Journal of Pathology
J Pathol May 2021; **254:** 70–79
Published online 16 March 2021 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/path.5638

ORIGINAL PAPER

Deep learning detects genetic alterations in cancer histology generated by adversarial networks

Jeremias Krause[1], Heike I Grabsch[2,3], Matthias Kloor[4], Michael Jendrusch[4], Amelie Echle[1], Roman David Buelow[5], Peter Boor[5], Tom Luedde[6], Titus J Brinker[7], Christian Trautwein[1], Alexander T Pearson[8], Philip Quirke[3], Josien Jenniskens[9], Kelly Offermans[9], Piet A van den Brandt[9] and Jakob Nikolas Kather[1,3,10]*

[1] *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*
[2] *Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands*
[3] *Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK*
[4] *Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany*
[5] *Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany*
[6] *Department of Gastroenterology, Hepatology and Infectious Diseases, University Hospital Duesseldorf, Duesseldorf, Germany*
[7] *Digital Biomarkers for Oncology Group (DBO), National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[8] *Department of Medicine, University of Chicago Medicine, Chicago, IL, USA*
[9] *Department of Epidemiology, Maastricht University Medical Center+, Maastricht, The Netherlands*
[10] *Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany*

Correspondence to: JN Kather, Department of Medicine III, RWTH University Hospital Aachen, Pauwelsstraße 30, 52074 Aachen, Germany. E-mail: jkather@ukaachen.de

Abstract

Deep learning can detect microsatellite instability (MSI) from routine histology images in colorectal cancer (CRC). However, ethical and legal barriers impede sharing of images and genetic data, hampering development of new algorithms for detection of MSI and other biomarkers. We hypothesized that histology images synthesized by conditional generative adversarial networks (CGANs) retain information about genetic alterations. To test this, we developed a 'histology CGAN' which was trained on 256 patients (training cohort 1) and 1457 patients (training cohort 2). The CGAN synthesized 10 000 synthetic MSI and non-MSI images which contained a range of tissue types and were deemed realistic by trained observers in a blinded study. Subsequently, we trained a deep learning detector of MSI on real or synthetic images and evaluated the performance of MSI detection in a held-out set of 142 patients. When trained on real images from training cohort 1, this system achieved an area under the receiver operating curve (AUROC) of 0.742 [0.681, 0.854]. Training on the larger cohort 2 only marginally improved the AUROC to 0.757 [0.707, 0.869]. Training on purely synthetic data resulted in an AUROC of 0.743 [0.658, 0.801]. Training on both real and synthetic data further increased AUROC to 0.777 [0.715, 0.821]. We conclude that synthetic histology images retain information reflecting underlying genetic alterations in colorectal cancer. Using synthetic instead of real images to train deep learning systems yields non-inferior classifiers. This approach can be used to create large shareable data sets or to augment small data sets with rare molecular features.

Keywords: digital pathology; microsatellite instability; deep learning; generative adversarial network; generative model; colorectal cancer; artificial intelligence; machine learning

Received 30 November 2020; Revised 10 January 2021; Accepted 5 February 2021

No conflicts of interest were declared.

Introduction

Colorectal cancer (CRC) is among the most prevalent cancer types worldwide and is a cause of many cancer-associated deaths [1]. To diagnose this disease, a tumor tissue sample is usually taken endoscopically, formalin-fixed, and paraffin-embedded. A tissue section is cut from the paraffin block by an experienced technician using a microtome; the tissue section is mounted onto a glass slide and stained with hematoxylin and eosin (H&E) before it can be examined microscopically. H&E-stained histology slides are available for almost any cancer patient [2] and for many cancer types, because they are required to make a diagnosis.

To select an appropriate treatment for CRC patients, further molecular tests may be needed. Currently, one of the most clinically relevant genetic alterations in CRC is microsatellite instability (MSI), a genetic abnormality

which affects approximately 10–15% of CRC patients [3]. MSI is a predictive biomarker for response to cancer immunotherapy in metastatic CRC [3] and is associated with Lynch syndrome. Therefore, some national guidelines recommend testing of all CRC patients for MSI [4,5], but in routine clinical practice this is not universally implemented, due to additional costs and tissue limitations [6].

Recently, several studies showed that deep learning-based analysis of digital slides can detect molecular alterations from routine histology slides [7–9], including MSI in colorectal cancer [10–12]. Ultimately, deep learning-based image analysis could pre-select patients for subsequent molecular testing or could be used as a definitive test [13]. The performance of deep learning systems in histology image analysis is dependent on the availability of large training data sets [11,14]. Using more images to train deep neural networks increased the performance of tumor detection [14] and molecular subtyping [11].

However, unlike in non-medical fields of deep learning research, histology image data suitable for deep learning-based analysis are not publicly available in abundance [15]. Histology images are only useful when combined with additional patient-specific metadata related to outcome or molecular alterations. Publicly sharing histology images with these metadata is problematic because these data are linked to an individual patient. Although histology images themselves are not considered protected health information (PHI) by the Health Insurance Portability and Accountability Act (HIPAA) in the United States of America, public sharing of patient-related data needs to comply with legal and ethical regulations in most countries. Therefore, in practice, publicly available matched histology and genetic data are very scarce. An exception is 'The Cancer Genome Atlas' (TCGA), a large-scale initiative which provides matched histology and genomics data for cancer patients [16]. However, data from this archive are limited and similarly molecularly comprehensive data are currently not available from any other resource. This lack of data is prohibitive for many researchers and may impair the development of new computer-based methods for detecting molecular alterations in cancer. In order to address this shortage of publicly available histology images, recent studies have used 'generative models' to generate synthetic histology images [17]. In particular, generative adversarial networks (GANs) seem to be able to synthesize histology images that are indistinguishable from real histology images for experts [17–20]. Unlike real histology images, synthetic images are not linked to a particular patient. Thus, synthetic images could be publicly shared with fewer legal or ethical difficulties [21]. However, it is unclear whether and how synthetic histology images retain information about molecular alterations in cancer. Also, it is unclear whether synthetic images can be used equally well to train deep learning-based predictors of molecular alterations. In the current study, we aimed to develop a conditional generative adversarial network capable of synthesizing histology images with associated genetic information. We investigated if these synthetic images could replace real images to train deep learning detectors of MSI status.

