, Chun H-K, Yun HR, Kim HC, Yun SH, Lee WY. Histological grade predicts survival time associated with recurrence after resection for colorectal cancer. *Hepatogastroenterology*. 2009;56(94-95):1335–1340.
53 Derwinger K, Kodeda K, Bexe-Lindskog E, Taflin H. Tumour differentiation grade is associated with TNM staging and the risk of node metastasis in colorectal cancer. *Acta Oncol*. 2010;49(1):57–62.
54 Barresi V, Bonetti LR, Ieni A, Domati F, Tuccari G. Prognostic significance of grading based on the counting of poorly differentiated clusters in colorectal mucinous adenocarcinoma. *Hum Pathol*. 2015;46(11):1722–1729.
55 Thakur N, Yoon H, Chong Y. Current trends of artificial intelligence for colorectal cancer pathology image analysis: a systematic review. *Cancers*. 2020;12(7):1884.
56 Chang X, Wang J, Zhang G, et al. Predicting colorectal cancer microsatellite instability with a self-attention-enabled convolutional neural network. *Cell Rep Med*. 2023;4(2):100914.
57 Ahmed FS, Ali L, Joseph BA, Ikram A, Mustafa RU, Bukhari SAC. A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit. *J Trauma Acute Care Surg*. 2020;89(4):736–742.
58 Ahmad FS, Ali L, Khattak HA, et al. A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs). *J Ambient Intell Hum Comput*. 2021;12(3):3283–3293.
59 Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med*. 2019;16(1):e1002730.
60 Foroughi Pour A, Loveless I, Rempala G, Pietrzak M. Binary classification for failure risk assessment. *Methods Mol Biol*. 2021;2194:77–105.
61 Argyrios Z, Boyd S, Candes E. Compressed sensing with quantized measurements. *IEEE Signal Process Lett*. 2009;17:149–152.
62 Laska JN, Baraniuk RG. Regime change: bit-depth versus measurement-rate in compressive sensing. *IEEE Trans Signal Process*. 2012;60(7):3496–3505.
63 Kleppe A, Skrede OJ, De Raedt S, et al. A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol*. 2022;23(9):1221–1232.
64 Edge SB, Edge SB, Cancer AJCo. *AJCC cancer staging manual*. 8th ed. New York, NY: Springer; 2017.
65 Lansdorp-Vogelaar I, Meester R, de Jonge L, Buron A, Haug U, Senore C. Risk-stratified strategies in population screening for colorectal cancer. *Int J Cancer*. 2022;150(3):397–405.
66 Stolzenbach LF, Nocera L, Collà-Ruvolo C, et al. Improving the stratification of patients with intermediate-risk prostate cancer. *Clin Genitourin Cancer*. 2021;19(2):e120–e128.
67 Li X, Jonnagaddala J, Yang S, Zhang H, Xu XS. A retrospective analysis using deep-learning models for prediction of survival outcome and benefit of adjuvant chemotherapy in stage II/III colorectal cancer. *J Cancer Res Clin Oncol*. 2022;148:1–9.
68 Soria F, D'Andrea D, Abufaraj M, et al. Stratification of intermediate-risk non–muscle-invasive bladder cancer patients: implications for adjuvant therapies. *Eur Urol Focus*. 2021;7(3):566–573.
69 Foroughi pour A, White BS, Park J, Sheridan TB, Chuang JH. Deep learning features encode interpretable morphologies within histological images. *Sci Rep*. 2022;12(1):1–12.