Materials and methods

Ethics statement

All experiments were conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects by the Council for International Organizations of Medical Sciences (CIOMS). Histology images with matched MSI status were derived from The Cancer Genome Atlas (https://portal.gdc.cancer.gov) [22] and the Netherlands Cohort Study (NLCS) study [23], as described previously [11]. The NLCS study was cleared by the institutional ethics board of the respective institutions, as described previously [23].

Image selection and preprocessing

We retrieved whole-slide digital histology images from the TCGA database as described previously [10]. For 398 colorectal cancer patients, H&E-stained slides from formalin-fixed, paraffin-embedded tissue with matched MSI status were available. These patients were randomly assigned to a training set ('training cohort 1') and a test set with a 2:1 split ensuring similar MSI incidence in each set. As in previous studies [8], the train–test split was performed at the patient level as opposed to at the tiles level. Thus, we ensured that the train and test set never contained tiles from the same patient. To increase the sample size of the training cohort in subsequent experiments, we acquired images from another cohort ('training cohort 2'): We randomly selected 1457 histopathological whole-slide images of CRC (152 MSI and 1305 non-MSI, one image per patient) from the NLCS cohort as described before [11]. We used this second set of images to create an additional training set. In all whole-slide images, tumor regions were manually outlined by a trained observer supervised by an expert histopathologist, as described previously [10]. Tissue within the tumor region was tessellated into tiles of 512×512 pixels corresponding to 256×256 μm and a magnification of 20×. All image tiles were color-normalized using the Macenko method [24].

Development of generative models

Based on the classical generative adversarial network (GAN) architecture [25] and previously described conditional GAN (CGAN) architectures [18], we developed a specialized CGAN to generate red, green, blue (RGB) images of 512×512 pixels. The CGAN consisted of a generator and a discriminator network (Figure 1A,B). It was trained for 50 000 iterations to generate synthetic images of microsatellite stable (MSS) and microsatellite instable (MSI) CRC (Figure 1C,D). CGANs were trained

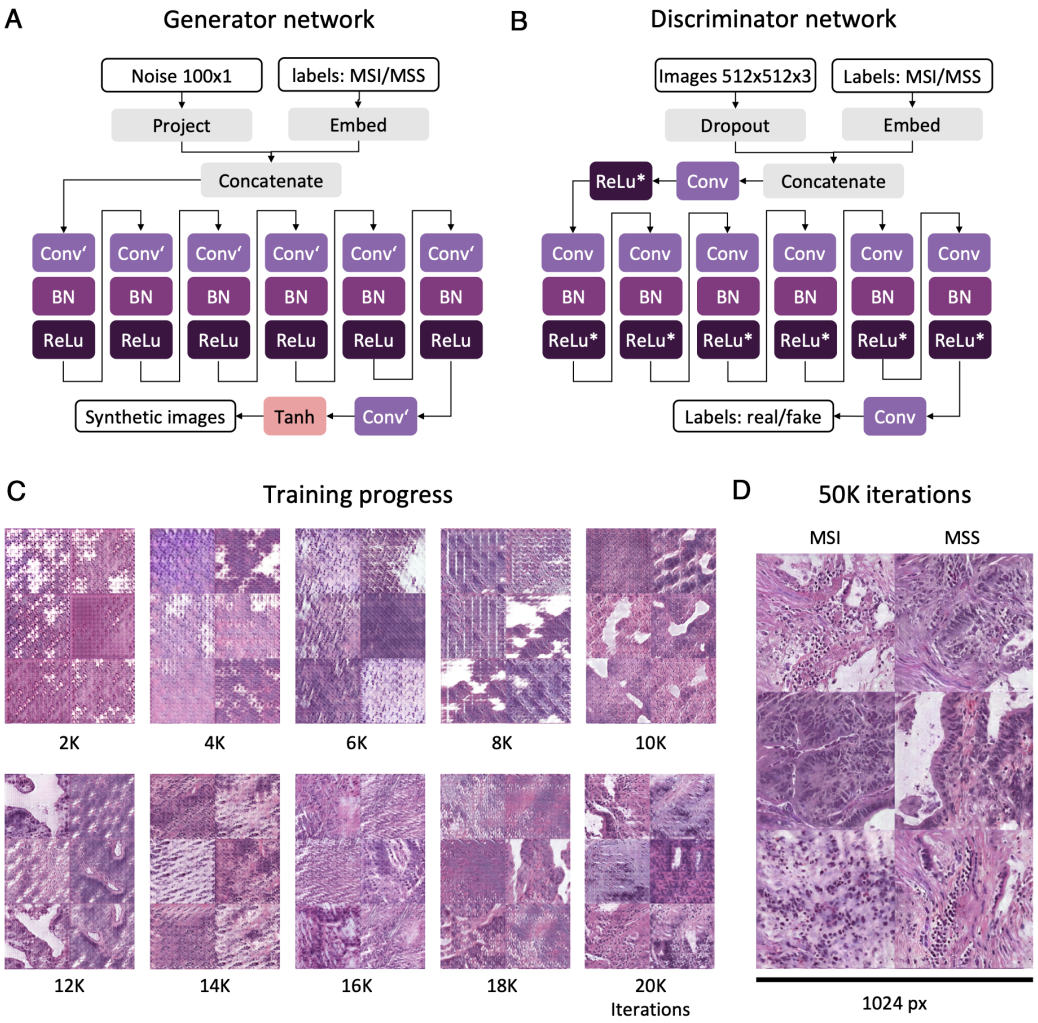

Figure 1. A conditional generative adversarial network (CGAN) for histology images with molecular labels. (A) Overview of the generator network for generation of synthetic histology image patches with 512 × 512 × 3 pixels. MSI, microsatellite instable; MSS, microsatellite stable; Conv', transposed convolution 2D layer; BN, batch normalization layer; ReLu, rectified linear unit layer. (B) Overview of the discriminator network for classifying images as real or fake (synthetic). Conv, convolution 2D layer; ReLu*, leaky rectified linear unit layer. (C) Progress of synthetic images from 2000 (2K) to 20 000 (20K) epochs. (D) Final output of the generator network after 50 000 (50K) epochs.

using a mini batch size of 32 and a label flip factor of 0.125. The CGAN was found to converge smoothly to realistic images without noticeable artifacts such as mode collapse. After the development phase, the CGAN was deployed to generate synthetic images of MSI and MSS colorectal cancer histology images. Two sets of synthetic image tiles were generated (supplementary material, Tables S1 and S2): first, a set of 10 000 images per class (SYNTH-CRC-10K), matching the size of the real training set (TCGA-CRC-TRAIN). Then a larger synthetic set of 75 000 image patches per class was generated (SYNTH-CRC-75K), roughly matching the size of other state-of-the art training sets for MSI detection [11]. Finally, images from SYNTH-CRC-75K were used to augment the real training set TCGA-CRC-TRAIN, yielding a large hybrid training set (MIXED-CRC-75K). All images from the TCGA archive which were used in this study are publicly available (supplementary material, Table S2). In addition, a CGAN was trained on images from $N = 1457$ patients in the 'training cohort 2' (NLCS), generating another synthetic image set (SYNTH-MULTI-10K). The flow of all samples is shown in supplementary material, Figure S1.

Classification network models

After generating synthetic images, a deep learning classifier was trained on each training set (supplementary material, Tables S1 and S3) and evaluated on a set of image tiles generated from held-out patients (TCGA-CRC-TEST). The architecture and hyperparameters used for this classifier network have been developed and optimized previously [8]. In brief, a modified ShuffleNet [26] was trained for four epochs. For data augmentation and to achieve rotational invariance, random horizontal and vertical flips were used in all experiments. To account for color variations in the training set, all tiles were color-normalized using the Macenko method before training [24]. No additional color-based data augmentation steps were used as previous work showed that this does not add a large benefit [27]. To limit overfitting, 10% of the training set were set aside and used as an

internal validation set, which was used to stop training as soon as the validation performance plateaued. The performance of all classifier networks was evaluated at patient level by pooling tile-level predictions as described previously [10]. Patient-level performance was evaluated by the area under the receiver operating curve (AUROC) with 10× bootstrapped pointwise confidence bounds.

Comparison of real and synthetic image tiles

To assess whether synthetic images are deemed 'realistic' by human observers, we performed a blinded experiment. We randomly selected 50 real and 50 synthetic image tiles with balanced MSI status. The images were presented to five observers who were asked to classify an individual image as either 'real' or 'synthetic.' One observer was a pathologist; the other four observers were non-pathologists with experience in reviewing colorectal cancer digital slides. The primary endpoint was accuracy, with 1.0 corresponding to perfect discrimination and 0.5 corresponding to perfect confusion. The experiment was first performed with synthetic images from the 'SYNTH-CRC-10K' image set and was repeated with images from the 'SYNTH-MULTI-10K' set (supplementary material, Table S1 and Figure S1).

To further quantify the similarity between real and synthetic image tiles, we extracted a feature vector for each image tile and used t-stochastic neighbor embedding (t-SNE) [28] to visualize clusters among the tiles. The feature vector was obtained from the last fully connected layer (penultimate layer activations) from the network which was trained on real images. The aim of this experiment was to find out whether real and synthetic images cluster in distinct groups and are thus distinguishable. Clustering was attempted using a range of commonly used values for t-SNE parameters' exaggeration and perplexity in Matlab R2020a. The respective documentation defines perplexity as 'the effective number of local neighbors of each point' and exaggeration as 'size of natural clusters in data'.

Statistical analysis

All statistical analyses were performed using Matlab R2020a (MathWorks, Natick, MA, USA). Classifier performance is always reported on a patient level by pooling tile-level predictions in a majority vote. The main statistical endpoint was the patient-level area under the receiver operating curve (AUROC), with the upper and lower bound achieved in a 10× bootstrapped experiment reported as confidence bounds as described in https://www.mathworks.com/help/stats/perfcurve.html.

Implementation and code availability

All experiments were implemented in Matlab R2020a (MathWorks) and were run on computer workstations with two NVidia Titan RTX graphics processing units. All source codes for CGANs are available at https://github.com/jnkather/histoGAN, translated to Python with PyTorch at https://github.com/mjendrusch/pytorch-histogan. All source codes for training and evaluating classifier models are available at https://github.com/jnkather/DeepHistology. The code for the observer study is available at https://github.com/JeremiasKrause/Histoquiz. Hyperparameters for CGAN training and a step-by-step explanation to reproduce the experiments are available in supplementary material, Tables S4 and S5, respectively.

Results

Conditional GANs generate realistic image patches with multiple tissue types

We trained conditional generative adversarial networks (CGANs) to generate synthetic histological H&E images of colorectal cancer (CRC), encoding microsatellite instability (MSI) status as a 'conditional' variable into the synthetic images. Comparing real and synthetic image tiles of MSI and microsatellite stable (MSS) tumors, we found that synthetic images were realistic and contained different types of tissue as expected (Figure 2A–D). In particular, synthetic images contained tumor epithelium, desmoplastic stroma, inflammatory cells, and mucus, without being explicitly trained to generate these types of tissue (Figure 2B,D). To quantify whether synthetic images are realistic to observers, we performed a study with five participants (Figure 3). In the first synthetic image set (SYNTH-CRC-10K; supplementary material, Table S1), the pathologist could distinguish real from synthetic images with an accuracy of 84%, while the non-pathologists achieved only an average accuracy of 65%. The participants reported that they mostly identified the synthetic images based on artifacts such as the presence of squares in whitespace background. Next, we investigated whether training a CGAN on a larger patient cohort can reduce these artifacts and thus yield more realistic synthetic images derived from $n = 1457$ patients. In this second user study, only the pathologist was able to detect synthetic images, with an accuracy reduced to 77%. All the other observers were unable to reliably detect synthetic images in this study and reached an average accuracy of 52% (individual accuracy was 47%, 58%, 55% and 48%, respectively) (Figure 3). Aiming at further quantifying possible differences between real and synthetic images, we assessed the local sensitivity of deep learning classifiers to regions in real and synthetic images. As shown in Figure 4A–D and supplementary material, Figure S2A–D, local sensitivity as measured by occlusion maps was highest in regions with tumor epithelium and in interface regions (epithelium/background, epithelium/mucus). In terms of local sensitivity, no obvious differences between real and synthetic images were detected. In addition, we visualized the clustering of real and synthetic tiles in a feature space. Dimensionality reduction with t-SNE revealed that these images were completely mixed in the feature space (Figure 4E). Taken together, this provides evidence that real and synthetic images are similar.

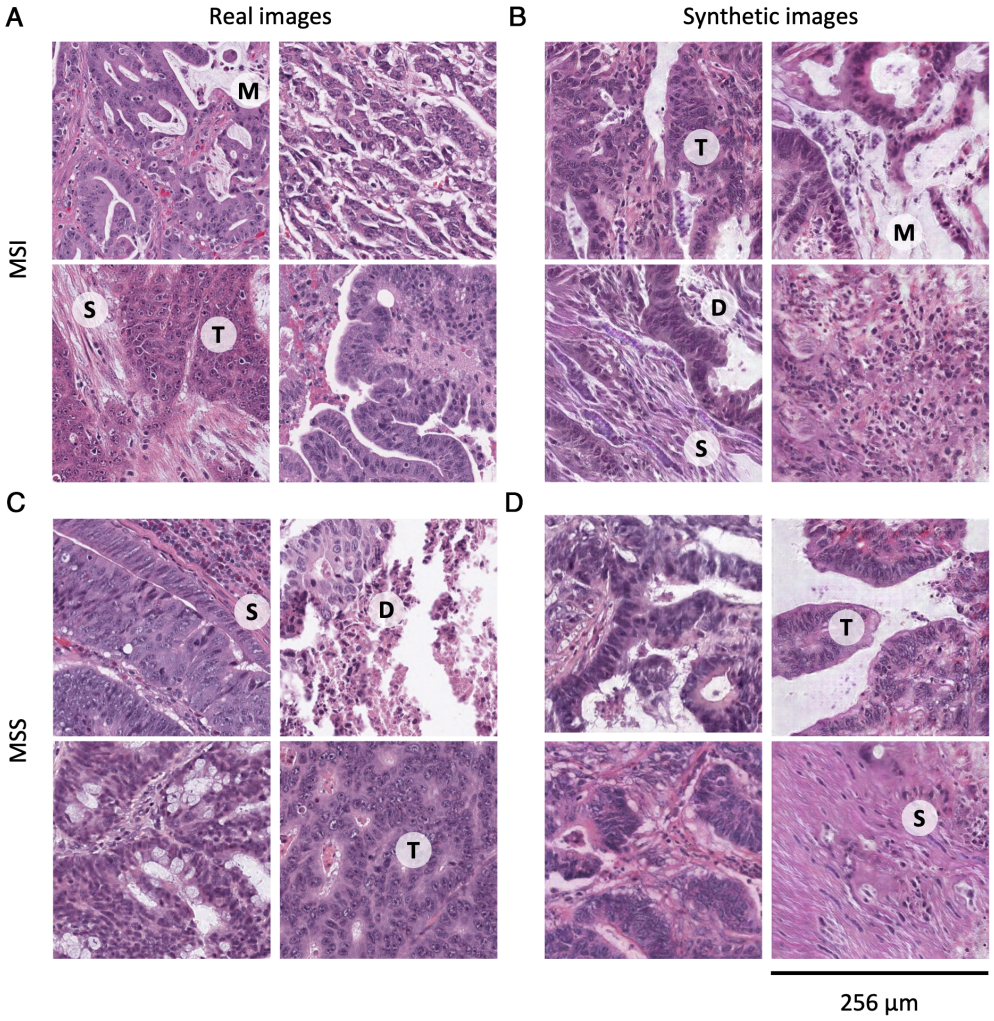

Figure 2. Synthetic histology images are realistic and contain multiple tissue types. These are representative images for MSI and non-MSI, real images and synthetic images. Visually, most of the synthetic images are indistinguishable from real images. Multiple tissue classes are present in real and synthetic images. T, tumor epithelium; S, desmoplastic stroma; M, mucus; D, debris or necrotic cells. (A) Real image patches of microsatellite instable (MSI) tumors. (B) Synthetic image patches of MSI tumors. (C) Real image patches of microsatellite stable (MSS) tumors. (D) Synthetic image patches of MSS tumors.

Synthetic images retain latent genetic information

Having generated synthetic histology images with realistic morphology, we investigated whether these images retain information about genetic subtypes of cancer. We trained a deep learning network for MSI detection on real and synthetic images and evaluated patient-level classification performance on a held-out test set (Figure 5A). When trained on real image tiles in TCGA-CRC-TRAIN (supplementary material, Figure S1), the classifier achieved a patient-level area under the receiver operating curve (AUROC) for MSI detection of 0.742 [0.681, 0.854] (Figure 5B) on the test set. Training the same classifier on synthetic images from the SYNTH-CRC-10K set for the same number of epochs yielded an AUROC of 0.710 [0.614, 0.753] (supplementary material, Table S1). Unlike real images, synthetic images can be generated in arbitrary numbers. Therefore, we trained a classifier for four epochs on a synthetic data set of 75 000 images per class (SYNTH-CRC-75K). When evaluated on the real test set, this classifier achieved a patient-level MSI detection AUROC of 0.743 [0.658, 0.801] (Figure 5C). In the clinically relevant [11] high-sensitivity, low-specificity area, this 'trained on synthetic' classifier was superior to the 'trained on real' classifier, reaching a specificity of greater than 0.6 at a sensitivity of 0.8 (Figure 5C). In addition, we trained additional classifiers on 25 000 image patches (SYNTH-CRC-25K) and on 100 000 image patches (SYNTH-CRC-100K; supplementary material, Table S1), without achieving a pronounced increase of the resulting performance. To quantify the upper limit of a test performance for a network trained on real images, we went beyond the pre-defined hyperparameter set and trained a network on TCGA-CRC-TRAIN for 30 epochs, reaching a higher AUROC of 0.787 [0.694, 0.860] (supplementary material, Table S1). This likely represents the upper limit of the performance that can be reached with the available data.

Improving the performance of genetic classifiers with generated images

Having shown that synthetic histology images can be used as a substitute for real histology images to train

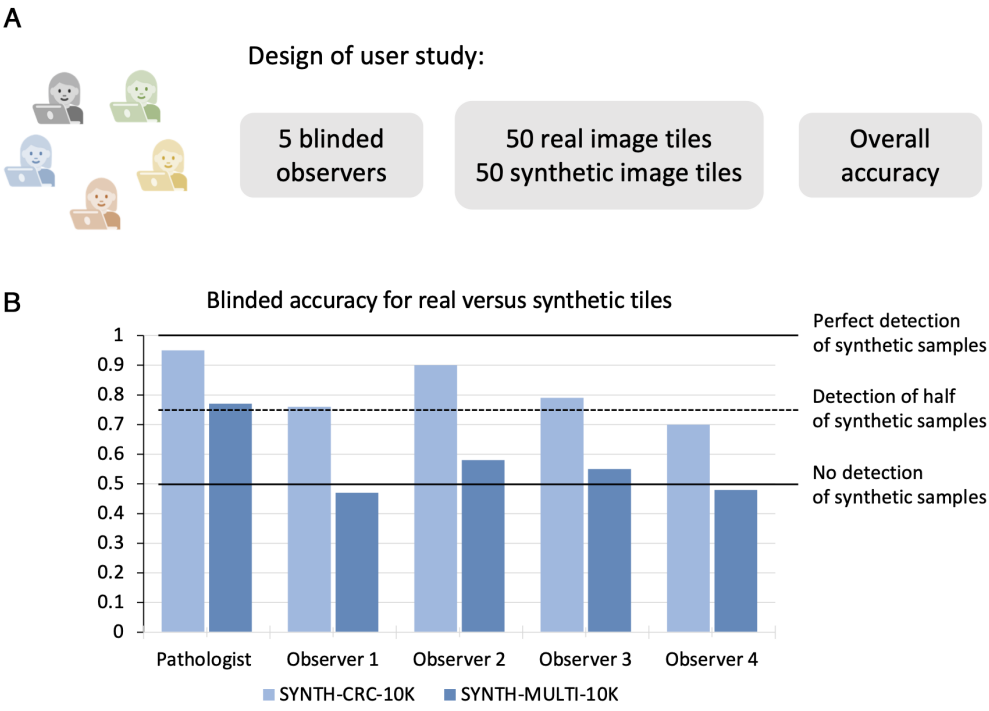

Figure 3. Results of the user study in which participants were asked to distinguish between real and synthetic images. (A) The group consisted of one histopathologist and four non-pathologists who were trained in histology. (B) Accuracy for each observer. The test set consisted of 100 pictures (50 real images and 50 synthetic images, balanced for MSI status). Icons in panel A are re-used under a CC-BY 4.0 license from Twitter Twemoji.

classifiers of molecular status, we asked whether synthetic images could augment real image data sets to improve classification performance. We used synthetic images from the SYNTH-CRC-75K data set to augment the TCGA-CRC-TRAIN data set until we reached 75 000 image patches per class (MIXED-CRC-TRAIN data set). Training on this mixed data set and evaluating on the benchmark test set, an MSI classifier reached a performance of AUROC 0.777 [0.715, 0.821] (Figure 5D). We concluded that synthetic histology image patches can be used to augment real training sets, potentially improving the detection of molecular subtypes in colorectal cancer with a deep learning classifier. Finally, we investigated whether classification performance increases if MSI detectors are trained on more realistic synthetic images in the MULTI-CRC-10K image set. In the user study, images in the MULTI-CRC-10K set were deemed much more realistic than images in the TCGA-SYNTH-10K set (Figure 3B). A likely reason for this superior quality is that these images were generated by a network which was trained on a larger cohort of images, providing a larger degree of biological variability to the CGAN. Indeed, training an MSI detector on this MULTI-SYNTH-10K set markedly improved prediction performance on TCGA-CRC-TEST, achieving an AUROC of 0.757 [0.707, 0.869] (supplementary material, Table S1 and Figure 5E). We conclude that optimizing for realistic appearance of synthetic images may also improve the classification performance for deep learning systems for detecting genetic alterations.

Discussion

Treating patients who have metastatic cancer may require genetic profiling of tumors. Recent studies have shown that this genetic profiling could potentially be supplemented or replaced by deep learning-based analysis of H&E-stained tissue sections [2]. However, large patient cohorts with matched histology, clinicopathological data, and molecular data are scarce, which is a limiting factor for further development of these methods [2]. Generative adversarial networks (GANs) have been proposed as a technology to solve data shortage in digital pathology [20]. However, while previous studies suggest that GANs may be able to generate realistic histology images [17], it is unclear whether these synthetic images retain information about molecular or genetic information.

In the present study, we investigated whether a deep learning classifier of microsatellite instability in colorectal cancer can be trained equally well on synthetic CRC images. Among all genetic biomarkers detectable by deep learning, microsatellite instability is the most extensively validated biomarker [11]. Our results from the present study suggest that when synthesizing histology images with a CGAN, image information able to predict molecular status is preserved. Classifiers trained on synthetic images alone, or hybrid data sets of real and synthetic images, appear to be able to infer molecular information from histology images in a benchmark test set. In terms of performance as measured by AUROC, training on synthetic images yields a classifier

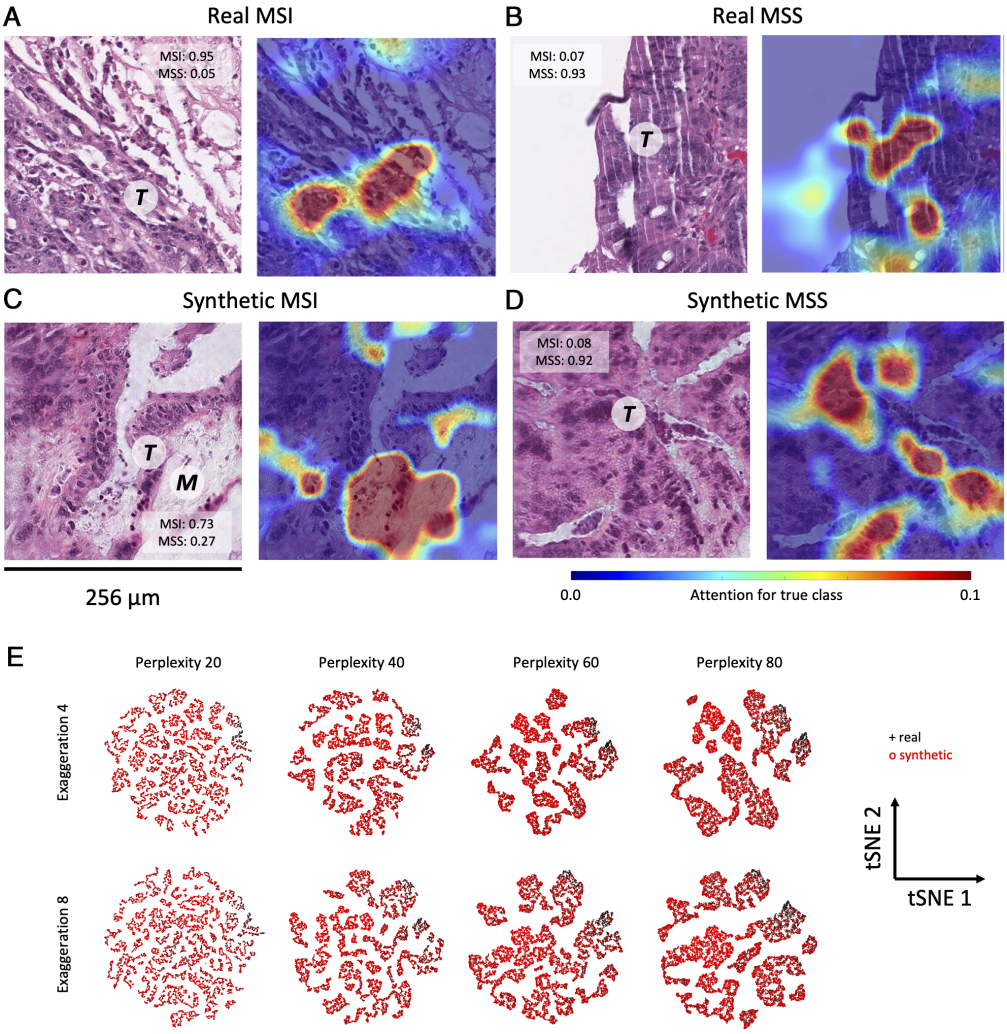

Figure 4. Quantifying similarity between real and synthetic images. (A–D) Representative occlusion maps of real and synthetic images generated by the model trained on real images. The model is sensitive to highly cellular regions in synthetic and real images alike. This suggests that the network processes synthetic images similarly to real images. Furthermore, the strength of activation in both real and synthetic images seems to be similar to one another in relevant regions. (E) t-Stochastic neighbor embedding (tSNE) of visual features in real (black) and synthetic (red) images, as extracted by a network trained on real histology images using a range of clustering parameters (perplexity and exaggeration).

which is on par with a classifier trained on real images. A classifier trained on the hybrid data set was found to be superior to pure real or pure synthetic approaches (Figure 5D). In summary, CGANs could be used as an effective method to generate purely synthetic or hybrid training sets for deep learning classifiers of molecular status in cancer histology. Known histological features associated with MSI in colorectal cancer include poor differentiation, intraepithelial lymphocytosis, and the presence of extracellular mucin, features which were present in synthetic images of MSI tumors (Figure 2B). Furthermore, we saw a performance improvement using the synthetic images as data augmentation in addition to commonly used augmentation methods. However, the benefit of this was only marginal, increasing AUROC from 0.742 to 0.777. Limitations of our study include the observation that although trained non-pathologists could not distinguish between real and synthetic images, a pathologist was able to detect synthetic images with some degree of certainty based on the presence of artifacts unique to the synthetic images (Figure 3B). Additional studies are needed to further improve the performance of CGANs in the context of histopathology. In addition, while this study shows that CGANs can synthesize images containing information about a pre-defined molecular alteration, it is unclear whether other molecular information is encoded in the images.

With the exception of the TCGA study, no large-scale histology data sets with matched molecular or genetic information are publicly available. This is in contrast to non-medical fields of research, where there is an abundance of data sets publicly available for re-use. From an ethical perspective, patient-related raw data should only be publicly shared if the patient explicitly consented or if an ethics board formally waived the need for patient consent. From a legal point of view, handling personal health data requires special caution, especially when the data can be linked back to a particular

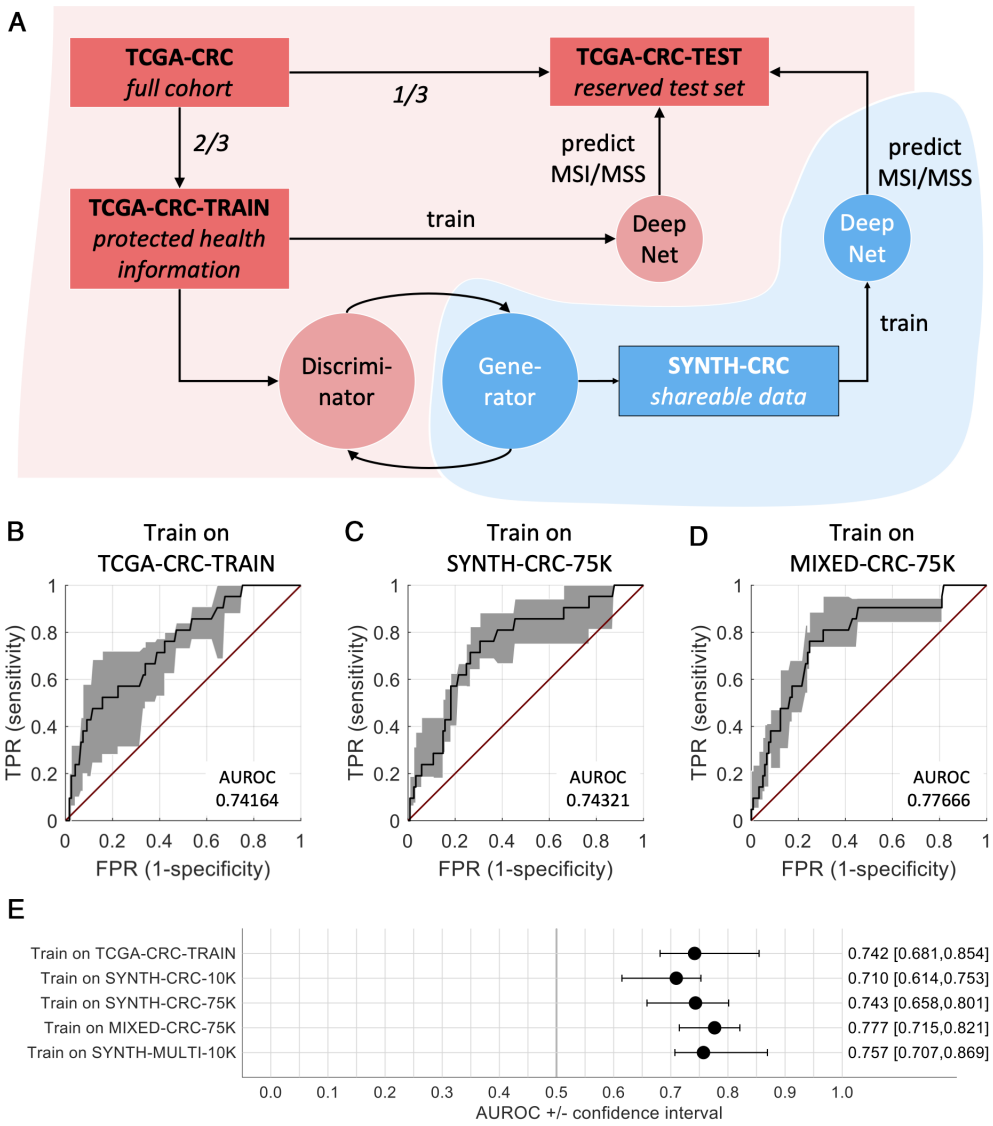

Figure 5. Experimental design and classification results. (A) This study aimed to investigate whether information about molecular status in histological images is preserved by synthesizing training images with a conditional GAN. The red area contains potentially protected health information, whereas the blue area contains shareable data which cannot be linked to any particular patient. (B) Receiver operating curve (ROC) with pointwise confidence bounds for a network trained on TCGA-CRC-TRAIN. All models were evaluated on TCGA-CRC-TEST and show patient-level statistics. Learning hyperparameters were pre-defined. (C) The corresponding ROC of a model trained on SYNTH-CRC-75K shows a similar performance. (D) Augmentation of the real training data set with synthetic images improves performance. (E) Area under the ROC (AUROC) shown for various classifiers trained for four epochs, evaluated on the same test set. Raw data are available in supplementary material, Table S1.

individual. While legal constraints differ between regions, publicly sharing histology slides with matched molecular, genetic or clinical information is problematic in most countries. Generating 'synthetic' data which can be shared publicly could mitigate this problem. While generative models learn patterns from actual patient data, the generated samples cannot be linked to any particular individual. By showing that synthetic data are non-inferior to real data for detecting molecular features in colorectal cancer, our study provides a potential solution to legal and ethical problems associated with sharing patient-related data. In other words, our study provides a new method to anonymize medical image data for subsequent deep learning analysis while retaining subtle information linked to molecular features. Future studies are needed to demonstrate the robustness of this method for other biomarkers. Also, all methods in the present study were limited to the analysis of image tiles extracted from whole-slide histology images. While this approach is currently the state of the art in computational pathology [29], some studies have provided evidence that analysis of larger regions or even whole-slide images is a feasible alternative [30]. Thus, further studies are needed to fine-tune generative networks to a range of clinically relevant applications in computational pathology. Our study provides a proof of principle and a benchmark data set for such approaches.

Acknowledgements

We acknowledge the support of the Rainbow-TMA Consortium, especially the project group: PA van den Brandt, A zur Hausen, HI Grabsch, M van Engeland, LJ Schouten, J Beckervordersandforth (Maastricht University Medical Center, Maastricht, The Netherlands); PHM Peeters, PJ van Diest, HB Bueno de Mesquita (University Medical Center Utrecht, Utrecht, The Netherlands); J van Krieken, I Nagtegaal, B Siebers, B Kiemeney (Radboud University Medical Center, Nijmegen, The Netherlands); FJ van Kemenade, C Steegers, D Boomsma, GA Meijer (VU University Medical Center, Amsterdam, The Netherlands); FJ van Kemenade, B Stricker (Erasmus University Medical Center, Rotterdam, The Netherlands); L Overbeek, A Gijsbers (PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive, Houten, The Netherlands); and Rainbow-TMA collaborating pathologists, among others: A de Bruïne (VieCuri Medical Center, Venlo); JC Beckervordersandforth (Maastricht University Medical Center, Maastricht); J van Krieken, I Nagtegaal (Radboud University Medical Center, Nijmegen); W Timens (University Medical Center Groningen, Groningen); FJ van Kemenade (Erasmus University Medical Center, Rotterdam); MCH Hogenes (Laboratory for Pathology Oost-Nederland, Hengelo); PJ van Diest (University Medical Center Utrecht, Utrecht); RE Kibbelaar (Pathology Friesland, Leeuwarden); AF Hamel (Stichting Samenwerkende Ziekenhuizen Oost-Groningen, Winschoten); ATMG Tiebosch (Martini Hospital, Groningen); C Meijers (Reinier de Graaf Gasthuis/SSDZ, Delft); R Natté (Haga Hospital Leyenburg, The Hague); GA Meijer (VU University Medical Center, Amsterdam); JJTH Roelofs (Academic Medical Center, Amsterdam); RF Hoedemaeker (Pathology Laboratory Pathan, Rotterdam); S Sastrowijoto (Orbis Medical Center, Sittard); M Nap (Atrium Medical Center, Heerlen); HT Shirango (Deventer Hospital, Deventer); H Doornewaard (Gelre Hospital, Apeldoorn); JE Boers (Isala Hospital, Zwolle); JC van der Linden (Jeroen Bosch Hospital, Den Bosch); G Burger (Symbiant Pathology Center, Alkmaar); RW Rouse (Meander Medical Center, Amersfoort); PC de Bruin (St. Antonius Hospital, Nieuwegein); P Drillenburg (Onze Lieve Vrouwe Gasthuis, Amsterdam); C van Krimpen (Kennemer Gasthuis, Haarlem); JF Graadt van Roggen (Diaconessenhuis, Leiden); SAJ Loyson (Bronovo Hospital, The Hague); JD Rupa (Laurentius Hospital, Roermond); H Kliffen (Maasstad Hospital, Rotterdam); HM Hazelbag (Medical Center Haaglanden, The Hague); K Schelfout (Stichting Pathologisch en Cytologisch Laboratorium West-Brabant, Bergen op Zoom); J Stavast (Laboratorium Klinische Pathologie Centraal Brabant, Tilburg); I van Lijnschoten (PAMM Laboratory for Pathology and Medical Microbiology, Eindhoven); and K Duthoi (Amphia Hospital, Breda).

JNK is funded by the Max-Eder-Programme of the German Cancer Aid (Bonn, Germany; grant #70113864) and the START Programme of the Medical Faculty Aachen (Aachen, Germany, grant #691906). PB is supported by the German Research Foundation (DFG; SFB/TRR57, SFB/TRR219, BO3755/3-1, BO3755/9-1, BO3755/13-1), the German Federal Ministries of Education and Research (STOP-FSGS-01GM1901A and DEFEAT PANDEMIcs-01KX2021), and Economic Affairs and Energy (EMPAIA). JNK, TL, and PB are funded by the German Ministry of Health ('Förderung aufgrund eines Beschlusses des Deutschen Bundestages durch die Bundesregierung'; grant DEEP LIVER, #ZMVI1-2520DAT111). PvdB is funded by The Dutch Cancer Society (KWF, Amsterdam; grant number 11044). Data from the NLCS study were provided by the Rainbow-TMA Consortium, which was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch Government (NWO 184.021.007 to PAvdB), and Maastricht University Medical Center, University Medical Center Utrecht, and Radboud University Medical Centre, The Netherlands.

Open Access funding enabled and organized by Projekt DEAL.

Author contributions statement

JK, HIG and JNK conceived the experiments. JK, JNK and MJ carried out experiments and analyzed data. HIG, MK, RDB and PB contributed pathology expert knowledge to experimental design and data interpretation. HIG, PB, TL, CT and PAvdB provided essential resources. All the authors were involved in interpreting the results, writing the paper, and had final approval of the submitted and published versions.

References

1. Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; **68:** 394–424.
2. Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. *Nat Rev Gastroenterol Hepatol* 2020; **17:** 591–592.
3. Kather JN, Halama N, Jaeger D. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. *Semin Cancer Biol* 2018; **52:** 189–197.
4. Stjepanovic N, Moreira L, Carneiro F, *et al*. Hereditary gastrointestinal cancers: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Ann Oncol* 2019; **30:** 1558–1571.
5. Molecular testing strategies for Lynch syndrome in people with colorectal cancer (section 4, Evidence) | Guidance. NICE. [Accessed 30 April 2020]. Available from: https://www.nice.org.uk/guidance/dg27/chapter/4-Evidence.
6. Snowsill T, Coelho H, Huxley N, *et al*. Molecular testing for Lynch syndrome in people with colorectal cancer: systematic reviews and economic evaluation. *Health Technol Assess* 2017; **21:** 1–238.
7. Fu Y, Jung AW, Torné RV, *et al*. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* 2020; **1:** 800–810.
8. Kather JN, Heij LR, Grabsch HI, *et al*. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* 2020; **1:** 789–799.

9. 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:** 1559–1567.
10. Kather JN, Pearson AT, Halama N, *et al*. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25:** 1054–1056.
11. Echle A, Grabsch HI, Quirke P, *et al*. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020; **159:** 1406–1416.
12. Schmauch B, Romagnoni A, Pronier E, *et al*. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun* 2020; **11:** 3877.
13. Echle A, Rindtorff NT, Brinker TJ, *et al*. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021; **124:** 686–696.
14. Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25:** 1301–1309.
15. Calderaro J, Kather JN. Artificial intelligence-based pathology for gastrointestinal and hepatobiliary cancers. *Gut* 2020. https://doi.org/10.1136/gutjnl-2020-322880 [Epub ahead of print].
16. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487:** 330–337.
17. Levine AB, Peng J, Farnell D, *et al*. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *J Pathol* 2020; **252:** 178–188.
18. Murali LK, Lutnick B, Ginley B, *et al*. Generative modeling for renal microanatomy. *Proc SPIE Int Soc Opt Eng* 2020; **11320:** 113200F.
19. Gadermayr M, Gupta L, Appel V, *et al*. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. *IEEE Trans Med Imaging* 2019; **38:** 2293–2302.
20. Safarpoor A, Kalra S, Tizhoosh HR. Generative models in pathology: synthesis of diagnostic quality pathology images. *J Pathol* 2020; **253:** 131–132.
21. Han T, Nebelung S, Haarburger C, *et al*. Breaking medical data sharing boundaries by using synthesized radiographs. *Sci Adv* 2020; **6:** eabb7973.
22. van den Brandt PA. Maastricht Pathology 2018. 11th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 19–22 June 2018. *J Pathol* 2018; **246**(suppl 1)**:** S1–S46.
23. van den Brandt PA, Goldbohm RA, van't Veer P, *et al*. A large-scale prospective cohort study on diet and cancer in The Netherlands. *J Clin Epidemiol* 1990; **43:** 285–295.
24. Macenko M, Niethammer M, Marron JS, *et al*. A method for normalizing histology slides for quantitative analysis. *Proc 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009; 1107–1110; DOI: https://doi.org/10.1109/ISBI.2009.5193250.
25. Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. v1. arXiv.org 2014. Available from: https://arxiv.org/abs/1406.2661. [Accessed 4 February 2021]. Not peer reviewed.
26. Zhang X, Zhou X, Lin M, *et al*. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. v2. arXiv.org 2017. Available from: https://arxiv.org/abs/1707.01083. [Accessed 4 February 2021]. Not peer reviewed.
27. Tellez D, Litjens G, Bándi P, *et al*. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal* 2019; **58:** 101544.
28. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9:** 2579–2605.
29. Coudray N, Tsirigos A. Deep learning links histology, molecular signatures and prognosis in cancer. *Nat Cancer* 2020; **1:** 755–757.
30. Deshpande S, Minhas F, Graham S, *et al*. SAFRON: stitching across the frontier for generating colorectal cancer histology images. v1. arXiv.org 2020. Available from: http://arxiv.org/abs/2008.04526. [Accessed 4 February 2021]. Not peer reviewed.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Visualization of the flow of all samples

Figure S2. Additional activation maps, related to Figure 4

Table S1. Overview of the results

Table S2. Overview of the image sets

Table S3. Step-by-step explanation of experiment #1 in Table S1

Table S4. Hyperparameter sets

Table S5. Step-by-step explanation of experiment #2 in Table S1